Reduction de dimensionalité et analyse des réseaux de voies de signalisation pour les données de transcriptome: Appliquation à la caractérisation des cellules T.

Christophe Bécavin

Dimensionality reduction and pathway network analysis of transcriptome data : Application to T-cell characterization.

THÈSE

présentée et soutenue publiquement le 6 Décembre 2010

pour l'obtention du grade de

Docteur de l'Ecole Normale Supérieure de Paris

(spécialité Interdisciplinaire Frontière du vivant, ED474)

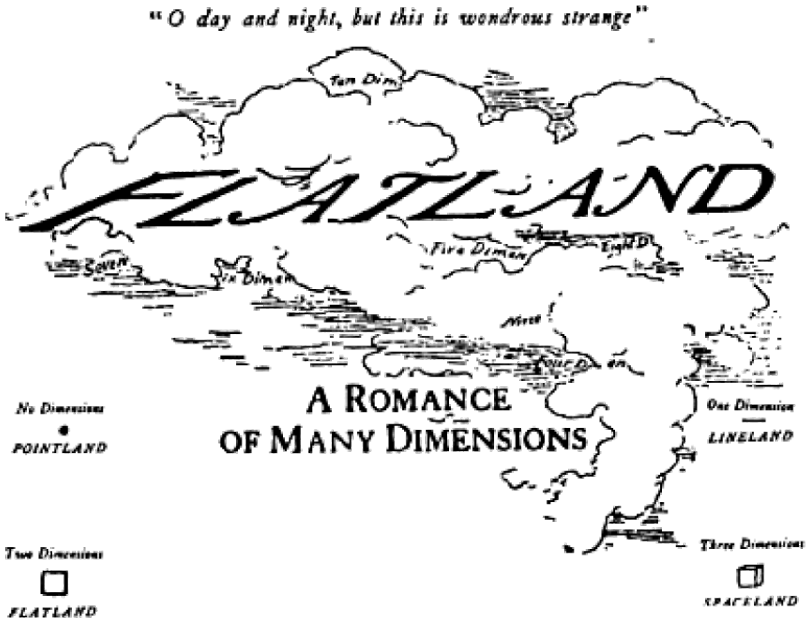
par

Christophe Bécavin

Composition du jury

Rapporteurs :	Alain Arneodo, DR1 CNRS Lars Rogge, CR Institut Pasteur
Examinateurs :	Jean-Marc Victor, DR2 CNRS Sylviane Pied, DR2 CNRS Andrei Zinovyev, IR1 Institut Curie
Directrice de Thèse :	Annick Lesne, CR1 CNRS
Co-Directeur de Thèse :	Arndt Benecke, CR1 CNRS

Mis en page avec la classe thloria.



"Fie, fie, how franticly I square my talk!"

Ewin A. Abbot, Flatland, 1884

Résumé

Dans le contexte de l'étude pan-génomique de données d'expression des gènes (transcriptome), différents outils existent déjà. Parmi eux, les techniques de réduction de dimensionnalité cherchent les formes remarquables et les composants importants du système qui peuvent aider à résumer les données. Au cours de ma thèse, j'ai étudié en profondeur les différentes techniques existantes dans ce domaine. Nous avons ensuite développé notre propre approche basée sur la combinaison de la décomposition en valeurs singulières (Singular Value Decomposition) et le Multidimensional Scaling. Nous avons prouvé son utilité et sa précision.

En plus des outils d'analyse de données spécifiques à l'étude de l'expression des gènes, nous avons développé un logiciel qui permet de correler l'expression des gènes à des réseaux d'interactions protéine-protéine. Et ceci afin de lier l'information sur l'expression des gènes à celle des interactions entre protéines (protéome) qui ont lieu au sein de la cellule.

Tous les outils venant d'être décrits et de nombreux autres ont été utilisés afin d'analyser différent types de données biologiques. La première application a été de corréler l'expression d'auto-anticorps et de cytokines dans le corps humain lors d'une infection au paludisme. Nous avons déterminé des marqueurs spécifiques du paludisme cérébral, permetant à termes de prévenir et détecter plus tôt la maladie. La plus grande analyse que nous avons réalisé visait à définir le profil du transcriptome des cellules T régulatrices (Treg). Ces cellules sont détruites au cours d'une infection par le VIH, une bonne caractérisation moléculaire de celles-ci permettrait par exemple de mieux suivre l'évolution des Treg au cours des traitements pour le SIDA. Parmi les nouveaux marqueurs moléculaires de Treg que nous avons étudié, un nouveau facteur de transcription FOXLF a été découvert, qui pourrait jouer un rôle important dans l'apparition du caractère de "regulation" chez les Treg.

Mots-clés: Génétique, Immunologie, Réduction de dimensionalité, Décomposition en valeur singulière, Echelonnement multidimensionelle, Voie de signalisation, Réseau, Expression des gènes, Transcriptome, Paludisme, VIH, Treg, Puce-ADN.

Abstract

In the context of whole-genome expression (transcriptome) data analysis, different tools already exist today. One class of tools, called dimensionality reduction techniques, seeks for general patterns and important components of the system which can help to summarize the data. During my thesis I extensively studied the different state-of-the-art techniques existing in this field. We then developed our own approach based on the combination of Singular Value Decomposition and Multidimensional Scaling. We proved its usefulness and accuracy.

In addition to gene expression-specific data analysis tools, we developed a software which allows to map different gene expression patterns to protein-protein networks. In order to link the gene expression scale to the protein scale (proteome). Those protein-protein networks are built based on curated ontology-based pathway models.

The tools developed here and many others were used in order to analyze different "omics" data. The first application was on the analysis of experiments measuring autoantibodies and cytokine expression in the human body during Malaria infection. We determined specific markers of Cerebral Malaria, which will help to better detect the disease. The larger analysis we have performed, consisted in defining the transcriptome profile of regulatory T-cell subsets (Treg). These cells are depleted during HIV infection, for this reason a good molecular characterization of the different subsets would help find more accurate markers to, for example, follow their evolution during the treatment with novel drugs to fight AIDS. Among the new molecular markers of Treg we identified, a new transcription factor FOXLF was discovered which may play an important role in the regulation of the "regulatory" function of those cells.

Keywords: Genetic, Immunology, Dimensionality reduction, Singular Value Decomposition, Multidimensional Scaling, Pathway, Network, Gene expression, Transcriptome, Malaria, HIV, Treg, microarray.

Remerciements

Tel Ulysse qui sans ses fidèles compagnons serait toujours sur les plages de Troie, ou Neil Armstrong qui sans l'appuie de la NASA n'aurait fait des petits et des grands pas que dans son jardin, je n'aurais jamais pu réaliser ma thèse sans l'aide de diverses personnes que je tiens à remercier, avant de développer plus loin les diverses projets réalisés.

Tout d'abord je tiens à remercier l'Agence Nationale de Recherche sur le Sida et les hépatites virales et le Génopole Evry, qui tous deux ont financé ma thèse. Je tiens à remercier l'Institut des Hautes Études Scientifiques et son directeur Jean-Pierre Bourguignon qui m'ont permis d'effectuer ma thèse dans un cadre de travail exceptionnel.

Durant toute ma thèse je n'ai jamais eu trop à me soucier des problèmes matériels et financiers, tout ce que dont j'avais besoin m'a été généralement fournis dans les plus brefs délais. J'ai aussi eu la chance d'avoir une totale liberté dans mes choix scientifiques. Je remercie pour ces deux choses essentielles et pour beaucoup d'autres, Annick Lesne et Arndt Benecke mes directeurs de thèse. Au sein du laboratoire, je remercie aussi les "informaticiens" : Nicolas, Felipe, Brice et Guillaume dont les conseils précieux m'ont souvent aidé à éviter des heures de programmation acharnées, et m'ont permis de combler beaucoup de lacunes que j'avais en informatique. Je n'oublierai pas de remercier les "biologistes" Helene et Sebastian, dont les récits épiques de la "vie à la paillasse" et les différentes précisions en biologie qu'ils m'ont donnés, m'ont grandement aidé à toujours rester humble devant ces mécanismes immensément complexes que sont les cellules.

Je remercie tous les membres du jury de thèse qui ont accepté d'évaluer mon travail. Parmi ceux-ci, je remercie Alain Arneodo et Lars Rogge qui ont accepté d'être rapporteur, et qui m'ont aussi accompagné tout au long de ma thèse par leur rôle de tuteur de thèse, me donnant de précieux conseils. Je remercie enfin Jean-Marc Victor, Sylviane Pied, et Andrei Zinovyev, qui ont accepté d'être examinateurs.

Je me dois de remercier encore une fois Lars Rogge qui m'a permis d'effectuer un stage au sein de son laboratoire et m'a donc permis d'avoir une autre idée de la biologie "du quotidien". Comme décrit dans cette thèse, nous avons travailler en collaboration avec Lars et les membres de son groupe que je remercies tous. Parmi eux, je remercie tout particulièrement Sylvie Maiella qui m'a apporté une aide précieuse pour la réalisation des divers analyses et du logiciel décrit au chapitre 3. Nous avons aussi travaillé en collaboration avec Sylviane Pied et les membres de son groupe, je les remercie tous.

Une autre grande aide matérielle et logistique me fut apportée par diverses personnes de l'école doctorale ED474, parmi lesquels François Taddei, Samuel Bottani, Laura Ciriani et Céline Garrigues, je les remercies tout chaleureusement. Cette école doctorale est soutenue par la fondation Bettencourt-Schueller que je tiens aussi à remercier pour l'aide financière qu'elle m'a apporté, me permettant d'aller dans diverses conférences en Europe et en Amérique. Au sein de cette école j'ai eu la chance de faire partie de deux *journal club*, je remercie tous les membres de chacun de ces deux clubs pour l'ambiance convivial dans lequel j'ai appris énormément de choses, surtout sur le plan biologique.

Enfin, *mens sana in corpore sano*, si l'esprit sain a était permis par les diverses personnes remerciées plus haut, le corps sain lui, doit tout à mon entourage qui m'a toujours soutenu et m'a permis de toujours être dans les meilleures conditions. Je remercie donc tout particulièrement ma mère, mon père, mon frère, ma soeur, ma grand-mère, ma belle-soeur, les cousins, et tous les nouveaux venus de ces trois dernières années.

L'entourage cela comprend aussi les amis. Je ne me risquerai pas à les nommer tous, la place disponible et surtout la peur de faire des laissé-pour-compte m'en empêche. Je remercierai quand

même tout spécialement Guillaume, mon deuxième frère, qui me remerciera en retour d'avoir son nom dans au moins une thèse. Je remercie mes deux muses qui chacune à leur tour ont eu l'immense honneur de pouvoir partager au quotidien mes multiples questions existentielles : L'univers pourquoi, comment ? Enfin, dans le désordre, je remercierai tous les autres amis qui m'ont accompagné tout au long de mon voyage scientifique de ces dernières années : Les Physiciens, les Issyens, les Tombe Issiens, les Volleyiens, et les autres.

Table des matières

Introduction

The research field I have chosen for my Ph.D. is Bioinformatics. The work of a Bioinformatician is mainly to create appropriate computational tools for answering a biological question. In order to do so, it is often mandatory to acquire a vast variety of interdisciplinary knowledge. This manuscript reflects the interdisciplinary journey I have made during the past three years, as it first deals with mathematical problems of dimensionality reduction, followed by computational problems of finding a proper algorithm for optimization, continues with pure informatics development, and finishes with applications of the developed methodology to specific biological problematics.

However, one has not to forget that the only goal of every bioinformatics research is to answer to a proper biological question. Even if a large amount of time is passed on the development of a model or a tool, the goal is always at the end to apply it to biological matters. In this thesis, the two main biological topics are : Genetics and Immunology. The former describes how a cell can live and reproduce using the different information encoded in DNA, the latter studies how an organism, such as the human body, can protect itself in an open environment with myriads of potentially lethal pathogens.

Among the different pathogens which can attack the human body : Plasmodium (caused Malaria) and HIV, are of the most dangerous and widespread. Even if treatments already exist to control both infections, no drugs have been found to totally cure them. One way of finding accurate drugs is to characterize at the molecular level the effect of the infection. But, this means also to understand perfectly the different mechanism within the a healthy organism. During my thesis, we try to perform this kind of study. First, we looked at the effect of Plasmodium Falciparum infection on the immunological system of the human body. Second, we characterized a very specific type of immunological cells, so-called regulatory T-cells, which are depleted during HIV infection.

Our analysis were generally performed on datasets containing a large amount of variables. In order to have a "global" view on these kinds of data one can look at clusters, using clustering method, or look at general patterns, using dimensionality reduction techniques. I focus during my thesis on the latter. A dimensionality reduction technique seeks for a representation of studied data in a lower-dimensional space. The reduction of the number of dimension induces a loss of information for the data, the goal of every dimensionality reduction technique is thus to reduce this loss using an optimization process.

In the first chapter of my thesis I will review the state-of-the-art techniques of dimensionality reduction. They can be divided in two groups, depending on whether the correlation within the data are linear or not. Consequently my review will be divided in two parts, focusing in each one on the strategy employed to retrieve the mapping function. The mapping function being the transformation which embeds a studied dataset in a low dimensional space.

It exists a vast number of dimensionality reduction techniques, each of them having their advantages and their drawbacks. For our different analysis we rapidly developed the need of

an accurate, computationally efficient, and easy to use technique of dimensionality reduction. We developed to this end our own approach based on the combination of two state of the art techniques : Singular Value Decomposition (SVD) and Multidimensional Scaling (MDS). We proved the usefulness of this combination in comparison to SVD or MDS only techniques. We also assess the quality of the results given by our approach, comparing them to other strategies of computing Multidimensional Scaling. All these results are described in the chapter 2.

The data I analyzed during my thesis were essentially transcriptome microarray data (measure of gene expression). In order to analyze them, I use not only the dimensionality reduction technique described in chapter 2 but also many other tools, all regrouped in a software developed in the host group and called Ace.map : Software which I have helped to develop, and which I describe in the first appendix.

In order to understand the overall functioning of a cell, studying only the gene expression scale (transcriptome) is not sufficient. One has to link gene expression to protein-protein networks which describe the different mechanisms controlling the fate of a cell. To allow this kind of multi-scale analysis I developed another software, described in chapter 3, which permits to map gene expression information to protein networks. These networks are constructed using information of molecular pathways given by a publicly available databases named Panther (Protein ANalysis THrough Evolutionary Relationships).

Finally, in the last chapter, after a review on the principles of Genetics and Immunology, I describe how I used and combined the different bioinformatics tools described in this thesis, in order to study two types of fundamental biological processes :

- First in collaboration with Sylviane Pied's group, at the Pasteur Institute in Lille, we investigated the fundamental mechanisms of Cerebral Malaria, a lethal disease due to Plasmodium falciparum. The goal of this analysis was to find good autoantibodies and cytokines markers of this disease, in order to better diagnose it ;
- The second analysis was performed with Lars Rogge's group, at the Pasteur institute in Paris. We studied the transcriptome of regulatory T-cell (Treg), searching for molecular markers that can characterize the different subtypes of Treg.

1

Dimensionality reduction

Today's biological datasets can easily contain thousands of instances (number of measures) with 10^5-10^9 variables (number of parameters measured). A prominent example for such datasets are microarray and so-called 'deep-sequencing' data generated in the field of functional genomics [1,2]. The high number of dimensions found, makes global types of analysis difficult for Biologists and Bioinformaticians even for pure computational methods. Before engaging much energy in analyzing your data, reducing the number of dimensions needed to describe your dataset is usually a good thing to do. Methods developed in this spirit are called : Dimensionality reduction techniques. These techniques are in most of the cases used for visualization purposes, as appropriate and faithful visualization of high-dimensional data is often a prerequisite for their analysis ; the human visual cortex being still one of the most powerful tools to detect and conceptualize structure in data [3]. Furthermore, communication of numerical and statistical results is greatly aided by the intuition arising from appropriate representations of data.

Dimensionality reduction is part of Multivariate Analysis which is based on the statistical principle of multivariate statistics, which involves observation and analysis of more than one statistical variable at a time. The correlations between these variables are desired, because the more correlations are observed, the lower number of components are needed to describe the system. The purpose of dimensionality reduction techniques is to take advantage of these correlations to find the best way to summarize the data using fewer components without losing too much information about them. Correlations between variables can be of two forms : linear and non-linear, in this context, methods of dimensionality reduction will also be similarly classified (see figure 1.1).

The mathematical principle of dimensionality reduction is to find a mapping function F from a high dimensional basis B_X with $dim(B_X) = m$ to a new dimensionality reduced basis B_Y with $dim(B_Y) = p$. Here we denote X the data matrix in the basis B_X and Y the transformed data matrix in the basis B_Y. As a first constraint on F reconstitution $p < m$ (ideally $p << m$), and as a second constraint minimization of statistical variables screening the evolution of one or more geometric properties between X to Y.

$$\begin{array}{rl} F: & \mathbb{R}^m \longmapsto \mathbb{R}^p \\ & X \longrightarrow Y \end{array} \tag{1.1}$$

In this chapter, I will give a non-exhaustive review of dimensionality reduction techniques which can be used [1]. I will not go through all the technical and algorithmic details of each

1. This chapter has lead to a review article published in Expert Review of Molecular Diagnostics in January 2011. See appendix C

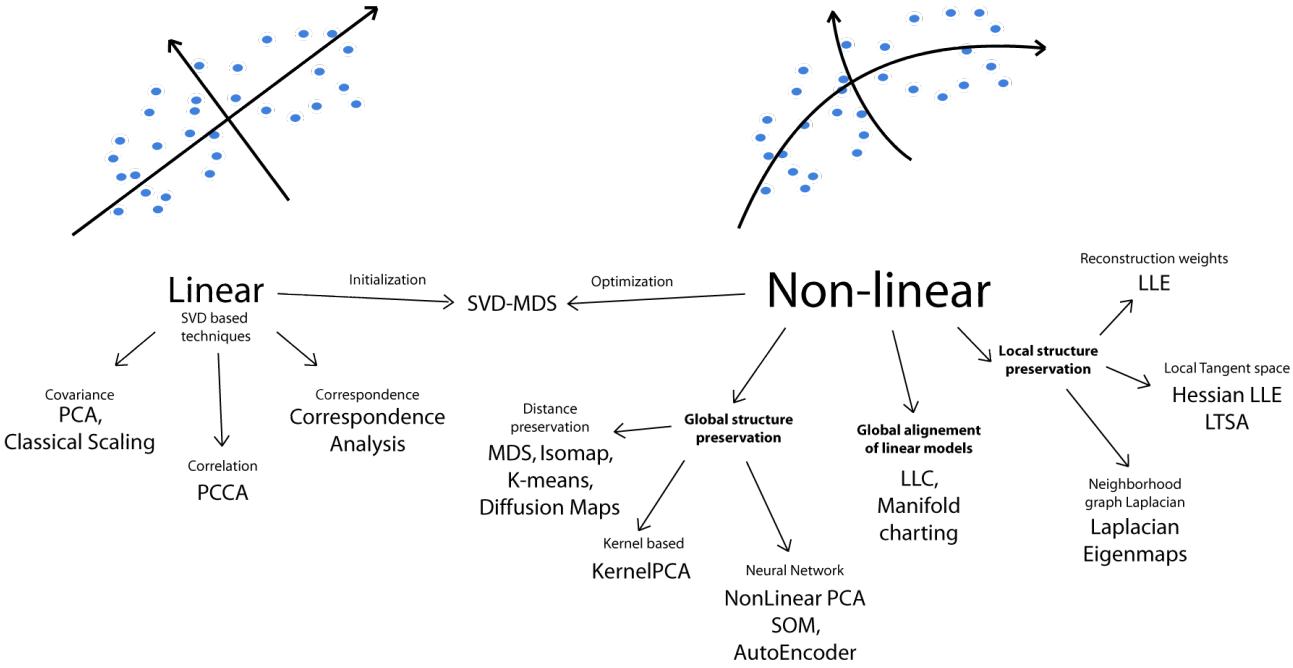


FIGURE 1.1 – *Scheme summarizing the different dimensionality reduction techniques discussed in this chapter.*

technique, avoiding a "handbook" style, but rather prefer a story-telling style where I discuss and compare the different mapping strategies adopted. This chapter will be divided into two parts, one dealing with linear techniques based on linear mapping, the second dealing with non-linear techniques. For two major techniques described below I will show examples of their utilization in the context of "omics" data (see appendix A), in order to demonstrate their usefulness for the type of biological data we are studying in the host group.

Data used here are quantitative and express a frequency. This type of frequency data can be summarized in a matrix X with n rows and m columns (in the following we will say X is $n.m$) and x_{ij} is its value in row i and column j. We denote $\bar{X}_i$ the m components row vector corresponding to row i of the matrix, and $\underline{X}_j$ the n components column vector corresponding to column j of the matrix. A set of vectors $\bar{X}_i$ is then a set of instances (*i.e. experiments*), whereas a set of vectors $\underline{X}_j$ is a set of variables (*e.g. microarray probes*). In all the following chapters I will use this notation for vectors extracted from a matrix X.

1.1 Linear methods

1.1.1 Principal Component Analysis

The first technique of dimensionality reduction invented by Pearson in 1901 was Principal Component Analysis. As he explained in [4] :

In many physical, statistical, and biological investigations it is desirable to represent a system of points in a plan, three, or higher dimensioned space by the "best-fitting" straight line or plane.

So his goal was to find the "best-fitting" principal components which induces the best representation of our data in a low dimensional basis (see fig 1.2-a). In order to find these components,

Pearson claimed that we have to not only consider the mean of statistical variables, but more importantly their standard deviations. The first component will be the one with the highest statistical deviation as shown in fig 1.2-b taken from Pearson's article. Pearson's technique has been reinvented and improved many times [5–7] ; I will present here the modern formulation extracted from [8].

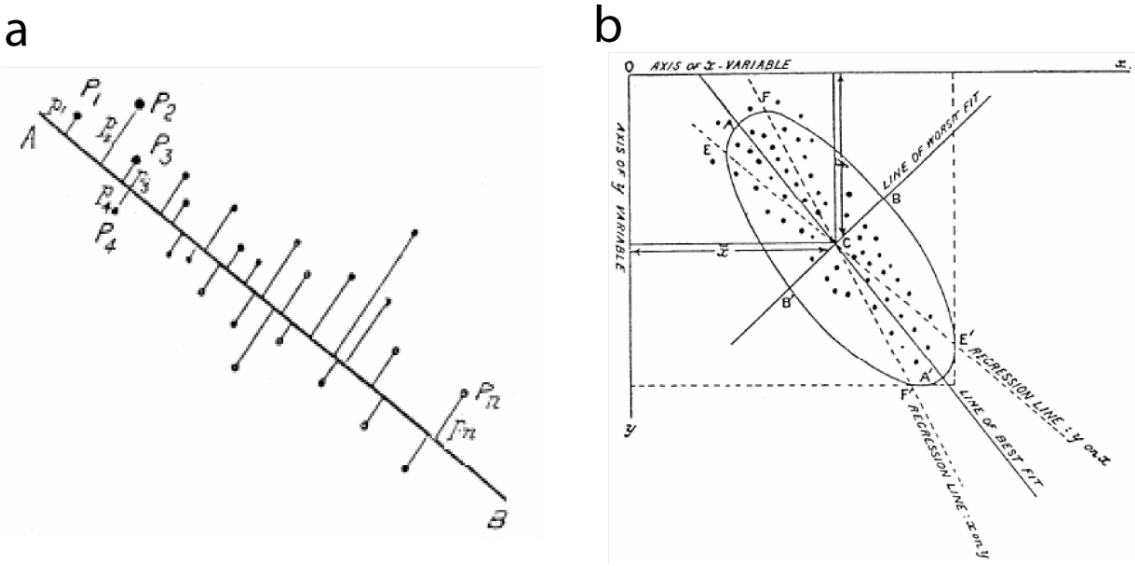


FIGURE 1.2 – *Principle of PCA, Figures extracted from original Pearson article [4]). (a) : Example of "best-fitting" line, with P_i a list of points and p_i their orthogonal distances from the line. The goal of PCA is to minimize the sum of p_i^2. (b) : The results of PCA on a two dimensional cloud of points.*

Principal Component Analysis relies on the concept of variance. The main idea is that if one wants to find the "best-fitting" line, one has to maximize the variance of the first principal components. Variance of all the variables are in the diagonal of the covariance matrix of centered data. To center data one applies the transformation

$$x_{ik} \to x_{ik} - mean(\underline{\mathrm{X}}_k) \tag{1.2}$$

for every element x_{ik} of the data matrix X. Consequently for every $i \in \{0, 1, .., n\}$, $mean(\underline{\mathrm{X}}_i) = 0$. After this transformation, all elements Cv_{ij} of the covariance matrix Cv are calculated as

$$Cv_{ij} = cov(\underline{\mathrm{X}}_i, \underline{\mathrm{X}}_j) = \frac{1}{m}\sum_{k=1}^{n}(\underline{\mathrm{X}}_{ik}.\underline{\mathrm{X}}_{jk}) \tag{1.3}$$

Cv's diagonal contains terms like $Cv_{ii} = \frac{1}{m}\sum_k \underline{\mathrm{X}}_j^2$ which is the variance on the i^{th} variable. As Cv is symmetric positive, for centered data one can find its Eigenvectors solving the following matrix equation

$$Cv = V\Lambda V^{-1} \tag{1.4}$$

where Λ is a diagonal matrix. Eigenvalues are the elements of Λ and are denoted λ_i, Eigenvectors are denoted $\tilde{e}_i$ and $\tilde{e}_i^T\tilde{e}_j = 0$ for $i \neq j$, so the new vectorial space, based on compounds of vectors $\tilde{e}_i$, is orthonormal. In this new basis, matrix Y, is given by,

$$Y = XV \tag{1.5}$$

will represent our data in the principal space, the covariance matrix in this space is equal to Λ. Every basis vector e_i of B_X has been transformed in a principal component $\tilde{e}_{i'}$ given by the matrix V, $\tilde{e}_{i'}$ is just a linear combination of e_i. V is here the matrix representation of the mapping function F, which is consequently linear.

Statistical deviations on each principal component $\tilde{e}_i$ are given by $\sqrt{\lambda_i}$. To obtain a list of principal components organized by their statistical deviation one has to reorganize Λ and U in order to verify $\lambda_1 > \lambda_2 > ... > \lambda_m$. If X has n rows and m columns, Y will have n rows and p columns with $p = rank(Cv) =\leq min(n-1, m)$, I will demonstrate this fact in paragraph 1.1.5.

In the case of data visualization one would like to obtain a matrix Y with only two or three columns. The only way of doing so if $p > 3$ is by simply deleting all the components from p to 3 (or from p to 2, in the case of a reduction to two dimensions). Two linked questions arise : How do we know that this operation will not delete too much valuable information on the data ? And, second, how do we know the effective dimensionality of the system ? The only information one can use for this purpose are the Eigenvalues, as they are directly linked to the statistical deviation. One defines a new parameter called inertia :

$$c_i = \frac{\lambda_i}{\sum_i \lambda_i} \tag{1.6}$$

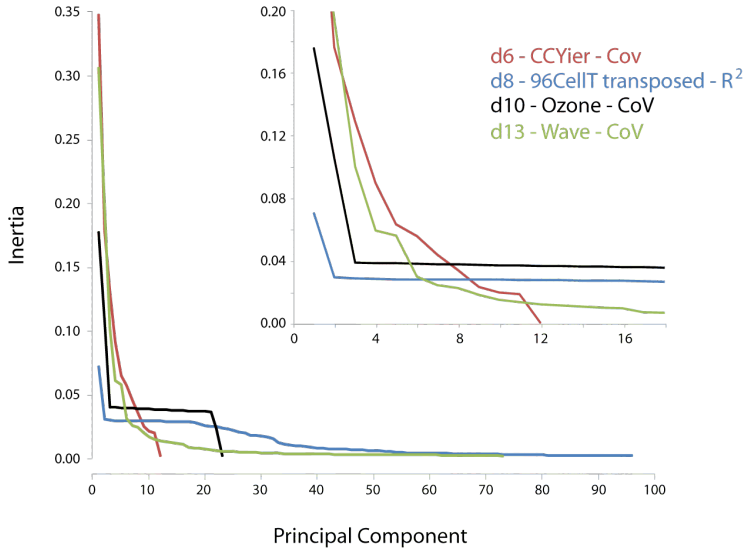


FIGURE 1.3 – Inertia distributions over principal components lead to an appreciation of the structure of the geometric object under study : Scree diagram for four different datasets "d6 — CCYier_T" in *covariance basis*, "d8 — 96Cell_T transposed_T" in *correlation basis*, "d10 — Ozone" in *covariance basis*, and "d13 — Wave" in *covariance basis*. The inset is a zoom of the same graph on the initial eighteen principal components.

Consequently $\sum_i c_i = 1$, so inertia of the i^{th} principal component indicates the proportion of total information enclosed in it. For example if $c_1 = 0.2$, $c_2 = 0.3$ and one decides to represent the data in two dimensions. The final representation will only explain 50% of the total form of the cloud of points. To have the best picture of the form of the cloud of points one can represent inertia versus the index of the corresponding component ; this kind of graphic is called

a scree diagram. Figure 1.3 gives an example on four different datasets, see chapter 2 for more information on these datasets.

It exists many examples of PCA utilization in the literature as it is still the best known technique of dimensionality reduction [9–11]. All those studies are based on the same principle : Using the information on covariance between variables or between groups of variables, one tries to find the first principal components which will best summarize, in terms of standard deviation variation, the general characteristics of the system under study. Thus, instead of studying all variables, one has only to look at the evolution of a few principal variables.

The major development in the field of dimensionality reduction was certainly the switch from the search for lines or planes to more complicated Euclidian geometric structures. Even if the linear condition of PCA is lost, the idea of "best-fitting" always remains prevalent, and form the basis of all the techniques developed during the century following the publication of Pearson's article in 1901. That is why, PCA which is the trivial application of the concept of search for "best-fitting" geometry, is still a major technique which has to be known and understood well if one wants to study more complicated dimensionality reduction techniques.

1.1.2 Principal Component Correlation Analysis

Principal Component Analysis uses the information of covariance, which is related to the variance of each statistical variable used. If two variables have very different variances they may have low covariance whereas they are in fact closely related. The resulting prediction error would not be related to a real property of the underlying data but to errors induced during the measurement, or insufficient quantity of data leading to spurious variances.

In the case in which one postulates that every variable should have the same mean equal to zero and the same variance equal to one (i.e. every variable has the same gaussian distribution), one rather should use the Pearson-correlation measure instead of covariance. Covariance and Pearson-Correlation between two real-valued random variables x and y are defined as :

$$cov(x, y) = E[(x - E[x])(y - E[y])] \tag{1.7}$$

$$corr(x, y) = \frac{cov(x, y)}{\sigma(x)\sigma(y)} \tag{1.8}$$

with $\sigma(x)$, and $\sigma(y)$ the standard deviation of x and y. Consequently for two centered statistical variables $\underline{X}$ and $\underline{Y}$, estimators of covariance and correlation are

$$cov(\underline{X}, \underline{Y}) = \frac{1}{m} \sum_{k=1}^{n} \underline{X}_k . \underline{Y}_k \tag{1.9}$$

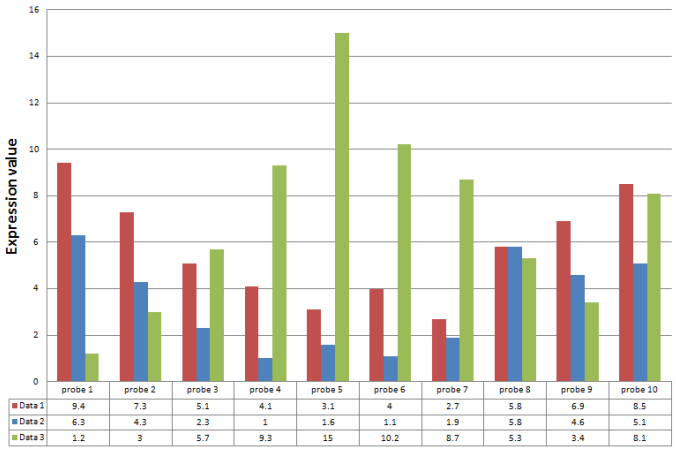
$$corr(\underline{X}, \underline{Y}) = \frac{1}{m} \sum_{k=1}^{n} \frac{\underline{X}_k . \underline{Y}_k}{\sigma(\underline{X})\sigma(\underline{Y})} \tag{1.10}$$

$$\tag{1.11}$$

This means that by a simple substitution $\tilde{\underline{X}} = \frac{\underline{X}}{\sigma(\underline{X})}$ Pearson-correlation between $\underline{X}$ and $\underline{Y}$ becomes the covariance between $\tilde{\underline{X}}$ and $\tilde{\underline{Y}}$.

A Principal Component Correlation Analysis (PCCA) of a matrix X will be a Principal Component Analysis of a matrix $\tilde{X}$ which is related to X using the previous substitution :

$\tilde{x_{ij}} = \frac{x_{ij}}{\sigma(\underline{X}_j)}$. So I will not develop more the principle of this technique because after the centering and normalization operation every step is identical to PCA. Instead I will rather focus on the reason why to prefer PCCA instead of PCA.



	probe 1	probe 2	probe 3	probe 4	probe 5	probe 6	probe 7	probe 8	probe 9	probe 10
Data 1	9.4	7.3	5.1	4.1	3.1	4	2.7	5.8	6.9	8.5
Data 2	6.3	4.3	2.3	1	1.6	1.1	1.9	5.8	4.6	5.1
Data 3	1.2	3	5.7	9.3	15	10.2	8.7	5.3	3.4	8.1

FIGURE 1.4 – *Randoms statistical distributions used : Profile and values used for Data 1, Data 2 and Data 3.*

The use of PCCA becomes relevant if one assumes that every variable should have the same variance, because the substitution operation between X and $\tilde{X}$ has the effect of giving the same variance to all variables. To illustrate this phenomenon, I randomly created three different variables *Data 1*, *Data 2* and *Data 3* which have all ten instances. The profile of *Data 1* and *Data 2* have been designed to be very similar whereas the one of *Data 3* is very different from the other two. Figure 1.4 shows these different profiles.

The first step of PCA and PCCA is to center variables, Figure 1.5-a shows the results of this operation on Data from Figure 1.4. Then if one applies PCCA, one has to perform the substitution by dividing each variable by its standard deviation. Figure 1.5-b demonstrates the effect of this normalization using standard deviation. It is clear that every variable has now the same variance. In the final Figure 1.5-c I show the difference between values of covariance and correlation, with both parameters it is apparent that *Data 3* is different from the other data, but the difference is not the same in both cases. Covariance will be lower between *Data 1* and *Data 3* than between *Data 2* and *Data 3* whereas correlation is the same in both case. This difference is a consequence of the difference between the standard deviation of *Data 1* (σ equal to 2.28) and *Data 2* (σ equal to 2.03).

If one adopts a pure geometric point of view and sees variables as vectors in an affine space : Covariance will be equal to the scalar product between the two variables up to a scalar $1/n$, whereas correlation will be equal to the cosinus of the angle between these two vectors. This becomes obvious by looking at these two equivalent definitions of the scalar product between two vectors :

$$\underline{X}.\underline{Y} = ||\underline{X}||.||\underline{Y}|| \cos(\underline{X}, \underline{Y}) \tag{1.12}$$

$$\underline{X}.\underline{Y} = \sum_{k=1}^{n} \underline{X}_k \underline{Y}_k \tag{1.13}$$

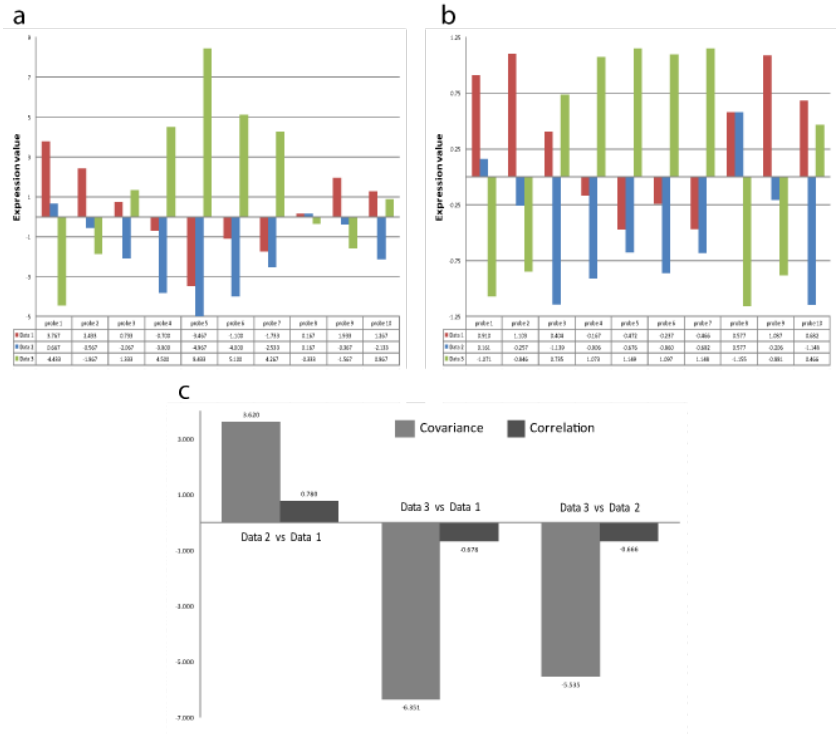


FIGURE 1.5 – *Effect of centering (a) and standard deviation normalization (b) on the different Data. (c) Present differences encounter between covariance and correlation.*

Given the link between magnitude and variance of a centered statistical vector,

$$\sigma(\underline{X}) = \sqrt{\frac{1}{n}\sum_k \underline{X}_k^2} = \frac{1}{\sqrt{n}}||\underline{X}|| \tag{1.14}$$

one can conclude without efforts :

$$\cos(\underline{X},\underline{Y}) = \frac{\sum_{k=1}^{n} \underline{X}_k \underline{Y}_k}{\sqrt{\sum_k \underline{X}_k^2}\sqrt{\sum_k \underline{Y}_k^2}} = \frac{n.cov(\underline{X},\underline{Y})}{\sqrt{n}\sigma(\underline{X}).\sqrt{n}\sigma(\underline{Y})} = corr(\underline{X},\underline{Y}) \tag{1.15}$$

1.1.3 Classical Scaling

The first techniques presented here focuses on the statistical deviation of each variable to find a dimensionality reduced vector space for embedding points. Another way of performing dimensionality reduction is to focus on the distances between data points. Multidimensional Scaling regroups a set of methods which have the same basic goal : Given a matrix of similarities or dissimilarities between points, one will embed points in a dimensionality reduced geometric space. As similarities or dissimilarities, named by the general term of "distances" in the following without link to the mathematical definition of distance, are the only information available, the goal of Multidimensional Scaling techniques will be to conserve this information. The final result consists in a matrix giving positions of our points in a p-dimensional space ; distances between

those points resulting from a transformation of input distances. This transformation can be linear or non linear. The first development of a Multidimensional Scaling algorithm found in the literature is linked to the linear transformation of distances.

Classical scaling (cMDS for classical multidimensional scaling), also named Principal Coordinate Analysis, is a linear Multidimensional Scaling method for finding the original Euclidean coordinates from the derived Euclidean distances. It has been developed by Young et Householder [12] and was popularized as an application for scaling by Torgerson [13]. The modern algorithm presented here is extracted from Cox & Cox's book on Multidimensional Scaling [14].

For a data matrix X with n rows and m columns, the Euclidean distances between two instances $\bar{X}_i$ and $\bar{X}_j$ are given by

$$d^2(\bar{X}_i, \bar{X}_j) = d_{ij}^2 = (\bar{X}_i - \bar{X}_i)(\bar{X}_i - \bar{X}_j)^T = \sum_{i=0}^{m}(x_{ik} - x_{jk})^2 \tag{1.16}$$

One can see that the Euclidean distance is linked to the inner product $\bar{X}_i.\bar{X}_j^T$ by the formula

$$d_{ij}^2 = \bar{X}_i\bar{X}_i^T + \bar{X}_j\bar{X}_j^T - 2\bar{X}_i\bar{X}_j^T \tag{1.17}$$

The link between Euclidean distance matrix and inner product matrix is a key property which permits Classical Scaling. This technique relies on two steps :

- First one recovers the inner product matrix from the Euclidean distance matrix. This operation is called *double centering* ;
- Second one calculates the Euclidean coordinates from the inner products.

To begin one has to move the centroid of the configuration of points to the origin by the following operation

$$x_{ik} \to x_{ik} - mean(\bar{X}_k) \tag{1.18}$$

Using equation 1.17 one can demonstrate that :

$$\sum_i d_{ij}^2 = \sum \bar{X}_i\bar{X}_i^T + n\bar{X}_j\bar{X}_j^T \tag{1.19}$$

$$\frac{1}{n^2}\sum_i\sum_j d_{ij}^2 = \frac{2}{n}\sum_i \bar{X}_i\bar{X}_i^T \tag{1.20}$$

and can conclude

$$b_{ij} = \bar{X}_i\bar{X}_j^T = \frac{1}{2}(d_{ij}^2 - \sum_i d_{ij}^2 - \sum_j d_{ij}^2 + \frac{1}{n^2}\sum_i\sum_j d_{ij}^2) \tag{1.21}$$

If one denotes matrix A given by $a_{ij} = -1/2d_{ij}^2$ one then obtains :

$$b_{ij} = a_{ij} - \frac{1}{n}\sum_i a_{ij} - \frac{1}{n}\sum_j a_{ij} + \frac{1}{n^2}\sum_i\sum_j a_{ij} \tag{1.22}$$

Which can be written in matrix terms as

$$B = HAH \tag{1.23}$$

where $H = I - n^{-1}\underline{\mathbf{I}}\underline{\mathbf{I}}^T$, I is the identity matrix, and $\underline{\mathbf{I}}$ is a n-dimensional column vector filled only with 1.

Once one obtains matrix B, one will search for its Eigenvectors and Eigenvalues

$$B = U\Lambda U^{-1} \tag{1.24}$$

where Λ is a diagonal matrix, and Eigenvalues λ_i are classified by order of their magnitude.

Hence as $B = XX^T$, the new matrix Y is given by :

$$Y = U.\Lambda^{1/2} \tag{1.25}$$

In conclusion, after the operation of *double centering*, the principle of Classical Scaling ressembles Principal Component Analysis. In fact, even if they came from two different fields of Data Analysis, we will see in chapter 1.1.5 that they are related, and that they give identical results. The only difference lies in the fact that cMDS uses the inner product matrix, whereas PCA uses the outer product matrix.

1.1.4 Correspondence analysis

One of the problems encountered when using Euclidean distances is close to the case of covariance used in PCA and which lead to the use of correlation in PCCA. A high value of distance does not always imply a great difference between two points. In the case of frequency, such as microarray, data sometimes large values are due to a problem of measurement. In one experiment which has worked very well a high number of each quantity is produced, whereas in another experiment a production of few quantities of each instance due to poor technical quality occured. In both cases the shape of the statistical profile of each experiment is close to the one which would have been observed in a idealized experiment, but its "height" is very different.

To tackle this problem one can assume that for each instance the number of quantities produced should be equal. In this context the use of Correspondence Analysis will become obvious. As many other data analysis techniques, Correspondence Analysis has been invented and reinvented many times. Its major developments have been provided by Benzécri et al. in France under the name of "Analyse Factorielle des Correspondances" [15]. A complete demonstration by Greenacre can be found in [16].

This technique is especially designed for the study of a two-way contingency table, which is a table containing frequencies of categorical data (see [8] for more information on contingency tables and categorical data). Typical examples of data that can be enclosed in a contingency table are histograms. Such types of tables with n rows and p columns, will correspond to n histograms, and p bins for each histogram. In this particular example Correspondence Analysis will be used to compare all histograms' distribution profiles, by defining a value of distance between them. Correspondence Analysis will represent the set of histograms in a low dimensional space.

For each row of a contingency table given by a matrix X (dim(X)=n.m) one calculates the sum of frequencies on a row, which is called the marginal row frequency, and one also calculates the marginal column frequency and the total sum of frequencies. Then, to calculate distances between points one will use χ^2, which is defined between two points $\bar{X}_i$ and $\bar{X}_j$ as :

$$(\chi^2(\bar{X}_i, \bar{X}_j))^2 = \sum_{k=1}^{m} \frac{1}{f_k} \left(\frac{f_{ik}}{f^i} - \frac{f_{jk}}{f^j} \right)^2 \tag{1.26}$$

where $W = \sum_{i=1}^{n} \sum_{j=1}^{m} x_{ij}$ is the total sum, $f_{ik} = x_{ik}/W$ is the relative frequency, $f^i = \sum_{l=1}^{m} f_{il}$ is the marginal row frequency, and $f_k = \sum_{l=1}^{n} f_{lk}$ is the marginal column frequency.

This distance is derived from the χ^2 statistical test, which evaluates weather or not two distributions have the same profile. One can develop χ^2 distance, in order to write it in terms of x_{ij}

$$(\chi^2(\bar{X}_i, \bar{X}_j))^2 = \sum_k W \left(\frac{x_{ik}}{(\sqrt{\sum_l x_{lk}})(\sum_l x_{il})} - \frac{x_{jk}}{(\sqrt{\sum_l x_{lk}})(\sum_l x_{jl})} \right)^2 \tag{1.27}$$

A straight forward substitution

$$\tilde{x}_{ik} = \frac{x_{ik}\sqrt{W}}{(\sqrt{\sum_l x_{lk}})(\sum_l x_{il})} \tag{1.28}$$

will prove that

$$(\chi^2(\bar{X}_i, \bar{X}_j))^2 = \sum_{k=1}^{m} (\tilde{x}_{ik} - \tilde{x}_{jk})^2 = (d(\tilde{\bar{X}}_i, \tilde{\bar{X}}_j))^2 \tag{1.29}$$

So χ^2 distance is in fact an Euclidean distance if the data matrix is properly rescaled.

The rescaling matrix is given as

$$\tilde{X} = \left(\frac{X_{ij}\sqrt{W}}{(\sqrt{\sum_k x_{kj}})(\sum_k x_{ik})} \right) \tag{1.30}$$

$$\tilde{X} = \sqrt{W} \begin{pmatrix} 1/\sum_k x_{1k} & 0 & ... & 0 \\ 0 & 1/\sum_k x_{2k} & ... & 0 \\ ... & ... & ... & ... \\ 0 & 0 & ... & 1/\sum_k x_{nk} \end{pmatrix} X \begin{pmatrix} 1/\sqrt{\sum_k x_{k1}} & 0 & ... & 0 \\ 0 & 1/\sqrt{\sum_k x_{k2}} & ... & 0 \\ ... & ... & ... & ... \\ 0 & 0 & ... & 1/\sqrt{\sum_k x_{km}} \end{pmatrix} \tag{1.31}$$

The effect of this rescaling is to give approximatively the same marginal row frequency to each instance. To illustrate this, I show in table 1.1, the contingency table resulting from the random data created in section 1.1.2. Then in table 1.2 I show the effects of rescaling on the data, and on the marginal row frequency.

Instances	Pb1	Pb2	Pb3	Pb4	Pb5	Pb6	Pb7	Pb8	Pb9	Pb10	Sum
Data 1	9.4	7.3	5.1	4.1	3.1	4	2.7	5.8	6.9	8.5	56.9
Data 2	6.3	4.3	2.3	1	1.6	1.1	1.9	5.8	4.6	5.1	34
Data 3	1.2	3	5.7	9.3	15	10.2	8.7	5.3	3.4	8.1	69.9
Sum	16.9	14.6	13.1	14.4	19.7	15.3	13.3	16.9	14.9	21.7	160.8

TABLE 1.1 – *Contingency table of the random data*

Instances	Pb1	Pb2	Pb3	Pb4	Pb5	Pb6	Pb7	Pb8	Pb9	Pb10	Sum
Data 1	0.510	0.426	0.314	0.241	0.156	0.228	0.165	0.314	0.398	0.407	3.158
Data 2	0.572	0.420	0.237	0.098	0.134	0.105	0.194	0.526	0.444	0.408	3.139
Data 3	0.053	0.142	0.286	0.445	0.613	0.473	0.433	0.234	0.160	0.315	3.154
Sum	1.134	0.988	0.837	0.784	0.903	0.806	0.792	1.075	1.003	1.130	9.451

TABLE 1.2 – *Contingency table of the random data after rescaling using equation 1.30*

I graphically demonstrate the effect of the rescaling on the random data in Figure 1.6. Before rescaling, instances contain different quantities of data (Figure 1.6-a), after rescaling the quantity of data in each distribution appears to be the same (Figure 1.6-b). I also plot Euclidean distance and χ^2 distance in Figure 1.6-c. The comparison of the two results show that χ^2 better discriminates the data from each other in our example , as $d_{31}/d_{21} = 2.307$, $d_{32}/d_{21} = 2.580$ and $d_{32}/d_{31} = 1.118$ are less than $\chi^2_{31}/\chi^2_{21} = 2.820$, $\chi^2_{32}/\chi^2_{21} = 3.374$ and $\chi^2_{32}/\chi^2_{31} = 1.196$.

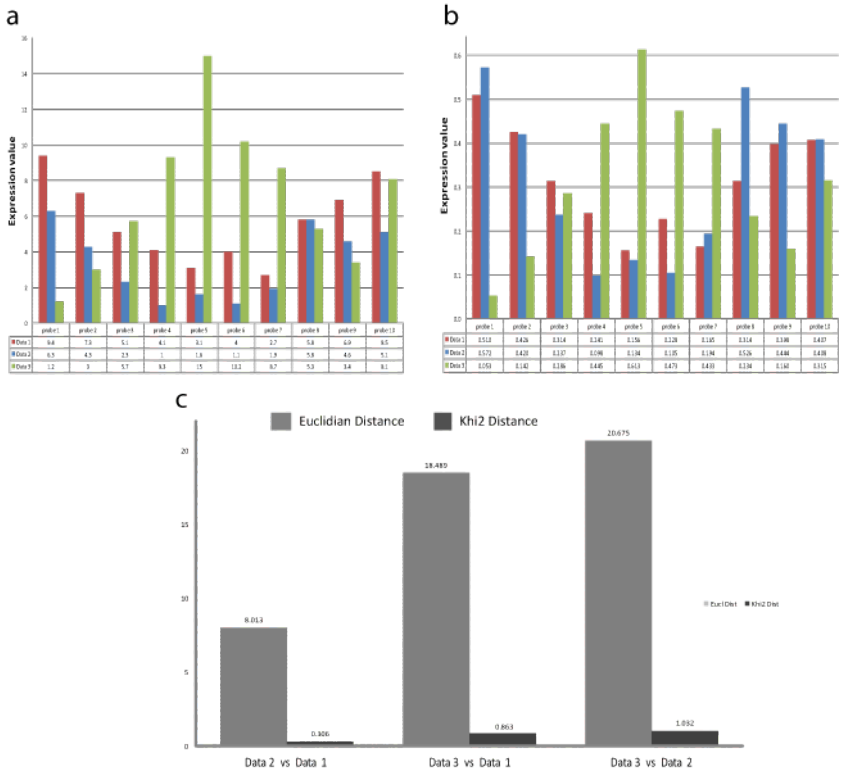


FIGURE 1.6 – *Effect of rescaling on different random Data. (a) Data before rescaling (b) Data after rescaling. (c) Present differences encounter between Euclidean distance and χ^2 distance.*

After rescaling the χ^2 distance becomes an Euclidean distance, the next step is to embed the data-points in a dimensionality reduced Euclidean space by performing Classical Scaling on these Euclidean distances. Thus, I will not explain in more details the algorithm of Correspondence Analysis because it is the same as the one I demonstrated in section 1.1.3.

One advantage of a contingency table is that it gives similar internal structures to instances and variables. In consequence, Correspondence Analysis with proper rescaling (see [17] allows to make a biplot or other types of representation which will include both instances and variables in the same graph.

Correspondence Analysis is specifically adapted to the study of contingency tables, which are tables of frequency data, making it particularly interesting for "omics" data which are often counts of the presence of particulars types of objects / entities. In this kind of study, one assumes that every instance should have the same number of cumulated "counts". In the case of for instance comparative genome hybridization (CGH) microarrays [18] will be equivalent to assuming the same number of genomic fragments for almost every locus. Despite its interest for the analysis of "omics" data, Correspondence Analysis is in this context not frequently utilized.

1.1.5 Singular Value Decomposition (SVD)

We proved that all the techniques explained in the previous sections are based on two distinct principles (i) search of Eigenvalues and Eigenvectors of the outer-product matrix, or (ii) the search of the Eigenvalues and Eigenvectors of inner-product matrix. With the use of a mathematical tool called Singular Value Decomposition I will demonstrate that these two different processes are in fact closely related.

It is known [19] that every rectangular matrix can be decomposed using its singular values

$$X = USV^t \tag{1.32}$$

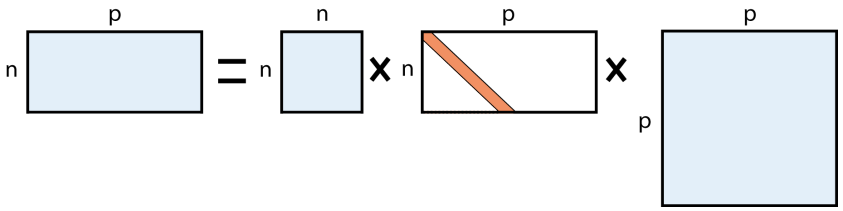


FIGURE 1.7 – *Scheme of the singular value decomposition $X = USV^t$. Rectangular boxes are matrices, and the red box is the diagonal of S containing singular values, the other elements of S are equal to zero.*

where U (matrix containing left singular vectors) and V (matrix containing right singular vectors) are both square orthogonal matrices ($UU^t = U^tU = Id$ and $VV^t = V^tV = Id$), and S is a rectangular matrix containing the singular values (s_i) which are positive. If n is the number of rows and p the number of columns of X, X is $n.p$, U is $n.n$, S is $n.p$, V is $p.p$ (see Figure 1.7). In the case of $n = p$, S is a square diagonal matrix, if $n \neq p$, $S_{ii} = s_i$ and $S_{ij} = 0$. This decomposition is very useful in dimensionality reduction because of its link with the Eigenvalue decomposition of the inner-product (XX^t) and outer-product (X^tX) of X. If $X = USV^t$ one obtains

$$XX^t = USV^t(VS^tU^t) = USS^tU^t \tag{1.33}$$

$$X^tX = (VS^tU^t)USV^t = VS^tSV^t \tag{1.34}$$

SS^t and S^tS are two diagonal square matrices, they do not have the same size but they have the same number of non-null values on the diagonal [19]. So SVD allows to demonstrate that the inner and the outer product have the same Eigenvalues λ_i, with $\lambda_i = s_i^2$. It also gives a matrix link between them, and it is very useful as all the classical techniques of dimensionality reduction are performed using one of these products.

In the previous sections, I demonstrated that in the case of a dimensionality reduction method based on the calculation of Euclidean distance, one uses this type of product $\bar{X}_i.\bar{X}_j$, therefore one uses the inner product matrix as in Classical Scaling (see 1.1.3) and Correspondence Analysis (see 1.1.4). In the case of a method based on statistical measures linked to covariance, one uses the product written in terms of $\underline{X}_i.\underline{X}_j$, hence one uses the outer product matrix as in PCA (see 1.1.1) and PCCA (see 1.1.2). The most important thing is that for the four well known data analysis techniques I described in the previous sections, after appropriate renormalization of the data, the problem can be reduced to SVD. That is why, in some journal articles, one uses the term SVD instead of PCA. These links are demonstrated in further detail in the next section.

The simplest way to find SVD, is to search first for the Eigenvalues and the Eigenvectors of the inner and outer products. As finding the Eigenvalues of a matrix X with n rows and p columns, is hard to perform for objects with a high number of variables, this step is only feasible if either n or p are small (typically inferior to 1000). If both, the number of rows and the number of columns were high, it is impossible to perform SVD with linear techniques, and one is obliged to use iterative Singular Value Decomposition techniques as shown in [19]. It has also been shown that Eigenvalue search methods of SVD resolution is not numerically stable [20], especially when Eigenvalues are close to zero, an heuristic method is more reliable. I never encountered this kind of instability during all of my work, which is why I only used Eigenvalue search methods whenever I used SVD.

However, if either n or p is small, which is often the case in biological datasets for evident experimental costs reasons, one will be able to perform SVD, again because of the close link between inner and outer products :

- If $n < p$ one will search the Eigenvalues of the inner product matrix which is n square dimensional (this step consists in fact of performing a Classical Scaling)

$$XX^t = U\Lambda U^t = USS^tU^t \tag{1.35}$$

  as one has $X = USV^t$, one determines V with the equation : $V = X^tUS^{t^{-1}}$;
- If $p < n$ one will search the Eigenvalues of the outer product matrix which is p square dimensional (this step consist in fact of performing a PCA)

$$X^tX = V\Lambda V^t = VS^tSV^t \tag{1.36}$$

  as one has $X = USV^t$, one determines U with the equation : $U = XVS^{-1}$.

Then U,S, and V, can be reorganized in order to have $s_1 > s_2 > ... > s_r$, with r being the rank of S. Generally before performing SVD the matrix X is centered, which means that the cloud of points is placed in the center of our axes, $(x_j^i)' = x_j^i - mean(X_j)$. Practically, this translates to all the means of the columns of the matrix being equal to zero after transformation. The inner product matrix can then be demonstrated to be a solution of the equation $XX^t.1_n = 0$ with $1 = (1, 1, ..., 1)^t$. As zero is one of the Eigenvalues of the inner-product : $rank(XX^t) \leq n - 1$; and in addition $rank(X^tX) \leq p$ holds, it follows

$$rank(X) = rank(S) \leq min(n - 1, p) \tag{1.37}$$

Finally the new matrix Y will then be given by

$$Y = XV = US \tag{1.38}$$

V is orthogonal so this transformation has conserved the distance information between points without any deformation.

Note also that missing values in data can be imputed using SVD, and please refer to appendix B for more information on the missing value imputation by this technique. If the number of missing values is relatively low, the Eckart Young theorem [21], which is the most commonly used theorem for matrix approximation by another matrix in the least-squares sense, assures that the result of the SVD will change only in the value of the last singular values. Hence, for a rapid imputation, the row average method [22], can be used which is sufficiently precise in most cases. In this case one replaces each missing value by the mean of the statistical sample the value is supposed to be part of.

The outer-product is the covariance matrix on all components up to a scalar (see 1.1.1), and $Y^tY = V^tXXV = S^tUUS = S^tS$. S^tS is an ordered square diagonal matrix, with $s_1 > s_2 > ... > s_r$, r being the rank of S. The covariance matrix diagonal is formed by the variance of each component, thus one can conclude that each singular value s_i is the standard deviation of the *i-th* component. The new basis one has obtained is also an orthonormal basis because the covariance matrix is diagonal, and every component is ordered by order of its relative contribution to the general pattern. An inertia parameter is defined to evaluate this relative contribution. For one component, inertia is $c_i = s_i / \sum_i s_i$, so $\sum_i c_i = 1$. The inertia vector is thus very important for us as it carries geometric information on the form of the cloud of points.

In conclusion, singular value decomposition provides three major types of information :

(i) A new data matrix Y, which represents the data points in a new orthonormal basis with a minimum number of components, and where distances between the instances are preserved.

(ii) Inertia parameters indicate the standard deviation and relative contribution of the cloud of points on each principal component.

(iii) The matrix V in which one obtains the contribution to each principal component of all the variables in the former basis.

Y is a representation of the data in a r-dimensional orthonormal basis (with $r = rank(S)$), this representation is well chosen as distances are conserved in the cloud of points (orthogonal transformation). Also, it has been demonstrated that Principal Component Analysis results are a very good choice for the initial state in order to perform K-means clustering [23]. In the new representation given by SVD cluster structure of the data will then naturally appear, and it thus provides a natural interpretation of clusters.

Alter *et al.* have made extensive use of SVD [24–26] for the study of cell-cycle transcriptome data. For example in [24] they applied SVD to a selection of 5980 genes expressions values measured by Spellman *et al.* [27] and classified them by their cell-cycle regulation of the budding yeast *Saccharomyces cerevisiae*. The expression of these genes has been measured at 14 different times (at $\sim 30min$ intervals) over approximately one cell cycle period ($\sim 390min$), in a yeast culture synchronized by elutriation, and relative to a reference mRNA from an asynchronous yeast culture. They found a specific cosinus and sinus oscillatory behavior in the first and second component of their analysis (see Figure 1.8). This example demonstrates how dimensionality reduction effectively helps to reveal global properties in data.

1.1.6 A general view on linear methods : From Factor Analysis to Independent Component Analysis

We saw in the previous sections that when one wants to know what is the most relevant orthonormal basis for summarizing data, one can use the results given by Singular Value Decomposition. The purpose of this method is to find the most convenient linear transformation of the original data matrix X such that :

$$Y = XV \tag{1.39}$$

This transformation is convenient as it gives an orthonormal basis, and the first principal variable $\underline{Y}_1$ has the maximum variance. The other principal variables then having continuously decreasing values of variances.

Techniques presented above are part of a class of methods which have as purpose the search of factors summarizing our data optimally. Factors come from linear combinations of input variables in order to maximize a statistical variable. In the case of PCA, it will maximize the value of variance. Another technique of this type is simply called Factor Analysis : it is both a

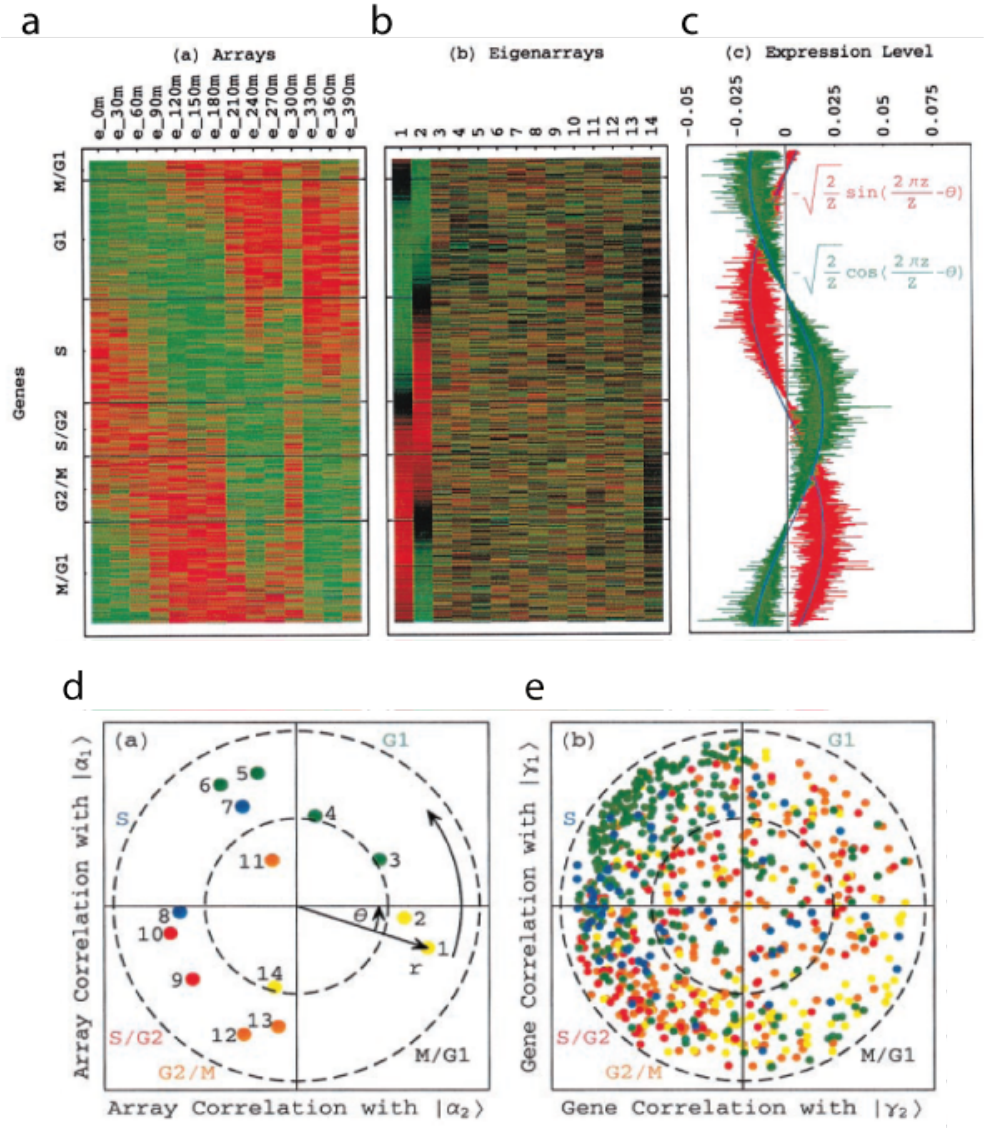


FIGURE 1.8 – *Results of SVD analysis on* 5980 *cell cycle genes measured at* 14 *different moments during a complete cell cycle of Saccharomyces cerevisiae [24]. (a) A heatmap of the expression of these genes over all times (arrays), (b) the same heatmap with the Eigenarrays, which are the principal components found with SVD. One can see the oscillatory behavior found only in the two first principal components,* 90% *of the total inertia is found in these two principal components. (c) Screen of the variation of expression of these two principal components during the cell cycle. The sinusoidal behavior of the first component (in red) and the cosinus-like behavior of the second principal component (in blue) is shown. (d) A 2-dimensional representation of the arrays according to their correlation to the first eigenarray* $|\alpha_1>$ *and the second* $|\alpha_2>$*. (e) A 2-dimensional representation of the* 5980 *genes according to their correlation to the first eigengene* $|\gamma_1>$ *and the second* $|\gamma_2>$*. The colorization of these two last representations is function of the corresponding cell-cycle phase of each array and gene. One can see again how the oscillatory behavior appears using the SVD.*

confirmatory technique and an exploratory one. One postulates that data may be summarized using few numbers of factors. In the case of confirmatory analysis the number of factors will be predetermined by the experiment, whereas in the case of exploratory analysis the number of factors will be determined by the data.

For a centered data matrix X which is $n.m$, one postulates that data can be described using p ($p < m$) factors $\bar{F}_k$

$$\bar{X}_i = l_{i1}\bar{F}_1 + l_{i2}\bar{F}_2 + ... + l_{ip}\bar{F}_p + \bar{\Psi}_i \tag{1.40}$$

Where l_{ij} are matrix terms of L which is the loading matrix (i.e. instances in the "factor basis"), and $\bar{\Psi}_i = (\psi_{i1}, \psi_{i2}, ..., \psi_{ip})$ is a vector of error terms with zero mean and finite variance, different for every i. The decomposition of each instance leads to the following matrix equation :

$$X = LF + \Psi \tag{1.41}$$

where F is $p.n$ and is the matrix of the linear transformation containing factor vectors, L being $p.n$ is the data embedded in the factor space, and Ψ is the matrix of noise being $m.n$. One also postulates that F and Ψ are independent, $E(F) = 0$ and $Cov(F) = Id$ the identity matrix. The major difference between Factor Analysis and PCA lies in the noise matrix Ψ. Because in PCA one only postulates that all the variance comes from the data itself, in Factor analysis one is supposed to have an idea on the underlying data structure including the noise which depends on the distribution of the random variables one expects to observe. The noise matrix has to be postulated by the user who is performing Factor Analysis (see Figure 1.9).

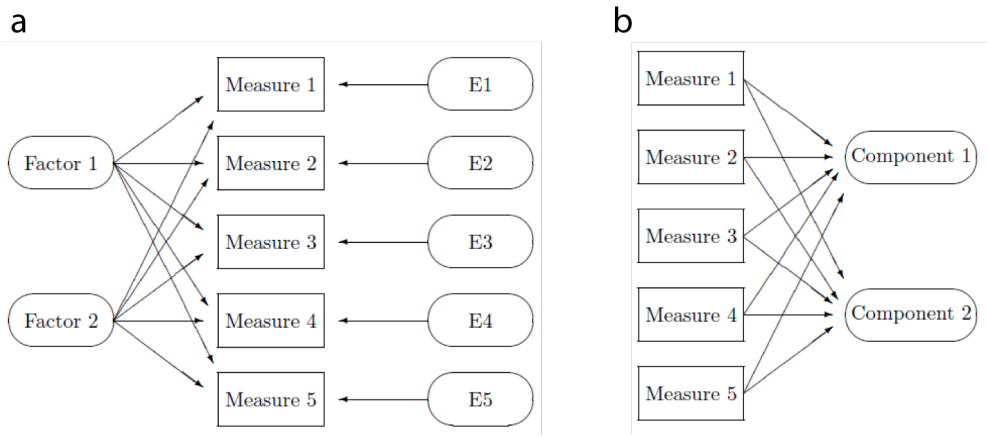


FIGURE 1.9 – *(a) Principle of Factor Analysis named Common Factor Model. Measures are assumed to originate from a set of factors and noise components (E in the Figure). (b) Principle of Principal Component Analysis. Here it is the components which originate from measures. Figures extracted from [28].*

The different spirit of this different way of analyzing your data is well summarized in [28] :

"When an investigator has a set of hypotheses that form the conceptual basis for her/his factor analysis, the investigator performs a confirmatory, or hypothesis testing, factor analysis. In contrast, when there are no guiding hypotheses, when the question is simply what are the underlying factors the investigator conducts an exploratory factor analysis. The factors in factor analysis are conceptualized as "real world" entities such as depression, anxiety, and disturbed thought. This is in contrast to principal components analysis (PCA), where the components are simply geometrical abstractions that may not map easily onto real world phenomena.

In practice $\Sigma = \Psi^t\Psi$, the covariance matrix of Ψ, is known. For performing Factor Analysis in this case one possesses two techniques. Either one searches the best factors having greater variances by maximum likelihood search, or one performs PCA using the modified covariance matrix $1/nX^tX - \Sigma$. If the covariance matrix of noise is not known one has to first estimate it with appropriate techniques. A general survey on Factor Analysis can be found here [29].

The common principle of all the techniques explained above is that they are all based on maximization of the variance. This principle becomes relevant when one deals with gaussian data, because when using centered variables, a gaussian distribution will be fully determined by its variance. Every linear method described previously is indeed well suited when normality of data is a reasonable approximation. In other cases, the concept of variance, and the concept of correlation which is derived from the first, become less relevant. That is why in the past years a novel method called Independent Component Analysis (ICA) has been developed. This technique is based on the principle of PCA, substituting uncorrelatedness by independence. One will now search for independent factors, instead of orthogonal (i.e uncorrelated) factors.

One has two random variables X and Y, they are said independent if one satisfies :

$$p(X \cap Y) = p(X)p(Y) \tag{1.42}$$

One can easily demonstrate [30] that for every measurable functions h_1 and h_2, one will have

$$E(h_1(X) \cap h_2(Y)) = \iint h_1(X)h_2(Y)p(X \cap Y)dXdY \tag{1.43}$$

$$= \iint h_1(X)h_2(Y)p(X)p(Y)dXdY = E(h_1(X))E(h_2(Y)) \tag{1.44}$$

Two random variables X and Y are independent if they verify this equation

$$E(h_1(X)h_2(Y)) = E(h_1(X))E(h_2(Y)) \tag{1.45}$$

for all measurable functions h_1 and h_2.

The first question one can ask when using Independent Component Analysis (ICA) is the difference between the concept of uncorrelatedness and independence between two random variables X and Y. One knows that $cov(X,Y) = E((X - E(X))(Y - E(Y))) = E(XY) - E(X)E(Y)$ It follows that for two uncorrelated random variables X and Y, one obtains :

$$cov(X,Y) = 0 \Leftrightarrow E(XY) = E(X)E(Y) \tag{1.46}$$

One can immediately conclude that independence implies uncorrelatedness, the opposite not being true. To illustrate this latter fact, one can use the following example extracted from [31] to explain how data can be uncorrelated but dependent. The two random variables $\bar{X} = (0, 0, 1, -1)$ and $\bar{Y} = (1, -1, 0, 0)$ are uncorrelated because $cov(\bar{X}.\bar{Y}) = 0$ but they are not independent as $E(\bar{X}^2\bar{X}^2) = 0$ and $E(\bar{X}^2)E(\bar{Y}^2) = 1/4$ are different so the condition 1.45 is violated.

The need for techniques which will discriminate instances based on other criteria than the measure of variance and correlation is not new. One of the first techniques invented in this spirit is called Projection Pursuit [32,33]. The goal here is to find the best projection that will unravel interesting structures in the data. Thus, one does not search for the projection which maximizes variance, but projections in the direction which has an interesting distribution. Figure 1.10 demonstrates the difference of results obtained between Projection Pursuit and variance based techniques. The example shows a cloud of points given by a bi-gaussian distribution. In the case of variance based techniques such as PCA, it is treated as a gaussian distribution, the first

principal component is passing between the two distributions showing no separation. Whereas with Projection Pursuit, in which the first component will pass through the distribution, a clear separation between the distributions appears.

As I said, the type of information that Projection Pursuit maximizes on principal components is not the variance, but other statistical variables such as differential entropy H of a random vector $\bar{X}$, its density being f :

$$H(\bar{X}) = -\int f(x) log f(x) dx \tag{1.47}$$

Once again, the use of this kind of techniques is relevant only for non-normal distributions. The search for statistical independence can be seen has a generalization of Projection Pursuit, because the search for "interesting" structures in data leads to the search of independent statistical components.

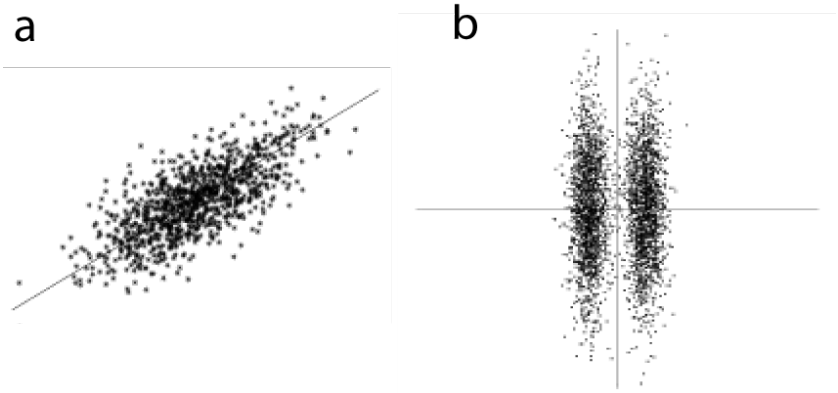


FIGURE 1.10 – *(a) Example of Principal Component Analysis which maximizes the variance on the first principal component (b) Projection pursuit results select the horizontal axes as being the one which reveals the more interesting structure. PCA would have selected the vertical axes as first principal component, thus not accurately revealing the bi-clustered structure of the data. This Figure is extracted from [34].*

The formal definition of ICA is close to Factor Analysis, as it is the search for a data matrix X of a set of components $\bar{S}_i$ which verify :

$$X = AS + \Psi \tag{1.48}$$

The difference with Factor Analysis is that components $\bar{S}_i$ will be determined to be as independent as possible. Perfect independence defined by equation 1.45 is not usually reachable with linear transformation, so in most of the cases one will have to approximate independence. In the common use of ICA the noise term Ψ is avoided as the task of evaluating it is as difficult as performing ICA and seems irrelevant as the application of noise-free ICA works in many real world data cases. The new general equation for ICA is then :

$$X = AS \tag{1.49}$$

As the condition 1.45 will not be satisfied in a vast majority of the cases, one needs to define its own independence condition and adapt the algorithm of ICA in consequence. There are two class of algorithms for performing noise-free ICA. The first uses the principle to search all the

independent components at the same time by minimizing a general function. The second searches first the most interesting components, and then the second most interesting which will be almost independent to the first one, and so one.

In the first class of techniques one relies on the maximum likelihood function. It is possible to define the log-likelihood L of the noise-free ICA model and then search for components that will maximize it.

$$L = \sum_{i=1}^{T} \sum_{j=1}^{m} \log f_j(\bar{X}_i \underline{W}_j) + T \ln |\det W| \tag{1.50}$$

with $W = A^{-1} = (\underline{W}_1, \underline{W}_2, ..., \underline{W}_m)$, f_i density functions of $\bar{S}_i$ which are assumed to be known. In the first class of techniques one can also use the mutual information function I between m random variables $\bar{X}_i$

$$I(\bar{X}_1, \bar{X}_2, ..., \bar{X}_i) = \sum_i H(\bar{X}_i) - H(X) \tag{1.51}$$

where H is the differential entropy defined in the Projection Pursuit paragraph.

In the second class of techniques for noise-free ICA the main type of statistical parameter one will use for finding the independent components one after the other is Negentropy. It is directly derived from mutual information and is defined as

$$J(X) = H(X_{gauss}) - H(X) \tag{1.52}$$

where X_{gauss} is a gaussian random vector with the same covariance matrix as X. Using this parameter ICA shows a strong link to Projection Pursuit, and will, as in the latter case, maximize the clusterization of data if they are far from normality.

In practice ICA is used in many fields; its historical application is in signal analysis for independent signal decomposition. It has been used for feature extraction, exploratory data analysis, blind decovolution, and so on [34]. It is a powerful extension of linear techniques based on second order measures like variance or covariance. In its simple form, it will only search for linear decomposition of the signal, but the possible extensions include using techniques of non-linear PCA. If one wants to go further and have a better insight into the data one needs then to suppose non-linear correlations between instances, which is often the case.

1.2 Non-linear methods

Linear dimensionality reduction techniques give perfect (that is without loss of information) results in a reduction to p dimensions (with $p = rank(X) = rank(S) \leq min(n-1, p)$). Only then the geometric structure of the representation is preserved when compared to the one given by the full number of dimensions. Often one wants to go further in the dimensionality reduction, for example down to two dimensions for proper visualization. The only choice available with linear methods consists in deleting all unused components and thereby inducing severe deformations in the representation. This is standard procedure in the cases of the linear dimensionality reduction techniques discussed above.

To overcome this problem one can use non-linear techniques that will search for non-linear correlations between data and then eventually lead to a lower number of principal, non-linear components. Alternatively, one will first impose the number of dimensions to which the data-objects shall be reduced and then one tries to find the best representation in this space minimizing a chosen geometric quality assessment parameter.

As said in the introduction of this chapter one searchs here for a non-linear mapping function F from the input data to its dimension-reduced representation. To retrieve this mapping non-linear dimensionality reduction techniques will be performed, consisting in defining one or more statistical parameters which evaluate the quality of the representation, and subsequent construction of an algorithm for minimizing the value of these parameters. There is two types of methods available seeking an explicit formulation of the mapping function F, or trying to retrieve the final representation directly leaving apart F. In practice, non-linear dimensionality reduction techniques are always based on optimization processes, for retrieving F as well as for finding the proper configuration of instances. In the following section, describing major non-linear techniques, we will in consequence often deal with optimization tools such as gradient descent, statistical machine learning theory and artificial neural networks [8].

1.2.1 Multidimensional Scaling (MDS)

The oldest non-linear dimensionality reduction technique is called Multidimensional Scaling (MDS) [14,35]. It is based on the principle of retrieving a proper configuration of points based only on information of similarity or dissimilarity (which we shall refer to with the general term of "distance") between the data-points. MDS stems from the observation that the most valuable information lies in the general structure of the data and not in the very values of the variables but rather in the information of similarity or dissimilarity between them.

We already discussed the first exact linear method of MDS which was proposed by Torgerson in 1952 [13], and is called nowadays Classical Scaling (see 1.1.3). It only works when Euclidean distances are provided to the algorithm. We show that due to the link given by SVD between the inner and outer products Classical Scaling gives the same results as PCA. Those methods are exact, if one reduces dimensionality to a r-dimensional basis, with r being the rank of the singular value matrix. If one goes further in the reduction, it induces a deformation in the distance between points, and one then needs an optimization method for the reduction. In conclusion, as for all optimization processes, the major problem in developping MDS techniques is to determine an initial state and choose a good optimization criterion and algorithm. For a detailed review on all the algorithms available for MDS one can refer to [14,35].

The MDS algorithm I have developed and used during my thesis is based on a spring analogy. The idea is to virtually connect all data-points to all other instances using springs. As a spring has an equilibrium length, if one runs a molecular dynamics simulation on it, it will tend to go back to its equilibrium state (see Figure 1.11). The equilibrium length for the spring between point i and point j will be defined as the Euclidean distance $d(\bar{X}_i, \bar{X}_j)$ as given by the initial state. So for each instance $\bar{X}_i$ a force is defined $F(\bar{X}_i)$, which is the sum of all spring interactions $F_{spring}(\bar{X}_i, \bar{X}_j)$ with the other instances $\bar{X}_j$, minus a friction term to avoid infinite oscillations of the spring network :

$$F_{spring}(\bar{X}_i, \bar{X}_j) = -k_{ij}(\delta(\bar{X}_i, \bar{X}_j) - d(\bar{X}_i, \bar{X}_j))(\bar{X}_j - \bar{X}_i) \tag{1.53}$$

$$F(\bar{X}_i) = \sum_{j \neq i} F_{spring}(\bar{X}_i, \bar{X}_j) - \gamma m_i \dot{\bar{X}}_i \tag{1.54}$$

with $\delta(\bar{X}_i, \bar{X}_j)$ being the distance between instances in the r dimensional space, k_{ij} the strength of spring ij, γ the friction parameter, and m_i the mass given to each point. We consider that every spring and all instances are equal in strength and weight so k_{ij} and m_i are the same for every i and j ($k_{ij} = k$ and $m_i = m$). It is, however, possible to use different parameters, for instance according to experimental precision different weights can be considered for the different instances.

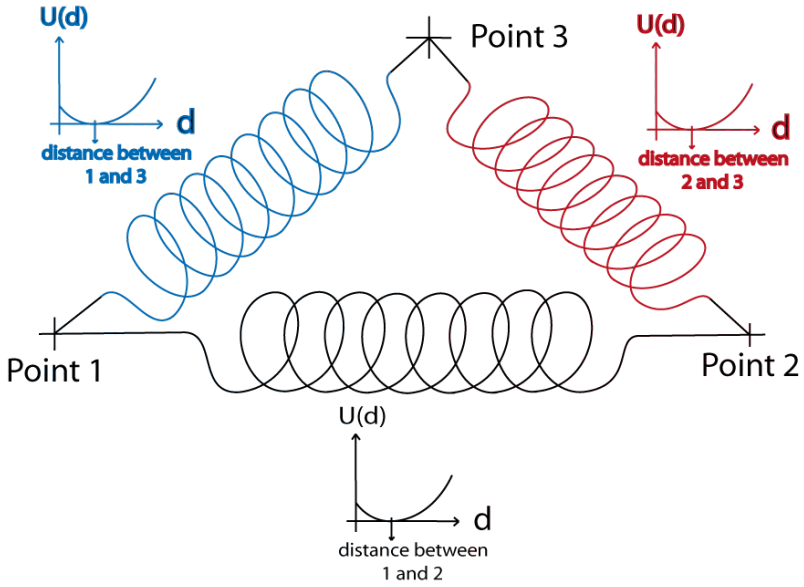


FIGURE 1.11 – *Principle of Molecular Dynamics Multidimensional Scaling. The three points to be embedded in a dimension reduced basis are linked thanks to spring. The energy minimum state will be searched using the equilibrium length of all springs which is the distance defined within the input data.*

A molecular simulation using the force vector is then executed. Following Newton's law : $m_i\ddot{\bar{X}}_i = F(\bar{X}_i)$, with $\ddot{\bar{X}}_i$ the double temporal derivative of vector $\bar{X}_i(t)$. In order to find the new position and velocity of our data points at the next simulation time step we then use simple Verlet integration :

$$\bar{X}_i(t + \Delta t) = 2\bar{X}_i(t) - \bar{X}_i(t - \Delta t) + A\Delta t^2 \tag{1.55}$$

$$\dot{\bar{X}}_i(t) = \frac{\bar{X}_i(t + \Delta t) - \bar{X}_i(t - \Delta t)}{2\Delta t} \tag{1.56}$$

with $\dot{\bar{X}}_i(t)$ being the temporal derivation of vector $\bar{X}_i(t)$. The algorithm is run with simulation time t increasing. To avoid divergence of the Verlet algorithm parameters of the simulation k, m, γ, and Δt have to be well chosen. In all the simulations we made we choose : $k = 1$, $m = 5$, $\gamma = 0.1$ $\Delta t = 0.02$, as an empirical study has shown those parameters to be the most appropriate. We also discover that a good way to provide divergence was to modify all initial states provided to the MDS algorithm by rescaling them to fit in a hypercube with a diameter of 6. This rescaling only consists in multiplying the initial state matrix by a scalar α. It will deform all distances between instances the same way, so it does not influence the organization of the cloud of points. The only initial state matrix we will consider from now on for MDS has been rescaled according to this operation.

To control the minimization process at each time step, a cost function termed the Kruskal stress is calculated according to [14] :

$$e = \sqrt{\frac{\sum_i \sum_j (\delta(i,j) - d(i,j))^2}{\sum_i \sum_j d(i,j)^2}} \tag{1.57}$$

this global parameter indicates how much the distance in the current cloud of points is different from the one in the input data matrix, and therefore a direct evaluation of the amount of energy in the system, and hence the loss of distance information, is possible. The optimization procedure thus minimizes the amount of lost information during the dimensionality reduction. Of course, Kruskal stress is not the only statistical parameter one can use for MDS. It exists a large variety of them [14], but we decided to use this one as it is the most common and the best adapted to MDS on Euclidean distances. Kruskal stress directly evaluates the distance information deformation. For example if every distance in the new cloud of data points is 10% different from the original distance in the input matrix, one obtains for every i and j : $(\delta(\bar{X}_i, \bar{X}_j) - d(\bar{X}_i, \bar{X}_j))^2 = (0.1 * d(\bar{X}_i, \bar{X}_j))^2$.
So $e = \sqrt{\frac{\sum_{i \neq j}(0.1 * d(\bar{X}_i, \bar{X}_j))^2}{\sum_{i \neq j} d(\bar{X}_i, \bar{X}_j)^2}} = 0.1$. A 10% difference on two large distances will therefore influence the Kruskal stress value more than a 10% difference between two small distances. This phenomenon has been studied by Graef et al. in 1979 [36], by looking at the influence of big distances compared to median and small distances on Kruskal stress using existing and generated datasets. MDS using Kruskal stress and molecular dynamics approaches will better conserve global organization then local one.

Our molecular dynamics approach is justify by Dzwinel et al. [37] who prove that every optimization process can be performed using a virtual particle paradigm. Molecular Dynamics represent an ideal solver for virtual particle problems. They also demonstrate great advantages of using particle simulations such as simplicity of the computer model used and its inherent parallel characteristic, and also the few numbers of free parameters. Following this study, they developed their own MDS algorithm using molecular dynamics and applied it to Geophysics problems [38].

In [39] a method for performing Multidimensional Scaling using Molecular Dynamics is also described. This method is a combination of the Molecular Dynamics based MDS we use here and the method of Simulated Annealing. The latter is supposed to find a global minimum by using stochasticity to choose between different probable states. The general Newton equation 1.54 is substitued to a stochastic Langevin equation

$$F(\bar{X}_i) = \sum_{j \neq i} F_{spring}(\bar{X}_i, \bar{X}_j) - \gamma m_i \dot{\bar{X}}_i + F_{stochastic} \tag{1.58}$$

In practice, this combination of methods is equivalent to adding a stochastic force to every data point $F_{stochastic}(\bar{X}_i) = -T * s(t)$ where $s(t)$ is a random number given by a generalized Gaussian stochastic distribution, and T is the temperature of the system. By beginning the simulation with a high temperature T and decreasing it across the simulation exponentially, one expects to reach the global minimum as the stochastic force avoids getting trapped in local minima.

Another derivative of MDS algorithms has been created by Andreas et al. [40]. They propose to run an interactive algorithm during which the user will be able to move the configuration away from a local minimum by hand. This development called interactiveMDS helps the user to directly assesses the rigidity of a configuration of points, to directly see if a particular structure is due to false local minima our real structure deduced from distance values between constituting points.

The first major application of MDS to "omics" data can be found in the works of Gray et al. [42] in which they demonstrate, using MDS, how one can trace back metastasis progression in different tissues. At that point they used binary data revealing for each type of tissue the presence or absence of metastases. In an article from Taguchi et al. [43] one can find the application of MDS to a large number of data-points. Their goal was to use MDS on cell-cycle gene expression data for obtaining a visual representation of the oscillatory phenomenon. They thereby prove the

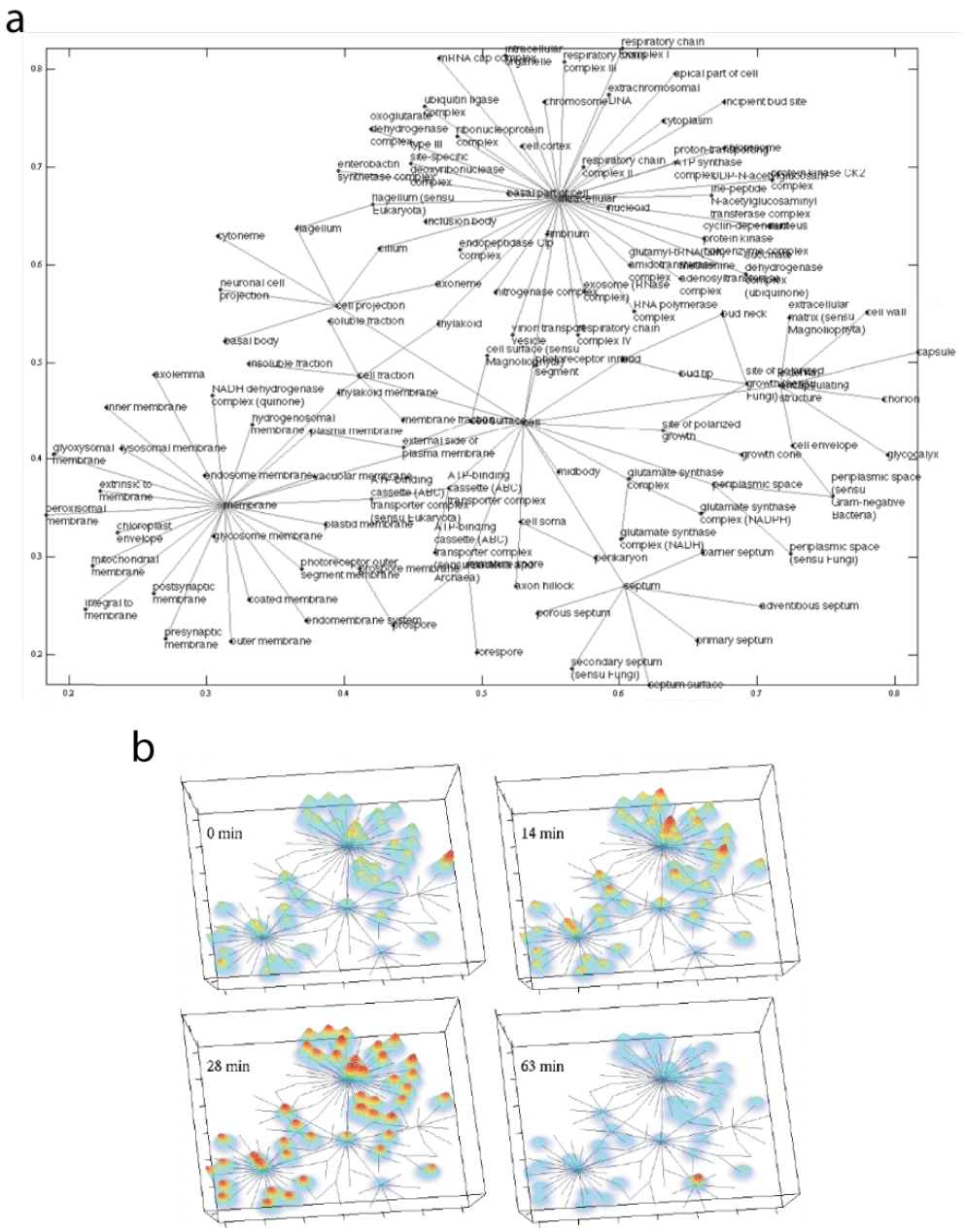


FIGURE 1.12 – *Figures extracted from [41] showing in (a) a part of the GO cellular component ontology visualized by the spring embedding algorithm. The ontology is rooted at the term "cell" and extends to a distance of 2 edges. The directions of the GO relationships are not shown to avoid unnecessary complication of the display. (b) The GO cellular component ontology of Figure 1 is shown (solid lines) with the mean expression values of all genes annotated to each node overlaid in the 3rd dimension (coloured landscape). Each panel shows expression data for a different time point in the cell cycle. Note that the height and color of the landscape in each panel has been scaled to maximize differences in mean expression between the nodes, and thus is not comparable between panels.*

significant advantage of MDS compared to linear techniques such as PCA, as the MDS process, due to its optimization procedure, assures a more accurate representation of the data. Other applications of MDS to "omics" data may also be found in [44, 45].

I show in Figure 1.12 the analysis of Ebbels *et al.* in [41]. They use an algorithm derived from MDS to represent networks of GeneOntology (GO) [46] components (see Figure 1.12-a). GO is a database of molecular-function terms which define the hierarchical link between all those terms. Once a network of GO molecular-functions is obtained using a spring embedding algorithm, they were able to map the information of gene expression over the *Yeast* cell cycle. Thus, they show in Figure 1.12-b the different types of molecular functions which are used during the cell-cycle. This example demonstrates how a good 2-dimensional representation obtained with efficient tools can help the human-mind to get a global view on a system, and then go deeper into the analysis based on this information.

1.2.2 K-means

Usually, a companion tool of Dimensionality Reduction in a data analysis is Clustering which means searching for groups of data sharing the same properties. Clusters are revealed by a low measure of distances between instances from one to each other within any group. Consequently every clustering algorithm consists in first defining a measure of distance between data points, and then use it for the regrouping of points according to their dissimilarities.

It exists three common ways of performing clustering : Hierarchical clustering, K-means clustering, and Self Organizing maps. The former corresponds to the basic idea of constructing a dendogram which represents a hierarchy of clusters. The second and third techniques will be described in the following. Both are based on optimization processes which seek for the best clusterization of the data. Also, in both cases the principle of the underlying methods relate them to techniques of dimensionality reduction. I will explain in the two following paragraphs the principle of these two typical clustering techniques, explaining why one can also consider them as dimensionality reduction methods. This latter fact does not seem fallacious as finding clusters which summarize data is pretty close to finding a mapping function which helps to represent data in a convenient way. Moreover, it is this very fact which favored the development of the two techniques Projection Pursuit and ICA, already discussed in section 1.1.6.

The first method I will illustrate is K-means clustering, which seeks for the best clusterization in K distinct groups. It was invented in 1956 by Steinhaus [47], an algorithm was proposed by Loyd just after in 1957 but not published until 1982 [48]. The principle is simple, given a data matrix X, one searches for K clusters C_k each characterized by m^k the centroid and n_k the number of points in this cluster. Finally, one has to find the clusters which will reach a global minimum of the function J_K

$$J_K = \sum_{k=1}^{K} \sum_{i \in C_k} (\bar{X}^i - m^k)^2 \tag{1.59}$$

This equation can be developed [23] in a form where the first term is constant, the second being subject to minimization.

$$J_K = \sum_i (\bar{X}^i)^2 + \sum_{k=1}^{K} \frac{1}{n_k} \sum_{i,j \in C_k} \bar{X}^i \bar{X}^{j^T} \tag{1.60}$$

Lloyd's K-means algorithm is still the prevalent one for minimizing this function; it has become a major algorithm in computer sciences due to its ubiquity. First, one decides an initial

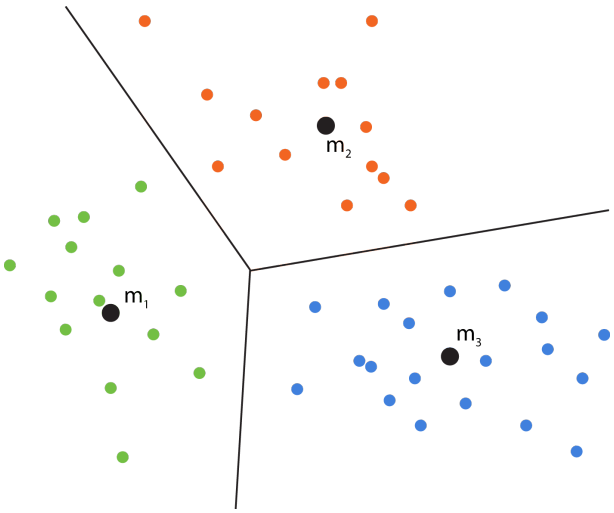


FIGURE 1.13 – *Illustration of K-means principle for $k = 3$ on a random dataset. Black lines represent the Voronoi diagram, black dots are the centroid of the three clusters.*

position for all centroids m^k, then a Voronoi diagram determines the best division of a cloud of points according to the m^k. Given the Voronoi separation one can calculate n_k, the number of points in each cluster, and thus evaluate the function J_K. For each cluster C_k, given by previous operation, one calculates the new centroid m^k. The Voronoi diagram provides new values for n_k, and consequently new clusters C_k. By repeating these steps one converges to a steady state, which may be the global minimum of J_K (see Figure 1.13). Multiple runs of the algorithm with different initial conditions will improve the chance of identifying the global minimum. Keep in mind that k-means is a supervised learning technique which depends on the choice of K. Fallacious choice of K can lead to spurious results, in consequence one has to run diagnostic checks before k-means to determine the ideal number of clusters in the data set.

Ding *et al* [23] show that this minimization problem can be improved using PCA for determining the initial configuration because principal components are the continuous solution of the discrete cluster membership indicators of K-means clustering. The simple example they gave for illustrating this fact is close to the one given in section 1.1.6 when discussing the Projection Pursuit. A convenient way to initialize a 2-means clustering algorithm is obviously to select the first cluster as the group of points placed above the first principal component, the second cluster will regroup points placed below (see Figure 1.10). This close link between PCA and k-means reinforces what was previously said on the link between clustering and dimensionality reduction. In this context, it is interesting to note that the "best" representation of data is generally the one which best discriminates clusters from each other.

1.2.3 Self Organizing Map (SOM)

Self Organizing Map (SOM) is a technique which stems from Artificial Neural Networks theory [8]. Its purpose is to map data in a 2-dimensional abstract space according to distances between instances. For this reason it is used both for dimensionality reduction and clustering. It is also called Kohonen maps as it was created by Teuvo Kohonen [49].

The SOM technique is based on an artifical neural network with competitive learning rules [8]. A special specification of the learning rule allows for each update of the weights of a neuron to

modify the weights of neighbors. This operation will consequently preserve the topology of the input space producing an abstract representation of it. The artificial neural network for SOM consists generally in a two-dimensional array of p neurons arranged on a squared or hexagonal lattice with no lateral connections (see fig 1.14. For each instances $\bar{X}_i$ of a matrix X, being $n.m$, one connects it to each neuron of the network according to the weight vector $\bar{W}_{i'}$ given by the weight matrix W which is $p.m$.

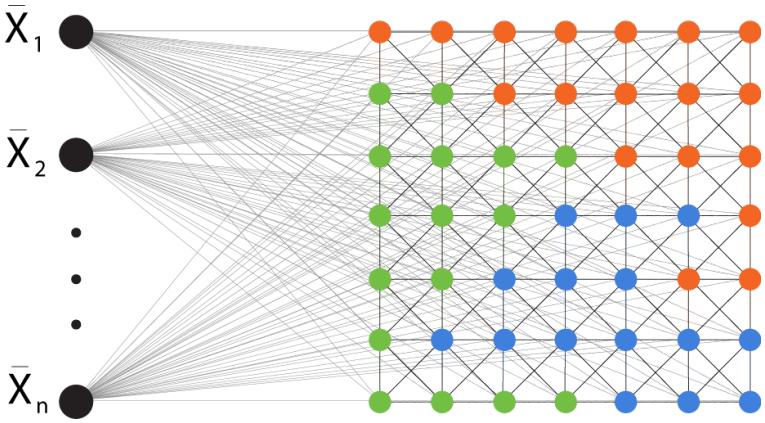


FIGURE 1.14 – *Figure representing the artificial neural network organization of a Self-Organizing Map. The colorization of the network is an imaginary result of the SOM where each neurons is colored if the map vector $\bar{W}_i$ is in one of the three known clusters of the data.*

The algorithm is initialized by randomly generating a matrix of weights W. Then for each time step t [50] :

- one will select the neurons $\eta_{i'}$ which best "represent" an input instance $\bar{X}_i$, following the rules

$$\eta_{i'} = \arg \min_{k \in 1,2,..,p} ||\bar{X}_i - \bar{W}_k|| \tag{1.61}$$

  the type of distance used is not specified here, but in the vast majority of cases one uses Euclidean distance between instances and weight-vectors of neurons.
- Once the winning neuron $\eta_{i'}$ is found, one has to update its weight vector $\bar{W}_{i'}$ and also the weight vectors of its neighbors. Neighbors' weights are updated following a different scheme, the most common being based on a gaussian rule, in which the update follows a bidimensional gaussian distribution centered in the neurons $\eta_{i'}$. The general updating equation is given by

$$\Delta \bar{W}_k(t) = \alpha(t) neigh(\eta_{i'}, k, t)(\bar{X}_i - \bar{W}_k) \tag{1.62}$$

  where $\alpha(t)$ is the learning rate, and $neigh(\eta_{i'}, k, t)$ is the neighborhood function which evaluates the propagation of changes in neuron $\eta_{i'}$ through its neighbors η_k. In the case of a gaussian rule, it will be of the form

$$neigh(\eta_{i'}, k, t) = \exp(-\frac{||\eta_{i'} - \eta_k||_m^2}{2\sigma(t)^2}) \tag{1.63}$$

  where $\sigma(t)^2$ is the changing effective range of the neighborhood, and $||\eta_{i'} - \eta_k||_m$ is the distance between the two neurons in the feature map (i.e. the neurons lattice). One denotes U the matrix of all distances between neurons in the map.

One has to repeat these two steps until an equilibrium is reached. To increase convergence, $\alpha(t)$ decreases monotonically, and $\sigma(t)$ increases monotonically. Furthermore, $\alpha(t)$ has to satisfy conditions (1) $0 < \alpha(t) < 1$, (2) $\lim_{t\to\infty} \sum_t \alpha(t) \to \infty$, and (3) $\lim_{t\to\infty} \sum_t \alpha^2(t) < \infty$.

Finally, two types of information will be provided by SOM. An abstract representation of the data given by the lattice of neurons. This representation can be completed by a colorization linked to instance properties such as known clusters (see Figure 1.14), values of an important variable, or the U matrix properties. Possibilities are infinite and depend of the information one searches to highlight. The second information given by SOM is the weight-vectors $\bar{W}_i$. They have the same dimensionality as instances $\bar{X}_i$, and they are optimized to be as close as possible to them. Thus the matrix of weights W is a set of "principal feature instances" representing the center of a particular topological zone of the input data.

To obtain a better representation of data using SOM, conservation of topology but also conservation of local features of the data is required. In this spirit a novel algorithm named ViSOM [50] has been developed. A constraint on the distance between the weight neuron vectors is added to the learning equation in order to conserve local organization. This new algorithm increases the link between SOM and dimensionality reduction, which is obvious as one maps data in a 2-dimensional lattice of neurons.

1.2.4 Non linear PCA

In the previous sections Multidimensional Scaling and two techniques of clustering were described. These methods do not seek for an analytic expression of the mapping function. They are also applicable to data with non-linear correlation. However, no *a priori* knowledge on the linearity of the data is required for using them. In the following I will discuss techniques which are specific for proved non-linearly correlated data, and seek to find the mapping function.

Before seeking non-linear mappings one has to assess the non-linearity of the data [51]. To do so, one has first to perform PCA on different centered regions of the data, and to demonstrate whether principal components are really best-fitting lines or, if using curves, a better results in terms of total variance minimization would be obtained. A statistical test is set to evaluate the change of results between each PCA an idea of the non-linearity. If one has pure linearly correlated data, a multiple run of PCA for different centered data would give the same results. This idea is linked to the fact that PCA and regression are closely related and seek the same objects. PCA searches the best lines which represent the linear correlation of the data, as does a linear regression. Figure 1.15, shows how with a random cloud of points, a polynomial function can best represent the data. This is compared to a linear regression to demonstrated the advantage of non-linear regression. When dealing with non-linear correlation, the same tools as non-linear regression need to be used. If multiple linear regressions on a cloud of points are performed in different regions of the data, and the results are different, non-linear correlation between points needs to be assumed.

As soon as the usefulness of using non-linear techniques of dimensionality reduction is realized, multiple alternative techniques can be chosen. I present here the three best known ones and make a brief review of other possible techniques which exist.

The oldest method is an extension of PCA to the non-linear world, it is simply called non-linearPCA (nPCA). As I demonstrate in section 1.1.1 the main equation of PCA which needs to be solved is :

$$\bar{X}_i = U\bar{Y}_i \tag{1.64}$$

for every instance $\bar{X}_i$, and instances in the principal basis $\bar{Y}_i$, with U being the loading matrix representing the linear transformation between X and Y. A non-linear extension of this equation

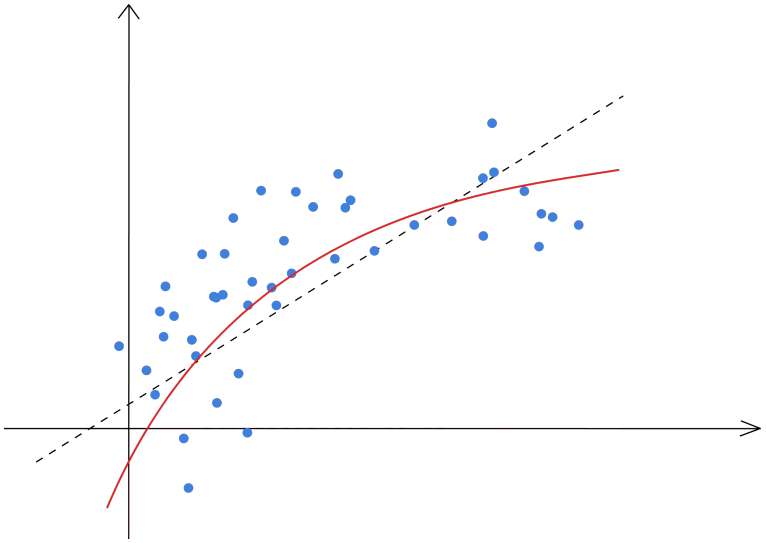


FIGURE 1.15 – *(black dashed line) Represents the result of linear regression. (red curve) represents the results of a polynomial regression. In this example the non-linear technique would be better suited.*

will be

$$\bar{X}_i = f(\bar{Y}_i) \tag{1.65}$$

where f is a non-linear function.

As exact solutions can not be found, an optimization algorithm needs to be employed. The cost function to minimize will be

$$C = \sum_{i=1}^{n} \|\bar{X}_i - f(\bar{Y}_i)\|^2 \tag{1.66}$$

It can be shown that Euclidean distance is a critical value for constructing f [51]. In fact, the key concept of the "best-fitting" line in PCA will now be replaced in the non-linear case by the concept of the principal curve [52]. The definition of the principal curve is linked to the definition of the distance between data points and the curve which is a representation of f. A smooth curve $f(\bar{Y}_i)$ is a principal curve if it corresponds to three properties. First it can not intersect itself, then it has to have finite length inside any bounded subset of $\mathbb{R}^n$, and finally it has to be self-consistent.

$$f(\bar{Y}_i) = E(\bar{X}_i | P_f(\bar{X}_i) = \bar{Y}_i) \tag{1.67}$$

with P_f being the projection function of points onto the curve f. This last property is the most important as it implies that every point $\bar{Y}_i$ on the curve is the mean of the points $\bar{X}_i$ which projections is $\bar{Y}_i$ itself, thus implying self-consistency. This definition is the mathematical translation of saying that the principal curve is the curve which passes in the line of gravity of the cloud of points. Figure 1.15 show that the red curve given by non-linear regression is a better representation of the principal curve, following the previous definition, than the black dashed line representing a simple linear regression, because the red curve is more often close to the line of gravity of the distribution than the black dashed line.

To find this principal curve, many algorithms have been developed. The best known [53] uses auto-associative neural networks, constituted of five layers, which infer the function f with

typical techniques of artificial neural networks training [8]. Another way of solving the problem is named KernelPCA [54], which uses Kernel functions that help to embed the manifold in a higher dimensional space, to simplify the structure of the manifold, and finally to retrieve it easily. A complete review of the problem of non-linear PCA can be found in [51].

With non-linear PCA and the search for a principal curve rises the idea of a principal manifold. A manifold is a general geometric object such as a curve, surface or hyperplane. One considers that data are noisy representations of p random variables. These random variables are correlated, linearly and non-linearly. A representation in $\mathbb{R}^p$ of these random variables will create a manifold, called principal manifold, for example if $p = 2$ this manifold will be a curve. To create data, one has to measure n different variables. These variables are not direct representations of the p random variables, and therefore the geometric object formed by the n variables is not similar to the principal manifold. The goal of techniques seeking the principal manifold is to retrieve it based solely on the data. In Non-linear PCA an optimization process searches the analytical expression of the mapping function which helps to define the manifold. In the two following sections another spirit is adopted : One assumes (i) that the principal manifold does exist, (ii) that it is non-Euclidean, and finally (iii) that a deformed representation of it is known (*e.g.* the data). Consequently, one will obtain a better dimensionality reduction of the data by using techniques which fulfill these criteria.

1.2.5 ISOMAP

One of the first techniques invented which is based on the assumption of existence of a principal manifold and reduces the dimensionality of data using this very manifold is called Isomap, and was published in 2000 by Tenenbaum et al. [55]. It can be seen as a development of MDS, but with a special definition of distance as it uses geodesic distances. To illustrate the usefulness of this approach, I use the example directly extracted from [55]. A geodesic is simply the shortest path in a manifold between two points. On a plane it is of course a line.

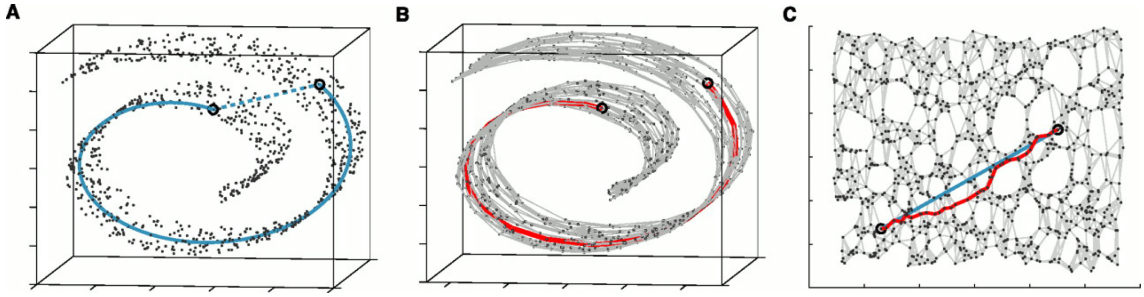


FIGURE 1.16 – *Isomap illustration using "Swiss roll" data, this Figure is extracted from [55]. (a) the dashed line represents the Euclidean distance, the solid curve represents the geodesic. It is obvious in this example that the geodesic better takes into account the intrinsic structure of the data (e.g. the principal manifold). (b) The "Swiss roll" data after the neighborhood graph G has been constructed using the $K = 7$ closest neighbors. The shortest path (in red) appears to be a good approximation of the geodesic. (c) Results of a two-dimensional embedding using K-Isomap. Euclidean distance in this representation represented by the straight blue line is now a good approximation of the shortest path distance represent by the red line.*

Assume that the data are a bad representation of a swiss roll, and thus the data are indeed embedded in this swiss roll. If one calculates the Euclidean distances between points the structure

of the manifold will be lost (see Figure 1.16-a). A definition of distance, following the structure of the manifold such as geodesic distance, will be more appropriate (see Figure 1.16-b).

Isomap, just as MDS, takes as input a matrix of distances $d(\bar{X}_i, \bar{X}_j)$ with $(i,j) \in \{1,2,..,n\}^2$. From these distances a graph representation of the data by defining the nearest neighbors of each instance (see Figure 1.16-c). For a given instance $\bar{X}_i$ its neighbor can be chosen following two methods :

- All instances $\bar{X}_j$ verifying $d(\bar{X}_i, \bar{X}_j) < \epsilon$ are chosen with ϵ being a constant selected by the user. This kind of selection is called ϵ-Isomap.
- Alternatively, the neighbors of the all closest K instances of $\bar{X}_i$ are chosen ; this is called K-Isomap and K has to be defined by the user.

Given an instance $\bar{X}_i$, edges are linked to its neighbors, and the lengths of each edge will be set using the classical Euclidean distance. Once all edges have been defined, the shortest path in the graph G is computed, thus an accurate representation of the geodesic is obtained. The geodesic distance $d_G(\bar{X}_i, \bar{X}_j)$ will be defined by adding the length of all edges constituting the shortest path between $\bar{X}_i$ and $\bar{X}_j$. Finally, a MDS algorithm is executed using the new geodesic distance matrix D_G, retrieving the proper representation of the data (see Figure 1.16-c). cMDS algorithm can be used as in the example given in Figure 1.16 or, a more general, non-linear MDS algorithm can be used.

This technique has been applied by Dawson *et al.* [56] in order to discover clusters of common phenotypes between different tissues of the rat, based on data generated using Affymetrix microarray technology.

1.2.6 LLE

In the same volume of the *Science* journal where Tenenbaum *et al.* described Isomap, Roweis *et al.* published a technique called Locally Linear Embedding (LLE) which takes a radically different approach [57]. It stems from the fact that the underlying principal manifold might be differential and therefore can locally be approximated by an hyperplane. For every instance which samples this manifold one will try to recover a hyperplane. Construction of a dimensionality reduced representation is attempted in accordance with the information retrieved from all the local hyperplanes.

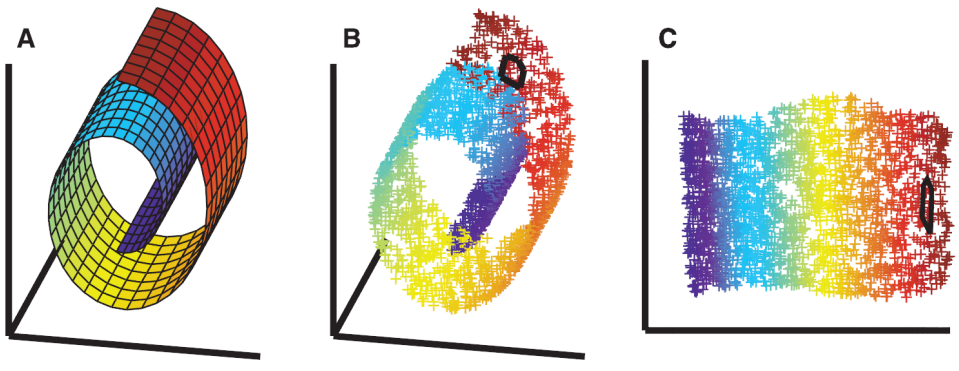


FIGURE 1.17 – *Results of Locally Linear Embedding on the "Swiss roll" dataset. This Figure is extracted from [57]. (a) Decomposition of the underlying manifold in a set of planes. (b) Selection of a set of neighbors for any instance (represented by a black circle). (c) Embedding of all instances in a 2-dimensional vector space.*

To describe and illustrate more in detail the algorithm, I will use again the "Swiss Roll" example as did Roweis *et al.*. Figure 1.17-a shows how a manifold can be approximated by a set of hyperplanes, more precisely it describes how a differential 3-dimensional manifold can be approximated by 2-dimensional planes. The first step of LLE is the same as in the Isomap approach, as it consists in finding the nearest neighbors of all instances $\bar{X}_i$ (see Figure 1.17-b). One seeks a linear link between $\bar{X}_i$ and its neighbors $\bar{X}_j$, which corresponds to characterizing the underlying hyperplane. To do so, the best weight matrix W has to be found which minimizes the cost function $\epsilon(W)$.

$$\epsilon(W) = \sum_i |\bar{X}_i - \sum_j w_{ij}\bar{X}_j|^2 \tag{1.68}$$

where W indicates the contribution of each of the neighbors of an instance. Enforcing $w_{ij} = 0$ for two instances $\bar{X}_i$ and $\bar{X}_j$ which are not identified as neighbors, and $\sum_j w_{ij} = 1$. The optimal weight matrix W is found by solving a least-square problem.

Once W is retrieved, the last step is to embed instances in an Euclidean space with fewer dimensions. To obtain this, one has to find the proper configuration of a set of points $\bar{Y}_i$ in the dimensionality reduced basis which solve the equation :

$$\Phi(Y) = \sum_i |\bar{Y}_i - \sum_j w_{ij}\bar{Y}_j|^2 \tag{1.69}$$

In this case, the parameter which needs to be found is not the weight matrix W, which has already been found in the previous step, but the data matrix Y corresponding to the position of instances in the representation space (see Figure 1.17-c). This last step can be solved by a sparse $n.n$ eigen-value problem which is described in detail in [57].

1.2.7 A general view on non-linear dimensionality reduction methods : Principal Manifolds Learning

We saw in previous sections different ways of reducing the dimensionality of data using non-linear techniques. I have only presented the state-of-the-art techniques. A vast number of alternative ways to reduce dimensionality exist. One can distinguish them into three classes according to their major objectives [58] : Global structure preservation, local structure preservation, and global alignment of linear models. In Figure 1.18 I summarize this classification, giving examples of techniques from each group.

The first class includes techniques which preserve global structure and contains more methods than the others. Among them one can find techniques which preserve distance such as MDS, Isomap [55], K-means, and Diffusion Maps [59] where random Markov chains are used to obtain a proximity-distance value. This group also includes techniques using kernel functions such as KernelPCA [54]. Finally, techniques using Neural Networks can be found to retrieve directly the mapping function like Non-linear PCA, SOM, or a recently published algorithm called AutoEncoder [60] which uses a N-layer "auto-trained" Neural Network.

The second class of techniques attempts to preserve the local structure of the manifold. The first one to be published was LLE [57] which defines weights between neighbors to determine the most useful ones for the reconstruction of any point. We can also cite techniques which determine the manifold by using the local tangent space of each datapoint, such as HessianLLE [61] or Local Tangent Space Analysis (LTSA) [62]. Finally, one technique called Laplacian Eigenmaps [63] uses Laplacian graphs only defined by the distance between a point and its nearest neighbors to retrieve a low dimensional representation.

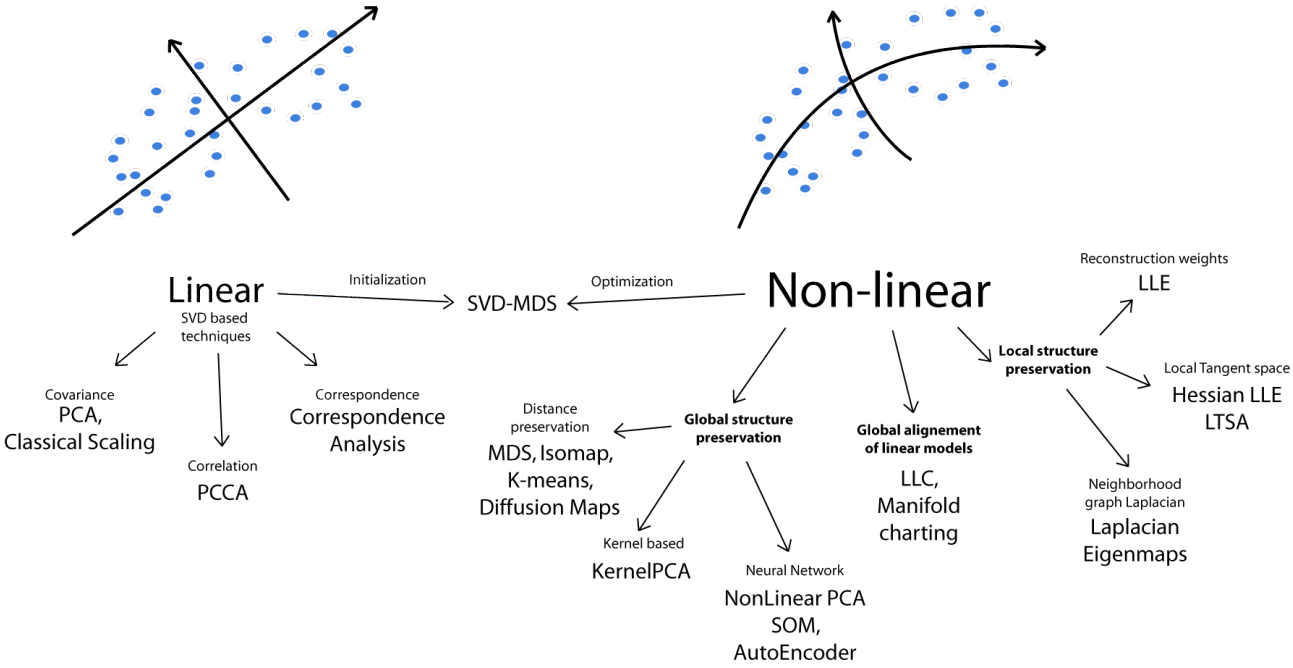


FIGURE 1.18 – *Scheme summarizing the different dimensionality reduction techniques discussed in this chapter.*

The last class includes Locally Linear Coordination (LLC) [64] and Manifold charting [65] and is based on computing a number of locally linear models (i.e. variants of LLE in the case of LLC) and subsequently performs a global alignment of the linear models.

Each class of techniques have their own advantages and drawbacks. For example, it is known [58] that local structure conservation techniques are very badly suited for noisy or intrinsically high-dimensional data. It is also known that for high-dimensional data, the number of instances needed to retrieve correctly its structure grows exponentially with the number of variables [66]. This phenomenon, called "the curse of dimensionality", forbids to retrieve correctly the manifold if the number of points is not sufficient in comparison to the intrinsic dimensionality of the manifold. Moreover, if points are noisy, the local structure will be deformed and so is the representation extracted from it. The example in Figure 1.19 shows the effect of noise on the results of different dimensionality reduction techniques. One can see that adding 10% of noise has more effect on the results of the local structure conservation method. Finally, the risk of over-fitting is important and can lead to false representations. On the contrary, for well defined datasets such as the "Swiss roll", local structure conservation techniques will be very well suited. Due to the specific properties of each technique, we cannot claim that any given technique is better than the others. The choice of the technique to be employed depends on the characteristics of the data on the one hand and on the geometric structure one wants to preserve on the other. To infer this fact, I show in Figure 1.20 different results in of dimensionality reduction techniques on four different types of datasets ; the examples being extracted from [67].

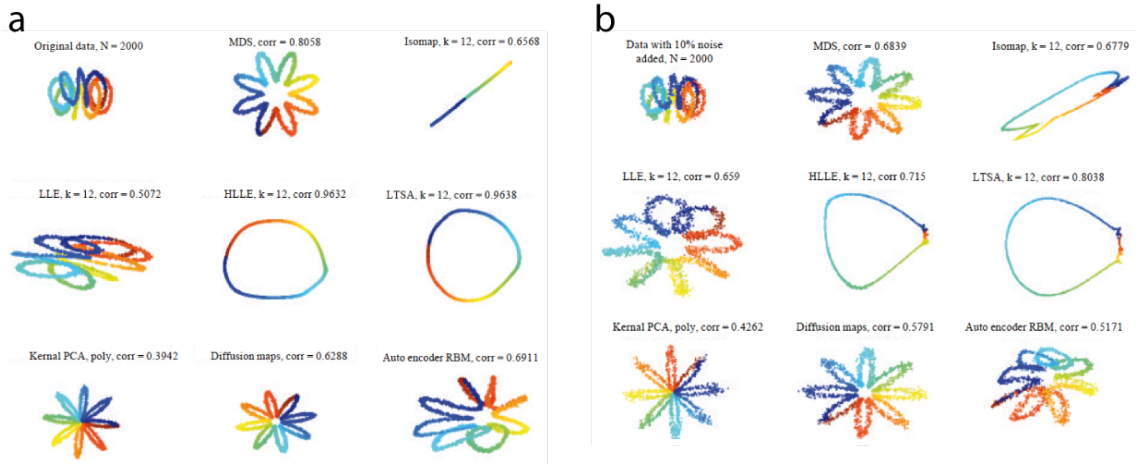


FIGURE 1.19 – *Effect of noise on the manifold retrieving. Utilization of different dimensionality reduction techniques on a virtual dataset (a), and the effect when adding* 10% *of noise (b). This example is extracted from [67].*

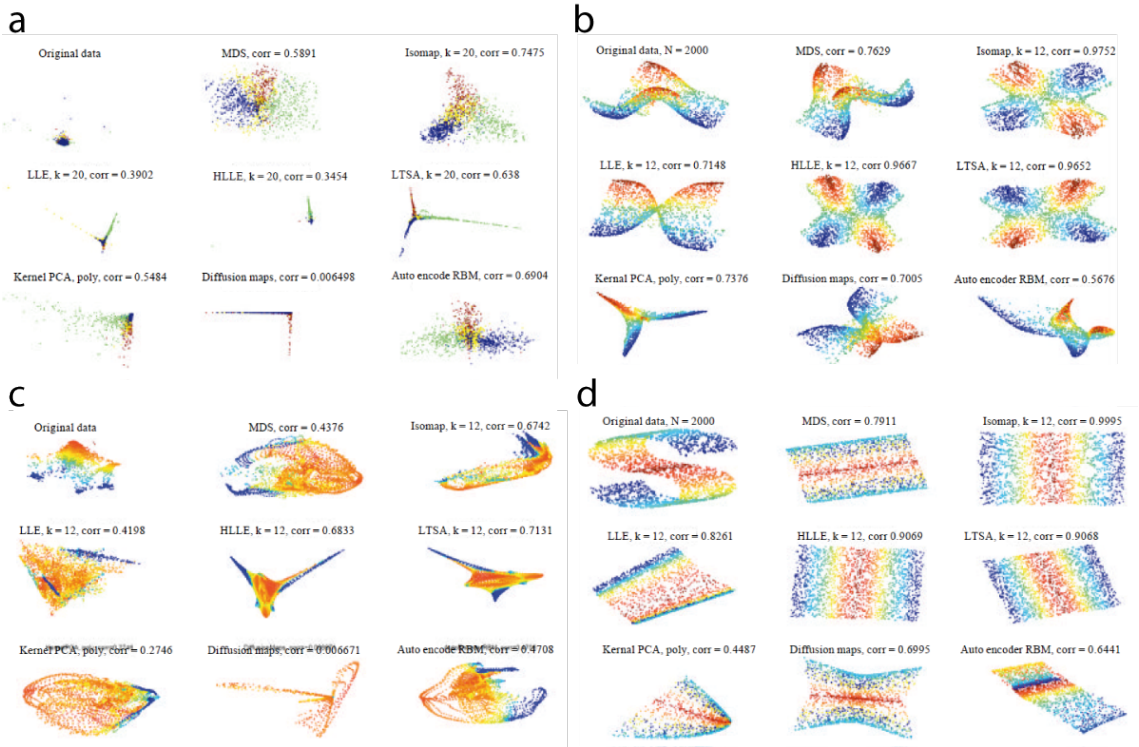


FIGURE 1.20 – *Comparative results of different dimensionality reduction techniques on four different artificial datasets. This example is extracted from [67]*

2

Multidimensional Scaling initialized by Singular Value Decomposition

In the host laboratory where this thesis was conducted, we were concerned by the analysis of transcriptome microarray data. In order to do so, we use a variety of techniques, from simple statistical test to assess differences between biological conditions, to clustering methods and of course dimensionality reduction. The latter was used almost entirely for obtaining an insight into the global structure of datasets, either by visualization in two or three dimensions or by looking at important properties of the principal components. Consequently, we rapidly developed the need to have a dimensionality reduction technique powerful, efficient, and also easy to implement and execute [2].

2.1 The search for a dimensionality reduction technique

The constraint of developing a fast and computationally efficient algorithm is understandable as one will have to run the dimensionality reduction many times during data analysis. Such an algorithm will have to be manipulated by experts and, more importantly, non-experts in the field of multivariate analysis. For this reasons, it should be simple to understand, and should not depend on many parameters.

Manifold learning techniques are not the best suited for the kind of biological dataset studied in the host group. Each one of these techniques seeks for a manifold using only statistical variables information, the reconstructed manifold is then dependent of noise and inner dimensionality of the data [67]. Furthermore, the reconstruction will become harder if the number of dimensions grows. The reason stems from the particular properties of high dimensional data, summarized in [66], which increase the risk of to spurious correlation, compression of the distribution of distances values, and the curse of dimensionality. All these problems infer over-fitting, estimation instability and local convergence.

Consequently, to try to retrieve the manifold using the data seems impossible with noisy and high-dimensional data such as one uses in biology nowadays, and thus it seems that utilization of Multidimensional Scaling is more preferable. Another argument in favor of this choice is that all Principal Manifold Learning techniques except MDS are dependent of one or more free parameters [58].

2. The result of this analysis was summarized in a journal article submitted to BioInformatics journal in October 2010. See appendix D

One drawback of MDS is the fact that points are embedded directly on the manifold without trying to retrieve a formulation of the mapping function, contrary for example to SVD where we have the matrix V representing the mapping function. This matrix V can be important to get information on the variables contributing at most to the first component (see section 1.1.5). From this conclusion stems the idea of combining in one technique the advantages of MDS and SVD.

We call this technique SVD-MDS, and it consists in first performing a linear reduction using SVD, and use the results of it as an initialization state for the non-linear reduction with MDS. Thus it is a technique in between linear and non-linear spirit, trying to combine advantages of both as shown in figure 1.18. As SVD and MDS are parameter-free and extensively used in the data analysis community, their combination will surely fit the goal of an easy to implement and performing technique even for the non-experts. In addition we have demonstrated that the combination of these two techniques is computationally efficient and allows a gain of performance in the process of dimensionality reduction.

To assess the efficiency of SVD-MDS compared to MDS only, we have first experimentally demonstrates the usefulness of using SVD for providing an initial configuration for MDS compared to other initialization strategies. Second, we tried to find the best algorithm for MDS, comparing a normal MDS algorithm to an iterative one and a stochastic one, and showing that these two other algorithms did not improve the convergence. Finally, using a statistical geometric parameter evaluating the local deformation of the representation, we demonstrate that SVD-MDS generally outperforms in terms of quality of the representation all the other algorithms.

As seen in section 1.2.1, Multidimensional Scaling (MDS) is a methodology that reduces dimensionality using only the information of similarities or dissimilarities between instances, hereafter regrouped in the general term of "distance". During the process a part of this distance information will be lost. It hence results an optimization problem of finding an arrangement of the instances in the lower dimensional space that reflects the least loss of distance information when compared to the original distances in the higher dimension. To evaluate the deformation of distance information one uses the Kruskal stress parameter (see section 1.2.1)

$$e = \sqrt{\frac{\sum_i \sum_j (\delta(i,j) - d(i,j))^2}{\sum_i \sum_j d(i,j)^2}} \tag{2.1}$$

In the following it will be our major parameter for evaluating the quality of the dimensionality reduction, and to compare different MDS algorithm.

On the contrary SVD, is not an optimization process, but a matrix operation. It gives a dimensionality reduced matrix Y of a matrix X. As $Y = XV$ with V an orthogonal matrix, distance between instances in X and Y are the same. Consequently, after a singular value decomposition the stress of Y is equal to zero. The problem with SVD (as all linear dimensionality reduction techniques) is that $rank(Y) = min(n-1, m)$ if X is $n.m$. And so when one wants to reduce dimensionality up to p dimensions, with $p < rank(Y)$, the only possibility is to delete the last principal components which have the least inertia. This operation will induce a deformation in the distance between instances, and so the Kruskal Stress will increase in function of the total inertia left in the principal components. The idea behind SVD-MDS is to take advantage of the SVD reduced cloud of points which has non zero Kruskal Stress, and use MDS to reduce the value of stress. As SVD organizes the cloud of points in order to put the maximum information in the first components, we hoped that MDS will be able to reduce the stress, induced by the deletion of the small inertia components, by deforming as little as possible the distribution of points of the first components. And that this organized SVD initial state will be more convenient

for initializing MDS than a random or centered state.

2.2 Datasets used in this study

ID	Dataset Name	No. of Instances	No. of Variables
d1	96Cell	96	32878
d2	96Cell_T	96	1553
d3	Iris	150	4
d4	Wine	178	13
d5	Stochast 200	200	50
d6	CCYier	516	12
d7	Pima	768	9
d8	96Cell_T transposed	1553	96
d9	Secom	1567	590
d10	Ozone	2565	72
d11	Stochast 3000	3000	300
d12	Ecoli	4288	7
d13	Wave	5000	22

TABLE 2.1 – The different datasets used in this study.

The experimental demonstration of the efficiency of SVD-MDS has been made by running the different to be compared algorithms on a variety of datasets. Our goal is to use this algorithm on biological datasets, but the quality of SVD-MDS is not restricted to this type of data, that is why we use a set of datasets with heterogeneous characteristics and not restricted to biological high-throughput data. We got the data used for our demonstration from several publicly available datasets of different origins.

First, we have used two different transcriptome datasets. Briefly, the cellular transcriptome is defined as the *ensemble* of RNA molecules resulting from gene expression in a cell. Using microarray technology, in the human case, some thirty-thousand different RNA species can be quantified simultaneously. The dataset here referred to "d1 — 96Cell" includes ninety-six transcriptome measurements generated from thirty-two individual human tissues under non-pathological conditions. This dataset was initially published by [68], and is available for download from : http ://mace.ihes.fr using accession number : 2914508814. The dataset here called "d6 — CCYier" ([69] ; mace access. no. : 2960354318), is composed of twelve human fibroblast transcriptome data points generated over twenty-four hours during the cell-cycle. Note that we eliminated one (Interleukin 8, IL8) of the 517 genes as an outlier from this dataset. The dataset "d2 — 96Cell_T" (*c.f. Table 2.1*), is a derivative of the initial dataset "d1 — 96Cell", where only genes were retained that are specific to one and only one human tissue as provided in [68], and removing again one outliers gene (Probe_ID : 162105). The dataset "d8 — 96Cell_T" (*c.f. Table 2.1*), is the transposed (Instances, Variables) dataset "d2 — 96Cell_T". All transcriptome datasets were median normalized in log2-space and processed according to standard procedures ([70], [2], [71]). Second, seven additional datasets with no relation to biology were used. They all originate from the Machine Learning Repository [72] :
http ://archive.ics.uci.edu/ml (1) "Iris" here "d3 — Iris", (2) "Wine" here "d4 — Wine", (3) "Pima Indians Diabetes" here "d7 — Pima", (4) "SECOM" here "d9 — Secom", (5) "Ozone

Level Detection" here "d10 — Ozone", (6) "E. Coli Genes" here "d12 — Ecoli", and (6) "Waveform Database Generator (Version 1)" here : "d13 — Wave". Please refer to the ML repository for details on these data.
Third, we generate two datasets stochastically :
(i) One with 200 instances and 50 variables between -6 and 6 here "d5 — Stochast 200", (ii) the other with 3000 instances and 300 variables between -6 and 6 here "d11 — Stochast 3000".
The number of instances and the number of variables for all thirteen datasets is given in Table 2.1.

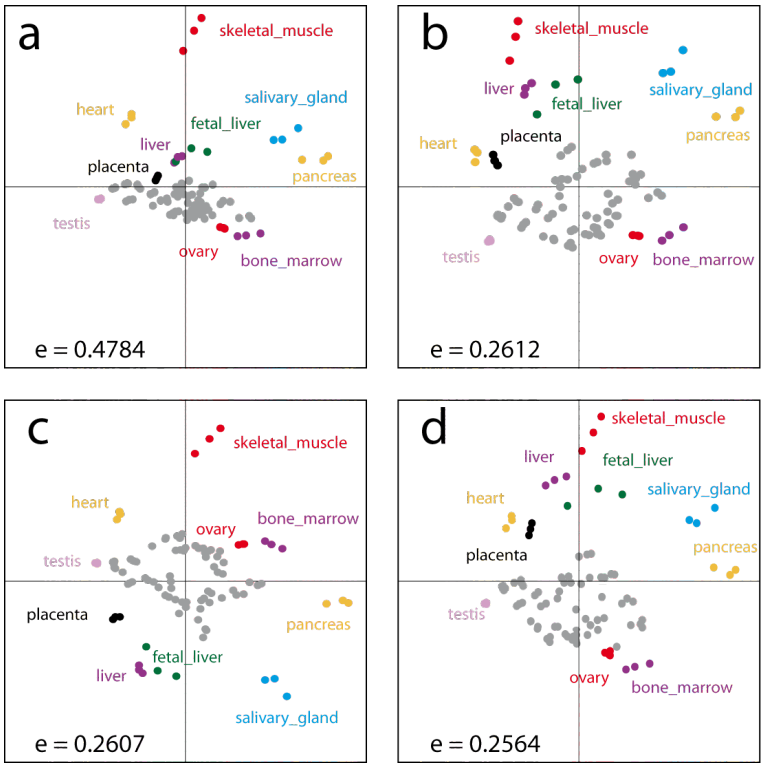


FIGURE 2.1 – *Comparison of the results of different dimensionality reduction techniques on the same dataset. The dataset "d1 — 96Cell", composed of ninety-six individual transcriptome profiles generated from thirty-two different human tissues (c.f. Table 2.1, and Section 2.2) was represented in 2D space using : (a) Singular Value Decomposition based on covariance, (b) random initialized Multidimensional Scaling ; (c) as in B using the same algorithm leading to a different random position matrix, and (d) Singular Value Decomposition-initialized Multidimensional Scaling. The peripheral data points were color coded and labeled according to the human tissue analyzed. The resulting Kruskal-Stress e for each of the dimensionality reductions is indicated. Similar computations were used to generate Table 2.2*

2.3 Comparison of different initialization methods

As already said in section 2.1, we postulated that the inconveniences associated with the MDS concerning the problems related to the choice of the initial condition for the simulation leading

to insufficient control of being trapped in local minima, as well as the large information loss when SVD techniques are used for dimensionality reduction, can be overcome when both methods are combined. We therefore created an SVD-MDS algorithm which uses SVD to compute the initial state of a molecular dynamics simulated MDS. This SVD-MDS approach was then compared to SVD and MDS on thirteen different datasets (see Table 1). Figure 2.1 well illustrates the shortcomings of SVD and MDS. The dataset "d1 — 96Cell" (see section 2.2 for a discussion of the different datasets used in this study) containing ninety-six different instances was used to compute a 2D representation using SVD (panel A), two examples of MDS initialized by random positions defining a 12 unit hypercube (panels B and C), and our combined SVD-MDS approach (panel D). The Kruskal stress was also computed for all four examples. As it is clear from the illustration and the Kruskal stress, MDS techniques (panels B-D) better preserve the distances between the instances and their relationship. The data cloud is better resolved and the global distance information loss (as estimated by the Kruskal stress) is lower than for SVD. Note that we chose to label only those tissues in the illustration that are sufficiently well resolved, all other instances are in gray color.

ID	Dataset Name	Metric	SVD	SVD-MDS		zeroMDS		stochastMDS	
			e	e	t	e	t	e	t
d1	96Cell	R^2	0.6472	0.3409	2500	0.352	2500	0.3478	2500
d2	96Cell_T	Cov	0.5001	0.1401	4500	0.146	4500	0.1503	4500
d3	Iris	Cov	0.0421	0.0344	509	0.0343	3554	0.0344	4059
d4	Wine	Cov	0.0010	0.0010	0	0.0064	4500	0.0061	4500
d5	Stochast 200	Cov	0.7513	0.4088	1500	0.4169	1500	0.4157	1500
d6	CCYier	Cov	0.1634	0.0765	400	0.0932	3500	0.1079	4500
d7	Pima	Cov	0.0964	0.0708	700	0.105	3500	0.1098	3500
d8	96Cell_T transposed	R^2	0.6954	0.1498	4500	0.1572	4500	0.1715	4500
d9	Secom	Cov	0.1801	0.1168	750	0.1217	4499	0.1283	4375
d10	Ozone	Cov	0.1223	0.0935	712	0.0935	2587	0.0951	2143
d11	Stochast 3000	Cov	0.9067	0.4353	130	0.4382	130	0.438	130
d12	Ecoli	Cov	0.1634	0.000	0	0.0202	4500	0.2484	4500
d13	Wave	Cov	0.2922	0.2132	324	0.2132	2252	0.2132	1998

TABLE 2.2 – Results from the different MDS algorithms applied to the various datasets (*c.f.* Table 2.1). CoV = covariance, R^2 = correlation, e = Kruskal Stress, t = time steps for molecular dynamics simulation in MDS.

In order to demonstrate generality of our approach we next analyzed the twelve other datasets (Table 1) using four different approaches : 1. using SVD only, 2. using SVD-MDS, 3. using MDS initialized with all data points placed at zero with minimal random noise (zeroMDS), and 4. MDS initialized with random positions (stochastMDS). The results of those analyses are reported in Table 2. In all cases, we reduced the dimensions to two. It becomes again apparent from the Kruskal stress that the MDS-based techniques systematically outperform the SVD. While stochastMDS, zeroMDS and SVD-MDS give similar results in terms of the final information loss, the number of time-steps needed to identify a minimum stress is greatly reduced using SVD-MDS (Table 2, and for four examples Fig. 2.2). Therefore, SVD-MDS approaches the final state (here defined as a Kruskal stress value) faster than either of the MDS methods. We show an example of stress evolution in Figure 2.4a where stochastMDS and zeroMDS are slow due to the existence of local minima, and SVD-MDS clearly outperform them. Basically, the initial steps in simulation

time need not be used to globally arrange the geometric object in space as in MDS, but already contribute to minimizing the Kruskal stress over the entire set of distances.

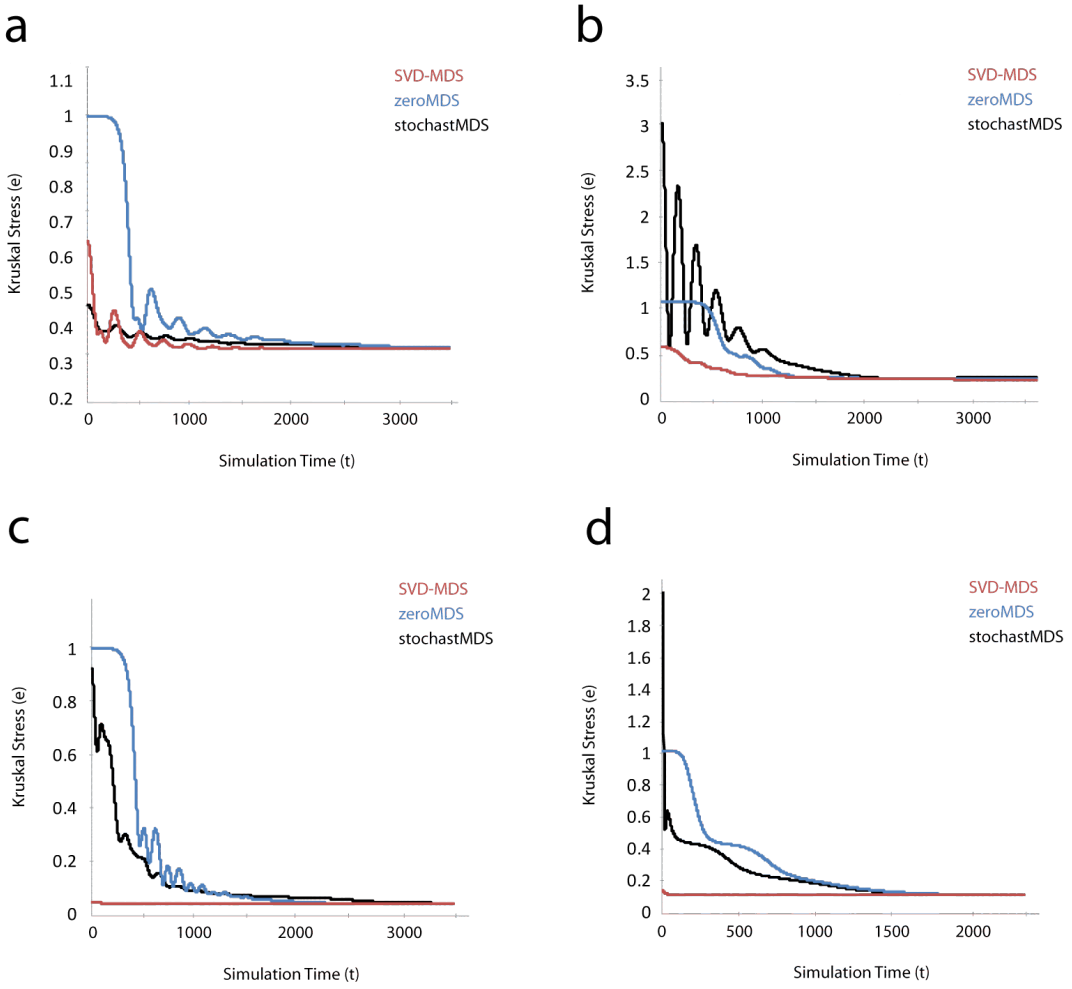


FIGURE 2.2 – *Comparison of the SVD-MDS, MDS initialized with all points in the center (zeroMDS), and MDS initialized by stochastic positions (stochastMDS) methods on different datasets (a) "d1 — 96Cell" in correlation basis, (b) "d2 — 96Cell_T" in covariance basis, (c) "d3 — Iris" in correlation basis, (d) "d10 — Ozone" in covariance basis.*

2.4 Iterative dimensionality reduction using iSVD-MDS

We next wondered whether the dimensionality reduction could be further improved by a step-wise reduction of one component after another, to this end we developed an algorithm called Iterative SVD-MDS (iSVD-MDS). The difference between SVD-MDS and iSVD-MDS will be that in the former one deletes all dimension of the SVD configuration and then reduces the stress using MDS, whereas in the latter one deletes each dimension one after another and reduce for each step the stress using MDS following by SVD for a proper reorganization of the configuration. Figure 2.3 shows the different steps of the algorithm iSVD-MDS with a comparison to SVD-MDS.

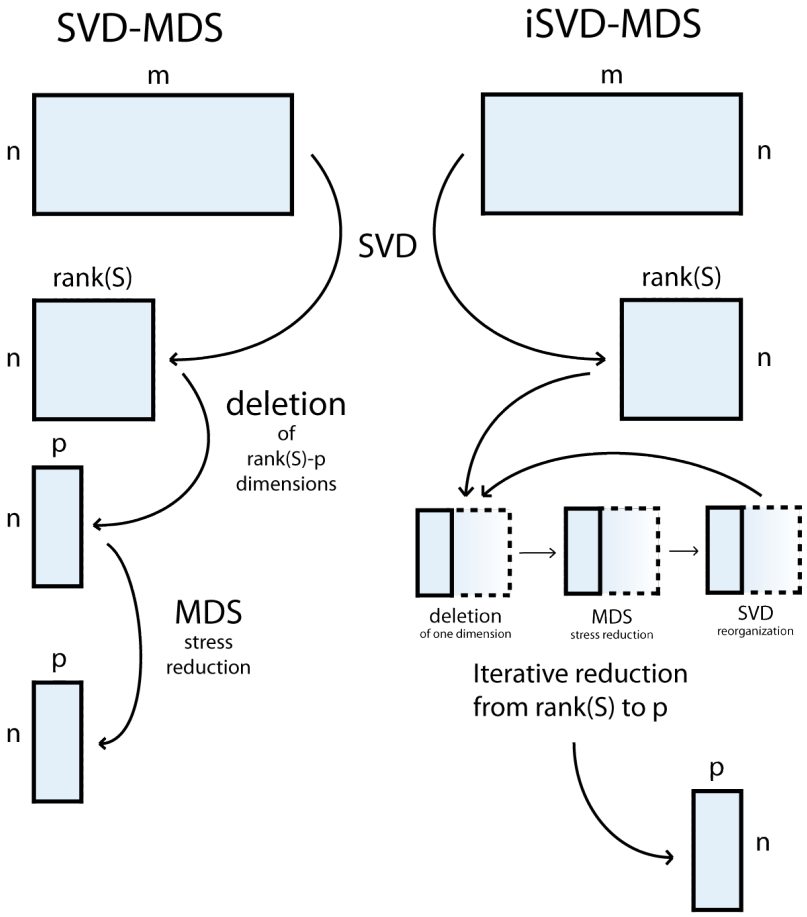


FIGURE 2.3 – *SVD-MDS and iSVD-MDS algorithms. Both schemes show the different steps encounter in the reduction of a n.m matrix to n.p, with rank(S) being the rank of the singular values matrix, thus the number of dimensions after performing SVD.*

We compared the performance of the three techniques SVD-MDS, MDS, and Iterative SVD-MDS (iSVD-MDS) on the different datasets. As can be seen in Figure 2.4a, SVD-MDS rapidly approaches a minimal Kruskal stress configuration over the simulation time. The previously described MDS procedure which uses stochastic initiation for the molecular dynamics simulation requires much more simulation time to find the same minimal stress configuration as the SVD-MDS algorithm. Finally, the iterative iSVD-MDS approach will also converge to the identical minimum obtained through the other methods, however, as for each component a separate simulation is performed the convergence time is greatly increased when compared to the former two methods. Albeit many different simulations on the different datasets we have never obtained a final configuration using iSVD-MDS were the Kruskal stress would allow to conclude on an improved performance when compared to SVD-MDS. Therefore, the iterative method does not allow for improved accuracy, but rather prolongs simulation time with no immediate gain (Table 3 summarizes the results). We next compared iSVD and iSVD-MDS methods to determine how the loss of information is distributed during iterative dimensionality reduction. As can be seen in Figure 2.4b, for both procedures the amount of stress or lost information increases both relatively

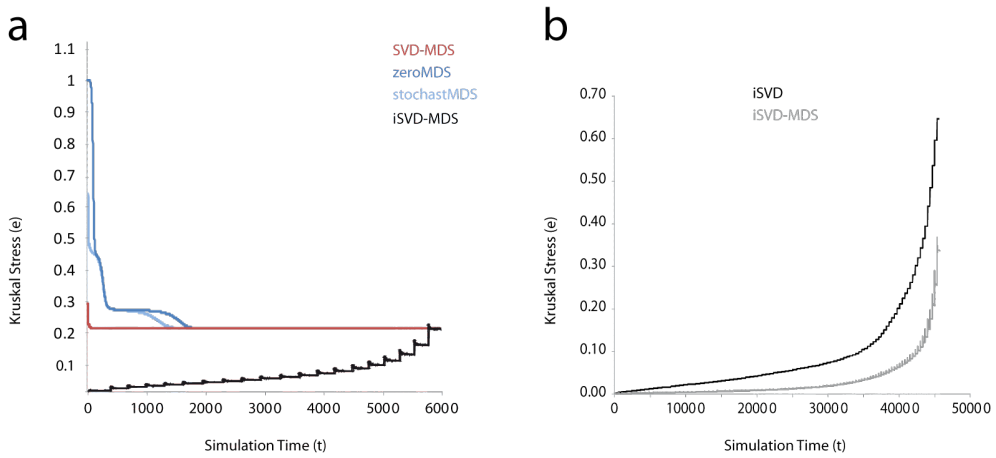


FIGURE 2.4 – *Iterative SVD-MDS algorithm assessment. (a) Comparison of the SVD-MDS, zeroMDS, stochastMDS, and iterative SVD-MDS (iSVD-MDS) methods on dataset "d13 — Wave" in covariance basis. (b) Comparison of the iterative SVD (iSVD) and iSVD-MDS methods on dataset "d1 — 96Cell" in correlation basis.*

and absolutely with the number of components removed. Note also, that the iSVD-MDS method better preserves at every consecutive iteration the distance information of the object (Figure 2.4b).

While computationally unattractive, and without any apparent bearing on the final result of the dimensionality reduction, the iterative iSVD-MDS method allows the analysis of the structural changes of the geometric object during dimensionality reduction. More precisely, and as seen in Figure 2.4, the absolute contribution of every single dimension can be evaluated. This might be of particular interest in the case of biological data, as the major distance information often is not necessarily restricted to a few dimensions. The iterative SVD-MDS method in conjunction with the *Entourage* parameter (see section 2.6) hence gives the user a potential control on when to stop the dimension reduction process. Based on the conjunction of iSVD-MDS and local structural control it might even be feasible to develop quantitative methods to define maximal compressibility at some defined distance information loss.

2.5 Molecular Dynamics dimensionality reduction with added stochasticity

In [39] an approach reminiscent of simulated annealing has been used to avoid getting trapped in local minima during the molecular dynamics simulation. Adding stochasticity to the molecular dynamics driven MDS is, after [39], required to insure reproducibility of the algorithmic performance. One can prove theoretically that by adding stochasticity convergence will improve, however, in practice this addition seems irrelevant as it will not lead to a different result.

To compare MD-MDS (SVD-MDS with stochasticity) with our SVD-MDS algorithm we have implemented different MD-MDS algorithms with stochastic energy. Our MD-MDS algorithm consists in adding a stochastic force to the MDS algorithm

$$F_{stochastic}(\bar{X}_i) = -T * s(t) \tag{2.2}$$

ID	Dataset Name	Metric	SVD-MDS		iSVD-MDS	
			e	t	e	t
d1	96Cell	R^2	0.3409	2500	0.3381	232097
d2	96Cell_T	Cov	0.1401	4500	0.1494	92536
d3	Iris	Cov	0.0344	509	0.0344	3008
d4	Wine	Cov	0.0010	0	9.0E-4	10003
d6	CCYier	Cov	0.0765	400	0.0753	22508
d7	Pima	Cov	0.0708	700	0.0692	27005
d8	96Cell_T transposed	R^2	0.1498	4500	0.1525	122059
d10	Ozone	Cov	0.0935	712	0.0935	66031

TABLE 2.3 – Results from iSVD-MDS algorithms and SVD-MDS algorithm applied to the various datasets (*c.f.* Table 2.1). CoV = covariance, R^2 = correlation, e = Kruskal Stress, t = time steps for molecular dynamics simulation of the MDS.

Where $s(t)$ is a random number generated here uniformly between -0.5 and 0.5. We thereby chose to linearly ("lin") and exponentially ("exp") (as in [39]) remove this extra energy from the system over simulation time. We thus use two types of temperature-decrease, the first linear, beginning with a temperature of $100J$ and decreasing linearly to $0J$ during 3000 steps of simulation ; we call this method MD-MDS linear. The second includes an exponential decrease from $100J$ to below $0.1J$ during 3000 steps of simulation ; we call this method MD-MDS exponential.

As can be seen in Figure 2.5, SVD-MDS as well as the two MD-MDS algorithms "lin" and "exp" always identify final configurations with the same amount of residual energy (*i.e.* stress). It can also be seen that SVD-MDS converges faster for these four examples than the MD-MDS methods. In Table 2.4 I show that both statements hold for the entire set of analyzed data.

ID	Dataset Name	Metric	SVD-MDS		MDMDSlinear		MDMDSexpo	
			e	t	e	t	e	t
d1	96Cell	R^2	0.3409	2500	0.3453	5500	0.3421	2500
d2	96Cell_T	Cov	0.1401	4500	0.1465	4500	0.1542	4500
d3	Iris	Cov	0.0344	509	0.0359	4500	0.0343	4000
d4	Wine	Cov	0.0010	0	0.0089	4500	0.0067	4500
d5	Stochast 200	Cov	0.4088	1500	0.4092	4500	0.4089	4500
d6	CCYier	Cov	0.0765	400	0.1346	5500	0.1162	5500
d7	Pima	Cov	0.0708	700	0.1128	5500	0.0986	5500
d8	96Cell_T transposed	R^2	0.1498	4500	0.1832	4224	0.1822	4500
d9	Secom	Cov	0.1168	750	0.1511	5500	0.1396	4500
d10	Ozone	Cov	0.0935	712	0.0944	4500	0.0951	3500
d11	Stochast 3000	Cov	0.4353	130	0.4353	200	0.4353	200
d12	Ecoli	Cov	0.0	0	0.312	5500	0.2273	5500
d13	Wave	Cov	0.2132	324	0.2132	3671	0.2132	2203

TABLE 2.4 – Results from the different MD-MDS algorithms and SVD-MDS algorithm applied to the various datasets (*c.f.* Table 2.1). CoV = covariance, R^2 = correlation, e = Kruskal Stress, t = time steps for molecular dynamics simulation.

We next asked whether or not similarly adding stochasticity to the SVD-MDS algorithm

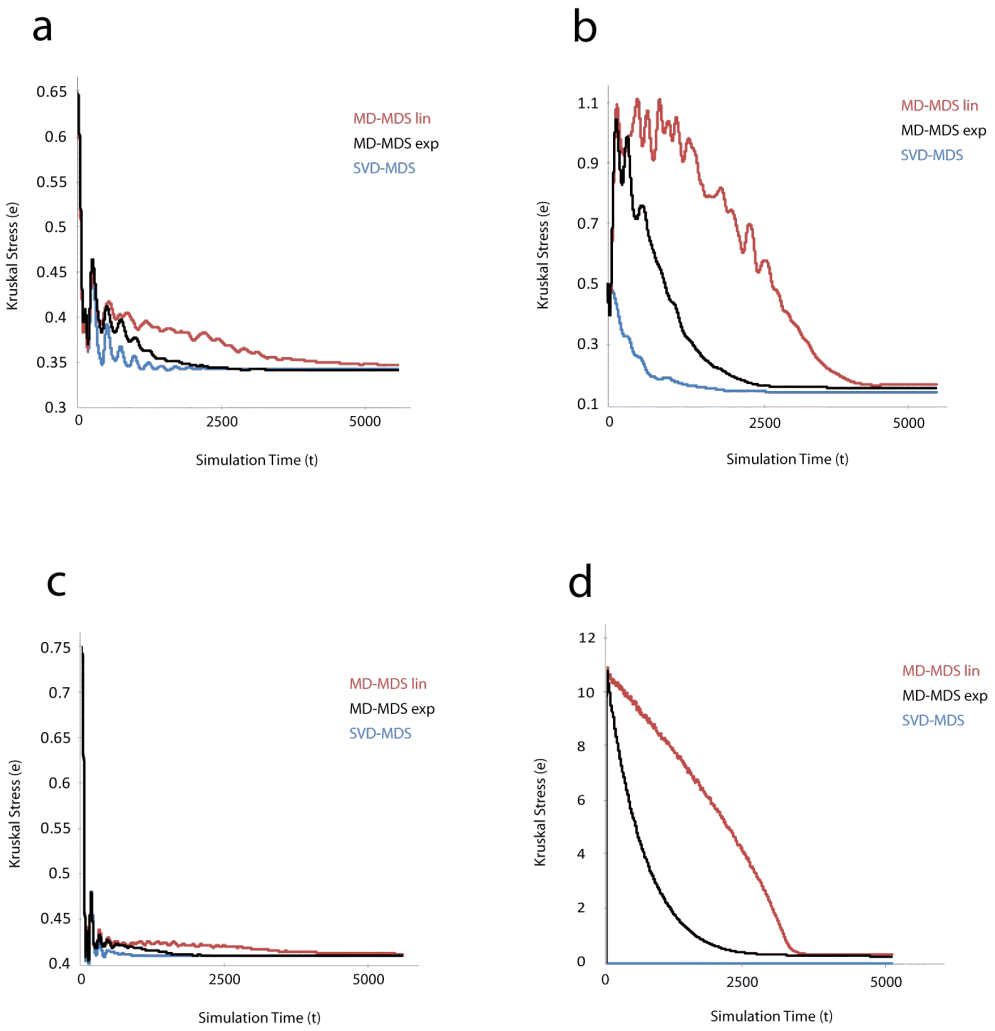


FIGURE 2.5 – *The influence of stochasticity on MD-MDS algorithms. Comparison of the SVD-MDS, Molecular Dynamics linear MDS (MD-MDS linear), Molecular Dynamics exponential (MD-MDS exponential) methods on different datasets (a) "d1 — 96Cell" in correlation basis, (b) "d2 — 96Cell_T" in covariance basis, (c) "d5 — Stochast 200" in covariance basis, (d) "d12 — Ecoli" in covariance basis.*

would improve its performance. Figure 2.6 illustrates the results we have obtained on three different datasets. Indeed, adding different amounts of energy at different times of the simulation (as indicated in the panels by arrows) does not lead to the "discovery" of lower energy minima during the simulation procedure. The SVD-MDS algorithm, similarly as the MD-MDS algorithms (Figure 2.5) always converges to the same energy state. This has been confirmed using other datasets with identical results (data not shown). Taken together, the results using MD-MDS-lin and MD-MDS-exp and SVD-MDS raise the question of whether indeed several minima exist or only a single ground-state is to be found. While we do not have any formal proof of the latter, we believe that the detailed analysis of the geometric structure of the data objects presented below strongly argues in favor of a global energy minimum.

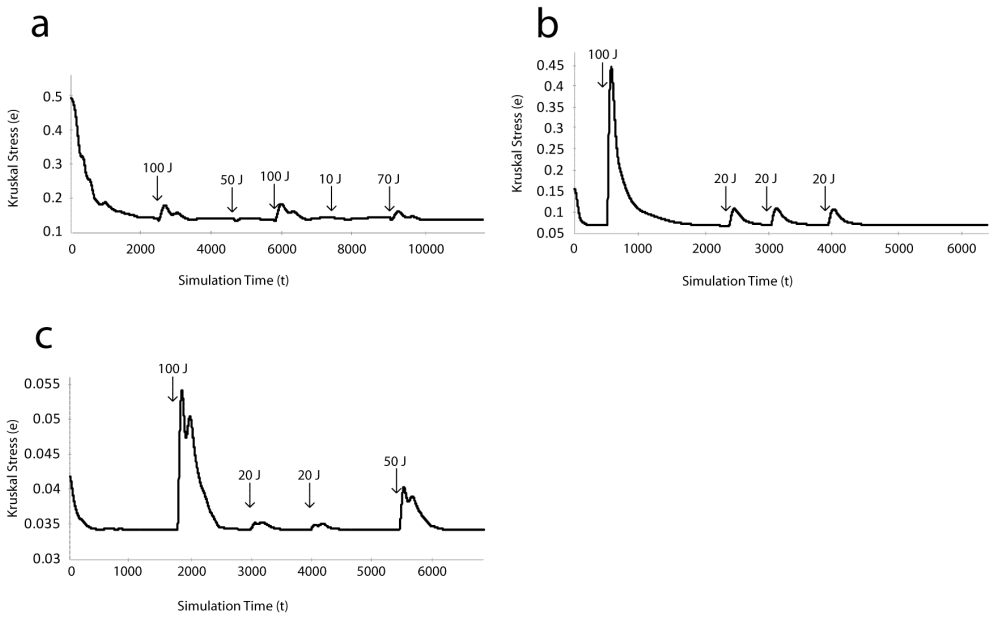


FIGURE 2.6 – *Robustness of SVD-MDS simulations. Evolution of stress over simulation time with injection of energy, on different datasets (a) "d2 — 96Cell_T" in covariance basis, (b) "d2 — CCYier" in covariance basis, (c) "d5 — Iris" in covariance basis.*

2.6 Geometric final structure assessment

Kruskal stress rather evaluates global deformation of the cloud of instances,as it is dependent of big distances [36]. Consequently it does not give any indication on how local distances are affected by the dimensionality reduction. In order to quantify the faithfulness of any representation of data in low-dimensional space it is thus required to define a new parameter.

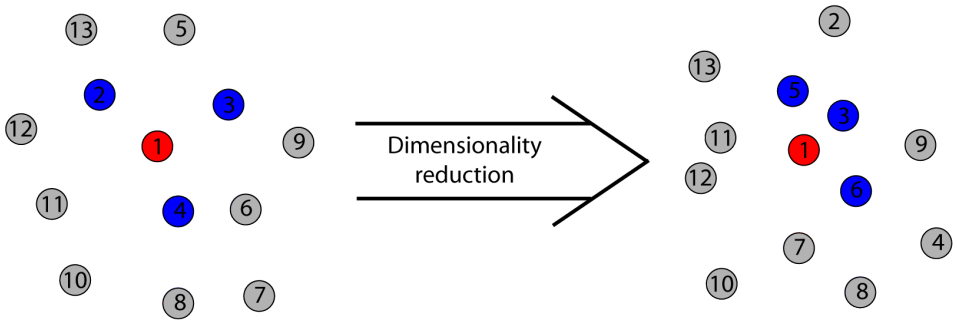


FIGURE 2.7 – *Principle of Entourage parameter. For each point (here the point with index 1, in red) one evaluates the change in the k (here k=3) nearest neighbors (in blue), and counts the number of common neighbors (here $G_i = 1$).*

This new parameter, *Entourage*, we chose to define is based on an analysis of the change in k nearest neighbors which evaluates local organization change. SVD leads to a undistorted representation of the instances in rank(X) dimensions, and can be used as reference representation

of the data points. Any representation in reduced dimension will have to be the most similar to this reference representation. For any one instance $\bar{X}_i$ in the reference distribution obtained through SVD we consider its k nearest neighbors : N_i^{ref}. In the new distribution obtain after dimensionality reduction, we also compute the k nearest neighbors for the same instance $\bar{X}_i$, and obtain a list : N_i^{new}. We then search for $G_i = card(N_i^{ref} \bigcap N_i^{new})$, which will be the number of instances common to those two lists. We repeat this operation for all instances i, and obtain the *Entourage* parameter

$$Ent_k = \frac{\sum_{i=1}^{n} G_i}{G} \tag{2.3}$$

with $G = nk$ a normalization parameter ($Ent \in (0, 1)$).

If $G_i = card(N_i^{ref} \bigcap N_i^{new}) \approx 0.01 card(N_i^{ref}) = 0.01k$ for every i then $Ent_k \approx \frac{0.01 \sum_{i=1}^{n} k}{nk} = 0.01$, a difference of 1% between two values of *Entourage* mean an average deformation of 1% in the local organization.

This parameter has only signification for a low number of considered neighbors k compared to the total number of points n. We arbitrary choose k=0.1 of n as the number of nearest neighbor instances considered, we obtained a good evaluation on how well the local organization is conserved.

A second type of analysis of the geometric properties of the objects under study can be developed by analyzing the behavior of the *Entourage* parameter, defined as the relative change in the k nearest neighbors, as a function of the k nearest neighbors considered. We have plotted the relationship of *Entourage* and k for six different methodologies : zeroMDS, stochastMDS, SVD-MDS, iSVDMDS, MD-MDS-lin, MD-MDS-exp in (Figure 2.8) for nine different datasets which represent the different behaviors one can observe. From the selected examples it becomes clear that again the SVD-MDS method outperforms the different types of MDS over a wide array of structures analyzed as the *Entourage* value is consistently higher no matter how many different k nearest neighbors are considered. The iterative iSVD-MDS method, due to the accumulation of small residual errors during the molecular dynamics simulation, and the MDS method give similar results. At the cost of increasing computational load, the iSVD-MDS better and better approximates the SVD-MDS method. In conclusion, the SVD-MDS method, under all conditions tested, better represents the geometric structure of the datasets in low-dimensional space when compared to the input object with $rank(S)$ components. Note that this holds even for objects with equal stress.

Figure 2.1 illustrates the problem of rotational variance when using stochastically initiated molecular dynamics simulations for MDS. It becomes apparent, when comparing panels B and C as well as comparing them to panel A and D that stochastMDS can result in different final configurations. The stochastMDS algorithm produces two near-optimal solutions (with respect to the Kruskal stress), the resulting orientation of the instances, however, is different (focus for instance on the relationship between "skeletal muscle" and "fetal liver"). The problem arising, if stochastMDS can lead to different representations despite using the same parameters for computation, is accuracy of the representation. Strikingly, SVD-MDS on the contrary only produces a single result. This observation, taken together with the results on the relevance of stochasticity in the simulation obtained above, argues for the existence of different equivalent energy minima that only differ in the rotational orientation of the object and at best only minimally in the Kruskal-stress. Therefore, taken together SVD-MDS not only reduces significantly the computational load when compared to other methods, but also insures uniqueness of the resulting representation as there is no randomness in the process unlike in the other methods used for initialization. The quality of this final and unique representation can be demonstrated using the

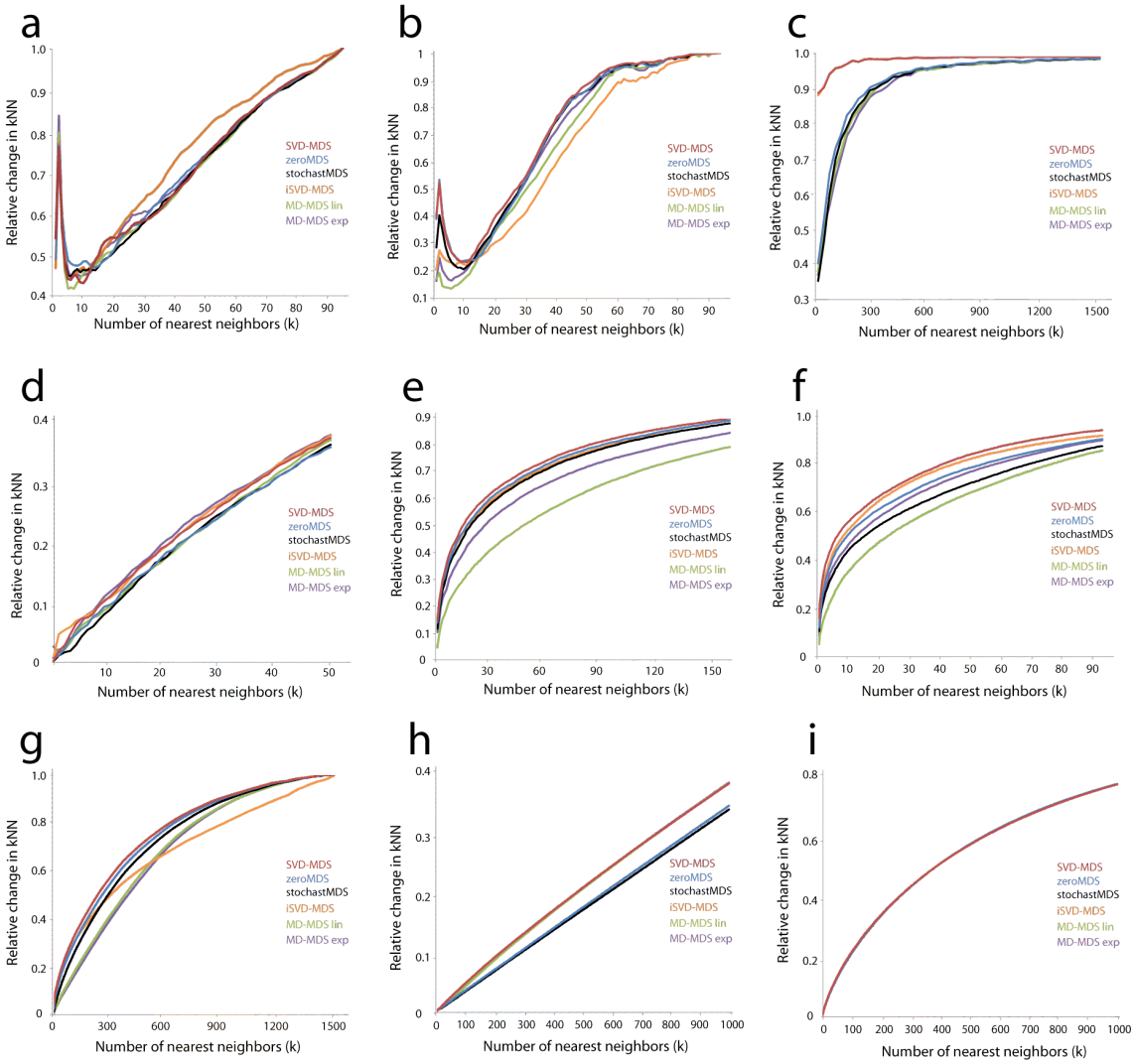


FIGURE 2.8 – *Relative changes in k nearest neighbors (Entourage) are local, structural measures of dimensionality reduction and thus assess quality of the procedure. As a function of the number of nearest neighbors k considered, the relative change in kNN between the initial high-dimensional space and 2D space is plotted for all the methods SVD-MDS, zeroMDS, stochastMDS, iSVD-MDS, MD-MDS linear, and MD-MDS exponential. The datasets used are in : (a) "d1 — 96Cell" in correlation basis, (b) "d2 — 96Cell_T" in covariance basis, (c) "d4 — Wine" in covariance basis, (d) "d5 — Stochast 200" in covariance basis, (e) "d6 — CCYier" in covariance basis, (f) "d7 — Pima" in covariance basis, (g) "d8 — 96Cell_T transposed" in correlation basis, (h) "d11 — Stochast 3000" in covariance basis, (i) "d13 — Wave" in covariance basis.*

Entourage parameter, as shown here. This increase in fidelity in the representation of data should not be underestimated (see Figure 1). Stress values should not be considered as the only relevant parameter to determine the performance of dimensionality reduction techniques as local and global structure considerations can effectively, as demonstrated here, allow to judge fidelity of the final representation. This is reminiscent to techniques of principal manifold searches (see section 1.2.7 were parameters describing topology, local organization or other geometric characteristics

are used.

3

Global Pathway Analysis.

In order to understand the molecular biology of the cell [73] one has to decipher the different networks [74] of interactions within each cell. For each specific scale a specific network of interactions exist, for example genome or transcriptome networks. Of course each one of this "ome" scales is connected to the other scales. Consequently, systems biology, which is the science seeking the identification of these networks, has two major goal nowadays : (i) the discovery and analysis of the different "omic" networks, (ii) the identification of connections between these networks.

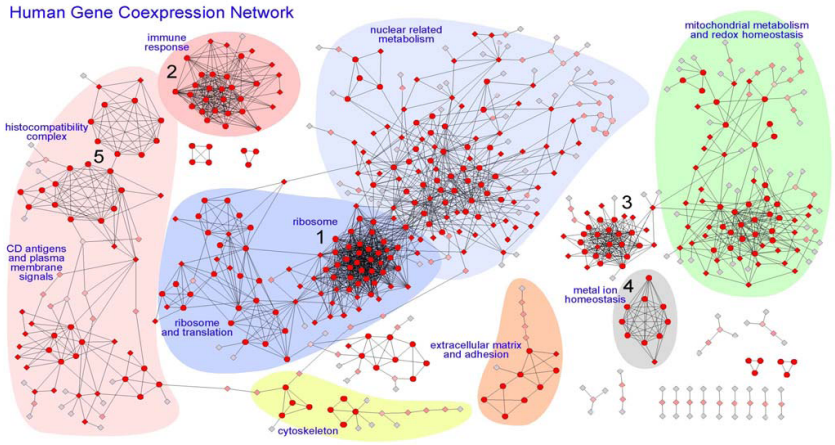


FIGURE 3.1 – *Co-expression network created by Prieto et al. of* 615 *genes, Figure extracted from [75].*

In the case of the transcriptome scale, which is the one I have studied during my thesis, many works have already been done to infer gene expression networks. Different types of networks exist, depending on the way of defining links (*e.g.* correlation) within nodes, and the most prevalent ones are gene co-expression networks because they are the simplest to obtain. They consist in looking at a set of genes and define a value of correlation between them whenever the corresponding expression values show similar characteristics. One great example of this kind of network is given by Prieto *et al.* in [75], where they have created a network of 3327 gene-nodes and 15841 co-expression links, using microarray transcriptome data. I show in Figure 3.1 this network for a selection of the 615 most highly connected nodes.

One interesting idea which has been developed in the past years, is to map the information of

gene expression to protein-protein interaction networks. In order to correlate the transcriptome scale to the proteome scale, different software has been developed such as Cytoscape [76], which allows to retrieve protein-protein interaction networks from popular databases and map gene expression information onto it.

In the host group, we have developed a tool named LEO which indicates for a set of microarray probes and their expression values the under and over-represented pathways of the analyzed biological condition. A pathway is a network of chemical interactions constituting major metabolic processes within a cell. Figure 3.2 provides an example of the pathway for 2-arachidonoylglycerol biosynthesis, in this specific case, the pathway consists only of protein and molecule interactions, but pathways can be constituted of a vast class of different species. It exist several databases which provide information on the different pathways. One can cite PANTHER [77] (Protein ANalysis THrough Evolutionary Relationships) which provides 165 different pathways, and also tools for correlating expression of genes to pathways [78]. The LEO tool is in fact derived from the Panther gene expression tool (see section 3.1.2).

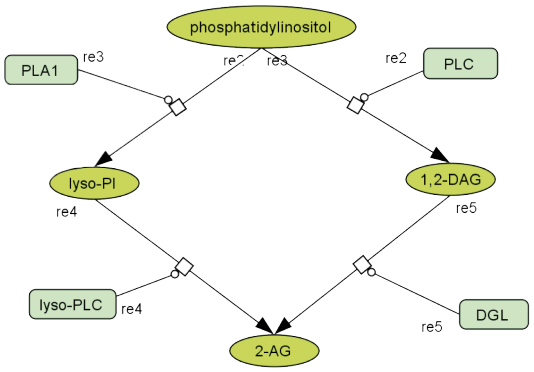


FIGURE 3.2 – *Pathway diagram of the pathway 2-arachidonoylglycerol biosynthesis.*

Despite the fact that our tool gives interesting information on the important pathways presumably active in any biological condition, it does not provide a network information on the pathways' components over-represented in a given gene expression setting. In addition, protein-protein networks of human cells are usually limited to a pathway or a reunion of few pathways. In consequence, we decided to correlate the different information of pathways given by Panther and to create a global pathway network of chemical interactions onto which the gene expression information can be mapped.

In order to do so, I developed *ex nihilo* a software which has two roles : (i) creation of the global interaction network using all the pathways downloaded from Panther, (ii) map the gene expression and LEO information onto this network. Figure 3.3 shows, as example, results we have obtained using this tool. The software does not contain challenging algorithms, the biggest difficulty here was to manipulate tables properly in order to assure having always the most accurate biological information. That is why in this following chapter, after a development of the principle of Panther and Leo, I will explain how I created the software focusing on the different table manipulation algorithms developed

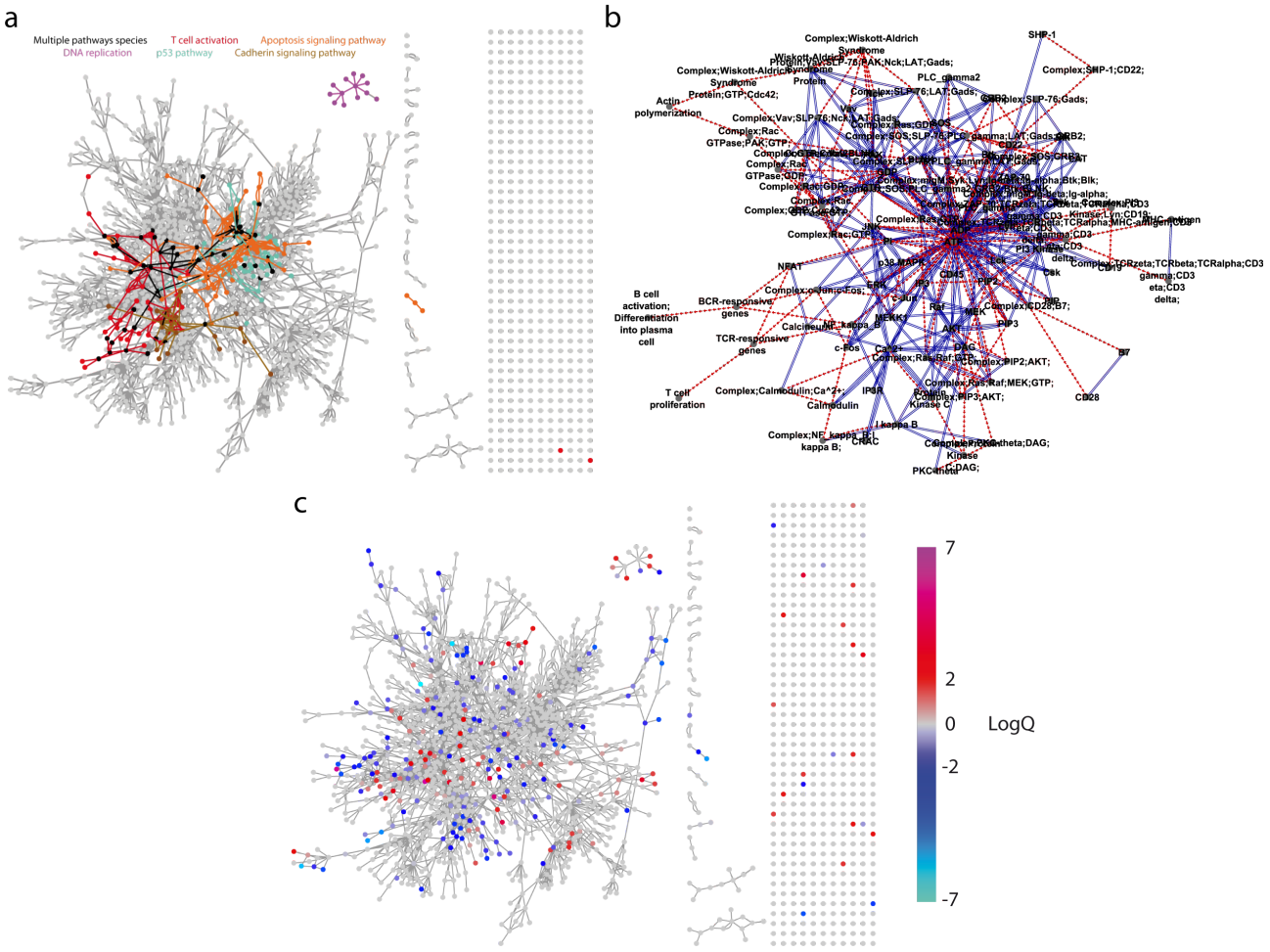


FIGURE 3.3 – *Example of results given by Pathway Analysis. (a) The different over-represented pathways found in activated Treg (see section 4.3.3) are colored on the GSP (Global Signaling Pathway) network (see section 3.2.6). (b) A network representation of the common components between B and T cell activation pathways. (c) Gene expression mapping of logQ information between Tconv and aTreg (see section 4.3.3) onto the ProbesNetwork (see section 3.2.6).*

3.1 Pathway Analysis

The two algorithms described below are based on the same principle, given a set of probes, over or under-represented ontologies compared to a random reference state are identified. Ontologies are of three types : Pathways, biological processes, or molecular functions.

3.1.1 PANTHER

The first goal of Panther is to provide to the science community a list of curated pathways taken from databases or entered on the Panther website by biologists. These pathways can be visualized directly on the website using a java applet or downloaded in SBML (systems biology markup language [79]) format, which is a standard format for systems biology data. These pathways in SBML format are enclosed in an XML file and can be opened by several

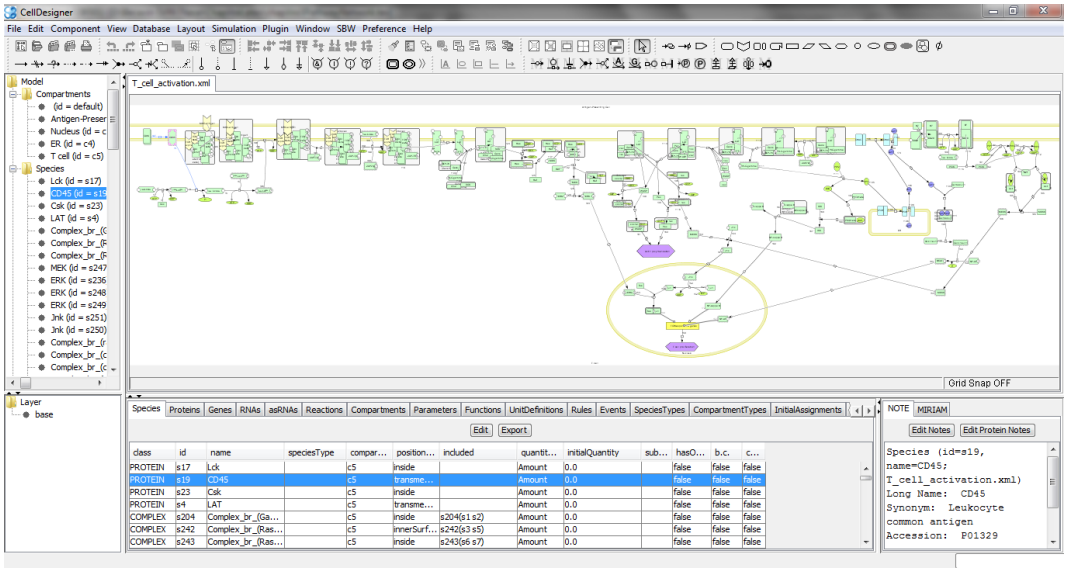


FIGURE 3.4 – *Printscreen of CellDesigner [80, 81] results for Tcell activation pathway.*

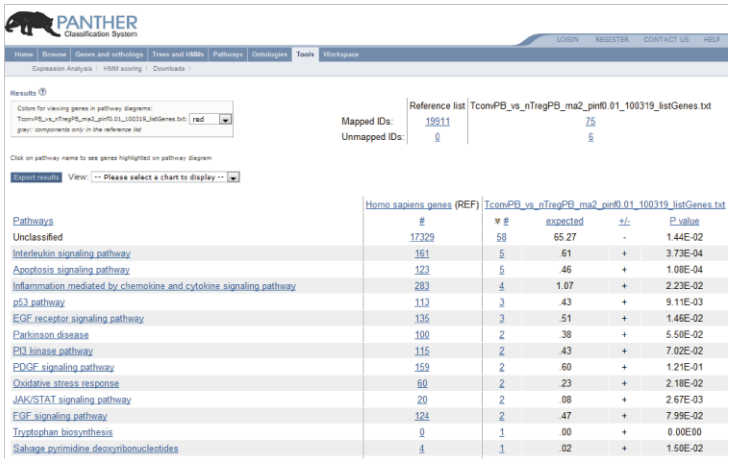
softwares such as Cell Designer [80]. This latter software helps to visualize and design pathways ; an example of utilization for the T cell activation pathway is provided in Figure 3.4.

The reason why we use Panther more than other similar websites is that it also provides links between Applied Biosystems microarray probes (the technology used in the laboratory) and the different ontologies considered on their website. They have defined internally for each ontology a list of Applied Biosystems probes which are relevant to it. They have also provided a tool (see Figure 3.5-a) which takes a set of genes provided by the user and compares it to the list of genes for each ontology they have defined. Thus, it indicates for these ontologies how much genes are in the reference file and how much in the file provided by the user. It calculates, given the number of genes present in the file and the known number of genes for each ontology, the number of genes expected, a binary "+/−" value indicates if the ontology is then over-or under-represented. Finally, a p-value indicates the relevance of the results in comparison to a random set of genes. This p-value associated to the ontology is calculated on a binomial law considering the values described just above. The parameters of the binomial law are : (i) number of success, thus the number of observed genes, (ii) the number of samples, thus the size of the submitted list of genes, and (iii) the probability of success which is the reference number of genes value divided by the reference probe number. This analysis can be enriched using values of expression or fold change between two values of expression (see Figure 3.5-b). The p-value is then corrected using a Mann-Withney-Wilcoxon statistical test.

3.1.2 Ace.map LEO

The LEO tool developed in the host laboratory gives the same information but it is based on probe information and not genes and it uses correction parameters to enhance the results. The corrections take account of the number of ontologies to which the considered probe is annotated and weight the value of the probe in proportion. The idea is to give less relevance to probes that are present in many ontologies. For each probe one applies a weight $W = 1/N$ with N the number of pathways to which the probe is annotated. This correction makes the analysis more

a



b

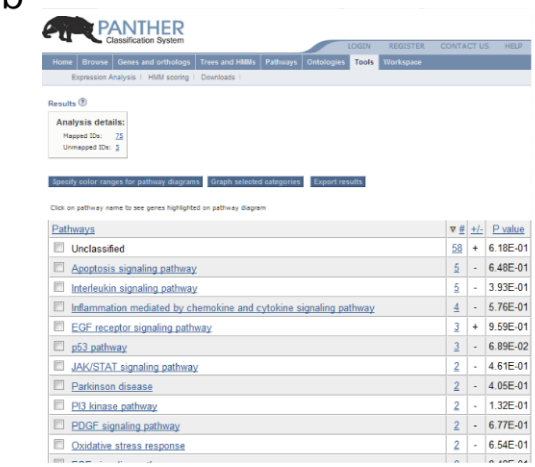
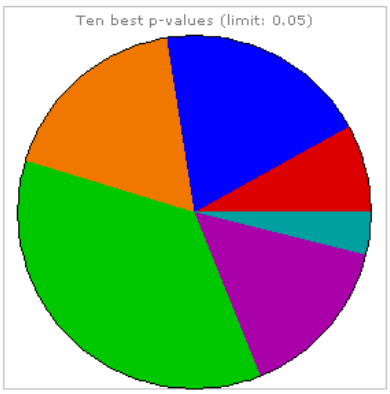


FIGURE 3.5 – *Example of results given by PANTHER [77] when providing 82 genes from TconvPB versus nTregPB and the corresponding logarithmic fold changes (see section 4.3.3 for more information on these data). (a) Comparison of the list of genes to the total list of genes of the human species. (b) List of preferential pathways corrected using PANTHER expression tools [78] on fold changes given in the data.*

specific to pathways. A pie chart is also provided with each analysis to illustrate the proportion of probes observed divided by the number of expected probes that are part of the ten ontologies having the least p-value.

In Figure 3.6, I show an example of LEO results on microarray data. The pie chart shows that DNA replication is the pathway which has the maximum number of probes observed in comparison to the number expected. The table below this chart has 5 columns :

- "Ref" column which shows the number of probes enclosed in the current ontology.
- "Obs" column indicates the number of probes actually observed.
- "Ref" column shows the number of probes expected.
- "+/-" column indicates if $Obs > Exp$ or $Obs <= Exp$
- "pValue" column shows the p-value enriched using corrections.



Ontologies	Ref	Obs	Exp	+/-	pValue
null	29917.00	3900.00	4033.69	-	9.01E-8
T cell activation	207.00	65.00	27.91	+	1.15E-7
Apoptosis signali...	187.00	54.00	25.21	+	3.35E-5
DNA replication	26.00	15.00	3.51	+	5.18E-4
p53 pathway	169.00	41.00	22.79	+	2.62E-2
Cadherin signalin...	220.00	13.00	29.66	-	4.34E-2

FIGURE 3.6 – *Results of a LEO analysis with Tconv versus aTreg file, see section 4.3.3 for more information on this data.*

The results of LEO analysis will be used in our software to highlight pathway components in the global pathway network.

3.2 Development of Global Pathway Analysis software

3.2.1 Overview

The idea behind the software I have developed (see Figure 3.7) is to make the link between gene expression and pathways. This will be done by first listing all species (*i.e.* pathway components) and interactions within each pathway, then constructing a "global pathway network" or network of a selection of interactions, and finally map onto the network the information of gene expression coming from the microarray experiments. For this reasons the software contains three separate parts :

- First, information contained in each pathway (see section 3.2.2) is used to create a list of all species and interactions. The major difficulty here was to be sure to obtain a reliable and biologically accurate list of species. The problem being that all pathways, even if they have been curated, are created by human users, and consequently, between two different pathways different names for the same species exist. Even sometimes acronyms for designing species are used. That is why I have developed an algorithm to correct these problems and to obtain an accurate list. This part of the software will be described in section 3.2.3 ;
- Second, a network needs to be constructed. This network should involve all the components of the pathways that can be identified, or selections of species within these pathways. Every possibility can be considered depending of the analysis, that is why I developed a solid user interface using the Java Rich Client Interface (java RCP) which allows to construct software easily with all user interfaces one might need. As shown in Figure 3.7, the software has

three parts, one for checking the files created during the analysis in the workspace, one for choosing elements for the network, and one for the console. The second one is the most important, as it allows to choose pathways, species, or reactions that will be included in the network construction. This part of the software will be described in section 3.2.6 ;

– Third, probes have to be connected to network nodes. This part was the most challenging as it consists in connecting all species of the different pathways to the corresponding probes. I have developed automatic association algorithms which I explain in section 3.2.5, but half of the associations were manually curated with great help of Sylvia Maiella, a former PhD student of Lars Rogge's group. I implemented import and export tools in the software to load expression data and LEO analysis data and export the network constructed using the software.

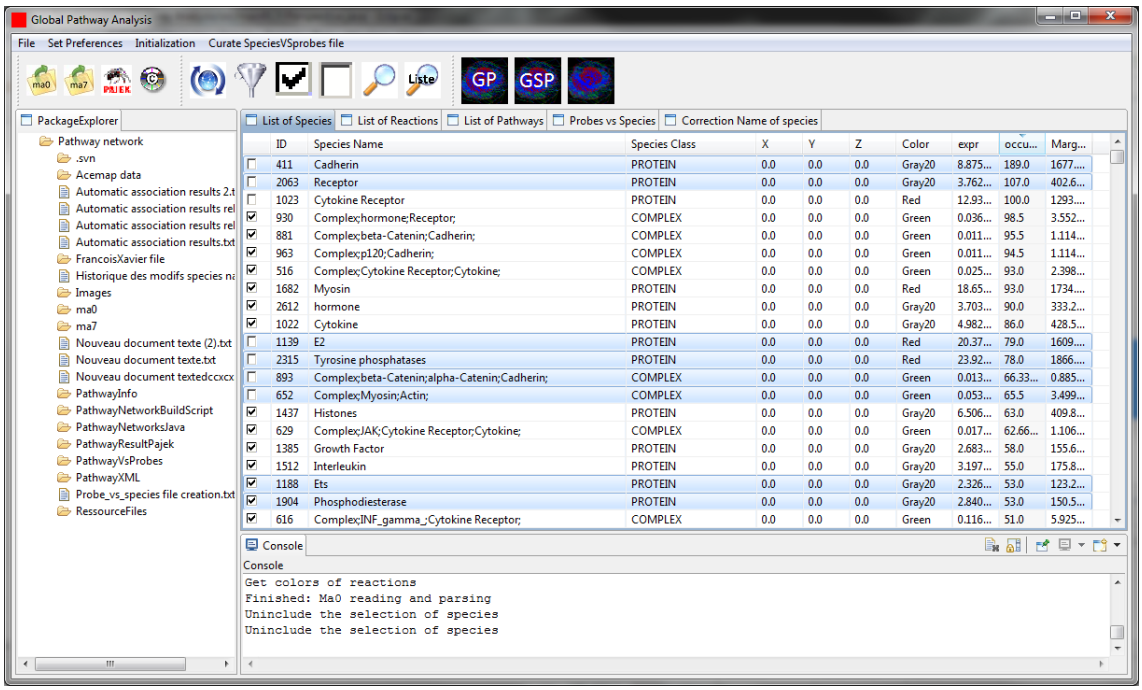


FIGURE 3.7 – *Printscreen of Global Pathway Analysis software.*

As all the algorithms of the software consist in manipulation of files, I extensively use preferences tools enclosed with RCP classes in order to save all paths used in the software. The software was developed in Java object oriented computer language in order to obtain a multi platform application.

3.2.2 Retrieving the pathways components

All data manipulated in the software stem from XML pathway files. In this object oriented file, formatted in the SBML format, there are many types of information, the most useful being the list of components of each pathway (named species) and the list of chemical reactions between these species. Consequently in the software all 165 XML pathways files, downloaded from Panther, are parsed. Two types of tables are extracted for each pathway : one giving the list of species with their characteristics (on the left of Figure 3.8), and another giving all reactions (on the right of Figure 3.8). Another table is extracted corresponding to the list of proteins. This last table is

only used in the software to retrieve all the species of a pathway.

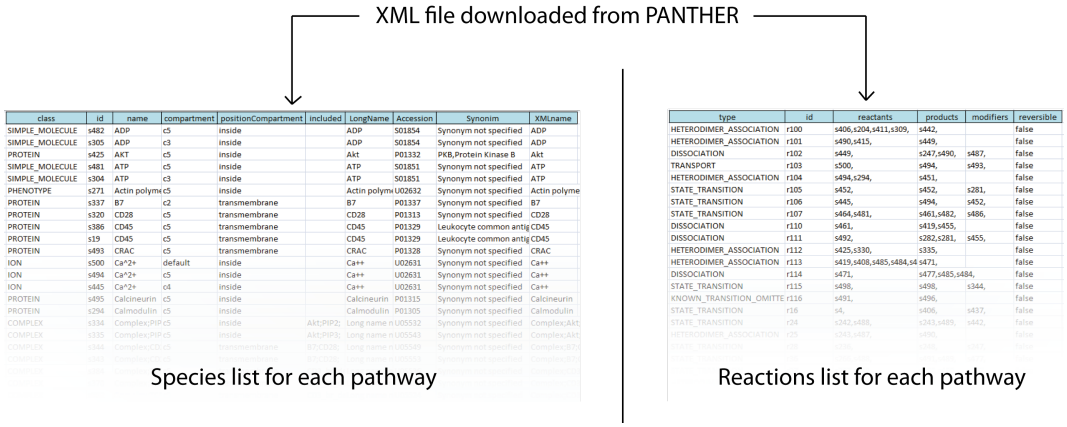


FIGURE 3.8 – *Principle of the XML parsing operation. (left) the table of species retrieved from the XML file. (right) the table of reactions retrieved from the XML file*

In these tables, the major information which needs to be obtained for each specie is its class (*e.g.* protein, molecule, ion, complex of protein, etc.), their id in the pathway, their name, their LongName which is the name given with more details, and the synonym which is a list of names that can also apply to the current species. All other species properties are not used by the software.

For the reactions, the information used by the software is only the list of reactants, products and modifiers for each reaction. I will show in the next sections how from the software will reconstruct the network with these two types of table.

3.2.3 Retrieving the species list

Once all the species tables have been created for all pathways, the software then creates a total list of species, which is a union of all species found in the 165 pathways. As each pathway has been implemented by different users, and so not all pathways have the same term for naming the same species a unique nomenclature has to be developed. Also, some terms derived from the graphical representation of the pathways are not well parsed. For example the "*_space_*" term is in CellDesigner automatically transformed to a blank space, whereas in the parsed species table it might be left unchanged despite the fact that I have implemented graphical information parsing functions in the software. To correct all these problems, the software uses a table (bottom-left of Figure 3.9) which indicates the new names to be given to species as required. This table has been created and updated manually by Sylvia Maiella and myself.

When users create the different pathways, some have added annotations to it, that it to say indications on some parts of the pathway which are considered as species by my software but are in fact biologically irrelevant. To delete these species the software uses a table file (bottom right of Figure 3.9) which has been manually created. This list of species to be deleted have been obtained by looking at species with are not in a reaction, because in all pathways all species should be a part of at least one reaction.

Finally, some species do not have the right classification ; for example species classified as *SIMPLE_MOLECULE* should be in the *ION* class. Also, as the goal of the software is to map gene expression probes to a network of pathway components, one gene and the corresponding

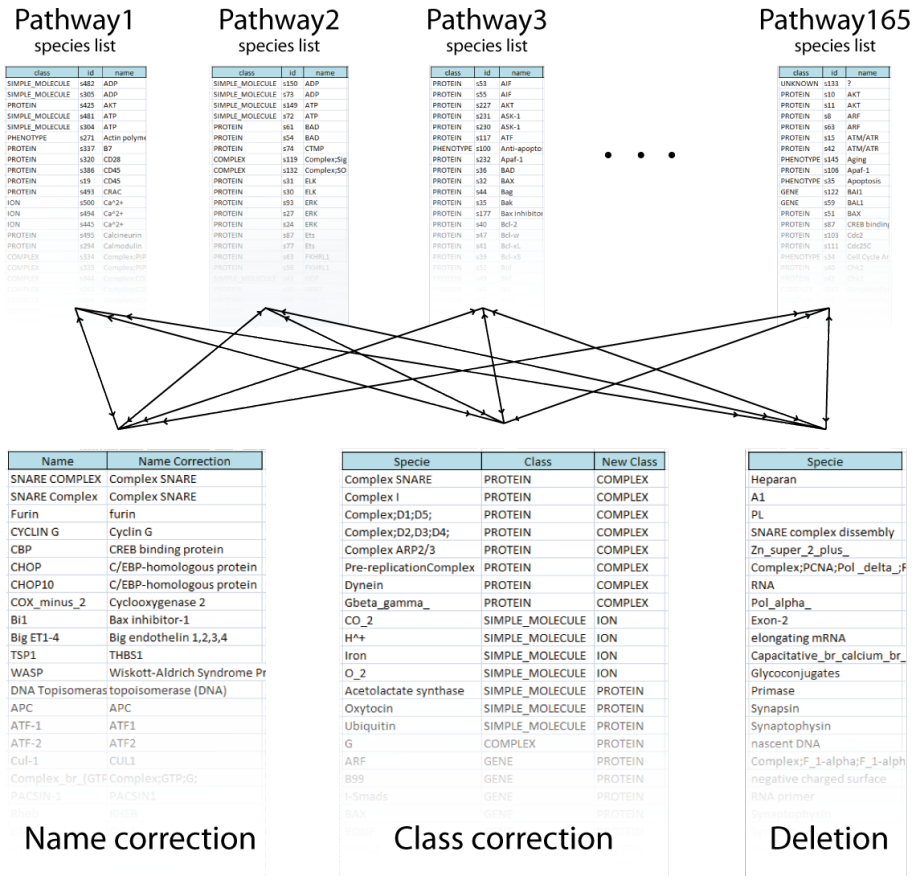


FIGURE 3.9 – *Principle of the species correction step. For each species table the software checks in the three system tables if some species have to be modified. The first table (bottom left) is used to change the names of certain species, the second (bottom center) is used for changing the class of certain species, and finally the last one (bottom right) is used for deleting some irrelevant species.*

protein should be considered has the same node in the network. Thus some species which are present with the two classifications : *GENE* and *PROTEIN* are modified to be part of only the *PROTEIN* class. Finally, for some "virtual" species, such as "gene transcription", their classification was changed into *ABSTRACTGENE*. All these class changes are performed by the software using a manually created table (bottom center of Figure 3.9).

To summarize, all changes in the species names and class tables are performed by going through all 165 species tables, and by modifying directly these tables accordingly. Figure 3.9 summarized this table manipulation for species correction. The former name extracted from the XML file of a modified species is conserved in the last column of the table.

Before getting the total list of species one has a specific type of species to look closer at : the *COMPLEX* class which correspond to complexes of species. Their name directly indicates which type of species are contained within. Consequently, the software then modifies the name of these complexes using different information.

There are different steps in the complexes name correction algorithm, which is run just after the species name and class corrections described above.

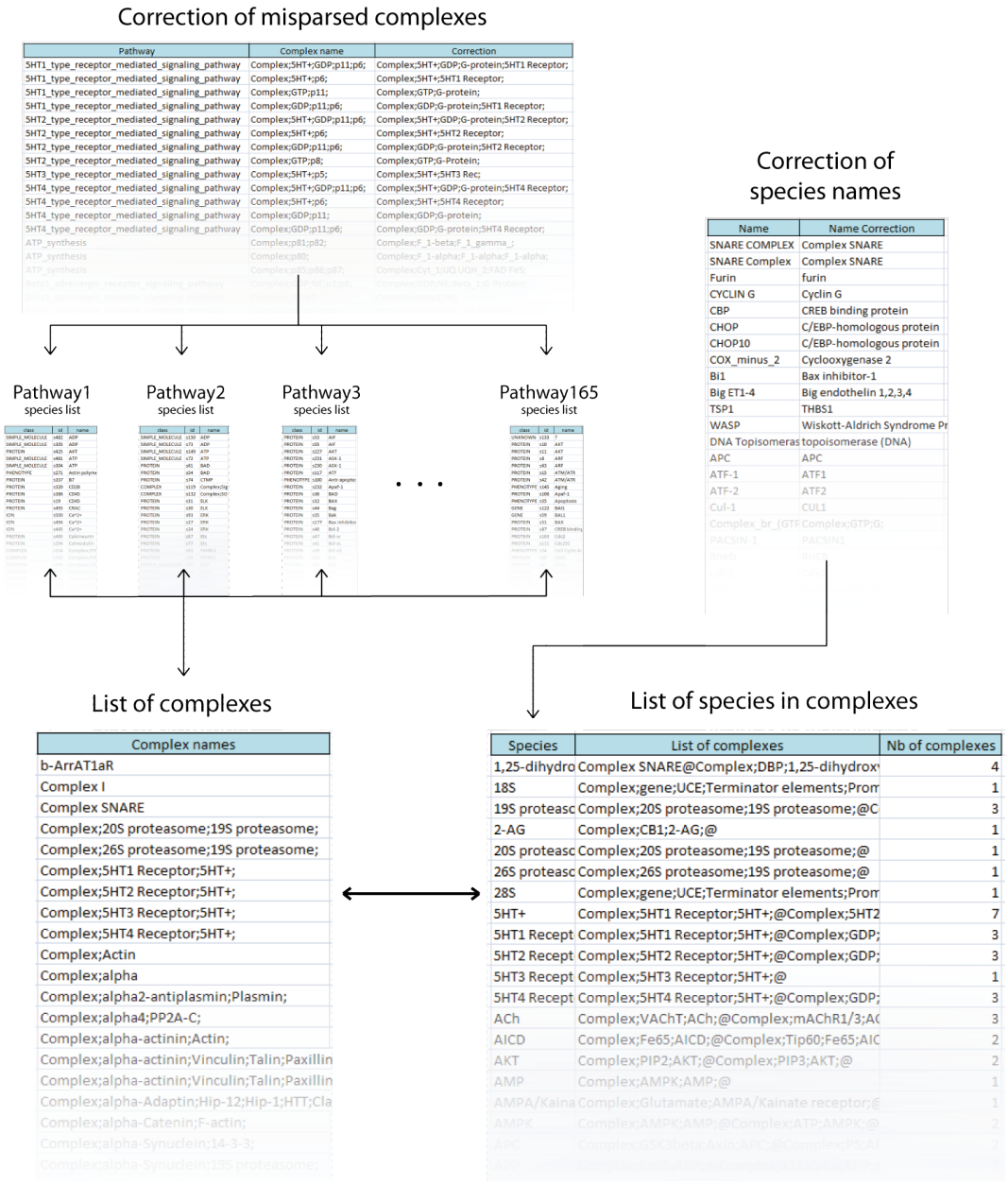


FIGURE 3.10 – *Principle of the complexes correction step. (top left) the table for correction of misparsed names of complexes. (bottom left) the list of complexes obtained from the union of all species lists. (bottom right) the list of all species that are in the complexes.*

- Some names of complexes in each species list are not parsed, for example some complexes designate species by their ID in the pathway and not by their name. To correct this, a table (top left of Figure 3.10) is manually created. After using this table, all complexes are formatted in the same way, that is to say their names are of the form *"Complex ;specie1 ;specie2 ;... ;specieN ;"* ;
- A list of all complexes present in all pathways is created (bottom left of Figure 3.10 ;
- From this list, a list of all species which are present in complexes is created by parsing the

complexe names (bottom right of Figure 3.10). For each specie the number of corresponding complexe and their names are written in the second and third columns ;
– All species names for the last table will be modified according to the same correction table used for the species corrections (see Figure 3.9) ;
– Given this correction of species names, the names of the complexes that they are part of are modified ;
– All new names for complexes are updated in all species lists of pathways.

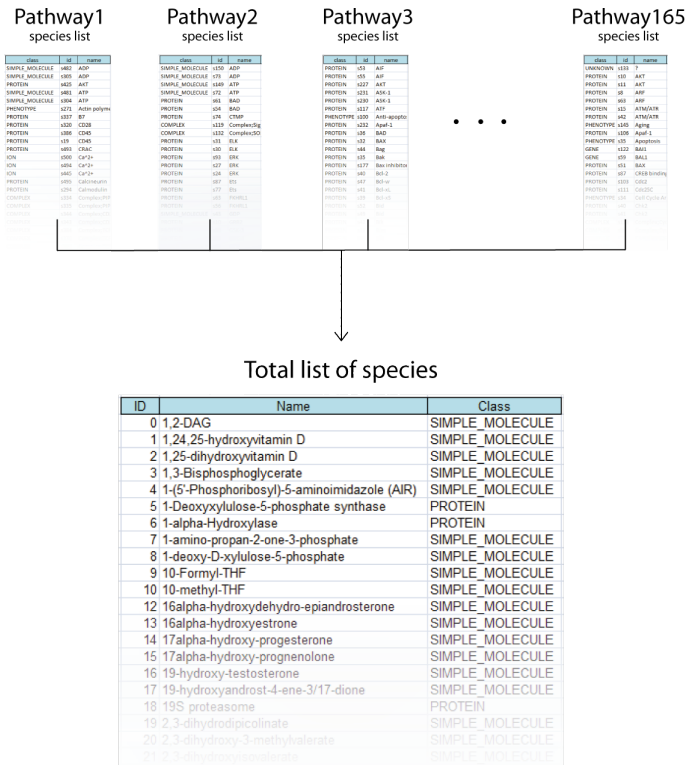


FIGURE 3.11 – *Principle of the total list of species creation.*

Once all name and class corrections have been performed, one can create a list of all species found in each pathway. This list is simply created (see Figure 3.11 by deleting all multi-occurrence of species with the same name and class. An Id is associated to each specie according to its position in the alphabetical order of the list. The list we obtained contains 2754 species, with 1345 proteins, 42 genes, 677 molecules, and 511 complexes.

3.2.4 Retrieving of the reaction list

As seen in section 3.2.2, all reactions are given by a list of reactants, modifiers, and products. That is because a reaction corresponds to a set of reactants which produce a set of products with the help of modifiers. There is hence a direct link from reactants to products, and modifiers interact bi-directionally with reactants and products. These reactions have to be transformed into a network link. We decided to create two types of network links following the reaction characteristics described above, the first one named Arcs which correspond to unidirectional links between reactants and products, and the second one named Edges which are bidirectional

links corresponding to the link between modifiers and others species present in the reaction. We also defined Edges between all elements of the same set (*e.g.* between all modifiers, between all reactants, and between all products). Figure 3.12 illustrates the creation of the network links we have chosen, using the simple pathway of 5-Hydroxytryptamine degredation.

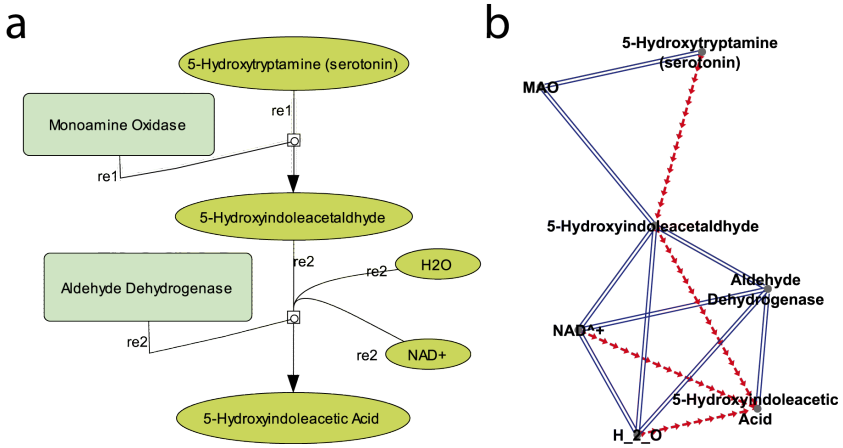
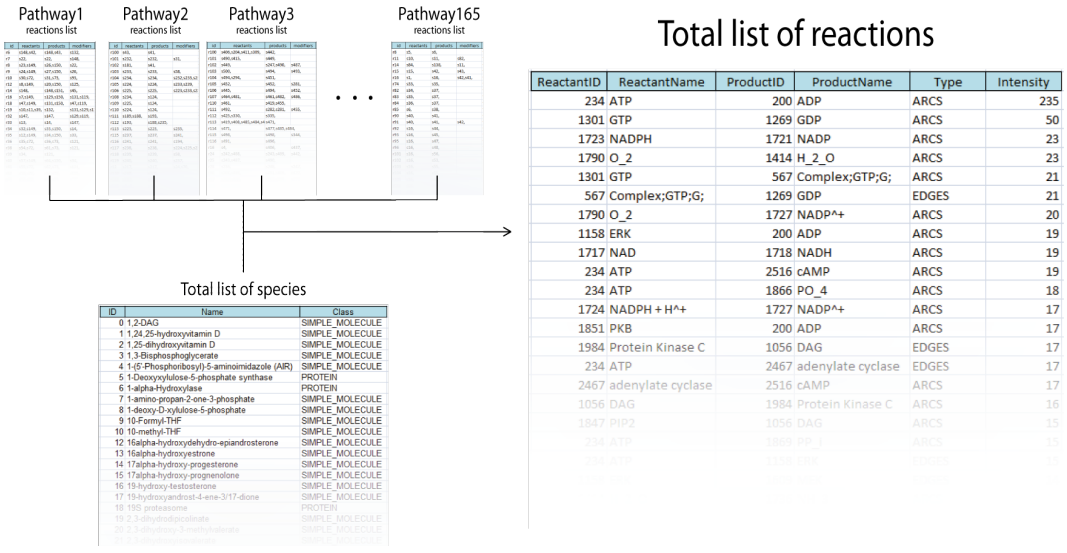


FIGURE 3.12 – *Principle of the creation of the Arcs and Edges, illustrated on the pathway 5-Hydroxytryptamine degredation. (a) Scheme of the pathway visualized with CellDesigner. (b) Network links obtained for this pathway. Red lines are Arcs (unidirectional links), blue lines are Edges (bidirectional links).*



ID	Name	Class
0	1,2-DAG	SIMPLE_MOLECULE
1	1,24,25-hydroxyvitamin D	SIMPLE_MOLECULE
2	1,25-dihydroxyvitamin D	SIMPLE_MOLECULE
3	1,3-Bisphosphoglycerate	SIMPLE_MOLECULE
4	1-(5'-Phosphoribosyl)-5-aminoimidazole (AIR)	SIMPLE_MOLECULE
5	1-Deoxyxylulose-5-phosphate synthase	PROTEIN
6	1-alpha-Hydroxylase	PROTEIN
7	1-amino-propan-2-one-3-phosphate	SIMPLE_MOLECULE
8	1-deoxy-D-xylulose-5-phosphate	SIMPLE_MOLECULE
9	10-Formyl-THF	SIMPLE_MOLECULE
10	10-methyl-THF	SIMPLE_MOLECULE
12	16alpha-hydroxydehydro-epiandrosterone	SIMPLE_MOLECULE
13	16alpha-hydroxyestrone	SIMPLE_MOLECULE
14	17alpha-hydroxy-progesterone	SIMPLE_MOLECULE
15	17alpha-hydroxy-pregnenolone	SIMPLE_MOLECULE
16	19-hydroxy-testosterone	SIMPLE_MOLECULE
17	19-hydroxyandrost-4-ene-3/17-dione	SIMPLE_MOLECULE
18	19S proteasome	PROTEIN
19	2,3-dihydrobiopterin	SIMPLE_MOLECULE
20	2,3-dihydroxy-3-methylvalerate	SIMPLE_MOLECULE

ReactantID	ReactantName	ProductID	ProductName	Type	Intensity
234	ATP	200	ADP	ARCS	235
1301	GTP	1269	GDP	ARCS	50
1723	NADPH	1721	NADP	ARCS	23
1790	O_2	1414	H_2_O	ARCS	23
1301	GTP	567	Complex;GTP;G;	ARCS	21
567	Complex;GTP;G;	1269	GDP	EDGES	21
1790	O_2	1727	NADP^+	ARCS	20
1158	ERK	200	ADP	ARCS	19
1717	NAD	1718	NADH	ARCS	19
234	ATP	2516	cAMP	ARCS	19
234	ATP	1866	PO_4	ARCS	18
1724	NADPH + H^+	1727	NADP^+	ARCS	17
1851	PKB	200	ADP	ARCS	17
1984	Protein Kinase C	1056	DAG	EDGES	17
234	ATP	2467	adenylate cyclase	EDGES	17
2467	adenylate cyclase	2516	cAMP	ARCS	17
1056	DAG	1984	Protein Kinase C	ARCS	16
1847	PIP2	1056	DAG	ARCS	15
234	ATP	1869	PP_I	ARCS	15

FIGURE 3.13 – *Principle of the creation of the list of reactions.*

Using all reaction lists for each pathway, the species ID given in the total list of species, and the network link creation rules described above, the software constructs a total list of reactions (Figure 3.13). Every reaction (*e.g.* network link), is annotated using the ID and name of the reactant and the product, the type of reaction (Arcs or Edges) and the intensity. This last parameter is given by the number of equivalent links found in the total list before deleting all

multi-occurrences of each reaction. It will help to give better representations of the network using intensity values as a weight for each network link.

3.2.5 Association of probes to species

One goal of the software is to map the information of gene expression onto the network of pathway component interactions. To this end, one needs to associate every species to one or more probes. In fact the only species that should be associated to probes are *GENE*, *PROTEIN* and species within *COMPLEX* as they all originate from gene transcription. On the contrary, species such as *ION* or *SIMPLE_ MOLECULE* originate from protein degradation or other sources, which is why they have to be left apart.

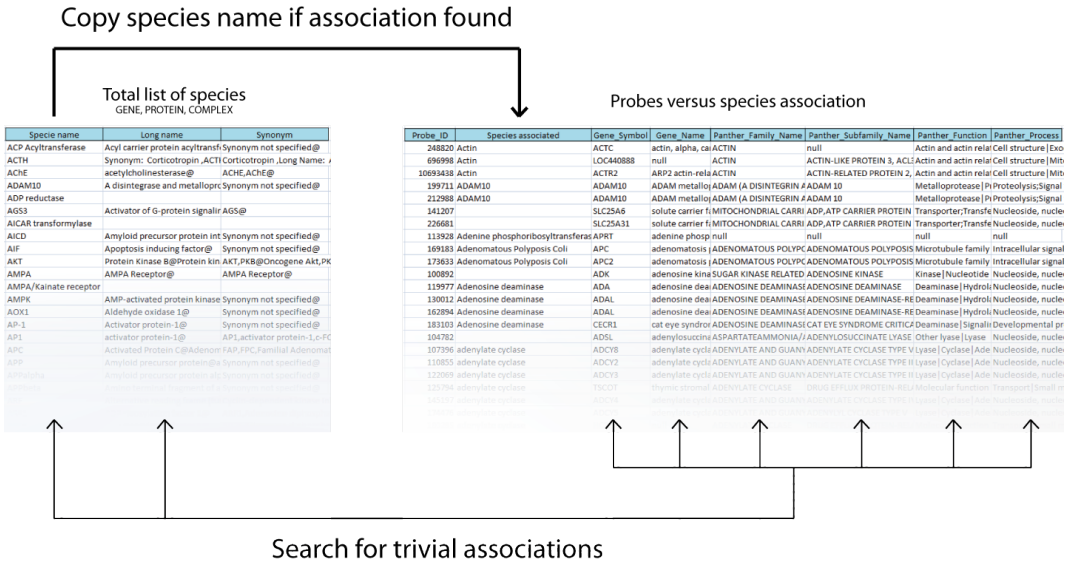


FIGURE 3.14 – *Principle of the automatic probe association step. (left) list of species and their LongName which have to be associated. (right) Probes_ Vs_ Species table in which the second column has to be filled with the species names associated to the probes.*

The technology used in our laboratory for human gene expression screening is the Applied Biosystems microarray (see appendix A), it contains 35517 probes. For this type of microarray, Applied Biosystems provides an annotation table in which for each probe one a variety of information such as : *Celera Gene ID*, *Entrez Gene ID*, *Gene_ Symbol*, *Gene_ Name*, and the *Chromosome_ Number*, is given. It does provide also five other classifications for each probe specific to Panther : *Panther_ Family_ Name*, *Panther_ Subfamily_ Name*, *Panther_ Function*, *Panther_ Process*. This last information is provided by Panther (*Protein ANalysis THrough Evolutionary Relationships* [77]) as its first goal was to create a protein classification system like Gene Ontology [46]. Our aim is to associate components of pathways extracted from Panther to Applied microarray probes, in order to do so, the only information one can use in the Applied Biosystem microarray annotation table is : *Gene_ Symbol*, *Gene_ Name*, *Panther_ Family_ Name*, *Panther_ Subfamily_ Name*, and *Panther_ Function*. To associate probes to species, this table is completed by writing in front of each probe the corresponding species : *Probes_ Vs_ Species* table is shown in the right side of Figure 3.14. As all probes are not annotated for some of them the five columns are void, we left apart these unannotated probes.

Finally, the *Probes_ Vs_ Species* table has eight columns, and 25877 rows corresponding to the list of probes with associated annotations. The number of species to associate to is 1549, consequently the association should not be done manually. I developed to this end an algorithm for making automatic associations (Figure 3.14).

The principle of this latter algorithm is quite simple : The entire list of species to be associated is parsed and a search not sensitive to the case of the species name is performed in the entire *Probes_ Vs_ Species* table. If the species name correspond for example to a *Panther_ Subfamily_ Name* of a probe, the corresponding *Species associated* cell will be updated. If this cell is already filled, the probe will be duplicated in order to have multiple species associated to the same probe. After the automatic association using species names is performed, one runs the same algorithm using the *Long Name* information of remaining species to complete the association. Figure 3.14 summarized this automatic association algorithm. Using it, we associated more than a quarter of the 1579 species.

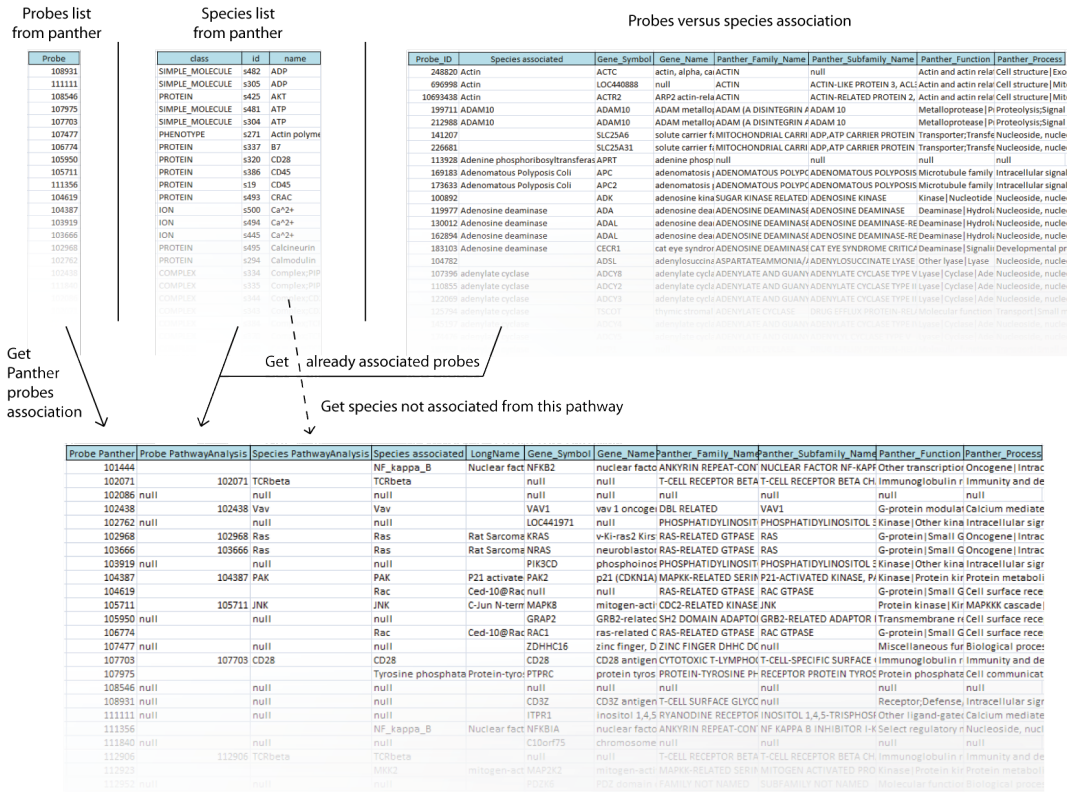


FIGURE 3.15 – *Principle of the creation of the PathwayProbesAssociation files for each pathway.*

The other associations should be found manually, to help us doing so I created different algorithms. The first one only creates three tables of species corresponding to the list of *GENE*, *PROTEIN* and species within *COMPLEX* that are not associated to probes, providing thus the number of species remaining.

The second algorithm helps to finish the association for important pathways. As our gene expression mapping will often be linked to LEO analysis, the quality of the network will be assured if one knows that species from important pathways are thoroughly associated. The idea is to correlate for each pathway the information given by the three different tables. As already mentioned, Panther provides for each pathway a list of probes as well as the list of species. The

algorithm will create then for each pathway a table, in which one can find the list of probes for each pathway given by Panther (top left of Figure 3.6), and the list of species contained in this pathway (top center of Figure 3.6). Using the *Probes_ Vs_ Species* table (top right of Figure 3.6) this information will be correlated, in consequence the *PathwayProbesAssociation* table (bottom of Figure 3.6 will indicate : (i) what are the probes from Panther that are not associated to species, (ii) which are the species from the pathway that are not associated to probes, and (iii) all the probes and species from this pathway which are already associated.

PathwayName	Nb_Common	Nb_Panther	Nb_PathwayAnalysis	Nb_Species
Wnt_signaling_pathway	239	215	210	49
Cadherin_signaling_pathway	188	32	163	10
Inflammation_mediated_by_chemokine_a	184	218	409	33
Integrin_signalling_pathway	178	128	188	27
Angiogenesis	156	140	245	34
Huntington_disease	128	123	152	49
T_cell_activation	93	114	95	46
Alzheimer_disease_presenilin_pathway	91	104	80	60
PDGF_signaling_pathway	85	149	54	16
FGF_signaling_pathway	84	112	32	16
Cytoskeletal_regulation_by_Rho_GTPase	83	92	193	19
Apoptosis_signaling_pathway	80	107	59	41
EGF_receptor_signaling_pathway	80	123	30	14
B_cell_activation	66	47	46	34
Ras_Pathway	64	60	208	15
Heterotrimeric_G_protein_signaling_pathv	63	129	86	22
Interleukin_signaling_pathway	63	361	86	20
ubiquitin_proteasome_pathway	63	73	84	6

FIGURE 3.16 – *Probes association table.*

A third algorithm creates a table (see Figure 3.16) which resumes all the information of association within each pathway. It contains five columns :

- The pathway name ;
- The number of probes which are both associated according to Panther, and associated according to the list of species correlated to the *Probes_ Vs_ Species* table ;
- The number of probes corresponding to the Panther probe lists which are not associated to species ;
- The number of probes which are associated according to our software, but are not included in the Panther lists ;
- The number of species in the pathway which are not associated to probes.

Using all these algorithms and also doing research in *Probes_ Vs_ Species* tables for elements containing species names, helps us to find the required associations. We finally reach the percentage of 64% of species associated, with 67% (908/1345) of proteins associated, 57% (24/42) of genes associated and 38% (63/162) of the species only within complexes. This last value is relatively low as the list of species contained only within complexes may be of any type (not only proteins or genes). On the list of 25877 probes, 7566 are associated.

One has to know that $PROTEIN$ class of species includes not only proteins such as Raf, or Synaptobrevin, but also big family of proteins. Consequently, one can find in the $PROTEIN$ class, species such as $IL2$, $Interleukin$ and $Cytokine$. The only difference between these species is the number of probes to which they are associated. In that particular case, $IL2$ is associated only to the gene with $GeneSymbol$: $IL2$ ($ProbeID$ 166840 and 190465), $Interleukin$ is associated to these probes to but also to 79 other probes, and $Cytokine$ is associated to $IL2$ and $Interleukin$ probes and 85 other probes. The network links defined between these different hierarchies of species depend only of the one defined in the Panther pathways. No algorithm has been created to link proteins to their parent (*i.e.* $IL2$ linked to $Interleukin$ itself linked to $Cytokine$). One possible development of the software in the future would be to take into account this hierarchy in the network construction.

3.2.6 Network construction

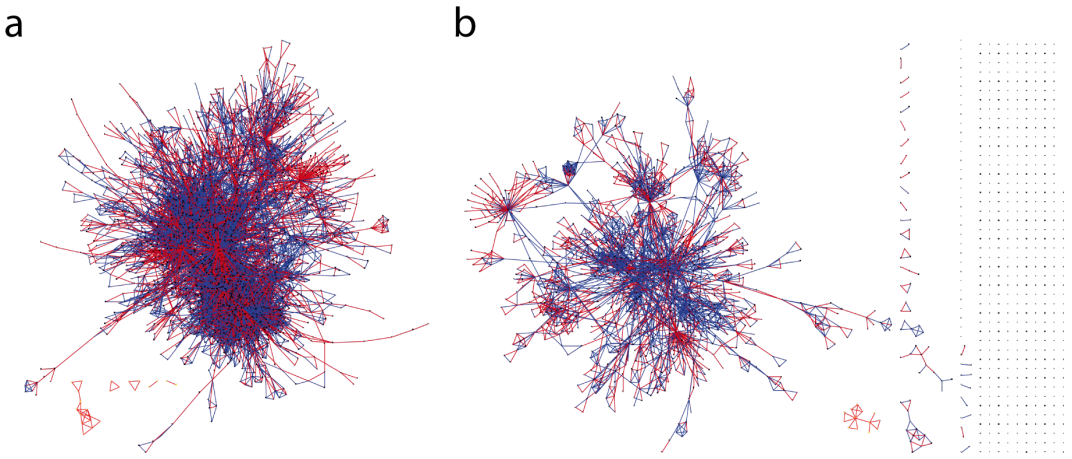


FIGURE 3.17 – *Networks of species are represented, with Arcs in red and Edges in blue. (a)* The SpeciesNetwork which contains all species found in all the pathways (2744 nodes and 10012 edges). There is two big hub nodes in the network which correspond to ATP and ADP, two important species, included in almost all pathways. (b) GSP (Global Signaling Pathway) which include all proteins, genes and complexes (1893 nodes and 3464 edges).

For the construction of a network, one requires two types of elements : nodes and edges. Consequently, the software allows to select a set of nodes (corresponding to a set of species), and then creates a list of edges (corresponding to a set of reactions containing the different selected species). To include species in the networks, (i) one can choose a list of pathways and the software will then include the different species contained in those pathways, or (ii) one can select directly a list of species within the user interface. Once the list of species is selected, the software automatically updates the list of reactions to include in the network.

For the node selection, possibilities are infinite, but there are indeed three types of global network that should be considered if one wants to compare different network mappings :

- SpeciesNetwork which includes all the species found in all the pathways (2744 nodes and 10012 edges). This network is the biggest one that can be obtained with the software (see Figure 3.17-a) ;
- GlobalSignalingPathway (GSP) network is created using all $PROTEIN$, $GENE$ and $COMPLEX$ species found in all the pathways (1893 nodes and 3464 edges). This network should be used for representing LEO results. The raw representation without information mapped onto it is shown in Figure 3.17-b ;
- ProbesNetwork is created using all $PROTEIN$, $GENE$ and $COMPLEX$ species which are associated to probes (1331 nodes and 2410 edges). This network is the best suited for mapping gene expression data onto it, as we are assured that every node is associated to a probe, as shown in Figure 3.18.

With the software, networks can be exported in two types of files, in order to be drawn and analyzed : Pajek and Cytoscape files. These two softwares are renowned systems biology tools, which allow to create networks, represent them, and analyze them easily. Cytoscape is the more powerful as it allows also to connect gene expression files to databases of protein-protein networks.

On networks, different information may be mapped : LEO and gene expression information. LEO results consist in a list of pathways over-or under-represented (see section 3.1.2). The software can load *ma*7 files, which are result of the LEO analysis, and it assigns a color for each pathway contained in the LEO table. To each species and reactions contained in the colorized pathways the corresponding color will be assigned. For species and reactions which are part of more than one colorized patwhays the black color is assigned. Finally, the network is exported in a file. An example of results of LEO analysis mapped onto a network is given in Figure 3.3-a.

The software can also load gene expression files by reading tables containing a list of probes and the "expression" value to be mapped onto the network (gene expression, fold change, p-value, etc.). For every probe, the corresponding list of species is searched, and the "expression" value of each specie is updated accordingly. As some species are associated to multiple probes (*i.e. Cytokine* specie which is associated to 166 probes), for the final calculus of species "expression" one has two choices :

- First,the "expression" value will be the average of all probe values associated to the species (see Figure 3.18-a). This type of node expression calculus is best suited when variables for evaluating differential gene expressions are used (comparison of "relative" values) ;
- Second, if multiple probes are associated to a species, the "expression" value will be the marginal sum of all probe values (see Figure 3.18-b). This calculus is best suited when parameter of gene expression are used (comparison of "absolute" values).

In the particular case of a probe associated to a specise which is part of a complex, the "expression" value of the complex will be updated using the formula $y = x/n$, with x the "expression" value of the probe, n the number of species in the complex, and y the new complex's expression value to add. Thus every species of one complex has same contribution.

Once values of "expression" are associated to each node, the network can be exported, and using a network analysis tool, such as Cytoscape. The node colorization will depend on the expression values, as shown in Figure 3.3.

The methodology developed here has been extensively used in different projects led by the host group. Among those is a project described in chapter 4 which has been part of my thesis work and concerns the definition of molecular transcriptome-based signatures for regulatory T-cell subpopulations. We have also made use of my algorithms in different projects of the host team, one of which has already led to a scientific publication found in Appendix G.

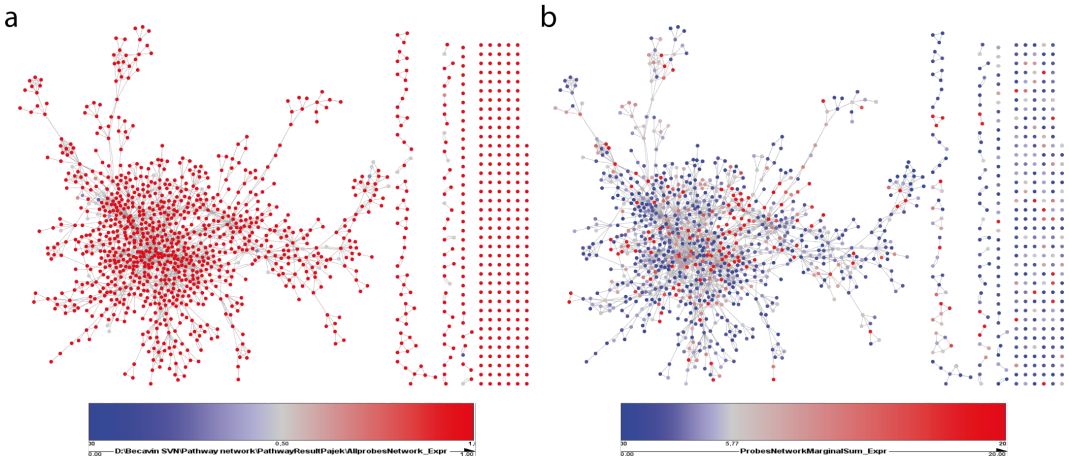


FIGURE 3.18 – *A gene expression table, in which every probe "expression" value is equal to 1, is mapped to the ProbesNetwork using (a) average of gene expression probe values, (b) marginal sum. In Figure (a) almost every node is red ("expression" equal to 1) which is reasonable as it is the average value of probes having an "expression" value of* 1. *Nodes having an "expression" value inferior to* 1 *are complexes in which not all of its constituting species are associated to a probe. In Figure (b) in which marginal sum is used, many nodes have an "expression" close to* 1, *which indicates that a vast majority of the species are associated to very few probes. Nodes which are solid red have a high "expression", and they are species associated to a lot of probes. Receptor, hormone and Cadherin are the most associated species, with more than* 200 *probes for each.*

4

Biological studies

In the previous sections I have presented the bioinformatics tools which I have designed for analyzing microarray data. For the moment, only the principle and the construction of them have been discussed and no complete utilization in an analysis has been shown. This is the topic of this chapter. Because during my thesis I worked not only on the development of tools for data analysis, but also passed a large amount of time using these tools as well as others not yet described in the analysis of biological datasets. These works were performed in cooperation with two different scientific groups :

- Sylviane Pied's group, at the Pasteur Institute in Lille, which is investigating the fundamental mechanism of Malaria, and tries to find good molecular discriminants for the different forms of Malaria severities ;
- Lars Rogge's group, at the Pasteur Institute in Paris, which works on the molecular characterization of regulatory T cells.

In a first section, I will briefly review on the principle of gene transcription in the context of cell biology, and then I will present the basics of Immunology, as both types of study I was involved in are related to this last topic. More importantly, despite development of bioinformatics tools being indeed a significant part of our work, one should not forget that the understanding of the mechanisms of information processing in the cell is our main scientific goal. I tried to always keep this kind of spirit during the development of all the tools described above.

After this review I will present the work done in collaboration with Sylvian Pied's group on the characterization of Malaria severity. Finally, the last section will be dedicated to the characterization of regulatory T-cell sub-populations.

4.1 Review of the biology relevant to our analysis

4.1.1 From genes to proteins

A typical eukaryotic cell such as the one shown in figure 4.1 is a complex machinery which involves many components from ions to big structures like mitochondria. In this diversity of elements, proteins are the major building blocks within a cell. Degradation or assembly of them form the vast majority of the component interactions in pathways. To create these proteins, one needs to transcribe and translate DNA which encloses the information. In the following, I will present quickly the different elements which allow the transformation of DNA-based information into proteins. My description of the phenomenon will be at some point specific to human cells, as all the studies presented in the following are human based, but the different mechanisms I

discuss are not specific to humans, for most of them, they are not even specific to eukaryotic cells.

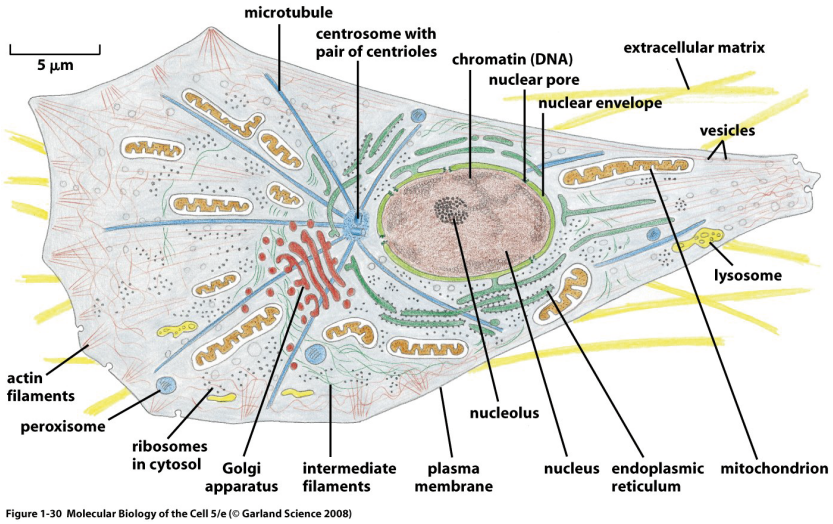


FIGURE 4.1 – *Schematic of a typical eukaryotic cell showing its different components, this figure is extracted from [73].*

The central dogma

Francis Crick, one of the fathers of the discovery of the structure of DNA [82], stated in 1958 the "central dogma of molecular biology" [83]. This dogma states that genetic information is transferred in three main modes of transfer between the three linear biological polymers : deoxyribonucleic acid (DNA) ribonucleic acid (RNA) and protein. The first transfer is from DNA to DNA through replication, reparation or recombination, the second is from DNA to RNA and is called transcription, the last one is from RNA to protein named translation (see figure 4.2).

Gene transcription

The transcription step , is the one I studied the most during my thesis. It involves the synthesis of a molecule of RNA which is complementary to the DNA sequence of gene. There are several forms of RNA of which one can extract three specific types which play a central role during the different processes involved in the central dogma :

- Messenger RNAs (mRNA) are those encoding for amino acids which are basics components of proteins, they are intended to be translated.
- Transfers RNAs (tRNA) bind to amino acids allowing the translation of mRNA into protein.
- Ribosomals RNAs (rRNA) are the most abundant, and are part of complexes nucleoprotein (ribosomes) which are the translation machineries.

In addition to these major RNAs, there are several other types of noncoding RNAs with enzymatic or regulatory functions. Table 4.1 shows a rather exhaustive list of all the RNAs one can encounter in an eukaryotic cell.

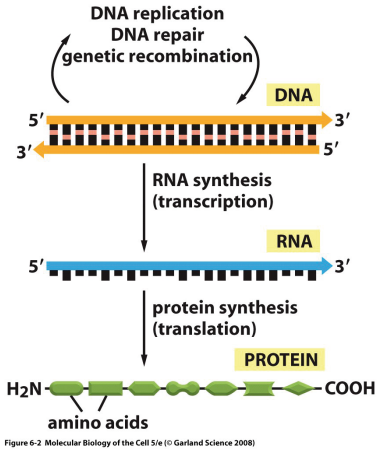


FIGURE 4.2 – *Scheme of the three types of exchange of information stated in the central dogma of molecular biology, this figure is extracted from [73].*

Based on the view of information transfer in the cell where genes encoded on the DNA are transcribed into RNA which themselves translated into proteins, one can consider RNA, and specifically the messenger RNA, as an intermediary between raw information stored as DNA and the product of this information (*e.g. protein*). This intermediary role in the flow of genetic information of the cell makes RNA a central element in the regulation of this flow of information. Indeed, through the various control mechanisms of transcription and the rate of degradation of RNA, these latter will determine which proteins are synthesized in the cell at a given time.

Transcriptome microarrays that we use in the laboratory measure the mRNA and therefore I will focus now on this type of RNA. The synthesis of mRNA is carried out in the nucleus by the complex DNA dependent RNA polymerase II and is typically described in three stages : initiation, elongation and termination. Initiation of transcription takes place in a specific DNA region called the promoter region, and is one of the main steps in transcription regulation. It begins with the recognition , through transcription factors, of the promoter region located upstream of the gene by the polymerase, to form the initiation complex of transcription. The transcript itself begins with a nucleoside $5' - triphosphate$ and will form a phosphodiester bond $5' - 3'$ with an identical nucleoside. The formation of similar bonds with all the nucleotides of the DNA sequence to transcript, following the principle of respecting complementarity with DNA, is the elongation step of RNA. The synthesis ends when a terminator sequence is encountered.

Then the just synthesized pre-messenger RNA just undergoes three types of post-transcriptional modifications (see figure 4.3) :

- $5'$ capping
- RNA splicing
- $3'$ polyadenylation

The first step corresponds to the addition of a $5'$ cap, that is to say the addition of a guanine by guanylyl-transferase with a connection type $5' - 5'$ [84]. This cap is used to regulate export of the mRNA from the nucleus, to ensure RNA stability, and to promote translation.

The third step correspond to a polyadenylation of the $3'$ part of the RNA, which means that a group of $\sim$ 200 adenine residues is added by the $poly - (A) - polymerase$, and the $Poly - ABindingProtein$ (PABP) binds to this polyadenylated sequence [85]. This $poly - A$ tail is also an important factor for the stability of the mRNA molecule as it protects from exonuclease

Type of RNA	Function
messenger RNA (mRNA)	Code for proteins.
ribosomal RNA (rRNA)	Form the basic structure of the ribosome and catalyze protein synthesis.
transfer RNA (tRNA)	Central to protein synthesis as adaptors between mRNA and amino acids.
small nuclear RNA (snRNA)	Function in a variety of nuclear processes, including the splicing of pre-mRNA.
small nucleolar RNA (snoRNA)	Used to process and chemically modify rRNA.
small cajal RNA (scaRNA)	Used to modify snoRNAs and snRNAs.
microRNA (miRNA)	Regulate gene expression typically by blocking translation of selective mRNAs.
small interfering RNA (siRNA)	Turns off gene expression by directing degradation of selective mRNAs and the establishment of compact chromatin structures.
Other noncoding RNA	Function in diverse cell processes, including telomere synthesis, X-chromosome inactivation, and the transport of proteins into the ER.

TABLE 4.1 – The different existing RNAs and their functions.

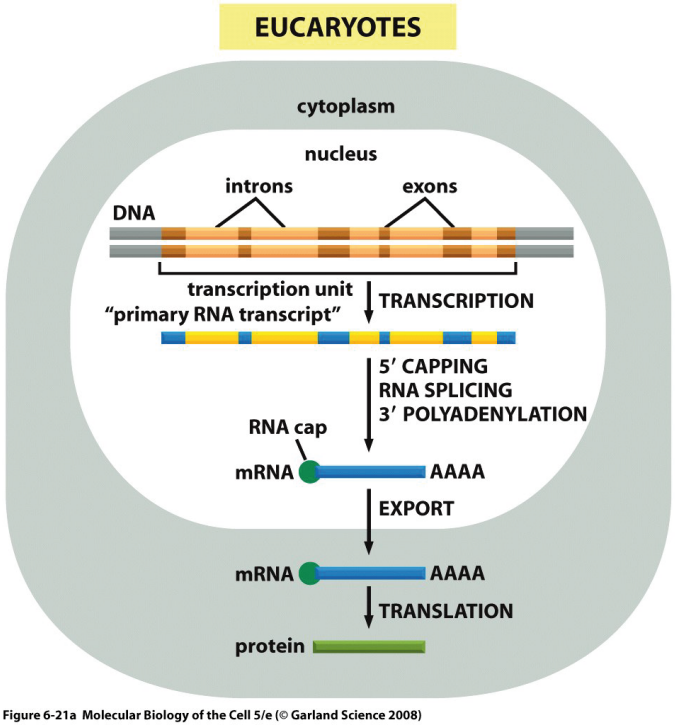


FIGURE 4.3 – *Scheme of the different processes involved in the transformation of DNA into proteins, this figure is extracted from [73].*

enzymes and degradation by the exosome complex in the cytoplasm and the nucleus [86].

The second step, RNA splicing, is important as it will separate the coding regions (exons) of the pre-messenger RNA from the non-coding region (introns). The latter are removed during splicing by the spliceosome, which is a complex of five small nuclear ribonucleoproteins (SnRNP),

U1, U2, U4, U5 and U6 [87]. It exists other minor splicing patterns involving other type of snRNP spliceosomes, autocatalytic introns or tRNA. Alternative splicing appears when some exons are removed or conserved in comparison of the main stream, consequently it induces genetic variability [88]. This phenomenon has to be considered in the design of the microarray chips (see appendix A).

Spliced mRNAs then travel through the nucleus, directly the cytoplasm. This set of mRNA plus all other types of RNA produced are called the transcriptome of a cell, and correspond to the list of all parts of the DNA which has been transcribed. This name is in opposition to the genome, which is the total list of nucleotides present in DNA, and the proteome, which is the set of proteins included in the cell.

Once transported into the cytoplasm via the nuclear pore complexes, mRNAs are translated into proteins by the ribosomes. A single molecule of RNA can be translated several times and the half-life of mRNA in the cytoplasm influences the frequency of this translation. Indeed, mRNA has an average life span of a few minutes [89] in the cytoplasm. The RNA degradation begins with the $poly - A$ tail and the cap, followed by the rest of the molecule.

Gene regulation

The transcription process just described is not as simple as it was enonciated in the premice of genetics. The reading of the DNA will be affected by several mechanisms of regulation. The first type of regulation involves a set of proteins acting at different levels on the transcription, the two major being :

- Transcription factors which act on specific areas of DNA and are capable of activating or inhibiting the expression of genes. They bind to DNA, usually upstream of the gene to facilitate or prevent transcription of this gene. These elements can also be regulated by transcription factors, it results a gene regulation network of DNA transcripts which has to be unraveled ;
- Epigenetic properties also participate to the transcription regulation. The best studied mechanism is the methylation of nucleotides in certain zones of the DNA [90]. These methylations are carried out by DNA methyltransferases, and occur most often at CpG islands. They may prevent transcription if they are placed in transcription sites or by recruiting chromatin remodeling proteins via $methyl - CpG - binding\ domains$ (MBDs) [91].
  Another type of epigenetic regulation may be induced by the different folded and chemical states of the chromatin both linked to the degree of compaction. It can directly affect the transcription by narrowing the access for the polymerase [92]. In section 4.1.1, I will described more in detail the different types of regulation which may be produced by the chromatin configuration.

Apart from all these possible regulations of transcription, a set of post-transcriptional events mentioned above may also affect the final composition of the transcriptome. They can intervene at the level of splicing, during the transport of RNA in the nucleus, and translation [93].

Human genome

The first draft of the human genome was published in 2001 jointly by the international human genome sequencing consortium [94] and Celera genomics [95]. A second publication of the genome corrected especially in the zones of compacted chromatin (heterochromatin), was published a few years after in [96]. With its lentgh of 3.2×10^9 base pairs (bp), the human genome is not particularly big or small, as a comparison *Saccharomyces cerevisiae* has $12 \times 10^6 bp$

Human genome	
DNA length	3.2×10^9 base pairs (bp)
Number of genes	~ 25000
Largest genes	2.4×10^6 bp
Mean gene size	27000 bp
Smallest number of exons per gene	1
Largest number of exons	178
Mean number of exons per gene	10.4
Largest exon size	17106 bp
Mean exon size	145 bp
Number of pseudogenes	more than 20000
Percentage of exons sequence in DNA	1.5%
Percentage of highly conserved sequence in DNA	3.5%
Percentage of high-copy repetitive elements in DNA	$\sim 50\%$

TABLE 4.2 – Some statistical variables of human genome, source [73].

and *Polychaos dubium* which is considered to have the largest genome of any known organism ($670 \times 10^9 bp$).

The human genome contains approximatively 25000 genes with a mean gene size of $27000bp$, table 4.2 provides more statistical information on the human genome. Among all the human DNA the quantity of coding regions is in fact pretty low, it is estimated at 1.5%. This phenomenon appears clearer when one performs a series of tenfold zooms of one chromosome such as in Figure 4.4. One can see that the size of a typical exon, is very small compared to the size of an intron, and thus all the exons together are a small proportion of genes which are themselves rare in the entire chromosome.

Chromatin structure

DNA is of course not linearly disposed in the nucleus such as illustrated in the different schemes I have shown in the above sections. It is formed of structures winded upon each other. However one can find distinct scales from DNA to chromosome structure, as shown in figure 4.5-a. One scale which seems to play role in gene regulation is the first chromatin scale. In nucleus, a family of proteins named histones assembles and attracts DNA which will wind around an histone complex forming a nucleosome [97] (see figure 4.5-b). The chromatin is formed by these nucleosomes and naked DNA sequences connecting them which are called linkers. The nucleosome positioning on the DNA sequence is quite mysterious. Some recent studies shown that a minority of the positions are sequenced driven, whereas the vast majority of nucleosomes are statistically positionned [98]. This question is important as nucleosome positioning is considered has one of the epigenetic processes which influence gene transcription. Moreover nucleosomes may be chemically modified into thousands of chemically possible states which will also modify gene transcription. These epigenetic modifications of nucleosomes form the "histone code" [99] which include acetylation by acetyl-transferases such as CREB binding protein and phosphorylation, methylation, ubiquitination, sumoylation and citrulline-conjunction.

The two types of chromatin dependent epigenetic control I just demonstrated proves the usefulness of unraveling chromatin structure in order to understand gene transcription. Before my thesis, I have done an internship in the group of Living media multi-scale modeling at the Paris 6 laboratory of Theoretical Physics of Condensed Matter (LPTMC) under supervision of Jean Marc Victor and Annick Lesne. In this group they try to unravel the structure of condensed

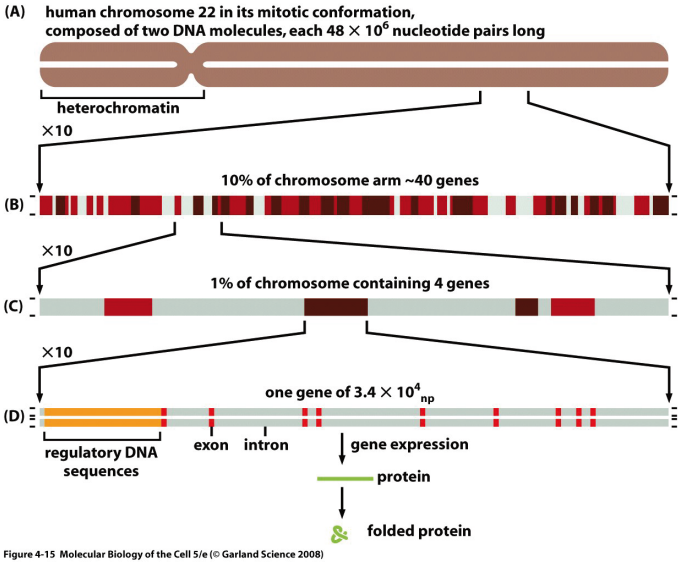


FIGURE 4.4 – *Succession of four tenfold zoom on the* 22^{th} *human chromosome. (a) View of the entire chromosome. (b) Tenfold expansion of the chrosmosome showing* ~ 40 *genes, in dark brown there is the genes annotated (known function) and in red the genes predicted which function is not yet known. (c) another tenfold expansion of the chromosome representing* 1% *of it. (d) the last tenfold expansion of the chromosome show a typical gene of* 34000*bp with its small proportion of coding region in red.*

chromatin (heterochromatin) by topological and physical constraint study. They found an all-atom coherent structure which follows some well established chromatin parameters [100] (see figure 4.5-c,d).

They have also been part of a discovery of a new inversed nucleosome structure called reversome [102]. This new component may appear when one applies a torque to the linker DNA of one nucleosome. If the right amount of energy is applied, the molecular conformation will pass the energy barrier and become a reversome, as shown in figure 4.6-a. Reversomes may play an important role in gene regulation as it is a less stable structure and then should allow the polymerase to pass through move easily.

Using the heterochromatin just described and this new type of nucleosome, we demonstrated a sequential model of transcription within heterochromatin (see figure 4.6-b) which is in general considered very difficult as the polymerase has few space to perform its task. We prove that in our model the transcription is not forbidden in heterochromatin but on the contrary facilitates. This model stems from work which began during my internship at LPTMC and finished in the middle of my thesis. It has been validated by a first author publication in the BioPhysical Journal in march 2009 [101] (see appendix E).

4.1.2 Immunological principles

Pathogens

The human body is not a closed complex system, it is always invaded by different types of the organisms. Sometimes it lives in symbiosis with them, but other times those organisms will disrupt the functioning the human host organism causing disease. Agents which cause infectious

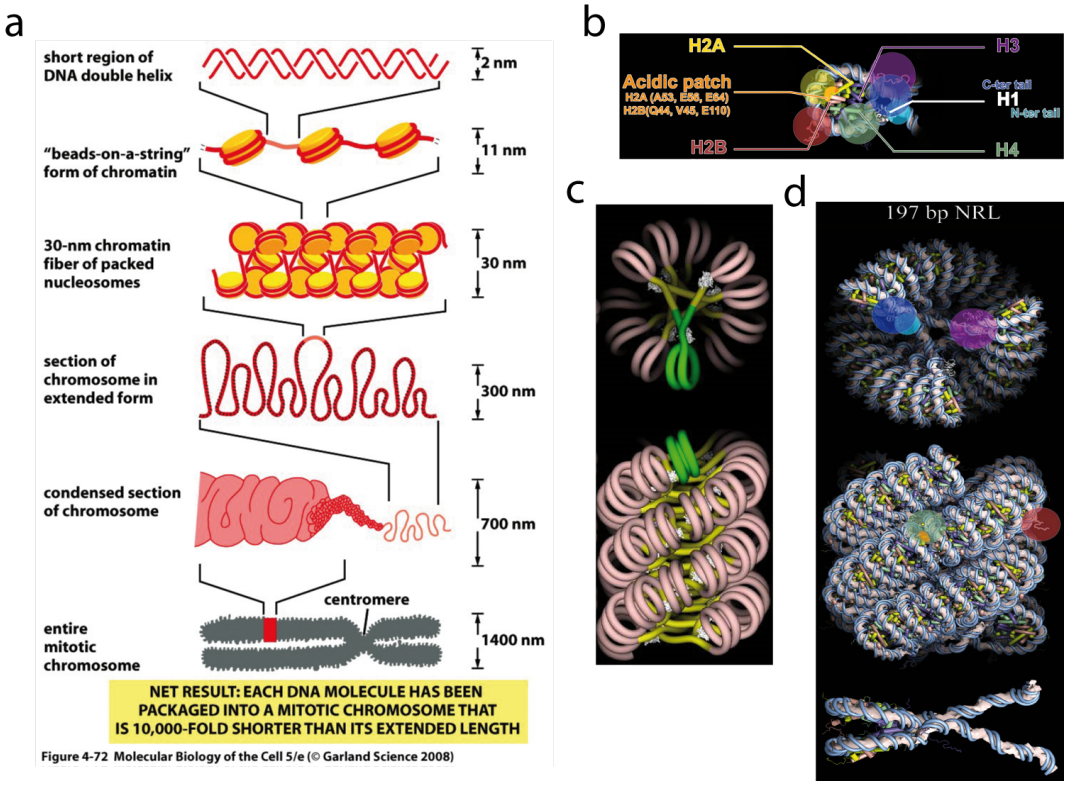


FIGURE 4.5 – *(a) The different scales of structures found from DNA to chromosomes, this figure is extracted from [73]. (b) Nucleosome representation showing all the histones proteins which compose it, extracted from [100]. (c) Schematic representation of the condensed chromatin structure described in [100]. (d) All-atom description of the same chromatin shown in (c). This condensed chromatin created using the N-start model, has 3-start and* 197*bp nucleosome repeat length (NRL).*

diseases are collectively called pathogens. They frequently exploit the biological attributes of their host's cells in order to infect them. Viruses, bacteria are the two main pathogens one will encounter.

Viruses are basically small encapsulated strands of DNA or RNA, much smaller than a normal cell, which will transfer their DNA or RNA into host cell. They enter host cells usually by membrane fusion, pore formation or membrane disruption. Once a viral genome is inside a cell, it uses the genetic machinery of the host cell to produce spurious proteins disturbing some molecular pathways causing cell malfunctions and especially reproduce itself.

Bacteria are in comparison bigger organisms, as they are prokaryotics cells. They enter a host cell by phagocytosis, which is the cellular process of engulfing solid particles by the cell membrane to form an internal phagosome by phagocytes and protists. Once a bacteria has invaded a host cell, different mechanism are possible, each of them will certainly modify the cell fate. Bacterial invasion are not always bad for a cell, sometimes the invader may improve its functioning. For example the mitochondria are supposed to have originated from prokaryotic cells which has fusioned with eukaryotics cells during the evolution. As this association was beneficial for the cell, providing a source of chemical energy, it was conserved.

The two pathogens types just described here, both take advantage of the cell machinery to

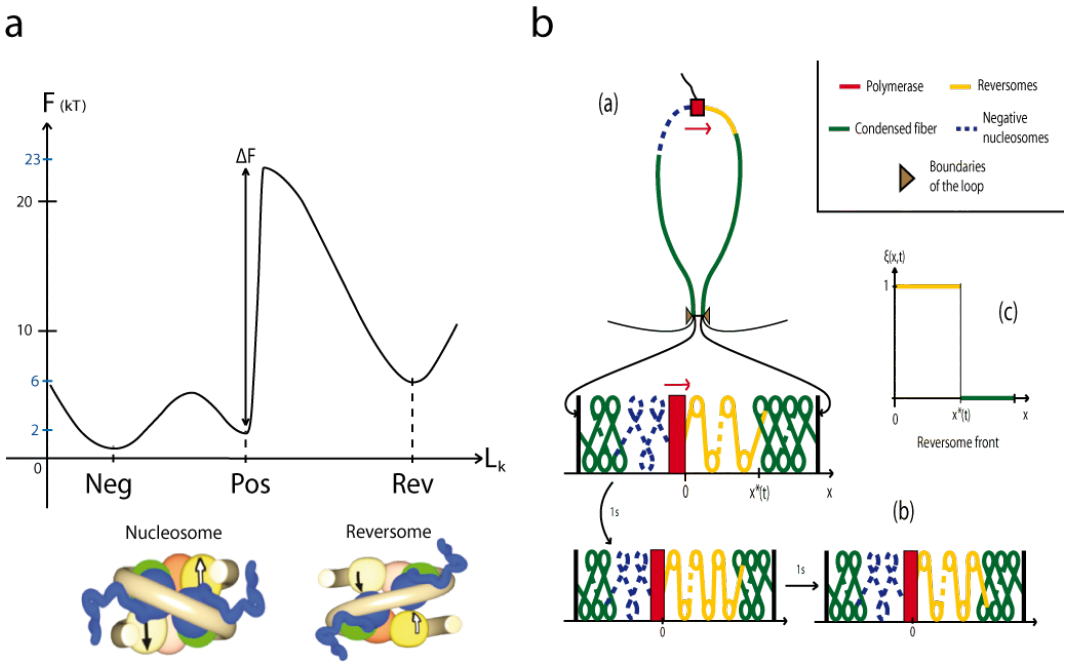


FIGURE 4.6 – *(a) Energy profile depending of torque applied to DNA linker. As the torque increases the nucleosome will go through negative to positive configuration, and then cross the energy barrier to transform in a reversome. (b) The sequential model of transcription (domino-effect) within loop of heterochromatin we have described in [101]. (b-a) The supercoiling generated by the polymerase activity is trapped within the loop delineated by topological boundaries (the thin black regions are outside the loop). The ensuing torsional constraints trigger the sequential transition of nucleosomes (in green) into reversomes (in yellow). (b-b) Illustration of the domino effect : after* $1s$*, the fifth nucleosome downstream of the polymerase (green in panel b-a) has turned into a reversome (yellow in panel b-b) ; after one more second, the sixth nucleosome has turned into a reversome. (b-c) Reversome density profile : in the bold yellow region* $[0, x*]$*, the reversome density* $x(x,t)$ *equals 1. The wavefront is located at* $x*$ *and propagates downstream* ~ 10 *times faster than the polymerase progression. In the polymerase wake, the nucleosomes turn to the negative state (dashed blue in panel b-a) to ensure the conservation of the total linking number of the loop.*

produce their own DNA or proteins. They can also alter the behavior of the host organism to facilitate the spread of the pathogen. The human body is in fact naturally armed to fight the vast majority of the invasions. Physical barriers provide the strongest armor against intruder. Barriers such as our outer layers of skin, and associated chemical defenses, or such as acid in the stomach, which prevents most microorganisms (microbes) from coming into contact with sterile tissues in our body. Secondly, individual human cells possess some intrinsic defensive capabilities, for example, cells aggressively degrade double-stranded RNA molecules, which are a hallmark of certain kinds of viral infections.

The immune system

Despite these many protections, some pathogens will enter sterile tissues. In order to fight them the human body has developed two types of defenses : the Innate immune system and the

adaptive immune system. The first type is basically composed of cells which kill intruders. But those killer cells have to fight only bad organisms, the adaptive immune system will help them by labeling organisms which are known to the organism to be pathogens.

Cells of both immune system are called white blood cells (Leukocytes), the other blood cells being the red blood cells (Erythrocytes) and platelets (Thrombocytes). All blood cells are produced during cell differentiation Hematopiesis (see figure 4.7). In developing embryos, blood formation occurs in aggregates of blood cells in the yolk-sac, called blood islands. As development progresses, blood formation occurs in the spleen, liver and lymph nodes. When bone marrow develops, it eventually assumes the task of forming most of the blood cells for the entire organism. However, maturation, activation, and some proliferation of lymphoid cells occurs in secondary lymphoid organs (spleen, thymus, and lymph nodes) as we will see in the following. In children, haematopoiesis occurs in the marrow of the long bones such as the femur and tibia. In adults, it occurs mainly in the pelvis, cranium, vertebrae, and sternum.

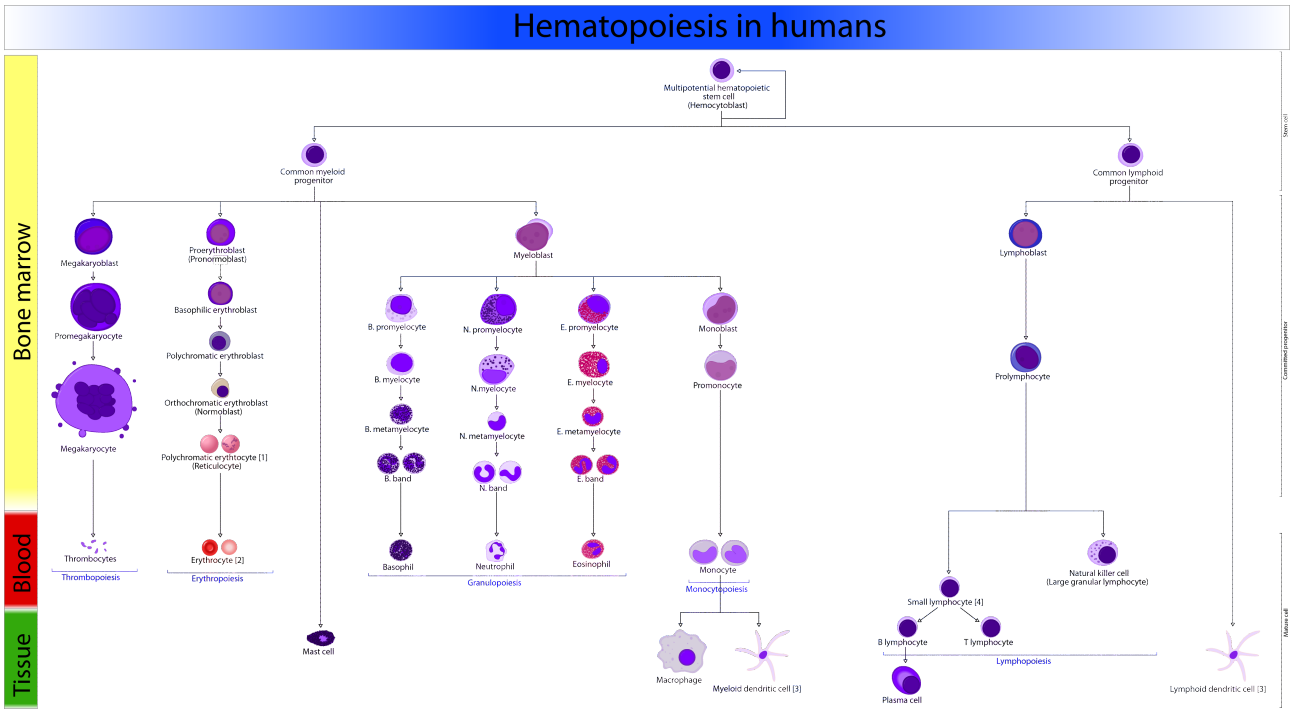


FIGURE 4.7 – *Hematopoiesis cell differentiation diagram. From Haematopoietic multipotent stem cells all different types of blood cells are produced. This figure is extracted from [103].*

Immune cells circulates in blood vessels but also in a specific network of organs which is called the Lymphatic system. This system illustrated in figure 4.8 allows the transportation of white cells from their site of maturation to the site of infection where intruders are. There are three types of components, first the organs of production of white blood cells such as bone marrow and thymus. Lymph nodes are small organs which make the link between blood stream and lymphatic stream. They act as blood filter which only let pass the constituent of the lymph. Lymph is a fluid composed mainly of white blood cells, lymphocytes are the most present in lymph. There is a great quantity of lymph nodes in a human body disposed in periphery of blood vessels. Finally, a huge network of lymphatic vessels connects every node to other lymphoid organs.

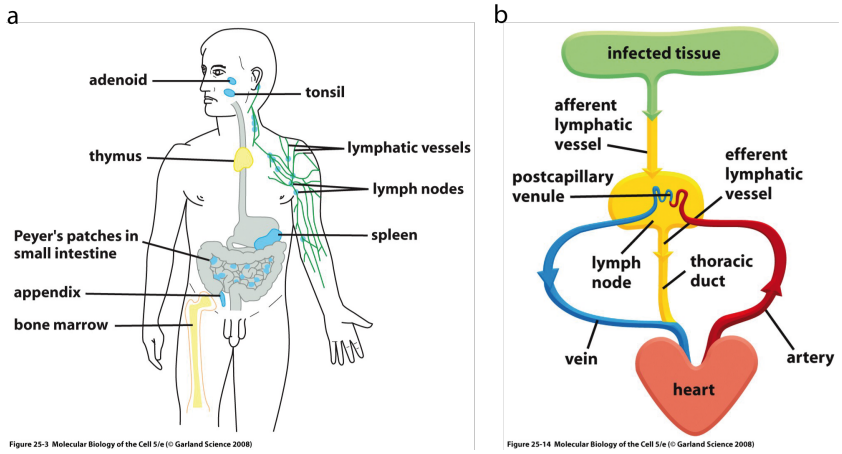


FIGURE 4.8 – *Overview of the Lymphatic system. (a) The different lymphoid organs. (b) A scheme of the lymphocytes circulation. Dendritic cells carry antigens from the infected tissue into the lymphatic system. In response, triggered lymphocytes are directed to lymph nodes where they will pass in the blood stream, and thus will attain the infected tissue. This figures are extracted from [73].*

Innate Immune system

As shown in the table 4.3, there is six categories of white blood cells. Almost three quarter of them are part of the innate immune system. As I said most of these cells are natural killer cells (NK). Neutrophils and macrophages destroy pathogens by phagocytosis. Even though their methods of killing cells are very similar, there is many differences between them. First, neutrophils are presents in a huge quantity in comparison to macrophages (monocytes) as shown in table 4.3. This difference is due to the fact that macrophages are long-lived cell whereas neutrophils are short-lived, thus in order to be efficient neutrophils as to be present in a great quantity. They operating mode for finding cells to kill is also different. Neutrophils are chemicals sensor, which are mainly directed to infected tissues by chemotaxis. On the contrary, macrophages target cells which has been triggered by the adaptive immune system. After a killing operation neutrophils often die, macrophages will often survived. Dead and dying neutrophils are a major component of the pus that forms in acutely infected wounds.

Type	Category	Proportion
Neutrophils	Granulocyte	$50-70\%$
Eosinophils	Granulocyte	$1-3\%$
Basophils	Granulocyte	$0.4-1\%$
Lymphocytes	Agranulocyte	$25-35\%$
Monocytes	Agranulocyte	$4-6\%$

TABLE 4.3 – The different types of white blood cells existing, with their category and their average proportion in human body.

Another type of killing cells are the Natural Killer cells (NK cells), they are lymphocites cells (see table 4.4, and their operating mode is different from phagocytosis. They monitor the level of class I major histocompatibility complex (MHC) on the surface of all host cells : high levels inhibit the killing activity of NK cells, so that NK cells selectively kill host cells expressing low

levels, which are mainly virus-infected cells and some cancer cells. MHC protein is a cell surface family of proteins which are expressed on every cell, and thus play a major role in the distinction between self and non-self. NK cells destroy virus-infected cells by inducing the infected cells to kill themselves by undergoing apoptosis.

The last type of innate immune cell are Dendritic cells. They play a major role in the connection between innate and adaptive immune system, as they recuperate infectious pathogens or molecules secreted by the pathogens which are called antigen (for antibody generator), and lead them to the adaptive immune cells through the lymphatic system.

Adaptive immune system

The two powerful characteristics of pathogens which help their species to survive are capacity of high duplication and capacity of evolving. This two capacities combined imply that in a case of an attack, a lot of them will die, few will survive thanks to some mutations, and these few pathogens will duplicated and finally win the war. Consequently The number of different pathogens that exists is extremely big. Considering this variety, how does an immune system discriminate good cells from bad cells ?

Every cell is an open system, different molecules enter and leave the cell at each moment. Some of this proteins are called Cytokines, they are small cell-signaling protein molecules that are secreted by the glial cells of the nervous system and by numerous cells of the immune system. They are a category of signaling molecules used extensively in intercellular communication. Pathogens also produces different types of molecules which leave them, they are usually called antigens (stands for antibody generator). Innate immune cells are not armed to recognized every existing antigen, the help in this recognition will be provided by the adaptive immune system.

The adaptive immune system is constituted of a pool of lymphocyte cells. In the early stage of the development of a person, every cells are naive, that is to say that they are not specific to a certain type of antigen. Then, the more antigens will be encountered the more lymphocyte cells training will occur. Once a lymphocyte is trained it will be specialized to the recognition of an antigen and may have two types of comportment :

- First, it can proliferate and differentiate into an effector cell, producing consequently an immune response.
- Secondly, it can differentiate in a memory cell. This type of cell is not immunologically active, but it is already antigen specific, and have the capacity of differentiate and proliferate into an effector cell more quickly then a naive cell when a later encounter with the same antigen occurs.

Effector cells have usually a short lives, whereas memory cells which have long lives. Thus, when a new antigen is detected within an organism it will induce the differentiation of naive cells into effector cells which will trigger an immunological response and die rapidly after it. It will also induce the differentiation of naive cell into memory cell which will rest for a long time in the bone marrow. In the case of a new encounter with this antigen, the pool of memory cells will rapidly differentiate into effector cells and induce immunological responses. The more the antigen is present in the system, the more memory cells will be stocked in the lymphatic system. But if this antigen is a molecule of the self, the different triggering mechanism of the immune system will provides the stock of memory cells to be created.

Adaptive immune cells are all lymphocyte cells. It does exists 5 types of lymphocytes classified in table 4.4. Helper T cells will be of great interest for us in the following as we have studied them extensively with Lars Rogge's group. They are the most produced and they play a role of "policeman". Indicating pathogens cells to the innate immune system, helping other lymphocytes

and regulating the number of immune cells in order to prevent auto-immune responses. Cytotoxic T cells have a killer role. As they are adaptive immune cells they are trained to recognize a certain antigen. Once they find it, they bind to the corresponding pathogen and kill it using the same process as natural killer cells. B cells have a role of creating of injecting antibody into the blood stream. These antibodies will bind to antigen and to the pathogen, thus triggering it for an innate immune response. Finally, $\gamma\delta$T cells represent a small subset of T cells that possess a distinct T cell receptor (TCR) on their surface. A majority of T cells have a TCR composed of two glycoprotein chains called $\alpha-$ and $\beta-$ TCR chains. However, in $\gamma\delta$T cells, the TCR is made up of one $\gamma-$chain and one $\delta-$chain. The role of this last type of immune cell is not clear for the moment, but it seems to be a part of both innate and adaptive immune system.

The different lymphocytes described here are discriminated using the different types of cell-surface markers they produced, for example T cells are cells which produce a big quantity of CD3 (cluster of differentiation 3). This cell-surface markers play major role in the function determination of a cell, as the entire immunologic system rely on cell communication using lymphokines (cytokines produced by lymphocytes).

Type	Function	Proportion	Phenotypic markers
Natural Killers cells	Lysis of virally infected cells and tumour cells	7% $(2-13\%)$	CD16, CD56 but not CD3
Helper T cells	Release cytokines and growth factors that regulate other immune cells	46% $(28-59\%)$	TCR$\alpha\beta$, CD3, CD4
Cytotoxic T cells	Lysis of virally infected cells, tumour cells and allografts	19% $(13-32\%)$	TCR$\alpha\beta$, CD3, CD8
$\gamma\delta$T cells	Immunoregulation and cytotoxicity	$< 5\%$	TCR$\gamma\zeta$, CD3
B cells	Secretion of antibodies	23% $(18-47\%)$	MHC class II, CD19, CD21

TABLE 4.4 – The different types of lymphocites, with their function, proportion and phenotypic cell surface markers (source [103]).

B cells and antibodies

Main important types of cells for adaptive immune system are helper T cells and B cells as they both play a major role in the triggering of the immune response. B cells role is to produce antibodies which will be released in the lymph and blood stream. Collectively called immunoglobulins (abbreviated as Ig), they are among the most abundant protein components in the blood, constituting about 20% of the total protein in plasma by weight. Mammals make five classes of antibodies, each of which mediates a characteristic biological response following antigen binding. An antibody is a Y shape protein, two termination are designed to bind with antigens, the other termination binds to macrophage, neutrophils and other killing cells in order to activate a phagocytosis. The role of antibody is either to neutralize antigens and to label an antigen presenting cell which have to be destroyed, and trigger a killing response of innate immune cells. All antibody molecules made by an individual B cell have the same antigen binding sites.

The process of B cell differentiation, described in figure 4.9-a, begin when an antigen (with the aid of a helper T cell) activates a naïve or a memory B cell, that B cell proliferates and differentiates into an antibody-secreting effector cell. During its life, an effector cell will growth to began a big plasma cell massively secreting antibodies. In the blood, most plasma cells die

after several days, but in the bone marrow some will survive much longer and keeps secreting antibodies in the blood, providing a long term defense against the pathogen.

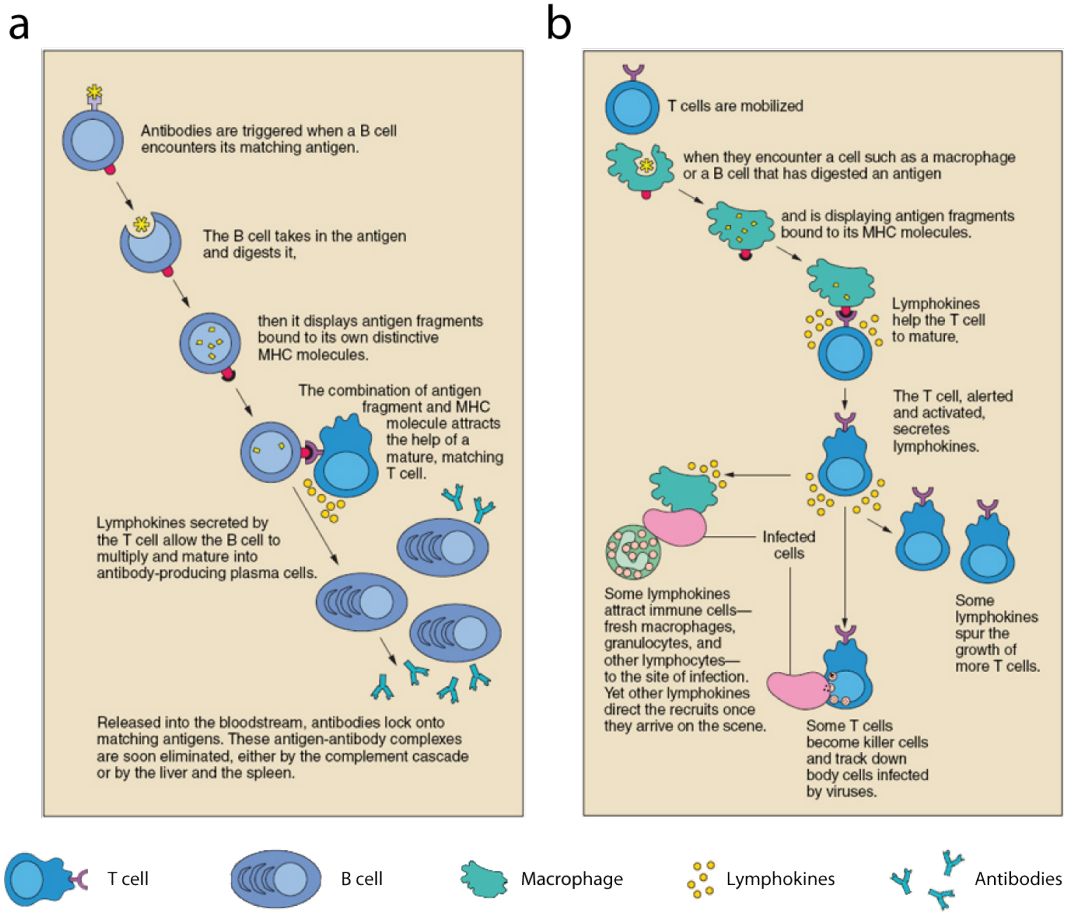


FIGURE 4.9 – *The operating mode of B cell and T cell. (a) A B cell encounter an antigen, and thanks to a T cell it differentiate into an antiboy-producing cell. (b) By contact with an antigen-presenting cell, T cell differentiate to become effector T cells which will trigger immune response, cytotoxic T cells which will kill, or regulatory T cells which regulate the quantity of T cell. These figures are extracted from [104].*

Helper T cell

Contrary to B cell which act as long-range, T cell only act as short-range. They need to be in contact of a cell to trigger an immune response. They can detect the presence of an antigen or even an antigen that have been partly degraded inside an antigen-presenting cell. Once the differentiation of T cells is engaged, different roles may be assigned to them. They can be transformed in cytotoxic T cell that will kill pathogens by triggering apoptosis. But the majority of naive T cells will become helper T cells. There is three types of helper T cell :

- Effector Tcells secrete cytokines, proteins or peptides that stimulate or interact with other leukocytes, including T cells ;
- Memory T cells retain the antigen affinity of the originally activated T cell, and are used

to act as later effector cells during a second immune response ;
- Regulatory T cells do not promote immune function, but act to decrease it instead. Despite their low numbers during an infection, these cells are believed to play an important role in the self-limitation of the immune system. They have been shown to prevent the development of various auto-immune diseases.

T cells emit lymphokines to trigger immune responses, but more importantly they bind to an antigen-presenting cell. They do it through their TCRs (T cell receptors) which bind to the MHC (major histocompatibility complex) of the cell. Depending on the class of TCR and MHC in presence the response of the T cell will be different. For this reasons, to characterize T-cells we extensively studied its cell surface complexes and the genes which produce them, as we will see in section 4.3 in which characterization of regulatory T-cell subsets.

4.2 First study : Characterization of Malaria severity

4.2.1 Malaria

Protozoan parasites are single-celled eucaryotes which require the "services" of one or more host, to live and reproduce. Malaria is the most common protozoal disease, infecting $200-300$ million people every year and killing $1-3$ million of them. It is caused by one of the four species of the Plasmodium parasite. It is transmitted to humans by the bite of the female of any of 60 species of Anopheles mosquito. Plasmodium falciparum the most intensively studied of the malaria-causing parasites exists in no fewer than eight distinct forms, and it requires both the human and mosquito hosts to complete its sexual cycle (see figure 4.10).

Gametocytes are formed in the bloodstream of infected humans, but they can only differentiate into gametes and fuse to form a zygote in the gut of the mosquito. Three of the Plasmodium forms are highly specialized to invade and replicate in specific tissues the insect gut lining, the human liver, and the human red blood cell. Even within a single host cell type, the red blood cell, the Plasmodium parasite undergoes a complex sequence of developmental events, reflected in striking morphological changes.

Sylviane Pied's group is studying the different immunological effects of Plasmodium falciparum invasion on the human body. Their main goal is to discover molecular markers of Malaria, in order to better prevent the development of the disease. To this end, they are working with the National Centre for Cell Science at Pune in India, which is an endemic country for this disease. In India different blood samples are extracted from a cohort of people having different types of malaria, and from three different control groups. The quantity of autoantibodies, and cytokines in patients are measured. We analyzed these data in cooperation with Sylviane Pied's group in Pasteur Institute of Lille. I will described now more in detail this analysis.

4.2.2 Discrimination of malaria severity using autoantiboy and cytokine measurements

There is different severity of Malaria, the most serious being Cerebral Malaria (CM), our analysis was mainly focus on this type of disease. It induces changes in mental status and coma. It is an acute, widespread disease of the brain which is accompanied by fever. The mortality ratio is between $25-50\%$, and can be fatal in 24 to 72 hours. The histopathological hallmark of this encephalopathy is the sequestration of cerebral capillaries and venules with parasitized red blood cells. Ring-like lesions in the brain are major characteristics. The key elements of Cerebral Malaria are [106] : (i) unrousable coma with no localizing response to pain persisting for more

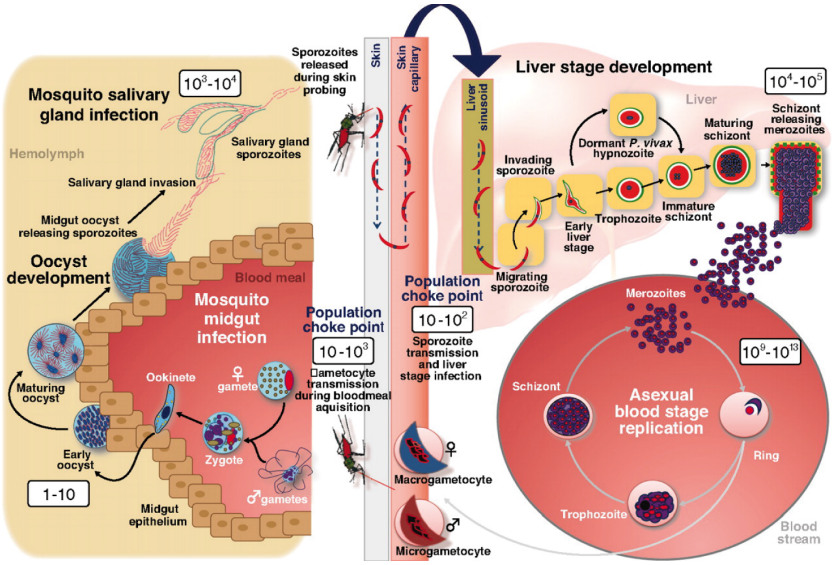


FIGURE 4.10 – *Overview of the sexual cycle of Plasmodium falciparum. This figure is extracted from [105].*

than six hours if the patient has experienced a generalized convulsion ; (ii) asexual forms of Plasmodium falciparum found in blood ; and (iii) exclusion of other causes of encephalopathy (*i.e.* viral or bacterial).

During a cerebral malaria infection, IgG autoantibodies to brain antigens are increased in patients as shown in a previous study on Gabonese childrens [107]. However, their role in the pathophysiology of cerebral malaria is not fully defined. To gain a better understanding, we studied the profile of IgG reactivity to brain proteins on different cohorts of patients. The quantity of IgG was measured using the PANAMA-blot method.

There were six different groups of patients analyzed, the first three were infected patients having different forms of malaria severities defined by the World Health Organization [108], the other three were control groups of healthy patients.

- Group 1, 42 patients : Cerebral malaria (CM) ;
- Group 2, 10 patients : Severe non-cerebral malaria (SM) ;
- Group 3, 16 patients : Mild malaria (MM) ;
- Group 4, 5 patients : Recovered cerebral malaria (ex-CM), constituted of subjects who had CM within the past 6 months and recovered ;
- Group 5, 14 patients : Endemic controls (EC), constituted of patient's relatives (brothers/sisters/parents) who accompanied the patient to the hospital and did not have malaria for at least the preceding 2 years, nor were they clinical asymptomatic carriers ;
- Group 6, 11 patients : Non-endemic controls (NEC), were the subjects residing in the Pune city with no history of malarial disease for 5 years.

The data analysis revealed that circulating IgG from CM patients highly reacts with recombinant beta tubulin III (TBB3) brain antigen, which was not already described as a discriminant for cerebral malaria. Autoantibodies were not the only molecular species studied in these different cohorts, quantity of specific cytokines were also measured. Their important role for the molecular characterization of Cerebral Malaria was demonstrated by Sylviane's Pied group and their collaborator [109] (see figure 4.11). We demonstrated in our analysis, using singular value

decomposition in the *correspondence basis*, a strong correlation between IgG anti-TBB3 and elevated concentration of cluster-II cytokines (IFNγ, IL1β, TNFα, TGFβ). All of them are now clearly identified as molecular markers of cerebral malaria. All the results of this study was published in Plos One journal in December 2009, the corresponding journal article can be found in appendix F.

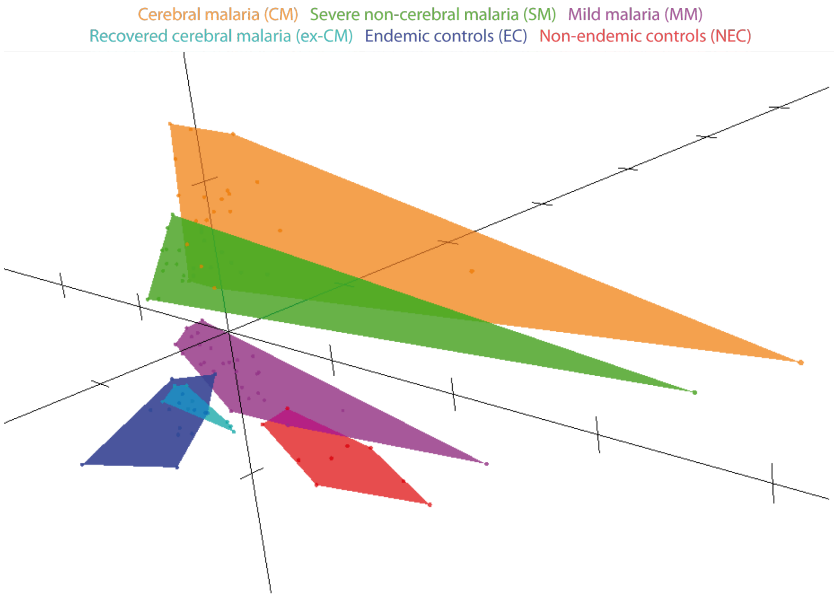


FIGURE 4.11 – *Representation of the different patients using their cytokine expression values.*

4.3 Second study : Characterization of regulatory T cell subpopulation

The data analysis on which I spent the most amount of time was in cooperation with the ImmunoRegulation group of Lars Rogge at Pasteur Institute in Paris. They are extensively studying the different populations of helper T-cell. Among the three existing types of differentiated helper T cell, we focused in our analysis on regulatory T cell (Treg). Those kind of cells regulate the number of Lymphocytes, to avoid auto-immune response. That is why, in the case of auto-immune diseases, such as AIDS, the amount of these cells is depleted. Then for a complete understanding of AIDS, characteristics and role of regulatory T-cells have to be clearly understand.

We studied in detail the transcriptome profile of regulatory T cell (Treg), confirming already established Treg specific genes and molecular markers, and searching for new ones. The first analysis I was involved in was to focus on the analysis of the Treg transcriptome evolution during a clinical trial of a new HIV treatment. Then, we characterized more deeply three types of Treg cells : naive Treg (nTreg), cytokine Treg (cTreg), and activated Treg (aTreg). Identified a novel transcription factor named FoxLF. Finally, we searched for ontologies specific of each type of Treg. Before developing the details of these analysis, I will briefly review the current knowledge on Treg molecular discriminants.

4.3.1 Specificity of Treg

T-cell play an immunological role mainly by cell communication using their different cell surface markers. For this reasons, for characterizing them, cell surface properties are extensively used as the different cell-surface receptors produced by a cell indicate its functioning. For example the quantity of CD4 (cluster of differentiation 4) produced is a major discriminant for Helper T-cells. CD4 is a co-receptor which help the TCR (T cell receptor) in activating the T-cell just after an interaction with an antigen-presenting cell has occurred. Cluster of differentiation (CD) proteins are usually receptors for interleukin (IL, *e.g.* a group of cytokines secreted by leukocytes), thus, this group of receptor proteins plays a major role in lymphocyte characterization.

As shown in figure 4.12, all T-cell lineages originate in the thymus and emigrate as naive $CD45RA+$ T cells. Activation of naive T cells in the periphery induces their differentiation into both conventional and regulatory subsets. Conventional T-cells can further differentiate into memory T-cells, which can then be reactivated. Although, conventional memory formation has not been described, T-reg cells have been shown to differentiate into terminal effector T-reg cells with unique cell surface marker expression. Also contributing to the $CD45RA-$ peripheral T-reg cell compartment are converted T-reg-like cells, which are derived from conventional T cells. These converted T-reg-like cells have cell surface marker expression similar to that expressed by natural T-reg cells.

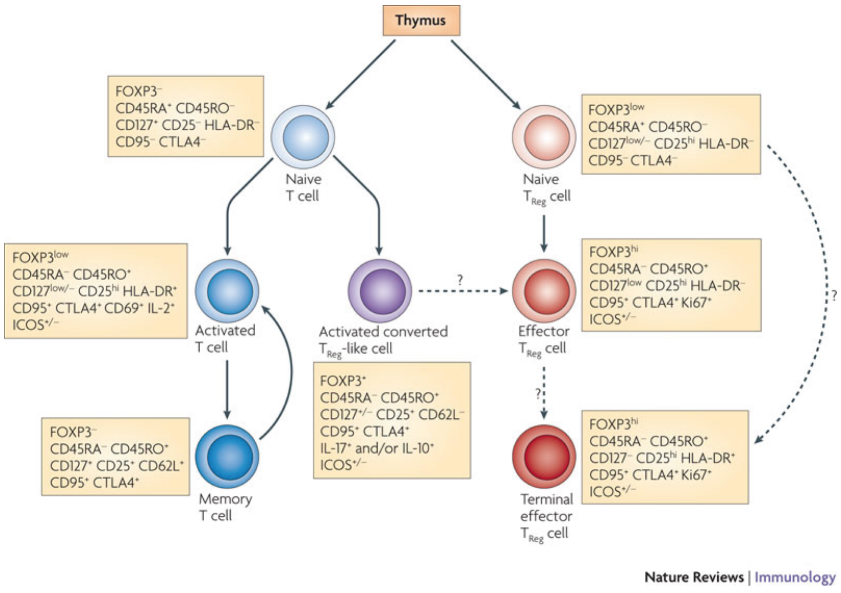
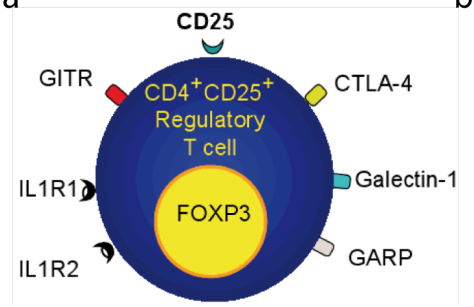


FIGURE 4.12 – *Phenotypic markers are indicated during CD4+ T-cell (helper T-cell) differentiation into the conventional T cell and regulatory T (T-reg) cell lineages. This figure is extracted from [110]*

The most important genetic marker used in Treg characterization is FOXP3. It is a transcription factor which is described since several years as a major regulator of the Treg cell lineage. Another discriminant between naive T cell and naive Treg is CD25. In figure 4.13, I present a list of all the others molecular discriminants of Treg cells. In the following we will characterize the different types of Treg using some of these markers.

a



b

Transcription factor	Activation and memory	Homing and origin	Suppressive and effector function	Apoptosis, survival or other
FOXP3	CD45RA CD45RO CD25 HLA-DR Lack of CD127 CD69	CD62L CCR4 CCR6 CCR9 CD103 CD304 CD31 Lack of CD49d	CTLA4 ICOS CD39–CD73 LAP Granzyme B Galectin 1 Galectin 10 TRANCE CD80 and CD86 IL-10 IL-17 CD2 Lack of IL-2	CD27 OX40 CD95 PD1 GITR Galectin 3 GARP MS4A4B IL-1R CD6

FIGURE 4.13 – *Presentation of the different Treg molecular markers. (a) Scheme of a Treg cell with its ten most important molecular markers. (b) A table showing the different known molecular markers of Treg extracted from [110].*

4.3.2 Effect of HIV on Treg

Human immunodeficiency virus (HIV) is a lentivirus (a member of the retrovirus family) that causes acquired immunodeficiency syndrome (AIDS). From its discovery in 1981 to 2006, AIDS killed more than 25 million people. HIV infects about 0.6% of the world's population. According to current estimates, HIV is set to infect 90 million people in Africa, resulting in a minimum estimate of 18 million orphans.

HIV infects primarily vital cells in the human immune system such as helper T cells , macrophages, and dendritic cells. HIV depletes the quantity of CD4+ T cells by three different mechanisms :

- First, direct viral killing of infected cells ;
- Second, increased rates of apoptosis in infected cells ;
- Third, killing of infected CD4+ T cells by CD8 cytotoxic T-cell which recognize infected cells.

As a results the number of helper T-cells decline drastically. When it reaches a critical level, cell-mediated immunity is lost, and the body becomes progressively more susceptible to opportunistic infections.

The Highly Active AntiRetroviral Treatment (HAART) is nowadays the most effective treatment against HIV. Even thought it does not eradicate totally the virus, it controls its replication. During acute infection (see figure 4.14-a), HIV replication is partially controlled by T cell responses, and depletion of the CD4+ T-cell compartment is limited. Because of viral cytopathic effects or immune-mediated killing, productively infected activated T-cells do not generally survive for long enough to revert to a memory state. A small pool of latently infected memory CD4+ T-cells harboring integrated HIV DNA, however, is established. HAART initiation during the acute phase (see figure 4.14-b) generally results in the normalization of CD4+ T cell counts and the preservation of memory T cell responses, which can subsequently contribute to the control of viral replication upon reactivation from stable reservoirs.

Chronic infection (see figure 4.14-c) is accompanied by depletion of the CD4+ compartment and exhaustion of HIV-specific T-cells, leading to uncontrolled viral production. HAART initiation during the chronic phase of the disease (see figure 4.14-d) generally abrogates viral replication, but CD4+ T-cell reconstitution is limited. This is associated with hyperimmune activation of T-cells of diverse specificities even in the absence of their cognate antigen. The profound depletion of memory CD4+ T-cells along with the exhaustion of HIV-specific CD8+ T-cells result

in the incapacity of the immune system to control sporadic reactivation events. Although viral dissemination is limited by HAART, *de novo* infection can occur and may contribute to HIV persistence.

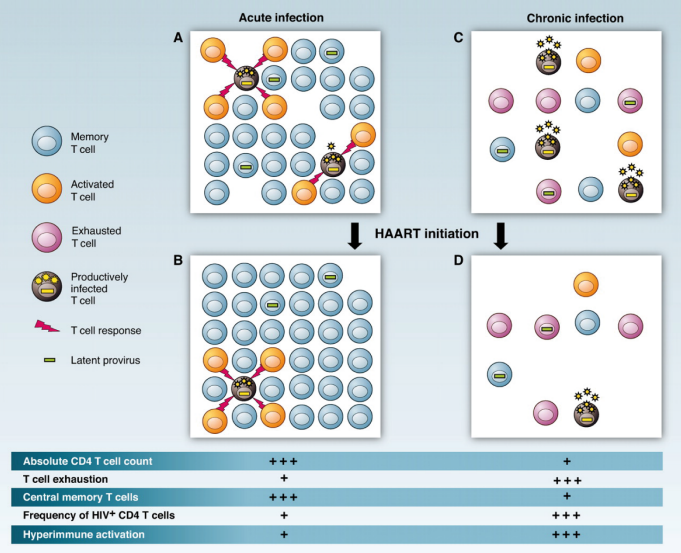


FIGURE 4.14 – *HAART effect on acute infection (a-b) and chronic infection (c-d). This figure is extracted from [111].*

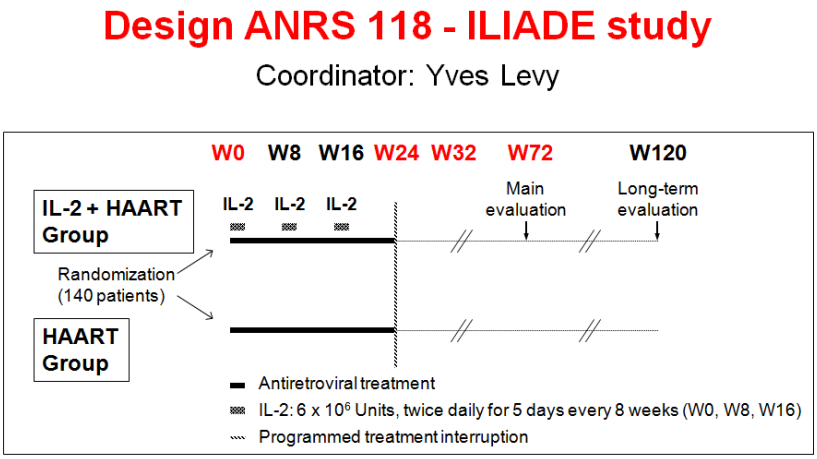


FIGURE 4.15 – *ANRS118 experimental design.*

Lars Rogge's group is part of an international consortium, supervised by Yves Levy, which has tested a new treatment for HIV patients. This treatment uses injection of IL-2 interleukin to increase the CD4+ T-cells count. This interleukin is known to be a growth and differentiation factor for T-cell. The clinical study (ANRS118 trial) consists in monitoring the evolution of 120 HIV-infected patients separated in two groups : one including patients treated with HAART and the other including patients treated with HAART+IL-2 therapy. Characteristics of each

cohort were screened at different weeks after the beginning of each treatment, as described in the experimental design in figure 4.15.

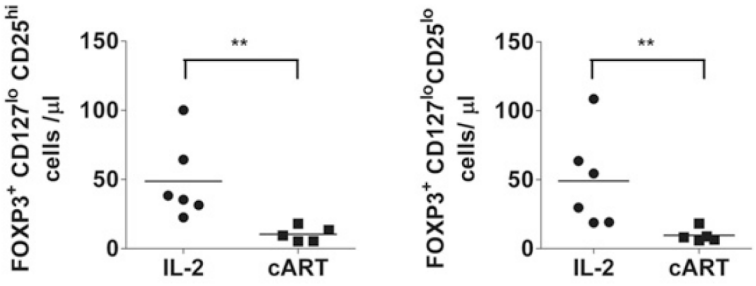


FIGURE 4.16 – *CD25+ and CD25- cell count measurement in (a) IL-2 and (b) HAART only (cART) treated patients, this figure is extracted from [112].*

The clinical study was not successful as treated patients with HAART+IL-2 did not reveal to have a significantly improvement in comparison to HAART only treated patients. However, it reveals to have a clear effect on Treg counts.

First, the absolute numbers of CD4+FOXP3+CD127lowCD25high cells (CD25+ cells, *e.g.* conventional Treg cells) and CD4+FOXP3+CD127lowCD25low (CD25- cells) in peripheral blood of 6 patients treated with HAART+IL-2 and 5 HAART only treated patients controls was measured, and showed a real difference between two groups, as shown in figure 4.16.
CD4+FOXP3+CD127lowCD25high (CD25+ cells) and CD4+FOXP3+CD127loCD25low (CD25- cells) are two types of cells that shared phenotypic markers of Treg but could be distinguished by the levels of CD25 and FOXP3 expression.

Second, we studied the transcriptome profile of CD4+CD25+FOXP3+ regulatory T-cell, and demonstrated no difference between those cells before and after IL-2 expansion. To make this statement we looked at differentially expressed genes using the substract tool of Ace.map (see section A.3.2), and we found no highly differentially expressed genes. This last results is illustrated in figure 4.17, in which, using GEO module of Ace.map (see appendix A) in *correspondence basis*, on can see differences between transcriptome profiles of CD25+ and CD25- cells, but no difference between these cells at week 0 and 24 of the clinical study. Moreover an *in vitro* study as shown that these IL-2 expanded cells have the same power of effector T-cell regulation.

Combining all results found, one can suppose that patients treated with HAART+IL-2 have only the cells with regulatory activities which are multiplied, in consequence, the number of regulatory T cell increase in comparison to other helper T-cells, but the total count of helper T-cells does not increase, and so does the health of patients. Some of these analysis were published in Proceedings of National Academy of Science in June 2010 [112].

4.3.3 Transcriptome profiling of Treg

Regulatory T-cell purification

As I already explained, three major molecular markers of regulatory T cells are already known : FOXP3 (Treg major transcription factor), CD4 (cell surface protein), and CD25 (cell surface protein). Recent studies have shown that two other discriminants exist which divide Treg into three sub-populations. First, Baecher-Allan *et al.* showed that CD25highHLADR+ cells were FOXP3high and highly suppressive whereas CD25hiHLADR- cells were FOXP3low and had lower

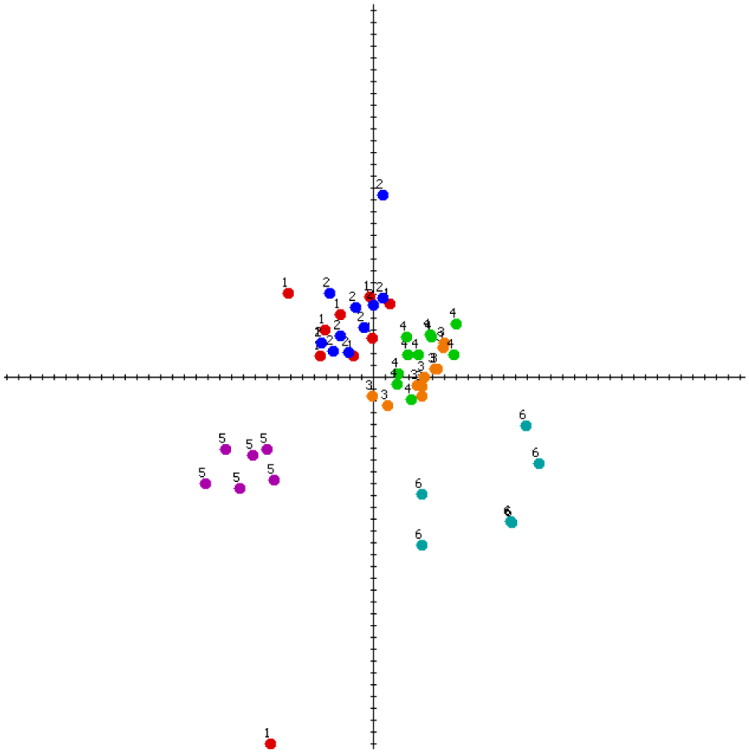


FIGURE 4.17 – *Different type of cells sorted from ANRS118 study are plotted using Geo module in correspondence basis and their gene expression values measured with transcriptome microarray. There are six groups of cells displayed : (in red) CD25+ cells at week 0, (in blue) CD25+ cells at week* 24*, (in orange) CD25- cells at week 0, (in green) CD25- cells at week* 24*, (in purple) CD25+HLADR+ peripheral blood cells extracted from the study described in the following section, (in dark green) CD25- cord blood cells. All instances were plotted using a selection of* 65 *probes being the most positively differentially expressed probes between CD25+ and CD25- cells.*

suppressive activity in vitro [113]. Second, Miyara *et al.* described three different subpopulations of CD4+FOXP3+ that could be identified based on the expression of CD45RA and CD25. CD25++CD45RA+FOXP3low resting Treg (rTreg), CD25+++CD45RA-FOXP3high activated Treg (aTreg), which represent different stages of Treg development and are both suppressive in vitro and CD25++CD45RA-FOXP3low cytokine secreting T cells which lack suppressive activity [114].

HLA-DR is a major histocompatibility complex of class II (MHC class II), and CD45RA is a form of the Protein tyrosine phosphatase receptor type C enzyme. To determine if both HLA-DR and CD45RA are discriminant of Treg sub-populations, we purified a set of CD4+ T-cell populations using their cell surface markers, to obtain 4 different types of T-cells :

- aTreg (activated Treg) CD4+CD25hiCD127-CD45RA-HLADR+ cells ;
- cTreg (cytokine emiter Treg) CD4+CD25hiCD127-CD45RA-HLADR- cells ;
- nTreg (naive Treg) CD4+CD25hiCD127-CD45RA+HLADR- cells ;
- CD4+CD25-CD45RA+ cells.

The population of cells came from buffy coats of 13 healthy donors. The purification of the cells was performed on a FACS Aria II platform. This instrument can sorted four different types of

cells in a single run. It uses cell surface fluorescent markers, to determine the quantity of each marker. Then the user defines cutoffs of selection on the amount of each markers, as shown in figure 4.18, the sorting is performed according to them. As an intracellular staining shows that the three T-reg populations were all CD25high, we were able to sort these three fractions to at least 95% purity from peripheral blood. Another sorting was also performed using 13 samples of CD34+ depleted cord blood samples. Cord-blood T-cell are usually not much differentiated, for this reasons we found almost exclusively Tconv and naive Treg cells.

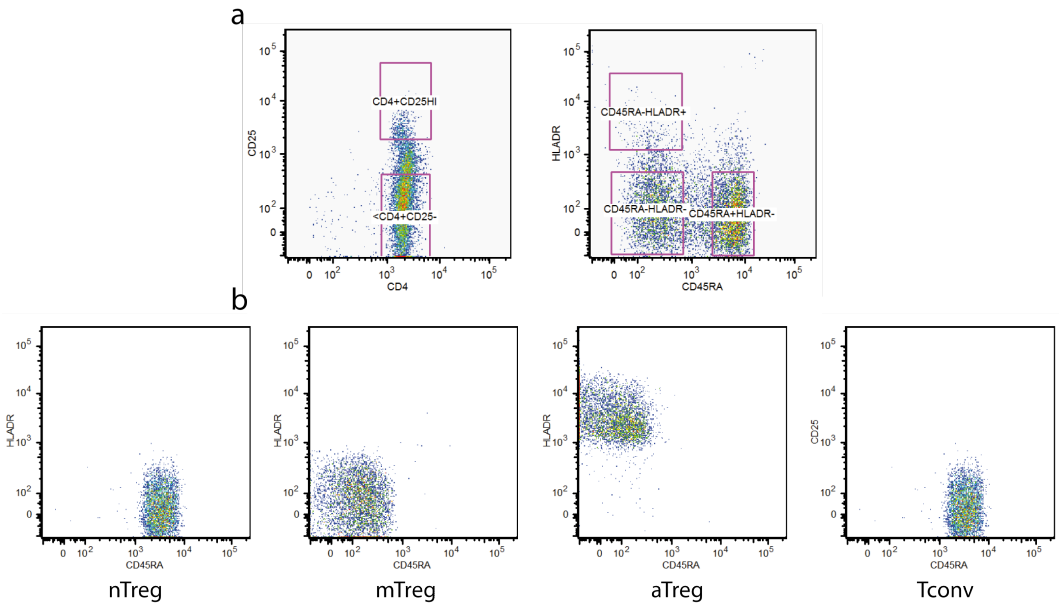


FIGURE 4.18 – *Results of the CD4+ cells sorting performed on FACS Aria II using CD4, CD25, HLA-DR and CD45RA cell surface markers. (a) The cell surface markers distribution of all cells for our samples. The* 5 *different purple rectangles are the cutoff defined for the selection of CD4+ sub-populations. (b) CD45RA and HLA-DR cell surface markers distributions of the* 4 *T-cell sub-populations.*

The defining feature of Treg cells is their ability to suppress pro-inflammatory immune function. Considering that in humans FOXP3 expression does not necessarily confer suppressive activity [115], we assess the "regulatory" power of the different sub-populations sorted. The method utilized to examine regulatory cell function was the *in vitro* co-culture suppression assay, in which potential Tregs are added, often in decreasing numbers, to CD4+CD25- responder cells (Tresp), and the cell proliferation is measured. We found that the three sub-populations have all regulatory functions three distinct sub-populations of Treg cells.

Gene expression profiling

The first step of the transcriptome analysis of the four groups of cells was to analyze by quantitative RT-PCR the amount of transcript of known Treg specific genes, in order to verify the difference of expression among all groups. After a 24hour stimulation with CD3/CD28 beads of the different sorted cells, the gene expression was measured using quantitative RT-PCR using Low density array techonology from Applied Biosystems, the results of this analysis are shown in figure 4.19. It confirms the role of Treg molecular markers already described in figure 4.13, such

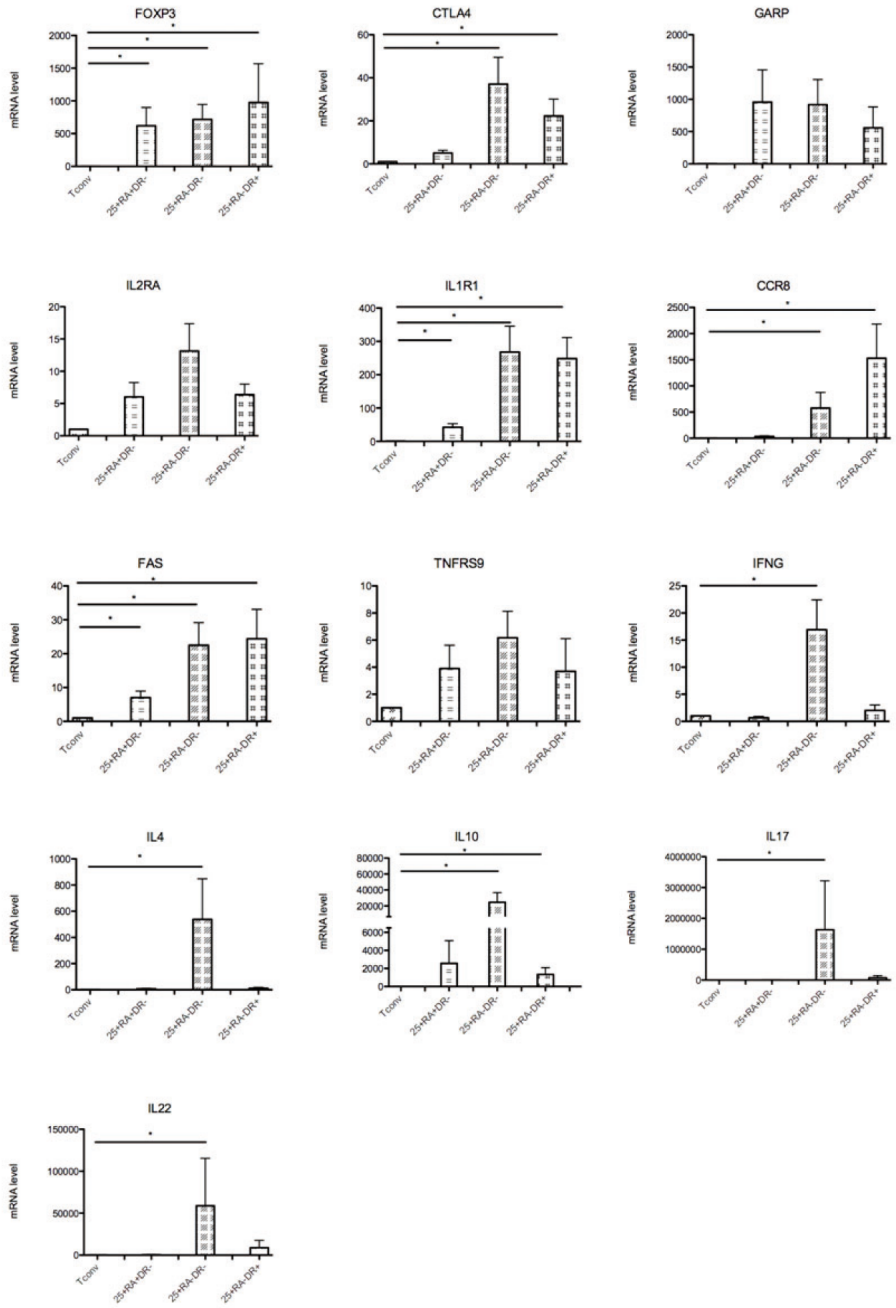


FIGURE 4.19 – *Results of quantitative RT-PCR on different Treg markers. Mean expression values plus SEM of six different donors are shown. Statistical comparison, indicated using* $*$*, were performed by Wilcoxon matched pair test.*

as FOXP3, IL-2RA, FAS, and CTLA4. Moreover it shows that CD4+CD25hiCD127-CD45RA-HLADR- cells are the only one to produce the different interleukin IL-10, IL-4, IL-17, and IL-22, confirming their role as cytokines emitter cells. And confirming results of previous study which have shown that activated regulatory T cell emits few cytokines, IL-10 being the only exception

as it is known to be a specific marker of Treg.

Then, we analyzed the transcriptome of each group of cells using Applied Biosystems microarray chips (see appendix A). For this analysis we used all modules of the Ace.map software described in section A.3. The goal was to determine sets of genes specific to the three types of Treg. To this end we extensively used the Substract and the Filter modules of Ace.map. The Substract module helped us to characterize the difference of expression between two set of biological conditions by calculating a fold change and a p-value. By selecting all probes having a p-value inferior to 0.01, with the Filter module we obtained different signatures. A signature is a set of genes differentially regulated between different biological conditions.

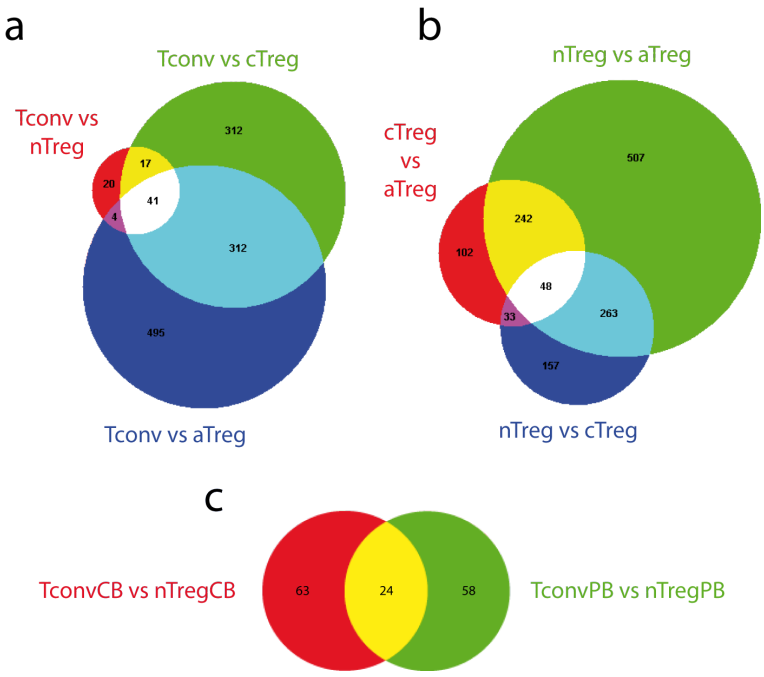


FIGURE 4.20 – *Three different Venn diagrams showing the number of probes extracted from different comparison of Tconv, nTreg, cTreg, and aTreg. (a) Selection of probes specific to Treg in comparison to Tconv, in white the CoreSignature, the sum of all circles gives the TotalSignature. (b) Characterization of probes specific to each Treg type, in white the CoreSignatureTreg, and the sum of all circles give the TotalSignatureTreg. (c) Comparison of probes specific to naive Treg cells in peripheral and cord blood.*

We determined different signatures (set of probes) specific to each type of Treg, first by comparing Tconv cells to all types of Tregs. Using these comparisons, we defined a CoreSignature of 41 probes which corresponds to genes highly differentially expressed between Tconv and all Tregs. We also defined a TotalSignature corresponding to 1201 genes highly differentially expressed between Tconv and at least one of the type of Treg. Then we determine signatures specific to each Treg type by comparing aTregs to nTregs, cTregs to nTregs and aTregs to cTregs. A CoreSignatureTreg was created including all probes which play a major role in the comparison of the Treg types. Figure 4.20 summarize with Venn diagrams the different signature created, indicating for each the number of probes. The last series of signatures we have created includes probes which are specific to naive Treg both in peripheral and cord blood, in order to assess the differentiation of naive cells which occurs during development. Those signatures were created

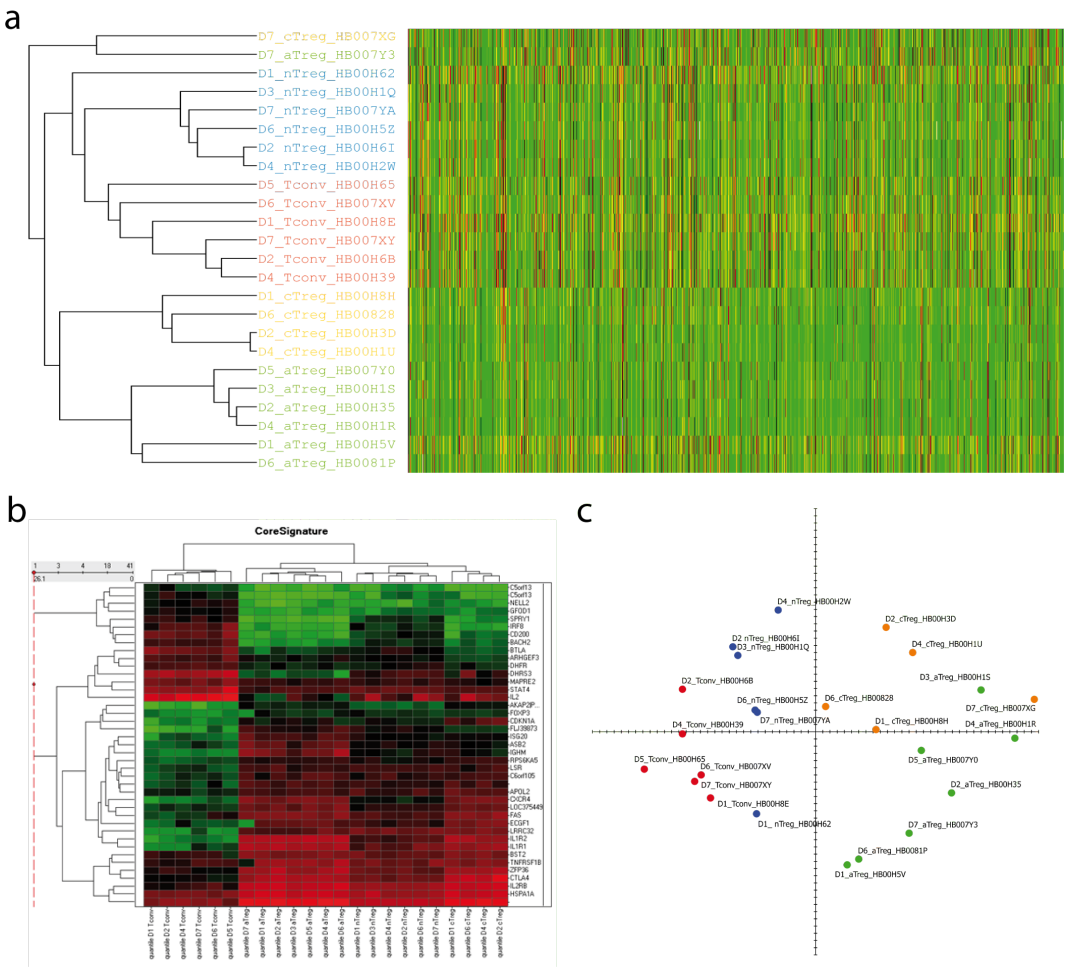


FIGURE 4.21 – *Clustering and GEO module utilization on the T-cell microarray data. (a) Hierarchical clustering of the different instances using the* 1201 *probes of the TotalSignature. Perfect clusterization of the different types of cell is revealed, excepted for D7 patient which was already detected has an outlier in the experiment. (b) Hierarchical clustering using the* 41 *probes of the CoreSignature also reveals perfect clusterization. The heatmap illustrated the different patterns of expression of the genes specific to Treg. (c) A GEO representation in the correspondence basis of all Tcells measured, using TotalSignature. The evolution of Treg differentiation appears by going from the left to the right of the representation.*

using comparisons of Tconv and nTreg microarray data for peripheral and cord blood (see figure 4.20-c).

These different signatures were used in different modules of Ace.map. For example, hierachical clustering of the different biological conditions were performed using only the TotalSignature composed of 1201 (see figure 4.21-a). It shows a perfect clusterization of the different groups, demonstrating the quality of the signature. Another hierchical clustering was performed using only the 41 probes of the CoreSignature (see figure 4.21-b). It also demonstrated a perfect clustering of these types of cells. Using the TotalSignature a GEO representation was calculated using *correspondence basis*. It clearly shows the differentiation process of Treg cells, as by going from left to right in the representation one sees the switch from Tconv to nTreg to cTreg and aTreg.

All the sets of genes determined to be specific to Treg will be soon validated by single cell PCR experiments, and other studies. The results will then be published with all the other analysis previously described in this section.

Discovery of a new transcription factor in regulatory T-cell

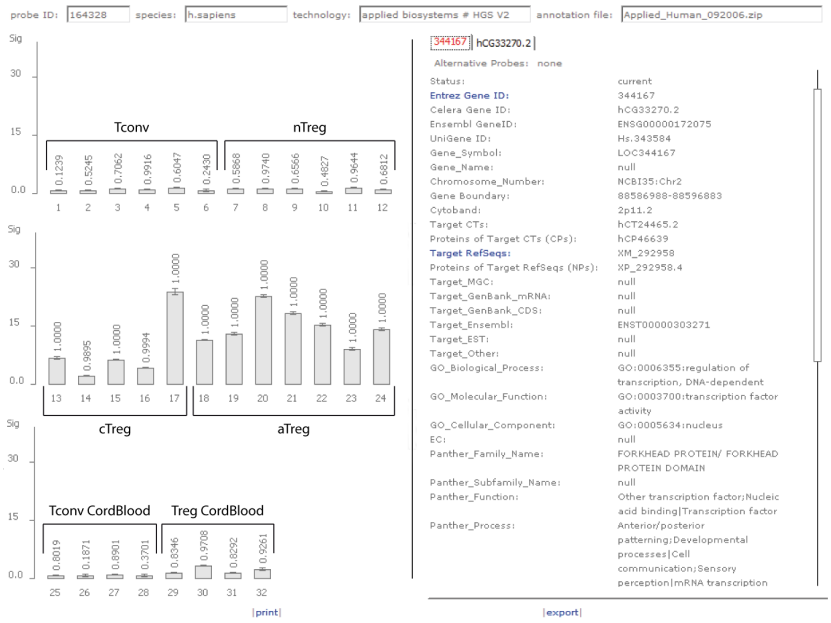


FIGURE 4.22 – *Print-screen of Ace.map's probe card showing the different expression values of FOXLF among all biological conditions. cTreg and aTreg instances revealed to be the only one in which FOXLF is over expressed.*

Among all the new Treg markers found using on the transcriptome analysis, is a predicted gene named LOC344167. The prefix *LOC* signifying that its function is not yet knwon (*e.g* the gene is not annotated). This gene has a sequence of 1263*bp*, is located on chromosome 2, and is predicted to encode for a protein of 420 amino acids with a forkhead DNA binding domain. Therefore this new gene seems apparent to the same family of FOXP3. The predicted secondary structure of the resulting protein shows also structure similarity with FOXP3. For these reasons we called it FOXLF (FOX like factor). This new transcription factor revealed to be over-expressed

in aTreg and cTreg, as shown in figure 4.22, thus it may play a great role as a major transcription factor for "regulatory" function of Treg, such as FOXP3.

a

Without specific signature	Tconv VS nTreg	Tconv VS cTreg	Tconv VS aTreg
T cell activation			1.1526235980759858E-7
Apoptosis signaling pathway	0.013750584832824642		3.3510596460201386E-5
DNA replication			5.1807955598034036E-4
p53 pathway			0.026173838882524152
Cadherin signaling pathway			0.043360707967435 79
Inflammation mediated by chemokine and cytokine signaling pathway		3.053120518081709 6E-5	
Interleukin signaling pathway		0.01647828725888609	

b

With Bonferroni correction	Tconv VS nTreg	Tconv VS cTreg	Tconv VS aTreg
Amino acid activation			5.7229600679563766E-6
Amino acid biosynthesis			0.023492858838202884
Amino acid metabolism			2.2285360536031723E-4
Apoptosis	7.216536750049253E-4	0.0010855199702934 60	8.029386046879083E-10
Carbohydrate metabolism			0.003298542150645061
Cell adhesion			0.007150752389849624
Cell cycle		1.429905124150376 3E-8	2.2279994217160926-19
Cell cycle control			2.2387661121038297E-6
Cell motility			
Cell proliferation and differentiation		0.002811388254339845	
Cell surface receptor mediated signal	1.786961253684435E-6	0.001727870523052443 7	
Chemosensory perception			5.664584186805076E-6
Chromosome segregation			0.005605874480184412 5
Coenzyme and prosthetic group metabolism			1.2163945095158168E-5
Coenzyme metabolism			0.003557804831254535
Cytokine and chemokine mediated s	1.9731616346536986E-11	1.4965568700704186E-1	0.015105541668897178
Cytokine/chemokine mediated immu	0.002482397847362445	1.3547579711273 3E-5	
DNA metabolism			8.472099751120622E-15
DNA repair			1.1231294022895534E-4
DNA replication		0.026101297180434908	7.888237293911471E-13
Extracellular transport and import			0.021874836907975155
General mRNA transcription activities			0.04539306455350542
G-protein mediated signaling			0.007197525565053106 5
Homeostasis		0.035233666370361684	
Immunity and defense	1.787479638360390 2E-7	1.5304104463254 7E-18	2.530259922508838-10
Induction of apoptosis	0.015492090789181 71		0.003492657360228962 3
Interferon-mediated immunity		4.150294883860384E-6	0.013574011593342934
Intracellular protein traffic		0.023918262940810 14	1.9029984976289E-7
Intracellular signaling cascade		4.507936288072982E-6	5.662912000164115 4E-5
JAK-STAT cascade		4.947062814290831 5E-4	
JNK cascade			0.010247437258046646
Ligand-mediated signaling		0.003968954806572774	
MAPKKK cascade		0.042803520918617 7	
MHCII-mediated immunity		0.008728563188562802	3.401904964755355E-7
MHCI-mediated immunity			0.014498434085809838
Mitosis			5.687157445663796E-4
mRNA splicing			3.860418949810831E-4
Nucleoside, nucleotide and nucleic acid metabolism			3.158070874055994 4E-35
Olfaction			1.633122947004169 7E-6
Oncogenesis		1.015052378113648 4E-4	0.001235054538979856 2
Other homeostasis activities		0.004170027546813767	
Other intracellular signaling cascade			0.027848073817892927
Other metabolism			2.307963356225047 6E-6
Pre-mRNA processing			3.962287523542985E-5
Protein biosynthesis		4.217899400347486 6E-4	
Protein complex assembly			9.574841857256122E-11
Protein folding			4.287135831216884E-8
Protein metabolism and modification			2.332815440853816 8E-8
Protein methylation			0.015004796964391479
Protein modification			3.325916044583111 7E-4
Purine metabolism			
rRNA metabolism			3.377279380014806 5E-18
Sensory perception			2.723519355834201 5E-6
Signal transduction	1.406460359554598E-8	6.938120000147474E-7	
Sulfur metabolism		1.567710594439381 5E-6	
Sulfur redox metabolism		0.014612297767131119	
T-cell mediated immunity		2.140362836140929 5E-6	1.991857992835177 5E-17
Tconv VS aTreg, specific signature	bonferonni correction		
Translational regulation			0.042264179540590245
Transport			
tRNA metabolism			6.994186433882453E-4

c

With Bonferroni correction	Tconv VS nTreg	Tconv VS cTreg	Tconv VS aTreg
Acetyltransferase			0.029899953774438003
Aminoacyl-tRNA synthetase			3.373386341058598 2E-6
Cadherin			1.965719132279947E-4
Cell adhesion molecule			0.028686098975654 85
Cell junction protein			0.017651739527849 21
Chaperone			4.559943359592756E-14
Chaperonin			2.275538984908941E-6
CREB transcription factor		0.034186189500707 2	
Cysteine protease			0.023010777821415328
Cytokine	0.001091561717864	1.211384697028447 2E-6	
Cytokine receptor	2.860232293414844	2.453201517820144 7E-5	
Damaged DNA-binding protein			0.007455501213453077
Defense/immunity protein		5.059188570049935E-5	4.220284906748917E-5
Dehydrogenase			7.070295206374445E-5
DNA helicase			0.048355959834432324
DNA-directed DNA polymerase			0.009180585535919474
Exoribonuclease			0.001733057169724982 9
Extracellular matrix			2.102733595124525E-8
Extracellular matrix glycoprotein			0.007436251360315362
G-protein			0.003280536937941129 4
G-protein coupled receptor			4.320638120043365E-5
G-protein modulator			0.029665213692132706
Helicase			1.450348635249482E-8
Hsp 70 family chaperone			0.004244519494180301
Hydrolase			8.128733595031018E-5
Immunoglobulin receptor family member		0.004351883557514573	0.001349880064180133
Interleukin		3.061167920664709E-7	
Interleukin receptor	3.568040237499064	0.021847854877139257	
Intermediate filament			0.001493412679789942
Ion channel			0.004259117830040 47
Kinase			0.005151834695338121
Kinase inhibitor	0.015100008597854	0.003869577235510958	0.047215818241787916
KRAB box transcription factor		0.004894926191727763	0.014530342045978053
Ligase			9.657246895672944E-7
Major histocompatibility complex antigen			1.816208995224767 4E-9
Membrane traffic protein			0.043439913444261 78
Methyltransferase			2.203005473373218 2E-8
Nuclease			0.004625318679581797
Nucleic acid binding			3.909127967070480 6E-20
Nucleotidyltransferase			0.001158774826919337 3
Other chaperones			1.937201358132572 1E-4
Other cytokine	0.002525629968823187		
Other G-protein modulator			0.010278680135438766
Other ligase			0.001442634371231016 1
Other miscellaneous function protein			4.909273796659E-5
Other nucleic acid binding			7.943397449909557E-4
Other RNA-binding protein			6.570543997844 37E-8
Other transcription factor		5.110578746716152E-6	0.008063810444888464
Phosphatase			0.048522786061001 29
Receptor	1.587257027521 05	5.655255199495299E-4	
Replication origin binding protein			0.032681681059428015
Ribonucleoprotein			1.849213873882440 1E-4
Ribosomal protein		1.078098918724966 6E-5	1.631698327302125 4E-6
RNA helicase			1.442322557684715 7E-8
RNA methyltransferase			2.639566996726958 6E-4
Select regulatory molecule		1.859761781202011 2E-4	1.335385734487818 5E-7
Serine/threonine protein kinase receptor			0.018757420398941606
Signaling molecule		0.003371848457431297	
Small GTPase			0.022326779111364 62
Storage protein		8.355098148127063E-6	2.106956469340 89E-4
Synthase and synthetase			3.221805196831719E-11
Synthetase			2.134094126365322 1E-10
Transfer/carrier protein			0.024695130565146054
Transferase			3.165172725070038E-13
Translation factor			8.377092956180828 5E-6
Translation initiation factor			9.804559451085858-9
Tubulin			0.011386284099835433

FIGURE 4.23 – *The different interesting ontologies found for highly differentially expressed probes between Tconv and the different groups of Tregs. For each type of Ontology : (a) Pathway, (b) biological process, (c) molecular Function, the p-value given by LEO module is indicated, the red color of cell's table indicates over-expressed ontology, whereas blue color indicates an under-represented ontology.*

Ontology analysis of regulatory T-cells

We finally extend our analysis to the protein and pathway levels. First, by performing a Leo analysis on the TotalSignature and the different groups of Tregs to search for Pathways, Molecular

Function and Biological processes over-represented. Result of this analysis are presented in figure 4.23. We found 6 important pathways which play major roles in the different types of Treg. Among those pathways, common T-cell specific pathways are found : T-cell activation, Interleukin signaling, Cadherin signaling. This latter fact proves once again the applicability of the signature we have constructed.

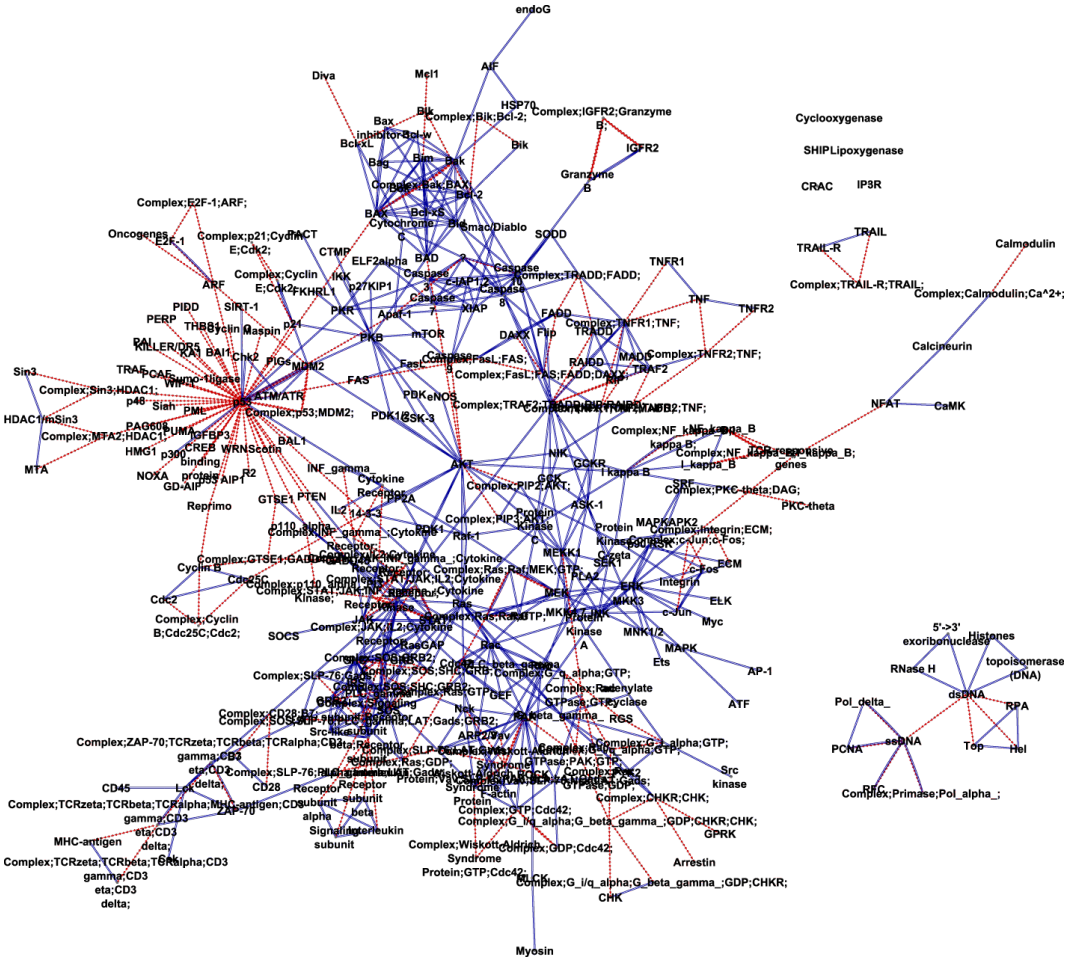


FIGURE 4.24 – *Representation of a sub-network of the ProbeNetwork constituted of species which are part of one of the* 6 *major pathways revealed by LEO analysis, and are associated to a probe. Red lines are unidirectional ARCS, blue lines are bi-directional EDGES*

To get a clear picture of the proteins which are over-expressed and which induce the over-representation of some pathways, we utilize the global pathway software to map LEO and Substract module results on two types of network. First, we mapped these data on the ProbeNetwork (see figure 4.25-a) which contains all proteins, genes and complexes associated to a probe (1331 nodes and 2410 edges, see section 3.2.6). Second, we mapped data on the Network constituted of the species which are both associated to a probe, and associated to one of the 6 important pathway described above. This last network is shown in figure 4.24, without any "expression" value mapped onto it, just to have a clear view on the proteins involved. On figure 4.25-b, the same network is represented, with LEO results mapped on it, and also the logarithmic fold change value of the differentially expressed probes. Only probes with a p-value inferior to 0.01 according

to Substract tool of Ace.map software are represented, that is to say probes already selected in the different signature described in section 4.3.3. This final analysis needs more development to extract reliable biological information, however it already gives a good general view of the Treg differentiation process at the protein scale.

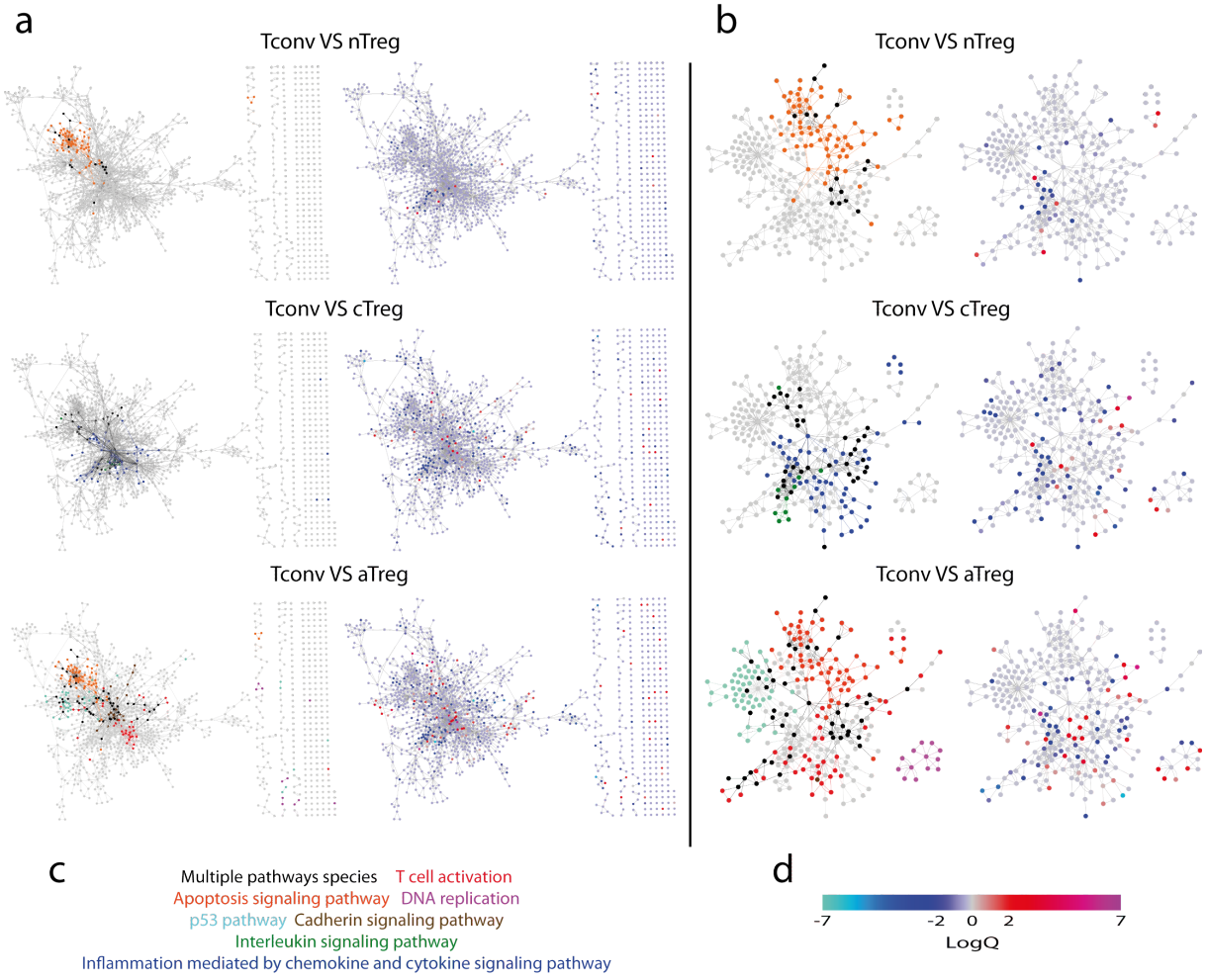


FIGURE 4.25 – *LEO and LogQ, obtained by comparing Tconv to the different groups of Treg, are mapped on ProbeNetwork (left of panel (a) and (b)) and the network of the* 6 *important pathways for Treg (right of panel (a) and (b)). (c) The color code of the different pathways represented is indicated. (d) The color code for logQ representation is indicated. LogQ is the logarithm of the fold change of a probe between two biological condition.*

Conclusion and Perspectives

In this manuscript I have shown the different projects on which I have worked during my thesis. I first described in a review all the state of the art techniques for dimensionality reduction. Then I have shown how by combining Singular Value Decomposition and Multidimensional Scaling one can obtain an accurate, computationally efficient, and easy to use technique of dimensionality reduction. In the third chapter I have described the creation *ex nihilo* of a software which allows to map different measured properties linked to gene expression onto protein-protein networks. Finally, the utilization of these tools and many others described in appendix 1, have lead to the discovery of new molecular markers for Cerebral Malaria and regulatory T-Cells.

The two biological studies described in chapter 4 are not finalized yet, more refinement are needed. For example, in the Treg characterization study, single cell quantitative RT-PCR will be performed in order to assess the role of the different genes specific to Tregs. Concerning, the SVD-MDS algorithm, some improvements may also be envisioned. Independent Component Analysis might be a good candidate to replace SVD in the initialization of MDS, it would give more discriminating power to the dimensionality reduction algorithm. Finally, Global Pathway Analysis software would be of better use if we increase its links with both Ace.map and Cytoscape. The integration with the former would help us to more easily embed gene expression data coming from the different Ace.map's modules. The integration with Cytoscape would allow to create networks of protein-protein interaction from other source then Panther, and map Acemap's results onto it.

I can continue to describe possible improvements indefinitely. As I have discovered during my thesis, this is in fact a characteristic shared by almost all bioinformatic tools, every one of them can be infinitely enriched. Consequently, a good approach when developing a computational tool is not to search for the more powerful one which can do many things, neither a very specialized one, but rather a balance in-between.

This latter fact is link to the general problem of developing a software, which is a pure informatics question which I had to ask myself many times. On the opposite, I always had to keep in mind the biological constraints which accompanied every measurement. On a personal point of view, this is for me the most important thing I have learned during my thesis : Making the bridge between pure computational and mathematics problems, and pure biological questions. This is for this kind of challenge that I enjoyed a lot my thesis, and that I will continue to work in the Interdisciplinary field of Systems Biology.

A

Microarray : Principle and analysis

For the understanding of the different elements and interactions involved in the fate of a cell, a large variety of powerful tools are available nowadays which allow to screen at the same time a huge number of elements. The class of such high-throughput hardwares which have known the biggest development in the past fifteen years are microarray chips. In the laboratory where I have done my thesis, we studied the transcriptome of different cells, to this end we extensively used specifics microarrays designed for screening all mRNAs which are produced in a cell. In this appendix, I will briefly describe the principle of these microarrays. For analyzing all the data coming from these chips, we developed in the laboratory a software called Ace.map. It allows different types of analysis, regroup all in modules. I will describe all of them in the second part of this appendix.

A.1 Transcriptome microarray principle

Transcriptomes microarrays all rely on the principle of oligo-nucleotide hybridization printed on a chip to complementary RNAs (cRNA) which are clone of messenger RNAs (mRNA) of the studied cell (see figure A.1). Each oligo-nucleotide, representing one important strand of DNA from the studied species, is compartmented in a given probes fixed on the microarray. Each cRNA is labeled with fluorescent or luminescent molecules, the more hybridization of mRNAs in a specific probe one has the more light will be created. Therefore, from the intensity of light one will know what was the quantity of mRNAs corresponding to the probe produced.

The first step of each microarray experiment is to get mRNAs from a giving sample of cells, and clone these strands in complementary DNA (cDNA), which will be replicated (RT-IVT process). To have a clear signal it is better to have cell with the same phenotype, otherwise the transcriptomic pattern will be an average on all the different phenotypes. But the more the cell sample is purified the more amplification processes are needed to obtain a sufficient quantity of cDNA. Consequently an equilibrium has to be found, otherwise the non purification of the cell or the wide amplification of cDNA will induce bad signal increasing the chance of having missing values.

As said, each microarray is a cluster of probes, in each a specified sequence of nucleotides is fixed on the microarray's plate. After having transform cDNA to cRNA, one injects them on the microarray, and just by hybridization, it will be targeted by the right probes, containing complementary sequence. Each manufacturer has its own design for probes, for example Affymetrix company chooses to have small oligo-nucleotide but put eleven different in one probes, whereas Applied Biosystems choose to use long oligo-nucleotide ($60 - mers$). For the choice of

the probe sequence manufacturer uses references genomes, but also other source of information. It exists less then ten different commercials chips which screen all human mRNAs, the most used being Affymetrix microarray. A vast number of scientific groups use microarray self-designed, that is to say they have chosen themselves the different probes oligo-nucleotide and printed it on chips using mainly ink-jet microarray printer. These laboratory made chips have the advantage of being specific to the experiment one wants to performed.

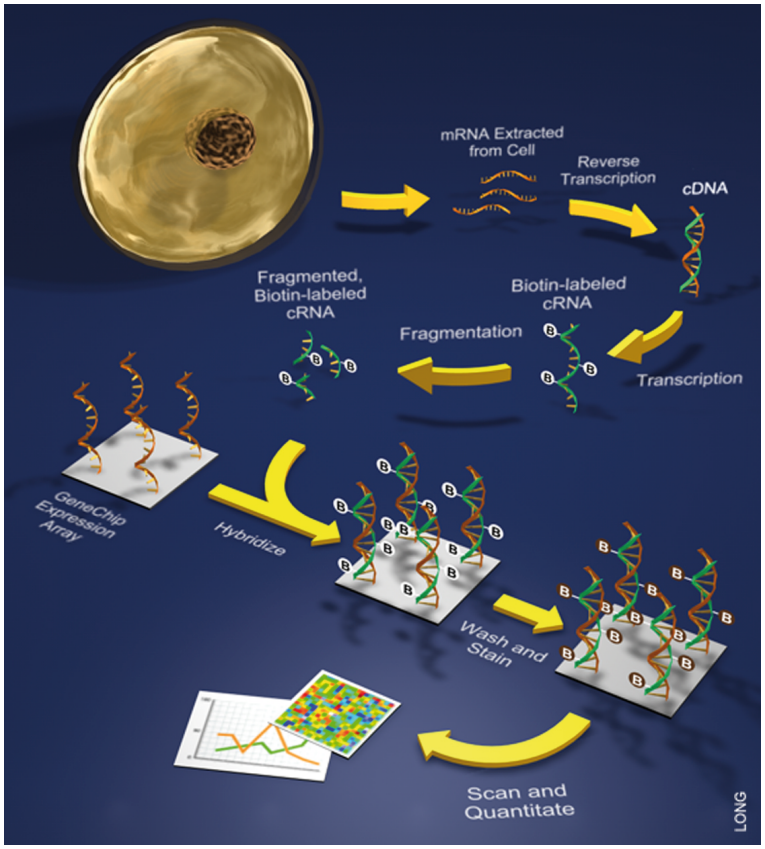


FIGURE A.1 – *The different step of a transcriptome microarray experiment are described here. This image extracted from [116].*

The choice of oligo-nucleotide probes is a huge problem, as several mechanisms may influence the sequence of mRNAs produced in a cell, such as alternative splicing or allele specificity of the cell. This variability may be a source of difficulty to identify specifically the different variants via microarrays. Indeed, for some transcripts, using a single probe per gene is not enough not differentiate between two alternative transcripts and it becomes necessary to design several probes for a transcript in order to measure unambiguous expression.

The design for the light label can also be different, fluorescence for Affymetrix, chemo-luminescence plus fluorescence for Applied Biosystems. At the end of the experiment one obtains an image of the quantity of light emit or detect in each probes giving an information on how much RNAs were produced in the studied cells. An example of microarray image directly extracted from an Applied Biosystems chip is shown in figure A.2.

FIGURE A.2 – *Image of microarray chemo luminescence extract from AB1700 technologies (Applied Biosystems).*

A.2 Applied Biosystems microarray technology

In this section, we will see more in detail the characteristics of Applied Biosystems microarrays, which are the type of chips used in the laboratory and by our collaborator during my thesis. The first main difference between Applied Biosystems chips and others rely in the probe design. These probes are oligonucleotides of 60 nucleotides in length ($60-mers$), which is longer than the probes of the Affymetrix chips for example, with only 25 nucleotides. The length of the probes influences the sensitivity and specificity of these. Longer probes are more sensitive but less specific because of the greater chance of partial mismatch. These chains of nucleotides are modified at their $3'$ end in order to bind covalently with the chip. In most cases, the probes correspond to a region of the gene located in its first 1500 nucleotides of the $3'$ end to facilitate hybridization. Moreover, they are often chosen in the non-coding $3'$ part ($3'UTR$), as this area is more specific for each gene, limiting the effects of alternative splicing. To address the problem that a gene can have many splice variants, the probes are selected so that their sequence corresponds to a zone of exons or $3'UTR$ shared by the largest possible number of variants. Other criteria are used in selecting the probes in order to cope with problems posed by genetic polymorphism in repetitive sequences and differences between the public version of human genome and the version of the Celera. Although 85% of genes are represented by a probe on Applied chips, the remaining 15% of genes are represented by several probes ($2-12$).

The annotation file used to describe the genes detected by the chip is obtained by correlating information from different public databases such as GenBank, Swiss-Prot, Medline, Unigene and RefSeq. A functional annotation is also present in the annotation file of each species, based on the databases of Gene Ontology [46] and PANTHER [77].

A second major difference between the chips from Applied Biosystems and that of competing technologies is the way of labeling and detection. Indeed, Applied chips are the only ones using chemiluminescence by cleavage of a derivative of $1,2-Dioxetane$ by alkaline phosphatase as a method of detecting nucleic acids hybridized to the probes. This system amplifies the signal from each molecule hybridized achieves high sensitivity by reducing considerably the number of

boundary molecules detected.

Applied technology has a more sensitivity due to its longer oligo-nucleotide and its different labeling [1]. Signals are thus more reliable with these chips, for this reason we decided to use it in the group.

A.3 Analysis of microarray data with Ace.map

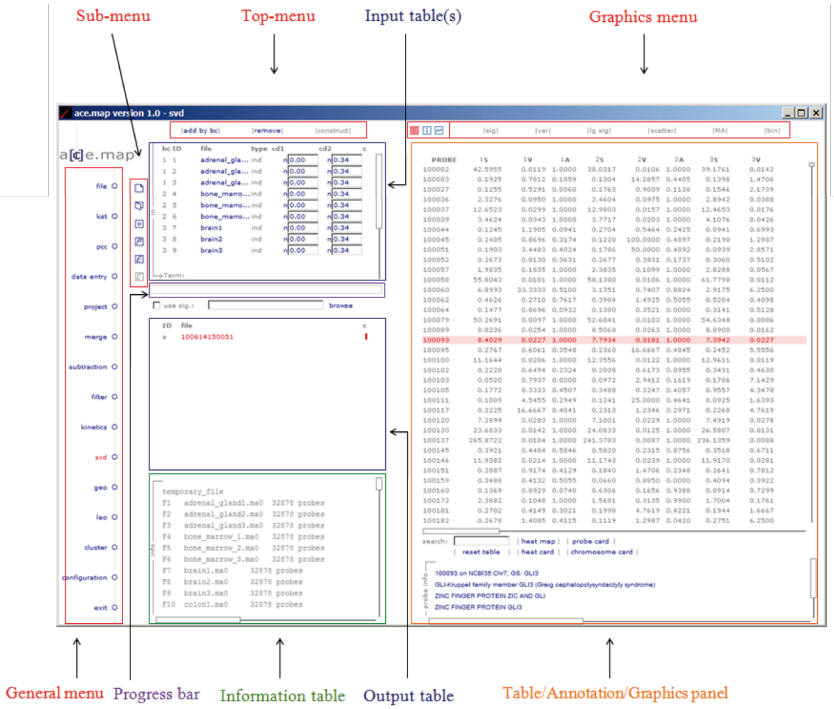


FIGURE A.3 – *Overview of Ace.map different panels, this figure is extracted from Ace.map user manual.*

When performing a microarray experiment there is two main things that we search for : Remarkable genes and links between our different experiments. Those two goals lead to two different approaches for analyzing data :

- The first one can be called "local study". One searches for biologically relevant genes, that is to say highly expressed in all different experiments, or differentially expressed between experiments, or maybe it evolves in a typical way during time-course experiments. This probe by probe study helps to constitute a set of interesting genes, on which the following analysis will be focused on.
- Sometimes to get insight into the data, one has to look at them on a global scale, searching for general organization and patterns. This second type of analysis may simply be called "global study".

Global studies are usually performed before local studies as they allow to get a global view of data, and give indication on interesting set of probes for local studies. But sometimes depending of the analysis one has to perform it might be reversed. Both types of study are complementary, thus for an accurate data analysis one needs a software which allow to switch from one type to the other easily. In the laboratory we developed a software called Ace.map in this sense. It allows

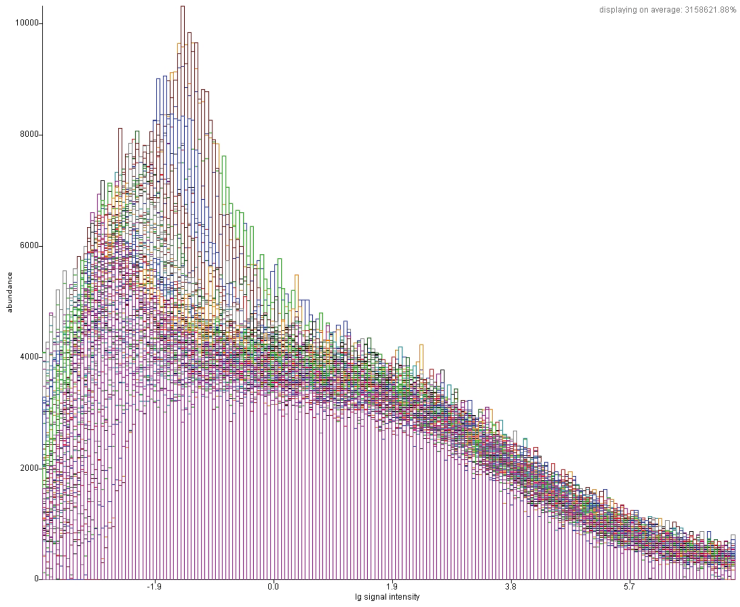


FIGURE A.4 – *Overview of microarray data. (a) Statistical distribution in logarithmic scale of a set of microarrays, each colored differently. (b) Microarray data table provided in a ma0 file type.*

for each dataset to perform different kind of study (see left panel in figure A.3), some which can be considered as local type of study (substraction, filter) other which are global type of study (Cluster, SVD, geo, etc.). I will described in the following the most important modules of this software.

Ace.map has been developed in Java language in order to facilitate its exportation to multiple types of Operating System. It is a fairly big software, which I have helped to develop. Consequently, it has given me a good insight on the different problems one encounters during the development of a software, which is something I was not familiar with at the beginning of my thesis.

A.3.1 Normalization of data

Microarray data are first normalized during the microarray scanning. This normalization is based on light spots quality, and several other parameters specifics to each chip manufacturer. Then a data table is retrieved a given to the users. Before doing any analysis, one has always to normalize again the data in order to be sure that every instance is on the same scale. That is to say that one assumes every instance should have the same statistical distribution. There is lots of way of normalizing data : mean, median, quantile normalization, and many other. We choose to use the median normalization for our microarray data. The median of a statistical distribution is described as the numeric value separating the higher half of a sample from the lower half. Every median is set to 1 by translating the all distribution. In figure A.4-a I show the statistical distributions of a set of microarray after normalization, the x-axis is in logarithmic scale, for this reasons all medians are equal to zero (*i.e.* $log(1)$).

All our microarray data are then prepared for the analysis. They are enclosed for each microarray in an *ma0* type file. In fact, Ace.map uses file type of the form *maX* with X a number from 0 to 9 which designated the Ace.map's module from which a file is extracted. Thus a basic

Ace.map data is a *ma0* type file, whereas the results of a clustering on this file will be a *ma8* type file. Each type of analysis may uses different type of *maX*, but everyone of these files types are derived from a set of *ma0* which are the normalized microarray experiment. In a *ma0* file one can find a table shown in figure A.4-b which contains the list of microarray probes, the signal of expression measured, the variance associated to the measure and a measure of quality named A which is given by the microarray scanner. An annotation file is provided with Ace.map to make the link between each probe and its corresponding gene. All these information will be used in the different analysis.

A.3.2 Substraction

One important search when performing microarray data analysis is differentially expressed genes. To this end the "Substract" module of Ace.map has been developed. It calculates several parameters for evaluating the difference of probe expression value between different *ma0s*. For example it calculates the fold change of every probe between two instances , which is the ratio of a probe expression value in the second instances divided by the value of this probe in the first instances, a $p-value$ is given to evaluate the reliability of the fold change. In a typical analysis, one uses most of the parameters calculated in Substract module to define a set of differentially expressed genes.

A.3.3 Filter

A typical workflow for defining a set of genes specific to a biological condition is to filter parameters like fold-change or p-value extracted from substract module. To select a specific set of probes which have relevant properties, one often needs to filter datasets using cutoff values on some variables. The "Filter" module is designed to this end. It allows for every important variable of a set of *maX* to define cutoff conditions, and then export the set of probes which verify these conditions.

A.3.4 Kinetics

A vast majority of microarray experiments are time-series, for analyzing those data a specific Ace.map module has been developed, which is called Kinetics (see figure A.5-a). It classifies every probe of an *maX* file into a *kinetic* class of comportment. Each *kinetic* class corresponds to a specific type of evolution of expression during a time-series experiment. For example, the group indexed by 1 represents the probes which are normally expressed at the beginning of an experiment and become highly expressed one time step after. There are 14 types of groups, described in figure A.5-b, which have been determined using Self Organizing Map (SOM) classification on training microarray data. An heatmap provided by the module gives for a set of probes of a time-series experiment, the different comportments observed, as shown in figure A.5-c.

A.3.5 Dimensionality reduction techniques

I already proves in chapter 1 the usefulness of dimensionality reduction techniques, thus Ace.map contains two types of dimensionality reduction module, which I have developed with the help of Nicolas Tchitchek. The first module is called SVD (see figure A.6) and it performs Singular Value Decomposition on a set of *maX* in every type of basis (*covariance basis*, *correlation basis*, and *correspondence basis*). It then gives six types of singular value decomposition (*e.g.* 3 type of basis for both biological condition and probes). One can export in table the different

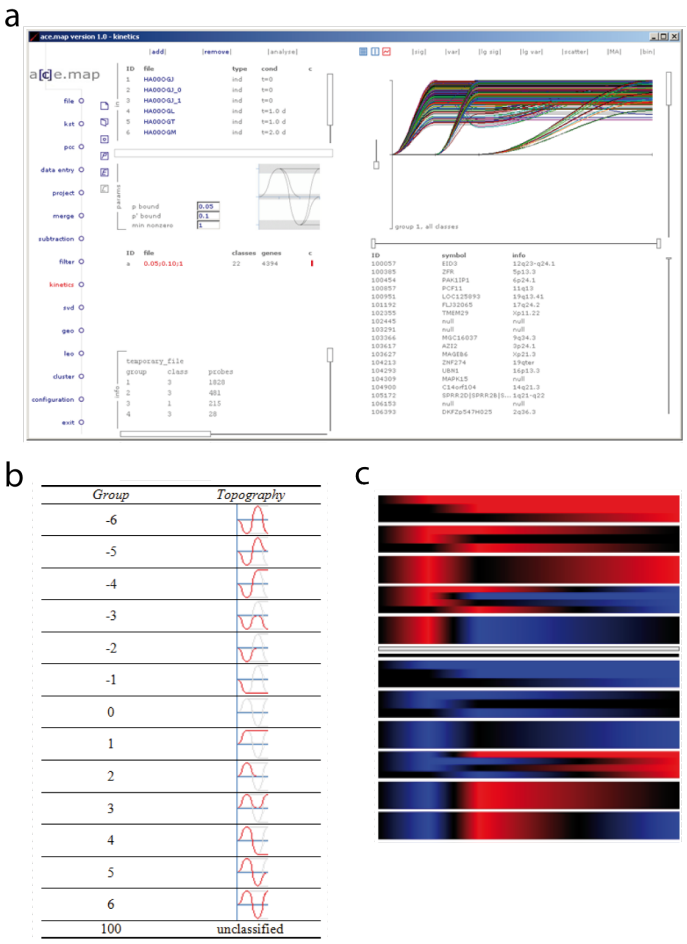


FIGURE A.5 – *Overview of Kinetics Ace.map's module. (a) The different Kinetics panels. (b) The different types of kinetics comportment which have been defined by SOM classification. For each group its index and topography is shown. (c) A Kinetics' heatmap showing the different types of evolution found in the time-series experiment. (red) parts represent highly expressed probes, (blue) is for low expression values. These figures are extracted from Ace.map user manual.*

matrices implies in the decomposition (see section 1.1.5), including of course inertia vectors. The module shows results in a biplot, two combo-box help to select the principal components to choose for the $2d$ representation. Another combo-box helps to select among the 6 types of results available, as shown in figure A.6-b.

The other dimensionality reduction module is called GEO (see figure A.6-c). It is in fact the direct implementation of the algorithm described in chapter 2. For every set of maX it reduces the dimensionality using SVD-MDS algorithm, to three and two dimensions, in order to represent biological conditions or probes accurately. A combobox allow to choose between $2d$ and $3d$ representation of data, as shown in figure A.6-d. A special algorithm has been implemented to display in the $3d$ representation the convex hull of groups of microarray. The convex hull of a set of points is the minimal convex set which contains all of them.

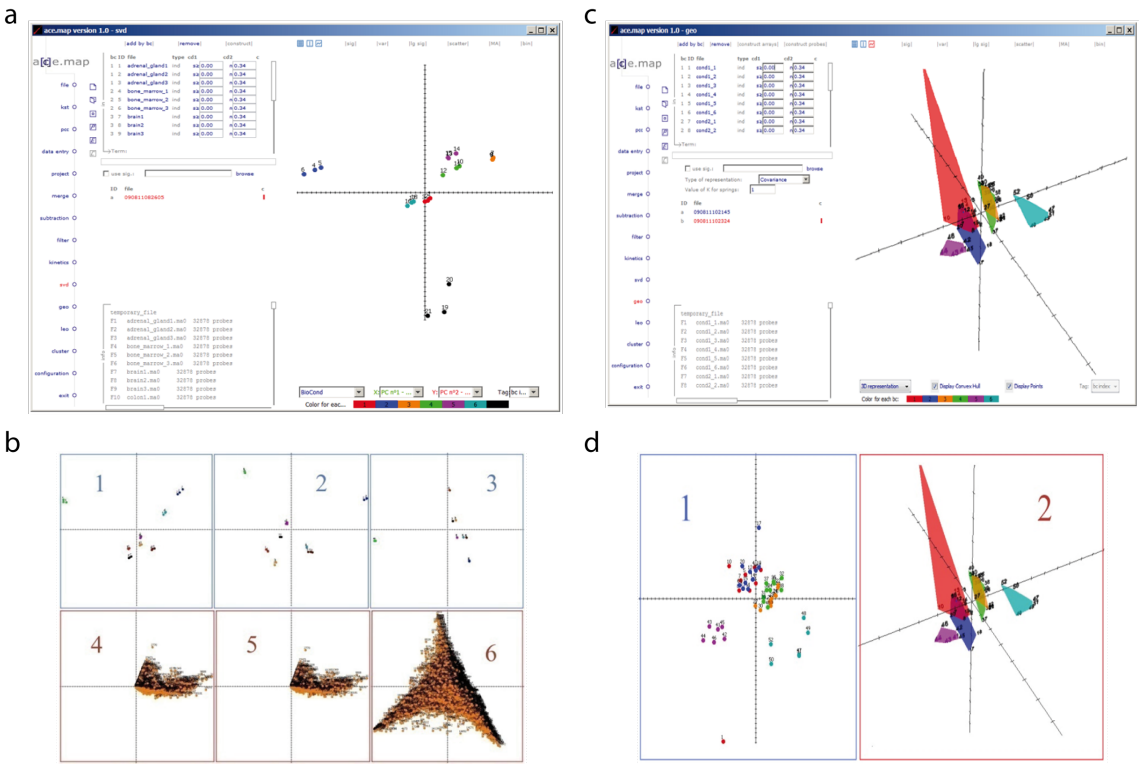


FIGURE A.6 – *Overviews of dimensionality reduction modules. (a) Print-screen of the SVD module in Ace.map. (b) The different 2d representation available in SVD module : (b-1) biological condition (bioCond) in covariance basis, (b-2) bioCond in correlation basis, (b-3) bioCond in correspondence basis, (b-4) probes in covariance basis, (b-5) probes correlation basis, and (b-6) probes in correspondence basis. (c) Print-screen of the GEO module in Ace.map. (d) Example of 2d and 3d representations of the same data obtained with GEO. All figures are extracted from Ace.map user manual.*

A.3.6 Clustering

In order to get an insight in the global structure of data, it is often useful to look at groups formed inside them. The Cluster module of Ace.map is used to this end (see figure A.7). It performs a hierarchical clustering [8] both on biological conditions and probes. One can export the results in a specific table called Cluster Map (see figure A.7-b), in which there are a heatmap showing the different expression values of the probes considered, and the clustering dendrogram of biological condition and probes.

A.3.7 LEO

The last important module has already been described in section 3.1.2. It helps to make the link between probes expression information and the different ontologies constitutive of a cell.

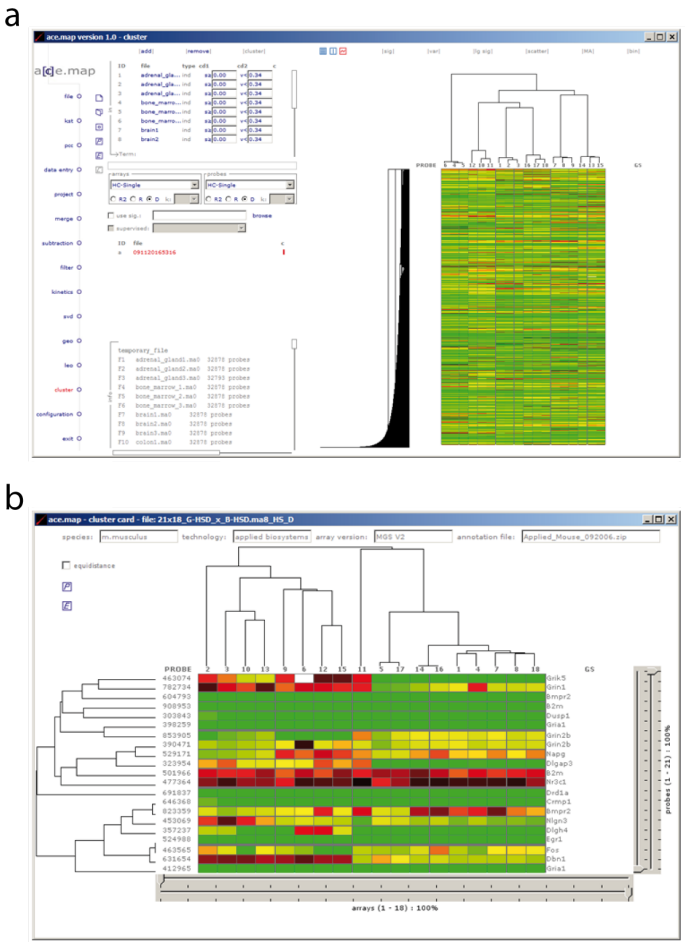


FIGURE A.7 – *Overviews of Clustering module. (a) The different panels of Clustering module. (b) The Cluster Map showing heatmap and dendrogram.*

B

Missing Value problem in microarray

B.1 Microarray and the creation of missing values

As seen in previous appendix, a microarray is a set of hybridization probes. All the processes which occur in these probes will not always performed perfectly. Sometimes, due to poor hybridization, probes' spot may be too small to be detected. Or two probes could have merged so one have one spot for two probes. Or maybe the global signal intensity in the entire microarray might be tiny so it is hard to differentiate all the spot from noise. Futhermore, one can have microarray fabrication errors and manipulation errors like contaminants on the chip including scratches, dust, and fingerprints.

For this many reasons some values of expression will not be taken account, as a result one have missing values in the microarray data matrix X (see figure B.1. In a general way we have from 0.5% to 10% of missing values [117]. From a missing values containing matrix extracting robust information from a data analysis will become more delicate [118]. Techniques of matrix imputation were invented in order to fill the blank induce by missing values in the result matrix.

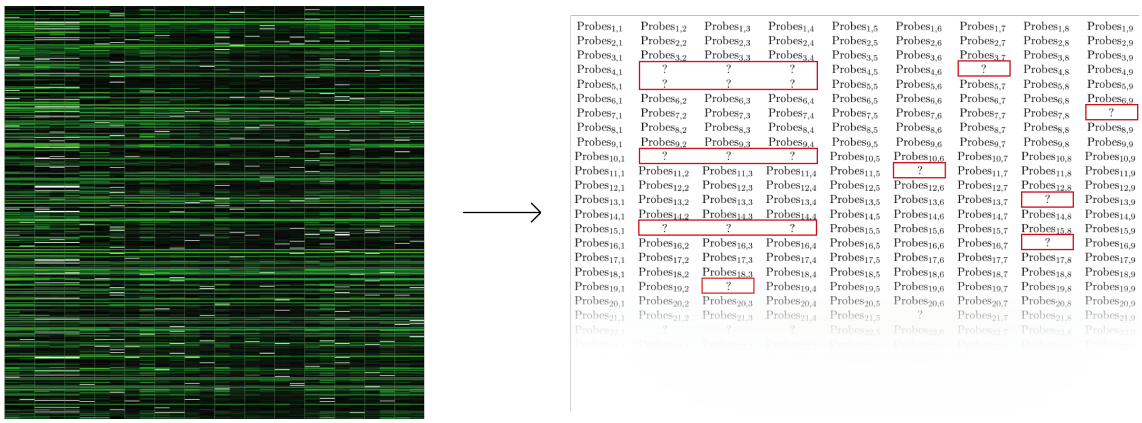


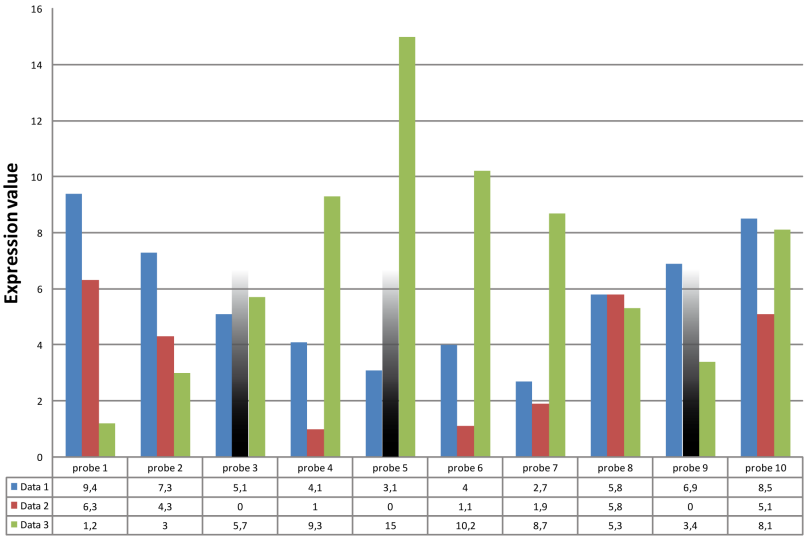
FIGURE B.1 – *(left) Hierarchical clustering and heatmap of microarray experiments, white point are missing values. (Right) Resulting matrix with missing values.*

In the following, I will review techniques of imputation that has been developed to tackle the problem of missing values. We have seen in previous appendix that it exists two types of analysis of microarray data : local and global. Techniques of missing values imputation were first designed using this distinction, the use of one being conditioned by the aim of the study. I show

secondly that nowadays techniques are developed using both local and global information. To finish I present recent techniques which use external information like gene ontology.

B.2 Local imputation

The first naive way of imputing missing values is to replace them by zero. A less trivial way is row averaging. When few expressions values of one probe is missing on a set of experiments, one may replace it by row average of the probe on all other experiments. This method is the most wide used, because of its simplicity. As replacing by zero means that corresponding probes have no signal, and inferred consequently real important biological meanings to data. Hence using row average to fill missing values holes in data matrix is the least one can do. The first study on microarray missing value estimation was performed by Troyanskaya *et al.* in 2001 [22]. They demonstrated that even though row average is an improvement upon replacing missing values with zeros and a common first step for lots of imputation techniques, this approach is not optimal because it neglects correlation within data. This has encouraged the development of more refined missing value procedures, which try to exploit instances relationships by using the information available in the whole dataset.



	probe 1	probe 2	probe 3	probe 4	probe 5	probe 6	probe 7	probe 8	probe 9	probe 10
Data 1	9,4	7,3	5,1	4,1	3,1	4	2,7	5,8	6,9	8,5
Data 2	6,3	4,3	0	1	0	1,1	1,9	5,8	0	5,1
Data 3	1,2	3	5,7	9,3	15	10,2	8,7	5,3	3,4	8,1

FIGURE B.2 – *Example of data with missing values using data from chapter 1. Probe* 3, 5 *and* 9 *of Data 2 are missing (black rectangle). Local imputation techniques will use probe which have the same profile as the uncomplete probes. Global imputation techniques will use the structure of the entire dataset.*

If one wants to take account of the structure underlying a probe containing missing values, one needs to use other probes which seem to have the same expression profile. That is the goal of the first local imputation techniques called the weighted k-nearest neighbors (kNN) method. kNN-based method selects probes with expression profiles similar to the probe of interest to impute missing values. To illustrate this method I use the three data examples from chapter 1, adding missing values to *Data 2* at probes 3, 5 and 9, and plotting their expression profile in figure B.2. If one consider probe 3 which has one missing value in *Data 2*, the kNN-based method goal is to find k other probes present in *Data 2* which seem to have the same "comportment" has

probe 3. To evaluate this last parameter one calculates euclidean distance or correlation between all probes using values in other instances (*Data 1* and *Data 3* in our example). If one search the nearest probe neighbor of probe 3 in our example, probe 8 will be a prefect candidate, probe 2 and probe 10 should be also selected in a 3-nearest neighbors search. A weighted average of values in *Data 2* from the k closest probes is then used as an estimate for the missing value. Here, the new value for probe 3 in *Data 2*, should be a weighted average (see equation B.2) of the value of probes 8, 2 and 10 in the case of a 3-nearest neighbors imputation. In the weighted average, the contribution of each probe is weighted by similarity of its expression to the considered probe. New value y_{ij} for the probe $\underline{\mathrm{X}}_j$ in the instances $\bar{X}_i$ is then be given by

$$y_{ij} = \sum_{j' \in \{k-nearest\ probes\}} \frac{x_{i,j'}}{d(\underline{\mathrm{X}}_j, \underline{\mathrm{X}}_{j'})} \tag{B.1}$$

where $x_{i,j'}$ is the expression value of probe j' (one of the nearest probe) in i^{th} instance, and $d(\underline{\mathrm{X}}_j, \underline{\mathrm{X}}_{j'})$ is the distance between j^{th} probe and j'^{th} probe.

The Weighted Nearest Neighbors method (WeNNI) [119] is really close to kNN but use information on the spot quality to calculate the weight average. Every analysis of microarray data is preceded by a filtering step, in which each spot is required to fulfill certain quality control criteria. If the spot fails to meet the quality requirements it is marked as a missing value. This is equivalent to accompanying each expression value with a binary weight, and enforces an abrupt cutoff in quality control criteria. A new parameter β is used, the more stricter spot quality control criterion are used the higher β is. Imputed values will be given by

$$y_{ij} = \sum_{j' \in \{k-nearest\ probes\}} W_{i,j'} \frac{x_{i,j'}}{d(\underline{\mathrm{X}}_j, \underline{\mathrm{X}}_{j'})} \tag{B.2}$$

with $W_{i,j'}$ a weight given by spot quality of probe j' in instance i in which there is a missing value.

Instead of selecting k-nearest neighbors one can used least square techniques, the missing value will be compute using only linear regression on similar probes profiles, this method is called : Ordinary Least Squares (OLS or LSimpute) [120,121]. Whereas, the two first techniques impute only expression in one instance at a time, using regression one will find a new complete expression profile for the probe where there is a missing value. A regression is the process of finding values that best fit an ensemble of statistics value. For example the typical linear regression is performed when you have an independent statistical variable $\bar{X} = (x_1, x_2, \ldots, x_n)$ and a dependant $\bar{Y} = (y_1, y_2, \ldots, y_n)$ variable. One wants to find the best linear function that fit data $\bar{Y} = a + b\bar{X} + \varepsilon\bar{Y}$, where a and b as to be estimated and ε is an error parameter. The estimation of $\bar{Y}$ is then given by

$$\hat{\bar{Y}} = \overline{\bar{Y}} + \frac{s_{xy}}{s_{yy}}(\bar{X} - \overline{\bar{X}}) \tag{B.3}$$

with s_{xy} being the correlation between $\bar{X}$ and $\bar{Y}$, s_{yy} being the standard deviation of $\bar{Y}$, $\overline{\bar{X}}$ and $\overline{\bar{Y}}$ are the mean of each statistical variable.

For multi-linear regression the goal is to fine the estimates of $(y_1, y_2, \ldots, y_k)$ given $(x_1, x_2, \ldots, x_k)$ with the assumption that :

$$y_i = a_i + b_{i1}x_1 + b_{i2}x_2 + \cdots + b_{ik}x_k \tag{B.4}$$

Since, multi-linear regression for probes correlations are not feasible for more than a few probe, one uses a weighted average of several single regression estimates of the same missing value. Given

a missing value in the data matrix for probe $\underline{X}_j$, only the k probes $\underline{X}_1, ..., \underline{Y}_k$ most correlated with $\underline{X}_j$ are included in the prediction model. In addition, none of $\underline{X}_1, ..., \underline{X}_k$ is allowed to have a missing value in the same experiment as the missing value to be estimated. When determining which are the most correlated probes, one uses the absolute correlation values since both positive and negative correlation between probe is equally well suited for regression. The correlation between probes $\underline{X}_j$ and $\underline{X}_{j'}$ is determined by only including experiment where both probes have non-missing values in the computation. Given the k closest correlated probes, k estimates $\underline{Y}_1, ..., \underline{Y}_k$ of the missing value are computed by single regression from each of $\underline{X}_1, ..., \underline{X}_k$. Choosing the value of k is one important question and it really depend on the dataset, but [121] find that $k = 10$ is a good value to choose. Finally, a weighted average of the estimates is computed. The weighting is designed to give the probes most correlated with $\underline{X}_j$ the largest weights, as these are expected to give the best estimates of the missing value. Given the estimated correlation $corr(\underline{X}_j, \underline{X}_{j'})$ between probes $\underline{X}_j$ and $\underline{X}'_j$, the weight w_j [121] assigned to the estimate $\underline{X}_j$ is

$$w_j = \left(\frac{corr^2(\underline{X}_j, \underline{X}_{j'})}{1 - corr^2(\underline{X}_j, \underline{X}_{j'}) + \varepsilon} \right)^2 \tag{B.5}$$

where $\varepsilon = 10^{-6}$. In this formula, the numerator approaches 1 with increasing absolute correlation, while the denominator approaches ε. Thus strong correlations will give large weights, and weak correlations will give small weights.

At the end of this process one has a new expression profile for the probe which is a weighted average of the entire set of estimates. The big difference between OLS and kNN is that the first one changes all the expression values for the considered probe, whereas kNN just change the value in the missing value part.

The Least Squares Adaptive (LSA) procedure of [121] combines probe-based and experiment-based imputation estimates, using an adaptive procedure to determine the weight of the two estimates. The probe-based estimates are determined as in OLS, and the experiment-based estimates are determined by multi-linear regressions based on the experiments, where the probe-based estimates are not used leaving missing values in the expression matrix. To determine the best weighting of the two estimates, known values in the data matrix are initially re-estimated, and the errors of the probe-based and experiment-based estimates are determined. The optimal weight is determined by minimizing the sum of the squared errors for the re-estimated. The weights are determined adaptively by considering the strength of the gene correlation in the gene-based estimates. That is only probes with similar values of the maximum gene absolute correlation used in the probe-based estimation which are used into the weight calculation.

The Local Least Square (LLS) procedure of [122] selects neighbors based on the Pearson correlation as in OLS, but instead of weighting univariate regressions they perform multiple regressions using all k-nearest neighbors. The missing values are imputed based on the least squares estimates, determined using the pseudo-inverse of the k-nearest neighbors expression matrix. If the percentage of missing values is relatively small, then neighbor probes with missing values are excluded from the least squares system, otherwise they will be initially estimate by the row average.

B.3 Global imputation

In all the previously described techniques, the methods involved usually the closest probes to impute missing values, and do not consider the global structure of the data. We will now describe

techniques that take this into account. They all try to find principal components or groups in the data, and infer missing values with the assumption that the global structure of the data has to be invariant through this imputation.

The first technique developed in this way is Partial Least Square (PLS). Regression selects linear combinations of probes (called components) exhibiting high covariance with the probe having missing values (the target probe). The first linear combination has the highest covariance with the target probe, and subsequent components have the greatest covariance with the target probe in a direction orthogonal to the previously selected components until a total number of c components are selected. The missing values are then imputed by regressing the target probe onto the PLS components. Missing values are first imputed by row average prior to PLS imputation [120]. This technique has one drawback because it consider principal components in term of high covariance with the considered probe. We have to leave this dependency, if we really want to take into account the global properties.

Principal Component Analysis (PCA) is a response to this problem [22,123], for performing it one uses Singular Value Decomposition (SVD) (see chapter 1. Singular value decomposition gives a set of mutually orthonormal components that can be linearly combined to approximate the expression of all probes in the data set. In term of matrix theory, one searches the decomposition of a matrix X

$$X = USV^t \tag{B.6}$$

with U and V two orthogonal matrices, and S a positive diagonal matrix.

Using SVD on microarray data will help us to find the most relevant components, the idea is to use them to retrieve missing values. which are the most significant. One selects k most important principal components, and estimates a missing value x_{ij} found on instance i of j^{th} probe by first regressing this probe against the k principal components and then use the coefficients of the regression to reconstruct x_{ij} from a linear combination of these k principal components. The i^{th} value of probe j and the i^{th} values of the k principal components are not used in determining these regression coefficients.

It should be noted that SVD can only be performed on complete matrices, therefore we originally substitute row average for all missing values in matrix X. We then utilize an expectation maximization method to arrive at the final estimate. Each missing value in X is estimated using the above algorithm, and then the procedure is repeated on the newly obtained matrix, until the total change in the matrix falls below the empirically determined threshold of 0.01. The optimization process is based on the first principal components, thus components with highest value of inertia. These high inertia components will not change during the optimization, the changes will appear in low inertia components. One can proves that this optimization tend to a minimum, so one will find an equilibrium [123].

SVD is the most used techniques for global imputation, because it was the first [22] taking account of the global scheme of data. But other techniques have now been developed. Some use Bayesian estimation to fit a probalistic PCA model (BPCA) [124]. A variational Bayes algorithm is used to iteratively estimate the posterior distribution of the model parameters and the missing values until convergence is reached. The key feature of this approach is that principal axes with small signal-to-noise ratios are shrunk toward zero, so that the algorithm automatically screens for those axes that are the most relevant, whereas with SVD where one needs to select a number of important component. As for SVD, missing values still have to be initially imputed by row or probe-wise average.

Other proposed techniques use cluster information to find the estimates [125]. Their method is

based on Gaussian Mixture Clustering (GMC) and model averaging. Microarray data are assumed to be generated by a Gaussian mixture of some number of components. The amplitudes of the estimates are computed. For a missing entry, the estimate is computed using a linear combination of the component-wise estimates, weighted by the probabilities that the probe belongs to the components.

Another totally different method for global imputation is based on Support Vector Machine (SVM), which is a powerful tool for general purpose machine learning problem [8,126]. This tool is usually seen as a classification technique, its purpose is to use a kernel function to embed our data matrix in a higher dimensional space, in which non-linear correlation will be transformed into linear correlation. Thus for classifying data in this new space, one will have to define hyperplan separator (linear separator) called support vector. SVM solves the "over-fitting" problem by using structure risk minimization principle, which minimizes both empirical risk and confidence interval. In practice, two kinds of SVMs are provided for different purpose : Support Vector machine for classification (SVC) and Support Vector Machine for regression (SVR) [127]. But for microarray data we will use just SVR. In few words, instead of making a regression with few selected probes or few principal components, one finds our estimates thank to a trained support vector machine network.

B.4 Comparative study of their efficiency

One can find only two studies which compare the different techniques of imputation presented above. First one was done by Troyanskaya et al. in 2001 [22]. They compared row average, kNN and SVD, using input dataset, deleting values in those data, and running the different techniques on them to see if the data are well retrieved. Their parameters for method quality assessment is called normalized root mean squared error :

$$LRMSE = \sqrt{\frac{\sum_{(i,j):x_{ij}} (y_{ij} - \hat{y}_{ij})^2}{Number\ of\ x_{ij}}} \tag{B.7}$$

where y_{ij} are the missing values, x_{ij} are the probe expression matrix value in input dataset and $\hat{x}_{ij}$ are the ones in impute dataset (input dataset with missing values retrieved). They found that row-average was never a good solution, and kNN was in general better than SVD in retrieving information.

The second was recently performed by Brocks et al. in 2008 [128]. Between these two studies, lots of new techniques have been developed as we just showed. And every research group that describes a new technique has claimed the importance and novelty of it by a comparison with a few earlier techniques, using LRMSE parameters and deletion of data in complete data-set, and claiming that their one was better.

Brocks et al. [128] have done a pretty exhaustive comparison. They have compared local imputation (kNN,OLS,LSA,LLS) and global imputation (PLS,SVD,BPCA) techniques using LRMSE and also a new parameters derived from LRMSE. They found that in general LSA, LLS and BPCA were better than the other techniques. This conclusion is reasonable because these are the last developed techniques for local (LSA, LLS) and global (BPCA) imputation. Those three methods compete for the title of best imputation method, although they perform differently on different kind of data structure.

Brocks et al. at the end of their paper claim that, as each technique has its own preferential data structure, one should use all three techniques at the same time, and employ only the better estimates. Brocks et al. use two different parameters to determine what a "best estimate" is. The

first one is based on Shannon entropy evaluation when imputing data, the second one uses self-training selection to determine, by deleting 5% of the data, what the best imputation techniques is in light of data structure. Combining these two parameters, one can select the best estimate matrix.

This idea of correlating local and global information was also developed by other research group. Some even imagine using external information to increase estimation. I describe those techniques in the following.

B.5 Imputation using a combination of local, global and external information

The first technique that used multiple techniques correlation was call LinCmb [129], and was developed in 2005. It employs row average, kNN, SVD, BPCA and GMC to define a set of estimation parameters. As this techniques use local and global information they perform better in retrieving information than each one of its components.

An other technique is called Projection Onto Convex Sets (POCS) [130] and was developed in 2006, using the same local-global imputation, but adds external biologic information. The main idea is to formulate every piece of prior knowledge into a corresponding convex set and then use a convergence-guaranteed iterative procedure to obtain a solution in the intersection of all these sets. One designs several convex sets, taking into consideration the biological characteristic of the data : first set mainly exploit the local structure among probe in microarray data, while the second set captures the global structure among arrays. The third set (actually a series of sets) exploits the biological phenomenon of synchronization loss in microarray experiments. In cyclic systems, synchronization loss is a common phenomenon and one constructs a series of sets based on this phenomenon for POCS imputation algorithm.

There is three other techniques which have been described and which use external information derived from biologic knowledge data base. Gene ontology (GO) [131], is a method in which semantic dissimilarity is employed as an external information on the functional similarity of two probes. The calculation of semantic dissimilarity starts by building an ontology tree created from the ontology downloaded from the GO website [46]. An annotation table from the annotation file (corpus) is also created and used to fetch all GO accession ids that are associated with a given probe. Based on these data structures the information content p-value for each node in the tree is calculated. This p-value will serve as a weight for average calculation of the estimates.

Histone acetylation [132] as also been used as external information. With the help of gene regulatory mechanism, an imputation method called Histone Acetylation Information Aided Imputation method (HAIimpute) has been created. It incorporates the histone acetylation information into kNN and LLS imputation algorithms for final estimation of the missing values.

All those techniques correlating local and global imputation or using external information, as revealed to be better than the imputation method they use. But that is an obvious thing to say, if you add information for estimation with no doubt you will find a better result, unless, this very information is not reliable.

B.6 Conclusion

We have seen in this appendix a non exhaustive list of techniques for imputing missing values (see table B.1 for a summary of all techniques presented). That imputation is not a minor problem. One has to take account missing values in order to make reliable inferences about

Imputation techniques	**Principle**	**Types**
Row average [22]	Replace missing values by row average.	Local
Weighted K-nearest Neighbors (KNN) [22]	Use k nearest probes (in term of profile) to calculate a weighted average to replace the missing value.	Local
Weighted nearest neighbors method (WeNNI) [119]	Use spot quality information to define a weighted average to replace missing value.	Local
Ordinary Least squares [120, 121]	Use k-nearest probes to estimate a new probability profile of the probe containing missing values.	Local
Least Squares Adaptive (LSA) [121]	Combine OLS on probes and Least square on instances to estimate probes containing missing values.	Local
Local Least Square (LLS) procedure of [122]	Multiple regression instead of one regression in OLS.	Local
Partial Least Square (PLS) [120]	Find orthogonal set of components which maximize covariance with probes containing missing values.	Global
Singular Value Decomposition (SVD) [24, 123]	Find principal components and estimates missing value thanks to global structure conservation.	Global
BPCA [133]	With Bayesian method principal components appear naturally and help to find missing values.	Global
GMC [125]	Retrieve values using natural cluster in our data.	Global
Support Vector Regression (SVR) [127]	Support vector machine regression estimates values by a learning process.	Global
LinCmb [129]	Combine row average, KNN, SVD, BPCA and GMC to find best estimates.	Local Global
POCS [130]	Combine Local, global and biological external knowledge to find good estimates	Local Global External
GO gene ontology [131]	Use gene ontology data-base to find estimates	External
HAImpute [132]	Use hystone acetylation to find estimates	External

TABLE B.1 – Table of all the techniques reviewed in this article.

biological significant from data. We see that replacing by zero our row average seems to be one solution but there exist lots of other solutions, relatively easy to implements, which always give better results. A choice has to be made on whether you want to enforce local properties, or global ones. One can go further, and combine local, global, and external information to retrieve information better. But these techniques takes a lot of computational time, and are not user friendly because of their complexity.

These general approaches of using external information will almost certainly develop further in the future, not only for microarray imputation, but also for all problems in systems biology, where all levels of biological organization are related. One has to use different information, from different types of experiences to better understand what the underlying phenomenon is.

C

Review article : Dimensionality reduction of "omics" data.

Christophe Bécavin, Arndt Benecke.
To appear in Expert Review of Molecular Diagnostics, January 2010.

Abstract

Omics data increase very rapidly in quantity and resolution and are more and more recognized as very valuable experimental observations in the systematic study of biological phenomena. The increase in availability, complexity, and non-expert interest in such data requires the urgent development of accurate and efficient dimensionality reduction and visualization techniques. To illustrate this need for new approaches we extensively discuss current methodology in terms of the limitations encountered. We then illustrate at a recent example how combinations of existing techniques can be used to overcome some of the present limitations, and discuss possible future directions for research in this important field of study.

Dimensionality reduction of "omics" data.

Christophe Bécavin [1,2], Arndt Benecke [1,2]

(1)Institut des Hautes Études Scientifiques - 35 route de Chartres - 91440 Bures sur Yvette - France
(2)Institut de Recherche Interdisciplinaire CNRS USR3078 Univ. Lille I, II - 50 avenue de Halley - 59658 Villeneuve d'Ascq - France arndt@ihes.fr

Summary

Omics data increase very rapidly in quantity and resolution and are more and more recognized as very valuable experimental observations in the systematic study of biological phenomena. The increase in availability, complexity, and non-expert interest in such data requires the urgent development of accurate and efficient dimensionality reduction and visualization techniques. To illustrate this need for new approaches we extensively discuss current methodology in terms of the limitations encountered. We then illustrate at a recent example how combinations of existing techniques can be used to overcome some of the present limitations, and discuss possible future directions for research in this important field of study.

Keywords

Multivariate Analysis, Dimensionality Reduction, Singular Value Decomposition, Multidimensional Scaling, Molecular Dynamics, Machine Learning, Genomics, Transciptomics, Proteomics

Expert commentary

Since over a decade now biologists have increasing access to a vast variety of "omics" experiments, each of them consisting in a parallel measurement of a very large number of variables. Development of data analysis techniques which help to obtain clear insights on the underlying correlations within the data is essential. Preliminary to the choice between different existing techniques one needs to first become aware of the particular properties of high-dimensional data where the number of measured variables is typically much higher than the number of measurements realized (i.e. instances). As was highlighted in a review by Clarke et al. [1], having a high number of data-points will influence the statistical distribution of the dimensionality reduction leading to the so-called 'curse of dimensionality', where also the meaningful differences between data-points are compressed. The structure of these spaces will be often complex, as the number of correlation within variables grows with the number of it. In any studies, a tiny set of variables will be selected to perform a local analysis, so this problem of high dimensionality will be irrelevant, but in global analysis where most of the variables are included one will need techniques of dimensionality reduction or at least adapted to high dimensional space to obtain more accurate biological insights. We will review here different types of techniques of dimensionality reduction which have been developed and used for the study of high dimensional "omics" data and discuss their advantages and shortcomings as well as hint at new developments in the field of study.

Multivariate analysis of data by reducing dimensionality can be summarized as the search of a map F from a high dimensional space X with $dim(X) = m$ to a new space Y with $dim(Y) = p$, with as a first constraint $p < m$ (ideally $p << m$), and as a second constraint the minimization of one or more parameters which will evaluate the changing properties of the space X.

$$F: \mathrm{R}^m \mapsto \mathrm{R}^p$$
$$X \to Y$$

Sometimes the mapping will be explicit as in Principal Component Analysis, and sometimes it will be implicit as in Multidimensional Scaling. One can distinguish two types of mappings: linear and non-linear ones, the difference between them is that in the first case on supposes a linear association

between variables whereas in the second one supposes non-linear relationships.

1 Linear methods

Linear methods of dimensionality reduction are all part of the generic field of Factor Search. One seeks to summarize data in a convenient way using some factors which are linear combinations of the variables of the data. These factors will regroup the information contained within the variables, and will be chosen in order to maximize their variances. The number of factors (dimensions) can be decided *a priori* (supervised analysis) or will be data-driven (unsupervised). We will only focus here on the second case, as the first is only relevant in some special cases where one already knows how many dimensions are best needed to maximize insight into the data, which is rarely the case in biological settings.

The best known technique of data-driven factor search is called Principal Component Analysis (PCA); basics of its having been described a century ago by Pearson [2] and since then PCA has been further developed and improved on several occasions. Different methods for linear dimensionality reduction have also been created based on similar ideas. At the end they have all proved to be closely linked by the mathematical tool of Singular Value Decomposition (SVD):

It is known [3] that every rectangular matrix can be decomposed using its singular values:

$$X = USV^t \qquad (2)$$

where U (matrix containing left singular vectors) and V (matrix containing right singular vectors) are both square orthogonal matrices ($UU^t = U^tU = Id$ and $VV^t = V^tV = Id$), and S is a rectangular matrix containing the singular values (s_i) which are positive. If n is the number of rows and p the number of columns of X, X is $n.\times p$, U is $n.\times n$, S is $n\times p$, V is $p\times p$. In case $n = p$, S is a square diagonal matrix, if $n \neq p$, $S_{ii} = s_i$ and $S_{ij} = 0$. U is $n.n$. This decomposition is very useful in dimensionality reduction because of its link with the eigenvalue decomposition of the inner- (XX^t) and outer-products (X^tX) of X. If $X = USV^t$ one obtains:

$$XX^t = USV^t(VS^tU^t) = USS^tU^t$$

$$X^tX = (VS^tU^t)USV^t = VS^tSV^t$$

SS^t and S^tS are two diagonal square matrices, they do not have the same size but they have the same number of non-null values on the diagonal. So SVD allows to demonstrate that the inner and outer product have the same eigenvalues λ_i, with $\lambda_i = s_i^2$. It also gives a matrix link between them, and it is very useful as all the classical linear techniques of dimensionality reduction are performed using one of these products.

Singular Value Decomposition provides three major types of information:

1. A new data matrix X_{new}, which represents the data points in a new orthogonal basis with a minimum number of components, and where distances between instances are preserved.

$$X_{new} = XV$$

2. Inertia parameters indicate the standard deviation and relative contribution of the cloud of points on each principal component. Each principal component will then be classified by its standard deviation value.

$$c_i = s_i \Big/ \sum_i s_i$$

3. The matrix V in which the linear contribution to each principal component of all the components in the former basis is given.

Alter et al. for instance have made extensive use of SVD for the study of transcriptome data, helping them to find a specific cosinus and sinus oscillator comportment in first and second component of their analysis of cell-cycle data [4-6].

The principle of PCA is to use the covariance matrix, find its eigenvalues, and then determine the best orthogonal space of lower dimension for embedding the data-points. As the covariance matrix is linked to the outer-product matrix, PCA can be performed using SVD [7]. It exists many example of PCA utilization in the literature as it is still the best known technique of dimensionality reduction [8-10]. All those studies are based on the same principle: Using the information on covariance between variables or between groups of variables, one tries to find the first principal components which will well

summarize, in term of standard deviation variation, the general characteristics of the system.

PCA is determined using covariance information, but sometimes the information on correlations between variables would be more useful for biologic interpretation. Correlation between two variables is equal to their covariance divided by the standard deviation of both, so the correlation suppresses the standard deviation of all variables, which is based on the hypothesis that they should all have the same variation and differences found in the data are only due to errors during measurement. In order to perform PCA using correlations (called Principal Component Correlation Analysis, PCCA) one only needs to rescale the data by dividing by the standard deviation of each variable before the analysis. One can find applications of Principal Correlation Analysis for instance in [11, 12].

Another alternative is to use the information of distance between instances rather than covariance between variables. Classical Scaling, a linear version of Multidimensional Scaling (MDS) described below, was designed for this purpose. Using the technique of double centering one can define a link between the Euclidean distance matrix and the inner product matrix and then by using SVD define the best minimal orthogonal space to embed the data-points [13, 14]. The link given by SVD between the inner- and outer-product matrices assures that the results of Classical Scaling will be identical to the ones obtained by PCA [13].

Another well known linear technique also linked to SVD is called Correspondence analysis. This technique is specifically adapted to the study of contingency tables, which are tables of frequency data, making it particularly interesting for "omics" data which are often counts of the presence of particulars types of objects / entities. In this kind of study one assumes that every instance should have the same number of cumulated "counts". In the case of for instance comparative genome hybridization (CGH) microarrays will be equivalent to assuming the same number of genomic fragments for every locus. Taking into account this hypothesis one defines a proper definition of distance between instances, equal to the χ^2 distance between them. It is easy to demonstrate [15] that with a proper rescaling of the data these distances will be equal to Euclidean distances in a rescaled space, and then using SVD one is able to perform Classical Scaling and find the proper orthogonal space. Despite its interest for the analysis of "omics" data, Correspondence Analysis is in this context not frequently utilized [16].

2 Non-linear methods

Linear dimensionality reduction techniques give perfect (that is without loss of information) results in a reduction to p dimensions (with $p = rank(X) = rank(S) \leq min(n-1, p)$). Only then the geometric structure of the representation is preserved when compared to the one given by the full number of dimensions. Often one wants to go further in the dimensionality reduction, for example down to two dimensions for a proper visual representation. The only choice available with linear methods consists in deleting all unused components and thereby inducing severe deformations in the representation. This current procedure in the cases of the linear dimensionality reduction techniques discussed above.

To overcome this problem one can use non-linear techniques that will search for non-linear correlations between data and then eventually lead to a lower number of principal non-linear components. Alternatively, one will first impose the number of dimension to which the data-objects shall be reduced and then one tries to find the best representation in this space minimizing a chosen geometric parameter.

The oldest non-linear dimensionality reduction technique is called Multidimensional Scaling (MDS) [13, 14]. It is based on the principle of retrieving a proper configuration of points based only on information of similarity or dissimilarity (which we shall refer to with the general term of "distance") between the data-points. We have already discussed above the linear derivative of MDS, Classical Scaling. MDS stems from the observation that the most valuable information for getting an insight in the general structure of the data is not the very value of variables but the information of similarity or dissimilarity between them. A dimensionality reduction induces a deformation of this distance information. In consequence, MDS is an optimization process trying to minimize a parameter evaluating the loss of distance information. The most used version of MDS is based on Euclidean distance information between data-points and the parameter to be minimize is called the Kruskal stress [17]

$$e = \sqrt{\frac{\sum_i \sum_j (\delta(i,j) - d(i,j))^2}{\sum_i \sum_j d(i,j)^2}} \quad (7)$$

with $\delta(i,j)$ Euclidean distance between point i and j in the input basis, and $d(i,j)$ Euclidean distance in the dimension reduced basis.

Major work on MDS is related to the development of proper algorithms. As it is an optimization process MDS comes with all the inherent drawback of optimization processes. The two major ones being: (i) its sensitivity to initialization, and (ii) the problem of local minima in the search of the global minimum. To overcome these problems the improvements that have been created concern essentially the algorithmic implementation of the method. Dzwinel et al. showed that using a model of Molecular Dynamics may improve the efficacy of MDS [18]. Their demonstration is based on the finding that every optimization process can be summarized using a virtual particle paradigm [19]. Another improvement can be found also in [20], where Andreas et al. proposed to run an interactive algorithm during which the user will be able to move the configuration away from a local minimum by hand. Finally, Andrecut proposed to add stochasticity [21] in the molecular dynamics process using the idea of decreasing temperature as in Simulated Annealing processes.

The first major application of MDS to "omics" data can be found in the works of Gray et al. [22] in which they demonstrate using MDS how one can trace back metastasis progression in different tissues. At that point they used binary data revealing for each type of tissue the presence or absence of metastases. In an article from Taguchi et al. [23] one can find the application of MDS to a large number of data-points. Their goal was to use MDS on cell-cycle gene expression data for obtaining a visual representation of the oscillatory phenomenon. They thereby prove the significant advantage of MDS compared to linear techniques such as PCA as the MDS process due to its optimization procedure assures a more accurate representation of the data. Other applications of MDS to "omics" data can be found in [24-26].

Multidimensional Scaling is not the only non-linear method which exists. A family of techniques, collectively referred to as "Manifold Learning" techniques has emerged more recently. The actual experimental data are considered in this context as being fragmented and error-prone higher dimensional representations of the random variables defining the manifold, however, can be used to retrieve this manifold.

that data are extracted from a set of random variables which define a manifold (i.e. geometric objects such as curve or surface). Data are looked at being just bad (fragmented, error-prone) representations of these random variables in a bigger dimensional space, however, can be used to retrieve the manifold.

This idea corresponds to the non-linear extension of linear regression where we suppose that data are a bad representation of several random variables which are linearly correlated. As one thus postulates non-linear correlation between random variables it becomes clear why the first techniques of Manifold Learning ever invented were called Principal Curves analysis [27] and Non-linear PCA [28]. Starting a decade ago such techniques are focus of intense research, resulting to the invention of Isomap [29] and Locally Linear Embedding (LLE) [30]. Isomap aims to determine the shape of the manifold by defining geodesic distances between points which are distances based on the graph formed by the configuration of points, and then use these distances with an algorithm resembling MDS to determine a representation of the data-points. LLE locally "unrolls" the cloud of points during dimensionality reduction.

Generally speaking, all techniques of Manifold Learning will define first a geometric property to conserve during the process, and then try to minimize the loss of this property during the dimensionality reduction process. A comprehensive review of these techniques can be found in Gorban et al. [31], with also some applications to microarray data, showing that these techniques can conserve better certain geometric properties such as the topology, or global and local organization than previous methods.

A major drawback of these techniques is their sensitivity to "complex" highly correlated data, making them hard to use for really high dimensional data such as transcriptome or proteome studies. A successful application of Manifold Learning has been shown in [32]. Manifold Learning techniques are also hard to implement and compute because they in fact belong to hard problems of Machine Learning [33], requiring very good (quantity and quality) training

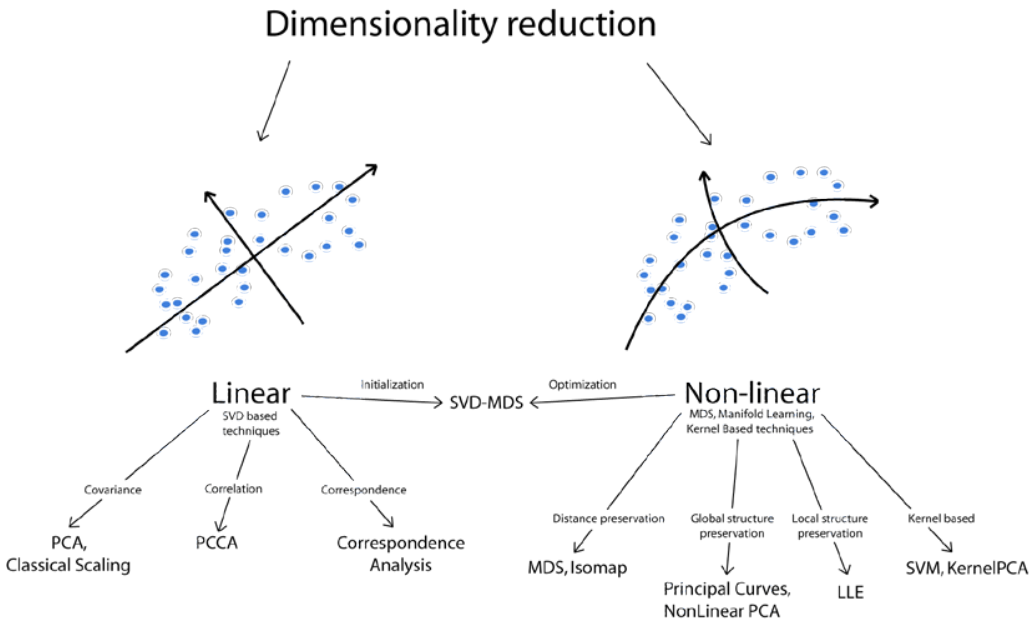


Figure 1: Scheme summarizing all the dimensionality reduction techniques described in this review.

data, and are in consequence very sensitive to over-fitting problems. The most appealing methods are based on Artificial Neural Networks, as for instance the one proposed by Hinton et al. [34], but are also the most sensitive to the over-fitiing problem.

To overcome the problems of training and over-fitting it has been proposed to use Kernel methods, such as Kernel PCA or Support Vector Machines [35]. The principle of these implementations is to use a well defined function called Kernel function which will transform the data-space first into a much higher dimensional space. The supposition is that non-linear correlations in the input space will be transformed into linear ones in the higher dimensional kernel-transformed space. Techniques of hypersurface separators such as Support Vector Machines can then be used. Such Manifold Learning techniques are not necessarily well known to non-experts and their application to the analysis of "omics" data is rare [36-40]. The major drawback of Kernel based methods is clearly the fact that to the use of a Kernel function prevents obtaining biological relevant insights into the underlying structure of the data.

3 Combining linear and non-linear methods: SVD-MDS

We discussed in the previous sections the state-of-the-art of techniques for dimensionality reduction (see Fig 1 for a summarizing scheme) using linear methods which will map the data by linear combination, and non-linear methods which search the best possible mapping considering constraints resulting from the conservation of geometric structures. As we have seen both families of techniques suffer from considerable drawbacks. While linear methods are relatively easy to implement and compute, much information of the initial data-structure is lost. Non-linear methods are more faithful in their representations (compare Fig 2A and 2B) as they produce lower resulting Kruskal stress values indicating better representation of the object in low dimensional space. However, they are sometimes prohibitive in computational cost, and always much more challenging to implement. In some cases the only available algorithmic solutions to the complexity problem lead to good representations which are, however, of reduced relevance to biological interpretation.

The need for a computational efficient technique which is easy to implement and does not rely on the fitting of parameters has lead more recently to the idea to combine both linear and non-linear methods in order to overcome their

respective shortcomings. We have introduced such a novel hybrid approach called SVD-MDS [41]. This method consists of using Molecular Dynamics-driven Multidimensional Scaling where the simulation is initialized by the results of *a priori* Singular Value Decomposition.

The basic idea of a Molecular Dynamics based MDS algorithm (MD-MDS) is to virtually connect each data point to all other instances with springs. As a spring has an equilibrium length during a molecular dynamics simulation it will tend to go back to its equilibrium state. The equilibrium length for the spring between point i and point j will be defined as the Euclidean distance $d(x^i, x^j)$ in the initial state. So for each instance x^i a force is defined $F(x^i)$, which is the sum of all spring interactions $F_{spring}(x^i, x^j)$ with the other instances x^j, minus a friction term to avoid oscillation of the spring network:

$$F_{spring}(x^i, x^j) = -k_{ij}(\delta(x^i, x^j) - d(x^i, x^j))(x^j - x^i) \quad (8)$$

$$F(x^i) = \sum_{j \neq i} F_{spring}(x^i, x^j) - \gamma m_i \dot{x}^i \quad (9)$$

with $\delta(x^i, x^j)$ being the distance between instances in the r dimensional space, k_{ij} the strength of spring ij, γ the friction parameter, and m_i the mass given to each point. We consider that every spring and all instances are equal in strength and weight so k_{ij} and m_i are the same for every i and j ($k_{ij} = k$ and $m_i = m$). It is, however, possible to use different parameters, for instance as a function of the quality of the individual measures. A molecular simulation using the force vector is then executed using a Verlet Algorithm. The Kruskal stress, which is linked to the free energy of the system, is calculated at each time step of the simulation. As physical systems tend to minimize free energy, the Kruskal stress is a measure of the quality of the data-representation given by the MDS procedure.

The initialization of MDS is usually done using randomized positions or by placing all data-point at the origin. We found, however, that using the results of SVD for the initialization increases significantly computational efficacy and overcomes the inherent problem of the non-uniqueness of the final representation [41]. This result is backed by theoretical considerations as the Eckart-Young theorem [42] demonstrates that an SVD-based representation best approximates the distance matrix. Our algorithm is thus divided in two steps for a dimensionality reduction of a matrix X which is $n \times m$ to p dimensions:

First one reduces the number of dimension from m to $rank(X)$ using SVD. One can perform this step even if n or m is big as long as the other is small. This operation will not induce deformation of Euclidean distances between points due to the conservation of inner-product matrices. Then one deletes dimensions from $rank(X)$ to p, and reduces the deformation induced by this operation using MD-MDS to a minimum.

The proposed algorithm was shown to be much more computationally efficient than with any other known method of initialization, and this by using results of stress minimization on thirteen different datasets, some of which are publicly available from gene expression experiments [43, 44] and from the machine learning reference website [45], others having been stochastically generated. It was also proven that adding stochasticity to MD-MDS algorithm does neither have similar increases in computational efficacy nor improve the final representations, contrary to what was previously suggested [21]. Finally it was possible to demonstrate that immediate removal of all but p dimensions (with the largest contributions) after SVD was more beneficial than going through an iterative cycle of step-wise dimension deletion plus reduction of deformation.

As the Kruskal stress is a parameter evaluating the statistical error between the Euclidean distances before and after dimensionality reduction, this parameter is more influenced by large distances then by small ones [46]. To evaluate the accuracy of the representation obtained by SVD-MDS, we propose to use in conjunction with Kruskal stress a parameter for evaluating the k-nearest neighbor relative changes which amounts to evaluating the change in local organization of the data-object.

Using this additional measure it can be easily appreciated that SVD-MDS gives in the vast majority of the cases better, or at least equivalent

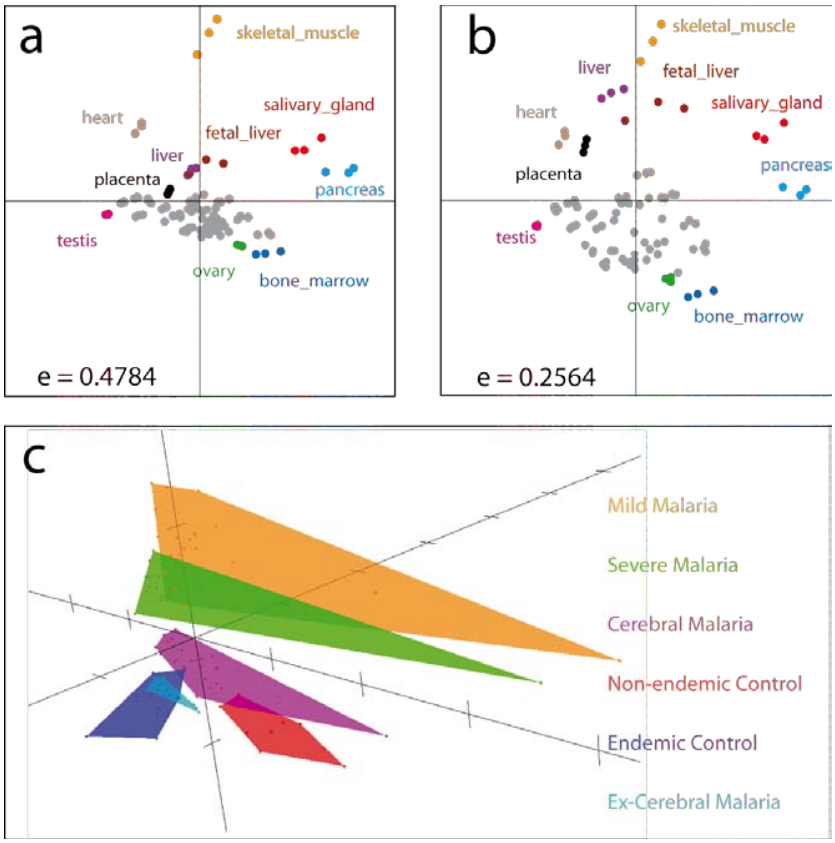


Figure 2: Comparison of linear and non-linear dimensionality reduction techniques and application to transcriptome and proteome data. (a) Singular Value Decomposition of a reference transcriptome dataset containing 32 human healthy tissues in triplicate recordings [44]. (b) SVD-MDS based representation of the same dataset. For both a subset of tissues was colored to allow visual comparison. The Kruskal stress e of both representations is indicated. (c) Application of the SVD-MDS algorithm to a dataset of multiple cytokine recordings consisting of six groups of Malaria patients and appropriate control subjects totaling 98 individuals is shown. In this 3D representation convex hulls are calculated for each of the six groups. The data are discussed in detail in Bensal et aL [47].

results than any other tested method. This improve in local organization results proves that SVD-MDS is a very good choice for obtaining accurate representation of your data and then get a clear insight in the underlying biological processes.

A first application of this novel combination of non-linear and linear methods to dimensionality reduction can be found in [47], where the SVD-MDS algorithm was successfully applied to the analysis of patterns of cytokine expression in a large cohort of patients suffering from different forms of Malaria. Here we show a new representation of these data generated from 98 indiviuals in three dimensions (Fig 2C) where a clear separation between the different groups can be obtained using the novel SVD-MDS approach. Thus, the cytokine activity measurements in the different patients, when analyzed using the SVD-MDS algotrithm, allows a classification of the different Malaria patiens and control groups according to the severity of the disease. Apart from the higher resolution (Fig 2A vs. 2B) that is obtained with the SVD-MDS algorithm, the main advantage of this new class of algorithms is the significantly reduced computational load when compared to standard non-linear techniques.

Five-year view

As both the quantity of available datasets and the number of variables per dataset are increasing exponentially, the need for faithful dimensionality reduction in the biological analysis will increase dramatically. As an increase in dimensionality of the input data translates to more relevant information also being contained in higher dimensions, linear techniques such as PCA will be less and less adaptable to the analysis of "omics" data. In the near

future we will thus certainly see an increasing use of non-linear dimensionality reduction techniques. As these techniques pose other complications such as high computational cost, uncertainty, and non-uniqueness of representation there is much room for novel theoretical and algorithmic development. Also, as biological phenomena are usually multifactorial both in their origin and consequences, methods using more than a single optimization parameter could emerge, leading to context-dependent [48] dimensionality reduction. The use of data generated from very large cohorts of patients, as is already common place in genome-wide association studies, also amplifies problems related to missing values, and by consequence should spur the incorporation of missing value imputation algorithms [49] into the dimensionality reduction process. A major challenge will be the development of representations of different biological scales / datasets in the same space as to render the results obtained for different datasets inter-comparable. Here, finding useful definitions of reference spaces will be both the most daunting task, however, will profit from the increasingly available data that start to well sample the entire space of for instance the human transcriptome. Finally, and most importantly, we will hopefully see the development of deep links between dimensionality reduction techniques and methods for the inference of biological networks, as only a combination of both technologies will lead to a deeper understanding of the biological phenomena at work.

Key issues

• Dimensionality reduction of high-dimensional "omics" data is a prerequisite for their analysis.

• Dimensionality reduction can be separated in linear and non-linear methods, the difference between them is that in the first case one supposes linear correlations between variables whereas in the second one supposes non-linear correlations.

• They already exists a lot of application of these techniques to "omics" data

• It is noteworthy, that so far the development of dimensionality reduction techniques, as they were inspired by other fields, has preceded the development of "omics" analysis methods, and they are thus little to not specifically adapted to biological data and hypotheses.

• Shortcomings of the existing approaches have been discussed and center on computational efficacy, information preservation, and usefulness for biologic interpretation. Only very recently, combined methods are starting to emerge which overcome some of these shortcomings [41].

• The future will bring techniques directly inspired by biological reality and the increasing amounts of data to be analyzed.

Financial disclosure/Acknowledgements

We are grateful to Brendan Bell and the anonymous reviewers for their helpful comments on the manuscript. This work was funded by the *Centre National de la Recherche Scientifique* (C.N.R.S.), the *Agence Nationale pour la Recherche contre le SIDA et les hépatites virales* (A.N.R.S.), the *Agence Nationale pour la Recherche* (A.N.R., ISPA project), and the *Genopole Evry*.
CB is recipient of a Ph.D. fellowship from the A.N.R.S..

References

[1] * Clarke R, Ressom HW, Wang A, Xuan J, Liu MC, Gehan EA, Wang Y The properties of high-dimensional data spaces: implications for exploring gene and protein expression data. *Nat Rev Cancer* 8, 37--49 (2008). *A review explaining the specifics properties of high dimensional data.*

[2] Pearson K On lines and planes of closest fit to systems of points in space. *Philosophical Magazine* 2(6), 572--559 (1901).

[3] * Schmidt E, Stewart GW, Stewart GW On the Early History of the Singular Value Decomposition. *University Of Maryland*, 1992. *A complete review on the development and mathematical foundations of SVD.*

[4] Alter O, Brown PO, Botstein D Singular value decomposition for genome-wide expression data processing and modeling. *PNAS* 97(18), 10101--10106 (2000).

[5] Alter O, Brown PO, Botstein D: Generalized singular value decomposition for comparative analysis of genome-scale expression data sets of two different organisms. *PNAS* 100(6), 3351--3356 (2003).

[6] Omberg L, Golub GH, Alter O A tensor higher-order singular value decomposition for integrative analysis of DNA microarray data from different studies. *PNAS* 104(47), 18371--18376 (2007).

[7] Wall M, Rechtsteiner A, Rocha L Singular Value Decomposition and Principal Component Analysis. In *A Practical Approach to Microarray Data Analysis*, Springer US, 91--109 (2003).

[8] de Haan JR, Wehrens R, Bauerschmidt S, Piek E, van Schaik

RC, Buydens LMC Interpretation of ANOVA models for microarray data using PCA. *Bioinformatics* 23(2), 184--190 (2007).

[9] Jonnalagadda S, Srinivasan R Principal components analysis based methodology to identify differentially expressed genes in time-course microarray data. *BMC Bioinformatics* 9(267), (2008).

[10] Hubert M, Engelen S Robust PCA and classification in biosciences. *Bioinformatics* 20(11), 1728--1736 (2004).

[11] Strickert M, Sreenivasulu N, Usadel B, Seiffert U Correlation-maximizing surrogate gene space for visual mining of gene expression patterns in developing barley endosperm tissue. *BMC Bioinformatics* 8(165), (2007).

[12] Culhane AC, Perriere G, Considine EC, Cotter TG, Higgins DG Between-group analysis of microarray data. *Bioinformatics* 18(12), 1600--1608 (2002).

[13] ** Cox TF, Cox MAA *Multidimensional scaling, Second Edition*. Chapman & Hall/CRC (2001). *Reference book on multidimensional scaling.*

[14] Borg I, Groenen PJF *Modern multidimensional scaling: theory and applications*. Springer (2005).

[15] Cuadras CM, Fortiana J Metric Scaling Graphical Representation of Categorical Data. *Penn State University* (1995).

[16] Fellenberg K, Hauser NC, Brors B, Neutzner A, Hoheisel JD, Vingron M Correspondence analysis applied to microarray data. *PNAS* 98(19), 10781--10786 (2001).

[17] Kruskal JB, Wish M *Multidimensional scaling*. SAGE 1978.

[18] Dzwinel WR, Blasiak JR Method of particles in visual clustering of multi-dimensional and large data sets. *Future Generation Computer Systems* 15(3), 365--379 (1999).

[19] Dzwinel W Virtual particles and search for global minimum. *Future Generation Computer Systems* 12(5), 371--389 (1997).

[20] Andreas B, Swayne DF, Littman ML, Nathaniel D, Hofman H Interactive Data Visualization with Multidimensional Scaling. Tech. rep., University of Pennsylvania (2004).

[21] Andrecut M Molecular dynamics multidimensional scaling. *Physics Letters A* 373(23-24), 2001-2006 (2009).

[22] Gray LC, Vaidya JS, Baum M, Badwe RA, Mittra I, Siddiqui T, Wiarda D Functional maps of metastases from breast cancers: proof of the principle that multidimensional scaling can summarize disease progression. *World Journal of Surgery* 28(7), 646--651 (2004).

[23] Taguchi Y, Oono Y Relational patterns of gene expression via non-metric multidimensional scaling analysis. *Bioinformatics* 21(6), 730-740 (2005).

[24] Ebbels TMD, Buxton BF, Jones DT springScape: visualisation of microarray and contextual bioinformatic data using spring embedding and an 'information landscape'. *Bioinformatics* 22(14), 99--107 (2005).

[25] Tzeng J, Lu HH, Li W Multidimensional scaling for large genomic data sets. *BMC Bioinformatics* 9(179), (2008).

[26] Deun KV, Marchal K, Heiser WJ, Engelen K, Mechelen IV Joint mapping of genes and conditions via multidimensional unfolding analysis. *BMC Bioinformatics* 8(181), (2007).

[27] Hastie T, Stuetzle W Principal Curves. *Journal of the American Statistical Association* 84(406), 502--516 (1989).

[28] Kramer M Nonlinear principal components analysis using auto-associative neural networks. *AIChE Journal* 37(2), 233--243 (1991).

[29] Tenenbaum JB, de Silva V, Langford JC A Global Geometric Framework for Nonlinear Dimensionality Reduction. *Science* 290(5500), 2319--2323 (2000).

[30] Roweis ST, Saul LK Nonlinear Dimensionality Reduction by Locally Linear Embedding. *Science* 290(5500), 2323--2326 (2000).

[31] ** Gorban AN, Kgl B, Wunsch DC, Zinovyev A *Principal Manifolds for Data Visualization and Dimension Reduction*. Springer Publishing Company, (2007). *A reference book on all techniques of non-linear dimensionality reduction.*

[32] Dawson K, Rodriguez RL, Malyj W Sample phenotype clusters in high-density oligonucleotide microarray data sets are revealed using Isomap, a nonlinear algorithm. *BMC Bioinformatics* 6(195), (2005).

[33] Vapnik VN *The nature of statistical learning theory*. Springer (2000).

[34] Hinton GE, Salakhutdinov RR Reducing the Dimensionality of Data with Neural Networks. *Science* 313(5786), 504--507 (2006).

[35] ** Berthold M, Hand DJ *Intelligent data analysis: an introduction*. Springer (2003). *A very well written reference book discussing all the different existing techniques for analysing data.*

[36] Furey TS, Cristianini N, Duffy N, Bednarski DW, Schummer M, Haussler D Support vector machine classification and validation of cancer tissue samples using microarray expression data. *Bioinformatics* 16(10), 906--914 (2000).

[37] Hua S, Sun Z A novel method of protein secondary structure prediction with high segment overlap measure: support vector machine approach. *Journal of Molecular Biology* 308(2), 397--407 (2001).

[38] Meyer D, Leisch F, Hornik K The support vector machine under test. *Neurocomputing* 55(1-2), 169--186 (2003).

[39] Pochet N, Smet FD, Suykens JAK, Moor BLRD Systematic benchmarking of microarray data classification: assessing the role of non-linearity and dimensionality reduction. *Bioinformatics* 20(17), 3185--3195 (2004).

[40] Komura D, Nakamura H, Tsutsumi S, Aburatani H, Ihara S Multidimensional support vector machines for visualization of gene expression data. *Bioinformatics* 21(4), 439--444 (2005).

[41] Becavin C, Tchitchek N, Mintsa-Eya C, Lesne A, Benecke A Molecular dynamics multidimensional scaling initialized by singular value decomposition leads to computationally efficient analysis of high dimensional data. *Submitted*.

[42] Eckart C, Young G The approximation of one matrix by

another of lower rank. *Psychometrika* 1(3), 211-218 (1936).

[43] Iyer VR, Eisen MB, Ross DT et al. The Transcriptional Program in the Response of Human Fibroblasts to Serum. *Science* 283(5398), 83-87 (1999).

[44] Dezso Z, Nikolsky Y, Sviridov E et al. A comprehensive functional analysis of tissue specificity of human gene expression. *BMC Biology* 6(49), (2008).

[45] Asuncion A, Newman D UCI Machine Learning Repository 2007, [[http://www.ics.uci.edu/ mlearn/MLRepository.html]] .

[46] Graef J, Spence I Using distance information in the design of large multidimensional scaling experiments. *Psychological Bulletin* 86, 60--66 (1979).

[47] Bansal D, Herbert F, Lim P, et al. IgG Autoantibody to Brain Beta Tubulin III Associated with Cytokine Cluster-II Discriminate Cerebral Malaria in Central India. *PLoS ONE* 4(12), e8245 (2009).

[48] Lesne A, Benecke A Feature context-dependency and complexity-reduction in probability landscapes for integrative genomics. *Theoretical Biology & Medical Modelling*, 5(21), (2008).

[49] * Brock G, Shaffer J, Blakesley R, Lotz M, Tseng G Which missing value imputation method to use in expression profiles: a comparative study and two selection schemes. *BMC Bioinformatics* 9(12), (2008). *Journal article discussing the problem of missing values.*

.

D

Journal article : Molecular dynamics multidimensional scaling initialized by singular value decomposition leads to computationally efficient analysis of high dimensional data.

Christophe Bécavin, Nicolas Tchitchek, Colette Mintsa-Eya, Annick Lesne, Arndt Benecke.
Submitted to BioInformatics journal in October 2010.

Abstract

Multidimensional scaling (MDS) is a well known multivariate statistical analysis method used for dimensionality reduction and visualization of similarities and dissimilarities in multidimensional data. The advantage of MDS with respect to Singular Value Decomposition (SVD) based methods such as Principal Component Analysis (PCA) is its superior fidelity in representing the distance between different instances specially for high-dimensional geometric objects. Molecular dynamics (MD) simulations have been recently shown [39] to be an ideal way of computing Multidimensional Scaling. Here we investigate the importance of the choice of initial conditions for MD-driven MDS (MD-MDS), and show that SVD is the best choice to initiate the MD simulation. Furthermore, we demonstrate that the use of the first principal components of SVD to initiate the MD-MDS algorithm is more efficient than an iteration through all the principal components. Therefore cycles of dimensionality reduction and reorientation of the object do not add further to the efficiency or the accuracy of the representation. Adding stochasticity to the molecular dynamics simulation, contrary to a previous suggestion [39], likewise does not increase accuracy. Finally, we introduce a k nearest neighbor method to analyse the local structure of the geometric objects and use it to control the quality of the dimensionality reduction. Structural analysis of the geometric objects also can be used to define lower bounds on the number of dimensions to be attained by the reduction technique with a controlled loss of distance information.

We demonstrate here, to our knowledge, the most efficient and accurate initialization strategy for MD-MDS algorithms, reducing considerably computational load. SVD-based initialization renders MDS methodology much more useful in the analysis of high-dimensional data such as functional genomics datasets.

Published online

BioInformatics, , Vol. , No. ***1–11***
doi:10.1093/nar/gkn000

Molecular dynamics multidimensional scaling initialized by singular value decomposition leads to computationally efficient analysis of high dimensional data.

Christophe Bécavin [1,2], Nicolas Tchitchek [1], Colette Mintsa-Eya [1,2], Annick Lesne [1,3], Arndt Benecke [1,2]
*To whom correspondence should be addressed. Tel: +33 1 60 92 66 65; Fax: +33 1 60 92 66 09; Email: arndt@ihes.fr

[1]Institut des Hautes Études Scientifiques - 35 route de Chartres - 91440 Bures sur Yvette - France and [2]Institut de Recherche Interdisciplinaire CNRS USR3078 Univ. Lille I, II - 50 avenue de Halley - 59658 Villeneuve d'Ascq - France and [3]Laboratoire Physique Théorique de la Matière Condensée CNRS UMR7600 Univ. Pierre et Marie Curie Paris 6 - 4 place Jussieu - 75252 Paris Cedex 05 - France

Received ; Revised ; Accepted

ABSTRACT

Multidimensional scaling (MDS) is a well known multivariate statistical analysis method used for dimensionality reduction and visualization of similarities and dissimilarities in multidimensional data. The advantage of MDS with respect to Singular Value Decomposition (SVD) based methods such as Principal Component Analysis (PCA) is its superior fidelity in representing the distance between different instances specially for high-dimensional geometric objects. Molecular dynamics (MD) simulations have been recently shown (1) to be an ideal way of computing Multidimensional Scaling. Here we investigate the importance of the choice of initial conditions for MD-driven MDS (MD-MDS), and show that SVD is the best choice to initiate the MD simulation. Furthermore, we demonstrate that the use of the first principal components of SVD to initiate the MD-MDS algorithm is more efficient than an iteration through all the principal components. Therefore cycles of dimensionality reduction and reorientation of the object do not add further to the efficiency or the accuracy of the representation. Adding stochasticity to the molecular dynamics simulation, contrary to a previous suggestion (1), likewise does not increase accuracy. Finally, we introduce a *k* nearest neighbor method to analyse the local structure of the geometric objects and use it to control the quality of the dimensionality reduction. Structural analysis of the geometric objects also can be used to define lower bounds on the number of dimensions to be attained by the reduction technique with a controlled loss of distance information.

We demonstrate here, to our knowledge, the most efficient and accurate initialization strategy for MD-MDS algorithms, reducing considerably computational load. SVD-based initialization renders MDS methodology much more useful in the analysis of high-dimensional data such as functional genomics datasets.

INTRODUCTION

The appropriate and faithful visualization of high-dimensional data is often a prerequisite for their analysis as the human visual cortex is still one of the most powerful tools to detect and conceptualize structure in data (2). Furthermore, communication of numerical and statistical results is greatly aided by the intuition arising from appropriate representations of data. Different methods for the required dimensionality reduction have been developed (3). An entire family of approaches, such as Principal Component Analysis (PCA) finds the minimal orthonormal basis using a mathematical tool called Singular Value Decomposition (SVD). These methods, using different similarity or dissimilarity measures such as covariance or correlation, order the *ensemble* of components by their statistical deviation, and for visualization only the first two or three components are retained. Thereby, the statistical information in the first components are entirely retained, whereas one of the subsequent components is entirely lost. Today's high-dimensional datasets can easily contain thousands of instances (number of measures) with 10^5-10^9 variables (number of parameters measured). A prominent example for such datasets are microarray and so-called 'deep-sequencing' data generated in the field of functional genomics (4, 5). Not only are such data high-dimensional, but they are also relatively homogeneous with respect to the repartition of information over the entire number of variables. Many small differences in many different variables define the distance between instances (experimental conditions), rather than a few large differences in few variables. In consequence, considering only the first components given by SVD based techniques is not necessarily the best choice. Multidimensional Scaling (MDS) is a methodology that reduces dimensionality of the geometric object by projecting the instances into a lower dimensional space. The only information needed for this technique is similarities or dissimilarities between instances, hereafter regrouped in the general term of "distance". During the process a part of this distance information will be lost. It hence results an optimization problem of finding an arrangement of the instances in the lower dimensional

space that reflects the least loss of distance information when compared to the original distances in the higher dimension.

As all optimization problems, the search for an optimal configuration, is reduced to finding the global minimum of a function. To be sure to find an acceptable minima, there is two things which have to be well chosen: (i) an initial state for the optimization algorithm, (ii) an optimization algorithm and the appropriate parameters. Recently it has been shown (1) that the best choice for the second is a Molecular Dynamics Multidimensional Scaling approach. We demonstrate here that the choice of the initial position is paramount to the quality of the representation and computationally efficient. By using SVD for providing an initial configuration for the MDS, we obtain a significantly increased computational efficacy. Furthermore, we provide a method to verify correctness of the representation. Interestingly, we also demonstrate that adding stochastic energy during the MD-MDS execution does not increase performance or reproducibility of the algorithm. In most cases SVD-MDS also outperforms such methods, and in all cases performs at least as efficiently.
We also investigate the dynamics of the global and local structure of the geometric objects during iterative dimensionality reduction using our methodology, and then evaluate the usefulness of the different approaches developed here. These investigations and the use of SVD to the initial state allow to better define and control the dimensionality reduction process for high-dimensional data.

MATERIALS AND METHODS

Singular value decomposition

Given a data matrix X with n rows and p columns and x_{ij} its value in row i and column j, we denote $\bar{X}_i$ the p components vector corresponding to row i of the matrix, and $\underline{X}_j$ the n components vector corresponding to column j of the matrix. A set of vector $\bar{X}_i$ is then a set of instances, whereas a set of vector $\underline{X}_j$ is a set of variables. In all the following article we will use this notation for vectors extracted from X.

One of the most used techniques for performing dimensionality reduction is principal component analysis (PCA). It is based on the principle of finding an orthonormal basis, with a minimal number of components, for embedding the data points (instances). PCA belongs to a family of methods based on this principle. In subsequent sections we show that those techniques are closely related to one another, and are based on a general matrix analysis technique called Singular Value Decomposition (SVD). It is known (6) that every rectangular matrix can be decomposed using its singular values:

$$X = USV^t \quad \textbf{(1)}$$

where U (matrix containing left singular vectors) and V (matrix containing right singular vectors) are both square orthogonal matrices ($UU^t = U^tU = Id$ and $VV^t = V^tV = Id$, with Id the unit matrix), and S is a rectangular matrix containing the singular values (s_i) which are positive. If n is the number of rows and p the number of columns of X, X is $n.p$, U is $n.n$, S is $n.p$, V is $p.p$. In case $n = p$, S is a square diagonal matrix, if $n \neq p$, $S_{ii} = s_i$ and $S_{ij} = 0$. U is n.n This decomposition is very useful in dimensionality reduction because of its link with the eigenvalue decomposition of the inner-product (XX^t) and outer-product (X^tX) of X.

If $X = USV^t$ we have :

$$XX^t = USV^t(VS^tU^t) = USS^tU^t \quad \textbf{(2)}$$

$$X^tX = (VS^tU^t)USV^t = VS^tSV^t \quad \textbf{(3)}$$

SS^t and S^tS are two diagonal square matrices, they do not have the same size but they have the same number of non null values on the diagonal (6). So SVD allows to demonstrate that the inner and outer product have the same eigenvalues λ_i, with $\lambda_i = s_i^2$. It also gives a matrix link between them, and it is very useful as all the classical techniques of dimensionality reduction are performed using one of these products, after the right renormalisation of the data.

Generally before performing SVD X is centered, so the mean of each column is equal to zero. In this context, $rank(X) = rank(S) \leq min(n-1, p)$ if X is $n.p$. Then, the simplest way to find SVD, is to search first for the eigenvalues and the eigenvectors of the inner and outer products. As finding the eigenvalues of a matrix X with n rows and p columns, is hard to perform for objects with a high number of variables, this step is only feasible if either n or p are small (typically inferior to 1000). If both, the number of rows and the number of columns were high, it is impossible to perform SVD with linear techniques, and one is obliged to use iterative Singular Value Decomposition techniques as shown in (6).

However, if either n or p is small, which is often the case in biological datasets, we will be able to perform SVD, again because of the close link between inner and outer products. Then U,S, and V, can be reorganized in order to have $s_1 > s_2 > ... > s_r$, with r being the rank of S. the new matrix will then be given by:

$$X_{new} = XV \quad \textbf{(4)}$$

V is orthogonal so this transformation has conserved the distance information between points without any deformation.

Note that missing values in data can also be imputed using SVD (7) (8, 9). If the number of missing values is relatively low, the Eckart Young theorem (10), which is the most commonly used theorem for matrix approximation, assures that the result of the SVD will change only in the value of the last singular values. Hence, for a rapid imputation, the row average method (7), can be used which is sufficiently precise in most cases. In this case we replace each missing values by the mean of the statistical sample the value is supposed to be part of.

The outer product is the covariance matrix on all components up to a scalar. The covariance matrix diagonal is formed by the variances of each component, thus we can conclude that each singular value s_i is the statistical deviation of *i-th* component. The new basis we have obtained is also an orthonormal basis, because the covariance matrix

is diagonal, were each component is ordered by order of its relative contribution to the general pattern. An inertia parameter is defined to evaluate this relative contribution. For one component inertia is $c_i = s_i / \sum_i s_i$, so $\sum_i c_i = 1$. The inertia vector is thus very important for us as it carries geometric information on the form of our cloud of points.

In conclusion, singular value decomposition provides three major types of information:

(i) A new data matrix X_{new}, which represent the data points in a new orthonormal basis with a minimum number of components, and where distance between the instances is preserved.

(ii) Inertia parameters indicate the standard deviation and relative contribution of the cloud of points on each principal component.

(iii) The matrix V in which we have the contribution to each principal component of all the component in the former basis. These different types of information have already previously been used in the literature to infere biological knowledge in various settings (11, 12, 13, 14).

X_{new} is a representation of the data in a r dimensional orthonormal basis (with $r = rank(S)$), this representation is well chosen as distances are conserved in the cloud of points (orthonormal transformation). Also, it has been demonstrated that principal component analysis results is a very good choice for the initial state in order to perform K-means clustering (15). In the new representation given by SVD, cluster structure of the data will then naturally appear, and thus provides a natural interpretation of clusters.

We discuss the links of SVD with Principal Component Analysis (PCA, section), Classical Scaling (cMDS, section), Correlation Analysis (section), and Correspondence Analysis (section) below.

SVD and PCA

Principal component analysis (PCA) goals are orthonormalization of the basis and classification by dimension's variance, in practice it corresponds to the search of the eigenvectors' covariance matrix. For a centered data matrix X, the corresponding covariance matrix corresponds to:

$$CovMatrix = \frac{1}{n} X^t X \quad \textbf{(5)}$$

hence, performing PCA reduces to finding the outer-product's eigen-vectors: $X^t X = V \Lambda V^t$. The singular-values of X are the square root of the outer-product's eigen-values, and V is the matrix of the right singular vectors. The link between PCA and SVD then becomes obvious (13).

Classical scaling

Classical scaling (cMDS for classical multidimensional scaling) was invented to embed a set of instances in the simplest space possible, with the constraint of preserving the Euclidean distance between data points. Euclidean distance can be written as a sum of inner-products $\bar{X}_i . \bar{X}_j$, and we can pass from an Euclidean distance matrix to an inner product matrix by a simple matrix manipulation called double centering. Torgerson (16) uses this link so that classical scaling can be performed using the eigen-value factorization of the inner-product matrix of a centered matrix X: $XX^t = U \Lambda U^t$. Singular-values of X are the square root inner-product's eigen-values, to determine embedding of the instances in the new space, one needs to find the right singular vector matrix V. It is given by: $V = X^t U^t \Lambda^{-1/2}$. The new data matrix will then become: $X_{new} = XV$. So PCA and Classical Scaling give the same results, a fact reflected by Classical Scaling sometimes being referred to Principal Coordinate Analysis.

SVD and Correlation Analysis

Correlation analysis is used to analyze data normalized with statistical deviation of each variable. Between two variables of X it is calculated as: $corr(\underline{X}_i, \underline{X}_j) = \frac{cov(\underline{X}_i, \underline{X}_j)}{\sigma(\underline{X}_i)\sigma(\underline{X}_j)}$, where $\sigma(\underline{X}_i)$ is the statistical deviation on variable i. So if we transform the centered data matrix X to: $\tilde{X} = \left(\frac{x_{ij}}{\sigma(\underline{X}_j)} \right)$, the correlation matrix will now be given as the outer-product $\tilde{X}\tilde{X}^t$. The new basis obtained after normalization will be called *correlation basis*. Finding the minimal space for representing this deviation rescaled data will then can be represented using PCA or cMDS on $\tilde{X}$, in other words SVD based methods.

SVD and Correspondence analysis

In statistics, correspondence analysis (17) is used in the case of a contingency table, which is a table obtained after an operation of counting on categorical data (see (3) for more information on contingency table and categorical data). A typical example of data that can be enclosed in a contingency table are histograms. Such sorts of tables with n rows and p columns, will correspond to n histograms, and p values of count for each histogram. In this particular example correspondence analysis will be used to compare all the histograms' distribution profile, by defining a value of distance between them. Correspondence Analysis will represent the set of histograms in a low dimensional space. This method can also be used for microarray data analyses (14) as each value of gene expression is in fact a count of the number of RNAs produced.

Generally speaking this technique is used to compare two vectors in terms of their distribution profile. To this end the chi-square distance is being used. When the distance is equal to zero, both vectors have the same statistical distribution. Between two vectors $\bar{X}_i$ and $\bar{X}_j$ of a data matrix X, the distance is defined as (17):

$$(\chi^2(\bar{X}_i, \bar{X}_j))^2 = \sum_{k=1}^{p} \frac{1}{f_k} \left(\frac{f_{ik}}{f^i} - \frac{f_{jk}}{f^j} \right)^2 \quad \textbf{(6)}$$

Where $W = \sum_{i=1}^{n} \sum_{j=1}^{p} x_{ij}$ is the total sum, $f_{ik} = x_{ik}/W$ is the relative frequency, $f^i = \sum_{l=1}^{p} f_{il}$ is the marginal row frequency, and $f_k = \sum_{l=1}^{n} f_{lk}$ is the marginal column frequency.

We can develop χ^2 distance, using the variables $\tilde{X}_{ik} = \frac{x_{ik}\sqrt{W}}{(\sqrt{\sum_l x_{lk}})(\sum_l x_{il})}$, we have at the end :

$$(\chi^2(\bar{X}_i,\bar{X}_j))^2 = \sum_{k=1}^{p}(\tilde{x}_{ik}-\tilde{x}_{jk})^2 = (d(\tilde{\bar{X}}_i,\tilde{\bar{X}}_j))^2 \quad (7)$$

So χ^2 distance is in fact an Euclidean distance if the data matrix is properly rescaled. To find the minimal space which embeds the data and conserves the information of χ^2 distance, one needs to perform a cMDS or PCA on the rescaled data matrix.

Multidimensional scaling

Multidimensional scaling (MDS) is a class of techniques their goal being to represent instances in an r dimensional space given an initial state and a similarity or dissimilarity matrix (18, 19). The first exact method of MDS was proposed by Torgerson in 1952 (16), and is called nowadays Classical Scaling (see previous section. It only works when euclidean distances are provided to the algorithm. We show in that due to the link given by SVD between inner and outer product, Classical Scaling gives the same results as PCA. Those methods are exact if we reduce dimensionality to a r dimensional basis, with r being the rank of the singular value matrix. If we go further in the reduction it will induce a deformation in the distance between points, and one then needs an optimization method for the reduction. In conclusion, as for all optimization processes, the major problem in developping MDS techniques is to determine an initial state and choose a good optimization criterion and algorithm.

Recently, Molecular Dynamics (MD) approaches have been used to perform MDS for high-dimensional objects drastically increasing quality of the dimensionality reduction (1). We had in parallel developed a similar approach based on a spring analogy to perform MDS. The idea is to virtually connect each data point to all other instances with springs. As a spring has an equilibrium length, if we run a molecular dynamics simulation on it, it will tend to go back to its equilibrium state. The equilibrium length for the spring between point i and point j will be defined as the Euclidean distance $d(\bar{X}_i,\bar{X}_j)$ in the initial state. So for each instance $\bar{X}_i$ a force is defined $F(\bar{X}_i)$, which is the sum of all spring interactions $F_{spr}(\bar{X}_i,\bar{X}_j)$ with the other instances $\bar{X}_j$, minus a friction term to avoid oscillation of the spring network:

$$F_{spr}(\bar{X}_i,\bar{X}_j) = -k_{ij}(\delta(\bar{X}_i,\bar{X}_j)-d(\bar{X}_i,\bar{X}_j))(\bar{X}_j-\bar{X}_i) \quad (8)$$

$$F(\bar{X}_i) = \sum_{j\neq i} F_{spr}(\bar{X}_i,\bar{X}_j) - \gamma m_i \dot{\bar{X}}_i \quad (9)$$

with $\delta(\bar{X}_i,\bar{X}_j)$ being the distance between instances in the r dimensional space, k_{ij} the strength of spring ij, γ the friction parameter, and m_i the mass given to each point. We consider that every spring and all instances are equal in strength and weight so k_{ij} and m_i are the same for every i and j ($k_{ij}=k$ and $m_i=m$). It is, however, possible to use different parameters — for instance according to experimental precision — if different weights shall be considered for the different instances. A molecular simulation using the force vector is then executed. Following Newton's law it follows: $m_i\ddot{\bar{X}}_i = F(\bar{X}_i)$, with $\ddot{\bar{X}}_i$ the double temporal derivation of vector $\bar{X}_i(t)$. In order to find the new position and velocity of our data points at the next simulation time we then use simple verlet integration:

$$\bar{X}_i(t+\Delta t) = 2\bar{X}_i(t) - \bar{X}_i(t-\Delta t) + A\Delta t^2 \quad (10)$$

$$\dot{\bar{X}}_i(t) = \frac{\bar{X}_i(t+\Delta t) - \bar{X}_i(t-\Delta t)}{2\Delta t} \quad (11)$$

with $\dot{\bar{X}}_i(t)$ the temporal derivation of vector $\bar{X}_i(t)$. The algorithm is run with simulation time t increasing. To avoid divergence of the Verlet algorithm parameters of the simulation k, m, γ Δt have to be well chosen. In all the simulations we made we choose: $k=1$, $m=5$, $\gamma=0.1$ $\Delta t = 0.02$, as an empirical study have shown those parameters to be the most useful. We also discover that a good way to provide divergence was modifying all initial states provided to the MDS algorithm by rescaling them to fit in a hypercube with a diameter of 6. This rescaling only consist in multiplying the initial state matrix by a scalar α. It will deform all distances between instances the same way, so it does not influence the organization of the cloud of points. The only initial state matrices we will consider from now on have been rescaled according to this operation.

To control the minimization process at each time step, a cost function termed the Kruskal stress is calculated according to (19):

$$e = \sqrt{\frac{\sum_i \sum_j (\delta(i,j)-d(i,j))^2}{\sum_i \sum_j d(i,j)^2}} \quad (12)$$

this global parameter indicates how much the distance in the current cloud of points is different from the one in the input data matrix, and therefore a direct evaluation of the amount of energy in the system and hence the loss of distance information. The optimization procedure therefore minimizes the amount of lost information during the dimensionality reduction.

Molecular Dynamics Multidimensional Scaling with stochastic force

In (1) a new method for performing Multidimensional Scaling is described. This method is a combination of the Molecular Dynamics based MDS we use here and the method of Simulated Annealing. The latter is supposed to find a global minimum by using stochasticity to pick between probable states. In practice this combination of methods is equivalent to adding a stochastic force to every data point $F_{stochastic}(\bar{X}_i) = -T * s(t)$ where $s(t)$ is a random number given by a generalized Gaussian stochastic distribution, and T is the temperature of the system. By beginning the simulation with a high temperature and decreasing it across the simulation exponentially, one expects to reach the global minimum, as the stochastic force avoids getting trapped in local minima.

In our study we implement this method by adding the stochastic force:

$$F_{stochastic}(\bar{X}_i) = -T * s(t) \quad \textbf{(13)}$$

Where $s(t)$ is a random number generated here uniformly between -0.5 and 0.5. We use two types of temperature-decrease, the first linear, beginning with a temperature of $100J$ and decreasing linearly to $0J$ during 3000 steps of simulation; we call this method MD-MDS linear. The second includes an exponential decrease from $100J$ to below $0.1J$ during 3000 steps of simulation; we call this method MD-MDS exponential.

The *entourage* parameter for local structure analysis

Kruskal stress directly evaluates the distance information deformation. For example if every distance in the new cloud of data points is 10% different from the original distance in the input matrix, one obtains for every i and j:
$(\delta(\bar{X}_i,\bar{X}_j) - d(\bar{X}_i,\bar{X}_j))^2 = (0.1 * d(\bar{X}_i,\bar{X}_j))^2$.
So $e = \sqrt{\frac{\sum_{i\neq j}(0.1*d(\bar{X}_i,\bar{X}_j))^2}{\sum_{i\neq j} d(\bar{X}_i,\bar{X}_j)^2}} = 0.1$. A 10% difference on two large distances will therefore influence the Kruskal stress value more than a 10% difference between two small distances. This phenomenon has been studied by Graef et al. in 1979 (20), by looking at the influence of big distances compared to median and small distances on stress using existing and generated datasets. Hence, Kruskal stress rather evaluates global deformation of the cloud of instances. It does not give any indication on how local distances are affected by the dimensionality reduction. In order to quantify the faithfulness of any representation of data in low-dimensional space it is thus required to define a new parameter. This new parameter, *Entourage*, we chose to define is based on an analysis of the change in k nearest neighbors which evaluates local organization change. SVD leads to a undistorted representation of the instances in rank(X) dimensions, and can be used as reference representation of the data points. Any representation in reduced dimension will have to be the most similar to this reference representation. For any one instance $\bar{X}_i$ in the reference distribution obtained through SVD we consider its k nearest neighbors: N_i^{ref}. In the new distribution obtain after dimensionality reduction, we also compute the k nearest neighbors for the same instance $\bar{X}_i$, and obtain a list: N_i^{new}. We then search for $G_i = card(N_i^{ref} \bigcap N_i^{new})$, which will be the number of instances common to those two lists. We repeat this operation for all instances i, and obtain the *Entourage* parameter:

$$Ent_k = \frac{\sum_{i=1}^{n} G_i}{G} \quad \textbf{(14)}$$

with $G = nk$ a normalization parameter ($Ent \in (0,1)$).

If $G_i = card(N_i^{ref} \bigcap N_i^{new}) \approx 0.01 card(N_i^{ref}) = 0.01k$ for every i then $Ent_k \approx \frac{0.01\sum_{i=1}^{n} k}{nk} = 0.01$, a difference of 1% between two values of *Entourage* mean an average deformation of 1% in the local organization.

This parameter has only signification for a low number of considered neighbors k compared to the total number of points n. We arbitrary choose k=0.1 of n as the number of nearest neighbor instances considered, we obtained a good evaluation on how well the local organization is conserved.

Datasets used in this study

ID	Dataset Name	No. of Instances	No. of Variables
d1	96Cell	96	32878
d2	96Cell_T	96	1553
d3	Iris	150	4
d4	Wine	178	13
d5	Stochast 200	200	50
d6	CCYier	516	12
d7	Pima	768	9
d8	96Cell_T transposed	1553	96
d9	Secom	1567	590
d10	Ozone	2565	72
d11	Stochast 3000	3000	300
d12	Ecoli	4288	7
d13	Wave	5000	22

Table 1. The different datasets used in this study.

To test and illustrate the algorithm discussed here, we have used several publicly available datasets of different origin. First, we have used two different transcriptome datasets. Briefly, the cellular transcriptome is defined as the *ensemble* of RNA molecules resulting from gene expression in a cell. Using microarray technology, in the human case, some thirty-thousand different RNA species can be quantified simultaneously. The dataset here referred to "d1 — 96Cell" includes ninety-six transcriptome measurements generated from thirty-two individual human tissues under non-pathological conditions. This dataset was initially published by (21), and is available for download from:
http://mace.ihes.fr
using accession number: 2914508814. The dataset here called "d6 — CCYier" ((22); mace access. no.: 2960354318), is composed of twelve human fibroblast transcriptome data points generated over twenty-four hours during the cell-cycle. Note that we eliminated one (Interleukin 8, IL8) of the 517 genes as an outlier from this dataset. The dataset "d2 — 96Cell_T" (*c.f. Table 1*), is a derivative of the initial dataset "d1 — 96Cell", where only genes were retained that are specific to one and only one human tissue as provided in (21), and removing again one outlier gene (Probe_ID: 162105). The dataset "d8 — 96Cell_T" (*c.f. Table 1*), is the transposed (Instances, Variables) dataset "d2 — 96Cell_T". All transcriptome datasets were median normalized in log2-space and processed according to standard procedures ((23), (5), (24)). Second, seven additional datasets with no relation to biology were used. Both originate from the Machine Learning Repository (25):
http://archive.ics.uci.edu/ml
(1) "Iris" here "d3 — Iris", (2) "Wine" here "d4 — Wine", (3) "Pima Indians Diabetes" here "d7 — Pima", (4) "SECOM" here "d9 — Secom", (5) "Ozone Level Detection" here "d10 — Ozone", (6) "E. Coli Genes" here "d12 — Ecoli", and (6) "Waveform Database Generator (Version 1)" here: "d13 — Wave". Please refer to the ML repository for details on these data.

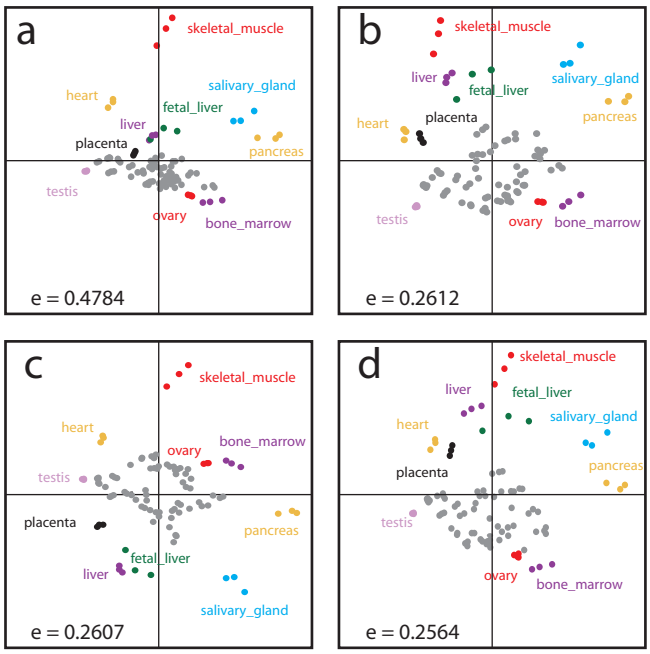


Figure 1. Comparison of the results of different dimensionality reduction techniques on the same dataset. The dataset "d1 — 96Cell", composed of ninety-six individual transcriptome profiles generated from thirty-two different human tissues (*c.f.* Table 1, and section) was represented in 2D space using: (a) Singular Value Decomposition based on covariance, (b) random initialized Multidimensional Scaling; (c) as in *B* using the same algorithm leading to a different random position matrix, and (d) Singular Value Decomposition-initialized Multidimensional Scaling. The peripheral data points were color coded and labeled according to the human tissue analyzed. Only sufficiently well resolved tissues are labeled, all other instances are in gray color. The resulting Kruskal-Stress *e* for each of the dimensionality reductions is indicated. Similar computations were used to generate Table 2.

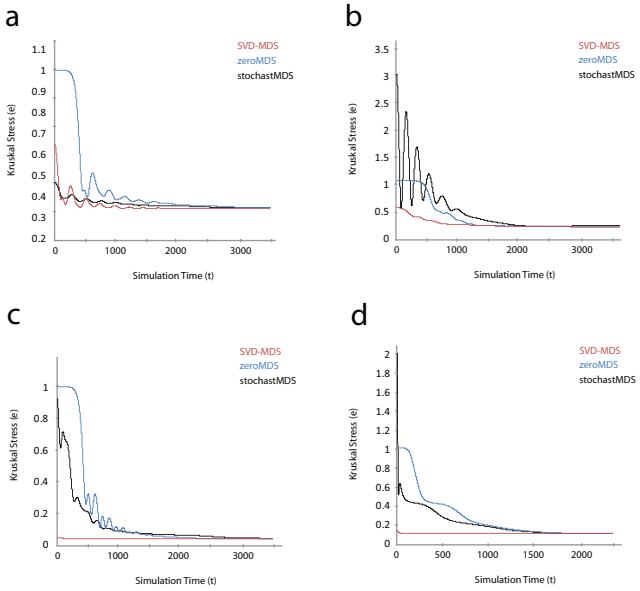


Figure 2. Kruskal Stress (*e*)-evolution over number of simulation iterations (*t*). Comparison of the SVD-MDS, MDS initialized with all points in the center (zeroMDS), and MDS initialized by stochastic positions (stochastMDS) methods on different datasets (a) "d1 — 96Cell" in *correlation basis*, (b) "d2 — 96Cell_T" in *covariance basis*, (c) "d3 — Iris" in *correlation basis*, (d) "d10 — Ozone" in *covariance basis*.

Third, we generate two datasets stochastically:
(i) One with 200 instances and 50 variables between -6 and 6 here "d5 — Stochast 200", (ii) the other with 3000 instances and 300 variables between -6 and 6 here "d11 — Stochast 3000".
The number of instances and the number of variables for all thirteen datasets is given in Table 1.

RESULTS

Comparison of different initialization methods for MDS

We postulated that the inconveniences associated with the combined Molecular Dynamics MDS techniques (hereafter simply: MDS) concerning the problems related to the choice of the initial condition for the simulation leading to insufficient control of being trapped in local minima, as well as the large information loss when SVD techniques are used for dimensionality reduction, can be overcome when both methods are combined. We therefore created an SVD-MDS algorithm which uses SVD to compute the initial state of a molecular dynamics simulated MDS. This SVD-MDS approach was then compared to SVD and MDS on thirteen different datasets (see Table 1). Figure 1 well illustrates the shortcomings of SVD and MDS. The dataset "d1 — 96Cell" (see section for a discussion of the different datasets used in this study) containing ninety-six different instances was used to compute a 2D representation using SVD (panel A), two examples of MDS initialized by random positions defining a 12 unit hypercube (panels B and C), and our combined SVD-MDS approach (panel D). The Kruskal stress was also computed for all four examples. As it is clear from the illustration and the Kruskal stress, MDS techniques (panels B-D) better preserve the distances between the instances and their relationship. The data cloud is better resolved and the global distance information loss (as estimated by the Kruskal stress) is lower than for SVD. Note that we chose to label only those tissues in the illustration that are sufficiently well resolved, all other instances are in gray color.

In order to demonstrate generality of our approach we next analyzed the twelve other datasets (Table 1) using four different approaches: 1. using SVD only, 2. using SVD-MDS, 3. using MDS initialized with all data points placed at zero with minimal random noise (zeroMDS), and 4. MDS initialized with random positions (stochastMDS). The results of those analyses are reported in Table 2. In all cases, we reduced the dimensions to two. It becomes again apparent from the Kruskal stress that the MDS-based techniques systematically outperform the SVD. While stochastMDS, zeroMDS and SVD-MDS give similar results in terms of the final information loss, the number of time-steps needed to identify a minimum stress is greatly reduced using SVD-MDS (Table 2, and for four examples Fig. 2). Therefore, SVD-MDS approaches the final state (here defined as a Kruskal stress value) faster than either of the MDS methods. We show an example of stress evolution in Figure 3a where stochastMDS and zeroMDS are slow due to the existence of local minima,

ID	Dataset Name	Metric	SVD	SVD-MDS		zeroMDS		stochastMDS	
			e	e	t	e	t	e	t
d1	96Cell	R^2	0.6472	0.3409	2500	0.352	2500	0.3478	2500
d2	96Cell_T	Cov	0.5001	0.1401	4500	0.146	4500	0.1503	4500
d3	Iris	Cov	0.0421	0.0344	509	0.0343	3554	0.0344	4059
d4	Wine	Cov	0.0010	0.0010	0	0.0064	4500	0.0061	4500
d5	Stochast 200	Cov	0.7513	0.4088	1500	0.4169	1500	0.4157	1500
d6	CCYier	Cov	0.1634	0.0765	400	0.0932	3500	0.1079	4500
d7	Pima	Cov	0.0964	0.0708	700	0.105	3500	0.1098	3500
d8	96Cell_T transposed	R^2	0.6954	0.1498	4500	0.1572	4500	0.1715	4500
d9	Secom	Cov	0.1801	0.1168	750	0.1217	4499	0.1283	4375
d10	Ozone	Cov	0.1223	0.0935	712	0.0935	2587	0.0951	2143
d11	Stochast 3000	Cov	0.9067	0.4353	130	0.4382	130	0.438	130
d12	Ecoli	Cov	0.1634	0.000	0	0.0202	4500	0.2484	4500
d13	Wave	Cov	0.2922	0.2132	324	0.2132	2252	0.2132	1998

Table 2. Results from the different MDS algorithms applied to the various datasets (*c.f.* Table 1). CoV = covariance, R^2 = correlation, *e* = Kruskal Stress, *t* = time steps for MD simulation.

and SVD-MDS clearly outperform them. Basically, the initial steps in simulation time need not be used to globally arrange the geometric object in space as in MDS, but already contribute to minimizing the Kruskal stress over the entire set of distances.

Iterative dimensionality reduction using iSVD-MDS

We next wondered whether the dimensionality reduction could be further improved by a step-wise reduction of one component after another. To this end we compared again the performance of the three techniques SVD-MDS, MDS, and iterative SVD-MDS (iSVD-MDS) on the different datasets. In iSVD-MDS, for each successive round of reducing the dimensionality of the geometric object by one, a SVD followed by a subsequent molecular dynamics MDS is performed. As can be seen in Figure 3a, SVD-MDS rapidly approaches a minimal Kruskal stress configuration over the simulation time. The previously described MDS procedure which uses stochastic initiation for the molecular dynamics simulation requires much more simulation time to find the same minimal stress configuration as the SVD-MDS algorithm. Finally, the iterative iSVD-MDS approach will also converge to the identical minimum obtained through the other methods, however, as for each component a separate simulation is performed the convergence time is greatly increased when compared to the former two methods. Albeit many different simulations on the different datasets we have never obtained a final configuration using iSVD-MDS were the Kruskal stress would allow to conclude on an improved performance when compared to SVD-MDS. Therefore, the iterative method does not allow for improved accuracy, but rather prolongs simulation time with no immediate gain (Table 3 summarizes the results). We next compared iSVD and iSVD-MDS methods to determine how the loss of information is distributed during iterative dimensionality reduction. As can be seen in Figure 3b, and in accordance with theoretical considerations (see Background section), for both procedures the amount of stress or lost information increases both relatively and absolutely with the number of components removed. Note also, that the iSVD-MDS method better preserves at every consecutive iteration the distance information of the object (Figure 3b).

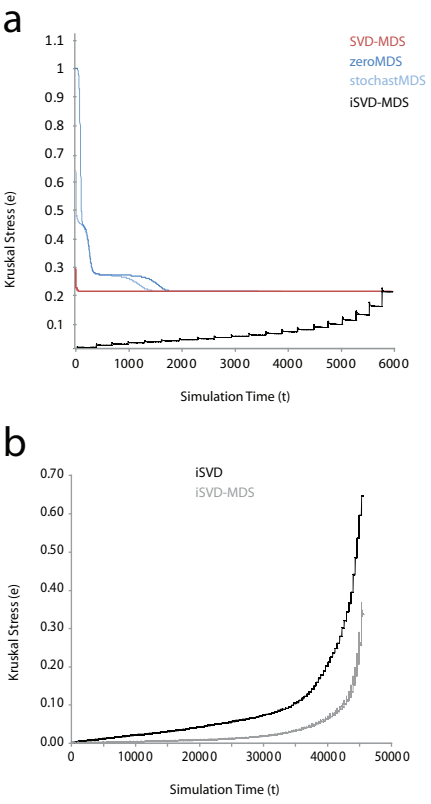


Figure 3. Iterative SVD-MDS. (a) Comparison of the SVD-MDS, zeroMDS, stochastMDS, and iterative SVD-MDS (iSVD-MDS) methods on dataset "d13 — Wave" in *covariance basis*. (b) Comparison of the iterative SVD (iSVD) and iSVD-MDS methods on dataset "d1 — 96Cell" in *correlation basis*.

ID	Dataset Name	Metric	SVD-MDS		iMDS	
			e	t	e	t
d1	96Cell	R^2	0.3409	2500	0.3381	232097
d2	96Cell_T	Cov	0.1401	4500	0.1494	92536
d3	Iris	Cov	0.0344	509	0.0344	3008
d4	Wine	Cov	0.0010	0	9.0E-4	10003
d6	CCYier	Cov	0.0765	400	0.0753	22508
d7	Pima	Cov	0.0708	700	0.0692	27005
d8	96Cell_T transposed	R^2	0.1498	4500	0.1525	122059
d10	Ozone	Cov	0.0935	712	0.0935	66031

Table 3. Results from iSVD-MDS algorithms and SVD-MDS algorithm applied to the various datasets (*c.f.* Table 1). CoV = covariance, R^2 = correlation, e = Kruskal Stress, t = time steps for MD simulation.

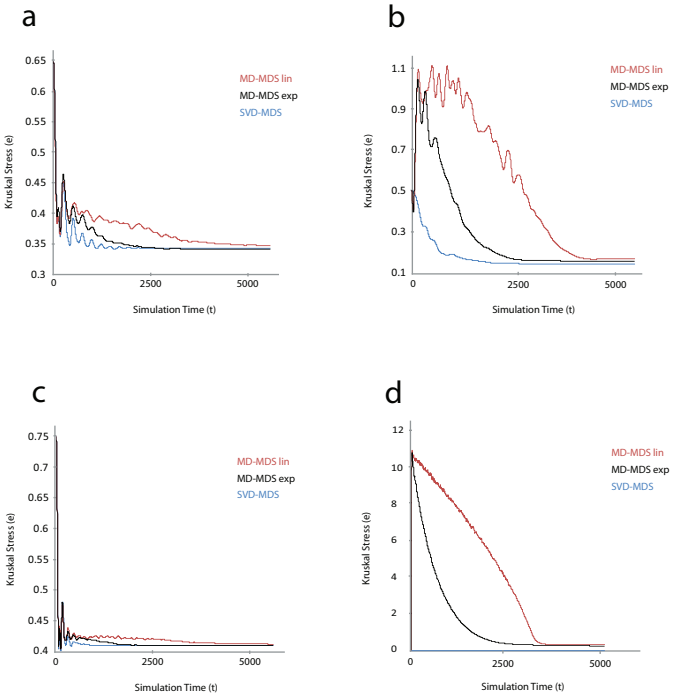


Figure 4. The influence of stochasticity on MD-MDS algorithms. Comparison of the SVD-MDS, Molecular Dynamics linear MDS (MD-MDS linear), Molecular Dynamics exponential (MD-MDS exponential) methods on different datasets (a) "d1 — 96Cell" in *correlation basis*, (b) "d2 — 96Cell_T" in *covariance basis*, (c) "d5 — Stochast 200" in *covariance basis*, (d) "d12 — Ecoli" in *covariance basis*.

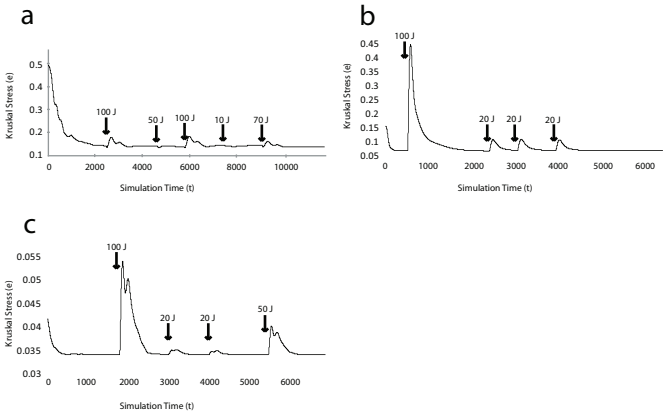


Figure 5. Robustness of SVD-MDS simulations. Evolution of stress over number of simulation iterations with injection of energy, on different datasets (a) "d2 — 96Cell_T" in *covariance basis*, (b) "d2 — CCYier" in *covariance basis*, (c) "d5 — Iris" in *covariance basis*.

Molecular Dynamics dimensionality reduction with added stochasticity

In (1) an approach reminiscent of simulated annealing has been used to avoid getting trapped in local minima during the molecular dynamics simulation. Adding stochasticity to the molecular dynamics driven MDS is, after (1), required to insure reproducibility of the algorithmic performance. One can prove theoretically that by adding stochasticity convergence will improve, however, in practice this addition seems irrelevant as it will not lead to a different result. To compare MD-MDS with our SVD-MDS algorithm we have implemented different MD-MDS algorithms with stochastic energy. We thereby chose to linearly ("lin") and exponentially ("exp") (as in (1)) remove this extra energy from the system over simulation time. As can be seen in Figure 4, SVD-MDS as well as the two MD-MDS algorithms "lin" and "exp" always identify final configurations with the same amount of residual energy. It can also be seen that SVD-MDS converges faster for these four examples than the MD-MDS methods. In Table 4 we show that both statements hold for the entire set of analyzed data.

We next asked whether or not similarly adding stochasticity to the SVD-MDS algorithm would improve its performance. Figure 5 illustrates the results we have obtained on three different datasets. Indeed, adding different amounts of energy at different times of the simulation (as indicated in the panels by arrows) does not lead to the "discovery" of lower energy minima during the simulation procedure. The SVD-MDS algorithm, similarly as the MD-MDS algorithms (Figure 4) always converges to the same energy state. This has been confirmed using other datasets with identical results (data not shown). Taken together, the results using MD-MDS-lin and MD-MDS-exp and SVD-MDS raise the question of whether indeed several minima exist or only a single ground-state is to be found. While we do not have any formal proof of the latter, we believe that the detailed analysis of the geometric structure of the data objects presented below strongly argues in favor of a global energy minimum.

ID	Dataset Name	Metric	SVD-MDS		MDMDSlinear		MDMDSexpo	
			e	t	e	t	e	t
d1	96Cell	R^2	0.3409	2500	0.3453	5500	0.3421	2500
d2	96Cell_T	Cov	0.1401	4500	0.1465	4500	0.1542	4500
d3	Iris	Cov	0.0344	509	0.0359	4500	0.0343	4000
d4	Wine	Cov	0.0010	0	0.0089	4500	0.0067	4500
d5	Stochast 200	Cov	0.4088	1500	0.4092	4500	0.4089	4500
d6	CCYier	Cov	0.0765	400	0.1346	5500	0.1162	5500
d7	Pima	Cov	0.0708	700	0.1128	5500	0.0986	5500
d8	96Cell_T transposed	R^2	0.1498	4500	0.1832	4224	0.1822	4500
d9	Secom	Cov	0.1168	750	0.1511	5500	0.1396	4500
d10	Ozone	Cov	0.0935	712	0.0944	4500	0.0951	3500
d11	Stochast 3000	Cov	0.4353	130	0.4353	200	0.4353	200
d12	Ecoli	Cov	0.0	0	0.312	5500	0.2273	5500
d13	Wave	Cov	0.2132	324	0.2132	3671	0.2132	2203

Table 4. Results from the different MD-MDS algorithms and SVD-MDS algorithm applied to the various datasets (*c.f.* Table 1). CoV = covariance, R^2 = correlation, *e* = Kruskal Stress, *t* = time steps for MD simulation.

DISCUSSION

Geometric structure

While computationally unattractive, and without any apparent bearing on the final result of the dimensionality reduction, the iterative iSVD-MDS method allows the analysis of the structural changes of the geometric object during dimensionality reduction. More precisely, and as seen in Figure 3, the absolute contribution of every single dimension can be evaluated. This might be of particular interest in the case of biological data, as the major distance information often is not necessarily restricted to a few dimensions. The iterative SVD-MDS method in conjunction with the *Entourage* parameter hence gives the user a potential control on when to stop the dimension reduction process. Based on the conjunction of iSVD-MDS and local structural control it might even be feasible to develop quantitative methods to define maximal compressibility at some defined distance information loss.

A second type of analysis of the geometric properties of the objects under study can be developed by analyzing the behavior of the *Entourage* parameter, defined as the relative change in the k nearest neighbors, as a function of the k nearest neighbors considered. We have plotted the relationship of *Entourage* and k for six different methodologies: zeroMDS, stochastMDS, SVD-MDS, iSVDMDS, MD-MDS-lin, MD-MDS-exp in (Figure 6) for nine different datasets which represent the different behaviors one can observe. From the selected examples it becomes clear that again the SVD-MDS method outperforms the different types of MDS over a wide array of structures analyzed as the *Entourage* value is consistently higher no matter how many different k nearest neighbors are considered. The iterative iSVD-MDS method, due to the accumulation of small residual errors during the molecular dynamics simulation, and the MDS method give similar results. At the cost of increasing computational load, the iSVD-MDS better and better approximates the SVD-MDS method. In conclusion, the SVD-MDS method, under all conditions tested, better represents the geometric structure of the datasets in low-dimensional space when compared to the input object with $rank(S)$ components. Note that this holds even for objects with equal stress.

Figure 1 illustrates the problem of rotational variance when using stochastically initiated molecular dynamics simulations for MDS. It becomes apparent, when comparing panels B and C as well as comparing them to panel A and D that stochastMDS can result in different final configurations. The stochastMDS algorithm produces two near-optimal solutions (with respect to the Kruskal stress), the resulting orientation of the instances, however, is different (focus for instance on the relationship between "skeletal muscle" and "fetal liver"). The problem arising, if stochastMDS can lead to different representations despite using the same parameters for computation, is accuracy of the representation. Strikingly, SVD-MDS on the contrary only produces a single result. This observation, taken together with the results on the relevance of stochasticity in the simulation obtained above, argues for the existence of different equivalent energy minima that only differ in the rotational orientation of the object and at best only minimally in the Kruskal-stress. Therefore, taken together SVD-MDS not only reduces significantly the computational load when compared to other methods, but also insures uniqueness of the resulting representation as there is no randomness in the process unlike in the other methods used for initialization. The quality of this final and unique representation can be demonstrated using the *Entourage* parameter, as shown here. This increase in fidelity in the representation of data should not be underestimated (see Figure 1). Stress values should not be considered as the only relevant parameter to determine the performance of dimensionality reduction techniques as local and global structure considerations can effectively, as demonstrated here, allow to judge fidelity of the final representation. This is reminiscent to techniques of principal manifold searches (26) were parameters describing topology, local organization or other geometric characteristics are used.

A major advantage of using SVD to define the initial state is that it provides the inertia of each principal component (Figure 7). As can be appreciated from this comparison of four different datasets the internal structure, as judged from the distribution of information over dimensions, of the

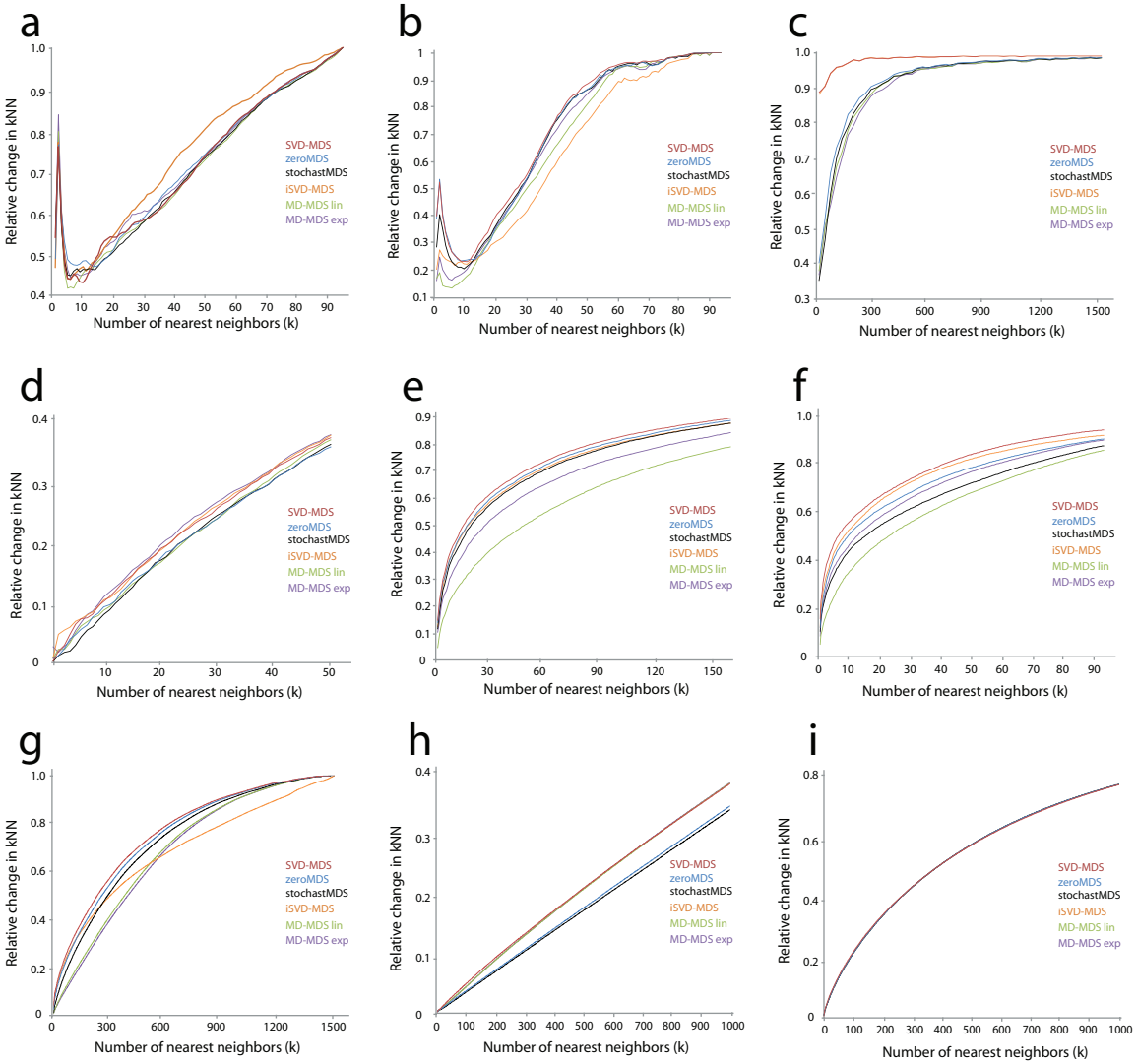


Figure 6. Relative changes in k nearest neighbors (*Entourage*) are local, structural measures of dimensionality reduction and thus assess quality of the procedure. As a function of the number of nearest neighbors k considered, the relative change in kNN between the initial high-dimensional space and 2D space is plotted for all the methods SVD-MDS, zeroMDS, stochastMDS, iSVD-MDS, MD-MDS linear, and MD-MDS exponential. The datasets used are in: (a) "d1 — 96Cell" in *correlation basis*, (b) "d2 — 96Cell_T" in *covariance basis*, (c) "d4 — Wine" in *covariance basis*, (d) "d5 — Stochast 200" in *covariance basis*, (e) "d6 — CCYier" in *covariance basis*, (f) "d7 — Pima" in *covariance basis*, (g) "d8 — 96Cell_T transposed" in *correlation basis*, (h) "d11 — Stochast 3000" in *covariance basis*, (i) "d13 — Wave" in *covariance basis*.

different datasets is quite different. Especially, biological data such as functional genomics data show a variety of different distributions of information (11, 12, 14, 21, 22). A good dimensionality reduction technique would ideally account for these differences. Taking into account the inertia, the stress and the *Entourage* during the MDS process will help to have an even more accurate representation of the data matrix in low dimensional space.

CONCLUSION

Dimensionality reduction of complex high-dimensional data is an important problem which becomes ever more complicated due to the increase of data concomitant with an increase in their dimensionality. This is particularly true for data from modern genomics analyses where more and more often data with thousands of instances each over millions of variables are generated. Different approaches for visualization of such data have been previously chosen. All those methods suffer from at least two shortcomings, (i) unfaithful representation, and / or (ii) computational inefficiency. We demonstrate here how a combined molecular dynamics simulation multi-dimensional scaling approach for dimensionality reduction of high-dimensional data can be improved by better defining the initial conditions for the molecular dynamics simulation. By defining an analytical parameter to study the geometric object during the dimensionality-reduction process we have shown the singular value decomposition is most effective to create an initial condition for the MD simulation based MDS. Using links between SVD and different standard data analysis methods, we demonstrate how our combined SVD-MDS method can be used to

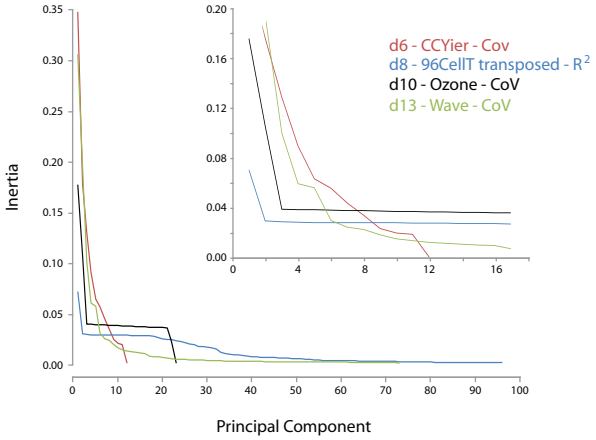


Figure 7. Inertia distributions over principal components lead to an appreciation of the structure of the geometric object under study. The inertia of each of the principal components were plotted for the four different datasets "d6 — CCYier_T" in *covariance basis*, "d8 — 96Cell_T transposed_T" in *correlation basis*, "d10 — Ozone" in *covariance basis*, and "d13 — Wave" in *covariance basis*. The inset is a zoom of the same graph on the initial eighteen principal components.

improve geometric representation in low dimensional space that are generally obtained with standard analysis methods (PCA, Correlation analysis, Correspondence analysis). We also show that the use of stochastic energy during the simulation process does not increase performance of the algorithms in terms of finding a optimal solution. Furthermore, the SVD-MDS approach developed here was shown to be computationally more efficient than other approaches. Finally, we have investigated different measures to better analyze and control the dimensionality reduction process. Overall, the methodology developed here should further advance our capacity to analyze high-dimensional data such as the ones produced by functional genomics approaches.

ACKNOWLEDGEMENTS

This work was funded by the *Centre National de la Recherche Scientifique* (C.N.R.S.), the *Agence Nationale pour la Recherche contre le SIDA et les hépatites virales* (A.N.R.S.), the *Agence Nationale pour la Recherche* (A.N.R., ISPA project), and the *Genopole Evry*.
CB is recipient of a Ph.D. fellowship from the A.N.R.S..

Conflict of interest statement. None declared.

REFERENCES

1. Andrecut M. Molecular dynamics multidimensional scaling. *Physics Letters A*. 2009;**373**:(23-24):2001–2006.
2. Holmes S. Visualising Data. *Statistical Problems in Particle Physics, Astrophysics and Cosmology*. 2006;*197*.
3. Berthold M, Hand D. *Intelligent Data Analysis*. Springer. 2003;2^d edition.
4. Noth S, Brysbaert G, Pellay F, Benecke A. High-sensitivity transcriptome data structure and implications for analysis and biologic interpretation. *Genomics, Proteomics & Bioinformatics*. 2006:**4**:212–229.
5. Benecke A. Chromatin code, local non-equilibrium dynamics, and the emergence of transcription regulatory programs. *The European Physical Journal E: Soft Matter and Biological Physics*. 2006;**19**:(3):353–366.
6. Schmidt E, Stewart G. On the Early History of the Singular Value Decomposition. *University Of Maryland*. 1992.
7. Troyanskaya O, Cantor M, Sherlock G, Brown P, Hastie T, Tibshirani R, Botstein D, Altman R. Missing value estimation methods for DNA microarrays. *Bioinformatics*. 2001;**17**:(6):520–525.
8. Candes E, Recht B. Exact Matrix Completion via Convex Optimization. *ArXiv*. 2008.
9. Brock G, Shaffer J, Blakesley R, Lotz M, Tseng G. Which missing value imputation method to use in expression profiles: a comparative study and two selection schemes. *BMC Bioinformatics*. 2008;**9**:12.
10. Eckart C, Young G. The approximation of one matrix by another of lower rank. *Psychometrika* 1936;**1**(3):211–218.
11. Alter O, Brown P, Botstein D. Singular value decomposition for genome-wide expression data processing and modeling. *Proc. Natl Acad. Sci. USA.*. 2000;**97**:(18):10101–10106.
12. Alter O, Brown P, Botstein D. Generalized singular value decomposition for comparative analysis of genome-scale expression data sets of two different organisms. *Proc. Natl Acad. Sci. USA.*. 2003;**100**:(6):3351–3356.
13. Wall M, Rechtsteiner A, Rocha L. Singular Value Decomposition and Principal Component Analysis. *A Practical Approach to Microarray Data Analysis*, Springer US 2003:91–109.
14. Fellenberg K, Hauser N, Brors B, Neutzner A, Hoheisel J, Vingron M. Correspondence analysis applied to microarray data. *Proc. Natl Acad. Sci. USA.*. 2001;**98**:(19):10781.
15. Ding C, He X. K-means clustering via principal component analysis. *Proceedings of the 21 st International Conference on Machine Learning*. ACM Press 2004;225–232.
16. Torgerson W. Multidimensional scaling: I. Theory and method. *Psychometrika*. 1952;**17**:(4):401–419.
17. Cuadras C, Fortiana J. Metric Scaling Graphical Representation of Categorical Data. *Penn State University*. 1995.
18. Kruskal J, Wish M. *Multidimensional Scaling*. SAGE Publications Inc. 1978.
19. Cox T, Cox M. *Multidimensional Scaling, Second Edition*. Chapman & Hall/CRC. 2000;2^d edition.
20. Graef J, Spence I. Using distance information in the design of large multidimensional scaling experiments. *Psychological Bulletin*. 1979;**86**:60–66.
21. Dezso Z, Nikolsky Y, Sviridov E, Shi W, Serebriyskaya T, Dosymbekov D, Bugrim A, Rakhmatulin E, Brennan R, Guryanov A, Li K, Blake J, Samaha R, Nikolskaya T. A comprehensive functional analysis of tissue specificity of human gene expression. *BMC Biology*. 2008;**6**:49.
22. Iyer V, Eisen M, Ross D, Schuler G, Moore T, Lee J, Trent J, Staudt L, Hudson J, Boguski M, Lashkari D, Shalon D, Botstein D, Brown P. The Transcriptional Program in the Response of Human Fibroblasts to Serum. *Science*. 1999;**283**(5398):83–87.
23. Benecke A. Genomic Plasticity and Information Processing by Transcription Coregulators. *Complexus*. 2003;**1**:(2):65–76.
24. Benecke A. Gene regulatory network inference using out of equilibrium statistical mechanics. *HFSP Journal*. 2008;**2**:(4):183–188.
25. Asuncion A, Newman D. *UCI Machine Learning Repository*. 2007.
26. Gorban N, Balzs K, Wunsch C, Zinovyev A. *Principal Manifolds for Data Visualization and Dimension Reduction*. Springer Publishing Company. 2007.

E

Journal article : Transcription within Condensed Chromatin : Steric Hindrance Facilitates Elongation.

Christophe Bécavin, Maria Barbi, Jean-Marc Victor, Annick Lesne.
Published in Biophysical Journal, March 2010, Volume 98, 824-833.

Abstract

During eukaryotic transcription, RNA-polymerase activity generates torsional stress in DNA, having a negative impact on the elongation process. Using our previous studies of chromatin fiber structure and conformational transitions, we suggest that this torsional stress can be alleviated, thanks to a tradeoff between the fiber twist and nucleosome conformational transitions into an activated state named "reversome". Our model enlightens the origin of polymerase pauses, and leads to the counterintuitive conclusion that chromatin-organized compaction might facilitate polymerase progression. Indeed, in a compact and well-structured chromatin loop, steric hindrance between nucleosomes enforces sequential transitions, thus ensuring that the polymerase always meets a permissive nucleosomal state.

Transcription within Condensed Chromatin: Steric Hindrance Facilitates Elongation

Christophe Bécavin,[†‡] Maria Barbi,[§] Jean-Marc Victor,[§*] and Annick Lesne[†§]
[†]Institut des Hautes Études Scientifiques, Bures-sur-Yvette, France; [‡]Institut de Recherche Interdisciplinaire, Centre National de la Recherche Scientifique, USR 3078, Universités Lille I and II, Villeneuve d'Ascq, France; and [§]Laboratoire de Physique Théorique de la Matière Condensée, Centre National de la Recherche Scientifique, UMR 7600, Université Pierre et Marie Curie, Paris, France

ABSTRACT During eukaryotic transcription, RNA-polymerase activity generates torsional stress in DNA, having a negative impact on the elongation process. Using our previous studies of chromatin fiber structure and conformational transitions, we suggest that this torsional stress can be alleviated, thanks to a tradeoff between the fiber twist and nucleosome conformational transitions into an activated state named "reversome". Our model enlightens the origin of polymerase pauses, and leads to the counterintuitive conclusion that chromatin-organized compaction might facilitate polymerase progression. Indeed, in a compact and well-structured chromatin loop, steric hindrance between nucleosomes enforces sequential transitions, thus ensuring that the polymerase always meets a permissive nucleosomal state.

INTRODUCTION

Transcription is a fundamental biological process during which a dedicated protein, the RNA-polymerase (RNAP), achieves the synthesis of a RNA stretch from a genomic DNA template. It can be divided into three phases: initiation, elongation, and termination. Initiation provides RNAP with an access to the promoter sequence. In eukaryotic cells, this requires the assembly of transcription factors together with RNAP into the transcription initiation complex. We here do not address problems related to the initiation phase, which are by far the most complex ones, inasmuch as they are involved at the heart of the transcriptional regulation. We focus on the elongation phase, which starts once the elongation complex has been completed, and progresses until a termination sequence is encountered. The elongation complex consists of a denaturation bubble of length ~10 nucleotides, enclosed within RNAP (1). During elongation, RNAP tracks along the genomic sequence, swallowing the DNA double helix.

However, in eukaryotic species, genomic DNA is wrapped around octamers of histone proteins, forming nucleosomes in turn organized at a higher level into a chromatin fiber. This complex architecture is bound to hinder both the initiation and the elongation phases (2). In the standard paradigm, transcription elongation requires a decondensed state of chromatin to take place. This is questionable for at least two reasons:

1. In vivo, chromatin decondensation remains elusive, all the more because chromosome structure is not yet elucidated. Whereas it is generally assumed that the fiber itself is decondensed in regions that have to be transcribed, the fiber structure has never been resolved, neither in condensed nor in decondensed chromatin—and we do not even know whether there is any difference between both structures.
2. In vitro, even in decondensed fibers, nucleosomes constitute nearly absolute obstacles to RNAP progression (3).

We wish to examine here whether elongation could take place within a condensed chromatin fiber, and if so, according to which scenario.

BIOLOGICAL SETTING

Our approach is based on a modeling study of the interplay between conformational dynamics of the chromatin fiber and RNAP processing along the fiber (4). We recall here the main biological features of eukaryotic transcription, focusing on recent biophysical results.

RNAP or DNA: which is moving?

There are three types of RNAP according to the type of RNA they synthesize. RNAP I is dedicated to ribosomal RNA synthesis and occurs in a particular environment—the nucleolus—probably devoid of nucleosomes because of its very high transcription rate. RNAP II transcribes RNA encoding proteins. The corresponding transcripts are much longer than the transcripts delivered by RNAP III, i.e., tRNAs and other small RNAs. Entanglement problems are therefore much more stringent for RNAP II than for RNAP III. As a matter of fact, RNAP progression along the genomic sequence requires a relative rotation of the RNAP together with its transcript around the DNA. Then there are two possibilities: either the DNA is kept fixed and the RNAP turns around it, thus following the DNA helical groove and producing a RNA strand coiled around the DNA double helix; or the RNAP is kept fixed and the DNA double helix has to screw inside it. In the first case, long RNA transcripts would have difficulty getting untangled and their further

Submitted August 30, 2009, and accepted for publication October 29, 2009.

*Correspondence: victor@lptmc.jussieu.fr

Editor: Laura Finzi.

0006-3495/10/03/0824/10 $2.00

doi: 10.1016/j.bpj.2009.10.054

migration would thus be impeded. That is why we favor the second case, where the RNAP is jammed into some nuclear structure (e.g., transcription factory (5)).

The twin-supercoiled-domain (TSD) model

The above assumption implies in turn a topological problem because eukaryotic transcription occurs within chromatin loops, i.e., genomic segments ~50–200 kilobases long, that partition chromatin into functionally independent domains (6); the loop ends are clamped by insulator elements (7), not necessarily tightly tethered to a matrix but enough constrained to make each loop a topologically insulated domain that traps DNA supercoiling and ensures the conservation of the linking number in the loop. We recall that the linking number is roughly the number of times one DNA strand is coiled around the other one (8). This topological quantity is conserved in the absence of topoisomerase activity, or before the topoisomerases act efficiently (see below). As the elongation complex progresses along the genomic sequence, the DNA double helix in front of it becomes overwound (positively supercoiled) whereas the DNA behind it becomes underwound (negatively supercoiled). This is the so-called twin-supercoiled-domain (TSD) model, first introduced by Liu and Wang (9) and extensively acknowledged since (for a review, see (10)).

Nucleosome conformations in a transcribing loop

The TSD model has been shown to be potentially relevant for eukaryotes as well (11,12). More recently Matsumoto and Hirose directly visualized (by fluorescence imaging) transcription-coupled negative supercoiling in chromatin even in the presence of active topoisomerases (13), thus strongly supporting the model. However, what kind of structural rearrangement of the chromatin loop should occur jointly with the absorption of positive (respectively, negative) supercoiling downstream (respectively, upstream)? We recently revisited the TSD model in the chromatin context by means of a single chromatin fiber nanomanipulation by magnetic tweezers and we proposed that nucleosomes may act as a topological buffer. This feature relies on the existence of three stable nucleosome states evidenced by the nanomanipulation, namely: *N* (negatively crossed), *O* (open), and *P* (positively crossed), according to the relative position and orientation of the linkers, one with respect to the other (14). In higher eukaryotes, linker histones H1/H5 presumably play a role both in stabilizing the states *N* and *P* against *O*, and channeling the transition in between them by acting as a pivot (15,16).

The reversome hypothesis

A convergent set of experimental observations (17–19) tends to indicate that RNAP II can transcribe through a nucleosome only if the nucleosome is in an activated conformation.

Using the same setup as in Bancaud et al. (14), we found that a fiber submitted to a large positive torsional stress can trap positive turns at a rate of two turns per nucleosome (20). This trapping has been shown to reflect a nucleosome chiral transition to a metastable state, called ''reversome'' (alternatively by the name of R-octasome (21)). This new state has been claimed to be a good candidate for the required activated conformation. Indeed, the transition to reversome is accompanied by the undocking of both H2A-H2B dimers from the $(H3\text{-}H4)_2$ tetramer (22) that relieves the hindrance to RNAP progression. The free-energy landscape of a nucleosome under physiological ionic conditions is schematically represented in Fig. 1. It presents three minima *N*, *P*, and *R*, corresponding respectively to the negative, positive, and reversome states, with $F_R > F_P \approx F_N$, and a maximum *B* corresponding to the top of the barrier encountered during the transition between states *P* and *R*, with a corresponding free energy F_B.

The less stable structure of the reversome arguably facilitates the RNAP progression through the reversome particles during transcription. Moreover, this auxiliary transcriptional mechanism avoids the need for a complete disassembling of the nucleosome into single histones, hence epigenetic marks can be preserved.

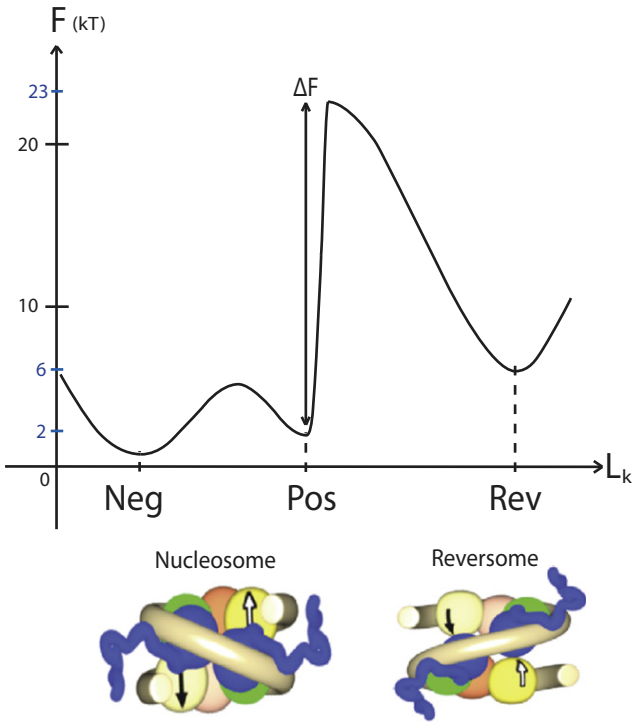


FIGURE 1 (*Color online*) Free-energy landscape for the nucleosome conformation. The reaction coordinate (abscissa Lk) is the linking number of nucleosomal DNA; this choice appears relevant to investigate the landscape changes when a torque Γ is applied to the DNA (see subsection Linking Number Conservation: Accounting for Mechanical Constraints). The two main states are sketched here: the current nucleosome, with two substates *N* and *P* according to the relative positions of the linkers (negative or positive crossing); and an activated state, the reversome, in which the histone core partially unfolds and the nucleosomal DNA adopts a right-handed path around the histone core. (Courtesy of Hua Wong.)

The fiber structure in a transcribing loop

After more than 30 years of effort, the structure of the chromatin fiber is still a matter of debate, both in vitro (where the path of the linker DNA remains elusive) and in vivo (where it is expected to vary considerably according to the cell cycle period and functional status of the fiber)—with possibly several different structures coexisting along the chromosome (23). The fiber structure is no better assessed in a transcribing loop.

Because we focus here on the transcription elongation within a condensed fiber, we favor regular fiber structures. These are indeed energetically favored by stacking interactions between nucleosomes and possibly functionally too, hence selected during (spontaneous) self-organization or (active) remodeling of the fiber. It has been shown in vitro that a small amount of nucleosome positioning is enough to get a regular structure (24). Accordingly, we shall consider as the generic setting the regular model structure of chromatin fiber established in a previous work (25), presenting a strong nucleosome stacking, hence strong steric hindrance (see Fig. 2).

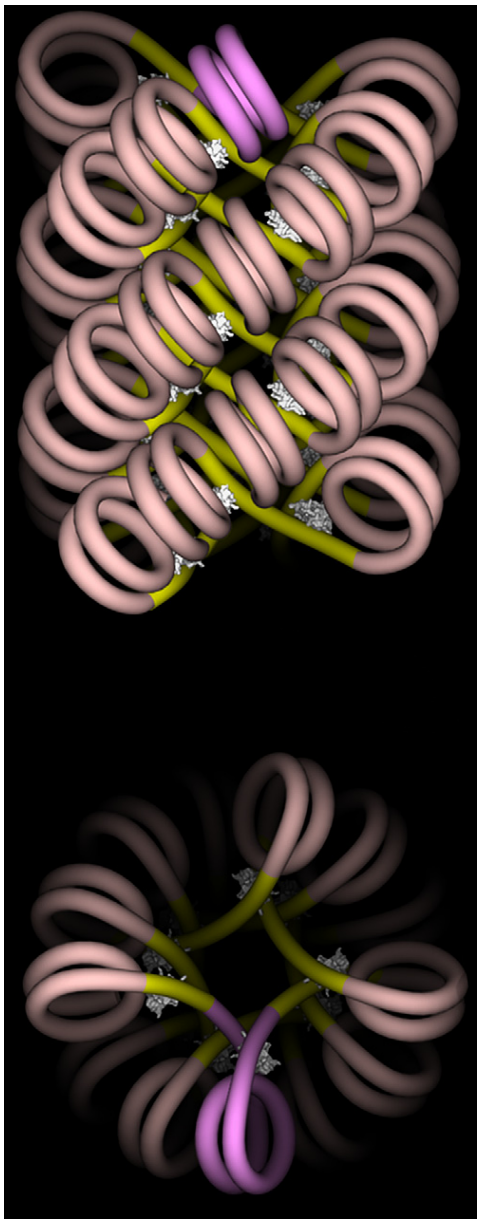


FIGURE 2 (*Color online*) The n-start fiber structure (with $n = 4$, corresponding to a repeat length of $n_{\text{repeat}} = 87$ bps (25)). Note the close and regular nucleosome stacking along each start, preventing the transition to reversome of a single nucleosome, and instead enforcing a concerted sequential transition. (Courtesy of Julien Mozziconacci.)

OUR MODELING FRAMEWORK

Let us sum up the biophysical bases of our model of transcription elongation in a chromatin loop:

1. RNAP is jammed into some nuclear structure and exerts a torque inducing the rotation of DNA on itself that can be estimated from experimental data to occur at a constant rate of $\omega_0 \approx 4\pi$ rad/s (two turns per second) (26), which provides a first boundary condition in our model.
2. The DNA is turning inside RNAP, inducing positive (respectively, negative) supercoiling in the downstream (respectively, upstream) part of the loop.
3. The chromatin fiber within the loop is assumed to be condensed enough to ensure nucleosome stacking.
4. Given that the average linking number of chromosomes in vivo has been evaluated to ~−1 (16), we assume that the starting nucleosome state in the fiber is an appropriate mix A of positive and negative nucleosome states, of linking number $\mathbf{Lk}^{\mathrm{A}} = -1$.
5. The positive torque exerted by RNAP on the loop downstream may induce the transition of nucleosomes into reversomes.

We shall adopt a continuous medium modeling of the fiber as a homogeneous elastic rod (27,28) and evaluate the role of chromatin fiber rigidity, the transmission of the torque exerted by RNAP along the fiber and the dissipation in the surrounding viscous medium. This continuous description is supported by the high and regular nucleosome density Λ along the fiber, varying between 0.5 and 1 nm^{-1}. To be valid, this framework mainly requires the description of the fiber behavior at the level of a few nucleosomes, with an elementary length dX along the fiber axis such that $1/\Lambda \ll dX \ll \mathrm{N}/\Lambda$, with N the number of nucleosomes per chromatin loop. This amounts to smoothing out single-nucleosome inhomogeneities and describing the average fiber behavior at a suprananometer scale. In this setting, the local state of the fiber is described by means of one or more continuous deterministic fields as, for instance, the local fraction of reversomes at point x and time t, denoted below $\xi(x, t)$ (29). We will switch to a discrete description in the section Transition Kinetics and Critical Torque to take into account the reversome transition

TABLE 1 Biological setting

Entity	Parameter	Typical value	Definition
RNAP	ω_0	4π rad/s	Angular velocity induced by RNAP to DNA (or equivalently to the fiber).
	V	20 bps/s	RNAP velocity, i.e., number of transcribed bps per second.
Loop	N	250	Number of nucleosomes per chromatin loop.
	L	250–500 nm	Loop length.
	l_0	100–500 nm	Length of the loop region downstream of the initiation site.
Fiber	Λ	0.5–1 nm^{-1}	Linear density of nucleosomes in a fiber.
	L_p	30–300 nm	Persistence length of the fiber.
	R	15 nm	Fiber radius.
DNA	l_{pitch}	3.4 nm	Pitch of the DNA-double helix in B-form.
	n_{pitch}	10.5 bps	Number of base pairs corresponding to the pitch.
Nucleosome	n_{repeat}	200 bps	Nucleosome repeat length, i.e., number of bps per nucleosome.
	l_{repeat}	70 nm	Length of DNA per nucleosome $l_{repeat} = n_{repeat}.l_{pitch}$.
	$\mathbf{Lk}^N$	−1.4	Linking number (per nucleosome) of the negative N state.
	$\mathbf{Lk}^P$	−0.4	Linking number (per nucleosome) of the positive P state.
	$\mathbf{Lk}^A$	−1.0	Linking number (per nucleosome) of the average A state in condensed fibers.
	$\mathbf{Lk}^B$	−0.25 to 0	Linking number at the barrier B position.*
	$\mathbf{Lk}^R$	1.0	Linking number (per nucleosome) of the reversome R state.
Energy and kinetics	F_N	0.7 kT	Free energy of the P state.
	F_P	2 kT	Free energy of the P state.
	F_B	23 kT	Free energy of the barrier between P and R.
	F_R	6 kT	Free energy of the R state.
	k_0	3 10^6 s^{-1}	Preexponential factor for spontaneous fluctuation between P and R states.

Summary of the notations and typical values of the parameters. The main parameters for the different states of the nucleosome have been obtained in Bancaud et al. (20).

*The value 0 is that used in Bancaud et al. (20). The value −0.25 is obtained by fitting the experimental hysteresis curves of Bancaud et al. (20) with a kinetic model similar to the one described in Appendix S1 in the Supporting Material (H. Wong, J. Mozziconacci, M. Barbi, J. M. Victor, unpublished results).

kinetics and to evaluate the torque exerted by RNAP on the fiber.

Typical values of the relevant parameters of the model are summed up in Table 1.

PRELIMINARY INVESTIGATIONS: SEVERAL TIME- AND SPACE SCALES

Kinematic notations

We shall denote X the arc-length (curvilinear abscissa) measured along the chromatin fiber denoting the position of the RNAP with respect to the transcription initiation site (TIS). If the relative DNA-RNAP angular velocity is $\omega_0 \approx 4\pi$ rad/s, then the distance $X(t)$ traveled by the RNAP measured along the fiber is

$$X(t) = Vt \quad \text{with} \quad V = \frac{\omega_0\, l_{pitch}}{2\pi\, \Lambda\, l_{repeat}} \approx 10 \text{ nm/s}, \qquad (1)$$

where l_{pitch} is the pitch of the DNA double helix, Λ the number of nucleosomes per nm along the fiber, and l_{repeat} the repeat length, i.e., the DNA length per nucleosome. A length ΔX along the chromatin fiber corresponds to a length $\Delta X\, \Lambda\, l_{repeat}$ along the embedded DNA. The length of the chromatin loop downstream of the RNAP is $l(t) = l_0 - X(t)$, with l_0 the length of the loop region downstream of the TIS. In the following, we will also introduce the variable x, defined as the arc-length downstream of the RNAP, again measured along the chromatin fiber (see Fig. 3).

Propagation of torsional stress

A preliminary issue is to investigate the propagation of the torsional stress generated by the polymerase through a fiber with a given local nucleosome state. Let us assume here that the fiber is exclusively composed of nucleosomes, with no allowed transition into reversomes (i.e., $\xi(x, t) = 0$ over the whole fiber at any time). At the chromatin scale, inertial effects can be ignored, hence it is relevant to restrict ourselves to the overdamped regime, in which external forces and torques are fully balanced by viscous dissipation. We introduce the torsional shear strain $\tau(x, t)$ and the integrated torsion

$$\Theta(x,t) = \int_x^{l_0} \tau(z,t)dz,$$

such that $\Theta(x, t)$ is the angle by which a fixed point on the chromatin fiber surface at abscissa x has turned around the fiber axis at time t. By equating the elastic torque (torsional shear stress) and the viscous (Stokes) torque, we get

$$\frac{\partial \Gamma}{\partial x}(x,t) = \eta R^2 \frac{\partial \Theta}{\partial t}(x,t), \qquad (2)$$

where $\Gamma(x, t)$ is the elastic torque exerted at x on the part of the loop downstream of x; $(\partial\Gamma/\partial x)(x, t)$ is the net elastic torque experienced by an element dx of the elastic rod of radius R modeling the chromatin fiber; and η is the dynamic viscosity of the surrounding solvent (water or crowded chromatin, but in any case, η does not exceed 10-times the viscosity of pure water $\eta \approx 10^{-3}$ N.s.m^{-2}). The elastic

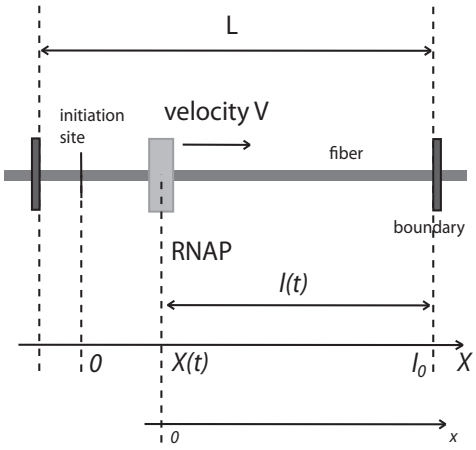


FIGURE 3 Relative positions along the chromatin fiber. The polymerase moves to the right, $X(t)$ being its position at time t. The value l_0 is the length of the loop region downstream of the initiation site $X(0) = 0$, and $l(t) = l_0 - X(t)$ is the length remaining at time t between the polymerase and the downstream boundary.

torque can be determined within linear response theory and is proportional to the torsional shear strain,

$$\Gamma(x,t) = kT\, L_p\, \tau(x,t), \tag{3}$$

where k is the Boltzmann constant, T the temperature, and L_p the twist persistence length of the fiber. The value of L_p varies from ~30 nm in a loosely condensed fiber ($\Lambda = 0.5\ nm^{-1}$) up to ~300 nm in a tightly condensed fiber ($\Lambda = 1\ nm^{-1}$) because of steric hindrance (30). Jointly, these two equations lead to a plain diffusion equation for $\Theta(x, t)$,

$$\frac{\partial \Theta}{\partial t} = D\frac{\partial^2 \Theta}{\partial x^2}, \tag{4}$$

where $D = kT\, L_p\, /\eta R^2$ (yielding $D = 1.8\ 10^{-10}\ m^2/s$ for $L_p \approx 30$ nm).

The relevant boundary conditions in our context are those of a finite chromatin fiber of length $l(t)$ with one end fixed (on the downstream boundary) and one end rotating (on the RNAP side) at constant angular velocity ω_0. In this case we have $\Theta(0, t) \equiv \omega_0 t$ and, on the downstream boundary, $\Theta(l(t), t) \equiv 0$. Anticipating that the torsional shear stress propagates much faster than the RNAP progresses, we start by keeping the RNAP fixed at $x = 0$ (quasistationary approximation), hence fixing $l(t) = l_0$. We can then look for a stationary solution in the form $\Theta(x, t) = f(x)\omega_0 t$: substituting into Eq. 4, we obtain

$$f(x) = \frac{\sinh\left[(l_0 - x)/\sqrt{Dt}\right]}{\sinh\left(l_0/\sqrt{Dt}\right)}. \tag{5}$$

The scaling form of this expression means that the torsional shear strain spreads along the fiber in a diffusive way roughly as $\sqrt{Dt}$. It thus takes no more than $t_0 \sim l_0^2/D \approx 10^{-2}$ s for the torsional strain to invade the whole loop, whereas the polymerase progresses by no more than $1.6\ 10^{-2}$ turn, i.e., 0.16 bp, during this time. This validates the quasistationary approximation made in investigating the stress propagation, while still considering that the RNAP stays fixed at $x = 0$.

For $t \gg t_0$, the function $f(x)$ reduces to the simple linear equivalent expression $f(x) \sim 1 - x/l_0$, leading to

$$\Theta(x,t) \approx \left(1 - \frac{x}{l_0}\right)\omega_0 t, \tag{6}$$

$$\tau(x,t) \approx \frac{1}{l_0}\omega_0 t, \tag{7}$$

$$\Gamma(x,t) \approx \frac{kT\, L_p}{l_0}\omega_0 t, \tag{8}$$

so that the torsional strain $\tau(x, t)$ and the torque $\Gamma(x, t)$ become very quickly homogeneous all along the fiber and then increase linearly with time. We conclude that the torque $\Gamma(0, t)$ that RNAP should exert on the downstream part of the loop, to progress at a constant angular velocity, would rapidly exceed its maximum value. This has been estimated on *Escherichia coli* RNAP to be <40 pN·nm (31). Considering the typical values given in Table 1, the maximum torque would be reached after RNAP has progressed by less than half a turn, i.e., 5 bp. This feature shows that RNAP cannot progress simply this way through a topologically constrained fiber, thus requiring either topoisomerase activity, if available, or a more sophisticated scenario involving conformational changes within the fiber, strain exchange, and ensuing stress relaxation. In the next section, we examine such a scenario and check its validity.

ELONGATION WITHIN A CONDENSED FIBER

Mechanical control of the nucleosome-reversome transition

Let us now consider the RNAP activity specifically within a condensed fiber. The most relevant feature of the fiber structure (25) is the regular and close nucleosome stacking into helical piles. (Helical piles are also known as ''starts''; the helical axis of each ''start'' being, by definition, transverse to the dyad axis of the stacked nucleosomes, there is only one way of decomposing the 30-nm fiber into a bunch of a variable number n of nucleosomal piles: one thus speaks of n-start fiber structure (25)). See Fig. 2. The closeness of stacked nucleosome faces along the start axis generates geometrical (hence mechanical) constraints on the conformational changes of single nucleosomes. The conversion of a single nucleosome into a reversome within a stacked pile is prevented due to steric hindrance. With the progression of the RNAP, the supercoiling constraint increases. The

torsional constraint is then essentially applied to the last nucleosome in the pile, although the rest of the fiber remains rigidly packed. Steric hindrance thus favors a domino effect where, under the effect of the applied torque, the nucleosomes pass to their altered reversome *R* state one by one, forming a progressive wavefront. The fiber response to the torsional constraint imposed by the RNAP activity is now controlled by the direct interaction between the border layer of the reversome wavefront and the adjacent layer of the stacked nucleosomes, and essentially by what happens in the linker relating the most downstream reversome and its neighboring nucleosome: here is the basic step in the propagation of the mechanical constraints that triggers the transition of the said nucleosome into a reversome and later stabilizes it in an irreversible way. In this model, steric constraints prevent the relaxation to chemical equilibrium and actually maintain the fiber in a far-from-equilibrium state.

Linking number conservation: a naive model

At which speed does the reversome wavefront progress? A naive model of the process can be introduced that immediately leads to an approximate but quite accurate estimation. The previous qualitative analysis leads us to assume, as a closure relation, a steplike profile for the local fraction of reversomes $\xi(x, t)$ (see Fig. 4),

$$\xi(x,t) = 1 \text{ for } x \leq x^*(t), \quad \text{else } 0 \tag{9}$$

with $x^*(t)$ the position of the reversome wavefront with respect to the RNAP location.

As RNAP moves forward, the linking number variation in the fiber, $\omega_0 t/2\pi$, i.e., the additional number of turns of one fiber end imposed by RNAP at time t, is mainly absorbed into the $A \rightarrow P \rightarrow R$ transitions that have occurred in the fiber region $x \leq x^*(t)$. Explicitly, the linking number conservation condition writes

$$\frac{\omega_0 t}{2\pi} \sim \Lambda \ \Delta\mathbf{Lk}^{\mathrm{RA}} \ x^*(t), \tag{10}$$

where we have introduced the linking number difference between the *R* and *A* states $\Delta\mathbf{Lk}^{\mathrm{RA}} = \mathbf{Lk}^{\mathrm{R}} - \mathbf{Lk}^{\mathrm{A}}$. This leads to an approximate estimation of the reversome wavefront motion,

$$x^*_{\mathrm{est}}(t) \sim \frac{\omega_0}{2\pi\Lambda \ \Delta\mathbf{Lk}^{\mathrm{RA}}} t, \tag{11}$$

which therefore progresses at constant speed

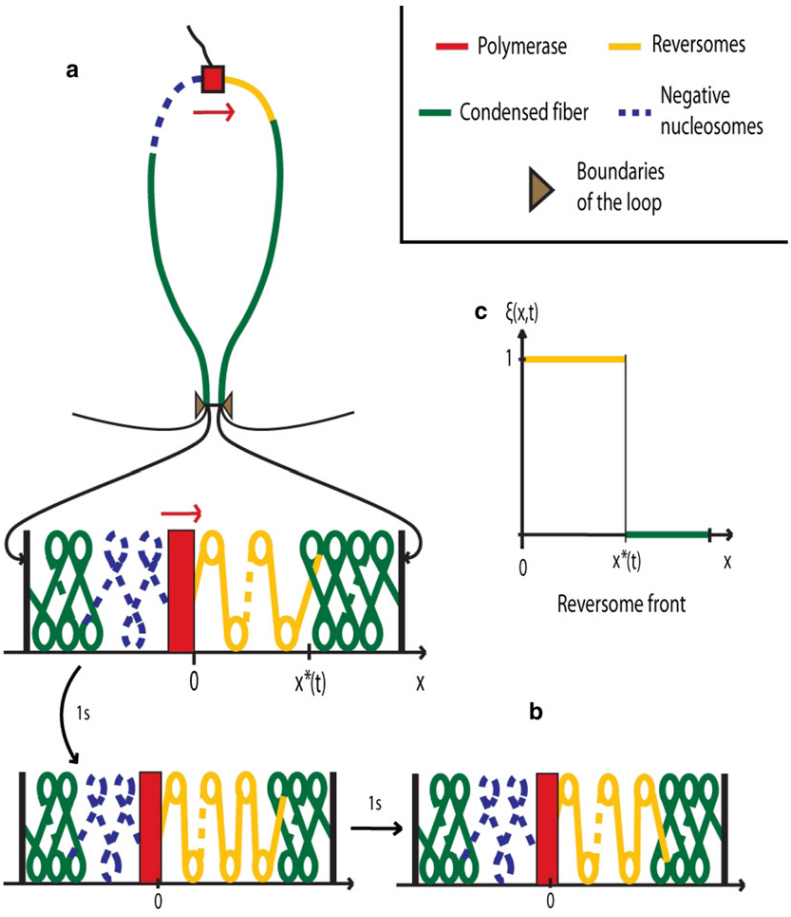


FIGURE 4 (*Color online*) RNA-polymerase processing within condensed chromatin fiber. (*a*) The supercoiling generated by the polymerase activity is trapped within the loop delineated by topological boundaries (the *thin black regions* are outside the loop). The ensuing torsional constraints trigger the sequential transition of nucleosomes (in *green*) into reversomes (in *yellow*). (*b*) Illustration of the domino effect: after 1 s, the fifth nucleosome downstream of the polymerase (*green* in panel *a*) has turned into a reversome (*yellow* in panel *b*); after one more second, the sixth nucleosome has turned into a reversome. (*c*) Reversome density profile: in the bold yellow region $[0, x^*]$, the reversome density $\xi(x, t)$ equals 1. The wavefront is located at x^* and propagates downstream ~10 times faster than the polymerase progression. In the polymerase wake, the nucleosomes turn to the negative state (*dashed blue* in panel *a*) to ensure the conservation of the total linking number of the loop.

$$v_{\text{est}} = \frac{\omega_0}{2\pi\Lambda\ \Delta\mathbf{Lk}^{\text{RA}}}\ [\text{nm/s}] = \frac{\omega_0}{2\pi\ \Delta\mathbf{Lk}^{\text{RA}}}\ [\text{nucl./s}] = \frac{\omega_0\ n_{\text{repeat}}}{2\pi\ \Delta\mathbf{Lk}^{\text{RA}}}\ [\text{bps/s}]. \quad (12)$$

By using the values of Table 1, the last two expressions give 1 nucl./s and 200 bps/s, respectively. Note that RNAP progresses in a much slower way, because V is ~10 times slower than v_{est}.

Of course, the previous derivation is oversimplified, insofar as it neglects the torsional shear strain induced by the applied torque. Nevertheless, as it will be confirmed through a more precise model, the obtained estimate for the reversome wavefront speed is rather good for a large choice of realistic fiber parameters.

Linking number conservation: accounting for mechanical constraints

Fiber torsion contribution

In naked DNA, the linking number conservation expresses itself in a balanced interchange between DNA twist and plectoneme formation (DNA writhe). In a chromatin loop, a different tradeoff will take place between chromatin fiber torsion and nucleosome conformational transitions. As discussed in the subsection Propagation of Torsional Stress, the applied torque spreads extremely rapidly through the whole fiber extent and becomes therefore homogeneous in a very short lapse of time. However, at this point, the torsional strain $\tau(x, t)$ now has a discontinuity at $x = x^*(t)$ because of the change in the persistence length of the fiber; indeed, the part of the fiber upstream of $x^*(t)$ is exclusively composed of reversomes whereas the part downstream is composed of nucleosomes. Inasmuch as reversomes have a more open structure than nucleosomes (see Fig. 1), reversome fibers are expected to be loosely condensed, with a persistence length close to 30 nm, ~10 times smaller than in a tightly condensed fiber. Hence, the torsional strain $\tau(x, t)$ downstream of $x^*(t)$ is negligible with respect to the one upstream and will be put to zero. We get, therefore,

$$\tau(x,t) \equiv \tau(t) = \Gamma(t)/\left(kT\ \mathrm{L_p}\right) \text{ for } x \leq x^*(t), \text{ else } 0. \quad (13)$$

This torsional strain should be accounted for in the fiber linking number balance.

The linking number of the fiber $\mathbf{Lk}^{\text{fiber}}$ can be decomposed into fiber writhe and twist contributions (8):

$$\mathbf{Lk}^{\text{fiber}} = \mathbf{Tw}^{\text{fiber}} + \mathbf{Wr}^{\text{fiber}}. \quad (14)$$

In practice, the fiber persistence length and the fiber diameter are such that the fiber axis can only bend smoothly, so that its writhe is practically negligible (8), at $\mathbf{Wr}^{\text{fiber}}$ ~0. Moreover, for a relaxed and homogeneous fiber, the twist can be written as the sum of the single nucleosome linking numbers (8): for a fiber of length $x^*(t)$ with N nucleosomes in the state X, $\mathbf{Tw}^{\text{fiber}} = N\,\mathbf{Lk}^{\text{X}} = \Lambda\mathbf{Lk}^{\text{X}}\,x^*(t)$. If such a fiber is now subjected to a torque $\Gamma(t)$, a torsional contribution $\tau(t)\,x^*(t)/2\pi$ should be added to the relaxed fiber twist, and we finally get

$$\mathbf{Lk}^{\text{fiber}} \simeq \mathbf{Tw}^{\text{fiber}} = \left[\Lambda\mathbf{Lk}^{\text{X}} + \frac{\tau(t)}{2\pi}\right]x^*(t). \quad (15)$$

The conservation of the linking number is therefore more correctly expressed by

$$\frac{\omega_0\,t}{2\pi} = \mathbf{Lk}^{\text{fiber}}(t) - \mathbf{Lk}^{\text{fiber}}(0) = \left[\Lambda\Delta\mathbf{Lk}^{\text{RA}} + \frac{\tau(t)}{2\pi}\right]x^*(t). \quad (16)$$

On the right-hand side of Eq. 16, we recognize the fiber twist (coming from the contribution of all the transitions into the reversome state) that has occurred in the loop at time t, which has added to the fiber torsion $\Theta[0, t]/2\pi$.

Equation 16 leads to a correction to our initial naive estimation of Eq. 10. To solve this equation for $x^*(t)$, we now need to calculate the torque $\Gamma(t)$ during RNAP progression. This requires a detailed description of the $A \rightarrow R$ transition kinetics.

Transition kinetics and critical torque

Whereas the $A \rightarrow P$ transition is rapid and occurs with almost no energetic cost (16), the $P \rightarrow R$ transition implies the crossing of a large free energy barrier. It is therefore a kinetic process, described by the rate equation

$$\frac{\partial P_{\text{R}}}{\partial t} = kP_{\text{P}} - k'P_{\text{R}}, \quad (17)$$

with P_{R} (respectively, P_{P}) the probability of being in the R (respectively, P) state. The forward and backward rate constants are given by $k = k_0\ \exp(-\ (G_{\text{B}} - G_{\text{P}})/kT)$ and $k' = k_0\ \exp(-\ (G_{\text{B}} - G_{\text{R}})/kT)$, respectively, with $G_{\text{X}} = F_{\text{X}} - 2\pi\mathbf{Lk}^{\text{X}}\Gamma$ the Gibbs potential for the X state (20). In practice, however, the reverse transition $R \rightarrow P$ is highly improbable for typical RNAP velocities (with k_0 given in Table 1 and the torque $\Gamma = \Gamma_{\text{c}}$ obtained in Appendix S1 in the Supporting Material, we get k ~6 s^{-1} and k' ~10^{-51} s^{-1}) so that the term $-k'P_{\text{R}}$ in Eq. 17 can be neglected, thus leading to the simplified kinetic equation

$$\frac{\partial P_{\text{R}}}{\partial t} \simeq kP_{\text{P}}. \quad (18)$$

Each $P \rightarrow R$ transition should occur within a typical transition time matching the RNAP velocity. Due to the kinetic character of the transition, the torque should therefore reach a critical threshold Γ_{c} (and the torsional strain a corresponding critical value τ_{c} to allow the transition into reversome to occur within this typical time). The critical torque Γ_{c} can be calculated following Evans (32), as done in Appendix S1 in the Supporting Material. It results to be constant with very good approximation, and writes

$$\Gamma_c \simeq \frac{1}{B}\left\{\Delta F + kT \ln\left(\frac{B\omega_0}{k_0}\right)\right\}, \quad (19)$$

where we have introduced the constants $\Delta F = F_B - F_P$ and $B = 2\pi(\mathbf{Lk}^B - \mathbf{Lk}^P)$, with $\mathbf{Lk}^B$ the linking number at the barrier position. Numerically, the value of the critical torque strongly depends on $\mathbf{Lk}^B$. Using the parameters listed in Table 1, we obtain Γ_c in the interval 3–9 kT, or, equivalently, 15–35 pN·nm.

The reversome wavefront velocity is slightly reduced with respect to the estimation v_{est} of Eq. 12, and writes

$$v = \frac{\omega_0}{2\pi\Lambda\Delta\mathbf{Lk}^{RA} + \tau_c} = \frac{v_{est}}{1 + \left(\tau_c/2\pi\Lambda\Delta\mathbf{Lk}^{RA}\right)}. \quad (20)$$

Equation 20 indicates that the additional fiber torsion introduced by the RNAP progression is not fully absorbed by nucleosome state transition, but partially used in twisting the fiber rod itself. As a consequence, the RNAP should apply greater than two turns for each $A \rightarrow P \rightarrow R$ transition, which leads to the observed decrease in the wavefront velocity. In any case, with Γ_c in the interval 15–35 pN·nm, the estimated v_{est} always matches the exact velocity within 6% (and down to 0.3% in the best case).

A complete picture of the stepping progression of the reversome wavefront, that takes into account its discrete nature, is given in Appendix S2 in the Supporting Material.

BIOLOGICAL INTERPRETATION AND PREDICTIONS

The torque exerted by RNAP

In the scenario that emerges from previous considerations, the transcription of every 20 bp induces two positive coils downstream that can be absorbed by the formation of one reversome. A reversome wavefront progresses downstream of an elongating RNAP II at a rate ~200 bp/s. Moreover, this reversome wavefront is expected to stop at boundary elements, because they act as topological insulators. To ensure the relevance of this model, however, RNAP should be able to exert a positive torque sufficient to trigger the transition. We have found in the section Transition Kinetics and Critical Torque that the maximum value of the torque predicted by the model amounts to Γ_c = 15–35 pN·nm.

The torque necessary to trigger the chiral transition of a nucleosome into a reversome has been recently measured (33). The authors reported a value close to 10 pN·nm in very low salt conditions (10 mM phosphate buffer). On the other hand, the torque exerted by *E. coli* RNAP has been estimated to be at least 6 pN.nm and always lower than 40 pN·nm (31). The interval obtained in our model is therefore included in the one proposed by Harada et al. (31). Moreover, recent experiments (34) gave the first in vivo evidence for torque generation by elongating RNAP II in eukaryotes, indicating that mechanical stresses, constrained by architectural features of DNA and chromatin, may broadly contribute to gene regulation. Transcription-generated dynamic DNA supercoiling may be propagated over thousands of basepairs through chromatin and contribute to the control of a variety of DNA transactions (34).

These data demonstrate that RNAP is a powerful molecular motor, likely to exert sufficiently high torque for inducing the $A \rightarrow R$ transition and generating a reversome wavefront, as described in this article. Of course, any measure of the torque exerted by RNAP in physiological conditions would be highly valuable and would, moreover, provide a critical test of our model.

Transcription in a compact fiber

An important feature of the presented model is that the progressing RNAP encounters only nucleosomes in an activated state (here identified with the reversome state). This process achieves twist relaxation and at the same time ensures that the RNAP progresses in a locally open and transcriptionally permissive configuration, encountering only transparent reversomes. Steric constraints prevent the chemical equilibrium from being reached (a kind of frustration phenomenon) and enforce the sequential transition of nucleosomes into reversomes.

We are thus led to the following quite counterintuitive prediction: RNAP progression is facilitated in a compact chromatin fiber, because steric constraints between nucleosomes enforce a steplike reversome profile, ensuring that the RNAP will always face reversomes during its progression. In other words, RNAP activity within a compact fiber modifies its surroundings in such a way as to ensure that each nucleosome encountered by the RNAP as it moves along the fiber will be in the reversome conformation. The spreading of the reversome phase appears as a precursor extending farther and farther ahead of the processing RNAP, and moving ~10 times faster than the RNAP.

Interestingly, there is evidence that transcription of siRNAs occurs in highly condensed chromatin (35).

The loop decondensation indirectly observed in vivo in yeast (36), where a chromatin locus moves toward a nuclear pore upon transcription, is a consequence of the conversion of fiber twist into fiber writhe (37). Arguably, only the decondensation associated with the conformational change of nucleosomes into reversomes is required for polymerase processing. It is nevertheless important to emphasize here that we consider only the elongation phase; a local decondensation of the 30-nm fiber is required for the transcription initiation.

Comparison with recent experiments

The reversome wavefront proposed by our model progresses downstream of an elongating RNAP II at a rate ~200 bp/s. Strikingly, recent experiments by Petesch and Lis (38) give evidence for a rapid wavefront of nucleosome disruption, progressing at a comparable rate and stopping at the loop

boundary. This wavefront arises immediately after heat shock induction and before productive elongation.

We propose that heat-shock transcription factor binding triggers a rapid productive elongation phase, during which RNAP II translocates over some genomic distance. Its progression in a topologically constrained environment creates positive torque in the downstream portion of the template, high enough to convert, at a distance, a fraction of nucleosomes into reversomes, through a domino effect. Arguably, this first productive elongation phase is too fast for topoisomerases to come into play. Reversomes are expected to be much less stable and to easily lose H2A/H2B dimers to form hexasomes or tetrasomes. Some reversomes may be lost altogether because H3/H4 tetramers prefer to bind negatively supercoiled DNA (21). Thus, the positive torque in front of the advancing RNAP will produce a complex (random) mixture of integral or partially disrupted reversomes, or will disrupt the nucleosome particles altogether. Our model thus explains straightforwardly why, in the Petesch and Lis experiments, nucleosome disruption observed downstream of the RNAP is much faster than the rate of elongation, and why it occurs over the entire downstream region and is limited to it.

Transcription initiation and RNAP pauses

Another interesting feature that has emerged from our analysis is the need of a DNA stretch free of nucleosomes at the beginning of the transcribed region (see Appendix S2). This free DNA length should ensure that at least two turns of supercoiling have been accumulated downstream of the RNAP before it arrives in front of the first nucleosome, so that the $A \rightarrow R$ transition can be achieved without inducing any negative torque. Relevant to our modeling, an initial region free of nucleosomes, immediately downstream of the TIS, is often observed (39,40). It is also interesting to note that, in the Petesch and Lis (38) experiment, and even under non-heat-shock conditions, the gene harbors a paused molecule of RNAP II, at position (+20)–(+40). RNAP stalling at this position occurs even after the gene is induced, even if its residence time dramatically decreases upon gene activation (41–43).

As soon as all the nucleosomes in the loop have turned into reversomes, the additional supercoiling due to further elongation fully accumulates in the form of fiber torsion; then the strain rapidly becomes too large, hence the resisting torque too strong, for the RNAP to progress any further, and a pause in the transcriptional activity is observed. Our model thus predicts that RNAP pausing will occur soon after the reversome wavefront has reached the loop boundary, i.e., when $t = t^*$, while the RNAP has traveled $\sim l_0/10$. For typical values of l_0 as given in Table 1, this leads to the rough estimate that pauses will occur after RNAP has transcribed 1–5 kb corresponding to a duration between 50 s and 250 s of nonstop elongation. This is in striking agreement with the elongation residence time recently evaluated by Darzacq et al. (44): these authors reported the first complete set of kinetic parameters of RNAP II transcription in physiological conditions (see Table 1 in their article); they found in particular an elongation residence time of ~30 s with pausing occurring ~1 kb downstream from the promoter. It remains to be seen whether there is a boundary 10 kb downstream from the TIS. More generally we suggest measuring nonstop elongation times for different loci together with the length of the corresponding genomic region downstream of the TIS.

CONCLUSION

Based on the facts that polymerase transcribes only through an activated nucleosome state and that its progression modifies the DNA linking number, we have proposed a scenario elucidating how transcription elongation can proceed within condensed chromatin. At odds with current views, this scenario does not require a decondensation of the 30-nm fiber. Our modeling study of the interplay between the RNAP activity and the chromatin fiber conformational dynamics evidences that, on the contrary, the presence of steric, mechanical, and topological constraints enforce an ordered preactivation of the fiber downstream of the RNAP. More precisely, within a condensed fiber loop with closely stacked nucleosomes, the very RNAP activity and the torsional constraints it generates at a distance along the chromatin fiber trigger the propagation of a conformational transition of the nucleosomes into a transcription-prone structure, more permissive to RNAP processing and transcriptional activity; we identify this nucleosomal structure with a recently proposed reversome conformation. Importantly, such an allosteric mechanism is relevant only in a condensed chromatin fiber. Obviously, alternative scenarios are to be searched for in other contexts likely to involve decondensed chromatin, e.g., for elongating RNAP I or III, or even RNAP II in highly transcribed genes.

Of note, we stress that all the relevant parameters—which are listed in Table 1—are taken from the literature, hence there are no fitted parameters in our model.

Finally, let us underline that it is a general fact that topological constraints induce long-range couplings along the fiber that coordinate fiber transactions and processes at the scale of a chromatin loop (typically embedding exons and introns associated to one gene); topological invariants play a channeling role in strongly constraining the possible deformations of the fiber. Conversely, functional constraints strongly condition the structure and dynamics of the fiber. Presumably, chromatin structure and function have coevolved so as to reach a good, if not optimal, consistency and efficiency.

SUPPORTING MATERIAL

Two appendices and one figure are available at http://www.biophysj.org/biophysj/supplemental/S0006-3495(09)01734-2.

We thank the reviewers for their recommendations, which improved the manuscript significantly.

C.B. holds a Doctoral Fellowship from the Agence Nationale de Recherche sur le Syndrome d'Immuno Déficience Acquise et les Hépatites Virales. Work at the Institut des Hautes Études Scientifiques has also been funded by a grant from the Génopole Evry, France. M.B. and J.-M.V. are supported in part by Agence Nationale de Recherche grant No. 05-NANO-062-03.

REFERENCES

1. Shilatifard, A., R. C. Conaway, and J. W. Conaway. 2003. The RNA polymerase II elongation complex. *Annu. Rev. Biochem.* 72:693–715.
2. Orphanides, G., and D. Reinberg. 2000. RNA polymerase II elongation through chromatin. *Nature*. 407:471–475.
3. Chang, C. H., and D. S. Luse. 1997. The H3/H4 tetramer blocks transcript elongation by RNA polymerase II in vitro. *J. Biol. Chem.* 272:23427–23434.
4. Wolffe, A. P. 1998. Chromatin: Structure and Function. Academic Press, New York.
5. Cook, P. R. 1999. The organization of replication and transcription. *Science*. 284:1790–1795.
6. Byrd, K., and V. G. Corces. 2003. Visualization of chromatin domains created by the gypsy insulator of Drosophila. *J. Cell Biol.* 162:565–574.
7. Labrador, M., and V. G. Corces. 2002. Setting the boundaries of chromatin domains and nuclear organization. *Cell*. 111:151–154.
8. Barbi, M., J. Mozziconacci, and J. M. Victor. 2005. How the chromatin fiber deals with topological constraints. *Phys. Rev. E Stat. Nonlin. Soft Matter Phys.* 71:031910.
9. Liu, L. F., and J. C. Wang. 1987. Supercoiling of the DNA template during transcription. *Proc. Natl. Acad. Sci. USA*. 84:7024–7027.
10. Lavelle, C. 2007. Transcription elongation through a chromatin template. *Biochimie*. 89:519–527.
11. Giaever, G. N., and J. C. Wang. 1988. Supercoiling of intracellular DNA can occur in eukaryotic cells. *Cell*. 55:849–856.
12. Ljungman, M., and P. C. Hanawalt. 1992. Localized torsional tension in the DNA of human cells. *Proc. Natl. Acad. Sci. USA*. 89:6055–6059, (Erratum in Proc. Natl. Acad. Sci. USA, 89:E9364).
13. Matsumoto, K., and S. Hirose. 2004. Visualization of unconstrained negative supercoils of DNA on polytene chromosomes of *Drosophila*. *J. Cell Sci.* 117:3797–3805.
14. Bancaud, A., N. Conde e Silva, …, J. L. Viovy. 2006. Structural plasticity of single chromatin fibers revealed by torsional manipulation. *Nat. Struct. Mol. Biol.* 13:444–450.
15. Sivolob, A., and A. Prunell. 2003. Linker histone-dependent organization and dynamics of nucleosome entry/exit DNAs. *J. Mol. Biol.* 331:1025–1040.
16. Prunell, A., and A. Sivolob. 2004. Paradox lost: nucleosome structure and dynamics by the DNA minicircle approach. *In* Chromatin Structure and Dynamics: State of the Art, New Comprehensive Biochemistry., Vol. 39. J. Zlatanova and S. Leuba, editors. Elsevier, Amsterdam, The Netherlands.
17. Kireeva, M. L., W. Walter, …, V. M. Studitsky. 2002. Nucleosome remodeling induced by RNA polymerase II: loss of the H2A/H2B dimer during transcription. *Mol. Cell*. 9:541–552.
18. Lee, M. S., and W. T. Garrard. 1991. Positive DNA supercoiling generates a chromatin conformation characteristic of highly active genes. *Proc. Natl. Acad. Sci. USA*. 88:9675–9679.
19. Bondarenko, V. A., L. M. Steele, …, V. M. Studitsky. 2006. Nucleosomes can form a polar barrier to transcript elongation by RNA polymerase II. *Mol. Cell*. 24:469–479.
20. Bancaud, A., G. Wagner, …, A. Prunell. 2007. Nucleosome chiral transition under positive torsional stress in single chromatin fibers. *Mol. Cell*. 27:135–147.
21. Zlatanova, J., T. C. Bishop, …, K. van Holde. 2009. The nucleosome family: dynamic and growing. *Structure*. 17:160–171.
22. Mozziconacci, J., and J. M. Victor. 2003. Nucleosome gaping supports a functional structure for the 30nm chromatin fiber. *J. Struct. Biol.* 143:72–76.
23. van Holde, K., and J. Zlatanova. 2007. Chromatin fiber structure: where is the problem now? *Semin. Cell Dev. Biol.* 18:651–658.
24. Weidemann, T., M. Wachsmuth, …, J. Langowski. 2003. Counting nucleosomes in living cells with a combination of fluorescence correlation spectroscopy and confocal imaging. *J. Mol. Biol.* 334:229–240.
25. Wong, H., J. M. Victor, and J. Mozziconacci. 2007. An all-atom model of the chromatin fiber containing linker histones reveals a versatile structure tuned by the nucleosomal repeat length. *PLoS One*. 2:e877.
26. Uptain, S. M., C. M. Kane, and M. J. Chamberlin. 1997. Basic mechanisms of transcript elongation and its regulation. *Annu. Rev. Biochem.* 66:117–172.
27. Ben-Haïm, E., A. Lesne, and J. M. Victor. 2001. Chromatin: a tunable spring at work inside chromosomes. *Phys. Rev. E Stat. Nonlin. Soft Matter Phys.* 64:051921.
28. Love, A. E. H. 1944. Treatise on the Mathematical Theory of Elasticity Theory. Dover, Mineola, NY.
29. Lesne, A., and J. M. Victor. 2006. Chromatin fiber functional organization: some plausible models. *Eur Phys J E Soft Matter*. 19:279–290.
30. Mergell, B., R. Everaers, and H. Schiessel. 2004. Nucleosome interactions in chromatin: fiber stiffening and hairpin formation. *Phys. Rev. E Stat. Nonlin. Soft Matter Phys.* 70:011915.
31. Harada, Y., O. Ohara, …, K. Kinosita, Jr. 2001. Direct observation of DNA rotation during transcription by *Escherichia coli* RNA polymerase. *Nature*. 409:113–115.
32. Evans, E. 2001. Probing the relation between force—lifetime—and chemistry in single molecular bonds. *Annu. Rev. Biophys. Biomol. Struct.* 30:105–128.
33. Celedon, A., I. M. Nodelman, …, S. X. Sun. 2009. Magnetic tweezers measurement of single molecule torque. *Nano Lett.* 9:1720–1725.
34. Kouzine, F., S. Sanford, …, D. Levens. 2008. The functional response of upstream DNA to dynamic supercoiling in vivo. *Nat. Struct. Mol. Biol.* 15:146–154.
35. Grewal, S. I., and S. C. Elgin. 2007. Transcription and RNA interference in the formation of heterochromatin. *Nature*. 447:399–406.
36. Cabal, G. G., S. Rodriguez-Navarro, …, U. Nehrbass. 2006. Molecular analysis of SAGA mediated nuclear pore gene gating activation in yeast. *Nature*. 441:770–773.
37. Mozziconacci, J., C. Lavelle, …, J. M. Victor. 2006. A physical model for the condensation and decondensation of eukaryotic chromosomes. *FEBS Lett.* 580:368–372.
38. Petesch, S. J., and J. T. Lis. 2008. Rapid, transcription-independent loss of nucleosomes over a large chromatin domain at Hsp70 loci. *Cell*. 134:74–84.
39. Vaillant, C., B. Audit, and A. Arneodo. 2007. Experiments confirm the influence of genome long-range correlations on nucleosome positioning. *Phys. Rev. Lett.* 99:2181035.
40. Miele, V., C. Vaillant, …, T. Grange. 2008. DNA physical properties determine nucleosome occupancy from yeast to fly. *Nucleic Acids Res.* 36:3746–3756.
41. Saunders, A., L. J. Core, and J. T. Lis. 2006. Breaking barriers to transcription elongation. *Nat. Rev. Mol. Cell Biol.* 7:557–567.
42. Lis, J. T. 2007. Imaging *Drosophila* gene activation and polymerase pausing in vivo. *Nature*. 450:198–202.
43. Nechaev, S., and K. Adelman. 2008. Promoter-proximal Pol II: when stalling speeds things up. *Cell Cycle*. 7:1539–1544.
44. Darzacq, X., Y. Shav-Tal, …, R. H. Singer. 2007. In vivo dynamics of RNA polymerase II transcription. *Nat. Struct. Mol. Biol.* 14: 796–806.

.

F

Journal article : IgG Autoantibody to Brain Beta Tubulin III Associated with Cytokine Cluster-II Discriminate Cerebral Malaria in Central India.

Devendra Bansal, Fabien Herbert, Pharath Lim1, Prakash Deshpande, Christophe Bécavin, Vincent Guiyedi, Ilaria de Maria, Jean Claude Rousselle, Abdelkader Namane, Rajendra Jain, Pierre-Andre Cazenave, Gyan Chandra Mishra, Cristiano Ferlini, Constantin Fesel, Arndt Benecke, Sylviane Pied.

Published in PLoS ONE, December 2009, Volume 4, Issue 12.

Abstract

Background : The main processes in the pathogenesis of cerebral malaria caused by Plasmodium falciparum involved sequestration of parasitized red blood cells and immunopathological responses. Among immune factors, IgG autoantibodies to brain antigens are increased in P. falciparum infected patients and correlate with disease severity in African children. Nevertheless, their role in the pathophysiology of cerebral malaria (CM) is not fully defined. We extended our analysis to an Indian population with genetic backgrounds and endemic and environmental status different from Africa to determine if these autoantibodies could be either a biomarker or a risk factor of developing CM.

Methods/Principal Findings : We investigated the significance of these self-reactive antibodies in clinically well-defined groups of P. falciparum infected patients manifesting mild malaria (MM), severe non-cerebral malaria (SM), or cerebral malaria (CM) and in control subjects from Gondia, a malaria epidemic site in central India using quantitative immunoprinting and multivariate statistical analyses. A two-fold complete-linkage hierarchical clustering allows classifying the different patient groups and to distinguish the CM from the others on the basis of their profile of IgG reactivity to brain proteins defined by PANAMA Blot. We identified beta tubulin III (TBB3) as a novel discriminant brain antigen in the prevalence of CM. In addition, circulating IgG from CM patients highly react with recombinant TBB3. Overall, correspondence analyses based on singular value decomposition show a strong correlation between IgG anti-TBB3 and ele-

vated concentration of cluster-II cytokine (IFNγ, IL1β, TNFα, TGFβ) previously demonstrated to be a predictor of CM in the same population.

Conclusions/Significance : Collectively, these findings validate the relationship between antibody response to brain induced by P. falciparum infection and plasma cytokine patterns with clinical outcome of malaria. They also provide significant insight into the immune mechanisms associated to CM by the identification of TBB3 as a new disease-specific marker and potential therapeutic target.

OPEN ACCESS Freely available online

PLoS one

IgG Autoantibody to Brain Beta Tubulin III Associated with Cytokine Cluster-II Discriminate Cerebral Malaria in Central India

Devendra Bansal[1☯], Fabien Herbert[1☯], Pharath Lim[1☯], Prakash Deshpande[2], Christophe Bécavin[3], Vincent Guiyedi[1], Ilaria de Maria[4], Jean Claude Rousselle[5], Abdelkader Namane[5], Rajendra Jain[6], Pierre-André Cazenave[1,7], Gyan Chandra Mishra[2], Cristiano Ferlini[4], Constantin Fesel[8], Arndt Benecke[3], Sylviane Pied[1]*

1 Equipe PIME CNRS, Inserm U547, Institut Pasteur de Lille, Pôle Universitaire Nord, France, **2** National Centre for Cell Science, Pune, Pune (Maharashtra), India, **3** Institut de Recherche Interdisciplinaire CNRS USR3078 Univ. Lille I, II, and Institut des Hautes Études Scientifiques, Bures sur Yvettes, France, **4** Laboratory of Antineoplastic Pharmacology, Università Cattolica Sacro Cuore, Rome, Italy, **5** Institut Pasteur, Plate-Forme de Protéomique, CNRS URA 2185, Paris, France, **6** K.T.S. Hospital, Gondia District, Maharashtra, India, **7** Université Pierre et Marie Curie–CNRS U7087, and Institut Pasteur, Paris, France, **8** Instituto Gulbenkian de Ciência, Oeiras, Portugal

Abstract

Background: The main processes in the pathogenesis of cerebral malaria caused by *Plasmodium falciparum* involved sequestration of parasitized red blood cells and immunopathological responses. Among immune factors, IgG autoantibodies to brain antigens are increased in *P. falciparum* infected patients and correlate with disease severity in African children. Nevertheless, their role in the pathophysiology of cerebral malaria (CM) is not fully defined. We extended our analysis to an Indian population with genetic backgrounds and endemic and environmental status different from Africa to determine if these autoantibodies could be either a biomarker or a risk factor of developing CM.

Methods/Principal Findings: We investigated the significance of these self-reactive antibodies in clinically well-defined groups of *P. falciparum* infected patients manifesting mild malaria (MM), severe non-cerebral malaria (SM), or cerebral malaria (CM) and in control subjects from Gondia, a malaria epidemic site in central India using quantitative immunoprinting and multivariate statistical analyses. A two-fold complete-linkage hierarchical clustering allows classifying the different patient groups and to distinguish the CM from the others on the basis of their profile of IgG reactivity to brain proteins defined by PANAMA Blot. We identified beta tubulin III (TBB3) as a novel discriminant brain antigen in the prevalence of CM. In addition, circulating IgG from CM patients highly react with recombinant TBB3. Overall, correspondence analyses based on singular value decomposition show a strong correlation between IgG anti-TBB3 and elevated concentration of cluster-II cytokine (IFNγ, IL1β, TNFα, TGFβ) previously demonstrated to be a predictor of CM in the same population.

Conclusions/Significance: Collectively, these findings validate the relationship between antibody response to brain induced by *P. falciparum* infection and plasma cytokine patterns with clinical outcome of malaria. They also provide significant insight into the immune mechanisms associated to CM by the identification of TBB3 as a new disease-specific marker and potential therapeutic target.

Citation: Bansal D, Herbert F, Lim P, Deshpande P, Bécavin C, et al. (2009) IgG Autoantibody to Brain Beta Tubulin III Associated with Cytokine Cluster-II Discriminate Cerebral Malaria in Central India. PLoS ONE 4(12): e8245. doi:10.1371/journal.pone.0008245

Editor: Mauricio Martins Rodrigues, Federal University of São Paulo, Brazil

Received July 16, 2009; **Accepted** November 10, 2009; **Published** December 14, 2009

Funding: This work was part of the Centre National de la Recherche Scientifique (CNRS)-Programme d'Incitation à la Mobilité d'Equipe (PIME): Malaria Immunopathophysiology. It was supported by the Indo-French Centre for Promotion of Advanced Research (grant 2103-3 and 3703-2). This work was also supported by grants from the Institut Pasteur and Pasteur-Genopole-Ile-de-France and "Fondation des Treilles". Work in the Benecke group is funded by grants from the Genopole Evry, the Agence Nationale pour la Recherche (ANR), and the Agence Nationale de Recherches sur le SIDA (ANRS) et les hepatites virales. CB is recipient of a Ph.D. fellowship from the ANRS. CF received a post-doctoral fellowship from the Fundation para a Ciencia e Tecnologia (Portugal).

Competing Interests: The authors have declared that no competing interests exist.

* E-mail: sylviane.pied@pasteur-lille.fr

☯ These authors contributed equally to this work.

Introduction

Malaria remains a major cause of morbidity and mortality in humans, resulting 350–500 million clinical cases and over one million deaths annually [1]. *Plasmodium falciparum* infection generates pleiomorphic clinical outcomes, from asymptomatic to severe syndromes depending on transmission intensity, age of the individuals and on the immunity and the genetic background of the populations [2,3,4]. Anemia and cerebral malaria (CM) are the most severe manifestations and deaths occur by CM in children and young adults in area of high transmission [5]. CM is characterized by a range of acute neurological manifestations including a diffuse encephalopathy, alteration in levels of consciousness, deep coma and seizure preceding death [6,7]. Sequestration of parasitized erythrocytes in cerebral blood vessels is often associated to CM [8]. Adhesion of blood stage parasite has been considered to lead to a

decrease of the blood flow and to contribute to the induction of brain damage and coma during CM [9,10]. Additionally, CM is also considered to be the result of an immunopathological process involving both lymphocytes and proinflammatory (Th1) cytokines such as TNFα, levels of which are increased in affected patients [11–13]. Thus, the outcome of *P. falciparum* infection may depend on a fine balance between appropriate and inappropriate immune responses [14,15]. Although the occurrence of numerous metabolic, pathological and physiological abnormalities has been demonstrated during CM, the mechanisms leading to progression into complicated disease have not been yet adequately explained. Particularly, pathogenic roles for autoantibodies are not defined in CM.

When exposed to *Plasmodium* parasite, the host immune response is characterized by a polyclonal B-cell activation and a hyper gammaglobulinemia [16,17]. Among antibodies produced some of them recognize autoantigens [17,18]. High levels of antibodies against phospholipids, cardiolipin, ssDNA, dsDNA, and rheumatoid factors are correlated with disease severity in *P. falciparum*-infected patients [19–22]. However, their role in pathophysiology of CM remains unclear. Recently, by studying several cohorts of children manifesting different disease spectrums induced by *P. falciparum* from a hyper endemic area of Gabon, we demonstrated that antibody mediated self-reactive response may contribute to the pathogenesis of CM. Thus, in these children we observed a significant increase of the repertoire of plasmatic IgG reacting with human brain antigens with disease severity [23]. Interestingly, CM patients developed a high IgG autoantibody response to brain α II spectrin which is significantly associated with increased plasma concentrations of TNFα [23]. These autoantibodies may or may not cause damage. The relationship between CM and antibody dependent autoimmune reactions has been also illustrated by the occurrence of autoantibodies against voltage-gated calcium channels in African populations [24]. Multiple mechanisms underlie the production of autoantibodies such as a polyclonal activation of B cells due to stimulation by parasitic mitogens [25], a stimulation of specific B cells by molecular mimetism [26,27], or even a deregulation of the B cells function [25,28]. Other mechanisms such as apoptosis of brain endothelial cells occurring during cerebral malaria could also be source of release of the auto antigens [29,30].

In this study, we extended our analysis to an Indian population with genetic backgrounds, endemic and environmental status different from the Gabonese population to determine if autoreactive antibodies specific to brain antigens are present in CM patients and could play a role in malaria pathogenesis. We used a multidisciplinary approach based on quantitative immunoprinting associated to biostatistics to study the autoantibody repertoire to brain antigens in several groups of *P. falciparum* infected patients from an epidemic area of central India manifesting different clinical spectra of the disease. We found that the different clinical malaria phenotypes can be discriminate according to their profile of IgG reactivity to brain antigens. Furthermore, we identified a novel discriminant brain antigen, the beta tubulin III (TBB3), targeted by circulating IgG in the prevalence of CM. TBB3, a cytoskeleton protein, is mainly expressed in neural tissue [31]. Finally, we show that IgG reactivity to TBB3 is strongly correlated with elevated levels of the previously described cytokine cluster II, composed of IL10, TNFα, TGFβ and IL1β, that characterized CM in the same group of patients [32].

Materials and Methods

Ethics statement

This study was conducted according to the principles expressed in the Declaration of Helsinki. The study was approved by the Institutional Review Board of NCCS, Institut Pasteur Paris and Gondia hospitals. The study design was also approved by the National health office ethics committee in India. All patients or relatives provided written informed consent for the collection of samples and subsequent analysis.

Study area and subjects

Blood samples were collected from the individuals living in villages in and around Gondia town, an epidemic region in central India. Gondia is a low transmission region, and know as endemic area for the last 20 years. *P. falciparum* appeared in Gondia over the last 10 years [33]. The subjects were divided into the following groups. Group 1 consisted of subjects, who had CM within the past 6 months and recovered (ex-CM), healthy malaria endemic controls (EC) were patient's relatives (brothers/sisters/parents) who accompanied the patient to the hospital and not had malaria for at least the preceding 2 years, nor were they clinical asymptomatic carriers; and malaria non-endemic controls (NEC) were the subjects residing in the Pune city with no history of malarial disease for ≥5 years. Group 2 consisted of patients infected with *P. falciparum* having different clinical status, according to the criteria defined by the World Health Organization [34] i.e. mild malaria (MM), severe non-cerebral malaria (SM) and cerebral malaria (CM). Samples from infected groups were collected during the period of high malaria transmission. Patients with MM (hemoglobin ≥8 g/dl, parasitaemia asexual ≥10000/µl, fully consciousness) were not hospitalized and SM patients were also complete conscious and displayed good verbal response to the doctor's questions. Patients with CM were at coma stages I and III. Patients with severe malaria (SM or CM) received intravenous quinine (25 mg/kg/day) with 5% or 10% glucose solution for non-hypoglycaemic or hypoglycaemic patients for five days. The patients with severe anaemia underwent blood transfusion. Most of the CM patients recovered from disease in one or two weeks and have been discharged from the hospital. The clinical history and informed written consent were obtained from all the subjects and a demographic profile was recorded.

A Pool of CM serum was constituted with samples from patients showing a high reactivity to brain antigens. An EC pool of 5 sera was randomly chosen as negative control.

Blood samples collection and parasite assessment

Ten milliliters of whole blood was collected from each subject by vein puncture in sterile EDTA tubes or in sterile vacutainers during 2001–2003 from different hospitals in and around Gondia town. Plasma was obtained by centrifuging the blood samples at 4500 *g* for 15 min and stored at −80°C until further use. Parasitemia was assessed, on thin blood smear, by counting asexual forms of *P. falciparum* under a light microscope after Giemsa staining. The total numbers of infected and uninfected erythrocytes from 10 fields (magnification, X100) were counted, and parasitemia were calculated.

Extraction of antigen

Parasites extract. Parasite antigen prepared from synchronous cultures of a field derived *P. falciparum* parasite line *FAN5HS* [33], ≥25% parasitemia, was used. The parasitized red blood cells (pRBC) were washed five times in sterile PBS and then lysed by lysis buffer containing protease inhibitors and briefly sonicated. The contents of the tube were agitated by cyclo-mixing and then centrifuged at 6,000 rpm for 30 min at 4°C. The supernatant was collected in a separate tube and the pellet was discarded. Aliquots of the antigen were frozen at −70°C until use. Parasite protein was quantified by the standard Bradford method

[35]. The concentration of the parasite line *FAN5HS* was 1.2 mg/ml.

Normal RBC extracts. Normal red blood cell (RBC) extract was prepared from the same batch of RBCs used for culturing the parasites [33] and followed the same procedure as previously described for pRBCs.

Human brain extract. The protein extraction of brain was done by homogenization of whole brain taken from a healthy Cuban national, who died accidentally and never had malaria [36,37]. The brain tissue was suspended in extraction buffer containing 60 mM Tris, 2% SDS, 100 mM Dithiothreitol (DTT) and protease inhibitors: 1 µg/ml Aprotinine, 1 µg/ml Pepstatine, 50 µg/ml n-α-todyl-L-lysine chloromethyl ketone (TLCK). After centrifugations at 10000 rpm at 4°C for 10 minutes, the supernatant was transferred into a clean tube and protein contents were estimated using a commercial available kit (BCATM protein assay kit, Pierce, France). The concentration of the brain extract was 3 mg/ml. Commercially available brain extract (Protein MEDLEY, Ozyme, France) was also used to compare auto reactivity to an external standard extract in the same samples.

Determination of IgG and IgM levels

Total IgG and IgM. The total IgG and IgM were quantified by "Sandwich ELISA" [23]. Briefly, 96 flat-bottomed plates were coated with monoclonal antibodies directed against human IgG or IgM (5 µg/ml) and left for adsorption at 4°C overnight. Plates were washed 5 times with PBS-0.1% Tween 20 and blocked with PBS-1% Gelatin at 37°C for 1 hour. Wells were incubated with serum samples diluted at 1:100 in PBS-1% Gelatin-0.1% Tween 20 for 1 h at 37°C. Excess antibody was removed by 5 PBS-0.1% Tween 20 washings and then plates were incubated with peroxydase-labeled anti-human IgG and anti-human IgM (1:2000 in PBS-1% Gelatin, 0.1% Tween 20) at 37°C for 1 h. The assay was developed by adding the enzyme substrate (O-phenylenediamine diluted to 0.3 mg/ml in Phosphate-citrate buffer in the presence of hydrogen peroxide). After appearance of yellow color in negative wells, the reaction was stopped with 10% SDS. The OD was measured at 450 nm using an Emax ELISA plate reader and results were analysed by the Sofmax software.

Specific anti-parasite and anti-brain IgG and IgM

The anti *P. falciparum* and anti-brain IgG and IgM were analyzed by direct ELISA. Flat-bottomed 96 well plates were coated overnight at 4°C with 5 µg/ml parasite line (*FAN5HS*) or brain antigen. After washing, the plates were saturated with PBS-1% Gelatin for 1 hour at 37°C. Subsequently the sera were diluted (anti-parasite and anti-brain IgG 1/1000 and 1/500 respectively and IgM 1/500 and 1/500 respectively) in PBS-1% Gelatin, 0.1% Tween 20 and added in duplicate to the wells and incubated at 37°C for 1 hour. The plates were washed five times in PBS 0-1% Tween 20 and incubated for 1 hour at 37°C following the addition of peroxydase - conjugated human anti-IgG or anti-IgM (1:4000 and 1:2000 for anti-parasite and anti-brain respectively in PBS-1% Gelatin, 0.1% Tween 20). The process of revelation is the same as the total IgG and IgM.

Cytokine quantification

The levels of cytokines (IL1β, IL2, IL4, IL6, IL10, IL12, TGFβ, TNFα and IFNγ) in plasma were estimated by use of Opti-EIA kits (BD-Pharmingen); the results of which are already published earlier [32].

Immunobloting using PANAMA–blot method

Patterns of recognition of brain proteins by plasma IgG were detected by quantitative immunoblotting as described earlier [23], using a protein extract from the brain of a healthy individual as the source of antigens and normal RBC as control as described above. Briefly, normal human brain and RBC protein extracts (300 µg protein/gel) were separated by a standard SDS-PAGE in a 10% polyacrylamide gel. The proteins were transferred onto nitrocellulose membranes (Schleicher & Schuell, Dassel, Germany) by semi-dry electro transfer method (Pasteur Institute, Paris, France). Membranes were then incubated with patient plasma samples diluted 1:20 in PBS-0.1% Tween 20 (non-adjusted assay) in a Cassette Miniblot System (Immunetics, Cambridge, MA, USA). The immunoglobulin reactivities were detected by incubation with γ chain-specific secondary rabbit anti-human IgG coupled to alkaline phosphatase (Sigma-Aldrich, France). Revelation was done by using BCIP/NBT. As described [36] dried membranes were then scanned by a high resolution scanner (600 DPI) using an 8-bit linear grayscale. Subsequently, transferred proteins on the membranes were stained with colloidal gold (Protogold, British-BioCell, Cardiff, GB), and the stained membranes scanned again. Using colloidal gold staining, immunoreactivity profiles were adjusted for migration inequalities, so that equivalent immunoreactivities could be rescaled to equivalent positions on a common standard migration scale within and between membranes. Intensities were adjusted between membranes by a standard, consisting of a pool of serum from Gabonese CM patients [23] that was replicated twice on each membrane.

Protein identification by mass spectrometry

Briefly, human brain extract was separated on a 10% SDS-PAGE. After Coomassie staining, the band analogous to section 10 was cut and analyzed by peptide mass fingerprinting. Bands were excised from gels using ProPic Investigator (Genomic Solutions, Ann Arbor, MI, USA) and collected in 96-well plate. Destaining, reduction, alkylation, trypsin digestion of the proteins followed by peptide extraction were carried out with the Progest Investigator (Genomic Solutions, Ann Arbor, MI, USA). After desalting step (C18-µZipTip, Millipore) peptides were eluted directly using the ProMS Investigator, (Genomic Solutions, Ann Arbor, MI,USA) onto a 96-well stainless steel MALDI target plate (Applied Biosystems/MDS SCIEX, Framingham, MA, USA) with 0.5 µl of CHCA matrix (5 mg/ml in 70% ACN/30% H2O/0.1% TFA) [38]. *MS and MS/MS analysis*: Raw data for protein identification were obtained on the 4800 Proteomics Analyzer (Applied Biosystems/MDS SCIEX, Framingham, MA, USA) and analyzed by GPS Explorer 2.0 software (Applied Biosystems/MDS SCIEX, Framingham, MA, USA). For positive-ion reflector mode spectra 3000 laser shots were averaged. For MS calibration, autolysis peaks of trypsin ([M+H]+ = 842.5100 and 2211.1046) were used as internal calibrates. Monoisotopic peak masses were automatically determined within the mass range 800–4000 Da with a signal to noise ratio minimum set to 30. Up to twelve of the most intense ion signals were selected as precursors for MS/MS acquisition excluding common trypsin autolysis peaks and matrix ion signals. In MS/MS positive ion mode, 4000 spectra were averaged, collision energy was 2 kV, collision gas was air and default calibration was set using the Glu1-Fibrino-peptide B ([M+H]+ = 1570.6696) spotted onto fourteen positions of the MALDI target. Combined PMF and MS/MS queries were performed using the MASCOT search engine 2.1 (Matrix Science Ltd., UK) embedded into GPS-Explorer Software 3.5 (Applied Biosystems/MDS SCIEX, Framingham, MA, USA) on the NCBInr database (downloaded 2008 10 22, 7135729 sequences;2462332163 residues) with the

following parameter settings: species: homo sapiens, mono charged peptides, 50 ppm peptide mass accuracy, trypsin cleavage, one missed cleavage allowed, carbamidomethylation set as fixed modification, oxidation of methionines was allowed as variable modification, MS/MS fragment tolerance was set to 0.3 Da. Protein hits with MASCOT Protein score ≥65 and a GPS Explorer Protein confidence index ≥95% were used for further manual validation.

Antibodies absorption experiments

The 96 wells flat-bottomed microtiter plates (NUNC, Denmark) were coated with 5 µg/ml of beta tubulin (TBB), beta tubulin III (TBB3) and Glial Fibrillary Acidic Protein (GFAP) and left for adsorption at 4°C overnight. The assay was performed on these plates after blocking with PBS-1% Gelatin and washing with PBS-0.1% Tween 20. Briefly, coated wells were incubated with serum samples (Pool of CM and EC sera) or monoclonal antibody (mAb) anti-TBB3 as positive control diluted at 1:100 and 1:500 respectively in PBS-1% Gelatin, 0.1% Tween 20 for 1 hour at 37°C. Following incubation, wells were washed 5 times with PBS-0.1% Tween 20 and the plates were then incubated with peroxydase-labeled anti-human IgG (1:10000 in PBS-1%Gelatine, 0.1% Tween 20) and anti-mouse IgG at 37°C for 1 h respectively. Each supernatant was consecutively submitted 40 times to the same treatment. The process of revelation was identical to the one for total IgG and IgM.

Following depletion assays, all serum samples including mAb TBB3 and control non-depleted sera and mAb TBB3 were blotted on membranes containing human brain antigens separated on 10% SDS-PAGE. The immunoglobulin reactivities were detected by incubation with a γ chain-specific secondary rabbit anti-human IgG and rabbit anti-mouse IgG coupled to alkaline phosphatase (Sigma-Aldrich, France). Revelation was done by using BCIP/NBT and then dried membranes were scanned with a high resolution scanner (600 DPI).

Statistical analysis

Immunoblot data were analyzed by multivariate statistical methods, using IGOR software (Wavemetrics, Lake Oswego, OR), including specially written software packages. The standard migration scale was divided into sections around individual peaks of immunoreactivity. Section-wise absorbance values were subjected to principal component analysis (PCA), based on covariance calculation. For quantitative comparisons between groups, we used either Mann-Whitney (between two groups) or Kruskal-Wallis tests (>2 groups). Qualitative association was tested by Pearson's χ^2 test. The association between continuous quantitative parameters was assessed by linear regression, with the exception of correlations between two different types of parameters such as reactivity and cytokine profiles, which were tested by Spearman's rank correlation. The p values<0.05 were considered significant.

Correspondence analysis (pcc) was performed after singular value decomposition (SVD) of the different distance matrices. Inertia of the dimensions are expressed as percentages. The results of the decomposition of the principal dimensions are expressed as relative contributions of each variable, or the relative contribution of a arithmetic mean of a group of variables, to the principal dimension under study. Two-way complete-linkage hierarchical clustering (HC) based on Euclidean distances was used to analyze the relationship between the clinical groups and section cross-reactivity of the antibody preparations. SDV, pcc, PCA, and HC, as well as plotting of the results were performed using proprietary software.

Results

Demographic profiles of malaria patient groups

Ninety eight *P. falciparum* infected individuals were included in the present study. Selection according to the clinical variants shown that among, 16 patients corresponded to the mild malaria (MM), 10 to the severe non-cerebral malaria (SM) and 42 to CM. In controls, 5 individuals were classified in the ex-CM, 11 in the non-endemic controls (NEC) and 14 in the endemic controls (EC) groups. The demographic characteristics of each group are shown in the Table 1. Males and females were 65 and 33 respectively; a median age was 30 years (range 7–70). The NEC individuals were from a non-endemic area of *P. falciparum* and are individuals from laboratory staff that did not contact disease during at least the 5 preceding years. No parasitemia was detected in the EC, NEC or in the ex-CM groups at the time of inclusion in the study. There were no mixed infections. The median level of *P. falciparum* in the blood of patients from the infected groups (MM, SM and CM) was 1.5, 1 and 2 respectively but no statistical difference was observed between infected groups.

Total and specific IgM and IgG responses to *P. falciparum* and brain antigens according to clinical groups

We assessed the levels of total IgG and IgM in sera of the different groups of patient by ELISA. Interestingly, EC, NEC and ex-CM groups exhibit similar levels of total IgM and IgG. Thus, they were considered as a unique control group of non infected patients. Median levels of total IgM in MM, SM and CM patients were significantly higher than in controls ($p = 0.018$, 0.02 and 0.04 respectively). It is noteworthy that no significant difference was

Table 1. Demographic profiles of the *P. falciparum* malaria patient groups.

Groups	Patients no. (%)	Median age (range)	Median Parasitemia % (range)	Sex (M/F)
NEC	11 (11,2%)	35 (25–63)	0	10/1
EC	14 (14,3%)	26 (23–37)	0	13/1
MM	16 (16,3%)	30 (15–45)	1,5 (0,1 – 4,25)	9/7
SM	10 (10,2%)	30 (8–65)	1 (0,1 – 15)	7/3
CM	42 (42,9%)	36 (9–70)	2 (0,25 – 60)	22/20
ex-CM	5 (5,1%)	24 (7–60)	0	5/0
Total	98 (100%)	30 (7–70)	0,5 (0,1 – 60)	65/33

NEC- non endemic control, EC- endemic control, MM- mild malaria, SM- severe non-cerebral malaria, CM- cerebral malaria, ex-CM- Ex-cerebral malaria.
doi:10.1371/journal.pone.0008245.t001

observed between infected and control groups for total IgG levels (Figure 1A and 1B). Then, we measured the concentrations of specific IgG and IgM to *P. falciparum (FAN5HS)*. A slight but non-significant increase in the rate of specific IgG and IgM to parasite was observed in infected compared to non-infected groups. However, in CM group of patients, we observed a significant

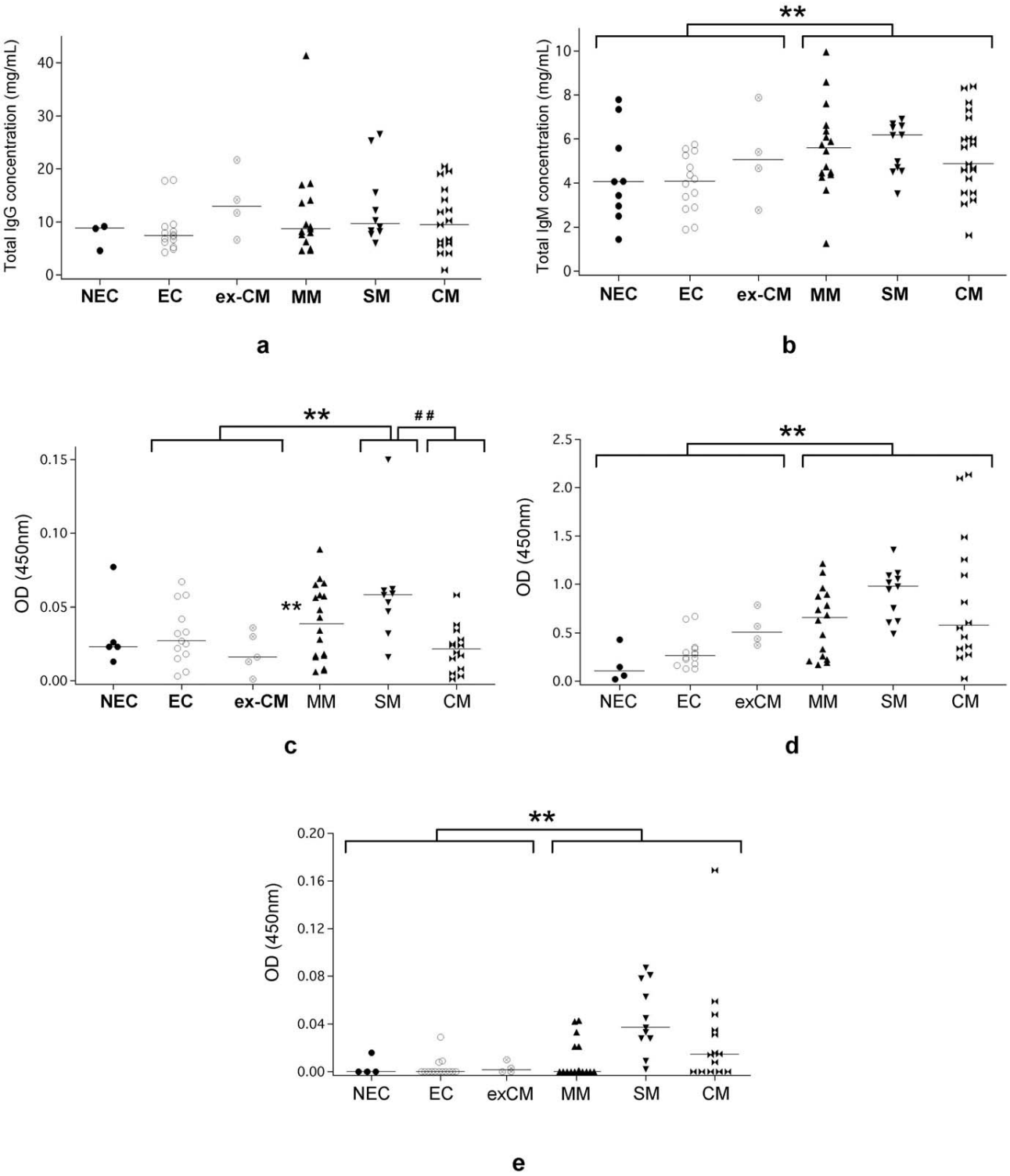


Figure 1. Total, brain-, and *P. falciparum*–specific IgG and IgM responses in different groups. Distribution of total levels of IgG (a) and IgM (b) in the different group of patients determined by Sandwich ELISA (** $p = 0.003$). Median level is indicated. Rate (optical density) of specific IgG against *P. falciparum* FAN5HS erythrocytic stage extract quantified by direct ELISA (** $p = 0.008$) (## $p = 0.001$) (c). Rate (optical density) of IgG (** $p < 0.001$) (d) and IgM (** $p = 0.002$) (e) recognizing the human brain extract quantified by direct ELISA.
doi:10.1371/journal.pone.0008245.g001

decrease of specific IgG to *P. falciparum* when compared to SM ($p = 0.001$) (Figure 1C).

Furthermore, when assessed the levels of IgG and IgM recognizing brain proteins in the different group of patients and controls, we found significant higher levels of antibody against brain proteins in infected groups than in the control ($p<0.001$ and $p = 0.002$ respectively) albeit their rates were significantly lower in the CM patients (Figure 1D and 1E).

Taken together, these data suggest that the efficiency to produce specific antibody response to either parasite or brain antigens is diminished in the CM patients group. Also, no significant correlations were observed between age, sex or parasitemia and the rate of total and specific IgG or IgM to *P. falciparum* and brain antigens and no relationship with disease severity and level of total or specific IgG or IgM to brain or to *P. falciparum* antigens.

We next used correspondence analysis to examine the relationship between the three types of specific antibody responses (total, parasite and brain) and clinical outcome. Correspondance analysis of the specific antibody responses measured in all patients reveals that the first principal component (pcc1, 55% inertia) separates the response to the parasite extract from the other two, and only the second principal component (pcc2, the remaining inertia) places the antibody response to the parasite roughly equidistant to both the brain and total IgG responses (Figure 2A). This result indicates that the production of antibodies against brain antigens is independent of this specific to the parasite. Decomposition of both dimensions and arithmetic averaging over patient groups reveals that pcc1 represents IgG response in CM and ex-CM, whereas pcc2 is in majority defined through response in ex-CM and EC groups (Figure 2B). Specially, the decomposition of pcc1 confirms that CM patients seem to develop lower measurable levels of antibody to parasite antigens, but broader in term of specificity as exemplified in the Figure 2C, than both SM and MM patients. Some of these *P. falciparum* specific antibodies are still detectable in ex-CM patients.

Analysis of the serum IgG autoantibody repertoire expressed against brain antigens in patients with distinct clinical forms of *P. falciparum* malaria

We first analyzed the reactivity patterns of IgG from the different groups of malaria patients to brain protein using PANAMA-BLOTs as previously described [23]. Reactivity against brain antigens expressed by Indian patients and by the standard consisting of a pool of serum from Gabonese CM patients are shown in Figure 3A. We found a high correlation between disease severity and an increased diversity of the repertoire of antigens recognized by circulating antibodies in *P. falciparum* patients. This was principally observed in CM patients. Those patients recognized more protein sections than the other groups of individuals tested. Healthy individuals completely lack reactivity against the brain extract. It is interesting to note that the link between CM pathology and the increase of IgG reactivity to brain antigens is reinforced by the low number of sections observed in the ex-CM patients (Figure 3B). These data are in agreement with our earliest observation in children from a hyperendemic area of Gabon [23]. Optical density analysis of the profiles of reactivity on the immunoblots allow defining peaks of density which corresponds to a section of brain protein recognized by IgG from a pool of sera of Gabonese CM children constituting our standard used for adjustment [23]. Profiles of antibody reactivities were separated into 18 sections as shown in the Figure S1. Next, profiles of reactivity from each patient group were compared by principal component analysis (PCA). In PCA, the components are identified in decreasing order of importance. Thus, by definition,

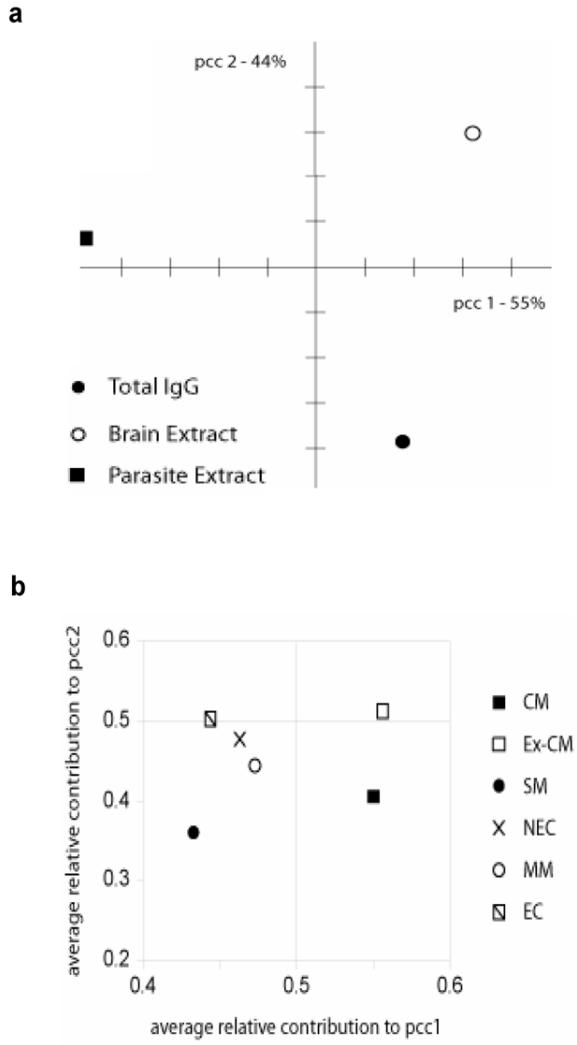


Figure 2. Correlation between total, brain-, and parasite-specific IgG responses. (A) Correlation-based principal component analysis of the levels of IgG responses to brain and *P. falciparum* FAN5HS antigens and total IgG determined by ELISA from all patients. (B) Relative average contributions of the different patient groups to the two principal components shown in (A). (C). Example of immunoblot showing the reactivity of IgG from patients sera from different groups with *P. falciparum* FAN5HS erythrocytic stage antigens.
doi:10.1371/journal.pone.0008245.g002

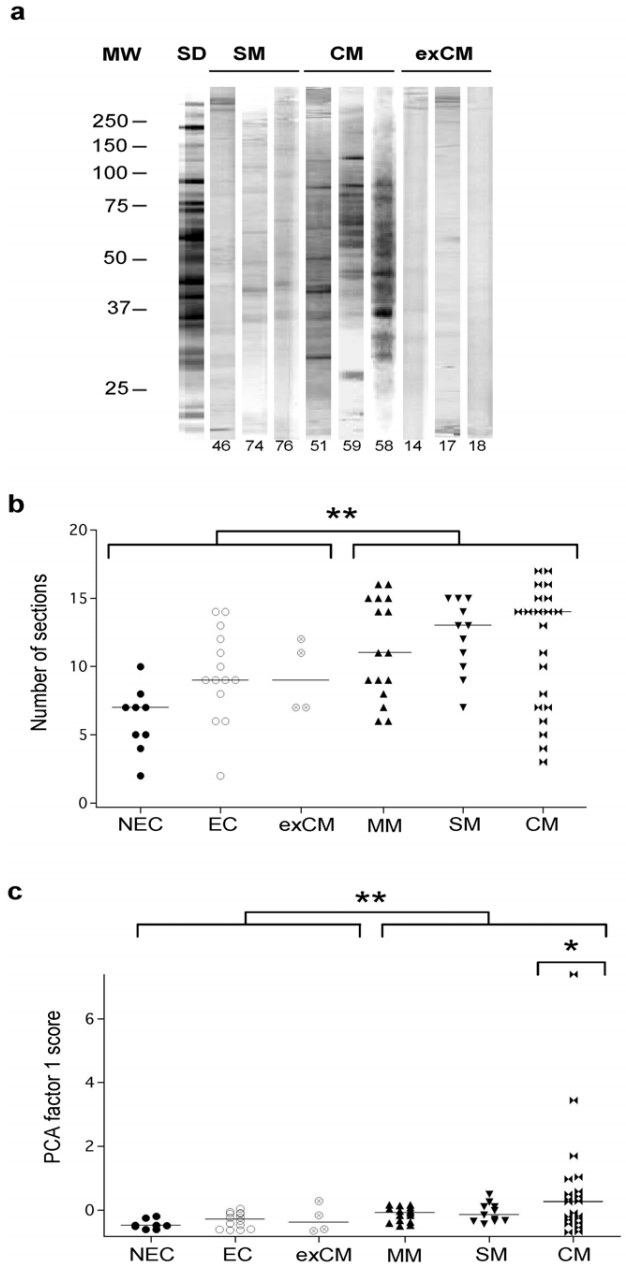


Figure 3. Profiles of IgG reactivities to brain antigens of the different *P. falciparum* infected groups. (A) Example of IgG reactivity from SM, CM, or ex-CM patients sera showing the increase with disease severity and the number of brain antigens (section) reacting with patient sera. (B) Median number of sections recognized by each patient from the different groups (** $p<0.001$). (C). PCA factor 1 score from unadjusted IgG reactiviy profiles. Groupwise distribution of PCA factor 1 scores. PCA1 score were significantly higher in infected than control groups (** $p<0.001$) and in CM than other groups (* $p = 0.01$).
doi:10.1371/journal.pone.0008245.g003

the first two components identified account for a large proportion of total reactivity. Factor 1 scores mostly reflected the recognition of one particular section and significantly higher in CM patients than the other groups with brain extract ($p = 0.01$) (Figure 3C). These results thus demonstrate a production of autoantibody to brain proteins in CM patients.

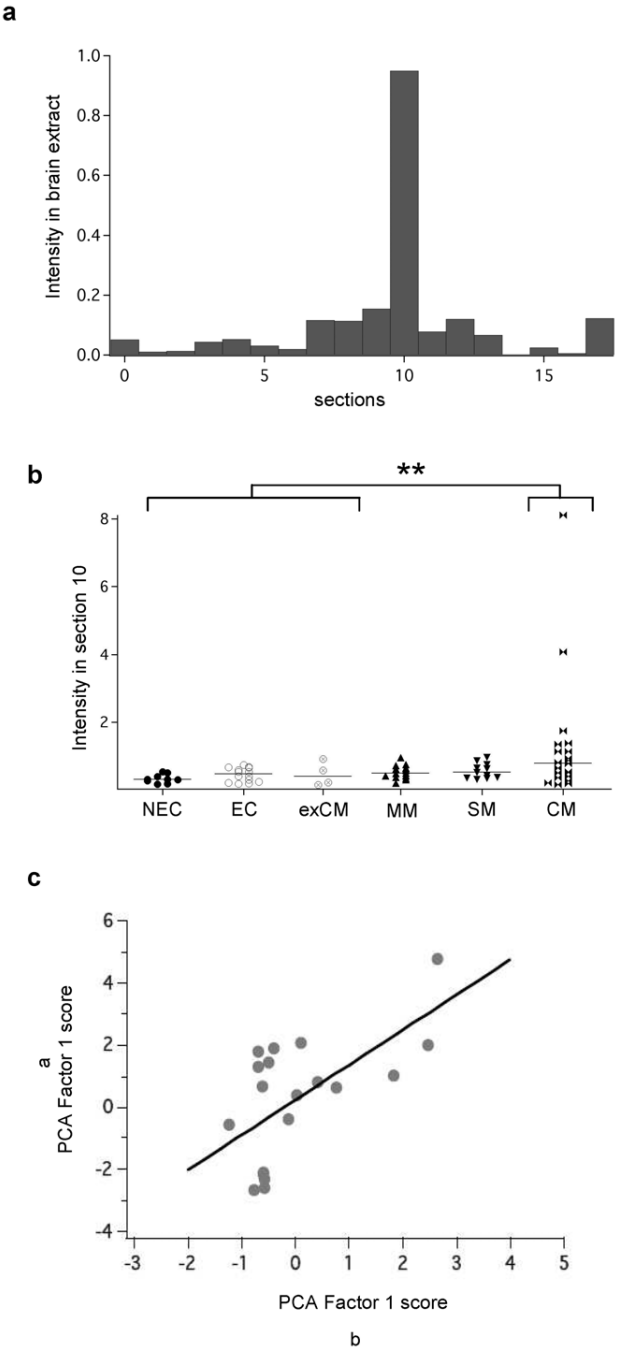


Figure 4. Reactivity to brain antigens of the different malaria patients group. (A) Distribution of mean intensity reactivities of IgG from CM patients with the different sections in brain extract. The section 10 is the most recognized among 18 sections. (B) IgG reactivity with section 10 is significantly higher in CM group than other groups (** $p = 0.006$). (C). Correlation of IgG reactvity with two different brain extracts. PCA factor 1 (α) correspond to the IgG reactivity from CM patients with Cuban healthy brain extracts (β) represent the reactivity of a commercial protein medley sample. Correlation coefficient: $R = 0.6237$, Regression: $p = 0.003$.
doi:10.1371/journal.pone.0008245.g004

In CM group, section 10 which corresponds to proteins of approximately 50 kDa, had maximum impact; more than 90% of total reactivity corresponds to PCA factor 1 (Figure 4A). In

addition, the mean intensity of serum IgG reactivity to brain against antigens of section 10 was significantly higher in CM patients than control groups ($p = 0.006$) (Figure 4B). In order to exclude alloreactivities, we used another source of brain antigens, the Medley brain protein extract. Results obtained were similar than those found with the previous extract. A high significant correlation between reactivities expressed by the different group of patients with the two brain antigen sources was calculated ($r = 0.623$, $p = 0.003$). This observation thus suggests that the IgG reactivity against brain proteins observed during malaria is irrespective of the brain donor (Figure 4C).

We also compared the average ability of sera from Indian and from our previous published Gabonese CM patients data to react with the same brain antigen extracts [23]. Interestingly, their profiles of reactivity to the brain extract overlaid suggesting that these different groups of patients originating from India or Gabon recognize the same spectrum of brain proteins. Nevertheless, the Gabonese CM group show a dominant reactivity with proteins of the section 0 while the Indian CM patients are distinguished by their predominant reactivity with proteins of the section 10 even if they also recognize section 0 (Figure S2).

To assess if the reactivity with section 10 in CM patients is specific to the brain tissue, we analysed the patterns of reactivity of same sera with RBC protein extract. An example of reactivity of patient serum IgG against RBC protein extract is shown in the Figure S3A. Statistical analysis reveal high differences between infected patients and controls ($p<0.001$) (Figure S3B). However, no statistical differences were observed when comparing groups of infected patients (MM, SM and CM). Profiles of reactivity of patient plasma samples with RBC proteins were separated into 20 sections according to the standard used for brain extract. However, PCA analyses allow us to identify the section 17 contributing to the difference between groups but not the section 10 identified in brain extract (Figure 5A). Moreover, the mean intensity of the IgG reactivity to section 17 was significantly higher in SM than control groups ($p = 0.002$) (Figure 5B). These results suggest that the IgG reactivity to antigens in brain section 10 could be a signature of CM patients. It is noteworthy that no significant correlation was observed between the reactivity profile to brain antigens represented by PCA factor 1 scores and age, parasitemia or sex. Besides, levels of total, brain or parasite specific IgG do not correlate with brain autoreactivity profiles.

Importantly, when arithmetically averaged over the patient groups, the reactivity of patient sera to the different sections on the blots can be used to classify the different patient groups using two-fold complete-linkage hierarchical clustering (Figure 6A). As would be expected, *P. falciparum* infected patients and the control groups form distinct clusters. Interestingly, among malaria affected groups, SM and MM are more closely related amongst each other than with CM. Also, it is worthy to note that the close relationship between EC and ex-CM sets them apart from the NEC group, indicating a possible contamination of the EC group with undetected ex-malaria cases. Correspondence analysis of IgG reactivity to brain antigens of the patient groups according to their average response to the eighteen different sections (Figure 6B) reveals that the second resulting principal component (representing 26% of total inertia) is almost solely responsible for separating the ex-CM and CM patient groups from the control and other *P. falciparum* infected groups. Decomposition of the first two principal components (Figure 6C) demonstrates that sections 10 and 17 account for the majority of pcc2 whereas section 17 has the least and section 10 the principal contribution to pcc1. Taken together, this dimensionality reduction analysis firmly establishes the predominance of section 10 and 17 in distinguishing CM from ex-CM and of section 10 in defining CM.

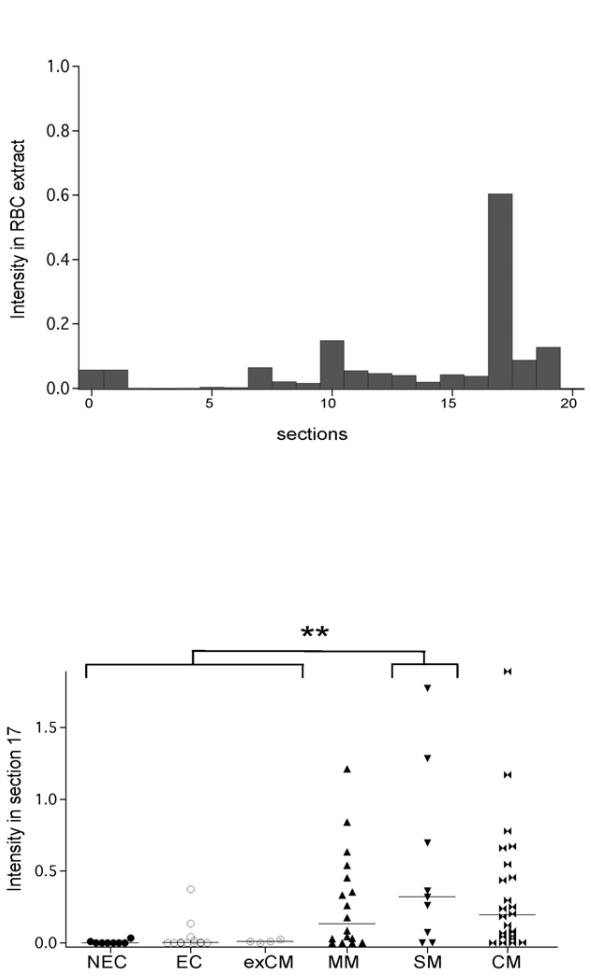


Figure 5. Reactivity to non-infected red blood cell antigens of different groups of patients. (A) Distribution of the mean intensity of IgG reactivity to RBC protein extract separated into sections (B) mean intensity of IgG reactivity to section 17 is significantly higher in SM than in the other groups (** $p = 0.002$).
doi:10.1371/journal.pone.0008245.g005

Note that the contributions of the individual sections are expressed as relative measures with the barycentre at (0,0). No significant contribution of any other section is observed. Therefore, IgG reactivity to antigens in section 10 and 17 could be biomarkers of ex-CM cases whereas IgG reactivity to section 10 could be used as a disease-marker for CM.

TBB3 is a major discriminant antigen recognized by serum IgG of CM patients

Furthermore, candidate proteins in section 10 were identified using mass spectrometry. In three independent experiments based on matching of peptide mass, the family of Beta Tubulin (TBB), in particular TBB3 specifically expressed in the brain and Glial Fibrillary Acidic Protein (GFAP) were identified as discriminant antigens using the Swiss-Prot database. However, due to the structural homologies between the several tubulin isotypes, it was not possible to distinguish by mass spectrometry if one or several isoforms of TBB were present in this section (Figure 7A). Therefore, additional analyses were performed to specifically analyze the involved tubulin isotypes. To validate the mass spectrometry results we further depleted sera samples from CM

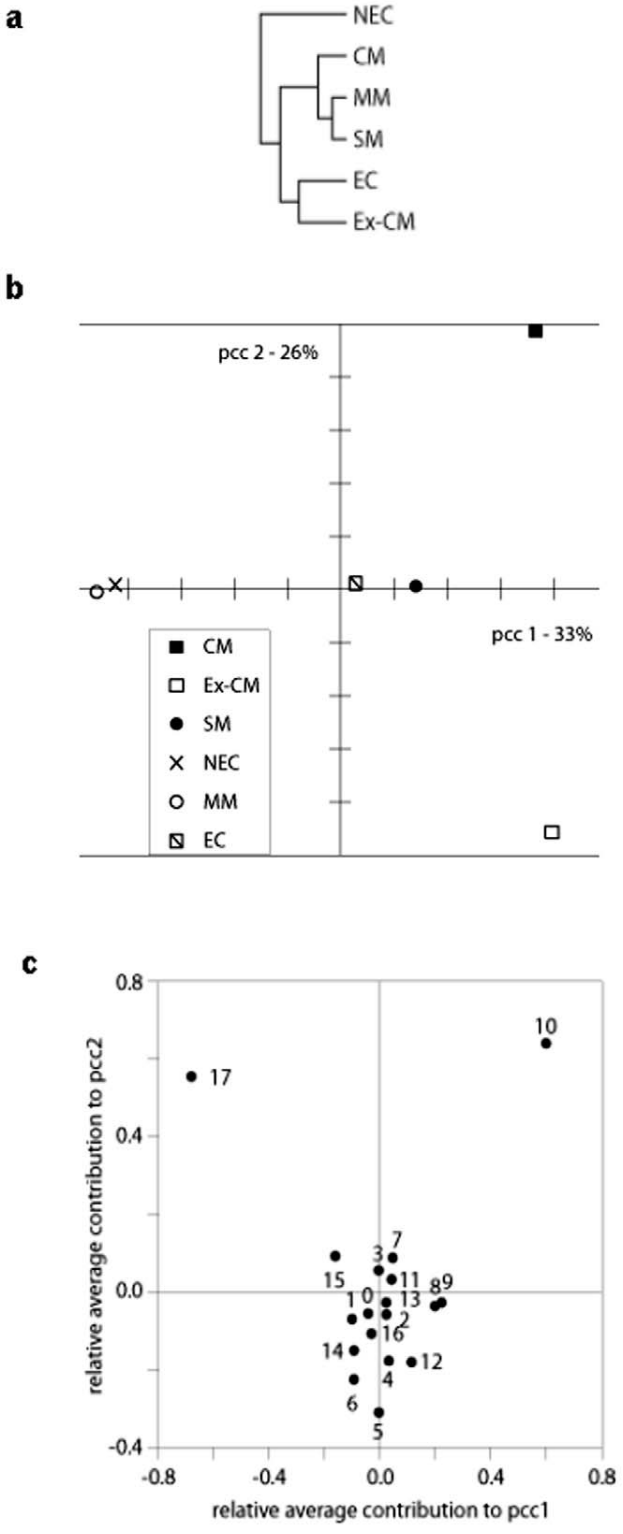


Figure 6. Reactivity to section 10 distinguishes cerebral malaria. (A) Hierarchical clustering of malaria patient groups according to their reactivity with brain proteins analyzed by PANAMA blots. (B) Correlation analysis of malaria patient group IgG reactivity. (C) Decomposition of correlation analysis.
doi:10.1371/journal.pone.0008245.g006

and EC with TBB3, TBB or GFAP proteins. After 40 rounds of depletion of 1 hour each, levels of antibodies recognizing TBB3, TBB or GFAP were quantified in the depleted samples by ELISA. A decrease of the level of specific antibodies with round number to the respective proteins was observed except for GFAP demonstrating that we can exclude GFAP as a candidate (data not shown). It is noteworthy that 40 rounds of depletion were necessary to remove TBB or TBB3 specific antibodies in CM sera whereas only 7 rounds were sufficient for EC samples. In addition, the recognition of section 10 by depleted serum samples was analysed by Western blot (Figure 7B). No signal at 46 kDa was detected in CM samples depleted with TBB3 or TBB after 40 rounds whereas a signal was still detectable when the membrane was blotted with GFAP depleted sera. In addition, no signal was seen in depleted EC sera and TBB3 monoclonal antibody (Figure 7B). Taken altogether these results indicate that TBB3 is a discriminant autoantigen targeted by IgG in CM patients.

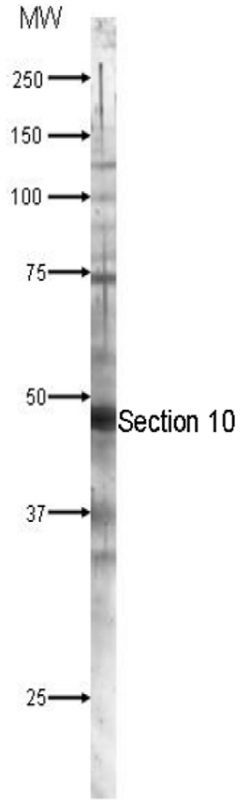
Relationship between IgG reactivity to brain proteins and cytokine activity patterns in *P. falciparum* malaria

Cytokines are thought to play an important role in malaria pathogenesis, particularly in CM (11). The relationship between the clinical severity of malaria and the complex pro- and anti-inflammatory cytokine network has been addressed in the same cohorts of *P. falciparum* infected patients and the results were previously published [32]. Among the 12 cytokines quantified, a coupled 2-way clustering and discriminant analyses allowed identification of a cluster cluster-II cytokines (IL1β, IL10, TNFα and TGFβ) that displays significantly increased ($p<10^{-6}$) activity in the CM group compared to other groups, and which can be used to differentiate between different clinical forms of malaria and control groups [32]. As we show here, similarly TBB3 autoantigen presence is a marker for CM and immune-reactivity of subjects can be used to classify different clinical forms of malaria as well as control groups. We therefore wanted to know whether or not both markers truly correlate and are surrogates. To this end we calculated Spearman rank correlations between cytokine reactivity and immune-activity towards the 18 sections on the blots for each subject in our patient cohorts. We considered the entire panel of cytokines used in our previous study [32]. When the resulting Spearman rank correlation distance matrix is singular value decomposed and analyzed for its principal covariance-based components, the cluster-II cytokine TGFβ, TNFα, IL10, and IL1β, also form a distinct cluster (Figure 8A). Therefore, the correlation between cluster-II cytokine levels and total immunoreactivity to the different sections on the immune-blots is sufficiently strong to allow distinction. The Spearman rank correlation for the cluster-II cytokine activities and the immune-reactivity to section 10 thereby is highly significant (Figure 8B). In conclusion, IgG reactivity to TBB3 in central Indian malaria patients is statistically significantly correlated with the cluster-II cytokine levels which we had previously shown to be a marker for cerebral malaria in the same population.

Discussion

Recently, we have shown autoantibodies to α-II spectrin of the brain in the serum of Gabonese *P. falciparum* infected children with CM [23]. Nevertheless, the exact nature of this response remains elusive. Considering the multifactorial character of malaria, the purpose of the present study was to validate these findings by generalizing our analysis to an Indian population with different genetic background, endemic and environmental status. The presence of autoantibodies in malaria patients has long been

a



Protein name	NCBI Accession number	Protein MW (kDa)	Protein Mascot score
GFAP protein	38566198	49.8	165
glial fibrillary acidic protein	16265836	49.8	163
glial fibrillary acidic protein variant	62896925	49.8	153
hypothetical protein	57997051	50.3	138
glial fibrillary acidic protein, isoform CRA_b	119571953	34.1	123
glial fibrillary acidic protein, isoform CRA_c	119571954	47.7	115
TUBB2C protein	14124960	26.0	94
tubulin, beta 2B, isoform CRA_c	119575532	33.3	93
tubulin beta	223486	49.8	89
tubulin, beta 4	21361322	50.0	148
TUBB protein	16198437	30.6	150
TUBB3 protein	38014278	46.0	82

b

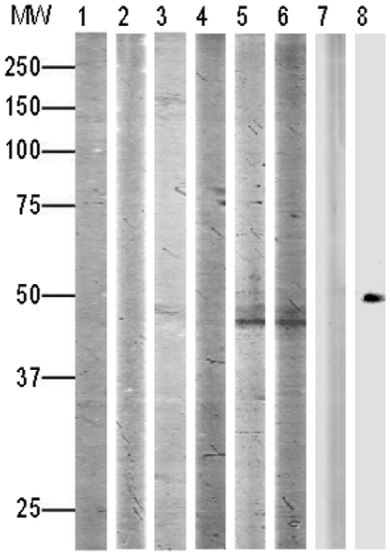


Figure 7. Identification and characterization of proteins contain in the section 10. (A) Protein identification by mass spectrometry. Twelve human proteins identified from the section 10. Identification of proteins was carried out as described (see material and methods). In this case, Mascot protein scores greater than 65 are significant ($p<0.05$). (B). Characterization of section 10 protein candidate by antibody depletion. Immunoprinting with sera depleted (d) or not depleted (nd) with TBB, TBB3, and GFAP proteins. No signal was detected at 46 kDa in CM sera depleted with TBB and TBB3 proteins in lane 3 and 4 respectively. MW, Molecular weight marker; 1, EC sera (nd); 2, EC sera (d); 3, CM sera (d) with TBB; 4, CM sera (d) with TBB3; 5, CM sera (d) with GFAP; 6, CM sera (nd); 7, TBB3 mAb (d) with TBB3; 8, TBB3 mAb (nd).
doi:10.1371/journal.pone.0008245.g007

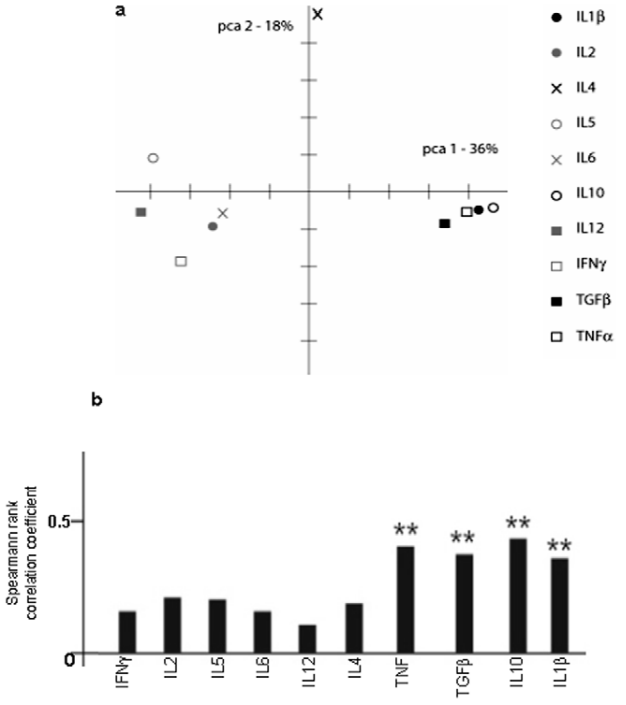


Figure 8. Correlation of total IgG reactivity against brain with cytokine profiles. (A) Principal component analysis (PCA) of the Spearman rank correlations between cytokine levels and IgG reactivity towards the eighteen sections of the Panama blots for each subject in our patient cohorts. We considered the entire panel of cytokines used in our previous study (32). (B) Distribution of Spearman rank correlations of PCA factor 1 scores between IgG self-reactivity in CM patients to brains proteins and the levels of cytokines from cluster 1 (IL2, IL5, IL6, IL12, and IFN-γ) and 2 (IL1β, IL10, TNFα, and TGFβ) quantified by ELISA.
doi:10.1371/journal.pone.0008245.g008

recognized but their role in the pathophysiology of CM is very little explored and not defined [22,23,28].

We have used a global approach aiming not only at studying the individual components involved, but also the complex interactions between these components, in order to elucidate the global nature of autoantibody response to brain antigens produced in patients with different clinical spectra of malaria. The population studied was from Gondia in the central India where *P. falciparum* malaria is epidemic [39]. In the groups of patients studied, the most severe form of the disease is developed for the greater part in 30-year-old adults on average. This could mainly due to the fact that the majority of these patients were seasonal workers staying in the region of Gondia only during the periods of harvest. Gondia is a zone where the spread of *P. falciparum* is rather recent. Most of the patients studied developed their first *P. falciparum* malaria episode and do not present a mixed infection. In our population of study and, in agreement with previous reports, the parasitemia rate alone was not enough to evaluate the severity of the disease since it was equivalent in CM than in SM and MM groups [40].

Polyclonal B cell stimulation through parasite mitogens coupled with the secretion of parasite specific antibodies can explain the higher amount of total antibodies observed in infected compared to control groups [25,26,28]. Similar observations have been made when analysing the autoantibody response to brain antigens among the various groups of patients. The group of CM presents the lowest specific IgG and IgM reactivities to brain proteins while those are increased when infected versus non infected group of patients are compared. As demonstrated and confirmed by the correspondence analysis, the antibody-mediated immune response to brain proteins detected in *P. falciparum* infected patients seems to be mainly due to a selective and inducible process during the infection. The disappearance of these autoantibodies in the ex-CM group reinforces this hypothesis. As revealed by pcc1 factor (Figure 1), the antibody response to brain antigens is largely independent of the parasite specific response. In addition, there is no relationship with disease severity and total antibody levels, neither with specific IgG or IgM to brain or to *P. falciparum* antigens. Overall, these observations suggest that the spontaneous autoantibody production against the brain during malaria carries the hallmark of a typical immune response induced by parasite infection.

Interestingly, reactivity to all 18 sections of the brain extract with circulating IgG from the different individuals of the cohorts is sufficient to comprehensibly cluster the different patient groups as demonstrated by the hierarchical clustering analysis. Decomposition analysis reveals that reactivity with section 10 in CM patients is mainly responsible for this classification capacity. Brain specificity of the IgG response to section 10 in CM patients has been demonstrated by the lack of reactivity to the same section when normal RBC protein extract has been used as antigen. Thus, IgG reactivities against human brain and RBC extracts strongly suggest that the development of autoimmune antibodies is more noticeable in patients who develop CM than in the other group of malaria patients.

This study highlights the important finding of the increase of the repertoire of brain antigens recognized by IgG of Indian CM patients. These results validate and extended our previous observations in Gabonese patients [23]. They also strongly support the hypothesis that an antibody-mediated self-reactivity to brain antigens triggered during *P. falciparum* infection is associated with cerebral malaria. However, we do not know yet if this antibody response is an aggravating factor that contributes to the development of cerebral malaria or is one of the consequences of the syndrome. Nevertheless, on the contrary to the Gabonese CM patients mostly characterized by an autoantibody response directed to α-II spectrin, Indian CM patients showed strong reactivity with the human brain proteins TBB3 identified by mass spectrometry in section 10. It is noteworthy that only some Indian CM patients recognize the α-II spectrin. This observation indicates a particularity of IgG self-reactive response to brain proteins in the Indian population. The correlation of the profiles of reactivity of CM patients to two different brains extracts point out that the profile of IgG reactivity to brain cannot be explained by an alloreactive response. Opposite to the observations made in the study with *P. falciparum* infected children from Gabon, no correlation was found between the age, the sex, parasitemia and concentration of total IgG, and the IgG autoreactivity to brain antigens.

TBB3 is a cytoskeleton protein, which is abundant in the central and peripheral nervous systems (CNS and PNS) and expressed during fetal and postnatal development. In adult tissues, TBB3 is mainly expressed in the brain and PNS and used as a neuron-specific marker molecule encoded by a gene located at the long arm of chromosome 16 in man [41] thereby highlighting a possible pathogenic role between such autoimmune response and the occurrence of CM. In support of this interpretation, no significant increase in anti-tubulin antibody levels was seen in sera of patients infected with *Plasmodium vivax* or with tuberculosis [42]. However, the level of serum anti-tubulin antibodies was significantly elevated during infectious diseases such as visceral or cutaneous leishmaniasis, onchocerciasis, schistosomiasis and leprosy, but it is unknown if such autoantibodies involve a reaction against TBB3 [42].

A two-way coupled cluster analysis revealed 2 clusters of cytokines relevant to clinical subgroups of disease in the same cohorts of malaria patients studied [32]. In particular, the significant abundant level of cluster-II cytokines (TGFβ, TNFα, IL10 and IL1β) was relevant to the discrimination of CM from SM. Importantly, we have shown that cluster-II cytokine levels strongly correlate with reactivity to TBB3 in CM. The fact that we have been able to classify the different malaria and control groups based on the statistically significant IgG reactivity to TBB3 associated with cluster II cytokines despite the relatively small size of our cohort, demonstrates the prevalence of this autoantibody-mediated reactivity in CM and therefore its clinical relevance.

To summarize, the IgG response against TBB3 found in CM could be a new biomarker of CM in the Indian population. While the molecular mechanisms of antibody production to TBB3 during *P. falciparum* infection remain unknown, the study of this phenomenon potentially leads to new avenues in the understanding of malaria physiopathology. Despite these findings, a longitudinal study of malaria clinical states, in conjunction with studies of cytokine production, specific and self-reactive antibody responses and several other biological parameters on largest populations from endemic and epidemic areas of India would appreciably add to our understanding of the role of immune responses in general in disease severity associated with *P. falciparum* infection. We have here established the basis for such a deeper investigation.

Supporting Information

Figure S1 Determination of sections. Localizations of the bands on Western blot profile of different groups obtain after the computer analysis of membrane N19 and sections are defined using the IgG reactivity of standard (pool of Gabonese CM patients). Bands are ordered from high to low molecular weight (between about 230 kDa and 20 kDa).
Found at: doi:10.1371/journal.pone.0008245.s001 (1.05 MB TIF)

Figure S2 Comparison of IgG reactivities within different clinical groups with section 0. The mean intensity of IgG reactivity in different groups of patients with section 0 (* $p = 0.012$) (** $p = 0.018$).
Found at: doi:10.1371/journal.pone.0008245.s002 (1.04 MB TIF)

Figure S3 Profiles of IgG reactivity in different clinical groups of patients with RBC extract. (a) A blot represents increase IgG immunoreactivity in CM patients than others (b) Groupwise distribution of PCA factor 1. The PCA1 score was significantly higher in infected than control groups (** $p<0.001$)
Found at: doi:10.1371/journal.pone.0008245.s003 (1.17 MB TIF)

Acknowledgments

We thank Pr. Shobhona Sharma, Pr. Laurent Rénia, and Dr. David Dombrowicz for critical reviews of the manuscript. We also thank Pr. Monique Capron for scientific advice and for providing facility and Jacques Roland for fruitful discussion.

Author Contributions

Conceived and designed the experiments: PD PAC GCM SP. Performed the experiments: DB FH PL JCR AN. Analyzed the data: DB FH PL CB VG JCR AN CF AB SP. Contributed reagents/materials/analysis tools: PD CB VG IdM RJ GCM CF CF AB. Wrote the paper: DB FH PAC GCM CF CF AB SP.

References

1. Korenromp E (2004) World Malaria Report: Roll Back Malaria. World Health Organization: Geneva.
2. Breman JG, Egan A, Keusch GT (2001) The intolerable burden of malaria: a new look at the numbers. Am J Trop Med Hyg 64: iv–vii.
3. Mazier D, Nitcheu J, Idrissa-Boubou M (2000) Cerebral malaria and immunogentics. Parasite Immunol 22: 613–23.
4. Miller LH, Baruch DI, Marsh K, Doumbo OK (2002) The pathogenic basis of malaria. Nature 415: 673–9.
5. Greenwood BM (1997) The epidemiology of malaria. Ann Trop Med Parasitol 91: 763–9.
6. Idro R, Jenkins NE, Newton CR (2005) Pathogenesis, clinical features, and neurological outcome of cerebral malaria. Lancet Neurol 4: 827–840.
7. Adams S, Brown H, Turner G (2002) Breaking down the blood-brain barrier: signaling a path to cerebral malaria? Trends Parasitol 18: 360–6.
8. Pongponratn E, Riganti M, Punpoowong B, Aikawa M (1991) Microvascular sequestration of parasitized erythrocytes in human *falciparum* malaria: a pathological study. Am J Trop Med Hyg 44: 168–75.
9. Gallien S, Milea D, Thiebaut MM, Bricaire F, Le Hoang P (2007) Brain and optic nerve ischemia in malaria with immune disorders. J Infect 54: 1–3.
10. Medana IM, Turner GD (2007) *Plasmodium falciparum* and the Blood-Brain Barrier-Contacts and Consequences. J Infect Dis 195: 921–923.
11. Hunt NH, Grau GE (2003) Cytokines: accelerators and brakes in the pathogenesis of cerebral malaria. Trends Immunol 24: 491–9.
12. Schofield L, Grau GE (2005) Immunological processes in malaria pathogenesis. Nat Rev Immunol 5: 722–35.
13. Grau GE, Taylor TE, Molyneux ME, Wirima JJ, Vassalli P, et al. (1989) Tumor necrosis factor and disease severity in children with *falciparum* malaria. N Engl J Med 320: 1586–1591.
14. Rénia L, Potter SM, Mauduit M, Rosa DS, Kayibanda M, et al. (2006) Pathogenic T cells in cerebral malaria. Int J Parasitol 36: 547–54.
15. Vigário AM, Gorgette O, Dujardin HC, Cruz T, Cazenave PA, et al. (2007) Regulatory CD4+ CD25+ Foxp3+ T cells expand during experimental *Plasmodium* infection but do not prevent cerebral malaria. Int J Parasitol 37: 963–73.
16. Freeman RR, Parish CR (1978) Polyclonal B-cell activation during rodent malarial infections. Clin Exp Immunol 32: 41–45.
17. Daniel-Ribeiro C, Druilhe P, Monjour L, Homberg JC, Gentilini M (1983) Specificity of auto-antibodies in malaria and the role of polyclonal activation. Trans R Soc Trop Med Hyg 77: 185–188.
18. Shaper AG, Kaplan MH, Mody NJ, McIntyre PA (1968) Malarial antibodies and autoantibodies to heart and other tissues in the immigrant and indigenous peoples of Uganda. Lancet 1: 1342–1346.
19. Adu D, Williams DG, Quakyi IA, Voller A, Anim-Addo Y, et al. (1982) Anti-ssDNA and antinuclear antibodies in human malaria. Clin Exp Immunol 49: 310–316.
20. Jakobsen PH, Morris-Jones SD, Hviid L, Theander TG, Hoier-Madsen M, et al. (1993) Anti-phospholipid antibodies in patients with *Plasmodium falciparum* malaria. Immunology 79: 653–657.
21. Consigny PH, Cauquelin B, Agnamey P, Comby E, Brasseur P, et al. (2002) High prevalence of co-factor independent anticardiolipin antibodies in malaria exposed individuals. Clin Exp Immunol 127: 158–164.
22. Soni PN, De Bruyn CC, Duursma J, Sharp BL, Pudifin DJ (1993) Are anticardiolipin antibodies responsible for some of the complications of severe acute *Plasmodium falciparum* malaria? S Afr Med J 83: 660–662.
23. Guiyedi V, Chanseaud Y, Fesel C, Snounou G, Rousselle JC, et al. (2007) Self-reactivities to the non-erythroid alpha spectrin correlate with cerebral malaria in Gabonese children. PLoS ONE 2: e389. doi:10.1371/journal.pone.0000389.
24. Lang B, Newbold CI, Williams G, Peshu N, Marsh K, et al. (2005) Antibodies to voltage-gated calcium channels in children with *falciparum* malaria. J Infect Dis 191: 117–121.
25. Minoprio P (2001) Parasite polyclonal activators: new targets for vaccination approaches? Int J Parasitol 31: 588–91.
26. Greenwood BM (1974) Possible role of a B-cell mitogen in hypergammaglobulinaemia in malaria and trypanosomiasis. Lancet 1: 435–6.
27. D'Império Lima MR, Alvarez JM, Furtado GC, Kipnis TL, Coutinho A, et al. (1996) Ig-isotype patterns of primary and secondary B cell responses to *Plasmodium chabaudi chabaudi* correlate with IFN-gamma and IL-4 cytokine production with CD45RB expression by CD4+ spleen cells. Scand J Immunol 43: 263–70.
28. Daniel-Ribeiro C, Druilhe P, Monjour L, Homberg JC, Gentilini M (1983) Specificity of auto-antibodies in malaria and the role of polyclonal activation. Trans R Soc Trop Med Hyg 77: 185–8.
29. Touré FS, Ouwe-Missi-Oukem-Boyer O, Bisvigou U, Moussa O, Rogier C, et al. (2008) Apoptosis: a potential triggering mechanism of neurological manifestation in *Plasmodium falciparum* malaria. Parasite Immunol 2008 Jan; 30(1): 47–51.
30. Tripathi AK, Sha W, Shulaev V, Stins MF, Sullivan DJ Jr (2009) *Plasmodium falciparum* infected erythrocytes induce NF-{kappa} B regulated inflammatory pathways in human cerebral endothelium. Blood 2009 Aug 27.
31. Jouhilahti EM, Peltonen S, Peltonen J (2008) Class III beta-tubulin is a component of the mitotic spindle in multiple cell types. J Histochem Cytochem 56: 1113–9.
32. Prakash D, Fesel C, Jain R, Cazenave PA, Mishra GC, et al. (2006) Clusters of cytokines determine malaria severity in *Plasmodium falciparum*-infected patients from endemic areas of Central India. J Infect Dis 194: 198–207.

33. Duarte J, Deshpande P, Guiyedi V, Mécheri S, Fesel C, et al. (2007) Total and functional parasite specific IgE responses in *Plasmodium falciparum*-infected patients exhibiting different clinical status. Malar J 6: 1.
34. WHO (1990) Severe and complicated malaria. Trans R Soc Trop Med Hyg 84: 1–65.
35. Bradford MM (1976) A rapid and sensitive method for the quantitation of microgram quantities of protein utilizing the principle of protein-dye binding. Anal Biochem 72: 248–54.
36. Haury M, Grandien A, Sundblad A, Coutinho A, Nobrega A (1994) Global Analysis of Antibody Repertoires. 1. an Immunoblot Method For the Quantitative Screening of a Large Number of Reactivities. Scand J Immunol 39: 79–87.
37. Nobrega A, Haury M, Grandien A, Malanchere E, Sundblad A, et al. (1993) Global Analysis of Antibody Repertoires. II. Evidence for Specificity, Self-Selection and the Immunological Homunculus of Antibodies in Normal Serum. Eur J Immunol 23: 2851–2859.
38. Saveanu C, Namane A, Gleizes PE, Lebreton A, Rousselle JC, et al. (2003) Sequential protein association with nascent 60S ribosomal particles. Mol Cell Biol 23: 4449–4460.
39. Kumar A, Valecha N, Jain T, Dash AP (2007) Burden of malaria in India: retrospective and prospective view. Am J Trop Med Hyg 77: 69–78.
40. Gendrel D, Kombila M, Martz M, Nardou M, Lecointre C, et al. (1992) Parasitemia in *Plasmodium falciparum* malarial attacks in children. Presse Med 21: 1805–8.
41. Katsetos CD, Herman MM, Mörk SJ (2003) Class III beta-tubulin in human development and cancer. Cell Motil Cytoskeleton 55: 77–96.
42. Howard MK, Gull K, Miles MA (1987) Antibodies to tubulin in patients with parasitic infections. Clin Exp Immunol 68: 78–85.

.

G

Journal article : HMGA1-dependent and independent 7SK RNA gene regulatory activity.

Sebastian Eilebrecht, Christophe Bécavin, Hélène Léger, Bernd-Joachim Benecke, and Arndt Benecke.

Submitted to RNA Biology journal in October 2010.

Abstract

The small nuclear 7SK RNA negatively controls transcription by inactivating positive transcription elongation factor b (P-TEFb) and is an integral component of TAT-dependent and independent HIV-1 transcription initiation complexes. 7SK RNA has recently been shown to also directly control HMGA1 transcription activity. HMGA1 is a master regulator of gene expression and its deregulation is associated with virtually any type of human cancer. The degree of HMGA1 overexpression thereby correlates with tumor malignancy and metastatic potential. 7SK snRNA directly interacts through its loop 2 (7SK L2) with the first A/T DNA binding hook of HMGA1. We have developed several 7SK L2 RNA chimera with the Epstein Barr Virus expressed RNA 2 (EBER2) to target HMGA1 function in transcription regulation. The efficiency of interfering with HMGA1 transcription activity by the chimeric 7SK L2-EBER2 fusions by large exceeds the efficiency of 7SK wild-type RNA due to the stronger EBER2 promoter activity. Furthermore, the 7SK L2-EBER2 chimera do not interfere with P-TEFb controlled transcription elongation or the formation of 7SK sn/hnRNPs. The comparison of the effects of wild-type 7SK RNA on cellular transcriptome dynamics with those induced by the two 7SK L2 mutants as well as the changes in gene expression following inhibition of HMGA1 allow the identification and characterization of HMGA1-dependent and independent effects of 7SK snRNA. We furthermore also present evidence for P-TEFb and HMGA1-independent 7SK RNA L2 regulatory activity.

HMGA1-dependent and independent 7SK RNA gene regulatory activity.

Sebastian Eilebrecht[1], Christophe Bécavin[1], Hélène Léger[1], Bernd-Joachim Benecke[2], and Arndt Benecke[1] *

[1]Institut des Hautes Études Scientifiques & CNRS USR3078; 35, route de Chartres; 91440 Bures sur Yvette; France.

[2]Department of Biochemistry; Ruhr University Bochum; Universitätsstr. 150; 44780 Bochum; Germany.

*To whom correspondence should be addressed: Tel: +33 1 60 92 66 65; Fax: +33 1 60 92 66 09; Email: arndt@ihes.fr

Abstract

The small nuclear 7SK RNA negatively controls transcription by inactivating positive transcription elongation factor b (P-TEFb) and is an integral component of TAT-dependent and independent HIV-1 transcription initiation complexes. 7SK RNA has recently been shown to also directly control HMGA1 transcription activity. HMGA1 is a master regulator of gene expression and its deregulation is associated with virtually any type of human cancer. The degree of HMGA1 overexpression thereby correlates with tumor malignancy and metastatic potential. 7SK snRNA directly interacts through its loop 2 (7SK L2) with the first A/T DNA binding hook of HMGA1. We have developed several 7SK L2 RNA chimera with the Epstein Barr Virus expressed RNA 2 (EBER2) to target HMGA1 function in transcription regulation. The efficiency of interfering with HMGA1 transcription activity by the chimeric 7SK L2 — EBER2 fusions by large exceeds the efficiency of 7SK wild-type RNA due to the stronger EBER2 promoter activity. Furthermore, the 7SK L2 — EBER2 chimera do not interfere with P-TEFb controlled transcription elongation or the formation of 7SK sn/hnRNPs. The comparison of the effects of wild-type 7SK RNA on cellular transcriptome dynamics with those induced by the two 7SK L2 mutants as well as the changes in gene expression following inhibition of HMGA1 allow the identification and characterization of HMGA1-dependent and independent effects of 7SK snRNA. We furthermore also present evidence for P-TEFb and HMGA1-independent 7SK RNA L2 regulatory activity.

Introduction

The small nuclear (sn) 7SK RNA belongs to the highly abundant non-coding regulatory RNAs in eukaryotic cells and is synthesized by RNA polymerase III (1-4). It has been shown to be a negative regulator of the transcription elongation reaction of RNA polymerase II by inhibiting the positive transcription elongation factor b (P-TEFb) (5,6) and thus to act as a global repressor for the expression of genes transcribed by RNA polymerase II. Beyond that, 7SK RNA is involved in the transcription regulation of diverse viral pathogens via its interaction with P-TEFb. During early HIV-1 transcription, 7SK RNA is released by the Tat protein and the TAR RNA from the inactive P-TEFb complex, resulting in efficient transcription of viral genes (7,8). Recent studies have shown that the amount of inactive 7SK/P-TEFb snRNP can be controlled by the human T-Lymphotropic Virus Type 1 (HTLV-1) Tax protein during HTLV-1 infection (9).

We recently discovered a second major role of this small nuclear RNA in the negative regulation of the function of the high mobility group protein HMGA1 by directly competing with DNA for binding to the first A/T-hook motif of the protein (10). The interaction of 7SK RNA with HMGA1 is mediated by the second stem-loop structure of the RNA (L2), which is not involved in the formation of the P-TEFb snRNP or 7SK hnRNP complexes (10-12). The main functions of HMGA1 is the regulation of gene expression consist in altering chromatin-structure, changing the DNA binding-affinity of transcription factors, or modifying nucleosome positions (13-16). Depending on the position of the target sequence within the corresponding gene, HMGA1 activates or inactivates gene expression. The majority of genes affected in their expression by the interaction between 7SK RNA and HMGA1 have been shown to be repressed by HMGA1 (10). HMGA1 over-expression has been detected in virtually every type of cancer and moreover, HMGA1 has been ascribed a role as a major factor during cancerogenesis, repressing the expression of tumor repressors on the one hand and enhancing the expression of oncogenes on the other (17-21). Given these key functions during malignant transformation, HMGA1 has been considered as a promising drug target for cancer therapy (22). Due to the negative regulatory role of 7SK RNA on HMGA1 function, the over-expression of 7SK RNA, or rather of the HMGA1 binding 7SK L2 substructure, has to be considered as a potential novel therapeutic approach to control HMGA1. Here we report on two important observations regarding the control of gene expression by 7SK snRNA. First, we have investigated the transcriptome dynamics induced by the overexpression of 7SK wild-type RNA and

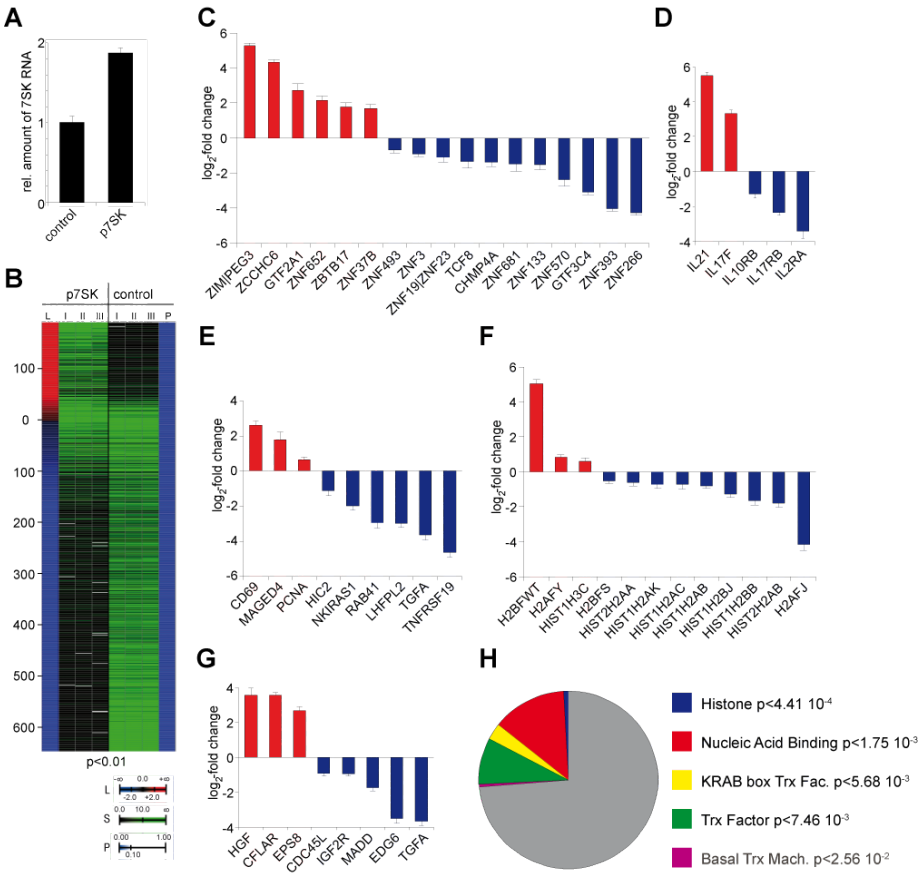


Figure 1: Changes in gene expression upon over-expression of full length 7SK snRNA.
(A) HEK293 cells were transfected with p7SK coding for full length human 7SK RNA controlled by the full 7SK gene promoter. The amount of 7SK RNA was quantified by RT-qPCR analyzes in comparison to mock transfected control cells.
(B) Heat-map of the subset of statistically significant ($p<0.01$) changes in gene expression between HEK293 cells over-expressing 7SK snRNA (p7SK) compared to mock transfected cells (control). L indicates the $\log_2$-fold change in expression, S indicates the signal, P indicates the p-value and the roman numbers indicate independent biological replicates.
(C) Genes related to transcription, that show a significantly increased (red) or decreased (blue) expression upon over-expression of full length 7SK RNA compared to the control.
(D) Interleukin (IL) related genes, that show a significant differential expression upon over-expression of full length 7SK RNA compared to the control.
(E) Genes related to cancer, that show a significant differential expression upon over-expression of full length 7SK RNA compared to the control.
(F) Histones, that show a significant differential expression upon over-expression of full length 7SK RNA compared to the control.
(G) Genes related to cell growth, differentiation or apoptosis, which are significantly differentially expressed upon over-expression of 7SK RNA compared to the control.
(H) Gene ontology enrichment analysis for 'molecular functions'. The graph represents the fraction of genes corresponding to one of the statistically significantly (binomial distribution) enriched ontologies compared to the entire set of 7SK RNA target genes.

compared those to our previous (10) and additional transcriptome profiles obtained with different 7SK L2 — EBER2 chimera in order to dissect 7SK RNA activity on HMGA1 from its activity on P-TEFb regulation. Using signaling network inference methodology we reconstruct and compare 7SK RNA, HMGA1, and 7SK L2 dependent changes on the inferred signalosome activity. Second, by comparing the transcription regulatory properties of two 7SK L2 mutants with both wild-type L2 and the entire 7SK RNA we identify a yet unknown activity of the 7SK L2 substructure in transcription regulation, which seems also HMGA1-independent.

Results

7SK snRNA induced transcriptome dynamics.

While we had previously focused our attention to the characterization of the specific role of 7SK RNA in regulating HMGA1-dependent chromatin dynamics and transcription regulation (10), we aimed now at understanding the relative contribution of HMGA1 to the regulatory activity of 7SK snRNA.

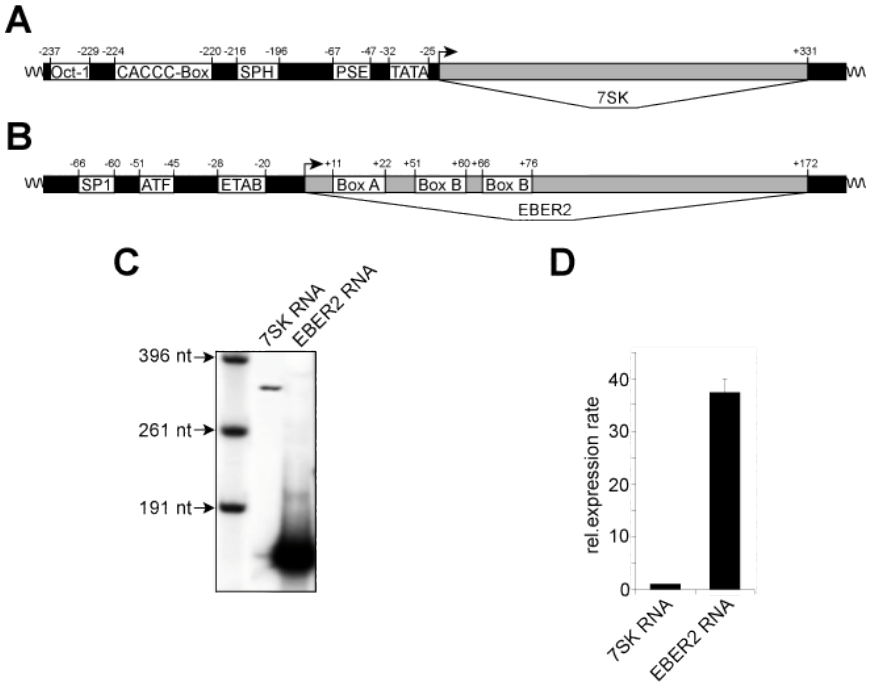


Figure 2: Comparison of the viral EBER2 gene promoter with the 7SK gene promoter
(A) Schematic representation of the 7SK gene promoter: The entirely gene external 7SK promoter contains an Octamer motif (Oct-1), a CACCC-Box, a Sph Post Octamer Homology (SPH) domain as well as a Proximal Sequence Element (PSE) and a TATA-Box. The positions of each element with respect to the transcription start (+1; arrow) is annotated. The coding sequence for 7SK RNA is highlighted in grey.
(B) Schematic representation of the viral EBER2 gene promoter: The gene external promoter elements consist of a Sp1 transcription factor- (SP1) and an Activating Transcription Factor (ATF)-binding site as well as an EBER TATA box (ETAB). The gene internal elements are Box A and Box B regulatory sequences. The positions of each element with respect to the transcription start (+1; arrow) is given. The coding sequence for EBER2 RNA is highlighted in grey.
(C) The transcription rate of 7SK RNA was compared to that of EBER2 RNA in *in vitro* transcription analyses using HeLa nuclear extract.
(D) The relative expression rate of EBER2 RNA was calculated in comparison to that of 7SK RNA using ImageJ software. The expression rate of 7SK RNA was arbitrarily set to 1.

We first set out to globally identify genes whose expression is impacted by over-expression of full length human 7SK RNA in HEK293 cells. We therefore expressed full length human 7SK RNA controlled by the wild type 7SK promoter in these cells and compared the downstream transcriptome with the gene expression profile of mock-transfected (empty pUC18) control cells. RT-qPCR analyses revealed a significant ($p<0.01$) over-expression of 7SK RNA (about 190%), when compared to the mock-transfected control (Figure 1A). Transcriptome analyses of these cells, using three independent biological replicates for either condition, revealed a total of 844 genes regulated by 7SK RNA in a statistically significant manner ($p<0.01$). 194 (or 23%) of these genes showed an enhanced expression after 7SK over-expression, whereas 650 genes (or 77%) were repressed in their expression (Figure 1B). Notably, the genes, whose expression was up-regulated upon over-expression of 7SK RNA were significantly enriched with HMGA1 target genes (10). Among the global set of genes regulated by 7SK RNA we were able to classify certain subgroups with regard to the cellular functions of the corresponding gene products. Genes products having known activities in mRNA transcription regulation (Figure 1C), genes related to Interleukins (IL) (Figure 1D), genes related to oncogenesis (Figure 1E), genes coding for different histones (Figure 1F) and genes, whose products are involved in cell growth, differentiation and apoptosis (Figure 1G) are regulated by full-length 7SK RNA. Notably, the interleukin receptor 2 alpha (IL2-Rα) gene, which is found to be significantly down-regulated upon 7SK RNA over-expression (Figure 1D), is also known to be positively regulated by HMGA1 (10, 16,23-29), agreeing with the idea of 7SK RNA as a negative regulator of HMGA1 function. These observations are also in perfect agreement with gene ontology enrichment analyses. We identify as sole biological process enriched in the 7SK RNA target genes 'mRNA transcription' ($p<2.53x10E-3$, binomial distribution). Furthermore, the enrichment analysis for molecular function identifies DNA binding proteins including histones and transcription factors as statistically significantly enriched (Figure 1H).

The 7SK gene promoter belongs to the RNA polymerase III type 3 promoters and is located entirely upstream of the transcription start point. It consists of an Octamer motif (Oct-1) (30-32), a CACCC-Box (33), a Sph Postoctamer Homology (SPH) domain (34,35) as well as a Proximal Sequence Element (PSE) (36) and a TATA-Box (Figure 2A). This promoter efficiency results in about 2x10E5 copies of 7SK RNA per individual cell in HeLa cells (2), making 7SK RNA one of the most abundant small nuclear RNAs in eukaryotic cells. Hence, it is no surprise that the over-expression of 7SK RNA controlled by its own promoter, as done here, results only in an about two-fold increase of expression of this RNA (Figure 1A), limiting the effectiveness of the approach. Nevertheless, 7SK RNA intra-cellular concentrations are apparently tightly controlled, and the over-expression of exogenous 7SK RNA sufficiently impacts on the dynamic equilibrium between bound and free 7SK RNA, as we are able to identify several hundreds of 7SK RNA target genes in the transcriptome profiling (Figure 1B).
As 7SK RNA acts as a global repressor of RNA polymerase II transcription elongation mediated by the stem-loop structures L1, L3 and L4 (11,12), the expression of full length 7SK RNA will target both, HMGA1 function and P-TEFb function. In order to better characterize and dissect the relative impact on these two distinct 7SK RNA target activities, we next set out to compare the wild-type 7SK RNA induced transscriptome dynamics to the previously reported effects on as well as newly generated 7SK substructure chimera-induced transcriptome dynamics.

7SK --- EBER2 chimera and EBER2 background activity.

One of the strongest promoters known among the genes transcribed by RNA polymerase III is the type 2 promoter of the EBER2 gene, which codes for the 172 nt long EBER2 RNA of the Epstein Barr Virus (37). It contains gene external as well as gene internal promoter elements, both being essential for efficient transcription by RNA polymerase III. To the former elements belong a binding site for the Sp1 transcription factor, an Activating Transcription Factor (ATF) binding site (38) as well as the EBER TATA-Box (ETAB) (Figure 2B). The gene internal promoter elements comprise A- and B-Box regulatory sequences (39,40) (Figure 2B). The efficient transcription of the EBER2 gene results in more than 10^7 copies of EBER2 RNA per infected cell (37).
We compared the transcription efficiency of the EBER2 gene with that of the 7SK gene by *in vitro* transcription analyses with HeLa nuclear extract (Figure 2C). The quantified signals reveal a more than 35-fold higher transcription rate of EBER2 RNA compared to 7SK RNA (Figure 2D), mirroring roughly the cellular copy numbers of the two transcripts determined in earlier studies (2,37). The high expression rate, as well as the composite secondary structure of EBER2 RNA make it a potentially very potent tool to over-express small hairpin-RNA structures. The EBER2 RNA mainly consists of a basal stem and the two stem-loop structures La and Lb (Figure 3A). The gene internal promoter elements of the EBER2 gene are located exclusively in the basal stem as well as the stem-loop structure La (Figure 3A), predestining the stem-loop structure Lb to be exchanged by the hairpin-RNA structure of interest.
We have previously constructed a chimeric RNA, in which the sequence from +81nt to +140nt of wild type EBER2 RNA was exchanged by the sequence from +113nt to +154nt (L2) of wild type 7SK RNA (EBER2 7SK L2, Figure 3D) (10). We were able to show that this chimera maintains the HMGA1 binding activity of wild type 7SK L2 RNA as well as the promoter activity of the wild type EBER2 gene (10). Transcriptome analyses following over-expression of this RNA chimera compared to a mutated 7SK L2 chimeric construct with a reduced HMGA1 binding activity (EBER2 7SK L2 m137, Figure 3E) revealed the regulation of a wide variety of HMGA1-controlled genes (10). A second mutated RNA chimera discussed here (EBER2 7SK L2 m122/m137, Figure 3F) presumably restores the 7SK L2 secondary structure, but does not restore HMGA1 binding activity when compared to EBER2 7SK L2 m137 or EBER2-7SK L2 (10). This finding indicates that HMGA1 binding is contingent on the actual sequence of the L2 of 7SK

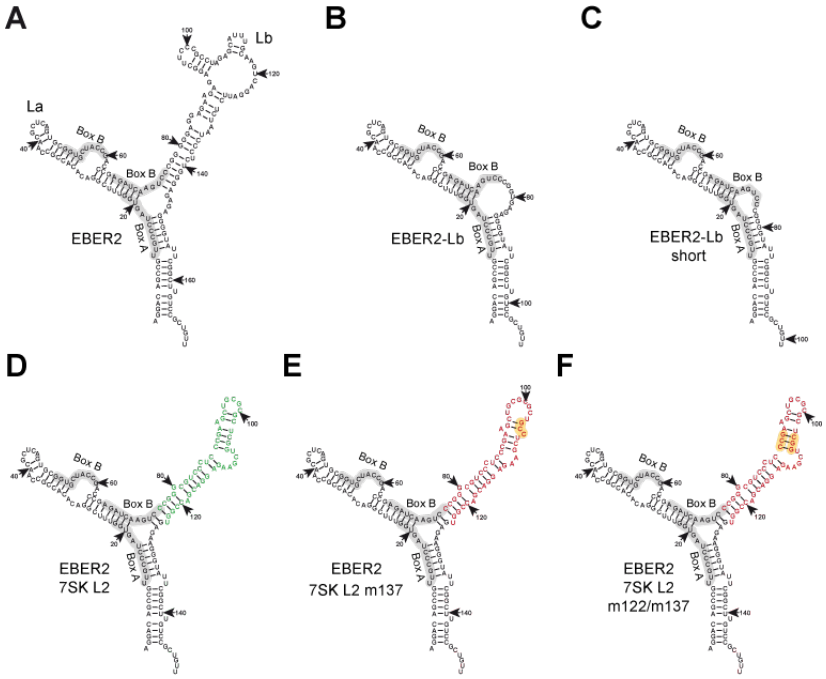


Figure 3: Secondary structures of EBER2 RNA, truncated control RNAs and EBER2 7SK L2 RNA chimera.
The secondary structure of each RNA was calculated using RNAstructure 4.4 software (45). Gene-internal box A and box B elements of the strong viral EBER2 promoter are highlighted in grey.
(A) The secondary structure of wild type EBER2 RNA consists of a basal stem and two stem-loop structures (La and Lb).
(B) In the EBER2-Lb RNA nt79 to nt141 of wild type EBER2 RNA have been deleted.
(C) A shortened version of the EBER2-Lb RNA is lacking nt79 to nt148 of wild type EBER2 RNA.
(D) In the EBER2 7SK L2 RNA chimera the Lb (nt80 to nt140) was exchanged by the loop 2 structure (nt113 to nt154) of wild type 7SK RNA. The 7SK sequence is marked in green.
(E) A mutated version of the EBER2 7SK L2 RNA chimera (EBER2 7SK L2 m137) carries a three nucleotide (CGG to GCC) mutation at position 104nt to 106nt (yellow), which corresponds to nt137 to nt139 of wild type 7SK RNA. The mutated 7SK sequence is marked in red.
(F) A second mutant of the EBER2 7SK L2 RNA chimera (EBER2 7SK L2 m122/m137) contains a three nucleotides mutation (GGC to CCG) at position nt89 to nt91 in addition to the mutation in the EBER2 7SK L2 m137 RNA chimera in order to restore the predicted secondary structure of the 7SK L2. The mutated 7SK sequence is marked in red here as well.

RNA rather than only the (predicted) secondary structure since otherwise the combined EBER2 7SK L2 m122/m137 mutant would restore binding to wild-type L2 levels (Figure 3D & 3F).

As we have previously reported, EBER2 as well as EBER2-Lb have profound effects by themselves on the cellular transcriptome (41), calling for very careful analysis and correction of the transcriptome activity of EBER2-based chimera. In the hope to construct EBER2-Lb backbones with even further reduced background activity, we designed an additional control RNA lacking the EBER2 Lb substructure (EBER2-Lb-short, Figure 3C, compare to 3B). This Lb deleted EBER2 mutant, however, also contains all gene internal promoter elements. In comparison to the wild type EBER2 RNA the longer control RNA (EBER2-Lb; Figure 3B) has been used previously to distinguish effects of the basal stem and the stem-loop structure La from those of the stem-loop structure Lb of EBER2 RNA, both having drastic effects on gene expression (41). We have thus here investigated and compared the effects of both EBER2-Lb and EBER2-Lb-short on the cellular transcriptome using as reference the effects of EBER2 wild-type RNA. As reported previously (41), EBER2-Lb has strong effects on the cellular transcriptome and acts quite differently than EBER2 alone, resulting in the statistically significant differential expression of some 3500 genes (Figure 4A). The mechanism by which EBER2 and EBER2-Lb interfere with cellular transcription or the measurement itself are at this point completely obscure. When comparing the effects of the new EBER2-Lb-short construct to EBER2 wild-type in five independent biological replications to the overlap of the effects with the corresponding comparison of EBER2-Lb and EBER2, two observations can be made. First, EBER2-Lb-short also has significant effects on the measured transcriptome when compared to EBER2 wt. Second, these effects are quite distinct from those induced by EBER2-Lb as the overlap is minimal (Figure 4A). Taken together, the EBER2-Lb-short construct confirms the not understood but highly reducible effects of EBER2 expression on the determination of the transcriptome, and

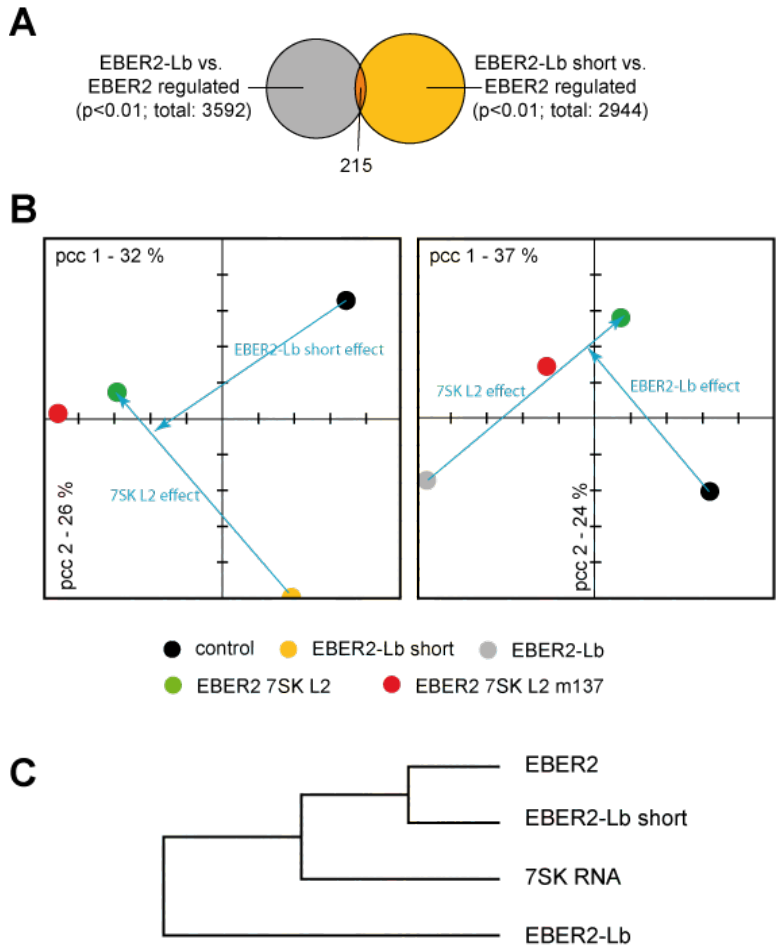


Figure 4: Comparison of the effects of EBER2-Lb and EBER2-Lb-short on the cellular transcriptome of HEK293 cells.(A) Venn diagram of EBER2-Lb versus wild-type EBER2 and EBER2-Lb-short versus wild-type EBER2 target genes at p<0.01. (B) Principal component analysis in correlation space (PCC) of the weighted averages of the transcriptome profiles from the indicated biological conditions. The inertia of the two first principal components are depicted along with their relative contributions. The arrows illustrate the effects of EBER2-Lb, EBER2-Lb-short, and the added effect of the 7SK L2 RNA expression. (C) Euclidean distance-based hierarchical complete linkage clustering of the variance-weighted averages of the indicated transcriptome profiles.

furthermore, also confirms that different EBER2 deletion mutants can have significantly different effects.

In order to decide which of the two EBER2-Lb back-bone constructs was best suited for the characterization of 7SK RNA substructure chimera, we directly compared their gene transcription activity to the ones previously reported for EBER2-7SK L2, the EBER2-7SK L2 m137 (both generated in the context of the EBER2-Lb) including control transfected cells (pUC18). The variance-weighted average transcriptome profiles calculated from the three to five biological replicates of each biological condition were visualized using principal component analysis in correlation space (PCC, Figure 4B). The arrows superposed onto the PCC results indicate the independent contributions of the EBER2 backbones and the 7SK L2 effects. In case of the EBER2-Lb, the effect of the 7SK L2 m137 mutant is intermediary to the 7SK L2 wild-type versus EBER2-Lb control, which is in accordance with our previous observations of the 7SK L2 m137 mutant being still able to target HMGA1 albeit with lowered affinity (10). Furthermore, Euclidean distance-based hierarchical clustering using a complete linkage method (Figure 4C), identifies the EBER2-Lb-short transcriptome effects to be much closer related to the EBER2 wild-type than EBER2-Lb and wild-type 7SK RNA. It is for these two observations that we have decided that the EBER2-Lb-short backbone is less well suited than the previously constructed EBER2-Lb. Consequently, for all following analyses, the EBER2-Lb has been used as control or as backbone for the different 7SK substructure chimera. The investigation of the different effects of EBER2 and EBER2 internal deletion constructs on the cellular transcriptome and its determination, however, will need to be thoroughly investigated further especially in any analysis of EBER2-based chimeric RNAs.

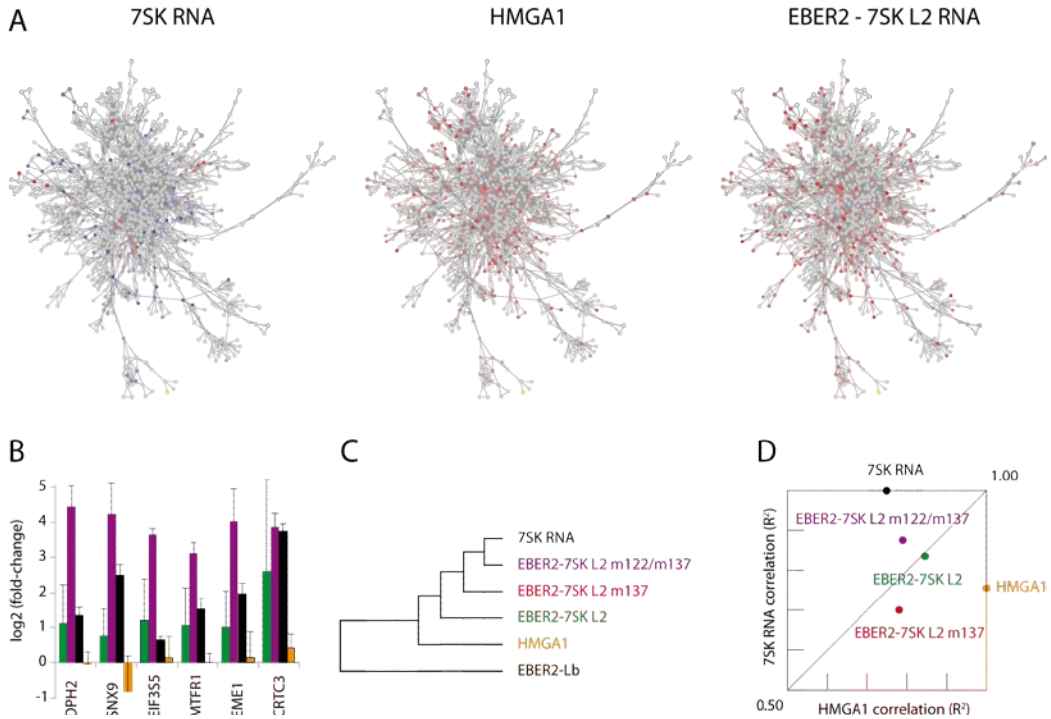


Figure 5: Inferred signalosome networks for wild-type 7SK RNA, HMGA1, and 7SK L2. (A) The gene product networks inferred from the transcriptome profiles using available PANTHER, Kegg, and GO annotation were depicted for wild-type 7SK RNA, HMGA1, and 7SK L2 using Cytoscape software (http://www.cytoscape.org/). Predicted activity of gene-products in the three networks is illustrated by a blue (diminished activity) to grey (unchanged activity) to red (increased activity) linear color gradient. Only connected components are drawn. (B) Examples of genes up-regulated by EBER2-7SK L2, EBER2-7SK L2 m122/m137, wild-type 7SK RNA, but not HMGA1 are shown. Logarithmic (base 2) fold-changes are given. (C) Euclidean distance-based hierarchical complete linkage clustering of the variance-weighted averages of the indicated biological conditions. The Euclidean distances were computed only for those genes identified to be regulated by HMGA1 and 7SK RNA. (D) Pearson correlation analysis of the variance-weighted averages of the indicated biologic conditions with respect to the 7SK RNA and HMGA1 average transcriptome profiles. Only the +0.50 to +1.00 space is shown, and the diagonal indicated.

Comparison of the inferred gene encoded networks downstream of 7SK RNA, HMGA1, and EBER2-7SK L2.

Large-scale gene expression information can be used to infer the effective networks of the presumably expressed gene products under the, arguably important hypothesis, that the gene expression changes translate linearly to changes in cellular protein levels. We have used the available gene product annotation information of the Kegg, PANTHER, and GO databases to reconstruct the HEK293 protein network and infer its activity based on the changes in the transcriptome dynamics following expression of wild-type 7SK RNA, as described here, as well as the previously reported HMGA1 and EBER2-7SK L2 dependent activities (Figure 5A) (10). The inferred networks have then been visualized using Cytoscape software (51) and colored using a linear gradient from blue (reduced activity) to grey (unchanged activity) to red (induced activity). Note that only the largest spanning and fully connected tree is depicted. As becomes immediately obvious from the reconstructed network and the inferred activities, the HMGA1 and EBER2-7SK L2 signalosomes are virtually identical in both the identity of the regulated nodes as well as their degree of inferred activity. This is in complete agreement with our previous proposition of all HMGA1 target genes being also 7SK RNA targets (10). The comparison of those two networks with the one inferred from wild-type 7SK RNA activity, however, reveals significant differences with respect to both the identity of regulated components as well as their level or even sign of regulation (Figure 5A). This is expected, as expression of full-length 7SK RNA also should display activity on P-TEFb target genes (5,6) in addition to HMGA1 target genes. The over-expression of wild-type 7SK RNA under control of its own promoter being far less efficient than the expression of the EBER2-7SK L2 chimera, dampen the overall effects of 7SK RNA when compared to the chimera. Note that we have also reconstructed the EBER2-7SK L2 m137 inferred signalosome activity. As expected, and in complete agreement with our previous gene-based analyses (10), it resembles closely the EBER2-7SK L2 and HMGA1 networks (data not shown, see also Figure 7). Given the fact that the EBER2-7SK L2 m137 mutant does not abrogate the effects of the 7SK L2 substructure

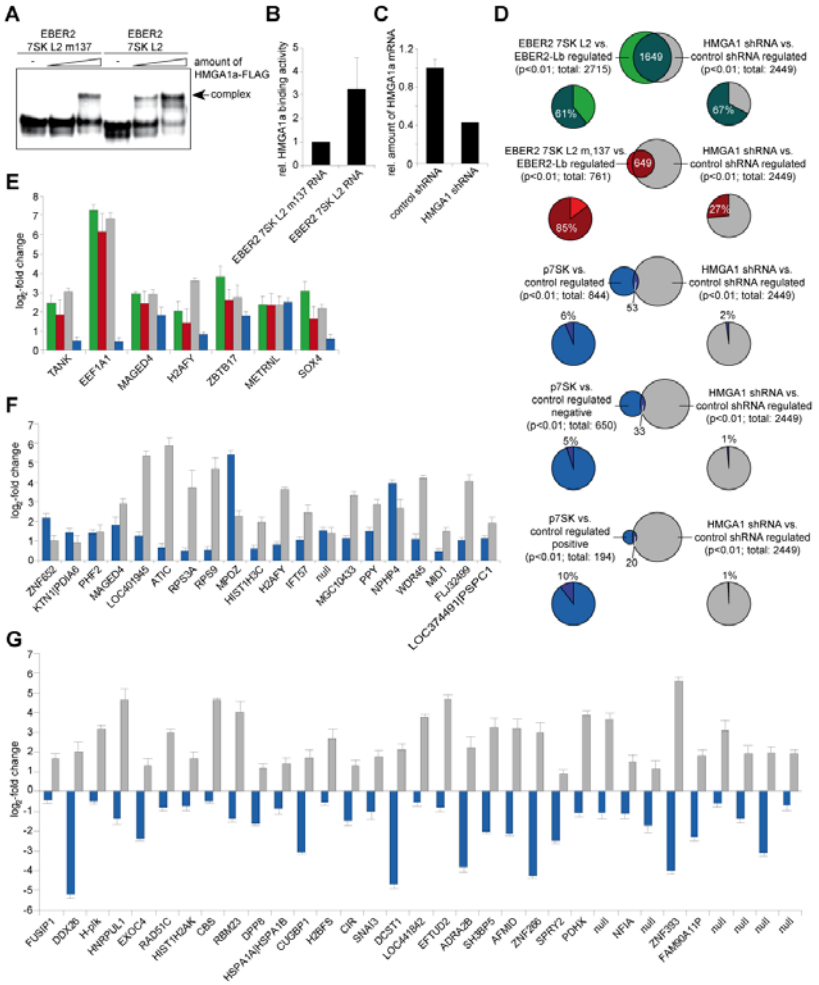


Figure 6: EBER2 7SK L2 RNA chimera are a potent tool to affect HMGA1-dependent gene expression. (A) 2ng of 32P-labeled EBER2 7SK L2 m137, EBER2 7SK L2 and wild type EBER2 RNA were incubated with increasing amounts of immuno-purified HMGA1a-FLAG fusion protein (approx. 5ng and 10ng). The resulting complexes were analyzed in EMSAs compared to the corresponding RNAs alone (-). (B) The HMGA1-binding activity of EBER2 7SK L2 RNA and EBER2 7SK L2 m137 RNA from (A) was calculated using ImageJ. The activity of EBER2 7SK L2 m137 RNA was set arbitrarily to 1. (C) The expression of endogenous HMGA1 in HEK293 cells was knocked down by shRNA in comparison to non-targeting control shRNA using a vector that allows the expression of both shRNAs under the control of the 7SK RNA promoter. The knockdown of HMGA1 mRNA was verified by quantitative PCR analyses. * indicates $p<0.05$ in a student's t-test. The error bars show the s. d.. (D) The genes expressed differentially in a significant manner ($p<0.01$) after over-expression of the EBER2 7SK L2 and EBER2 7SK L2 m137 RNA chimera in comparison to the over-expression of EBER2-Lb RNA, which are shown in green and red, respectively, were compared to the genes with a significant ($p<0.01$) change in expression after shRNA.mediated knockdown of HMGA1 (grey). The fraction of genes significantly differentially expressed in both conditions is shown in dark green (dark red). The genes significantly ($p<0.01$) differentially expressed after full length 7SK RNA over-expression (blue) were compared to the genes affected by shRNA-mediated knockdown of HMGA1 (grey). The common subset between these conditions is shown in dark blue. The genes differentially expressed after full length 7SK RNA over-expression are divided into negatively and positively regulated genes and compared to the genes affected by shRNA-mediated knockdown of HMGA1. The common subsets of all comparisons are shown as percentage of the genes affected by each condition in total as well as by the shRNA-mediated knockdown of HMGA1 (grey). (E) Different genes, regulated by all four conditions shown in (C) The error bars show the s. d.. (F) Genes, whose expression is increased upon both, 7SK RNA over-expression and HMGA1 knockdown. (G) Genes, whose expression is decreased upon 7SK RNA over-expression and increased after HMGA1 knockdown.

on HMGA1-dependent transcription control, we decided to also determine the transcriptome dynamics induced by the expression of the EBER2-7SK L2 m122/m137 double mutant (Figure 3). This mutant, as discussed above, preserves the predicted secondary structure of the 7SK L2 substructure, while changing the nucleotide sequence of the L2 on two blocks of three consecutive positions. The HEK293 transcriptome has thus been recorded in four independent biological replications following the over-expression of this double mutant (see Materials and Methods).

Evidence of HMGA1-independent 7SK L2 activities on target genes.

When comparing the gene expression profiles obtained using the EBER2-7SK L2 m122/m137 double mutant with those of the other EBER2 chimera, wild-type 7SK RNA, and HMGA1 shRNA-mediated knock-downs, we made the intriguing observation of several genes being statistically significantly ($p<0.05$) regulated by wild-type 7SK RNA, the chimeric EBER2-7SK L2, and the double mutant m122/m137, but not HMGA1 and the single mutant EBER2-7SK L2 m137 (Figure 5B). Despite the small number of genes displaying this regulatory phenotype - a dozen, of which half are of unknown function - this finding is of particular importance as it suggests that there are other, yet unidentified, factors being targeted by the 7SK L2 in addition to HMGA1 in HEK293 cells. Those must be either of low abundance, or have very restricted in terms of amplitude, or specific in terms of identity, effects on gene expression to result in such low penetration in the present analyses. Furthermore, this finding suggests that the interaction of these unknown factors (directly or indirectly) with 7SK L2 RNA depends on different properties of the L2 substructure. HMGA1, as our previous and present experiments show, displays sequence specificity towards 7SK RNA in that the nucleotides mutated in the 7SK L2 m137 mutant reduce significantly binding to HMGA1 (10). The double mutant 7SK L2 m122/m137 has similarly reduced HMGA1 binding activity despite having a 'rescued' secondary structure (10). Therefore, the predicted secondary structure of 7SK L2 – at least at this very position – does not impact HMGA1 recognition. The, by these comparisons predicted unknown factor, however, displays a distinct binding preference. While its activity can be altered by the wild-type 7SK L2 structure as well as the entire 7SK RNA, it seems no longer targeted by the 7SK L2 m137 mutant (data not shown). The 'rescue' double mutant 7SK L2 m122/m137, however, again seems to affect this yet to be identified activity, as the effects of EBER2-7SK L2 m122/m137 over-expression are similar if not stronger than those of the wild-type 7SK L2 (Figure 5B), suggesting that in this case the secondary structure of the 7SK L2 is a critical component of interaction.

These observations are further supported when average Euclidean distances are computed for the HMGA1 and 7SK target genes and used to hierarchically cluster the variance-weighted, averaged transcriptome profiles of the wild-type 7SK RNA, the EBER2-7SK L2 chimera, and the HMGA1 transcriptomes (Figure 5C). The resulting cluster structure reveals a closer relationship of the double/rescue mutant with 7SK RNA when compared to the single mutant. Furthermore, using correlation analysis based on the Pearson correlation coefficients calculated for the same sets of target genes it can be shown that the EBER2-7SK L2 wild-type over-expression results in gene expression changes equally well correlated with the effects of HMGA1 knock-downs and wild-type 7SK RNA over-expression, whereas the 'rescue' mutant is slightly closer related to the wild-type 7SK RNA effects than to the HMGA1 effects (Figure 5D).

Comparison of the 7SK RNA and HMGA1 target genes.

In order to better quantify the effect of the 7SK L2 single mutant on HMGA1 binding, we performed additional experiments to detect the HMGA1 binding activity of EBER2 7SK L2 RNA in comparison to EBER2 7SK L2 m137 RNA in EMSAs with increasing amounts of immuno-purified HMGA1a-FLAG fusion protein (Figure 6A). The relative affinity of each RNA to HMGA1a was then quantified, revealing, that the EBER2 7SK L2 RNA chimera exhibits a 3,5-fold higher affinity to HMGA1 than the mutated EBER2 7SK L2 m137 RNA chimera (Figure 6B). To further substantiate our findings obtained on the network level (Figure 5A) we next compared the transcriptome dynamics induced by wild-type 7SK RNA over-expression to those previously reported for HMGA1 knock-down and EBER2-7SK L2 / EBER2-7SK L2 m137 over-expression on the gene level (Figure 6D). As reported (10), the expression of the EBER2 7SK L2 RNA chimera resulted in the significantly ($p<0.01$) differential expression of 2715 genes (Figure 6D, green), whereas the expression of EBER2 7SK L2 m137 RNA led to a significant ($p<0.01$) change in expression of 761 genes (Figure 6D, red). Noteworthy, the number of genes regulated by each RNA roughly mirrors the ability of the corresponding RNA to bind HMGA1 and correlates with the relative binding affinities determined here (Figure 6A & 6B). To check the effects of the over-expressed RNA chimera for HMGA1 targets, we compared each set of differentially expressed genes with the gene expression profiles recorded after shRNA-mediated knockdown of endogenous HMGA1 (Figure 6C). We have previously shown that a total of 1649 (or 61%) of the genes affected by EBER2 7SK L2 RNA expression were also found to be HMGA1 target genes (10). More interestingly, even though the total number of differentially expressed genes is much lower downstream of EBER2 7SK L2 m137 RNA

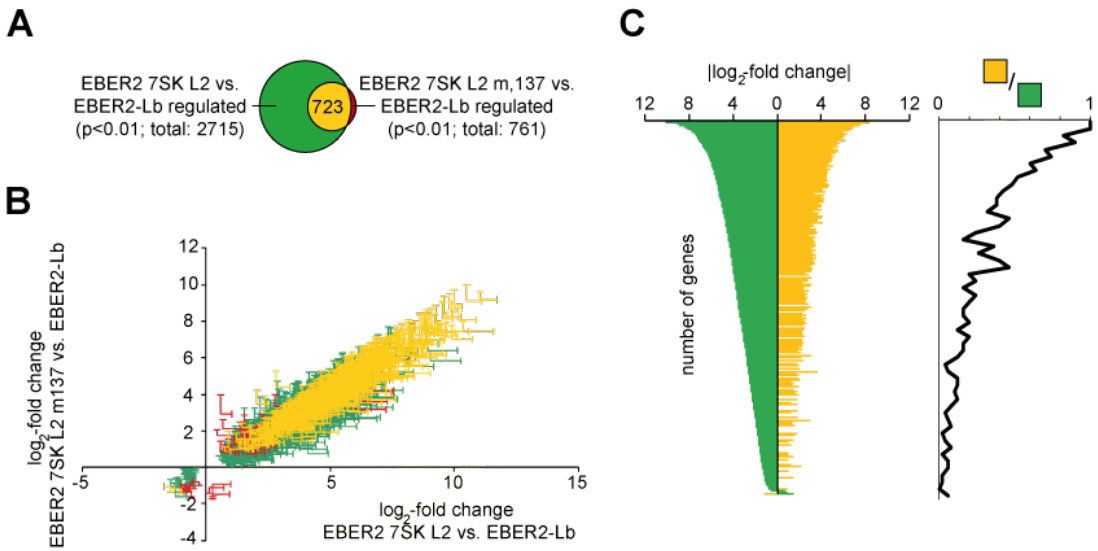


Figure 7: EBER2 7SK L2 RNA and EBER2 7SK L2 m137 RNA target the same set of genes, but to different extents. (A) Comparison of the genes differentially expressed in a significant manner (p<0.01) after EBER2 7SK L2 RNA expression compared to expression of EBER2-Lb RNA (green) with the genes significantly (p<0.01) differentially expressed after EBER2 7SK L2 m137 RNA expression compared to EBER2-Lb RNA expression (red). The common subset is marked in yellow. (B) The log2-fold change of gene expression after over-expression of EBER2 7SK L2 RNA (x-axis) is compared to that after over-expression of EBER2 7SK L2 m137 RNA (y-axis). Genes significantly regulated by both conditions are marked in yellow. Those, which are solely regulated significantly by the over-expression of EBER2 7SK L2 RNA are marked in green and those solely regulated by the over-expression of EBER2 7SK L2 m137 RNA are marked in red. The error bars show the positive s. d.. (C) The common subset of genes expressed differentially after over-expression of EBER2 7SK L2 RNA compared to the control (EBER2-Lb) was ordered by the log2-fold changes (2715 genes in A; green). Those genes in the same time expressed differentially after EBER2 7SK L2 m137 RNA expression in comparison to the expression of EBER2-Lb RNA are plotted in yellow (723 genes in A; yellow). The ratio of the number of genes differentially expressed in the overlap (yellow) per every 50 genes differentially expressed after EBER2 7SK L2 RNA over-expression in comparison to the control (green) is plotted against the log2-fold change in gene expression.

expression in comparison to EBER2 7SK L2 RNA, the common subset with HMGA1 targets contains 649 genes (85%) of the total number of genes expressed differentially in this condition (Figure 6D, red). Among the 844 significantly (p<0.01) regulated full length 7SK RNA targets, we observe a common subset with HMGA1-controlled genes of 53 (6%) (Figure 6D, blue). With 20 genes out of 194 genes (10%) having increased expression the histone H2AFY, the differentiation regulator METRNL and the transcription factor SOX4 (Figure 6E). All 20 genes in the common subset of genes, whose expression is increased upon 7SK RNA over-expression, and which are known HMGA1 target genes, also show an increased expression upon shRNA-mediated knockdown of HMGA1 (Figure 6E). Finally, all 33 genes in the common subset between 7SK RNA and HMGA1 target genes, are over-expressed after shRNA-mediated knockdown of HMGA1 (Figure 6F).

Taken together, these comparative analyses establish two important facts. First, as expected, wild-type 7SK RNA and HMGA1 target genes are only partially overlapping. While, based on the comparison with the much higher expressed EBER2-7SK L2 chimera, we believe that most, if not all, HMGA1 target genes can also be affected levels upon 7SK over-expression, the common subset with HMGA1 targets is significantly elevated in comparison to the 33 genes out of 650 genes (5%), whose expression is decreased upon 7SK over-expression (Figure 6D, blue). Among the genes with statistically significant (p<0.05) elevated expression in all of the biological conditions compared here, we find e.g. the NFκB activator TANK, the melanoma antigen MAGED4, by 7SK RNA, the reverse is not true. Some 7SK RNA targets are either independent of HMGA1 or regulated in the opposite manner by 7SK RNA and HMGA1 (Figure 6G). This likely is explained by opposing roles of HMGA1 and P-TEFb in the expression of these genes, and both activities being simultaneously being targeted by full-length 7SK RNA. Second, 7SK RNA also has positive effects on gene expression which are independent of HMGA1 and P-TEFb inhibition, as not all of the genes induced in their expression levels are also HMGA1 target genes (Figure 6D) and can not be P-TEFb target genes as no inhibitory activity of free (not 7SK-bound) P-TEFb has been identified. Those genes, including the ones identified through the comparison with the EBER2-7SK L2 m122/m137 'rescue' mutant (Figure 5B), might indicate that other, yet to be identified, roles of 7SK

RNA in transcription regulation might exist. Note that since the target cells analyzed here do neither contain the HIV1 promoter nor express TAT protein, this (these) additional regulatory functions of 7SK RNA would indeed be novel.

To further substantiate and illustrate the close relationship between 7SK L2 and HMGA1 target gene sets, and thereby indirectly lend further support to the analyses presented with respect to the 'rescue' mutant (Figure 5B) as well as the conclusions drawn here, we further investigated the overlap between the EBER2 7SK L2 and the EBER2-7SK L2 m137 target genes. 723 genes (or 95%, Figure 7A, yellow) of those genes whose expression is affected in a statistically significant manner after over-expression of the EBER2 7SK L2 m137 RNA chimera (Figure 7A, red) are also found to be among the EBER2 7SK L2 RNA target genes (Figure 7A, green). Almost all genes statistically significantly ($p<0.01$) differentially expressed by either of these conditions are also affected by the corresponding other condition (Figure 7B). The Pearson correlation coefficients of either set (Figure 7B, R^2=96, green and R^2=97, red) are comparably high with respect to the joint subset (Figure 7B, R^2=97, yellow), indicating, that both RNA chimera target the same set of genes, solely differing in the efficacy of gene expression regulation which appears to be a direct function of HMGA1 binding activity (Figure 6B). Therefore, the EBER2-7SK L2 m137 targets are a (complete) subset of the EBER2-7SK L2 targets.

Discussion

With the recent discovery of HMGA1 as being a major cellular target of the 7SK snRNA (10) in addition to the positive transcription elongation factor P-TEFb (5,6), it has become feasible to understand the transcription regulatory effects of this highly abundant and essential polymerase III transcript. Notably, the fact that P-TEFb is a non-essential component of the basal transcription machinery, as well as P-TEFb-independent effects of 7SK RNA depletion (52), have for long strongly argued for other activities of 7SK RNA in transcription regulation. While having previously focused our attention exclusively on the effects of

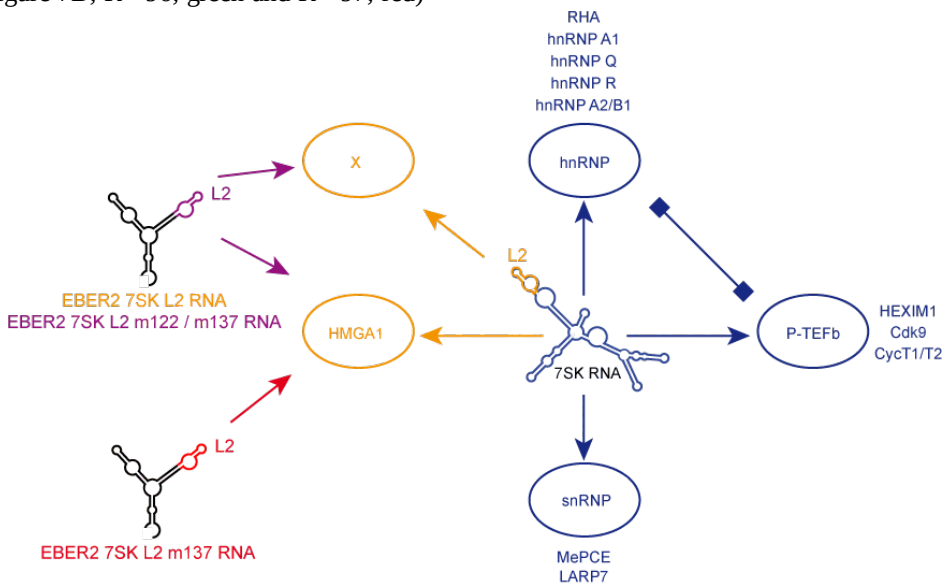


Figure 8: Schematic drawing of 7SK RNA and its different cellular interaction7SK snRNA has been shown to be stabilized by both, a γ-monomethylephosphate-GTP cap on its 5'-end, which is synthesized by the methylephosphate capping enzyme (MePCE), and by the interaction with the La-related protein LARP7. One function of 7SK RNA is the inactivation of the positive transcription elongation factor b (T-TEFb) by its interaction through L1, L3, and L4 with HEXIM1 and P-TEFb (CdK9 and CycT1/T2), which results in an inhibition of the RNA polymerase II transcription elongation reaction. 7SK RNA, which is released from P-TEFb, has also been shown to be captured in a complex with diverse heterogenous nuclear ribonuclear particles (hnRNPs), among those RHA, hnRNP A1, hnRNP Q, hnRNP R and hnRNP A2/B1. In both of these interactions, with P-TEFb and with the different hnRNPs, only the stem-loop structures L1, L3 and L4 are involved. Recently we were able to demonstrate an interaction of 7SK RNA with the chromatin master regulator and oncogenic marker HMGA1, a member of the high mobility group proteins. This interaction takes place via the stem-loop 2 (L2) substructure of 7SK RNA and has profound effects on HMGA1-dependent gene expression. The EBER2 7SK L2 RNA chimera discussed in this study carry solely the 7SK L2 substructure and thus are a potent tool to directly target HMGA1 function, avoiding off-target effects on P-TEFb function, that would be expected for full length 7SK RNA over-expression. Furthermore, the two different L2 mutants affecting either the sequence (m122/m137) or the predicted secondary structure in addition to the sequence (m137) display different activities with respect to HMGA1 and lead to the postulation of a yet to be identified activity ('X') associated with 7SK RNA L2.

the 7SK L2 substructure on HMGA1 (10), we have here investigated the effects of wild-type 7SK RNA on the cellular transcriptome in the same model system, and systematically compared the regulatory effects of full-length 7SK RNA and the effects of the isolated 7SK L2 substructure on HMGA1 target gene expression. These comparisons have allowed to establish HMGA1-dependent and HMGA1-independent effects of 7SK RNA on HMGA1 target genes (Figure 6G) as well as genes not affected by HMGA1 at all (Figure 5B). These findings have been substantiated by the investigation of the effects of two 7SK L2 mutants on the cellular transcriptome dynamics (Figures 5-7). As suggested by the previous anlyses (10), the 7SK L2 m137 mutant targets HMGA1 in a manner identical to wild-type 7SK L2 (Figure 7), however, displays lowered binding affinity and concomitantly lowered inhibitory activity on the positive and negative regulatory actions of HMGA1. In contrast, the double 7SK L2 m122/m137 'rescue' mutant with restored (predicted) secondary structure, and previously not investigated for its effects on transcriptome dynamics for displaying similar binding affinity to HMGA1 as the single mutant (10), reveals HMGA1-independent 7SK L2 substructure activity on a few target genes. This finding does not only potentially explain the existence of 7SK L2-dependent gene regulatory phenomena which are not detectable by HMGA1 knock-downs, but importantly also leads to the hypothesis of yet unidentified factors being targeted by 7SK RNA and in particular the L2 substructure (Figure 8). Importantly, the comparative investigation of both mutants in HMGA1 binding and their transcriptome activities, has lead to the observation that while HMGA1 binding strength is affected by the nucleotide sequence of the 7SK L2 substructure but not necessarily by the secondary structure of this precise part of 7SK snRNA. This seems to be different for the unknown activity targeted by 7SK RNA, as different behavior is observed whether or not the predicted secondary structure is destroyed (L2 m137) or not (L2 m122/m137, Figure 8). Finally, on a more technical note, we have extended previous observations on the various background activities of different EBER2 internal deletion mutants (10, 41), and demonstrated the need for carefully controlling the activity of chimera constructed using EBER2 as an expression vehicle for non-translated RNA substructures. These latter findings, albeit awaiting a molecular explanation, should further aid in the dissection of specific and non-specific effects of EBER2 chimera which turn out to be a very valuable tool for instance in the characterization of HMGA1 activity as reported on here and elsewhere (10).

The further characterization of HMGA1-dependent and independent effects of 7SK snRNA on cellular gene expression provide a means to better understand the functional roles of both 7SK RNA and HMGA1 in physiology. HMGA1, being a 'hub' of nuclear function and through its activity tightly linked to oncogenesis and metastasis (13, 18-22), might be targeted by the expression of the 7SK L2 substructure in a therapeutically meaningful manner. It remains to be seen whether or nor 7SK snRNA also is capable of directly affecting other, yet unidentified, cellular factors as the analyses reported here suggest.

Acknowledgments

We are grateful to Brendan Bell for his helpful comments on the manuscript. Nicolas Tchitchek is thanked for instructions on the use of the microarray and relevant statistical analysis methodologies. This work was supported by the Centre National de la Recherche Scientifique (C.N.R.S.) and the Genopole Evry (to A.B.) as well as the Deutsche Forschungsgemeinschaft (BE 531/19-3, to: BJB). C.B. is a recipient of a pre-doctoral fellowship from the Agence Nationale de recherches sur le SIDA et les hépatites virales (ANRS). The authors declare to have no competing interests.

Materials and Methods

Cell culture

HEK293 and HeLa cells were cultured as described previously (41). Transfections were either performed by the calciumphosphate coprecipitation method (42) or by FugeneHD (Roche) according to the manufacturers instructions.

Molecular biology

Extract Preparations. The preparation of nuclear extract for *in vitro* transcription analyses was performed as described previously (43). To prepare extracts for immunopurifications, transfected cells were reconstituted in isotonic IP-lysis buffer (50 mM Tris/HCl, pH 7.4, 150 mM NaCl, 1 mM EDTA, 1% TRITON^{TM}X-100, protease inhibitor) and homogenized in a dounce-homogenizer before cellular debris was removed by high speed centrifugation (16.000 x g). The extract was subsequently dialyzed against IP-buffer (20 mM HEPES/KOH, pH 7.9, 100 mM KCl, 0.2 mM EDTA, 20% glycerol).

Immunopurification of HMGA1a-FLAG. Immunoprecipitations of FLAG fusion proteins were performed with M2 anti-FLAG agarose (Sigma) as recommended by the manufacturer. Briefly, anti-FLAG agarose was incubated with extracts containing FLAG fusion proteins in IP-buffer (20 mM HEPES/KOH, pH 7.9, 100 mM KCl, 0.2 mM EDTA, 20% glycerol) for 2 hours at 4°C. The agarose was pelleted by centrifugation and the supernatant was removed. After washing three times with the 20-fold volume of IP-washing buffer (20 mM HEPES/KOH, pH 7.9, 100 mM KCl, 0.2 mM EDTA, 0,025% Tween 20), the HMGA1a-FLAG was eluted with the 5-fold volume of elution buffer (100 mM Glycine/HCl, pH 3.5). The elution fraction was subsequently dialzed against IP-buffer before use in EMSAs.

Electrophoretic Mobility Shift Assays. For electrophoretic mobility shift assays (EMSAs), 32P-labeled RNA was denatured for 5 minutes at 75°C and renatured by cooling slowly down to room temperature. The nucleic acid was incubated with immuno-purified protein in EMSA-buffer (10 mM HEPES/KOH, pH 7.9, 100 mM KCl, 5 mM $MgCl_2$, 0.25 mM DTT, 0.1 mM EDTA, 10% glycerol and 2 µg yeast tRNA per lane) for 10 minutes at 37°C and subsequently for 10 minutes on ice. The resulting complexes were separated by 10% native PAGE and visualized by autoradiography.

Total RNA preparation. Total RNA from transfected cells was either prepared using the RNeasy Midi Kit (Qiagen) as recommended by the manufacturer or as described previously (44).

In vitro transcription. In vitro transcription analyses to detect promoter activity were performed using nuclear extract from HeLa cells (43). RNA for EMSAs was prepared using T7 RNA Polymerase (NEB) on a linearized template DNA containing the coding sequence of the RNA of interest controlled by a T7 promoter, as recommended by the manufacturer. To label RNA with ^{32}P, the reaction mixture contained a 10-fold excess of ^{32}P-labeled α-UTP.

RNA structure prediction

The secondary structure of RNA was calculated by RNAstructure (Version 4.4) (45).

Quantitative RT-PCR analyses

qRT-PCR analyses were performed using SuperScript^{TM} 2 Reverse Transcriptase (Invitrogen), Lightcycler (Roche) and Fast Start DNA MasterPLUS SYBR Green I reaction mix (Roche) as recommended by the manufacturer. Primers used to quantify the housekeeping gene Ubiquitin were 5'-GTT GAG ACT TCG TGG TGG TG-3' (sense) and 5'-TCT CGA CGA AGG CGA CTA AT-3' (antisense). Primers used to quantify endogenous HMGA1 were 5'-TCC CAG CCA TCA CTC TTC-3' (sense) and 5'-CTC CTT CTG ACT CCC TAC C-3' (antisense). Primers to detect endogenous 7SK RNA were 5'-CAT CCC CGA TAG AGG AGG AC-3' (sense) and 5'-GCC TCA TTT GGA TGT GTC TG-3' (antisense).

Transcriptome analyses

For microarray analyses, RNA amplification, labeling, hybridization and detection were performed following the protocols supplied by Applied Biosystems using the corresponding kits (Applied Biosystems, ProdNo: 4339628 and 4336875). The data obtained were analyzed as described previously (10, 41,46-49). Transcriptome data, annotated to MIAME I+II standards, were deposited in the public database MACE (http://mace.ihes.fr):

1. 7SK wild-type RNA overexpression: MACE Acc. No.: 2655275478
Reviewer login: 2655275478 password: ppEjkv9IJ_

2. EBER2-Lb-short & EBER2- 7SK L2 m122/m137: MACE Acc. No.: 2147641814
Reviewer login: 2147641814
password: 9vBB0Nut4l

In addition we used previously published transcriptome data:

3. EBER2 and EBER2-Lb: MACE Acc. No.: 3034287118
4. EBER2-7SK L2 and EBER2-7SK L2 m137: MACE Acc. No.: 2979879726
5. HMGA1 knock-down: MACE Acc. No.: 2970966830

The statistical analysis of the transcriptome data has been described previously (10, 41). Gene ontology enrichment analyses were also described previously (41, 49). The signalosome network inference will be described in detail elsewhere (Bécavin et al. in preparation), and the network representation tool used was Cytoscape (51).

Constructs used in this study

EBER2 7SK L2 fusion transcripts. The EBER2 gene from -171 nt to +226 nt containing the complete viral promoter sequence was cloned into the EcoRI and HindIII restriction sites of puC18 vector (Fermentas). This clone was used to construct all EBER2 7SK L2 RNA chimera. To obtain the EBER2 7SK L2 construct, the sequence from +81 nt to +140 nt of wild type EBER2 RNA was exchanged by the sequence from +113 nt to +154 nt of wild type 7SK RNA. To obtain the EBER2 7SK L2 m137 construct nucleotides +104 to +106 (CGG) of the EBER2 7SK L2 construct were mutated to GCC. In the construct EBER2 7SK L2 m122/m137 in addition the nucleotides +89 to +91 (CCG) were mutated to GGC. To obtain the EBER2-Lb (EBER2-Lb-short) construct, nucleotides +80 to +145 (+78 to + 148) of wild type EBER2 RNA were deleted.

HMGA1a-FLAG. To obtain a vector for the expression of FLAG fusion protein, the hybridized oligonucleotides 5'-CTA GAG GGC GAC TAC AAA GAC GAT GAC GAC AAA GGA TGA A-3' (sense) and 3'-TC CCG CTG ATG TTT CTG CTA CTG CTG TTT CCT ACT TGG CC-5' (antisense) were inserted into the XbaI and AgeI restriction sites of the
vector pcDNA4/TO/myc-his B (Invitrogen). The coding sequence of wild type HMGA1a from +1 nt to +321 nt was inserted into the PstI and XbaI restriction sites of the vector mentioned above.

ShRNA knockdown of HMGA1. To knock down the expression of endogenous HMGA1, we used the expression of short hairpin RNA containing a functional siRNA sequence targeting HMGA1 mRNA (Liau et al.; 2006). The full sequence of the HMGA1 targeting shRNA was 5'-CAA CUC CAG GAA GGA AAC CAA GCG AUU GGU UUC CUU CCU GGA GUU G-3'. For control a non targeting (scramble) shRNA was used with the sequence 5'-AAC AGU CGC GUU UGC GAC UGG GCG ACC AGU CGC AAA CGC GAC UGU U-3'. For downstream transcriptome analyses these shRNAs were expressed using a vector coding for the corresponding shRNA under control of the 7SK promoter.

P7SK. To overexpress full length 7SK RNA, the complete human 7SK gene including the full length promoter was cloned into the puC18 vector (Fermentas) (50).

References

1. Matera, A.G. and Ward, D.C. (1993) Nucleoplasmic organization of small nuclear ribonucleoproteins in cultured human cells. *J Cell Biol*, **121**, 715-727.
2. Gurney, T., Jr. and Eliceiri, G.L. (1980) Intracellular distribution of low molecular weight RNA species in HeLa cells. *J Cell Biol*, **87**, 398-403.
3. Zieve, G., Benecke, B.J. and Penman, S. (1977) Synthesis of two classes of small RNA species in vivo and in vitro. *Biochemistry*, **16**, 4520-4525.
4. Zieve, G. and Penman, S. (1976) Small RNA species of the HeLa cell: metabolism and subcellular localization. *Cell*, **8**, 19-31.
5. Nguyen, V.T., Kiss, T., Michels, A.A. and Bensaude, O. (2001) 7SK small nuclear RNA binds to and inhibits the activity of CDK9/cyclin T complexes. *Nature*, **414**, 322-325.
6. Yang, Z., Zhu, Q., Luo, K. and Zhou, Q. (2001) The 7SK small nuclear RNA inhibits the CDK9/cyclin T1 kinase to control transcription. *Nature*, **414**, 317-322.
7. Barboric, M. and Lenasi, T. Kick-sTARting HIV-1 transcription elongation by 7SK snRNP deporTATion. *Nat Struct Mol Biol*, **17**, 928-930.
8. D'Orso, I. and Frankel, A.D. RNA-mediated displacement of an inhibitory snRNP complex activates transcription elongation. *Nat Struct Mol Biol*, **17**, 815-821.
9. Cho, W.K., Jang, M.K., Huang, K., Pise-Masison, C.A. and Brady, J.N. Human T-Lymphotropic Virus Type 1 Tax Protein Complexes with P-TEFb and Competes for Brd4 and 7SK snRNP/HEXIM1 Binding. *J Virol*.
10. Eilebrecht, S., Brysbaert, G., Wegert, T., Urlaub, H., Benecke, B.-J. and Benecke,A. (2010) 7SK small nuclear RNA is a direct affector of HMGA1 function in transcription regulation. *Nucleic Acids Research*.

11. Egloff, S., Van Herreweghe, E. and Kiss, T. (2006) Regulation of polymerase II transcription by 7SK snRNA: two distinct RNA elements direct P-TEFb and HEXIM1 binding. *Mol Cell Biol*, **26**, 630-642.
12. Van Herreweghe, E., Egloff, S., Goiffon, I., Jady, B.E., Froment, C., Monsarrat, B. and Kiss, T. (2007) Dynamic remodelling of human 7SK snRNP controls the nuclear level of active P-TEFb. *EMBO J*, **26**, 3570-3580.
13. Reeves, R. (2001) Molecular biology of HMGA proteins: hubs of nuclear function. *Gene*, **277**, 63-81.
14. Reeves, R. (2003) HMGA proteins: flexibility finds a nuclear niche? *Biochem Cell Biol*, **81**, 185-195.
15. Reeves, R. (2004) HMGA proteins: isolation, biochemical modifications, and nucleosome interactions. *Methods Enzymol*, **375**, 297-322.
16. Reeves, R. and Beckerbauer, L. (2001) HMGI/Y proteins: flexible regulators of transcription and chromatin structure. *Biochim Biophys Acta*, **1519**, 13-29.
17. Chiappetta, G., Avantaggiato, V., Visconti, R., Fedele, M., Battista, S., Trapasso, F., Merciai, B.M., Fidanza, V., Giancotti, V., Santoro, M. *et al.* (1996) High level expression of the HMGI (Y) gene during embryonic development. *Oncogene*, **13**, 2439-2446.
18. Chiappetta, G., Bandiera, A., Berlingieri, M.T., Visconti, R., Manfioletti, G., Battista, S., Martinez-Tello, F.J., Santoro, M., Giancotti, V. and Fusco, A. (1995) The expression of the high mobility group HMGI (Y) proteins correlates with the malignant phenotype of human thyroid neoplasias. *Oncogene*, **10**, 1307-1314.
19. Cleynen, I. and Van de Ven, W.J. (2008) The HMGA proteins: a myriad of functions (Review). *Int J Oncol*, **32**, 289-305.
20. Fusco, A. and Fedele, M. (2007) Roles of HMGA proteins in cancer. *Nat Rev Cancer*, **7**, 899-910.
21. Hess, J.L. (1998) Chromosomal translocations in benign tumors: the HMGI proteins. *Am J Clin Pathol*, **109**, 251-261.
22. Reeves, R. and Beckerbauer, L.M. (2003) HMGA proteins as therapeutic drug targets. *Prog Cell Cycle Res*, **5**, 279-286.
23. Himes, S.R., Reeves, R., Attema, J., Nissen, M., Li, Y. and Shannon, M.F. (2000) The role of high-mobility group I(Y) proteins in expression of IL-2 and T cell proliferation. *J Immunol*, **164**, 3157-3168.
24. John, S., Reeves, R.B., Lin, J.X., Child, R., Leiden, J.M., Thompson, C.B. and Leonard, W.J. (1995) Regulation of cell-type-specific interleukin-2 receptor alpha-chain gene expression: potential role of physical interactions between Elf-1, HMG-I(Y), and NF-kappa B family proteins. *Mol Cell Biol*, **15**, 1786-1796.
25. John, S., Robbins, C.M. and Leonard, W.J. (1996) An IL-2 response element in the human IL-2 receptor alpha chain promoter is a composite element that binds Stat5, Elf-1, HMG-I(Y) and a GATA family protein. *EMBO J*, **15**, 5627-5635.
26. Lehn, D.A., Elton, T.S., Johnson, K.R. and Reeves, R. (1988) A conformational study of the sequence specific binding of HMG-I (Y) with the bovine interleukin-2 cDNA. *Biochem Int*, **16**, 963-971.
27. Magnuson, N.S., Spies, A.G., Nissen, M.S., Buck, C.D., Weinberg, A.D., Barr, P.J., Magnuson, J.A. and Reeves, R. (1987) Bovine interleukin 2: regulatory mechanisms. *Vet Immunol Immunopathol*, **17**, 183-192.
28. Reeves, R., Elton, T.S., Nissen, M.S., Lehn, D. and Johnson, K.R. (1987) Posttranscriptional gene regulation and specific binding of the nonhistone protein HMG-I by the 3' untranslated region of bovine interleukin 2 cDNA. *Proc Natl Acad Sci U S A*, **84**, 6531-6535.
29. Reeves, R., Leonard, W.J. and Nissen, M.S. (2000) Binding of HMG-I(Y) imparts architectural specificity to a positioned nucleosome on the promoter of the human interleukin-2 receptor alpha gene. *Mol Cell Biol*, **20**, 4666-4679.
30. Danzeiser, D.A., Urso, O. and Kunkel, G.R. (1993) Functional characterization of elements in a human U6 small nuclear RNA gene distal control region. *Mol Cell Biol*, **13**, 4670-4678.
31. Murphy, S., Pierani, A., Scheidereit, C., Melli, M. and Roeder, R.G. (1989) Purified octamer binding transcription factors stimulate RNA polymerase III--mediated transcription of the 7SK RNA gene. *Cell*, **59**, 1071-1080.
32. Sturm, R.A., Das, G. and Herr, W. (1988) The ubiquitous octamer-binding protein Oct-1 contains a POU domain with a homeo box subdomain. *Genes Dev*, **2**, 1582-1599.
33. Kleinert, H., Bredow, S. and Benecke, B.J. (1990) Expression of a human 7S K RNA

gene in vivo requires a novel pol III upstream element. *EMBO J*, **9**, 711-718.
34. Kunkel, G.R., Cheung, T.C., Miyake, J.H., Urso, O., McNamara-Schroeder, K.J. and Stumph, W.E. (1996) Identification of a SPH element in the distal region of a human U6 small nuclear RNA gene promoter and characterization of the SPH binding factor in HeLa cell extracts. *Gene Expr*, **6**, 59-72.
35. Schaub, M., Myslinski, E., Schuster, C., Krol, A. and Carbon, P. (1997) Staf, a promiscuous activator for enhanced transcription by RNA polymerases II and III. *EMBO J*, **16**, 173-181.
36. Dahlberg, J.E. and Blattner, F.R. (1975) Sequence of the promoter-operator proximal region of the major leftward RNA of bacteriophage lambda. *Nucleic Acids Res*, **2**, 1441-1458.
37. Lerner, M.R., Andrews, N.C., Miller, G. and Steitz, J.A. (1981) Two small RNAs encoded by Epstein-Barr virus and complexed with protein are precipitated by antibodies from patients with systemic lupus erythematosus. *Proc Natl Acad Sci U S A*, **78**, 805-809.
38. Howe, J.G. and Shu, M.D. (1989) Epstein-Barr virus small RNA (EBER) genes: unique transcription units that combine RNA polymerase II and III promoter elements. *Cell*, **57**, 825-834.
39. Geiduschek, E.P. and Tocchini-Valentini, G.P. (1988) Transcription by RNA polymerase III. *Annu Rev Biochem*, **57**, 873-914.
40. Rosa, M.D., Gottlieb, E., Lerner, M.R. and Steitz, J.A. (1981) Striking similarities are exhibited by two small Epstein-Barr virus-encoded ribonucleic acids and the adenovirus-associated ribonucleic acids VAI and VAII. *Mol Cell Biol*, **1**, 785-796.
41. Eilebrecht, S., Pellay, F.X., Odenwalder, P., Brysbaert, G., Benecke, B.J. and Benecke, A. (2008) EBER2 RNA-induced transcriptome changes identify cellular processes likely targeted during Epstein Barr Virus infection. *BMC Res Notes*, **1**, 100.
42. Graham, F.L. and van der Eb, A.J. (1973) Transformation of rat cells by DNA of human adenovirus 5. *Virology*, **54**, 536-539.
43. Dignam, J.D., Martin, P.L., Shastry, B.S. and Roeder, R.G. (1983) Eukaryotic gene transcription with purified components. *Methods Enzymol*, **101**, 582-598.
44. Chomczynski, P. and Sacchi, N. (1987) Single-step method of RNA isolation by acid guanidinium thiocyanate-phenol-chloroform extraction. *Anal Biochem*, **162**, 156-159.
45. Zuker, M. (1989) Computer prediction of RNA structure. *Methods Enzymol*, **180**, 262-288.
46. Noth, S. and Benecke, A. (2005) Avoiding inconsistencies over time and tracking difficulties in Applied Biosystems AB1700/Panther probe-to-gene annotations. *BMC Bioinformatics*, **6**, 307.
47. Noth, S., Brysbaert, G. and Benecke, A. (2006) Normalization using weighted negative second order exponential error functions (NeONORM) provides robustness against asymmetries in comparative transcriptome profiles and avoids false calls. *Genomics Proteomics Bioinformatics*, **4**, 90-109.
48. Noth, S., Brysbaert, G., Pellay, F.X. and Benecke, A. (2006) High-sensitivity transcriptome data structure and implications for analysis and biologic interpretation. *Genomics Proteomics Bioinformatics*, **4**, 212-229.
49. Wilhelm, E., Kornete, M., Targat, B., Vigneault-Edwards, J., Frontini, M., Tora, L., Benecke, A. and Bell, B. TAF6delta orchestrates an apoptotic transcriptome profile and interacts functionally with p53. *BMC Mol Biol*, **11**, 10.
50. Surig, D., Bredow, S. and Benecke, B.J. (1993) The seemingly identical 7SK and U6 core promoters depend on different transcription factor complexes. *Gene Expr*, **3**, 175-185.
51. Shannon P, Markiel A, Ozier O, Baliga NS, Wang JT, Ramage D, Amin N, Schwikowski B, Ideker T. (2003) Cytoscape: a software environment for integrated models of biomolecular interaction networks. Genome Research 13(11):2498-504.
52. Haaland, R.E., Herrmann, C.H. and Rice, A.P. (2005) siRNA depletion of 7SK snRNA induces apoptosis but does not affect expression of the HIV-1 LTR or P-TEFb-dependent cellular genes. *J Cell Physiol*, **205**, 463-470.

.

Bibliographie

[1] Noth S, Brysbaert G, Pellay F, Benecke A : **High-sensitivity transcriptome data structure and implications for analysis and biologic interpretation**. Genomics, Proteomics & Bioinformatics / Beijing Genomics Institute 2006, **4**(4) :212–229, [http://www.ncbi.nlm.nih.gov/pubmed/17531797].

[2] Benecke A : **Chromatin code, local non-equilibrium dynamics, and the emergence of transcription regulatory programs**. The European Physical Journal E : Soft Matter and Biological Physics 2006, **19**(3) :353–366, [http://dx.doi.org/10.1140/epje/i2005-10068-8].

[3] Holmes S : **Visualising Data**. In Statistical Problems in Particle Physics, Astrophysics and Cosmology. Edited by Lyons L, Karag M 2006 :197.

[4] Pearson K : **On lines and planes of closest fit to systems of points in space**. Philosophical Magazine 1901, **2**(6) :572, 559.

[5] Hotelling H : **Analysis of a complex of statistical variables into principal components.** Journal of Educational Psychology 1933, **24**(6) :417–441.

[6] Hotelling H : **Analysis of a complex of statistical variables into principal components.** Journal of Educational Psychology 1933, **24**(7) :498–520, [http://search.ebscohost.com/login.aspx?direct=true&db=pdh&AN=edu-24-7-498&site=ehost-live].

[7] Karhunen K : **Uber lineare methoden in der wahrscheinlichkeitsrechnung**. Ann. Acad. Sci. Fennicae, ser. A1,Math. Phys. 1946, **37**.

[8] Berthold M, Hand D : Intelligent Data Analysis. Springer second ed, 2003.

[9] de Haan JR, Wehrens R, Bauerschmidt S, Piek E, van Schaik RC, Buydens LMC : **Interpretation of ANOVA models for microarray data using PCA**. Bioinformatics 2007, **23**(2) :184–190, [http://bioinformatics.oxfordjournals.org.gate1.inist.fr/cgi/content/abstract/23/2/184].

[10] Jonnalagadda S, Srinivasan R : **Principal components analysis based methodology to identify differentially expressed genes in time-course microarray data**. BMC Bioinformatics 2008, **9** :267–267. [PMID : 18534040 PMCID : 2435549].

[11] Hubert M, Engelen S : **Robust PCA and classification in biosciences**. Bioinformatics 2004, **20**(11) :1728–1736, [http://bioinformatics.oxfordjournals.org.gate1.inist.fr/cgi/content/abstract/20/11/1728].

[12] Young G, Householder A : **Discussion of a set of points in terms of their mutual distances**. Psychometrika 1938, **3** :19–22.

[13] Torgerson W : **Multidimensional scaling : I. Theory and method**. Psychometrika 1952, **17**(4) :401–419.

[14] Cox T, Cox M : Multidimensional Scaling, Second Edition. Chapman & Hall/CRC, 2 edition 2000.

[15] Benzécri JP : L'analyse des données vol.2 : L'anaylse des correspondances. Dunod Paris 1976.

[16] Greenacre MJ : Theory and applications of correspondence analysis. London Academic Press 1984.

[17] Cuadras C, Fortiana J : **Metric Scaling Graphical Representation of Categorical Data**. Penn State University 1995, [http://citeseerx.ist.psu.edu/viewdoc/summary?doi=10.1.1.56.958].

[18] Fellenberg K, Hauser N, Brors B, Neutzner A, Hoheisel J, Vingron M : **Correspondence analysis applied to microarray data**. Proceedings of the National Academy of Sciences 2001, **98**(19) :10781.

[19] Schmidt E, Stewart G : **On the Early History of the Singular Value Decomposition**. Univeristy of Maryland 1992, [http://www.lib.umd.edu/drum/bitstream/1903/566/4/CS-TR-2855.pdf].

[20] Golub G, Kahan W : **Calculating the singular values and pseudo-inverse of a matrix**. Journal of the Society for Industrial and Applied Mathematics : Series B, Numerical Analysis 1965, :205–224.

[21] Eckart C, Young G : **The approximation of one matrix by another of lower rank**. Psychometrika 1936, **1**(3) :211–218.

[22] Troyanskaya O, Cantor M, Sherlock G, Brown P, Hastie T, Tibshirani R, Botstein D, Altman RB : **Missing value estimation methods for DNA microarrays**. Bioinformatics 2001, **17**(6) :520–525, [http://bioinformatics.oxfordjournals.org/cgi/content/abstract/17/6/520].

[23] Ding C, He X : **K-means clustering via principal component analysis**. In Proceedings of the 21 st International Conference on Machine Learning, ACM Press 2004 :225–232, [http://www.aicml.cs.ualberta.ca/banff04/icml/pages/papers/262.pdf].

[24] Alter O, Brown P, Botstein D : **Singular value decomposition for genome-wide expression data processing and modeling**. Proceedings of the National Academy of Sciences 2000, **97**(18) :10101–10106, [http://www.pnas.org/content/97/18/10101.abstract].

[25] Alter O, Brown P, Botstein D : **Generalized singular value decomposition for comparative analysis of genome-scale expression data sets of two different organisms**. Proceedings of the National Academy of Sciences 2003, **100**(6) :3351–3356.

[26] Omberg L, Golub G, Alter O : **A tensor higher-order singular value decomposition for integrative analysis of DNA microarray data from different studies**. Proceedings of the National Academy of Sciences 2007, **104**(47) :18371.

[27] Spellman PT, Sherlock G, Zhang MQ, Iyer VR, Anders K, Eisen MB, Brown PO, Botstein D, Futcher B : **Comprehensive identification of cell cycle-regulated genes of the yeast Saccharomyces cerevisiae by microarray hybridization**. Molecular Biology of the Cell 1998, **9**(12) :3273–3297, [http://www.ncbi.nlm.nih.gov/pubmed/9843569]. [PMID : 9843569].

[28] DeCoster J : **Overview of Factor Analysis**. from http ://www.stat-help.com/notes.html 1998.

[29] Harman HH : Modern factor analysis. University of Chicago Press 1976.

[30] Papoulis A : Probability, random variables, and stochastic processes. McGraw-Hill 1991.

[31] Hyvärinen A, Oja E : **Independent component analysis : algorithms and applications**. Neural Networks 2000, **13**(4-5) :411–430, [http://www.sciencedirect.com.gate1.inist.fr/science/article/B6T08-43X2MFY-2/2/e16b1ec17aa9eec8247cd937c631e846].

[32] Friedman J, Tukey J : **A projection pursuit algorithm for exploratory data analysis**. Graphics : 1965-1985 1988, :149.

[33] Huber PJ : **Projection Pursuit**. The Annals of Statistics 1985, **13**(2) :435–475, [http://www.jstor.org/stable/2241175].

[34] Hyvärinen A : **Survey on Independent Component Analysis**. Neural Computing Surveys 1999, **2** :94–128, [http://citeseerx.ist.psu.edu/viewdoc/summary?doi=10.1.1.45.3488].

[35] Borg I, Groenen PJF : Modern multidimensional scaling : theory and applications. Springer 2005.

[36] Graef J, Spence I : **Using distance information in the design of large multidimensional scaling experiments**. Psychological Bulletin 1979, **86** :60–66.

[37] Dzwinel W : **Virtual particles and search for global minimum**. Future Generation Computer Systems 1997, **12**(5) :371–389, [http://www.sciencedirect.com/science/article/B6V06-3VV6MK2-5/2/732a9b64702a52b4c2326fb6f9c4ba07].

[38] Dzwinel W, Blasiak J : **Method of particles in visual clustering of multidimensional and large data sets**. Future Generation Computer Systems 1999, **15**(3) :365–379, [http://www.sciencedirect.com.gate1.inist.fr/science/article/B6V06-3XWYWG5-S/2/c0fb6341dfd01be2ee038008dbeafc95].

[39] Andrecut M : **Molecular dynamics multidimensional scaling**. Physics Letters A 2009, **373**(23-24) :2001–2006, [http://www.sciencedirect.com/science/article/B6TVM-4W1JVY3-2/2/b9378b51366c7be277cecdbe86c8e8ce].

[40] Andreas B, Swayne DF, Littman ML, Nathaniel D, Hofmann H : **Interactive Data Visualization with Multidimensional Scaling**. Tech. rep., University of Pennsylvania 2004.

[41] Ebbels TMD, Buxton BF, Jones DT : **springScape : visualisation of microarray and contextual bioinformatic data using spring embedding and an 'information landscape'**. Bioinformatics (Oxford, England) 2006, **22**(14) :e99–107, [http://www.ncbi.nlm.nih.gov.gate1.inist.fr/pubmed/16873528]. [PMID : 16873528].

[42] Gray LC, Vaidya JS, Baum M, Badwe RA, Mittra I, Siddiqui T, Wiarda D : **Functional maps of metastases from breast cancers : proof of the principle that multidimensional scaling can summarize disease progression**. World Journal of Surgery 2004, **28**(7) :646–651, [http://www.ncbi.nlm.nih.gov.gate1.inist.fr/pubmed/15185001]. [PMID : 15185001].

[43] Taguchi Y, Oono Y : **Relational patterns of gene expression via non-metric multidimensional scaling analysis**. Bioinformatics 2005, **21**(6) :730–740.

[44] Tzeng J, Lu HH, Li W : **Multidimensional scaling for large genomic data sets**. BMC Bioinformatics 2008, **9** :179–179.

[45] Deun KV, Marchal K, Heiser WJ, Engelen K, Mechelen IV : **Joint mapping of genes and conditions via multidimensional unfolding analysis**. BMC Bioinformatics 2007, **8** :181–181. [PMID : 17550582 PMCID : 1904247].

[46] Ashburner M, Ball CA, Blake JA, Botstein D, Butler H, Cherry JM, Davis AP, Dolinski K, Dwight SS, Eppig JT, Harris MA, Hill DP, Issel-Tarver L, Kasarskis A, Lewis S, Matese JC, Richardson JE, Ringwald M, Rubin GM, Sherlock G : **Gene ontology : tool for the unification of biology. The Gene Ontology Consortium**. Nature Genetics 2000, **25** :25–29, [http://www.ncbi.nlm.nih.gov.gate1.inist.fr/pubmed/10802651]. [PMID : 10802651].

[47] Steinhaus H : **Sur la division des corps materiels en parties.** Bull. Acad. Pol. Sci., Cl. III 1957, **4** :801–804.

[48] Lloyd S : **Least squares quantization in PCM**. IEEE transactions on information theory 1982, **28**(2) :129–137.

[49] Kohonen T : Self-organizing maps. Springer 2001.

[50] Yin H : Principal Manifolds for Data Visualization and Dimension Reduction, Springer Publishing Company, Incorporated 2007 chap. Learning Nonlinear Principal Manifolds by Self-Organizing Maps.

[51] Kruger U, Zhang J, Xie L : Principal Manifolds for Data Visualization and Dimension Reduction, Springer Publishing Company, Incorporated 2007 chap. Developments and applications of Nonlinear Principal Components Analysis - a Review.

[52] Hastie T, Stuetzle W : **Principal Curves**. Journal of the American Statistical Association 1989, **84**(406) :502–516.

[53] Kramer M : **Nonlinear principal components analysis using auto-associative neural networks**. AIChE Journal 1991, **37**(2) :233–243.

[54] Schalkopf B, Smola A, Maller KR : **Nonlinear Component Analysis as a Kernel Eigenvalue Problem**. Neural Computation 1998, **10**(5) :1299–1319, [http://dx.doi.org/10.1162/089976698300017467].

[55] Tenenbaum JB, de Silva V, Langford JC : **A Global Geometric Framework for Nonlinear Dimensionality Reduction**. Science 2000, **290**(5500) :2319–2323, [http://www.sciencemag.org.gate1.inist.fr/cgi/content/abstract/290/5500/2319].

[56] Dawson K, Rodriguez RL, Malyj W : **Sample phenotype clusters in high-density oligonucleotide microarray data sets are revealed using Isomap, a nonlinear algorithm**. BMC Bioinformatics 2005, **6** :195, [http://www.ncbi.nlm.nih.gov/pubmed/16076401]. [PMID : 16076401].

[57] Roweis ST, Saul LK : **Nonlinear Dimensionality Reduction by Locally Linear Embedding**. Science 2000, **290**(5500) :2323–2326, [http://www.sciencemag.org.gate1.inist.fr/cgi/content/abstract/290/5500/2323].

[58] Van der Maaten L, Postma E, Van den Herik H : **Dimensionality reduction : A comparative review**. Published online 2007.

[59] Lafon S, Lee AB : **Diffusion maps and coarse-graining : A unified framework for dimensionality reduction, graph partitioning, and data set parameterization**. IEEE Transactions on Pattern Analysis and Machine Intelligence 2006, **28**(9) :1393–1403, [http://www.ncbi.nlm.nih.gov.gate1.inist.fr/pubmed/16929727]. [PMID : 16929727].

[60] Hinton GE, Salakhutdinov RR : **Reducing the Dimensionality of Data with Neural Networks**. Science 2006, **313**(5786) :504–507, [http://www.sciencemag.org.gate1.inist.fr/cgi/content/abstract/313/5786/504].

[61] Donoho DL, Grimes C : **Hessian eigenmaps : Locally linear embedding techniques for high-dimensional data**. Proceedings of the National Academy of Sciences of the United States of America 2003, **100**(10) :5591–5596, [http://www.ncbi.nlm.nih.gov.gate1.inist.fr/pubmed/16576753]. [PMID : 16576753].

[62] Zhang Z, Zha H : **Principal Manifolds and Nonlinear Dimension Reduction via Local Tangent Space Alignment**. cs/0212008 2002, [http://arxiv.org/abs/cs/0212008].

[63] Belkin M, Niyogi P : **Laplacian Eigenmaps for Dimensionality Reduction and Data Representation**. Neural Computation 2003, **15**(6) :1373–1396, [http://dx.doi.org/10.1162/089976603321780317].

[64] Teh YW, Roweis S : **Automatic Alignment of Local Representations**. Advances in Neural Information Processing Systems 2003, **15** :841—848, [http://citeseerx.ist.psu.edu/viewdoc/summary?doi=10.1.1.13.897].

[65] Brand M, Brand M : **Charting a Manifold**. Advances in Neural Information Processing Systems 2003, **15** :961—968, [http://citeseerx.ist.psu.edu/viewdoc/summary?doi=10.1.1.11.418].

[66] Clarke R, Ressom HW, Wang A, Xuan J, Liu MC, Gehan EA, Wang Y : **The properties of high-dimensional data spaces : implications for exploring gene and protein expression data**. Nat Rev Cancer 2008, **8** :37–49, [http://dx.doi.org/10.1038/nrc2294].

[67] Tsai F : **Comparative Study of Dimensionality Reduction Techniques for Data Visualization**. Journal of Artificial Intelligence 2010, **3**(3) :119–134, [http://www.scialert.net/fulltext/?doi=jai.2010.119.134&org=11].

[68] Dezso Z, Nikolsky Y, Sviridov E, Shi W, Serebriyskaya T, Dosymbekov D, Bugrim A, Rakhmatulin E, Brennan R, Guryanov A, Li K, Blake J, Samaha R, Nikolskaya T : **A comprehensive functional analysis of tissue specificity of human gene expression**. BMC Biology 2008, **6** :49, [http://www.biomedcentral.com/1741-7007/6/49].

[69] Iyer V, Eisen M, Ross D, Schuler G, Moore T, Lee J, Trent J, Staudt L, Hudson J, Boguski M, Lashkari D, Shalon D, Botstein D, Brown P : **The Transcriptional Program in the Response of Human Fibroblasts to Serum**. Science 1999, **283**(5398) :83–87, [http://www.sciencemag.org/cgi/content/abstract/283/5398/83].

[70] Benecke A : **Genomic Plasticity and Information Processing by Transcription Coregulators**. Complexus 2003, **1**(2) :65–76, [http://content.karger.com/ProdukteDB/produkte.asp?doi=10.1159/000070463].

[71] Benecke A : **Gene regulatory network inference using out of equilibrium statistical mechanics**. HFSP Journal 2008, **2**(4) :183–188.

[72] Asuncion A, Newman D : **UCI Machine Learning Repository** 2007, [http://www.ics.uci.edu/$\sim$mlearn/{MLR}epository.html].

[73] Alberts B, Johnson A, Walter P, Lewis J, Raff M, Roberts K : Molecular Biology of the Cell. Garland Publishing Inc, 5th revised edition edition 2007.

[74] Barabasi A, Oltvai ZN : **Network biology : understanding the cell's functional organization**. Nat Rev Genet 2004, **5**(2) :101–113, [http://dx.doi.org.gate1.inist.fr/10.1038/nrg1272].

[75] Prieto C, Risueno A, Fontanillo C, Rivas JDL : **Human Gene Coexpression Landscape : Confident Network Derived from Tissue Transcriptomic Profiles**. PLoS ONE 2008, **3**(12). [PMID : 19081792 PMCID : 2597745].

[76] Cline MS, Smoot M, Cerami E, Kuchinsky A, Landys N, Workman C, Christmas R, Avila-Campilo I, Creech M, Gross B, Hanspers K, Isserlin R, Kelley R, Killcoyne S, Lotia S, Maere S, Morris J, Ono K, Pavlovic V, Pico AR, Vailaya A, Wang P, Adler A, Conklin BR, Hood L, Kuiper M, Sander C, Schmulevich I, Schwikowski B, Warner GJ, Ideker T, Bader GD : **Integration of biological networks and gene expression data using Cytoscape**. Nat. Protocols 2007, **2**(10) :2366–2382, [http://dx.doi.org/10.1038/nprot.2007.324].

[77] Thomas PD, Campbell MJ, Kejariwal A, Mi H, Karlak B, Daverman R, Diemer K, Muruganujan A, Narechania A : **PANTHER : A Library of Protein Families and Subfamilies Indexed by Function**. Genome Research 2003, **13**(9) :2129 –2141, [http://genome.cshlp.org/content/13/9/2129.abstract].

[78] Thomas PD, Kejariwal A, Guo N, Mi H, Campbell MJ, Muruganujan A, Lazareva-Ulitsky B : **Applications for protein sequence-function evolution data : mRNA/protein expression analysis and coding SNP scoring tools**. Nucleic Acids Research 2006, **34**(Web Server) :W645–W650, [http://nar.oxfordjournals.org/content/34/suppl_2/W645.full].

[79] Hucka M, Finney A, Sauro HM, Bolouri H, Doyle JC, Kitano H, , the rest of the SBML Forum :, Arkin AP, Bornstein BJ, Bray D, Cornish-Bowden A, Cuellar AA, Dronov S, Gilles ED, Ginkel M, Gor V, Goryanin II, Hedley WJ, Hodgman TC, Hofmeyr J, Hunter PJ, Juty NS, Kasberger JL, Kremling A, Kummer U, NovÃ¨re NL, Loew LM, Lucio D, Mendes P, Minch E, Mjolsness ED, Nakayama Y, Nelson MR, Nielsen PF, Sakurada T, Schaff JC, Shapiro BE, Shimizu TS, Spence HD, Stelling J, Takahashi K, Tomita M, Wagner J, Wang J : **The systems biology markup language (SBML) : a medium for representation and exchange of biochemical network models**. Bioinformatics 2003, **19**(4) :524 –531, [http://bioinformatics.oxfordjournals.org.gate1.inist.fr/content/19/4/524.abstract].

[80] Funhashi A, Matsuoka Y, Jouraku A, Morohashi M, Kikuchi N, Kitano H : **CellDesigner 3.5 : A Versatile Modeling Tool for Biochemical Networks**. Proceedings of the IEEE 2008, **96**(8) :1254–1265, [http://cat.inist.fr/?aModele=afficheN\&cpsidt=20592668].

[81] Kitano H, Funahashi A, Matsuoka Y, Oda K : **Using process diagrams for the graphical representation of biological networks**. Nat Biotech 2005, **23**(8) :961–966, [http://dx.doi.org/10.1038/nbt1111].

[82] Watson JD, Crick FH : **Molecular structure of nucleic acids ; a structure for deoxyribose nucleic acid**. Nature 1953, **171**(4356) :737–738, [http://www.ncbi.nlm.nih.gov.gate1.inist.fr/pubmed/13054692]. [PMID : 13054692].

[83] Crick FH : **On protein synthesis**. Symposia of the Society for Experimental Biology 1958, **12** :138–163, [http://www.ncbi.nlm.nih.gov.gate1.inist.fr/pubmed/13580867]. [PMID : 13580867].

[84] Kennedy TD, Lane BG : **The probable 'capping' of wheat leaf messenger ribonucleates by 7-methylguanosine**. Canadian Journal of Biochemistry 1975, **53**(12) :1346–1348. [PMID : 1220858].

[85] Takagaki Y, Ryner LC, Manley JL : **Separation and characterization of a poly(A) polymerase and a cleavage/specificity factor required for pre-mRNA polyadenylation**. Cell 1988, **52**(5) :731–742. [PMID : 2830992].

[86] Schilders G, van Dijk E, Raijmakers R, Pruijn GJM : **Cell and molecular biology of the exosome : how to make or break an RNA**. International Review of Cytology 2006, **251** :159–208, [http://www.ncbi.nlm.nih.gov.gate1.inist.fr/pubmed/16939780]. [PMID : 16939780].

[87] Grabowski PJ, Seiler SR, Sharp PA : **A multicomponent complex is involved in the splicing of messenger RNA precursors**. Cell 1985, **42** :345–353, [http://www.ncbi.nlm.nih.gov.gate1.inist.fr/pubmed/3160482]. [PMID : 3160482].

[88] Murphy WJ, Watkins KP, Agabian N : **Identification of a novel Y branch structure as an intermediate in trypanosome mRNA processing : evidence for trans splicing**. Cell 1986, **47**(4) :517–525. [PMID : 3779835].

[89] Yang E, van Nimwegen E, Zavolan M, Rajewsky N, Schroeder M, Magnasco M, Darnell JE : **Decay rates of human mRNAs : correlation with functional characteristics and sequence attributes**. Genome Research 2003, **13**(8) :1863–1872, [http://www.ncbi.nlm.nih.gov.gate1.inist.fr/pubmed/12902380]. [PMID : 12902380].

[90] Bird A : **Perceptions of epigenetics**. Nature 2007, **447**(7143) :396–398, [http://www.ncbi.nlm.nih.gov.gate1.inist.fr/pubmed/17522671]. [PMID : 17522671].

[91] Fuks F, Hurd PJ, Wolf D, Nan X, Bird AP, Kouzarides T : **The methyl-CpG-binding protein MeCP2 links DNA methylation to histone methylation**. The Journal of Biological Chemistry 2003, **278**(6) :4035–4040, [http://www.ncbi.nlm.nih.gov.gate1.inist.fr/pubmed/12427740]. [PMID : 12427740].

[92] Lee TI, Young RA : **Transcription of eukaryotic protein-coding genes**. Annual Review of Genetics 2000, **34** :77–137, [http://www.ncbi.nlm.nih.gov.gate1.inist.fr/pubmed/11092823]. [PMID : 11092823].

[93] Boyes J, Bird A : **DNA methylation inhibits transcription indirectly via a methyl-CpG binding protein**. Cell 1991, **64**(6) :1123–1134, [http://www.ncbi.nlm.nih.gov.gate1.inist.fr/pubmed/2004419]. [PMID : 2004419].

[94] international human genome sequencing consortium : **Initial sequencing and analysis of the human genome**. Nature 2001, **409**(6822) :860–921, [http://www.ncbi.nlm.nih.gov.gate1.inist.fr/pubmed/11237011]. [PMID : 11237011].

[95] Venter C, celera Genomics : **The sequence of the human genome**. Science (New York, N.Y.) 2001, **291**(5507) :1304–1351, [http://www.ncbi.nlm.nih.gov.gate1.inist.fr/pubmed/11181995]. [PMID : 11181995].

[96] international human genome sequencing consortium : **Finishing the euchromatic sequence of the human genome**. Nature 2004, **431**(7011) :931–945, [http://www.ncbi.nlm.nih.gov.gate1.inist.fr/pubmed/15496913]. [PMID : 15496913].

[97] van Holde K, Zlatanova J : **Chromatin fiber structure : Where is the problem now ?** Seminars in Cell & Developmental Biology 2007, **18**(5) :651–658, [http://www.ncbi.nlm.nih.gov.gate1.inist.fr/pubmed/17905614]. [PMID : 17905614].

[98] Milani P, Chevereau G, Vaillant C, Audit B, Haftek-Terreau Z, Marilley M, Bouvet P, Argoul F, Arneodo A : **Nucleosome positioning by genomic excluding-energy barriers**. Proceedings of the National Academy of Sciences of the United States of America 2009, **106**(52) :22257–22262, [http://www.ncbi.nlm.nih.gov.gate1.inist.fr/pubmed/20018700]. [PMID : 20018700].

[99] Jenuwein T, Allis CD : **Translating the histone code**. Science (New York, N.Y.) 2001, **293**(5532) :1074–1080, [http://www.ncbi.nlm.nih.gov.gate1.inist.fr/pubmed/11498575]. [PMID : 11498575].

[100] Wong H, Victor J, Mozziconacci J : **An all-atom model of the chromatin fiber containing linker histones reveals a versatile structure tuned by the nucleosomal repeat length**. PloS One 2007, **2**(9) :e877. [PMID : 17849006].

[101] Bécavin C, Barbi M, Victor J, Lesne A : **Transcription within Condensed Chromatin : Steric Hindrance Facilitates Elongation**. Biophysical Journal 2010, **98**(5) :824–833.

[102] Bancaud A, Wagner G, e Silva NC, Lavelle C, Wong H, Mozziconacci J, Barbi M, Sivolob A, Cam EL, Mouawad L, Viovy J, Victor J, Prunell A : **Nucleosome Chiral Transition under Positive Torsional Stress in Single Chromatin Fibers**. Molecular Cell 2007, **27** :135–147, [http://www.sciencedirect.com.gate1.inist.fr/science/article/B6WSR-4P48CFV-D/2/acffb9e0ebb1c1b691372f09bfab2a52].

[103] Abbas AK, Lichtman AH, Pillai S : Cellular and molecular immunology. Saunders Elsevier 2007.

[104] **Understanding the Immune System : How It Works**.

[105] Kappe SHI, Vaughan AM, Boddey JA, Cowman AF : **That Was Then But This Is Now : Malaria Research in the Time of an Eradication Agenda**. Science 2010, **328**(5980) :862–866, [http://www.sciencemag.org.gate1.inist.fr/cgi/content/abstract/328/5980/862].

[106] Newton CR, Krishna S : **Severe falciparum malaria in children : current understanding of pathophysiology and supportive treatment**. Pharmacology & Therapeutics 1998, **79** :1–53, [http://www.ncbi.nlm.nih.gov.gate1.inist.fr/pubmed/9719344]. [PMID : 9719344].

[107] Guiyedi V, Chanseaud Y, Fesel C, Snounou G, Rousselle J, Lim P, Koko J, Namane A, Cazenave P, Kombila M, Pied S : **Self-reactivities to the non-erythroid alpha spectrin correlate with cerebral malaria in Gabonese children**. PloS One 2007, **2**(4) :e389, [http://www.ncbi.nlm.nih.gov.gate1.inist.fr/pubmed/17460756]. [PMID : 17460756].

[108] Organization WH : **Severe and complicated malaria. World Health Organization, Division of Control of Tropical Diseases**. Transactions of the Royal Society of Tropical Medicine and Hygiene 1990, **84 Suppl 2** :1–65, [http://www.ncbi.nlm.nih.gov.gate1.inist.fr/pubmed/2219249]. [PMID : 2219249].

[109] Prakash D, Fesel C, Jain R, Cazenave P, Mishra GC, Pied S : **Clusters of cytokines determine malaria severity in Plasmodium falciparum-infected patients from endemic areas of Central India**. The Journal of Infectious Diseases 2006, **194**(2) :198–207, [http://www.ncbi.nlm.nih.gov.gate1.inist.fr/pubmed/16779726]. [PMID : 16779726].

[110] Sakaguchi S, Miyara M, Costantino CM, Hafler DA : **FOXP3+ regulatory T cells in the human immune system**. Nat Rev Immunol 2010, **10**(7) :490–500, [http://dx.doi.org.gate1.inist.fr/10.1038/nri2785].

[111] Trono D, Lint CV, Rouzioux C, Verdin E, Barre-Sinoussi F, Chun T, Chomont N : **HIV Persistence and the Prospect of Long-Term Drug-Free Remissions for HIV-Infected Individuals**. Science 2010, **329**(5988) :174–180, [http://www.sciencemag.org.gate1.inist.fr/cgi/content/abstract/329/5988/174].

[112] Weiss L, Letimier FA, Carriere M, Maiella S, Donkova-Petrini V, Targat B, Benecke A, Rogge L, Levy Y : **In vivo expansion of naive and activated CD4+CD25+FOXP3+ regulatory T cell populations in interleukin-2 treated HIV patients**. Proceedings of the National Academy of Sciences 2010, **107**(23) :10632 –10637, [http://www.pnas.org.gate1.inist.fr/content/107/23/10632.abstract].

[113] Baecher-Allan C, Wolf E, Hafler DA : **MHC class II expression identifies functionally distinct human regulatory T cells**. Journal of Immunology (Baltimore, Md. : 1950)

2006, **176**(8) :4622–4631, [http://www.ncbi.nlm.nih.gov.gate1.inist.fr/pubmed/16585553]. [PMID : 16585553].

[114] Miyara M, Yoshioka Y, Kitoh A, Shima T, Wing K, Niwa A, Parizot C, Taflin C, Heike T, Valeyre D, Mathian A, Nakahata T, Yamaguchi T, Nomura T, Ono M, Amoura Z, Gorochov G, Sakaguchi S : **Functional delineation and differentiation dynamics of human CD4+ T cells expressing the FoxP3 transcription factor**. Immunity 2009, **30**(6) :899–911, [http://www.ncbi.nlm.nih.gov.gate1.inist.fr/pubmed/19464196]. [PMID : 19464196].

[115] Gavin MA, Rasmussen JP, Fontenot JD, Vasta V, Manganiello VC, Beavo JA, Rudensky AY : **Foxp3-dependent programme of regulatory T-cell differentiation**. Nature 2007, **445**(7129) :771–775, [http://www.ncbi.nlm.nih.gov.gate1.inist.fr/pubmed/17220874]. [PMID : 17220874].

[116] Coe B, Antler C : **Spot your genes - an overview of the microarray** 2004, [http://www.scq.ubc.ca/spot-your-genes-an-overview-of-the-microarray/].

[117] de Brevern AG, Hazout S, Malpertuy A : **Influence of microarrays experiments missing values on the stability of gene groups by hierarchical clustering**. BMC Bioinformatics 2004, **5** :114, [http://www.ncbi.nlm.nih.gov/pubmed/15324460]. [PMID : 15324460].

[118] Scheel I, Aldrin M, Glad IK, Sørum R, Lyng H, Frigessi A : **The influence of missing value imputation on detection of differentially expressed genes from microarray data**. Bioinformatics (Oxford, England) 2005, **21**(23) :4272–4279, [http://www.ncbi.nlm.nih.gov/pubmed/16216830]. [PMID : 16216830].

[119] Johansson P, Häkkinen J : **Improving missing value imputation of microarray data by using spot quality weights**. BMC Bioinformatics 2006, **7** :306, [http://www.ncbi.nlm.nih.gov/pubmed/16780582]. [PMID : 16780582].

[120] Nguyen DV CR Wang N : **Evaluation of missing value estimation for microarray data**. Journal of Data Science 2004, **2** :347–370.

[121] Bø TH, Dysvik B, Jonassen I : **LSimpute : accurate estimation of missing values in microarray data with least squares methods**. Nucleic Acids Research 2004, **32**(3) :e34, [http://www.ncbi.nlm.nih.gov/pubmed/14978222]. [PMID : 14978222].

[122] Kim H, Golub GH, Park H : **Missing value estimation for DNA microarray gene expression data : local least squares imputation**. Bioinformatics (Oxford, England) 2005, **21**(2) :187–198, [http://www.ncbi.nlm.nih.gov/pubmed/15333461]. [PMID : 15333461].

[123] Candes E, Recht B : Exact Matrix Completion via Convex Optimization. ArXiv 2008, [http://arxiv.org/abs/0805.4471].

[124] Oba S, aki Sato M, Takemasa I, Monden M, ichi Matsubara K, Ishii S : **A Bayesian missing value estimation method for gene expression profile data**. Bioinformatics (Oxford, England) 2003, **19**(16) :2088–2096, [http://www.ncbi.nlm.nih.gov/pubmed/14594714]. [PMID : 14594714].

[125] Ouyang M, Welsh WJ, Georgopoulos P : **Gaussian mixture clustering and imputation of microarray data**. Bioinformatics (Oxford, England) 2004, **20**(6) :917–923, [http://www.ncbi.nlm.nih.gov/pubmed/14751970]. [PMID : 14751970].

[126] Vapnik VN : The nature of statistical learning theory. Springer 2000.

[127] Wang X, Li A, Jiang Z, Feng H : **Missing value estimation for DNA microarray gene expression data by Support Vector Regression imputation and orthogonal coding scheme**. BMC Bioinformatics 2006, **7** :32, [http://www.ncbi.nlm.nih.gov/pubmed/16426462]. [PMID : 16426462].

[128] Brock G, Shaffer J, Blakesley R, Lotz M, Tseng G : **Which missing value imputation method to use in expression profiles : a comparative study and two selection schemes**. BMC Bioinformatics 2008, **9** :12, [http://www.biomedcentral.com/1471-2105/9/12].

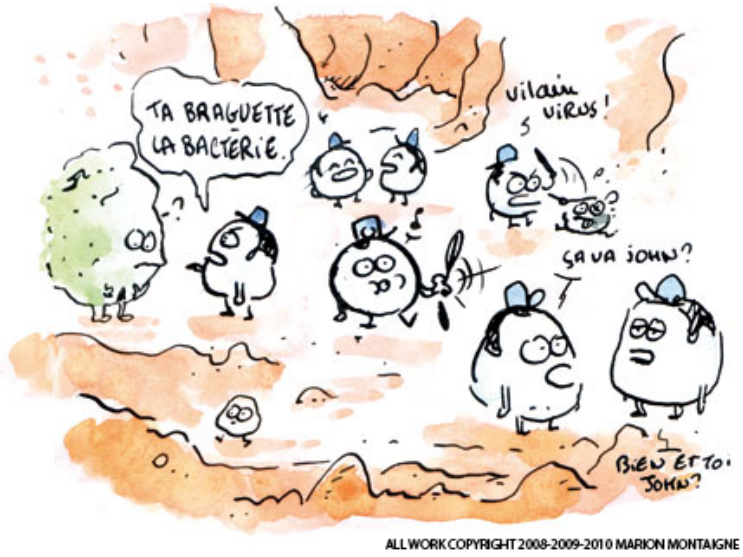
[129] Jörnsten R, Wang H, Welsh WJ, Ouyang M : **DNA microarray data imputation and significance analysis of differential expression**. Bioinformatics (Oxford, England) 2005, **21**(22) :4155–4161, [http://www.ncbi.nlm.nih.gov/pubmed/16118262]. [PMID : 16118262].

[130] Gan X, Liew AW, Yan H : **Microarray missing data imputation based on a set theoretic framework and biological knowledge**. Nucleic Acids Research 2006, **34**(5) :1608–1619, [http://www.ncbi.nlm.nih.gov/pubmed/16549873]. [PMID : 16549873].

[131] Tuikkala J, Elo L, Nevalainen OS, Aittokallio T : **Improving missing value estimation in microarray data with gene ontology**. Bioinformatics (Oxford, England) 2006, **22**(5) :566–572, [http://www.ncbi.nlm.nih.gov/pubmed/16377613].

[132] Xiang Q, Dai X, Deng Y, He C, Wang J, Feng J, Dai Z : **Missing value imputation for microarray gene expression data using histone acetylation information**. BMC Bioinformatics 2008, **9** :252, [http://www.ncbi.nlm.nih.gov/pubmed/18510747]. [PMID : 18510747].

[133] Zhou X, Wang X, Dougherty ER : **Missing-value estimation using linear and non-linear regression with Bayesian gene selection**. Bioinformatics (Oxford, England) 2003, **19**(17) :2302–2307. [PMID : 14630659].



Ce manuscrit est dédié aux globules blancs pour leur travail assidu qui a rendu possible cette thèse. Une mention spéciale à mes propres globules blancs, qui par leur dévouement à toute épreuve, m'ont permis de survivre à ces derniers mois de rédaction.