Modèles markoviens graphiques pour la fusion de données individuelles et d'interactions : application à la classification de gènes

Matthieu Vignes

UNIVERSITE JOSEPH FOURIER
SCIENCES. TECHNOLOGIE. SANTÉ

THÈSE

en vue de l'obtention du diplôme de

Docteur de l'Université Joseph Fourier

présentée par

Matthieu Vignes

Spécialité : Mathématiques Appliquées

Modèles markoviens graphiques pour la fusion de données individuelles et d'interactions : application à la classification de gènes

---

réalisée sous la direction de Florence Forbes et Gilles Celeux
au sein de l'équipe Mistis de l'Inria Rhône-Alpes

---

soutenue publiquement le 30 octobre 2007 devant le

Jury :

Stéphane Robin	Rapporteur
Jean-Philippe Vert	Rapporteur
Olivier François	Examinateur
Didier Piau	Examinateur
Florence Forbes	Directrice
Gilles Celeux	Directeur

TABLE DES MATIÈRES

REMERCIEMENTS

Je tiens à remercier en premier lieu mes directeurs thèse Florence Forbes et Gilles Celeux. Florence a su pendant ces quelques années à la fois m'apporter son soutien scientifique et m'encourager à aller de l'avant tout en respectant mes choix. Gilles depuis Orsay n'a pas manqué de marquer son intérêt dans la bonne progression de ma thèse. Ils ont su me faire profiter d'une condition vraiment très confortable pendant la thèse. En particulier, je leur suis extrêmement reconnaissant pour la rapidité avec laquelle ils ont accepté de me filer un coup de main dans la dernière ligne droite de cette thèse. Je profite aussi de cette occasion pour remercier Alain Viari qui a encadré mon stage de DEA. Son accueil chaleureux au sein de l'équipe Helix, un étage au-dessus de Mistis, a été un facteur indéniable pour bien lancer la thèse. Merci aussi à Fabien Campillo et Alice Guionnet pour m'avoir donné un bon avant-goût du monde de la recherche alors que les Calanques étaient si proches.

Un immense merci à Stéphane Robin et Jean-Philippe Vert d'avoir accepté de rapporter mon manuscrit de thèse. La qualité de leurs remarques sur ce mémoire est très appréciable. Leurs travaux scientifiques sont en outre un excellent exemple pour un chercheur débutant. Je remercie aussi Olivier François et Didier Piau d'avoir accepté d'être membres du jury de cette thèse. Je ne sais pas s'ils se souviennent de létudiant qui assistait à leurs cours avec intérêt (si si) parfois m eme à 8h. Du matin !

Merci aussi Juliette pour le travail effectué ensemble, et ta détermination pour que nous menions à bien des projets scientifiques de qualité. Bon et il n'y a pas que le boulot dans la vie. *I also would like to say a sincere thank you to Jim* McNicol *and those within BioSS who trusted me and hired me when the thesis wasn't over. I particularly appreciated the latitude they allowed me to finish my PhD work in excellent conditions.* Un gros merci aussi à tous les enseignants-chercheurs qui m'ont toujours super bien mis sur les rails pour mes services d'enseignement en étant des collègues accessibles : Stéphane Girard, Serge Dégerine, Carole Desprez-Durand, feu Jean-Claude Paumier, Marianela Fornerino, Marc Humbert, tous les autres chargés de cours, TD et TP que j'ai rencontrés et bien sûr tous «mes» étudiants à qui je souhaite bon vent. Je n'oublie évidemment pas les aides moins directes mais tout aussi essentielles des assistantes de projets ou des documentalistes : Françoise, Élodie, Marie, Alba, Chantal,...

Un salut et merci pour les pauses café, les visites impromptues de bureau, les discussions en s'étirant, les blagues de couloir avec beaucoup des gens sus-cités

mais encore : Guillaume, Charles, Fred, les Jér ome, Mohammed, Gérard, Julien, Émilie, Greg, Olivier, Franck, Henri, Christian, Radu, Édmond, Ben, Alain, Éric, les Anne, Hidde, Grégory, Juho,... Je n'oublie surtout pas aussi les gonzes qui aiment à courir après un ballon de penalty tiré à l'envers, à courir si possible loin du bitume, à vérifier au petit matin que la montagne est toujours aussi belle, à s'organiser un bon Bouchon, à siroter un pastis,...Citons Dav, Barth, Franck, Yvon, Rémi, Didier, Marcello, Julien, petit Charles, Jeanne et Pierre-Emmanuel, Nono, Cathy et Melvil, Laurence, Ludo, Damoune, Stéph, les Leyriolands et Crémolands, les Bonneton, les DeRidder, Pierrot, Séverine, Kelig et Émilie, Peluche et Delphine, Pierre, Fanny et Héloïse, Roger, Jane, Eve and Sandy, le club de hand de Vénissieux et les *Dundee Hawkhill Harriers*. Et pardon à ceux que j'oublie ici je ne sais pas pourquoi ou comment !

Allez c'est presque fini, mais je ne peux pas ne pas m'acquitter d'un merci infini à ma famille qui m'aime et me soutiens sans que je leur témoigne toujours la gratitude qu'ils méritent. Et un gros bisous à mes deux petits franco-écossais (futurs canadiens, australiens, neo-zélandais ? On ne sait pas encore) que j'aime.

0. BRÈVE INTRODUCTION

Le sujet de cette thèse n'a pas coulé de source. Il aura fallu plusieurs discussions avec des chercheurs à l'interface entre les statistiques, la biologie et l'informatique pour dégager une problématique pertinente. Le travail ne m'a jamais semblé aussi colossal que lorsqu'il a fallu rédiger de façon cohérente l'exploration statistique de données post-génomiques qui a été entreprise.

Nous apportons une contribution à l'étude de données biologiques en distinguant des attributs propres à chacun des individus/gènes et des données d'interaction entre ces individus. Nous avons porté notre choix de modélisation sur les champs de Markov cachés et plus généralement sur les modèles à données manquantes. Dans un premier temps, nous nous sommes focalisés sur l'apport des modèles markoviens pour la prise en compte d'interactions. Ceci correspond principalement au premier article à l'annexe C.1 de ce document. Dans un second temps, nous nous sommes intéressés plus particulièrement au problème des observations manquantes, très courant en génomique, et à la façon de prendre en compte efficacement cette particularité et de l'intégrer dans les modèles markoviens envisagés. Cela correspond essentiellement au second article mis dans l'annexe C.2. Notre travail comprend ainsi à la fois une réflexion sur les aspects mathématiques et sur la compréhension des données biologiques utilisées.

Organisation du mémoire

Un premier chapitre présentera la problématique biologique. Nous mettrons en valeur l'apport d'outils statistiques pour les données actuelles à haut débit présentées dans le chapitre 2. Suivra au chapitre 3 la présentation du modèle de champ de Markov cachés dédié à la classification de gènes et les principes de sa mise en œuvre dans notre cadre. Le chapitre 4 proposera des généralisations qui nous sont apparues nécessaires notamment le traitement des données manquantes (section 4.1). Le dernier chapitre présentera les expériences que nous menées aussi bien sur données simulées (partie 5.1) que sur données réelles (5.3). En particulier, ce chapitre présentera notre stratégie de validation. Puis nous ferons le bilan de ce travail de thèse dans le dernier chapitre 6. Des annexes donneront aux lecteurs une vision de l'arrière-plan biologique (A) pour permettre une meilleure compréhension des mécanismes mis en jeu et des précisions sur certains aspects mathématiques que nous utilisons ou mentionnons (B). Enfin nous adjoindrons deux articles qui résument significativement notre contribution.

1. MOTIVATIONS BIOLOGIQUES ET CHOIX DE LA MODÉLISATION

1.1 Que d'informations

Quelques points de repère

Le tournant du millénaire a été décrit comme l'aube d'une nouvelle révolution scientifique. Son impact sur la société devrait être comparable aux révolutions industrielles et à l'apparition des ordinateurs. Cette révolution a été annoncée en juillet 1995 quand le premier effort de **séquençage** à grande échelle a porté ses fruits : les quelques $1,8$ millions de paires de bases du génome de la bactérie *Hæmophilus influenzae* furent alors publiées. C'est la première fois qu'on arrivait à ce type de résultat pour un organisme du monde vivant. Depuis, la quantité de séquences génomiques disponibles dans les bases publiques n'a eu de cesse de croître à un taux exponentiel. Ainsi en avril 2003 fut achevée une étape importante et symbolique de ce qui est connu sous le nom de Projet Génome Humain : la séquence des paires de bases qui composent les vingt-trois chromosomes de l'espèce humaine a été identifiée. Pourtant on ne peut pas dire que le génome humain soit une entité connue loin s'en faut ! Certaines zones sont difficiles à cartographier.

Aujourd'hui on dénombre [1] plus de 10^{11} paires de bases dans les banques de données (*e.g. Genbank*). Cela correspond à presque 30 *HUman Genome Equivalent* (un *HUGE* correspond à la taille du génome humain). À titre de comparaison un *huge* correspond à six années d'édition complète du *New York Times* ([145]). A cette information de séquences ADN viennent s'ajouter des informations de structure -environ 35 000 structures précisant les coordonnées tridimensionnelles de protéines de longueur moyenne 400 résidus dans la *Protein Data Bank*-, des informations de relations entre objets -par exemple homologie entre séquences ou interactions entre protéines-, ou des données reflétant l'activité de la cellule d'un organisme : les données «omiques». Le problème du stockage, de l'accès, de l'indexation et de la mise en relation pour des données aussi nombreuses, précises et variées est une première problématique. En particulier leur disponibilité et leur compréhensibilité pour des chercheurs issus d'horizons pluriels.

[1] Les chiffres donnés ici sont publiés début 2007. Leur évolution est très rapide. Les ordres de grandeur resteront probablement valables sur une période d'un peu moins de deux ans, data à laquelle on peut envisager un doublement par exemple.

Génomique fonctionnelle

Le **gène** est l'entité biologique qui contient l'information relative à une action précise de la cellule. Cette information est exprimée lors de la **transcription** qui provoque la production d'**ARN messager**. À son tour cet ARN messager est traduit en une **protéine** qui est l'agent de base des organismes vivants. Souvent son action n'aura de sens et ne sera d'ailleurs activée qu'en regard de celle d'autres protéines avec lesquelles elle collabore. Nous ne prétendons ici aucunement apporter une définition biologique de la fonction d'une protéine (ou d'un gène par abus de langage). Pour plus de précision, le lecteur pourra se référer à l'annexe A consacrée aux problèmématiques biologiques. On y trouvera aussi une description plus complète des mécanismes du vivant tels qu'ils sont connus actuellement. Nous espérons cette section accessible pour un initier un lecteur sans connaissance préalable particulière en biologie. Pour un lecteur désireux d'aller plus loin, nous conseillons par exemple le traité de génomique de Benjamin Lewin [146].

La **Biologie moléculaire** est caractérisée par l'étude d'objets (les biomolécules) à l'aide d'un ensemble de pratiques ayant certaines propriétés communes. Elle a pourtant bien évolué depuis la redécouverte des travaux de Mendel au début du XX^{e} siècle jusqu'à la mise en œuvre de la technique PCR d'amplification des gènes au cours des années 1980. On pourra lire avec intérêt [168] pour un bon aperçu ou plus modestement la partie A.1 de l'annexe consacrée à la facette biologique de cette thèse. Le **dogme central** de cette «jeune» discipline stipule que la succesion des quatre lettres A, T, C et G qui se succèdent le long de la molécule d'ADN contient toute l'information génétique (voir les figures A.2 et A.4 de l'annexe A). Elle définit structures et fonctions présentes dans un organisme. Ce flux d'information génétique sert de base à la compréhension et l'explication du vivant.

La suite particulière des nucléotides d'un individu, son **génotype** présente une grande variabilité. A ce sujet un projet entier (*HapMap*) cherche à décrire les schémas de variations qui sont communs à tous les individus. Par exemple, les *Single Nucleotide Polymorphism* (SNPs à prononcer [snips]) identifient les sites où les génomes diffèrent d'une seule base. En premier lieu il faut donc prendre en compte ces variations entre individus pour élucider le problème de la recherche de **séquence codante**. Une séquence codante est un fragment de la molécule d'ADN qui donne lieu à la production d'une ou plusieurs protéines, les acteurs dans le monde vivant. Une large majorité de la séquence ADN n'est pas codante. Pourtant de telles zones peuvent revêtir une importance capitale sans que leur rôle exact soit bien identifié pour le moment [2]. On parle de d'ADN «en bazar» (*junk DNA*) ce qui ne signifie surtout pas que son contenu soit inutile : *junk DNA is not rubbish*!

[2] Certains pensent qu'il s'agit en partie de restes de séquences qui furent codantes, les pseudogènes. D'autres le voient à l'inverse comme un réservoir pour l'apparition de nouveaux gènes. Il pourrait aussi servir de garde-fou contre des mutations dommageables du matériel génétique. D'autres interprétation sérieuses ou ésotériques lui sont accordées.

L'ADN tel qu'il nous apparaît est la superposition de nombreux phénomènes. Les régularités, structures sous-jacentes irréductibles entre elles au premier abord peuvent être masquées. Voyons le *junk DNA* plutôt comme la pile d'articles, livres, mémoires et copies d'étudiants, vieil écran d'ordinateurs, formulaires non utilisés voire périmés, cartes postales souvenir ou actes de conférences d'une autre décénie,...qui peuplent le bureau d'un chercheur. Sans être son sujet de travail de l'instant, ils eurent un jour leur utilité et peut-être serviront-ils (à nouveau ?) un jour. Je suis résolument optimiste quant à l'utilisation du *junk DNA* par la nature...

Une fois les séquences codantes (ou gènes) détectées, il faudra accéder à leur rôle au sein du fonctionnement de l'organisme. C'est exactement le thème de la **génomique fonctionnelle**. Cette tâche est loin d'être triviale.

On met souvent en balance génotype (caractères hérités d'un ou plusieurs des ancêtres) et **phénotype** . Le phénotype est l'ensemble des caractéristiques physiques, biochimiques ou physiologiques propres à un individu. Il est conditionné par le patrimoine génétique et par l'environnement qui ont prévalu dans le passé de l'individu. En revanche le phénotype d'un individu n'influence pas son génotype. Plusieurs génotypes peuvent donner naissance à un même phénotype (mutations silencieuses ou combinaisons d'**allèles** [3] particulières). Il n'y a pas de correspondance univoque entre information contenue dans l'ADN et l'activité biologique effective dans la cellule. Seuls les changements dans le génotype peuvent être hérités d'une génération à la suivante. On parle alors d'asymétrie entre génotype et phénotype qui est un moteur de l'évolution, vue comme une altération dans la composition et la distribution du *listing* des gènes dans une population.

On peut expliquer ainsi en partie le déséquilibre fort entre quantité de génomes séquencés et connaissance sur les fonctions des gènes. Accéder à la fonction d'un gène peut être résolu de plusieurs façons.

Elle peut être prédite par homologie de séquence avec un gène pour lequel on dispose de davantage d'information fonctionnelle. On parle alors de **génomique comparative** qui est encore un champ largement actif notamment pour la problématique de similarité de séquences ([73]). Nous renvoyons à l'annexe A.2 pour une présentation détaillée de cette approche et à notamment à [67] ses apports et ses faiblesses.

Une autre façon est de faire des mesures directement sur les produits des gènes : ARN messagers, protéines ou métabolites par exemple puis de les analyser pour chercher une structure dans les données ainsi produites. On procède de manière renversée en analysant les données issues d'un organisme pour inférer

[3] Des allèles sont différentes formes -séquences- pour un même gène. Tout organisme qui possède plus d'une copie d'un gène (qui n'est donc pas haploïde) peut soit contenir plusieurs fois le même allèle (homozygote) indexhomozygote|seeallèle ou des versions différentes (hétérozygote). Si un allèle est dominant et que le gène a un contrôle exclusif sur le trait observé, des individus homozygotes et hétérozygotes auront le même phénotype.

son fonctionnement.

Une nouvelle ère pour la génomique fonctionnelle

En marge des projets de séquençage de nombreux organismes, de nouvelles technologies expérimentales ont donné lieu à une abondance de données nouvelles : transcriptome, protéome ou métabolome par exemple. Cette explosion de **données «omiques»** a provoqué un bouleversement du paradigme [4] de la biologie moléculaire.

Le génome donne la liste des instruments de l'organisme et leur mode d'emploi (dans un langage certes difficile à appréhender pour notre cerveau). Les données «omiques» sont une mesure instantanée de certaines actions de ces instruments ; chacun des types de données donne un angle de vue différent sur ces actions. Une image pour comprendre ceci est de voir un organisme comme un orchestre qui joue une partition. Peut-être lors d'un déchifrement de la partition ? L'observation d'une donnée omique particulière serait l'écoute d'un pupitre. Si on redemande au même pupitre de rejouer la même partition il pourra y avoir des variations (dans le *tempo* ou alors un couac !). Deux pupitres peuvent aussi jouer la même partition simultanément, autrement dit à l'unisson, sans que cela ait la même influence. Enfin l'œuvre ne prend en général son sens que lorsque toutes les parties sont jouées de façon harmonieuse. Il en est de même pour le fonctionnement d'un organisme vivant. Rappelons que nous n'avons que des connaissances très partielles sur la partition, les instruments utilisés voire la finalité du procédé ! Une différence toutefois est que les différentes données «omiques» ne sont pas indépendantes les unes des autres loin s'en faut !

L'approche réductionniste de la biologie a prouvé son efficacité jusqu'au XIX^{e} siècle en particulier dans la deuxième moitié de ce dernier. Elle a permis la description majoritairement qualitative de petits systèmes qui composent les organismes vivants. Elle les considérait isolés. Elle se basait sur des collections de spécimens, l'observation des systèmes et des expériences de laboratoire. Elle organisait la connaissance de façon systématique pour autoriser la formulation de concepts ([45]). Par exemple, on partait d'un phénotype (comme la couleur des fleurs). On imaginait ensuite des expériences qui pourraient mettre en évidence les gènes (avec une conception du gène qui n'avait rien à voir avec celle actuelle) dont les produits sont responsables du phénotype observé.

Cependant la compréhension du tout est dans de nombreux domaines autrement plus complexe que la descrition de tous les éléments constitutifs d'un système. La biologie moléculaire est hautement concernée par cette remarque. On ne pourra pas comparer un organisme à une machine que l'on sait démonter mais bien souvent pas remonter à l'identique (pourtant elle peut fonctionner

[4] Matrice disciplinaire qui comprend la théorie de base qui fait l'unanimité (non remise en question donc), les questions légitimes et les outils aussi bien conceptuels que pratiques qui permettent d'y apporter des éléments de réponse - source : Thomas Kuhn.

alors que des pièces jonchent encore le sol - source : ma Clio). Ou alors il faut essayer de concevoir une machine tellement complexe que la connaissance de très nombreuses personnes sera nécessaire pour assembler petit à petit les pièces d'un gigantesque puzzle, éparpillées sur le sol sans être sûr de toutes les avoir ni que ces pièces proviennent exactement du même modèle.

En fin de compte, on voudra donc **adopter une vue holiste dont l'exploration est suggérée par les données**. Les efforts portent désormais davantage dans l'intégration raisonnée de facteurs pertinents pour mieux interpréter des sytèmes biologiques. Les données génomiques et post-génomiques sont une opportunité pour révolutionner notre compréhension des processus cellulaires qui gouvernent la vie : la **biologie des systèmes** (*System Biology*) procède à l'intégration et à l'analyse de ces données. Elles doivent en effet contenir les moyens nécessaires pour évaluer la contribution des différents gènes *via* l'activité de leurs produits aux processus de fonctionnement des cellules et des organismes. Encore faut-il disposer des méthodes adaptées au traitement de telles masses de données. C'est là une porte ouverte à la collaboration entre le monde de la biologie et des domaines tels que les mathématiques ou l'informatique. La **biostatistique** est née et explore déjà de nombreux sous-problèmes.

La statistique dans tout ça

Les statistiques offrent de nombreux outils qui permettent de traiter rapidement, efficacement et rigoureusement de grosses masses de données dans un contexte stochastique. En effet, nous serons confrontés aux processus aléatoires intrinsèques des phénomènes biologiques ainsi qu'à la forte variabilité des mesures dans les procédures expérimentales. Le lecteur pourra lire l'annexe B.1 pour une vision historique de l'objet des statistiques et la justification de son emploi face à des masses de données telles que celles proposées par la biologie moléculaire. Il pourra aussi consulter la partie consacrée aux données de puces ADN (2.1.3) et se rendre compte que presque toutes les étapes techniques sont source d'une variabilité des mesures : extraction de l'ARN, efficacité de la rétro-transcription, hybridation sur les lames plus ou moins réussie alors que les conditions (température, humidité) semblent bien contrôlées, défaut de la lame elle-même ou lors de la phase d'acquisition,.... Celle-ci vient se superposer à celle émanant des variations biologiques ([2,215]). Autant de facteurs qu'un *design* expérimental adéquat pourrait essayer de limiter ou au moins d'en contrôler le biais systématique. Ainsi les statistiques ne sont pas seulement nécessaires en aval pour l'analyse des données post-génomiques mais en amont pour un meilleur traitement d'effets systématiques (les biais) ou aléatoires (bruit cellulaire).

Il faut aussi souligner la haute dimensionalité des données à haut débit. On voudra en revanche fournir des explications comparativement simples ([125]) à des niveaux d'organisation plus élevés. Par exemple pour dire si un gène est concerné ou non par une propriété de tolérance au froid.

Le souci du biostatisticien sera de fournir des méthodes statistiques, des réponses quantitatives adaptées aux vastes quantités de données de biologie des systèmes. Par exemple, une première tâche consistera en un pré-traitement des données ; une **normalisation** adéquate rendra les données standards pour que des comparaisons raisonnables puissent être effectuées (mais ne pas confondre normalisation qui transforme les données pour qu'elles varient toutes entre 0 et 1 en divisant la différence à la plus petite par l'étendue et la **standardisation** qui transforme les données pour qu'elles soient globalement de moyenne nulle et d'écart-type 1). Il faudra comprendre les procédés de production des données et les processus biologiques mis en jeu pour être à l'aise lors de l'interprétation des analyses. Leur utilisation ultérieure pour des applications biomédicales est aussi un facteur important.

Nous avons établi le cadre général dans lequel s'inscrit notre travail. Il reste désormais à situer précisément notre angle d'approche et expliciter plus avant le sujet de ce mémoire.

1.2 Notre choix de modélisation

Des données bruitées

Les connaissances théoriques actuelles sur l'organisation de la vie au niveau cellulaire sont plutôt limitées. On voudrait les compléter. Nous avons mis en évidence des progrès technologiques récents qui permettent de disposer de nombreuses données «omiques». Ces dernières sont de véritables indicateurs du fonctionnement des organismes vivants. Elles sont des mesures instantanées de l'état de la cellule dans un environnement bien identifié. Elles sont en quelques sortes des clichés qui donnent des points de vue complémentaires sur tous les acteurs des processus biologiques (ADN, ARN, protéines, métabolites,...) mis en jeu dans l'expérience menée ainsi que sur leurs interactions.

Une difficulté majeure est d'extraire l'information issue de ces grandes quantités de données à haut débit [5] qui ont pour trait inhérent d'être bruitées. En effet il est maintenant reconnu que les technologies pour acquérir les données ([2,145,160]) mais aussi les processus biologiques ([63,119,188]) eux-mêmes ont un comportement stochastique. Un modèle probabiliste sera alors particulièrement adapté pour rendre compte de tous ces signaux superposés. Un processus d'**inférence** (*i.e.* d'apprentissage) statistique sera utilisé pour faire de l'ajustement de modèle et de l'apprentissage à partir d'exemples.

[5] On parle communément de données à haut débit (*high-throughput data*) pour les données «omiques» ; les technologies modernes permettent en effet d'avoir accès simultanément à tous les niveaux d'abondance d'un certain type de biomolécules (*e.g.* séquences d'ARNm pour les puces à ADN, séquences protéiques ou métabolites par spectrométrie de masse,...)

Des voies moléculaires sous-jacentes pour rendre compte des données

Les **voies moléculaires** sont des séquences de réactions biochimiques successives. Souvent catalysées par des protéines appelées enzymes, elles sont à la base de la plupart des fonctions essentielles des cellules d'organismes vivants. C'est en tout cas l'hypothèse largement admise au niveau de la modélisation biologique que nous adopterons. Pour comprendre les mécanismes de transcription des gènes et de synthèse des protéines ainsi que leur finalité, on devra donc modéliser les interactions entre ces composants. La tâche ne se limite pas à mettre en évidence les réseaux responsables de la complexité observée du monde vivant. On veut aussi les interpréter.

Au niveau le plus précis, on pourra décrire les réseaux de régulation par des systèmes d'équations différentielles couplées qui traduisent les équilibres chimiques mis en jeu. On pourra consulter [63] pour une revue de telles approches dynamiques et [164] pour des simulations de données métaboliques. Indispensables pour une description complète des réseaux de régulation, elles requièrent un niveau de détail très élevé. Les relations entre les entités interagissantes doivent être caractérisées et les paramètres d'équilibre des réactions précisés. Non seulement ces valeurs sont difficiles à évaluer mais encore cette méthode ne peut être mise en œuvre que pour des réseaux où peu d'acteurs interviennent (*i.e.* au plus de l'ordre de la dizaine). Citons avec un peu d'avance la base de données la plus généraliste sur les voies métaboliques : KEGG (pour *Kyoto Encyclopedia of Genes and Genomes*, [120], `http://www.genome.jp/kegg/`).

Une approche alternative de ces réseaux consiste à utiliser dans un premier temps des méthodes de classification. Elles consistent à rassembler les individus (gènes) en un certain nombre de groupes. Les éléments de chacun de ces groupes seront plus semblables entre eux qu'à des éléments de groupes différents. Éventuellement de telles similarités internes permettront de donner une interprétation biologique aux groupes ainsi formés. En tant que méthodes d'exploration ou visualisation raisonnée des données, elles sont nettement plus simples à mettre en place qu'une modélisation exhaustive de la dynamique des voies métaboliques. En contre-partie, leur résolution est moins fine. Depuis les travaux pionniers de classification hiérarchique de [74], on a assisté à l'utilisation d'un vaste panel d'approches de classification parfois spécifiquement conçues pour les données d'expression : algorithme *k-means* [225] (ou [64] pour l'algorithme *fuzzy c-means* lui même généralisé par méthode de Gustafson-Kessel dans [128]), ACP (Analyse en Composantes Principales [117]) et généralisations dans [106, 109, 257], cartes auto-organisatrices (ou SOM pour *Self-Organizing Maps* utilisées dans [134, 221]. D'autres références incluent [91, 175, 254] qui utilisent les modèles de mélange ([159]), [150] basée sur un algorithme de recuit simulé ([131]), [99,246,263] basées sur la théorie des graphes (*e.g. minimum spanning tree* ou extraction de la plus grande composante connexe). Des travaux préconisent l'utilisation d'une distance particluière dans les méthodes de classification à base de distance comme [219]

avec l'information mutuelle. Des solutions propres à l'analyse de matrices de données d'expression sont proposées : *clustering* hiérarchique à noyaux ([183]), classification *diamétrale* ([69]) ou quantique avec décomposition en valeurs singulières ([107]), outil CLICK & EXPANDER ([211]). Parmi les approches supervisées, nous pouvons citer les SVM (*Support Vector Machines*) par [42] ou les méthodes à noyaux ([134, 234]). La classification de gènes est un moyen peu coûteux d'extraire l'information des puces à ADN en supposant que la co-expression implique la co-régulation ; les gènes à expression semblables sont en acceptant cette hypothèse reliés fonctionnellement. De ce point de vue, le profil d'expression moyen d'un *cluster* peut être bien informatif.

Une hypothèse sous-jacente de nombreux algorithmes est que les gènes sont considérés indépendants. Nous préfèrerons renforcer le réalisme du modèle en ne faisant plus cette hypothèse. Nous pensons que cela aidera à identifier des groupes de gènes à fonctions similaires ou impliqués dans des processus proches. Des travaux ont le souci d'utiliser les données d'interaction mais la plupart du temps *a posteriori* pour valider une classification de manière externe ([87, 138]). Les données individuelles peuvent aussi être transformées en tableau de similarité ce qui diminue l'information initiale et pose en outre le problème du choix d'une mesure adaptée. Par exemple [98] fait l'utilisation d'un algorithme de *co-clustering*.

En ce sens, les approches développées jusqu'alors sont sous-optimales et des paramètres doivent être spécifiés selon le type d'analyse. Nous voudrons alors proposer une méthode qui aura pour but de **prendre en compte dans un même modèle statistique des mesures attribuées individuellement et des informations d'interactions**.

Nous devons mentionner les méthodes à noyaux ([134, 139, 250]) qui ont la bonne caractéristique d'intégrer tous les types de données dans une même procédure. Mais il s'agit d'analyse dans un cadre supervisé. Nous voulons faire l'hypothèse qu'aucune autre information *a priori* que celle utilisée dans le modèle n'est disponible. On pourra consulter [125] pour une discussion comparative des approches supervisées ou non dans le cadre de la génomique fonctionnelle.

Modèles probabilistes graphiques

Afin d'intégrer au mieux traits individuels et données d'interaction, nous nous appuierons sur des modèles probabilistes graphiques. Leur développement est toujours un thème de recherche très actif ([30, 54, 147, 238]). Leur souplesse permet de forger un modèle adapté au problème considéré en le décrivant à l'aide d'une structure qui capte les dépendances entre les variables mises en jeu de façon relativement intuitive.

On pourra par exemple considérer simultanément des données individuelles (*e.g.* données d'expression) et des données d'interactions (*e.g.* en provenance d'un ou plusieurs réseau(x) biologique(s)). Nous nous intéressons plus précisément au modèle de champ de Markov caché dans lequel une distribution de probabilités

paramétrique rend compte de la distribution des données individuelles. Les données d'interaction qui peuvent refléter une distance ou une mesure de similarité entre gènes sont représentées par un graphe. Chaque nœud représente un gène tandis que chaque arête peut être pondérée selon qu'une information quantitative sur l'interaction est disponible ou non.

Notre choix d'adopter un modèle probabiliste a l'avantage de fournir un cadre théorique intéressant. Sur le plan pratique, des expériences aussi bien sur données simulées que sur données réelles nous ont fourni des résultats encourageants quant au gain dans l'utilisation de notre approche pour le type de problématique biologique envisagé. Avant d'attaquer le corps de ce mémoire, nous voulons préciser les relations qui nous semblent pertinentes entre les entités biologiques.

1.3 Le contexte d'un gène

Dans *La Barque de Delphes* ([58]), Antoine DANCHIN pose un peu la même question que celle adressée à l'oracle. La putréfaction naturelle des planches du bâteau d'un pêcheur requiert leur remplacement sporadique. Le bâteau se trouve donc légèrement transformé après une telle d'opération. À partir de quand peut-on dire que l'on n'a plus affaire au même bâteau ? Aussi incongrue que peut sembler cette question dans notre contexte, elle semble être au sein de la biologie moderne. Si chaque molécule d'une cellule est remplacée, a-t-on toujours la même cellule ? Si chaque cellule d'un organisme est remplacée, est-on confronté au même organisme ? Si un grand nombre d'individus d'un groupe se renouvellent après une génération, a-t-on la même population ?

Ces questions peuvent sembler éloignées et trop philosophiques pour notre problématique ; pourtant les réponses qui y sont apportées jouent un rôle central dans bon nombre d'axes de recherche bioinformatique. Ce ne sont pas les éléments constitutifs (les planches) qui explique le mieux un système (la Barque de Delphes). Ils peuvent être remplacés sans que l'aspect ou la fonction du bateau ne soient bouleversés. En revanche les relations entre les planches sont primordiales. Elles rendent la barque spécifique et opérationnelle.

De même, les interactions entre les objets du vivant rendent mieux compte du fonctionnement d'un organisme que l'énumération de ces objets et de certaines de leurs propriétés. L'approche réductioniste est éclipsée au profit d'une approche intégrative naturelle ; l'analyse des objets est conjointe à celle d'un réseau qui fait le lien entre eux. Par exemple une relation entre composants chimiques peut être décrite par des réactions. Elles régissent tous les équilibres en matière et en charge électrique des constituants de la cellule. Le taux relatif de contribution de chaque réaction à cet équilibre global est thermodynamiquement contraint par des conditions cellulaires comme le pH ou la température. Aussi une réaction est conditionnée par la présence (ou l'absence) d'une enzyme qui la catalyse. L'équilibre de la réaction est immédiatement affecté par la quantité ou l'état de cette

enzyme. On voudra alors dans cette optique comprendre le réseau d'interactions des voies métaboliques. C'est un défi de taille à l'échelle d'un organisme. La complexité du réseau grandit très vite avec son nombre d'éléments. La multitude de phénotypes observables résulte de l'utilisation imbriquée de ces voies différemment pour chaque individu. La diversité est permise par dérive génétique. Elle est pourtant limitée par la «sélection optimale» de certains états qui sont seuls observés.

En traitement automatique des langues (TAL), [6], le contexte sera primordial en biologie moléculaire. Explorer un génome est un peu comme explorer un nouveau continent : plutôt que de chercher à vérifier des hypothèses préétablies, on cherchera à en dresser un premier aperçu ([194]). On parlera de biologie des systèmes pour l'étude globale d'un système vivant complexe *via* l'intégration de connaissances et de données de différents types ([132]).

Une hypothèse peu évidente *per se* mais très bien mise à jour depuis l'étude des mécanismes de l'hérédité est que le plan d'organisation de la cellule est accessible sur le(s) chromosome(s). Ceux-ci sont en effet le seul matériel susceptible de contenir toute cette information transmise d'une génération à l'autre. Il faudra déterminer une séquence d'ADN et l'emplacement des gènes qu'elle contient : le **séquençage**. Pour simplifier on considèrera que chaque gène code pour une protéine (voir le dogme de la biologie moléculaire à ce propos dans l'annexe A).

Suit l'analyse fonctionnelle de chacune des protéines. Quand son rôle précis est déterminé, une protéine (ou par abus le gène) est dite annotée. Une telle tâche est loin d'être anecdotique. Le traitement expérimental exhaustif est impossible. Souvent la fonction biochimique remplie par une protéine dépend de son contexte. Cette notion de contexte au sens encore large reste à préciser. Il pourra s'agir de sa localisation cellulaire, des autres molécules avec lesquelles elle interagit, de sa position sur le chromosome, *etc.* ou d'un voisinage combiné qui prend en compte plusieurs de ces facteurs. Il en résulte que ce contexte ne peut pas être ignoré. La fonction d'une protéine est une notion revue : plutôt qu'un composant chimique isolé, elle devient acteur d'un réseau d'interactions. On se placera donc à un niveau cellulaire par l'intégration d'interactions entre entités moléculaires : voies métaboliques, cascades de signalisation, processus cellulaires,...Nous utiliserons un cadre précis, quantitatif au besoin de représentation des interactions. L'outil mathématique apparaît là aussi nécessaire.

Une étape préalable à notre étude est donc la construction du réseau biologiquement pertinent et exploitable. Pour cela nous nous servirons de bases de données d'interaction entre composants biologiques. Ce travail est décrit à la partie 3.7.

[6] Comment ranger un texte où le mot «avocat» est souvent répété ? Le terme «centrale» est-il important (s'il s'agit du nom comme dans «centrale nucléaire») ou peut-il être oublié (s'il est adjectif) ? Ne doit-on considérer que les lexèmes (astre, astronome, astrologie ne diffèrent que par leur suffixe) mais alors ne confondra-t-on cosmonaute et cosmétique ?

De nombreux indices suggèrent un rapport entre distribution des gènes et architecture de la cellule ([195]) ; la sélection naturelle peut l'expliquer. L'agencement des gènes sur le(s) chromosome(s) devrait être optimisé en vue de la production de leurs formes opérationelles : ARN et protéines. On rapproche donc le point de vue séquentiel du point de vue fonctionnel ! Une fonction sera vue comme une succession de processus organisés de façon à coopérer vers un but commun ([59]).

En fait de simples considérations physiques rendent l'hypothèse d'une importante organisation au sein de la cellule quasiment indispensable. Un chromosome peut être vu comme un brin très fin dont la longueur est de l'ordre de mille fois la taille typique d'une cellule. Les nécessaires repliements doivent donc être contrôlés afin d'éviter tout enchevêtrement. Plus essentiel encore sera la possibilité de lecture des chromosomes pour produire des milliers d'ARN messagers qui gouverneront à leur tour la production des protéines correspondantes ([167]).

En somme, nous voulons pénétrer l'organisation des génomes à l'aide de paramètres propres aux gènes et des interactions mises en jeu. On dépasse la notion de «sac de gènes» ([194]) puisque l'expression du programme génétique dépend des conditions dans lesquelles se trouve l'organisme. Une idée maîtresse est que plus un gène partage d'interacteurs communs avec un autre, plus leur rôle doit être semblable. L'information pourra être captée directement sur la séquence (si on mesure le biais en codons par exemple, voir la page 174) ou avec des mesures propres au gènes (telles que le niveau d'expression de chacun des ARN messagers). On parlera alors de **données individuelles** dans notre modèle. On fera aussi intervenir des **données pairées** pour élaborer le réseau. C'est à dire qu'on ne va plus seulement opérer un transfert des connaissances entre des objets similaires. Notre objectif est d'intégrer dans un même modèle ces types de données complémentaires pour faire de la classification de gènes. Nous verrons que cela permettra d'interpréter les regroupements occasionnés.

Nous décrivons dans le chapitre à venir les différentes données disponibles. En particulier nous expliquons leur intérêt individuel ou en complément les unes des autres conformément à des travaux dont les applications biologiques sont variées : reconstruction de réseaux de régulation [88, 98, 112], prédiction d'unités de transcription [35, 56, 199], recherche de groupes de gènes d'une même voie métabolique [104, 209], prédiction de fonctions de protéines [84, 139, 154, 177] ou d'interactions [75, 113, 222, 248], *etc.*

2. PROBLÉMATIQUE ASSOCIÉE AUX DONNÉES

Le point de départ de notre étude est l'examen des données. Il nous faut préciser lesquelles. Le lecteur désireux d'avoir un aperçu plus approfondi des mécanismes biologiques de production des données est invité à se réfèrer à l'annexe A ou à un livre de biologie moléculaire ([146]). À notre niveau, nous accèderons à de volumineux fichiers informatiques en libre accès sur des sites de bases de données, (nous avons éventuellement accepté un enregistrement en tant qu'utilisateur). Ces fichiers rassemblent un grand nombre d'informations sur les organismes concernés ainsi que sur les différentes mesures effectuées selon les diverses préoccupations. Une base de données peut être spécifique à un organisme. Par exemple pour *Escherichia coli* et *Bacillus subtilis*, on pourra regarder *Indigo* ([173]). Si on souhaite étudier levure de boulanger, la base *Saccharomyces Genome Database* (SGD, `http://www.yeastgenome.org/`) est incontournable. Les bases peuvent aussi se rapporter à un type particulier de données : (i) d'interactions (*Molecular INTeraction*, MINT, `http://mint.bio.uniroma2.it/mint/`) ou *Database of Interacting Proteins*, DIP, `http://dip.doe-mbi.ucla.edu`), (ii) de métabolisme (*Kyoto Encyclopedia of Genes and Genomes*, KEGG, `http://www.genome.ad.jp/kegg/`) ou (iii) de puces ADN (*Stanford Microarray Database*, SMD, `http://genome-www5.stanford.edu/`).

Enfin des bases de données sont à vocation de rassemblement de plusieurs types d'informations. Ce sont des interfaces vers plusieurs autres bases de données souvent avec une recherche centralisée : au *National Center for Biotechnology Information* (NCBI, `http://www.ncbi.nlm.nih.gov/`) ou à l'*European Bioinformatics Institute* (EBI, `http://www.ebu.ac.uk/Databases/`). Ces dernières illustrent d'ailleurs bien les fortes inter-dépendances entre tous les objets contenus dans ces bases : gènes, protéines, *Single Nucleotide Polymorphism*, *Expressed Sequence Tags*, *etc.* Ou entre les données concernées : séquence, usage du codon, localisation cellulaire, expression d'ARNm, voie métabolique,... Une première étape sera de structurer ces données sous une forme pertinente et utilisable pour mener nos expériences. Pour cela il a fallu les comprendre ; ce fut une occupation non négligeable pendant ces quelques années de thèse.

Les données disponibles sont d'abord les séquences ADN des organismes étudiés. Ces séquences biologiques constituent une masse considérable d'informations sur lesquelles nous ne possédons que des bribes de connaissances. En effet les données acquises n'ont pu être que partiellement traitées. Pourtant les enjeux sont importants et de différentes natures :

- fondamentaux du point de vue biologique : on voudrait comprendre les mécanismes de la vie, les relations de parentés entre les espèces par exemple. Ce type d'informations est supposé être déposé dans les génomes puisque ces derniers sont à la base de la conception des organismes et transmis d'une génération à la suivante.
- médicaux : comme en atteste toutes les applications des thérapies géniques ciblées. On cherche par exemple à caractériser un cancer au niveau moléculaire ([6]). Cela mènerait de façon idéale à un nouveau traitement dirigé vers une aberration «cellule-spécifique».
- agricoles : on voudrait disposer de plantes «améliorées» en contrôlant le mieux possible les changements induits.
- pharmaceutiques : on voudrait comprendre l'efficacité de certains médicaments face à certaines maladies. Bien sûr on n'agit pas sur les gènes au hasard mais conformément au plan d'organisation de la cellule.
- industriels : en vue de produire et utiliser des organismes vivants qui facilitent la production de certaines molécules.

Une question première est de savoir sous quelle(s) forme(s) utiliser les données. Au premier abord, elles sont stockées dans de vastes fichiers texte assez diffus. Il se dégage une volonté récente de structurer le mieux possible ces données, de normaliser le plus possible leur format et surtout de les commenter et les relier les unes aux autres de façon pertinente. Mais c'est encore un *eldorado* lointain (sinon inaccessible de notre modeste point de vue).

Nous assistons encore aujourd'hui au séquençage toujours plus rapide de nombreux génomes. Cette phase peut être effectuée grâce à la segmentation de l'ADN en zones homogènes pour organiser les extraits de séquences élucidés ; une molécule d'ADN ne se détermine pas en un seul morceau ([41]). La segmentation se base sur l'hypothèse que la distribution des bases constitutives de la séquence est proche dans des domaines. Des changements de régime sont observés aux discontinuités sur la séquence nucléique. Le chapitre 4 de [58] est consacré à cette aventure du séquençage. Cette partie a déjà été largement étudiée. Les progrès techniques font que les projets «génomes» ont des productivités sans cesse améliorées. Mais l'obtention du texte brut d'un génome est de loin insuffisante : on n'obtient rien d'autre en soi qu'un long texte constitué de quelques centaines de milliers à quelques dizaines de millions de caractères dans un alphabet à 4 lettres : A, C, G et T. Des traitements ultérieurs sont indispensables : on veut donner un sens au texte obtenu.

On voudra détecter les zones qui jouent un rôle dans les processus biologiques : zones codantes ou gènes mais aussi promoteurs, terminateurs, *etc.* devront être identifiés. On pourra se référer à [158] pour une revue actuelle de nombreuses méthodes dans le domaine de la **détection de gènes**. Mentionnons seulement des modèles de Markov cachés adaptés à de telles problématiques. Ils ont prouvé

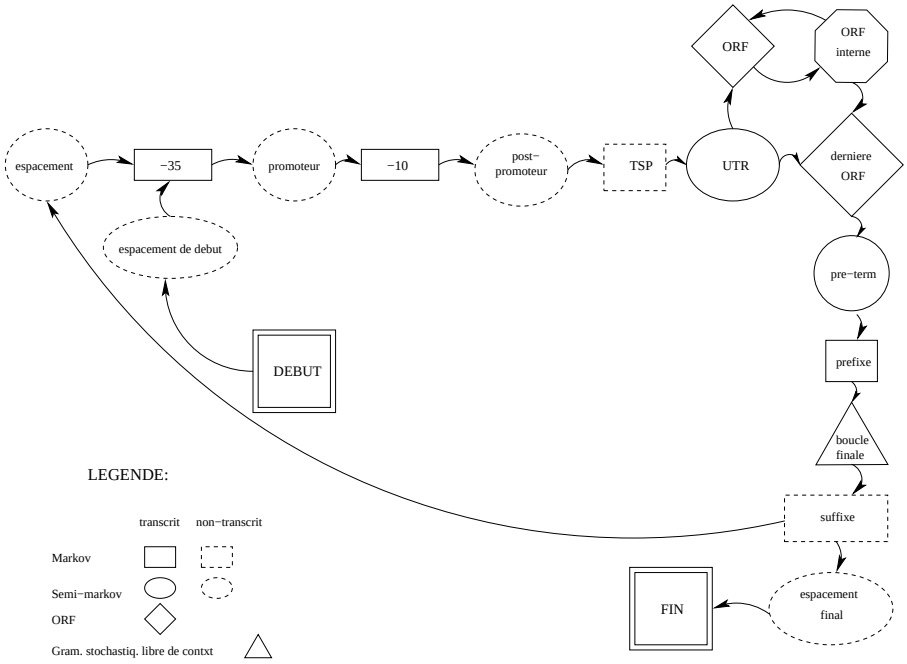


Fig. 2.1: Modèle global de structure d'un gène (assemblage de modèles locaux)

leur efficacité ([149, 200]) et un apport complémentaire des méthodes basées sur l'homologie de séquences (décrite dans l'annexe A.2). Les connaissances biologiques actuelles permettent de modéliser avec un bon niveau de détail la structure physique de l'environnement d'un gène (voir la figure 2.1). Aujourd'hui, de tels éléments sont intégrés pour une meilleure détection des gènes ([247] par exemple).

Puis on voudra obtenir le maximum d'informations sur les gènes ainsi localisés : comment ils sont activés, leur influence les uns sur les autres, la tâche biochimique accomplie dans l'organisme,... On veut comprendre la réalité physique du génome. On ne se limite pas au simple catalogage des gènes d'un organisme. On commence ici à aborder pleinement la problématique qui nous concerne : **entrevoir l'organisation du génome** *via* **des mesures traduisant les niveaux de son fonctionnement**. En plus d'avoir des informations sur les éléments qui le compose, on voudra pénétrer les relations entre ses éléments. La biologie intégrative ainsi présentée fera aussi un usage adéquat des techniques récemment développées : transcriptome, protéome,...Ces informations donneront un angle d'éclairage complémentaire et enrichiront les accès aux propriétés globales des organismes.

L'annotation est actuellement un traitement en aval du séquençage plus lent que ce dernier et conceptuellement moins avancée. Elle provoque un goulot d'étranglement pour la production d'informations génomiques opérationnelles ([194, 261]) : nous croulons sous l'information mais nous manquons cruellement de

connaissances. On pensait connaître la quasi-totalité des fonctions du vivant il y a une trentaine d'années juste avant les balbutiements du séquençage. On estime aujourd'hui qu'au moins 35% des protéines parmi les génomes actuellement connus ont encore une fonction complètement indéterminée ([155]). Mais il faut dire que cette étape est particulièrement ardue mais cruciale. D'une part elle demande de créer plutôt que de déterminer. En effet on ne connaît pas forcément les termes pour décrire la connaissance sur des objets dont on vient juste d'apprendre l'existence lors du séquençage. D'autre part cette information devra être de bonne qualité puisqu'elle conditionne les nombreuses analyses en aval. Notons que la qualité de la connaissance extraite varie grandement d'un projet à l'autre.

De manière générale, les méthodes les plus fiables pour l'annotation sont celles qui font intervenir des expérimentations biochimiques directes sur le gène ou la protéine à déterminer. Elles sont dites *in vivo* pour l'observation de phénomènes au sein des organismes ou *in vitro* pour l'observation des mêmes phénomènes en environnement artificiel. Des *artefacts* peuvent bien entendu apparaître en milieu artificiel qui ne peut pas reproduire toutes les conditions du vivant. Mais c'est aussi une de ses richesses en proposant une meilleure maîtrise des paramètres expérimentaux.

Cependant de telles méthodes laborantines sont incomparablement plus lourdes à mettre en place et/ou prennent énormément de temps. Pour traiter des données en très grandes quantités on préfèrera des prédictions automatiques. Celles-ci utiliseront un critère pour décider si une annotation convient ou non en regard des données disponibles. La précision de telles prédictions est difficile à évaluer en général. Pour juger de la qualité des méthodes, celles-ci sont testées sur des jeux de données connus. Cela sera notre point de vue et nous espérons que les résultats obtenus par notre méthode encourageront une utilisation ou des dévelopements ultérieurs sur des jeux de données originaux. Nous présenterons des méthodes pour évaluer les algorithmes de classification de données biologiques lors de nos expériences au chapitre 5.

Nous utiliserons et développerons des outils bioinformatiques. C'est à dire qu'on est au cœur des méthodes assistées par ordinateur des données en provenance « des » biologies cellulaire et moléculaire. Ces expériences seront dites *in silico* ([58]). Le chapitre 5 présente les expériences menées pour évaluer les prédictions qui découlent de notre modélisation.

Nous l'avons déjà mentionné : une des difficultés rencontrées provient du fait que la masse considérable de données est issue d'expériences et de laboratoires variés. Leur format -séquences, images, fichiers textes, citations bibliographiques, *etc*- ainsi que leur qualité, leur traçabilité voire leur fiabilité sont assez hétérogènes. Il faudra donc trouver une méthode de traitement à grande échelle qui prenne en compte ces disparités. C'est encore un avantage du choix d'un modèle probabiliste. Notre travail s'inscrira comme une partie de la génomique fonctionnelle. On voudra intégrer de gros ensembles de données valid(é)es issus d'expé-

riences variées. Une vision plus globale des processus biologiques pourra alors être accessible à condition de comprendre au moins en partie les phénomènes mis en jeu dans les expériences. On n'applique pas uniquement des méthodes numériques à des données, il faut saisir l'information extractible.

Un des apports de cette thèse a été de traiter une partie de ces données. En les extrayant directement des bases de données biologiques, nous leur avons cherché une structure. Nous proposons d'ailleurs de mettre à la disposition de chercheurs les outils pour se créer ces jeux de données. Ils peuvent intéresser ceux qui souhaitent tester les performances de leur approche comme nous l'avons fait ou des biologistes désireux d'explorer leurs propres données.

Quelques informations biologiques disponibles

Nous distinguons la présentation des données selon qu'elles sont du type «mesure individuelle» à une entité biologique étudiée ou de type «donnée d'interaction». Tandis que les premières émanereront principalement de données «omiques», les secondes seront issues d'expériences qui révèleront les dépendances entre ces entités. Nous verrons d'ailleurs qu'aussi bien ces données individuelles que celles d'interaction peuvent être relatives à des éléments de nature différente. Nous choisirons de trouver les correspondances adéquates pour que toutes les données se rapportent aux gènes.

2.1 Données propres à chacune des entités

2.1.1 Lues directement sur la séquence

La séquence ADN qui porte un gène est composée d'une succession de nucléotides parmi 4, symbolisés par les lettres A, C, G ou T. Caractériser cette séquence par sa composition particulière en ces quatre bases est une démarche naturelle. Par exemple, on pourra donner les fréquences relatives d'utilisation de ces quatre nuclétotides. Certaines zones présentent en effet un fort biais envers certains de ces nucléotides. Compte tenu de la complexité biologique d'un brin d'ADN, l'information portée par ces fréquences relatives ne constituera que la partie visible de l'iceberg. On observe par exemple que la succession des nucléotides a une importance particulière dans des régularités recherchées. Par exemple les *CpG islands* ([29]) sont des régions caractérisées par une proximité inhabituelle de la Cytosine et de la Guanine. Elles sont souvent localisées à proximité de promoteurs de gènes essentiels pour des fonctions cellulaires importantes. De même le cadre de lecture influence beaucoup la loi de répartition des bases. On peut par exemple se servir du biais d'usage du codon (voir page 174) dans les *Open Reading Frame* (ou cadres le lecture) pour prédire quels gènes seront fortement exprimés ([122, 171]). Des éléments précisant la régulation (voir la page

174) peuvent être très informatifs. Mais de telles informations ne sont pas toujours disponibles (en grand nombre). En outre, on verra qu'il est surement plus pertinent de les intégrer dans les données d'interactions.

Concernant l'**usage du codon** , notons que l'enchaînement des propriétés physico-chimiques des acides aminés [1] fait que certaines combinaisons sont possibles/sélectionnées alors que d'autres sont impossibles/défavorisées. L'environnement acceptable de la protéine résultante s'en trouve grandement conditionné. En conséquence sa fonction aussi. Ces informations sont intrinsèques à la séquence protéique. Par exemple elle peuvent servir à séparer les protéines, prédire les séquences transmembranaires ou la structure secondaire des protéines.

Pour caractériser simplement une séquence d'ADN, on pourra se servir d'une mesure telle que le *Codon Adaptative Index* ([212]). Il résume simplement l'usage du codon de la séquence sur laquelle il est calculé. Il fournirait une mesure individuelle unidimensionnelle. D'autres travaux sur l'usage du codon peuvent par exemple être trouvés dans [162]. D'autres ([151, 192]) étudient la distribution de certains motifs. Il faudrait une mesure associée à un gène hors ces travaux ne concerne souvent pas forcément les séquences codantes bien au contraire. Une étude statistique de l'utilisation des acides aminés peut être trouvée dans [180]. [196] examine des contenus élevés en acides aminés soufrés dans certaines partie du génome et lien éventuel avec une pression sélective locale pour se préserver d'actions destructrices de gaz ou radicaux. Transférer l'information disponible sur une protéine au(x) gène(s) dont elle est issue ne pose pas trop de problème et est justifié biologiquement.

Nous regarderons dans la section suivante des données portées par la séquence mais pour lesquelles la connaissance d'autres organismes vivants apporte toute sa valeur. On met à profit le maximum de connaissances disponibles et utiles.

2.1.2 Information sur la séquence..en s'aidant d'autres génomes

Les **profils phylogéniques** ([179]) font référence à l'étude comparative de plusieurs génomes selon leur évolution relative. Parmi les organismes actuellement connus, on trouve environ 70% de gènes qui ne sont pas spécifiques à l'organisme mais peuvent être retrouvés (ou un gène qui lui est fortement similaire [2], voir A.2).

Plus précisément, pour un gène i, il s'agit d'un vecteur binaire a_i dont la dimension est le nombre de génomes considérés. Pour le génome j, on a :

[1] Ils peuvent être hydrophiles (acide aspartique, lysine, sérine) ou hydrophobes (isoleucine, leucine, valine), leur point isoélectrique, leur taille, le fait d'être porteur d'un groupe aromatique ou aliphatique, leur charge,...

[2] On parle de similarité plutôt que d'homologie qui fait plutôt référence au concept d'évolution

$$a_{i,j} = \begin{cases} 1 & \text{si le gène "i" est présent dans le génome "j",} \\ 0 & \text{sinon.} \end{cases}$$

On pourrait aussi imaginer remplacer ces valeurs 0 ou 1 par des valeurs reflétant une probabilité que le gène "i" soit bien présent dans le génome "j" ou plus généralement un score de similarité.

Pour un ensemble de gènes et de génomes, on rassemble ces valeurs dans une matrice :

$$\begin{array}{c} \\ \text{gene i} \end{array} \begin{array}{c} \text{genome j} \\ \rightarrow \end{array} \begin{pmatrix} & \vdots & \\ \cdots & a_{i,j} & \cdots \\ & \vdots & \end{pmatrix}$$

Une ligne contient le profil d'un gène tandis qu'une colonne liste l'ensemble des gènes d'un génome (et ceux qui en sont absents). Ces données prennent donc la forme classique des données analysées par les outils statistiques sous la forme d'une matrice avec les individus en lignes (pour nous) et les conditions en colonnes.

Soulignons qu'alors, la quantité sans cesse croissante de génomes séquencés et d'informations disponibles à leur sujet devraient impliquer une précision accrue de la méthode. Il faudra toutefois prendre garde à d'éventuels biais de la représentativité des différents organismes. En effet, tout l'éventail du monde vivant est loin de nous être révélé. Ainsi on pourrait croire que certains profils sont peu fréquents parce que nous n'observons pas suffisamment certaines branches de l'évolution. A l'inverse, on pourrait conclure hâtivement à l'abondance de certains profils dont l'importance pourrait n'être que toute relative : soit ils sont peu caractéristiques parce que présents dans la plupart des génomes (voir plus bas quant à la richesse d'une telle information), soit ils sont présents dans un ensemble de génomes qui a divergé récemment mais qui se trouvent sur-représentés. En fait cela fixera l'étendue de la validité des conclusions tirées de l'étude de telles données.

Un profil phylogénétique peut être construit à partir des COG (voir l'annexe A.2). Cela n'est pas la seule méthode mais elle est bien adaptée : des gènes orthologues qui voient leur association confirmée par association géographique auront leurs séquences plus semblables que s'ils ne sont pas voisins ([60]). Aussi la construction ne dépendra pas d'un seuil de similarité.

L'hypothèse sous-jacente est que deux génomes sont d'autant plus similaires du point de vue de leur profil qu'ils ont divergé récemment : on ne considère pas l'évolution convergente (qui fait qu'oiseaux et chauve-souris ont des ailes). Des arbres évolutifs peuvent être construits à l'aide d'une telle similarité pourvu que

l'ensemble des gènes sur lequel se base la comparaison soit pertinent. La présence d'un gène ou d'une classe de gènes spécifiques reflète la présence d'un processus cellulaire commun. On peut aussi imaginer identifier le génotype conditionnant un certain phénotype en différenciant son contenu par rapport à un autre génotype «proche». [169] prédit des unités de transcriptions (ou opérons) à l'aide de telles données.

L'esprit de notre travail sera davantage de regarder la distribution des lignes. Nous nous rapporcherons alors du travail de [179] qui postule que des protéines d'une même voie métabolique ou d'un même complexe structural vont évoluer de façon corrélée : soit co-conservées soit co-éliminées. Cette co-occurrence devrait permettre des prédictions fonctionnelles. Des statistiques peuvent être construites pour donner des niveaux de confiance sur le caractère informatif de la co-présence/absence de deux gènes dans l'ensemble des génomes.

Remarquons que la présence d'un groupe de gènes dans un nombre important (par rapport à une distribution aléatoire binomiale) ou trop faible parmi les génomes considérés est que celui-ci a été probablement mal choisi. En tout cas cela apporte une information pauvre. Des corrections peuvent aussi être nécessaires pour équilibrer les distances phylogénétiques des organismes. En effet, les génomes actuellement connus sont dans des familles particulières et certainement pas pris de façon suffisamment équilibrée parmi l'éventail de la vie. On cherchera donc à corriger ce biais.

L'intersection de tous les génomes devrait laisser un petit nombre de gènes nécessaire à la vie de la grande majorité des organismes ([37]).

Pour finir la présentation de telles données, il est intéressant de faire le lien avec le début la partie 2.2. En effet en s'intéressant aux colonnes de la matrice des profils, on pourra dégager des groupes de gènes qui sont certainement fonctionnellement liés parce que solidaires quant à leur présence dans un jeu de génomes significatifs. Nous avons aussi vu que la proximité chromosomique jouait un rôle (*cf.* le lien avec les opérons analysés dans [169]) ; combiner ces deux types de données a donc certainement un sens. Mais laissons pour le moment en suspens cette question de l'interaction entre les gènes pour nous intéresser à un autre type de données.

Nous avons mentionné dans l'introduction que nous mettrions à contribution les nouveaux types de données dites «omiques». Pour simplifier, on les verra comme des mesures instantanées de la quantité d'un certain type d'entité biologiques dans la cellule : ARN messagers pour le transcriptome, protéines pour le protéome, metabolites pour le métabolome,... Tandis que le génome d'un organisme peut être considéré comme figé à notre échelle, les données «omiques» évoluent rapidement d'une cellule à l'autre ou même au sein d'une même cellule avec le temps. Ceci par une interaction incessante entre la cellule, son patrimoine génétique et son environement.

2.1.3 Puces à ADN

Les méthodes de **puces ADN** impliquent une mesure qui n'est pas lue directement sur la séquence. On pourra regarder [66] pour la description du déroulement d'une manipulation concernant la levure et le type d'informations biologiques mises en jeu. L'idée en est simple même si son implémentation pratique, bien qu'aujourd'hui maîtrisée, ne le soit pas tant. Cette technique permet des mesures rapides et une surveillance à l'échelle du génome. Nous en décrirons les étapes majoritaires pour aider à comprendre la forme des données.

Très rapidement, il s'agit de greffer sur une surface de quelques centimètres carrés des fragments synthétiques d'ADN : les **sondes**. Ils représentent les différents fragments des gènes que l'on souhaite étudier. Leur nombre peut aller jusqu'à l'ordre de la dizaine de milliers. Ils sont espacés de quelques micromètres. Ils ont en général été obtenus par amplification (PCR). Ils sont greffés électrostatiquement sur une lame par un revêtement spécial. Ces ADN devront être dénaturés (*i.e.* les deux brins complémentaires seront séparés). Ainsi l'étape d'hybridation à suivre est rendue possible. Notons à cette étape une première difficulté technique à l'origine de mesures douteuses : la localisation précise des sondes sur une lame de si petite dimension. Il faut pouvoir identifier précisément ces sondes pour dire à quel gène correspond quelle mesure.

Ce dispositif est ensuite mis au contact des acides nucléiques à analyser : ARNm ou ADNc (ADN **cibles** obtenus par rétro-transcription) préalablement couplés avec un marqueur radioactif ou fluorescent (Cy3 vert ou Cy5 rouge) différent par conditions à tester. Ceux-ci ont été produits par un dispositif expérimental propre au phénomène dont on veut étudier l'impact sur l'expression des gènes. La mise en relation sondes-acides nucléiques produits par expérimentation conduit à l'hybridation quasi simultanée et la formation de couples que l'on localise et dont on quantifie l'intensité de la liaison par la lecture du signal : radioactif ou fluorescent. On pourra regarder la figure 2.2 pour une illustration simple de l'étape d'hybridation. Il y a concurrence entre les cibles provenant de l'échantillon à d'intérêt et celles de l'échantillon de contrôle. Ces deux échantillons sont mélangés en quantités égales. Les abondances relatives des ARNm pourront être détectées grâce à leur marqueur respectif. Par exemple dans le cas des marqueurs Cy3 vert et Cy5 rouge, on obervera une couleur verte, rouge (évidemment) ou jaune (si les ARNm des deux échantillons sont hybridés dans des quantités proches sur la lame).

Les choix de support, types de sondes conditionnent grandement la précision (sensibitié, spécificité) des résultats ; l'équilibre doit être trouvé entre budget disponible et qualité des données résultantes. Il existe de nombreux autres détails techniques qui limitent la qualité des échantillons ainsi mesurés.

Un laser excite la lame et les marqueurs fluorescents (les seuls que nous avons rencontrés dans nos expériences) en chacun des points où ont été déposées des sondes émettent un signal lumineux. Un dispositif optique permet de capter

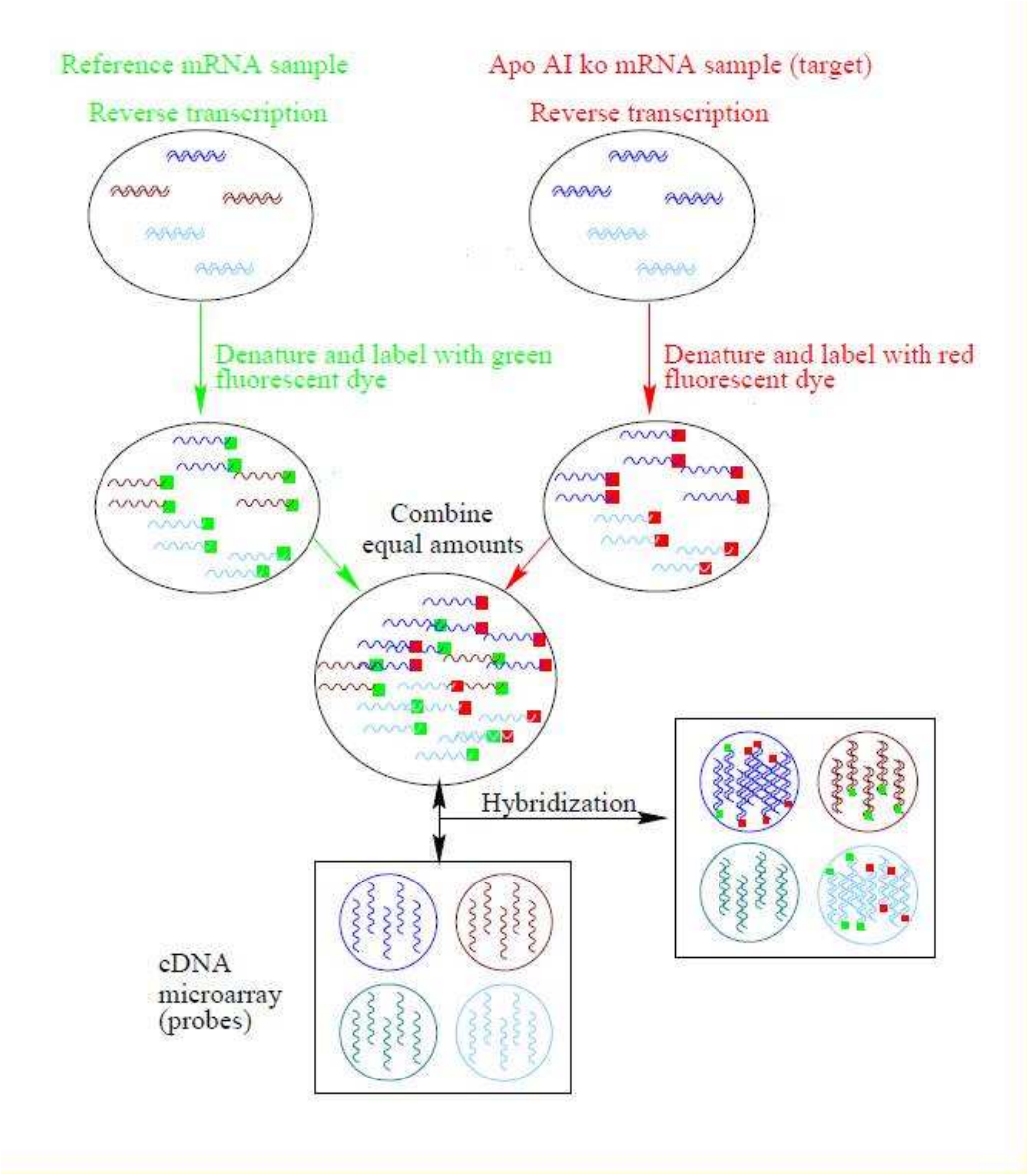


Fig. 2.2: Hybridation des deux cibles fluorescentes rouge (Cy5) et verte (Cy3) sur la sonde.

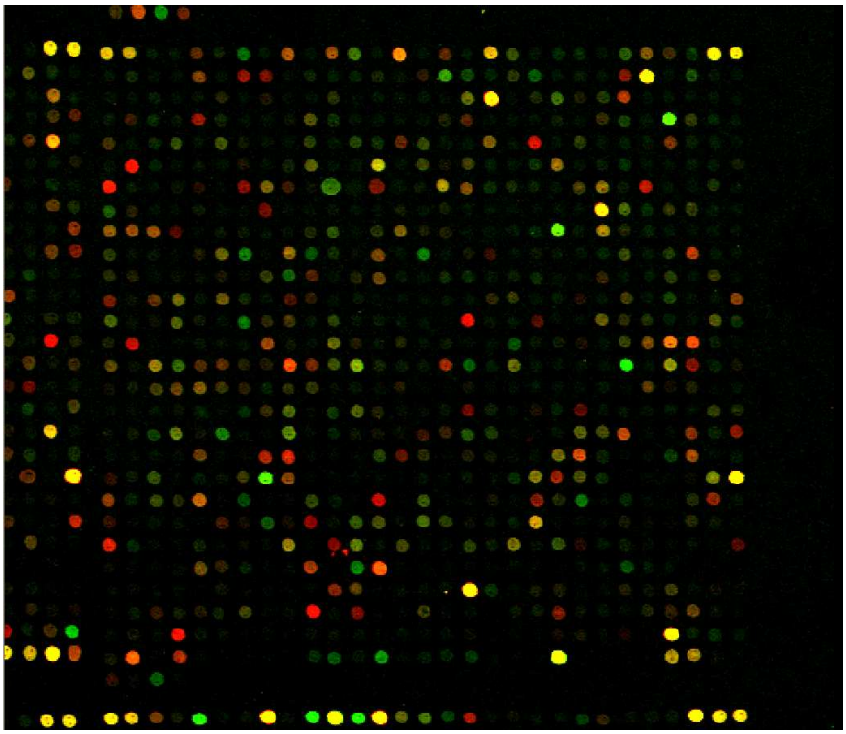


Fig. 2.3: Exemple d'image de puces ADN. Les colorations rouges et vertes intenses correspondent à des gènes fortement exprimés dans une des deux conditions respectivement. La coloration jaune correspond aux gènes exprimés dans les deux conditions (superposition de rouge et de vert).

l'image. Une nouvelle difficulté technique est la segmentation de l'image ainsi obtenue : il faut déterminer les pixels qui correspondent à un signal et ceux qui sont de l'arrière-plan (*background* c'est à dire du bruit autour du signal). En chacun des points où ont été déposées les sondes, on observe une coloration rouge, verte ou jaune plus ou moins intense. Extraire l'intensité de ces images n'est pas non plus une opération évidente. Par exemple, les deux marqueurs ne donneront pas les mêmes intensités brutes pour des mêmes quantités d'ARNm. Nous ne pouvons pas entrer dans les détails ici même si de tels facteurs peuvent jouer un rôle important quant aux conclusions basées sur les données. Par exemple la **normalisation** intra- ou inter-plaques, la qualité des images, des spots sont des questions statistiques légitimes. De bonnes références pour le lecteur désireux de se documenter plus avant dans ces domaines ne manquent pas ([156], [160],...). Mais aussi les sites de certains laboratoires qui détaillent très bien ce processus).

À cette étape on dispose donc de données brutes d'intensité de fluorescence pour un gène pour les marqueurs Cy3 et Cy5 ainsi que pour le bruit de fond. Ceci pour chacune des conditions expérimentales (par exemple les instants d'une série temporelle).

On notera immédiatement la difficulté à capter une information aussi riche que celle qui contrôle la machinerie de production des protéines à partir de la lecture de brins d'ADN : on veut accéder à des informations spatiales (voire à quatre dimensions si on prend en compte le réglage temporel de production des protéines) à partir d'informations linéaires.

Mais ceci est très certainement possible puisque l'organisme, sauf dysfonctionnement, sait se construire *via* les ARN messagers ; la connaissance du niveau d'expression des gènes est bien une donnée primoridale pour connaître leur rôle dans la cellule : dynamique cellulaire ou réponse à un *stimuli* externe (substance toxique, infection virale, appauvrissement nutritif, *etc.*). Ces derniers orchestrent tout le mécanisme grâce à leurs mouvements, leurs quantités et leur lieu de sécrétion dans la cellule. On accède à ces mesures pour des milliers d'entre eux simultanément. Comme mesure du produit de la transcription, on parle de **transcriptome**. La relation entre les quantités d'ARNm et de protéines n'est pas directe (voir l'annexe A ou [146]). Toutefois l'absence d'un ARNm implique un niveau très faible de la protéine correspondante. Aussi les concentrations nécessaires à l'action d'une protéine dépendent beaucoup de sa nature ou de la longueur de la séquence concernée par exemple.

On se réfèrera utilement à [184] pour une revue qui ne se contente pas de présenter les données, les traitements auxquels elles sont soumises (normalisation, algorithmes de classification,...) mais précise aussi combien le choix d'une méthode d'analyse doit être conditionné par l'information que l'on voudra capter dans les données. En d'autres termes la question biologique à laquelle on voudra répondre.

Le tracé des niveaux en ARNm effectué à partir des mesures dans différents tissus ou sous différentes conditions environnementales pour chaque gène constitue le profil d'expression. Chaque profil est un vecteur dont la dimension dépend de l'expérience d'intérêt. Pour ce qui nous concerne, on s'intéresse aux travaux qui cherchent à distinguer les différents types de profils qui se dégagent des données. On voudra répondre à des questions pertinentes aussi bien biologiquement que statistiquement : quels gènes sont co-régulés dans le phénomène étudié ? Positivement, négativement ? Quels gènes ont les mêmes comportements ? Des comportements opposés ? Ne donne pas de réponse particulière ? Quels gènes sont sur-sous exprimés ? Quelle confiance dans les réponses aux questions posées ?

Une normalisation voire une transformation préalable des données est nécessaire ; à ce sujet on pourra lire [182, 185, 254]. En effet, la nature biologique de l'expérience et le processus de mesure (préparation des échantillons ou variations de l'intensité du balayage optique d'une micropuce à l'autre par exemple) sont des effets systématiques qu'on voudra filtrer pour n'observer que les variations d'expression dues au phénomène étudié. [156] précise que les mesures d'expression obtenues par puces ADN nécessitent la réplication des données et l'utilisation de modèles adéquat pour pallier leur variabilité importante. Mais de telles méthodes

sont encore très peu explorées tant leur mise en œuvre est difficile et les jeux de données peu répandus ([256]). Les méthodes d'analyse ont une manche d'avance.

En résumé, les données d'expression se représentent souvent sous la forme d'une matrice $N \times D$. N est la taille des données (le nombre de gènes). D représente leur dimension c'est à dire le nombre de mesures : nombre de tissus, de mesures temporelles dans un processus cellulaire ou environnemental par exemple. Une entrée correspond au rapport des intensités rouge sur verte corrigées par rapport au bruit. Nous considérerons que les biais spatiaux, vis-à-vis de l'intensité, dus au type de lame utilisé, à la manipulation,...sont corrigés. On fait donc l'hypothèse (un peu idéale) que les valeurs reflètent les variations biologiques entre les échantillons. De nombreux travaux traitent de ces méthodes de normalisation indispensables sans qu'aucun consensus ne se dégage ([157, 160, 182, 185, 215]).

Nous nous intéresserons à la distribution des profils lignes plutôt qu'à celle des profils colonnes. Les profils colonnes servent à étudier les conditions expérimentales : sain ou malade (cancéreux) pour un tissu par exemple tandis que les profils lignes donnent l'évolution du niveau des ARNm transcrits d'un gène selon les conditions expérimentales. C'est dans cette direction que nous chercherons une structure aux données. Tandis que la classification définit des groupes sur des critères statistiques, les biologistes préfèrent une interprétation en terme de fonction biologique ([157]). Nous voulons réconcilier ces deux aspirations entre algorithme de classification et interprétabilité du modèle vis-à-vis des mécanismes réels de production des données. Notons ici simplement que l'étude des matrices d'expression selon leur colonne mène à des problèmes statistiques intéressants. En effet, la plupart du temps, $N >> D$. Ainsi, on veut faire des statistiques avec des données dont la dimension est bien plus grande que leur taille. Cela pose des problèmes d'estimation des paramètres par exemple. C'est tout un champ d'étude actif que nous n'abordons pas dans ce mémoire.

Nous voulons aller au-delà de simples mesures classiques comme la corrélation entre les deux profils d'expression. Ou même d'un éventail de mesures comparables (distance euclidienne, information mutuelle,...) proposées par certains auteurs ([102]). Sans réelles explications, certaines similarités semblent conduire à de meilleurs résultats : par exemple la distance euclidienne, pourtant sensible aux données extrêmes semble très bien convenir aux rapports logarithmiques des données d'expression pour être utilisée dans des procédures de classification à base de distance ([68, 92]).

En fait on ne veut pas transformer la donnée individuelle en une donnée de paire comme on vient de le présenter. La quantité d'ARNm donnera directement une information subrogée à la quantité de protéines ou au moins des variations dans leur production. Ceci grâce à la différence de fluorescence entre une condition de référence et une condition expérimentale à l'étude.

Signalons que nous aurons à simuler des données d'expression. Un profil (ou un mélange de profils-types) devra être choisi pour chaque gène. De telles don-

nées ne prétendront pas bien sûr refléter la complexité de données biologiques réelles. Mais cela nous sera grandement utile pour comparer les mérites relatifs de différentes méthodes d'analyse de données dans un cadre supervisé (comme l'étude de [103]). Nous simulerons des données semblables à celles de [254]. On aura des comportements périodiques et linéaires des profils d'expression.

Plus simplement, on pourrait aussi simuler des données selon un modèle probabiliste. Mais cette solution nous semble trop favorable envers les modèles que nous testons sans que cela puisse refléter leur bon comportement en ce qui concerne de vrais problème biologiques. Nous avons aussi essayé d'utiliser des données «plus réalistes» comme celles utilisées dans [214]. Elles ont l'agréable caractéristique de prendre en compte un réseau interprétable biologiquement. Or notre travail s'appuie sur un hypothèse selon laquelle la structure des données est fortement influencée par un réseau d'interactions biologiques. Mais elles ne permettent pas de produire des données d'expression pour un nombre suffisant de gènes compte tenu des applications que nous prévoyons pour l'analyse exploratoire de données sur un organisme entier et éventuellement en grande dimension. La partie 5.1 abordera toutes les expériences sur données simulées qui ont été réalisées dans le cadre de ce travail de thèse.

Mentionnons deux autres applications classiques des puces ADN. Tout d'abord la découverte de gènes différentiellement exprimés vise à connaître l'impact de traitements sur l'expression des gènes (conditions d'appauvrissement en nutriments, dosages d'une certaine substance,...). Cela fait appel à des outils statistiques appelés tests d'hypothèses. Des choix de modélisation spécifiques à l'analyse de données d'expression doivent être faits, notamment quant à la variabilité de l'expression des gènes. Une compréhension des contrôles d'erreurs inhérentes à de tels tests est aussi nécessaire. La deuxième application concerne l'aide au diagnostic médical. On veut apprendre des règles de décision dans une phase d'apprentissage. Elles permettront de distinguer des profils d'expression d'individus malades et d'individus sains par exemple. Alors on saura prédire le caractère malade ou sain d'un nouvel individu non étiqueté. La classification supervisée permet de répondre à de telles situations. Là encore une estimation la plus précise du taux d'erreur et une interprétabilité de la règle de classement sont souhaitables. Nous n'aborderons pas ces problématiques mais nous renvoyons le lecteur désireux aux chapitres traitant ces sujets dans des travaux tels que [156,157,160,182]. Les fondements des traitements statistiques y sont très bien détaillés.

Comprendre et rassembler des données d'expression cohérentes s'est révélé un travail important à mener en priorité pour établir notre proposition. Nous nous sommes également intéressés à d'autres types de données individuelles qui pourraient être incorporées au modèle probabiliste que nous proposons. Nous ne les avons pas effectivement utilisées (faute d'expertise biologique pour leur emploi simultané pertinent) mais nous les présentons en annexe A.3.

2.2 Données d'interaction

La similarité de séquence (débattue dans l'annexe A.2) est une méthode qui a été et est toujours très utilisée pour avoir une idée de l'ensemble des gènes et des fonctions concernées dans un nouveau génome. L'information est inférée à partir de celle connue sur un autre système par une proximité des gènes du point de vue de leur séquence (ou de celle des protéines). Cette méthode naturellement introduite à la suite de la détection de gènes séduit par la simplicité de son concept. Mais ses hypothèses fondamentales sont aujourd'hui largement remises en question. La découverte d'un gène dont la séquence ne ressemble à celle d'aucun autre déjà étudié est par exemple exclue. En outre, il paraît alors difficile d'éviter la propagation d'éventuelles erreurs d'annotation. Enfin nous soulignons que si l'idée que deux séquences similaires contiennent une information semblable, la façon de quantifier cette propriété est loin d'être triviale. Elle pose par exemple le problème de l'alignement de deux séquences nucléotidiques ou plus, de la fonction score qui donne le niveau de similarité et du seuil pour considérer un ensemble de séquences semblables. Néanmoins il est souvent difficile de se passer du concept de similarité de séquence. De nombreux travaux se sont intéressés à ce sujet toujours d'actualité (*e.g.* [166]) et l'utilise ([76,115,153,176,179,199],...). Mettons en avant l'ouvrage [73] qui fait le point sur de nombreuses approches envisagées et fait le bilan des algorithmes fondamentaux.

L'idée va alors aussi être d'utiliser le **contexte** d'un gène. En effet, ce dernier n'est pas isolé mais acteur au sein d'un important réseau fonctionnel. Il n'y a pas de correspondance univoque entre un gène et une fonction au sein de l'organisme mais plutôt entre un groupe de gènes et un type d'action(s) accomplissant une certaine tâche. D'où l'idée de rassembler des gènes sur des critères les rapprochant aussi bien dans leurs propriétés que du point de vue de leurs interacteurs dans un ou plusieurs réseaux pertinents.

2.2.1 Interactions révélées par les génomes

Ce qui mit la puce à l'oreille

L'idée de base qui a amené à regarder des données d'interactions est que l'agencement spatial des gènes le long d'un chromosome n'est pas du tout aléatoire ([195]). Des pressions sélectives fortes semblent en effet conditionner la place des gènes bactériens et leur implication dans l'architecture cellulaire. Donc ils seraient bien reliés par contrainte fonctionnelle. On prendra alors en compte la proximité physique ou **voisinage géographique** (figure 2.4) : sur un même brin (direct ou inverse) en se basant sur la notion d'**opéron** : entité de gènes co-trancrits sous différentes conditions chez les bactéries. On profite d'avoir une information plus riche que la liste des gènes grâce à leurs positions relatives sur le chromosome ([115]). Le concept d'opéron est *a priori* valable uniquement pour les bactéries. Mais il peut être généralisé à des organismes comme la levure ou même l'homme

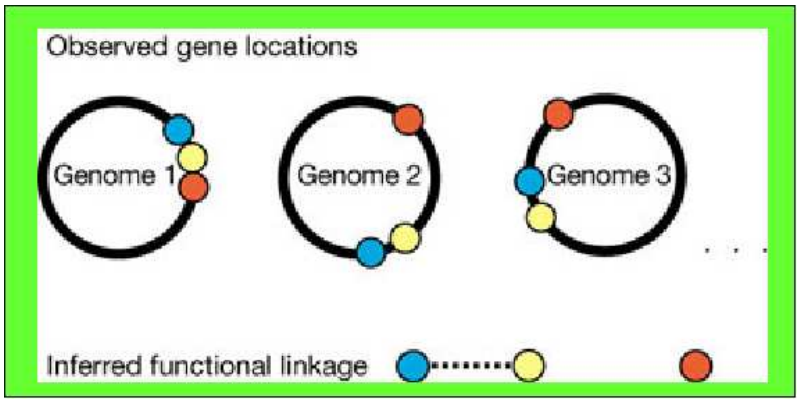


Fig. 2.4: Le voisinage de deux gènes (bleu et jaune sur la figure) dans plusieurs génomes peut indiquer un lien fonctionnel entre les protéines pour lesquels ils codent ([75]). En revanche le lien avec le gène symbolisé en rouge dans le premier génome semble fortuit : il n'est pas présent dans deux génomes sur trois. Bien entendu ce type de conclusion est d'autant plus valable qu'elle repose sur un panel de taille importante et suffisamment diversifié.

où des groupes ressemblant à des opérons ont été découverts. De tels groupes peuvent s'expliquer par plusieurs mécanismes (voir page 32).

Des agrégats locaux de gènes co-orientés interviennent dans des rôles fonctionnels concomittants. Une bonne mesure de ces agrégats est la **persistance** : le nombre moyen de gènes co-orientés qui se succèdent le long d'un chromosome. Il est connu que la distance intergénique dans un opéron est notablement plus faible qu'en dehors ([199]). Cela favorise une conservation plus aisée des voisinages. Un biais sélectif s'opère pour favoriser la conservation locale de certains groupes. Cela permet une méthode de détection en calculant la vraisemblance qu'un ensemble de gènes forme un opéron ([77]). On peut aussi trouver des raisons chimiques à de tels regroupements comme les îlots soufrés qui protègent le génome de certaines actions destructives ([196]). Inversement, les travaux de Tamames ([220]) sur *H. influenzæ* tendent à indiquer que le voisinage implique la liaison fonctionnelle. Un filtrage est nécessaire : inférer une liaison fonctionnelle sur le seul argument d'une co-localisation conduit à trop de fausses détections. Une modèle attribue un score pour des groupes «les plus semblables» entre plusieurs génomes ([176]). Cette donnée de distance fait bien sûr référence à au moins deux gènes et n'est pas une donnée individuelle comme celles de la section 2.1. Nos données de paires sont constituées !

Au delà du voisinage géographique : la fusion

On peut aussi assister à la **fusion** de deux gènes ou davantage. Distincts chez un organisme, ils seront systématiquement associés sur un autre organisme en un

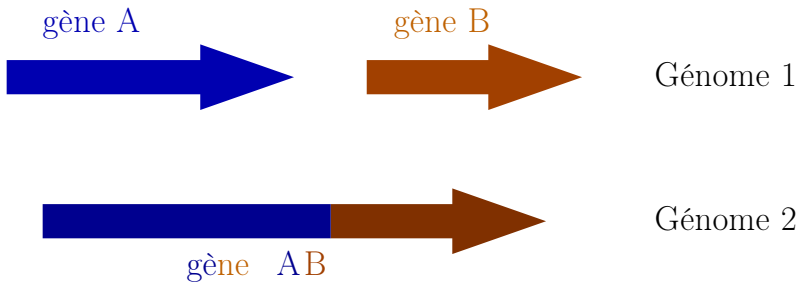


Fig. 2.5: Illustration du concept de fusion

gène composite. Une illustration simple de ces *chimères* du vivant est donnée à la figure 2.2.1.

Ce concept est fort bien décrit dans [153], [154] et [252]. De manière simplifiée, deux gènes fusionnent par réarrangement génétique aléatoire. Les domaines du complexe ainsi formé sont plus concentrés puisque leur enveloppe sur le chromosome est réduite. L'énergie libre du système diminue et favorise sa sélection thermodynamiquement. Un gène composite "AB" du génome 2 est détecté comme étant le plus semblable aux gènes "A" et "B" du génome 1. La comparaison de séquence inverse en ouvrant le gène composite est ensuite effectuée. Il faut aussi contrôler la ressemblance de "A" et "B" pour éviter une fausse prédiction ; ils ne doivent pas être paralogues.

Plutôt que de compliquer le paysage des protéines en allongeant leur liste, les évènements de fusion sont très informatifs. Sur un principe similaire à celui qui servit aux égyptologues pour déchiffrer les hiéroglyphes à l'aide de la pierre de Rosette, des gènes indépendants dans un génome peuvent fusionner dans un autre et sont souvent reliés fonctionnellement. Quand un des composants est à fonction inconnue, cela permet une inférence de fonction putative. La figure 2.6 donne un exemple d'un tel cas de figure.

Il est à remarquer que les composants peuvent être très éloignés dans un organisme où ils ne sont pas fusionnés ([76]). Cette dernière référence montre qu'alors, la fusion sur un autre génome peut servir à détecter des interactions fonctionnelles des protéines concernées avec seulement trois génomes bactériens. La fusion présente donc un apport indéniable par rapport au simple voisinage. En outre elle ne nécessite pas l'observation dans de nombreux génomes. Il suffit d'avoir un génome où le phénomène est identifié.

Les limitations majeures de cette méthodes sont (i) la trop faible occurrence de tels évènements, (ii) la difficulté de détecter de façon suffisamment fiable les fusions prédites à tort (tout repose sur de l'homologie de séquence...) et (iii) la capacité de certains gènes qualifiés de *hub* (plaque tournante) à tisser de nombreux liens avec des interacteurs agissant dans des réactions très variées. Ces derniers sont donc non spécifiques et ne permettent pas d'obtenir d'information sur le rôle

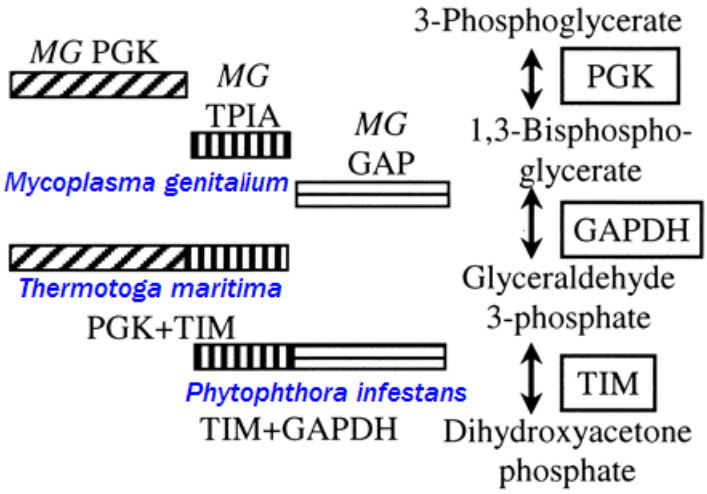


Fig. 2.6: Trois gènes de *M. genitalium* : PGK, $TPIA$ et GAP sont élucidés comme agents séquentiels au sein de la glycolyse grâce à des évènements de fusion chez d'autres organismes.

du gène associé.

Un ordre au delà de l'opéron

On a vu les concepts d'opéron ou de régulon ci-dessus et leur importance dans la prédiction des interactions entre protéines concernées. Si on regarde plus généralement des **segments conservés** sans se soucier de l'ordre dans le groupe conservé, on trouve une bien plus grande diversité d'interaction pertinente. L'ordre des gènes est quant à lui informatif en ce qui concerne les relations de proximité ou éloignement phylogénétique des espèces concernées. Il est aussi d'un bon apport quant à la prédiction de fonction quand l'ordre dans certains segments est conservé ([60]). On parlera d'*über operon* dans le cas où le biais sélectif s'opère au-delà de l'opéron. Des réarrangements génomiques sont en effet envisageables à proximité d'un gène spécifique. Mais on observe quelques groupes de gènes très bien conservés à l'échelle des temps de l'évolution ([37, 220]). La régulation ne suffit en effet pas à expliquer ce biais. Plusieurs modèles évolutifs (voir [9]) ont été proposés pour expliquer la tendance d'observer des ensembles de gènes qui restent regroupés d'un organisme à l'autre :

- le modèle natal : il y a eu un phénomène de duplication ancestrale et puis on a assisté à une divergence graduelle ; elle est peu vraisemblable parce qu'elle ne procure aucun bénéfice particulier pour l'organisme,
- le modèle de Fisher : postule qu'on évite de rompre des complexes de gènes co-adaptés,
- la co-régulation : suppose que l'expression coordonnée et la régulation des

gènes sont alors facilitées ; la proximité simplifie la tâche de l'organisme ; l'ARNm a moins de chance de se dégrader [169] et

- l'opéron égoïste : modélise la diffusion simultanée (par transfert horizontal) de gènes reliés par un même processus. Le gène «protège» la survie de sa fonction en restant groupé avec son contexte.

L'intégrité du regroupement est tout bonnement importante pour le bien-être de la cellule : interaction protéique, co-localisation d'ARNm,... Par exemple, deux opérons $\{A, B\}$ et $\{C, D\}$ d'un organisme pourront être regroupés en un *űber operon* si on trouve deux opérons $\{B, C\}$ et $\{A, D\}$ dans un autre génome. Cela sert à indiquer la proximité des gènes A, B, C et D au sein d'un processus cellulaire ou d'un lien entre les processus dans lesquels ils interviennent.

On veut généraliser ces concepts purement physiques et étendre les interactions à des voisinages plus généraux entre les acteurs du monde vivant ([59]). Comme nous l'avons fait pour les données individuelles, examinons des interactions qui ne sont pas accessibles directement sur la séquence des chromosomes.

2.2.2 Biologie des systèmes

Définir le nouveau champ de biologie des systèmes (*System Biology*) n'est pas évident. Les nouvelles disciplines scientifiques, de même que les espèces, voient le jour grâce à l'évolution de plus anciennes ([132]).

La biologie moléculaire, son point de départ peut maintenant être analysée avec du recul. Elle apporte un renouveau dans son paradigme et son cadre de pensée par rapport à la biochimie par exemple : des problèmes nouveaux sont soulevés ainsi que ses approches technologiques pour y répondre. La situation a grandement évolué depuis l'hypothèse «un gène, une enzyme» de Beadle et Tatum qui mit en évidence le lien entre génotype et phénotype au niveau moléculaire. Aujourd'hui bien que nous puissions encore associer une structure à un gène, on ne peut pas lui assigner de finalité dans la cellule ou dans l'organisme. Il y en a trop. L'action d'un protéine est définie par son contexte. Ce contexte inclut l'historique (dévelopemental ou physiologique). L'exemple du système immunitaire à ce propos est frappant. Les nouveaux projets génomes qui arrivent à terme alimentent le *listing* des gènes que nous connaissons. Mais cette liste est finie et sa taille est même dérangeante : la diversité des gènes ne suffit pas à expliquer la richesse des fonctions d'un organisme. Il faut donc autoriser une utilisation combinatoire de ces composants pour générer la diversité observée. Dès lors on suppose que les produits de gènes sont reliés. On peut ainsi expliquer leur co-régulation, leur localisation respective, leur association, leur modifications post traduction, leur dégradation,...

Finalement la grande question pour comprendre la biologie en ces termes n'est pas le lien régulatoire mais la nature des systèmes biologiques qui autorisent les interactions, pourvu que celles-ci mènent à une combinaison vivable voire

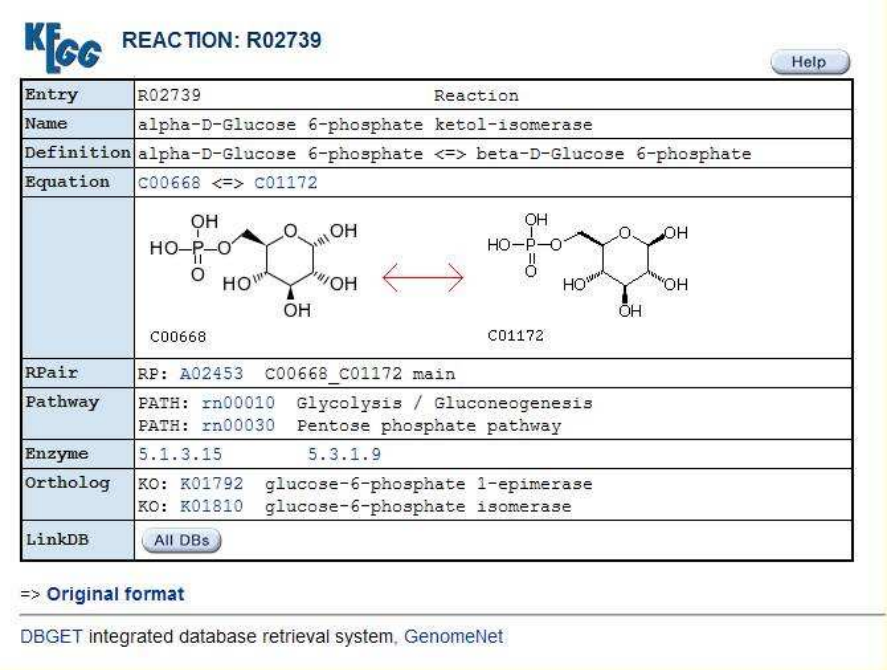
utile à l'organisme. La disponibilité des données «omiques» offre l'opportunité de déchiffrer comment le génotype peut générer le phénotype. Le fait que ces données à haut débit soient bruitées et aujourd'hui encore partiellement incomplètes (donc n'ont pas toute leur cohérence) a provoqué un changement radical dans la façon dont on aborde le problème. On ne sépare plus les effets des acteurs du réseau. Par exemple comprendre les bases génétiques d'une maladie ne peut se limiter à donner la listes des gènes concernés. Il faudra aussi comprendre comment le phénotype est créé et comment les interactions entre variations génétiques et environnement de l'organisme (du patient dans la cas humain) l'influencent.

Les buts poursuivis par la biologie des systèmes sont multiples : reconstruction de processus complexes à partir de l'évolution de ses composants et à un niveau supérieur, le développement et l'étude de modèle qui reproduisent le plus fidèlement possible la dynamique cellulaire. De plus en plus d'outils sont à notre disposition. La biologie des systèmes peut mener sa tâche à bien : étudier le comportement de l'organisation biologique complexe et de ses constituants moléculaires. Ses atouts sont les mesures quantitatives, la modélisation, les outils de reconstruction et la biologie théorique. Nous allons présenter ici les mécanismes pour rendre compte des interactions dont il est question et la façon dont nous allons les intégrer à la description du système auquel nous nous intéressons.

Les voies métaboliques

Les **voies métaboliques** sont l'ensemble des réactions qui prennent place dans un organisme. Le métabolisme est le flot de molécules et d'énergie régi par les voies métaboliques. Beaucoup de réactions métaboliques impliquent protéines, acides nucléiques, acides aminés et sucres. Des protéines appelées **enzymes** (protéine ou plus rarement un fragment d'ARN non transformé par le bilan de la réaction ; sa synthèse se fait donc à partir de l'inormation génétique) **catalysent** (permettent une réaction et en contrôlent la vitesse) des réactions chimiques qui transforment des petites molécules présentes dans la cellule (rappel : les métabolites) en d'autres. Les réactifs premiers des voies métaboliques sont ceux fournis par l'environnement et ceux en sortie sont ceux utilisables par l'organisme. C'est ainsi que sont produits les nucléotides pour l'ADN et l'ARN, les acides aminés pour les protéines, les lipides pour les membranes ou beaucoup d'autres composants essentiels : les cellules sont de vraies petites usines chimiques complexes et très performantes.

L'ensemble des voies métaboliques forment un réseau complexe. Certaines parties sont plutôt linéaires comme la synthèse du tryptophane. D'autres forment des boucles de rétroaction comme le cycle de Krebs (ou acide tricarboxyl). En outre tous ces chemins s'entrecroisent dans un enchevêtrement dense. La structure de l'ensemble des réseaux métaboliques, c'est à dire ses connectivité et topologie, ainsi que ses schémas de fonctionnement peuvent être analysés en terme mathématiques au moyen des graphes.



KEGG REACTION: R02739

Help

Entry	R02739 Reaction
Name	alpha-D-Glucose 6-phosphate ketol-isomerase
Definition	alpha-D-Glucose 6-phosphate <=> beta-D-Glucose 6-phosphate
Equation	C00668 <=> C01172
	C00668 C01172
RPair	RP: A02453 C00668_C01172 main
Pathway	PATH: rn00010 Glycolysis / Gluconeogenesis PATH: rn00030 Pentose phosphate pathway
Enzyme	5.1.3.15 5.3.1.9
Ortholog	KO: K01792 glucose-6-phosphate 1-epimerase KO: K01810 glucose-6-phosphate isomerase
LinkDB	All DBs

=> Original format

DBGET integrated database retrieval system, GenomeNet

Fig. 2.7: Exemple d'information sur une réaction métabolique de la glycolyse extractible

Nous nous sommes servis du module REACTION de KEGG (`http://www.genome.ad.jp/kegg`, [120]). C'est une partie de la base LIGAND. Elle contient des réactions biochimiques qui sont annotés (voir la figure 2.7). Fin juin 2007, cette base contenant 7205 réactions sur quelques 14838 réactifs répartis dans plus de 57000 voies métaboliques.

Les informations qui nous intéressent sont des descripteurs des éléments mis en jeu : réactions, composés réactifs et produits, enzyme(s), gènes associés à ces enzymes dans l'organisme étudié. Le métabolisme donne un cadre aux échanges de matière et d'énergie. Mais il ne livre pas entièrement l'information sur les contrôles qu'exerce l'organisme sur ces flux. Il faut pour cela faire appel aux voies de régulation qui connectent protéines et métabolites.

Graphe d'interactions métaboliques entre des gènes de la levure

Comment relier les données métaboliques qui rassemblent des informations sur des relations entre réactions successives à des gènes ? Une façon de faire est illustrée à la figure 2.8 avec des composés intervenant dans le cycle du `Citrate`.

À une réaction peut correspondre une ou plusieurs enzyme(s). Soit elles sont toutes nécessaires à son déroulement, soit elles agissent spécifiquement dans un certain milieu ou pour un sens donné d'une réaction réversible. Aucun problème dans le cas de correspondance univoque. Si plusieurs enzymes sont reliées à une même réaction, autant de nœuds dans la figure 2.8 (B) que d'enzymes sont créés

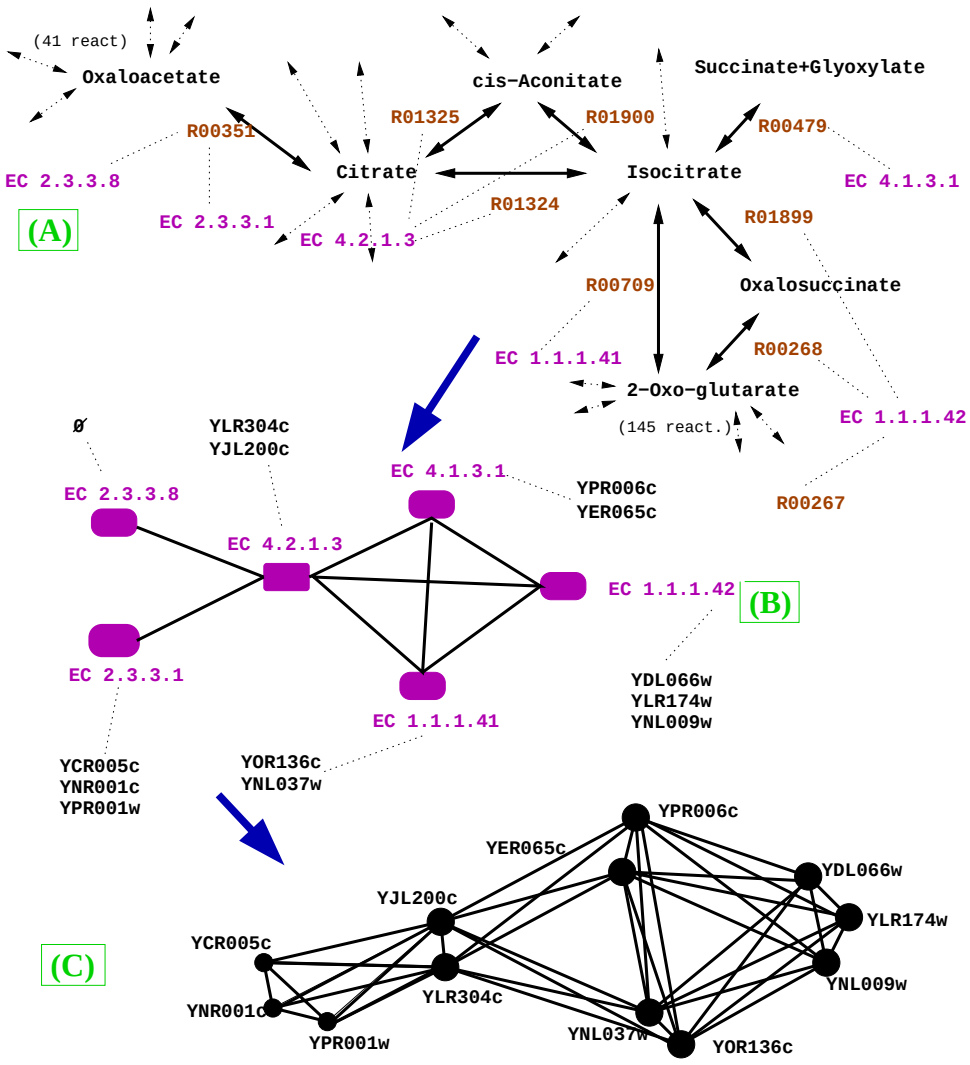


Fig. 2.8: Construction du graphe d'interactions entre gènes (C) à partir du réseau métabolique (A) à partir des relations entre les enzymes (B)

en multipliant les voisinages. Aucune information n'est perdue. En revanche on s'attend à des comportements similaires de ces nœuds : ils auront les mêmes interacteurs et ils interviennent pour une même réaction. Des nuances peuvent cependant apparaître si les enzymes sont spécifiques de certaines conditions environnementales par exemple ; leur fonction peut alors aussi l'être. Dans le cycle du citrate, on aura les enzymes $EC\,2.3.3.1$ et $EC\,2.3.3.8$ pour la même réaction : $R\,00351$. Leur classe est la même mais elles diffèrent par leurs annotations : le rôle pour lequel elles sont repérées n'est pas identique. Une même enzyme peut aussi intervenir dans plusieurs réactions. Par exemple les réactions $R\,02739$ et $R\,03321$ de la glycolyse font toutes les deux intervenir le glucose-6-phosphate isomérase mais sous des formes différentes (α et Δ). Une même enzyme les catalyse : $EC\,5.3.1.9$. La pluri-action de l'enzyme est une information qui ne paraît que dans l'ajout de liaisons pour le même nœud enzyme dans la figure 2.8 (B). Mais on ne sait pas si celles-ci résultent du voisinage d'une seule réaction ou du fait que plusieurs réactions sont catalysées par la même enzyme.

De même, plusieurs gènes peuvent correspondre à une enzyme. Par exemple, cette enzyme peut être l'assemblage de plusieurs protéines. Ou des gènes différents sont sollicités selon certains facteurs. Les protéines $HXK1$ ($YFR053C$) et $HXK2$ ($YGL253W$) de la levure sont toutes deux associées à l'enzyme $EC\,2.7.1.1$. Il est connu que les expressions des gènes correspondant sont fortement anticorrélées pendant le *shift* diauxique ([66]). Cela s'explique par le fait que l'hexokinase (le nom de l'enzyme) participe dans bon nombre de réactions dont $R\,01786$ et $R\,01961$ qui apparaîssent respectivement dans la voie de la glycolyse et dans celle du métabolisme des acides aminés. Notre méthode produit des nœuds différents pour les deux gènes dans le graphe 2.8 (C). On espère donc pouvoir apporter un complément d'informations sur l'intervention de l'une ou de l'autre dans un contexte fonctionnel ou un autre. Un gène peut aussi correspondre à plusieurs enzymes (comme module commun). De même que pour le passage de plusieurs réactions catalysées par une même enzyme, on va perdre une partie des renseignements originaux. Les liens seront conservés sans que l'on puisse dire s'ils sont hérités de voisinage d'enzyme dans 2.8 (B) ou de la correspondance d'un même gène pour plusieurs enzymes. Par exemple, on a le gène $YLR028C$ ($ADE\,16$) qui est associé aux enzymes $EC\,2.1.2.3$ et $EC\,3.5.4.10$ dans le métabolisme de la purine. Ces dernières n'ont pas du tout la même fonction (transfert d'un groupe carboné contre hydrolyse de liaison carbone-azote autre que peptide respectivement) mais interviennent dans le même contexte.

Il faut aussi rappeler que des erreurs d'annotations peuvent bien sûr substister dans les bases et que notre méthode automatique de construction de graphe les prendra en compte.

Nous placerons une arête entre deux enzymes si et seulement si elles catalysent deux réactions partageant un composé en commun. On veut ainsi prendre en compte la possibilité d'enchaînement des deux réactions dans une voie métabolique. On obtient alors le graphe 2.8 (B). Pour ne prendre en compte que des

relations ayant un sens, on pourra se concentrer sur une base regroupant les «composés principaux» ou ne pas prendre en compte les composés intervenant dans de trop nombreuses réactions. La première situation évite de lier des réactions artificiellement parce qu'elles partagent un composé qui n'a rien de spécifique comme l'eau ou le dioxyde de carbone. La seconde donne un critère quantitatif pour une limitation quant aux composés qui ne sont pas suffisamment (par rapport à un seuil à choisir) spécifiques. La majeure différence entre ces deux possibilités est que la seconde ne réclame pas une intervention experte sur toute la base mais uniquement pour le choix du seuil. En revanche elle peut écarter des composés-clef même s'ils sont présents dans de nombreuses réactions/voies métaboliques.

Puis on joindra deux gènes si ils correspondent à deux enzymes voisines dans le graphe précédent ou à une même enzyme. On obtient alors les relations de dépendances entre les gènes 2.8 (C). La contruction d'un tel graphe sera décrite formellement à la partie 5.2, page 122. On donnera les détails techniques alors que nous nous sommes ici intéressés au côté biologique.

Des poids peuvent être assignés aux arêtes. Ils peuvent refléter un score entre les deux nœuds, une facilité thermodynamique d'enchaînement des réactions [11,144]. Leurs coefficients dictant la thermodynamique d'une réaction quantifieront cette aspect. Des travaux récents se penchent sur ce type de problèmes ([11] par exemple). Mais le réseau des réactions bio-chimiques est probablement à l'heure actuelle trop complexe pour des considérations de thermodynamique et cinétique qui permettraient de donner des poids aux arêtes. [63] souligne qu'une des difficultés rencontrées par les méthodes de réseaux de régulation est d'intégrer d'aussi grands réseaux avec leur dynamique. La pondération des arêtes peut aussi refléter une distance qui sépare les objets. [98] exploite plusieurs distances simultanées qu'il combine par exemple.

Une limitation évidente de la construction du graphe ci-dessus est qu'aucune information n'existera pour des gènes pour lesquels il n'existe pas de correspondance enzymatique. Et ils sont nombreux ! Pour la levure par exemple, on arrive à ce jour à construire un réseau comprenant environ 700 gènes et 15 000 arêtes alors que le génome contient un peu plus de 6000 gènes. Beaucoup de gènes (*e.g.* les régulateurs, voir la page 174) jouent un rôle primordial dans les voies métaboliques mais se pose encore la question de la prise en compte de plusieurs types de données.

Et voies de régulation

Structure et dynamique de ces relations de régulations sont différentes des voies métaboliques. Ce qui correspond à la succession des transformations enzymatiques dans le métabolisme est l'assemblage des cascades de signalisation (*signaling cascades*).

La biologie des systèmes décrit les interactions métaboliques et de régulation en terme de **réseau d'interaction**. Ces deux types de connection entre les

acteurs du vivant agissent en parallèle : un réseau physique de complexes protéine-protéine ou protéine-acide nucléique et un réseau logique de cascades de contrôle. Les voies métaboliques participent à ces deux réseaux. Des interactions physiques entre protéines jouent le rôle d'intermédiaire et des interactions logiques régulent de nombreuses voies métaboliques.

On peut donner des exemples de plusieurs sortes. L'assemblage des centres de réactions photosynthétiques sont exclusivement basés sur des interactions physiques tandis que les boucles de rétroactions sont plutôt des interactions logiques. L'augmentation de la quantité de produits de certaines réactions chimiques peut inhiber la catalyse enzymatique d'une réaction antérieure par la production de petites molécules qui délivrent un signal aux autres cellules. Aussi, un facteur de transcription peut très bien ne jamais interagir physiquement avec toutes les ARN messagers dont il contrôle l'expression simplement en se fixant sur un brin d'ADN. Il existe bien entendu des exemples où les fonctionnalités de ces deux réseaux s'entrecroisent pour l'accomplissement d'une tâche telle que les changements alostériques dans l'hémoglobine (réponse à un changement de niveau d'oxygène par un changement de conformation qui altère l'affinité à l'oxygène) ou la transmission d'un signal depuis la surface de la cellule vers l'intérieur en traversant la membrane. Cela peut déclencher des changements importants de l'expression de certains gènes. L'important est de bien garder à l'esprit ces deux types d'actions qui peuvent coexister simultanément dans un organisme, notamment au moment de l'interprétation des phénomènes.

Les bases de notre compréhension de l'organisation et de la régulation de la vie dans une cellule se résume donc par des interactions centrées sur les protéines. Portons notre attention dessus.

Complexes protéiques

Les méthodes à haut débit permettent aujourd'hui de détecter les interactions des protéines avec des acides nuclétiques ou entre elles. L'étendue des structures et fonctions concernées est importante.

Souvent les protéines oligomériques [3] stables contiennent plusieurs copies d'une protéine (plutôt en nombre impair alors) ou en combine différentes. Les oligomères présentent habituellement une symétrie. L'insuline (figure 2.9) est par exemple un hexamère qui présente un axe de symétrie et une invariance par rotation de 120 degrés autour de son centre.

Parfois des composants d'un oligomère sont remplacés dans des espèces différentes (notamment entre procaryotes et eucaryotes) par des sous-unités semblables mais différentes. Ces évènements surviennent par duplication de gène puis divergence. Les complexes de protéines varient grandement selon le nombre (de

[3] Un oligomère est une molécule qui est formée de quelques (peu) petites molécules identiques assemblées. De même un monomère est composé d'une unité et un dimère de deux.

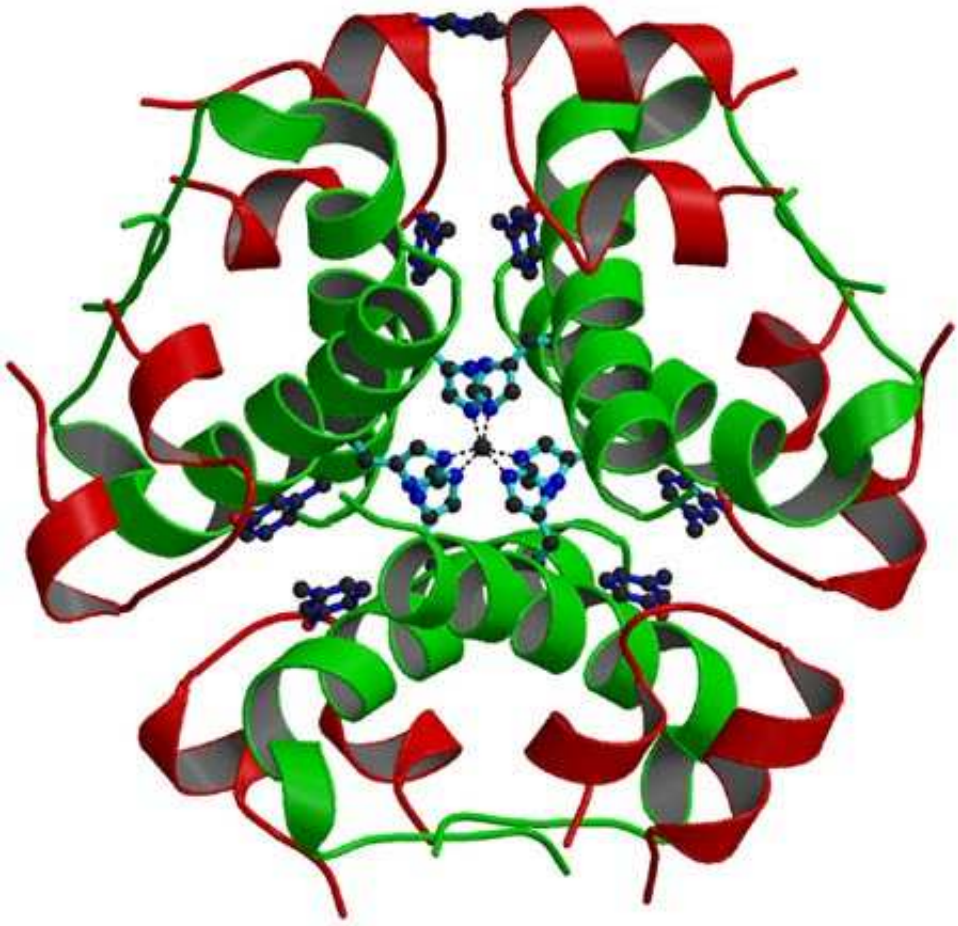


Fig. 2.9: Exemple de structure de l'insuline (R6)

quelques unités à l'ordre du millier pour une capside virale) et la variété des molécules qu'ils contiennent.

2.2.3 Autres données d'interactions

Certains auteurs ([10,218]) se sont intéressés à la **co-occurrence bibliographique**. Des gènes qui sont cités dans un même papier le sont peut-être parce qu'une relation (physique, fonctionnelle,...) existe entre eux. Si le nombre de références reportant cette association grandit, la confiance dans cette prédiction d'interaction fait de même. Cette idée est héritée du TAL. Un biologiste peut commenter un processus sans pour autant publier expressément cette interaction précise parce qu'elle n'est pas centrale dans sa problématique du moment à ses yeux. Nous avons ainsi une manière détournée d'enrichir la connaissance biologique globale. Ces données sont disponibles dans une base comme STRING développée à Heidelberg (`http://string.embl.de/`, [237]). Nous les avons utilisées dans certaines de nos expériences (voir la section 5.3).

Nous avons gardé les données d'interactions les plus connues pour la fin : les interactions protéine-protéine ([236]). Elles correspondent à des interactions entre deux protéines. Elles peuvent être mises à jour par des techniques laborantines longues mais très fiables (*e.g.* co-immunoprécépitation) ou par des techniques à haut débit (par exemple le système double-hybride chez la levure). Ces données pourront être trouvées dans la base spécifique DIP (`http://dip.doe-mbi.ucla.edu/`) ou dans des bases plus générales comme STRING sus-citée. Cette dernière base a d'ailleurs l'avantage de proposer plusieurs types d'interactions (donc plusieurs graphes). Elle propose une méthode propre pour en déduire un réseau composite avec des niveaux de confiance : plus on a d'indices «indépendants» (en fait ils ne le sont pas puisque le fait que les gènes ou protéines interagissent constitue le point commun à dénouer ! Mais les méthodes pour y arriver sont différentes), plus on est confiant dans l'interaction. On pourrait aussi se servir de la méthode des composantes connexes de [39] pour combiner ces graphes. Le gros avantage de cette dernière base de données réside dans la grande quantité d'interactions (vérifiées ou prédites). Notamment cela permet d'avoir une grande partie des gènes qui seront connectés alors qu'avec le graphe métabolique page 36, la partie connectée du graphe est plus limitée.

Nous voulons d'ailleurs terminer ce chapitre en commentant les ressources disponibles pour toutes les données que nous venons de présenter.

2.3 Ressources Internet

Il existe un grand nombre de bases de données d'intérêt biologique. Il est de coutume de distinguer deux types de banques : (i) les bases **généralistes** qui correspondent à une collecte des données la plus exhaustive possible et qui offrent

finalement un ensemble plutôt hétérogène d'informations et (ii) les bases **spécialisées** qui correspondent à des données plus homogènes établies autour d'une thématique et qui offrent une valeur ajoutée à partir d'une technique particulière ou d'un intérêt suscité par un groupe d'individus (*e.g.* bases de motifs ou de facteurs de transcription, de structure, d'expression, de voies métaboliques,). Nous ne nous intéresserons pas à ce raffinement (qui n'est d'ailleurs pas universel). Nous aurons une approche pragmatique pour trouver des données aussi fiables et annotées que possible.

Nous présenterons rapidement un petit historique des bases (ou banques) de données avant d'en sélectionner quelques unes qui nous semblent significatives selon les différents types de données que nous avons introduits dans les sections précédentes.

Historique

C'est au début des années 80 que sont apparues les premières bases de données publiques. Elles étaient alors consacrées aux acides nucléiques. Très rapidement avec les évolutions techniques du séquençage, la collecte et la gestion des données ont nécessité une organisation plus conséquente. Ainsi, plusieurs organismes ont pris en charge la production de telles bases de données.

En Europe, financée par l'EMBO (*European Moleculary Biology Organisation*), une équipe s'est constituée pour développer une banque de séquences nucléiques (*EMBL data library*) et en assurer la diffusion. Cette équipe travaille au sein du laboratoire européen de Biologie Moléculaire qui a longtemps été basé à Heidelberg et qui se trouve actuellement près de Cambridge au sein de l'EBI (*European Bioinformatics Institute*). Du côté américain, soutenue par le NIH (*National Institute of Health*), une banque nucléique (GenBank) a été créée à Los Alamos. Cette base de données était distribuée par la société IntelliGenetics et est diffusée maintenant par le NCBI (*National Center for Biotechnology Information*). La collaboration entre ces deux banques a commencé relativement tôt. Elle s'est étendue en 1987 avec la participation de la DDBJ (*DNA Data Bank*) du Japon pour donner naissance finalement en 1990 à un format unique dans la description des caractéristiques biologiques qui accompagnent les séquences dans les banques de données nucléiques.

Côté protéines, deux banques principales coexistent. La première, sous l'influence de la *National Biomedical Research Foundation* (NBRF) à Washington, produit maintenant une association de données issues du MIPS (*Martinsried Institute for Protein Sequences*), de la base japonnaise JIPID (*Japan International Protein Information Database*) et des données propres de la NBRF. Elle se nomme la *Protein Identification Ressource* (PIR-NBRF). La deuxième, *Swissprot* a été constituée à l'Université de Genève à partir de 1986 et regroupe entre autres des séquences annotées de la PIR-NBRF ainsi que des séquences codantes traduites de l'EMBL.

Pendant longtemps, la principale distribution fut l'envoi postal de bandes magnétiques aux personnes ayant souscrit un abonnement. Progressivement le CD-ROM a remplacé ce support de stockage et a permis une plus grande diffusion des données. Depuis le début des années 90, avec l'installation massive des réseaux informatiques à hauts débits qui permettent d'atteindre une machine située à plusieurs milliers de kilomètres de son terminal, beaucoup de laboratoires rapatrient les bases de données *via* ces réseaux à partir de serveurs publics. Ces réseaux informatiques rapides et les services qui en découlent permettent une large diffusion des bases. Ainsi beaucoup de serveurs mettent gratuitement à disposition de nombreuses bases, dont les grandes banques de séquences généralistes comme l'EMBL avec une mise à jour quotidienne des données, mais également un grand nombre d'autres bases dont la diffusion était auparavant plus restreinte. De ce fait, il résulte une banalisation de l'accès à l'information. Il n'est même plus nécessaire d'avoir localement les bases de données ou de se connecter par des procédures complexes à un centre serveur privilégié pour pouvoir exploiter aisément le contenu de ces bases. En 1997, on estimait le nombre total d'utilisateurs de la banque EMBL à plus de 50 000.

Qualité des données

Il faut avoir conscience que l'information contenue dans ces bases présente un certain nombre de lacunes. Une des plus importantes est le manque de vérifications des données soumises ou saisies surtout pour les plus anciennes. Les auteurs ont parfois du mal à restituer les connaissances qu'ils détiennent à propos de leurs données ou bien n'ont pas fait un certain nombre de vérifications. Dans le cas des séquences par exemple, il arrive que l'on retrouve des segments de vecteurs de clonage ou des incohérences dans les caractéristiques biologiques (parties codantes, définition des espèces ou des mots-clefs,...). Des informations biologiques peuvent aussi être incomplètes, voire erronées.

De ce point de vue l'établissement d'une sorte de thesaurus précis pour les mots-clefs faciliterait la vérification comme cela a été permis avec la définition d'arbres des espèces utilisés par plusieurs banques de données. Les organismes responsables de la maintenance de ces banques ont pris conscience de ces problèmes et maintenant de nombreuses vérifications sont faites systématiquement dès la soumission des données. Ceci n'élimine pas la totalité des imprécisions comme par exemple l'existence de doublons. Il s'agit là de séquences extrêmement similaires qui correspondent à des entrées différentes dans la banque et dont il est souvent difficile de savoir s'il s'agit de polymorphisme, de gènes dupliqués ou tout simplement d'erreurs établies lors de la détermination des séquences. Les retours des utilisateurs peuvent corriger les défauts décelés. Un autre problème important est le retard de l'insertion d'un nouveau résultat dans une banque, lié souvent au volume à traiter qui engendre des priorités ou des choix. Il peut y avoir une dizaine de mois entre une détermination expérimentale et son insertion dans une

banque.

Malgré cela, il faut souligner l'énorme richesse que représentent ces banques de données. La réunion en un seul ensemble de telles informations parfois fort différentes (séquences, annotations, mesures,...) est un élément fondamental pour l'accessibilité de principe aux données. D'autre part, la grande diversité d'organismes qui y est représentée permet d'aborder des analyses de type évolutif. Par exemple, on peut extraire les séquences d'un même gène issu de plusieurs espèces. Puis relier les données relatives à différents organismes grâce à ce rapprochement de séquence. Des outils en ligne pour ce faire sont directement disponibles. Un autre intérêt de ces bases réside dans l'information qui accompagne les séquences (annotations, expertise, bibliographie), même si celles-ci sont souvent de qualité inégale. Ces dernières peuvent parfois constituer les rares annotations disponibles sur certaines séquences. Enfin la présence de références à d'autres bases permet d'avoir un accès à d'autres informations non répertoriées localement. Ainsi on peut connaitre l'entrée dans une base protéique de la protéine qui correspond au gène que l'on a repéré dans une base nucléique. La banque SWISSPROT est particulièrement riche en références croisées avec d'autres banques et en annotations (par exemple, la notion de "prouvé (ou pas) expérimentalement" est introduite dans la table des caractéristiques biologiques). C'est un exemple de la qualité des données que l'on peut retrouver dans les différentes banques de séquences généralistes de ces dernières années.

Quelques exemples choisis

Nous fournissons ici quelques sites qui donnent accès à des bases de données qui nous semblent incontournables en bioinformatique...Ou tout simplement qui ont eu notre préférence sur des critères parfois subjectifs (table 2.1 sur plusieurs pages).

Nom	Type de données	Site	Rapide description
EBI	Base généraliste	`http://www.ebi.ac.uk/Databases/`	Cet institut européen s'est donné pour mission de construire, maintenir et fournir l'accès à plusieurs bases de données. Nous apprécions tout particulièrement : ArrayExpress. Des outils intéressants sont aussi présents (*e.g.* BioConductor).
NCBI	Base généraliste	`http://www.ncbi.nlm.nih.gov/`	Pendant nord-américain de l'EBI comme ressource pour la biologie moléculaire ; il héberge notamment PubMed (orientation biomédicale) et PubChem (orientation biochimique), le moteur de recherche qui permet l'accès à Medline et Entrez qui permet une requête parmi toutes les bases de données. COG/KOG fournit aussi des comparaisons de séquences de protéines parmi de nombreux génomes pour détecter des groupes d'orthologues.
ExPASy	Serveur protéomique suisse	`http://www.expasy.ch/`	Dédié à l'analyse de séquences et structure protéiques ; on retrouve des informations dans la base UniProt de l'EBI.

GO	Ontologie dédiée aux gènes	http://www.geneontology.org/	*Gene Ontology* est la plus grande ontologie (vocabulaire contrôlé et ordonné) libre internet d'une terminologie reliés à des gènes ; les termes sont rangés dans une hiérarchie initialement séparée en trois classes principales : processus biologiques, composants cellulaires et fonction moléculaire (voir aussi la page 129).
MIPS	Centre regroupant des informations sur les protéines	http://mips.gsf.de/	Nous avons surtout utilisé la base CYGD spéciale au projet sur la levure et le catalogue fonctionel proposé.
SGD	Base de données organismes spécifique	http://www.yeastgenome.org/	Information de biologie moléculaire et génétique centrées sur la levure du boulanger *Sacharomyces cerevisiae*. On pourra consulter la base plus généraliste européenne *Ensembl* (http://www.ensembl.org/) qui produit et maintient des annotations sur plusieurs génomes eucaryotes.
PDB	Structure des protéines	http://www.rcsb.org/pdb/	Outils et ressources pour l'étude de structures de molécules biologiques ainsi que les relations qu'elles ont avec la séquence, les fonctions ou des maladies (une version simplifiée pour les non-experts biologistes -nous- est proposée : PDB lite).
RegulonDB	Réseau génique	http://regulondb.ccg.unam.mx/	Modèle d'initiation régulée de la transcription ou du réseau de régulation de la cellule et de l'organisation des gènes en unités de transcription, opérons, régulons.

BOND (ex-BIND grossi)	Réseau d'objets biomoléculaires	http://bond.unleashedinformatics.com/	Cette base de données regroupe séquences et données relatives aux interactions des objets concernés.
KEGG	Réseaux biochimiques	http://www.genome.ad.jp/kegg/	Base de données systémique intégrant blocs d'informations moléculaires (gènes, ligands,...) et informations d'organisation (voies métaboliques ou autres relations entre objets).
DIP	Réseau protéique	http://dip.doe-mbi.ucla.edu/	Catalogue d'interactions entre protéines déterminées expérimentalement : manuellement par expert ou automatiquement par approches assistées par ordinateurs. (voir aussi STRING qui contient ces données http://string.embl.de//).

Tab. 2.1: Quelques exemples de bases de données bioinformatiques

Rappelons que ces informations sont rapidement périssables même si nous avons essayé de fournir des bases qui nous semblent stables (à l'échelle de notre expérience). Le lecteur aura peut-être tout intérêt à faire ses propres recherches à l'aide de mots-clefs ou il pourra consulter les deux volumes annuels de *Nucleic Acids Research* depuis 2004 consacrés aux bases de données, aux serveurs internet et à leur mise à jour. Des bases de données de bases de données sont aussi disponibles : *Sequence Retrieval System* (SRS, `http://srs.ebi.ac.uk/`) ou DBGET (`http://www.genome.jp/dbget/`) par exemple. Enfin, certains chercheurs font un travail considérable en maintenant à jour des listes impressionnantes d'adresses pour toutes ces bases de données. Le lecteur curieux pourra aller voir `http://www.cbs.dtu.dk/biolinks/index.php` (je ne suis pas certain que Jan HANSEN maintienne toujours ce site) [4].

Nous nous plaçons donc dans le contexte de la classification d'objets biologiques, les gènes, pour lesquels nous disposons de données individuelles et de données de paire. Nous voudrons dégager des interactions concertées entre les gènes. Peu d'approches, à notre connaissance proposent la prise en compte simultanée des deux types d'information au sein d'une même procédure (sinon l'article récent [189] qui se sert de la décomposition en valeurs singulières de la matrice d'expression selon des fonctions propres du réseau ; mais la majeure partie de ce travail se place dans le cadre supervisé). Pourtant cette démarche améliore nettement les performances dans l'analyse des phénomènes biologiques à l'aide de plusieurs types de données post-génomiques. [222] identifie ainsi des modules de gènes avec un comportement fortement corrélé parmi les différentes sources de données considérées à l'aide d'un graphe bipartie. [249] présente une méthode pour la découverte de réseaux d'enzymes de la levure en se servant de plusieurs sources de données post-génomiques (expression de gènes, localisation, profils phylogénétiques et compatibilité chimique ce qui constitue l'originalité de l'approche par rapport à des travaux antérieurs : [234, 250]). La différence entre cette approche et la nôtre est qu'elle se situe dans un cadre partiellement supervisée où des connaissances sur certaines parties du réseau est requise pour la construction de leur outil de discrimination basé sur des noyaux.

Le chapitre à venir introduit le formalisme des champs de Markov cachés pour les intégrer naturellement. Les mesures individuelles seront exprimées à travers des distributions de probabilités paramétriques tandis que les données d'interaction seront modélisées à l'aide d'un graphe.

[4] Signalons ici la mort récente d'un portail bioinformatique français qui rendait pourtant de grands services : INFOBIOGEN.

3. MODÈLES DE MARKOV CACHÉS POUR LA CLASSIFICATION DE GÈNES

Classification de gènes

Le problème général de la **classification** est de construire une procédure permettant de grouper des objets en les associant à des **classes** selon un ou plusieurs critères. Les applications ou domaines d'utilisation sont légions (*marketing*, génétique, segmentation d'images satellites, reconnaissance de formes, exploration de bases de documents,...) et le terme utilisé peut alors sensiblement varier selon le contexte : apprentissage non supervisé en reconnaissance de formes, taxonomie en écologie, typologie en sciences sociales,...

Comme précisé dans la partie 1.2, un certain nombre d'algorithmes de classification ont été proposés dans la littérature bioinformatique pour l'analyse des données d'expression. De telles méthodes de classification reposent sur des hypothèses peu réalistes d'indépendance entre les gènes ou sous-optimales en transformant les données individuelles en distances entre les gènes. Nous préfèrerons nous tourner vers un modèle probabiliste intégrant une structure de réseau de gènes. Une structure explicite des données sera proposée : leur distribution probabiliste reflètera leur niveau d'organisation. Nous tolérerons ainsi une variabilité des mesures liées aux mécanismes de production des données. On optimisera une fonction (la vraisemblance) qui mesure l'écart d'une structure solution aux données.

Nous présentons donc ici **les champs de Markov** (parfois nommés moins précisément modèles de Markov) **cachés**.

Leurs applications vont de la physique statistique, au *machine learning* en passant par les statistiques spatiales (*e.g.* l'épidémiologie [53]), l'économie ([79]), le traitement du signal et l'analyse d'images ([24, 90]), la sociologie ([240]) *etc.* Et bien sûr la biologie assistée par ordinateur (*computational biology*) ! On pourra consulter [130] ou [30] pour une description de certaines d'applications caractéristiques et [137] pour une construction motivée sur des exemples de la généralisation des chaînes de Markov où l'aspect temporel est remplacé par du spatial (sans que cela interdise par la suite de considérer des processus dynamiques sur le graphes). Nous appliquerons ce modèle à la classification de gènes et décrirons donc le modèle de la façon la mieux adaptée à cette problématique.

3.1 Notations employées

Soit S un ensemble d'individus de cardinal N ($|S| = N$). Ces points seront ceux à classer. Dans une application en vision il pourra s'agir des pixels d'une image sur une grille régulière ou de «points d'intérêt» munis d'un voisinage qui n'est pas forcément régulier. En biologie, on regardera plutôt un ensemble d'entités telles que des gènes ou des protéines (on peut aussi s'intéresser aux expériences auquel cas les individus sont les conditions expérimentales). Plus généralement, ce sont les objets ou individus soumis au traitement statistique.

Un jeu d'observations par individu

Pour chacun de ces individus $i \in S$, on observe une donnée individuelle x_i. On est en général confronté à des données multidimensionnelles. On notera la dimension des données D : $x_i \in \mathbb{R}^D, \forall i \in S$ pour des mesures réelles. On pourra aussi envisager des mesures binaires ou même qualitatives (voir [61]). Notons que la nature de cette mesure est la même pour tous les individus de S. En effet, les mêmes mesures sont conduites sur toute la population. Cependant, on pourra considérer le cas où cette donnée n'est pas ou est partiellement observée dans la modélisation plus générale de la partie 4.1. D représentera alors la quantité maximale de mesures disponibles. Certaines des dimensions du vecteur x_i pourront alors être manquantes.

L'hypothèse de base de notre méthode probabiliste est de considérer que ce vecteur est la réalisation d'une **variable aléatoire** X_i. Nous verrons plus bas de quoi peut dépendre la loi suivie par cette variable, c'est à dire la façon dont se répartissent les valeurs prises par X_i. Une manière standard de traiter le problème est de considérer que les X_i sont distribuées selon une gaussienne dont les paramètres sont à préciser. En toute généralité, nous n'aurons pas besoin de cette particularité mais seulement de quelques hypothèses à satisfaire pour la loi de la variable représentant la donnée individuelle (par exemple l'identifiabilité au sein d'un mélange).

Une donnée cachée : la classe d'un individu

Il reste à définir les classes de chacun des objets. Rappelons que notre approche a pour but de ranger les individus en groupes. On rassemble cette information dans le vecteur $\mathbf{z} = (z_1, \ldots, z_N)$ où chaque z_i est à valeur dans $C = (c_1, \ldots, c_K)$. On essaiera d'éviter l'abus de notation qui consiste à écrire les éléments de C : $1, \ldots, K$, pour éviter des confusions avec les objets eux-mêmes qui sont étiquetés $1, \ldots, N$. Cependant cela pourra parfois alléger les notations quand aucune confusion n'est à craindre.

En fait chaque z_i sera un vecteur binaire où une seule valeur sera non nulle, la k^{e} si $z_i = c_k$. Cette notation sera généralisée à un vecteur de nombres positifs dont

la somme vaut 1. La k^{e} composante reflètera alors une (probabilité d') appartenance à la classe c_k. Ce passage au continu ne pose pas de problème puisqu'on considère un nombre de classes fini. Nous n'envisagerons une généralisation de cette hypothèse autorisant un individu à être membre de plus d'un groupe dans la discussion.

De même que pour les observations x_i, nous considérerons que les z_i sont les réalisations d'une variable aléatoire Z_i dont la forme de la distribution sera précisée très vite (partie 3.3, équations (3.2) et (3.4) pour les impatients). Comme nous sommes dans un cadre non supervisé, les variables Z_i ne sont pas observées mais **cachées**. Elles permettent de rendre compte d'un processus qui n'a pas forcément de signification, d'interprétation physique directe. Elles sont en revanche les indicatrices d'un phénomène qui organise ou d'un régime de fonctionnement des mécanismes qui font produire aux individus de la population les observations x_i des expériences. Une autre différence par rapport aux observations est que nous ne considérerons que des variables Z_i à valeurs dans un ensemble discret de cardinal fini.

Des interactions entre les individus

Des outils classiques traitent les objets de S comme indépendants pour étudier les données $(x_i)_{i\in S}$. Nous envisageons des applications pour lesquelles les relations entre les objets ou les mesures observées semblent avoir une importance capitale. Il nous a donc semblé naturel d'inclure ces relations dans la modélisation.

Pour ce faire, nous supposons la donnée d'un graphe G dont les sommets sont les points de S. Une arête entre deux nœuds de G représentera une interaction entre les éléments de S correspondants. Une **relation de voisinage** ν statuera quelle(s) dépendance(s) existe(nt) entre les éléments du graphe. On pourra regarder [30] pour la traduction exacte d'un modèle graphique en terme d'indépendances conditionnelles. Par exemple, dans la figure 3.1, le couple (X_1, X_2) est indépendant des deux variables (X_5) et X_6 conditionnellement (*i.e.* si on le connaît) à (X_3, X_4). $\nu(i)$ décrit l'ensemble des voisins du site i dans le graphe G. Un voisinage ν est donc une application de S vers l'ensemble des parties de S. Elle est symétrique dans le sens où : j est dans le voisinage de i est équivalent au fait que i soit dans le voisinage de j ; on considère des graphes non orientés. Par définition, un site n'appartient pas à son voisinage.

Examinons dans un premier temps la configuration d'indépendance (section 3.2) puis par la suite les modèles de champ de Markov (section 3.3) qui en sont une généralisation.

3.2 Le modèle de mélange pour la classification

Une façon de prendre en compte l'information probabiliste simultanée sur les obervations et sur les classes est le **modèle de mélange**. Il apparaît en effet

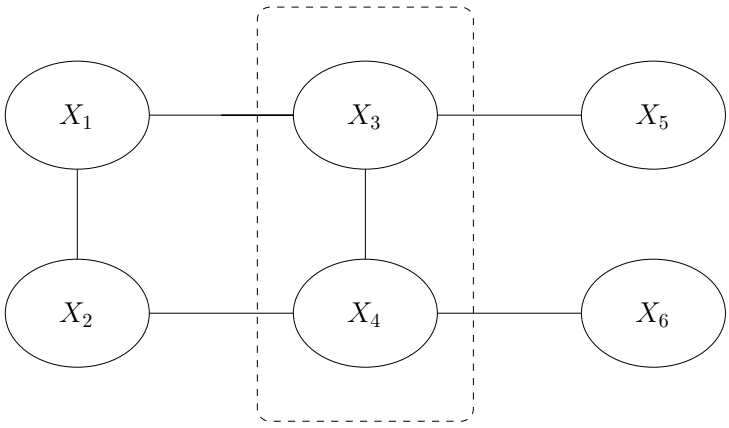


Fig. 3.1: Exemple d'indépendance conditionnelle graphique de (X_1, X_2), X_5 et X_6 sachant (X_3, X_4).

adapté de présenter les observations issues de K distributions homogènes dans l'objectif d'aboutir à une classification. Il permet ainsi de modéliser des données modernes et complexes dont la distribution ne peut pas être l'archétype d'une loi classique. Nous nous placerons ici dans le cadre de distributions paramétriques. Nous nous restreindrons dans nos applications aux données aux mélanges de lois normales (ou gaussiennes). Cependant, nous conserverons une présentation générale tant que cela n'implique pas de complexité supplémentaire. Un modèle de mélange ([159]) se base sur deux hypothèses :

- la classification $\mathbf{z}$ est tirée pour chacun de ses composants z_i de façon indépendante selon une loi multinômiale dont les paramètres sont les proportions du mélange :

$$P(\mathbf{Z}, \Delta) = \prod_{i=1}^{N} P(Z_i, \Delta),$$

  où Δ est l'ensemble des paramètres associés à la loi de $\mathbf{Z}$. Ici, Δ se limite donc aux proportions du mélange c'est à dire aux probabilités de répartition des objets dans chacune des composantes (ou classes du mélange) : $(\pi_1, \ldots, \pi_K)$, avec $\pi_k \in]0; 1[, \forall k \in \{1, \ldots, K\}$ (si un des π_k est nul, une classe est vide et le modèle s'écrit avec au plus $K-1$ groupes) et $\sum_{k=1}^{K} \pi_k = 1$. Les π_k sont les proportions (inconnues en classification non supervisée) du mélange. Une fois cette classification fixée,

- chaque x_i est indépendamment des autres, issu d'une loi paramétrée par $\theta_{z_i} := \theta_k$ si $z_i = c_k$. On notera aussi $\Theta = (\theta_1, \ldots, \theta_K)$ les paramètres de classe et $\Psi = (\Theta, \Delta)$ le jeu total des paramètres du modèle dans toute la suite de ce mémoire.

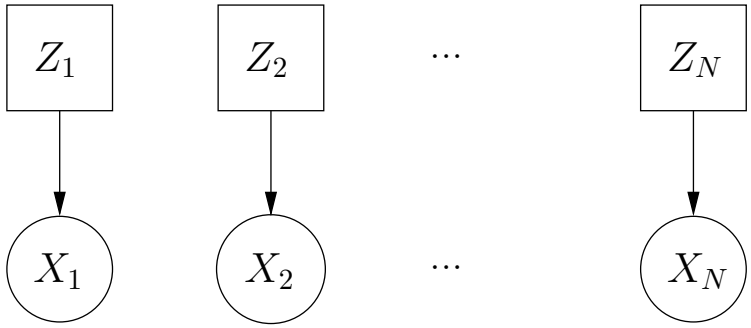


Fig. 3.2: Modèle graphique à variables indépendantes : c'est le graphe du mélange.

Cette situation est illustrée dans la figure 3.2. On a souvent besoin de la probabilité marginale d'observation de la variable en un point particulier $y \in \mathbb{R}^D$. Elle s'écrit alors :

$$P(y|\Psi) = \sum_{k=1}^{K} \pi_k f(y|\theta_k),$$

où $f(.|\theta_k)$ est la densité de probabilité des variables dont les objets sont assignés à la classe c_k.

En effet, on a que $P(\mathbf{x}|\mathbf{Z}, \Theta) = \prod_{i=1}^{N} P(x_i|Z_i, \theta_{Z_i})$, les dépendances entre les observations ne provenant que des classes qui les engendrent. La loi jointe s'écrit $P(\mathbf{x}, \mathbf{Z}|\Psi) = P(\mathbf{x}|\mathbf{Z}, \Theta)P(\mathbf{Z}|\Delta) = \prod_{i=1}^{N} P(x_i, Z_i|\Psi)$ avec l'équation précédente. En sommant sur les valeurs possibles prises par la classification $\mathbf{z}$ pour cette dernière expression, on retrouve bien que la loi marginale (ou vraisemblance des paramètres Ψ) pour $\mathbf{X}$ considère ces variables comme indépendantes :

$$P(\mathbf{x}|\Psi) = \prod_{i=1}^{N} P(x_i|\Psi) = \prod_{i=1}^{N} \sum_{k=1}^{K} \pi_k f(x_i|\theta_k). \tag{3.1}$$

On estimera les paramètres par *maximum de vraisemblance*. Outre son interprétation de meilleure estimateur pour expliquer les données dans le cadre probabiliste, cet estimateur est en effet sans biais consistant (c'est à dire converge en probabilité vers la vraie valeur du paramètre quand la taille de l'échantillon grandit) sous certaines conditions. Sa distribution est asymptotiquement gaussienne ([57]).

Notons aussi l'importance de la probabilité *a posteriori* de la classification conditionnellement aux données : $P(Z_i|x_i)$. Dans le cadre du modèle de mélange elle ne dépend pas du site i et s'obtient par formule d'inversion de Bayes :

$$P(Z_i = c_k|x_i) = \frac{\pi_k f(x_i|\theta_k)}{\sum_{j=1}^{K} \pi_j f(x_i|\theta_j)}.$$

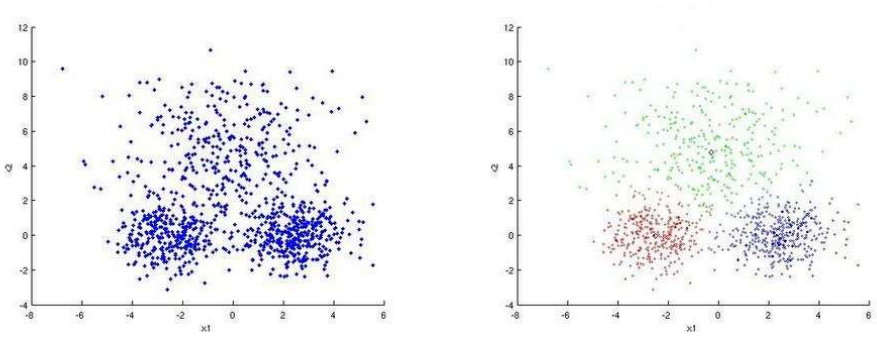


Fig. 3.3: À gauche : données issues d'un modèle de mélange gaussien bi-dimensionnel à trois composantes et à droite : mêmes données étiquetées.

Tandis que π_k sera vu comme la probabilité *a priori* de $Z_i = c_k$, on dira que la probabilité *a posteriori* est la «responsabilité» ([30]) de la classe c_k pour expliquer l'observation x_i.

Cas particulier des densités normales

Lorsqu'on est en présence de données quantitatives continues, sans autre connaissance particulière, il est fréquent de supposer que chaque classe suit une **loi normale** (ou **gaussienne**) multivariée de densité :

$$f(y|\theta_k) = \frac{1}{\sqrt{(2\pi)^D|\Sigma_k|}} \exp\left\{-\frac{(y-\mu_k)^t\Sigma_k^{-1}(y-\mu_k)}{2}\right\},$$

Chaque classe gaussienne est paramétrée par le vecteur moyen $\mu_k \in \mathbb{R}^D$ et la matrice de covariance Σ_k ($D \times D$, symétrique définie positive). Les valeurs des variables y dans cette classe s'écartent de la moyenne avec des fréquences de plus en plus faible conformément à une échelle dictée par la matrice de covariance Σ_k. Σ_k^{-1} –où M^{-1} est la notation pour l'inverse d'une matrice carrée inversible M- est parfois appelé la matrice de précision de la loi gaussienne. $|M|$ est le déterminant de la matrice carrée M. La figure 3.3 donne l'exemple de données simulées à partir d'un modèle de mélange gaussien bivarié à trois composantes.

De manière classique, on applique des **transformations sur les données** pour que l'hypothèse de distribution gaussienne soit jugée admissible lorsque cela est nécessaire. Par exemple, pour des données trop dissymétriques (*skew* en anglais), on utilise une transformation logarithmique. Cela réduit aussi l'influence de grandes valeurs atypiques. Ceci se justifie devant des données réelles parce que certains phénomènes sont mieux modélisés par des facteurs multiplicatifs plutôt qu'additifs. Dans ce dernier cas en effet, leur échelle de variation est bien modélisée selon une certaine fonction. On pensera par exemple au rapport entre la puissance auditive et définition du décibel.

On peut aussi opérer un recentrage des données. Celles-ci se présentant sous la forme d'une matrice $N \times D$, il faut alors savoir s'il est plus pertinent de centrer en lignes, en colonnes, ou les deux. Cela permet de prendre en compte la dispersion des données indépendamment de leur origine.

A la suite de ce recentrage (par la moyenne par exemple), il peut être utile de réduire les données. L'unité de mesure (ou l'échelle) n'est pas forcément à prendre en compte si elle n'est pas équivalente d'une variable à l'autre. Les variances peuvent en effet être très différentes sans que cela ait une réelle signification sur la variabilité des individus concernés. Encore une fois il faut avoir une idée du phénomène mis en jeu et de la pertinence d'une telle opération vis-à-vis de la question soulevée. Les deux opérations de recentrage puis réduction des données portent le nom de **standardisation**. On mentionnera aussi les divisions par les sommes marginales dans les tableaux représentant un nombre d'individus. La division par l'étendue (différence entre la valeur maximale et la valeur minimale) porte le nom de **normalisation**. Une dernière transformation existante est de remplacer les valeurs par leur rangs quand aucune transformation usuelle (log éventuellement itéré, puissance) n'est efficace.

La formule de la densité gaussienne peut s'interpréter comme une distribution ellipsoïdale pour le *cluster* centré en la moyenne μ_k et dont les caractéristiques géométriques sont précisées dans Σ_k. La décomposition spectrale de $\Sigma_k = \lambda_k D_k A_k D_k^t$ qui est symétrique définie positive fait intervenir $\lambda_k := |\Sigma_k|^{\frac{1}{D}}$, le volume de la classe, D_k une matrice orthogonale de taille $D \times D$ constituée des **vecteurs propres** de la matrice de covariance qui donne donc l'orientation de la classe et A_k une matrice diagonale dont les termes diagonaux $(a_{k1}, \ldots, a_{kD})$ sont les **valeurs propres** de Σ_k rangées par ordre décroissant et normalisées de telle manière que $\prod_{j=1}^{D} a_{kj} = 1$. Cette matrice mesure la forme de l'ellipsoïde.

Selon les libertés ou les contraintes qu'on impose à ces trois paramètres, on maîtrise la forme de chacune des classes et le fait qu'on peut les considérer identiques ou pas selon certaines caractéristiques. Cela peut permettre de réduire le nombre de paramètres lorsqu'on a affaire à des données en grande dimension. On parle alors de modèles gaussiens parcimonieux ([16, 48]). D'autres contraintes pour faire face aux données de grandes dimensions ont été proposées et ont prouvé leur efficacité pour la reconnaissance d'objets basée sur les descripteurs d'une image ([38]). Ce travail a noté que les modèles parcimonieux de [16] les plus employés parce qu'efficaces utilisaient encore trop de paramètres ou étaient trop simples notamment pour capter la structure complexe des données. L'auteur propose alors de combiner réduction de dimension, modèles parcimonieux et régularisation en évitant les écueils de ces approches. Cela nous a motivé pour étudier l'intégration de cette approche dans notre modèle markovien. Cette option de modélisation sera précisée à la partie 4.2.

Supposer la matrice de covariance diagonale, revient à faire l'hypothèse que les composantes des données sont indépendantes. Par exemple, on peut choisir

des classes sphériques (la force des interactions est identique dans toutes les directions) : $\Sigma_k = \lambda_k I$ et même forcer toutes les classes à avoir le même rayon : $\lambda_k = \lambda, \forall k = 1, \dots K$. On estime ainsi seulement 1 ou K paramètre(s) contre $K\frac{D(D+1)}{2}$ pour la matrice de covariance avec le modèle le plus général.

Dans le cas de modèles ayant les mêmes variances pour toutes les classes, on parle de modèle linéaire.

Nous avons présenté la forme des données dans chacune des classes en supposant les données mutuellement indépendantes. Nous précisons maintenant la distribution des classes. Nous ne la supposerons plus multinômiale comme dans le cas des modèles de mélange. La forme de cette distribution sera spécifiée par un **Champ de Markov** (*Markov Random Field*) que nous présentons dans les sections à venir. Leur caractéristique est de permettre de modéliser les dépendances entre individus.

3.3 Champs de Markov

Propriété(s) de Markov

La collection de variables aléatoires $\mathbf{Z} = (Z_1, \dots, Z_N)$ dont la loi est définie par un ensemble de paramètres Δ est un champ de Markov si et seulement si sa distribution vérifie les deux propriétés :

$$P(Z_i = z_i | \mathbf{Z}_{S\setminus i} = \mathbf{z}_{S\setminus i}, \Delta) = P(Z_i = z_i | \mathbf{Z}_{\nu(i)} = \mathbf{z}_{\nu(i)}, \Delta), \ \forall \mathbf{z} \tag{3.2}$$

$$P(\mathbf{Z} = \mathbf{z} | \Delta) > 0, \ \forall \mathbf{z}, \tag{3.3}$$

où $\mathbf{Z}_A$, $A \subset S$ désigne $\{Z_i, i \in A\}$. La propriété markovienne (3.2) traduit que l'information utile du graphe sur i se limite à celle portée par les voisins définis par ν sur G. La seconde propriété (3.3) est une propriété de positivité du champ. Elle interdit à toute configuration d'avoir une probabilité nulle d'observation. Nous avons ici donné une version **locale** de la propriété de Markov : Z_i est indépendant de $\mathbf{Z}_{S\setminus\nu(i)\cup\{i\}}$ conditionnellement à $\mathbf{Z}_{\nu(i)}$. Il est intéressant de noter que cette caractérisation est équivalente (dans le cas où S est fini) à une propriété **globale** : tout $\mathbf{Z}_A$ est indépendant de $\mathbf{Z}_C$ conditionnellement à $\mathbf{Z}_B$ où A, B et C sont des sous-ensembles de S et où B sépare A et C au sens de ν. Une dernière caractérisation par paires stipule que pour tout $(i, j) \in S^2$, Z_i et Z_j sont indépendants conditionnellement à $\mathbf{Z}_{S\setminus\{i,j\}}$ ([140, 241]).

La propriété markovienne a l'avantage d'être traduite localement mais définir une distribution de probabilité comme en (3.2) mène à des difficultés de calcul. On sait qu'une chaîne de Markov est entièrement déterminée par la donnée d'une distribution initiale et des probabilités de transition. Qu'en est-il pour un champ de Markov ? Une idée naturelle au vu de la formule (3.2) est qu'il faut préciser des distributions conditionnelles de la forme $P(Z_i = z_i | \mathbf{Z}_{\nu(i)} = \mathbf{z}_{\nu(i)}, \Delta)$.

On imagine facilement que pour des systèmes complexes sans même aborder des graphes issus de problèmes réels, pour de telles lois conditionnelles données, il est difficile de vérifier qu'elles définissent bien une probabilité sur l'espace C^N des valeurs de $\mathbf{Z}$ ([54, 147]). [12] donne une condition nécessaire et suffisante pour le cas de deux jeux de variables disjoints. [89] donne une condition suffisante pour que des probabilités marginales conditionnelles définissent l'unicité de la probabilité jointe...A condition de trouver un réarrangement qui permette de définir cette probabilité jointe et qu'elle soit à support positif.

Forme opérationnelle des champs de Markov

Ce problème de définition d'un champ de Markov est en partie résolu par un caractérisation pratique de la forme prise par sa loi de distribution.

Le **théorème de Hammersley-Clifford** dit qu'un champ de Markov est équivalent à un système de variables aléatoires régi par une **distribution de Gibbs** (dont la forme est donnée ci-dessus dans l'équation (3.4)). Ce théorème date de 1968 mais ses auteurs ont longuement retardé sa publication. Ils pensaient que l'hypothèse de positivité était trop restrictive [1]. On l'énonce ainsi en s'inspirant de l'exposé dans [25] (qui en plus fournit une preuve simplifiée de l'argument de «circuité») :

Un champ aléatoire $\mathbf{Z}$ est un champ de Markov si et seulement si sa distribution est gibbsienne *i.e.* :

$$P(\mathbf{Z} = \mathbf{z}|\Delta) = P_G(\mathbf{Z} = \mathbf{z}|\Delta) := \frac{\exp\left[-H(\mathbf{z}, \Delta)/T\right]}{W(\Delta)}, \tag{3.4}$$

où la fonction d'énergie H se décompose selon une somme de **potentiels** V_c sur les **ensembles complets** [2] c du graphe G :

$$H(\mathbf{z}, \Delta) := \sum_c V_c(\mathbf{z}_c, \Delta), \tag{3.5}$$

et où $W(\Delta)$ est la constante de normalisation aussi appelée **fonction de partition**. Son calcul est dans la plupart des cas problématique puisqu'il implique

[1] [170] donne d'ailleurs un contre exemple à 4 variables du fait que la positivité est bien requise.

[2] Un **ensemble complet** est un sous-ensemble dont les éléments sont des nœuds de S et tel que deux éléments distincs sont voisins deux à deux. Une **clique** dans un graphe est un ensemble complet de taille maximale *i.e.* tel que l'ensemble de ces nœuds soient voisins les uns des autres. Un ensemble dont deux objets ne seraient pas voisins n'est pas complet donc pas une clique. Si on trouve au moins un élément voisin de tous les éléments d'un ensemble complet, alors ce dernier n'est pas de cardinalité maximale ; il ne s'agit donc pas non plus d'une clique. Certains auteurs appellent clique ce que nous avons décrit comme ensemble complet et parlent de max-cliques pour décrire les cliques. Il pourra nous arriver de confondre ces deux définitions mais il est important de noter que la factorisation dans (3.4) se fait sur les ensembles complets

la somme de numérateur de l'équation (3.4) sur toutes les valeurs prises par $\mathbf{Z}$ possibles. Cette quantité augmente exponentiellement avec la taille N.

T est une constante appelée température. Cette grandeur a une existence physique réelle dans les problèmes de mécanique statistique. Les distributions de Gibbs voient le jour dans cette discipline pour capturer les interactions entre particules de systèmes physiques à l'echelle moléculaire. Les états que ce système peut prendre sont alors décrits par la variable $\mathbf{Z}$. En effet, la forme de la distribution d'énergie H maximise l'entropie $\mathbf{S}(P) = -\sum_{\mathbf{Z}} P(\mathbf{z}) \ln P(\mathbf{z})$ tout en gardant son énergie moyenne $U(P) = \sum_{\mathbf{Z}} P(\mathbf{z})H(\mathbf{z})$ constante. Ces deux principes contradictoires sont arbitrés par la température du système de façon à ce que ce dernier minimise globalement son **énergie libre** $F = U - T.\mathbf{S}$. Ce minimum (de valeur $-T \ln W(\Delta)$) est justement atteint pour la distribution gibbsienne.

L'interprétation que nous pouvons faire de la température est qu'elle contrôle la largeur de la distribution ; à températures élevées, on tend vers des configurations équiprobables quelle que soit la fonction d'énergie. C'est le terme d'entropie qui est dominant tandis que pour des températures faibles, le système reste très proche d'un état fondamental : la distribution se concentre près des configurations correspondant aux *minima* de la fonction d'énergie. En sociologie par exemple, ce paramètre indique si une société est plus ou moins de nature libérale. Un système totalitaire, à basse température verra le phénomène où une attitude qui apparaît par hasard se répand dans toute la société avec seulement peu d'opposants. A l'inverse un système libertaire à plus haute température est plus proche de l'indépendance. Entre les deux, le cas de la température critique rencontré en physique statistique serait celui d'une société très polarisée où de grands groupes de gens ont des attitudes opposées. Ce type de modélisation peut servir à étudier des comportements lors d'un scrutin.

T ne présente pas vraiment d'intérêt dans notre cadre et nous la supposerons égale à 1. Son utilisation est en revanche intéressante dans le cas d'un algorithme de type recuit-simulé qui chercherait à fournir le champ $\mathbf{z}$ qui conduit aux *minima* d'énergie. Ceci est possible car des sauts autorisent des remontées d'énergie. Ces dernières sont petit à petit supprimées au fur et à mesure qu'on se rapproche de l'*optimum* avec la décroissance de la température.

Un champ de Markov est dit **homogène** si la valeur des potentiels ne dépend pas de la position relative des ensembles complets. C'est à dire que les potentiels ne dépendront plus des indices des sites concernés mais uniquement de leur classe. Si le potentiel ne dépend pas de l'orientation de l'arête dans l'ensemble complet, on le dit **isotropique**. Cette notion est surtout utile sur des grilles régulières avec par exemple pour les ensembles complets d'ordre (ou cardinal) 2 où on peut distinguer (champ anisotrope) ou non la valeur du potentiel selon l'orientation Nord, Est, Sud, Ouest. Pour plus de détails, on pourra consulter [147].

On ramène donc le problème de la définition du champ de Markov à la spécification des fonctions-potentiels sur les ensembles complets. Plus précisément, on

devra donner la **forme canonique du potentiel** qui est normalisée (voir [130] ou [147]). L'unicité de la forme canonique du potentiel est assurée en considérant un état «favorisé» parmi $c_1, \ldots, c_K$. Dès lors qu'un site de c sera dans cet état favorisé on contraint $V_c(\mathbf{z}_c)$ à valoir 0. Ainsi on précisera les dépendances spatiales locales.

Remarquons que si le champ $\mathbf{Z}$ est le vecteur des N assignations de chacun des objets parmi une des K classes, le calcul de la constante de normalisation $W(\Delta) = \sum_{\mathbf{Z}} \exp\left[-H(\mathbf{z}, \Delta)\right]$ fait intervenir K^N termes. Les simplifications sont rares en pratique du fait des interactions entre les Z_i et cette grandeur est donc généralement difficile (ce qui signifie impossible dans la pratique) à calculer. Pourtant quand elle est calculée, tout le travail est fait ou presque. Par exemple, l'énergie moyenne s'en déduit puisque $U(\Delta) = -\frac{\partial \ln Z(\Delta)}{\partial \Delta}$. Une solution est de recourir à la pseudo-vraisemblance introduite par Julian Besag [25] et définie comme :

$$PL(\mathbf{z}) = \prod_{i=1}^{N} P_G(z_i | \mathbf{z}_{\nu(i)}, \Delta).$$

On peut l'utiliser pour estimer les paramètres du champ de Markov. Elle autorise la simplification de la constante de normalisation dans chacun des termes du produit :

$$P_G(z_i | \mathbf{z}_{\nu(i)}, \Delta) = \frac{\exp\left(-\sum_{c,\, i \in c} V_c(\mathbf{z}_c, \Delta)\right)}{\sum_{z_i'} \exp\left(-\sum_{c,\, i \in c} V_c(\mathbf{z}_c', \Delta)\right)},$$

où $\mathbf{z}_c' = \{z_i'\} \cup \mathbf{z}_c$. La somme sur les z_i ne contient donc que K termes. Insistons cependant sur le fait que cette pseudo-vraisemblance n'est pas une loi de probabilité à moins qu'il n'y ait réellement indépendance entre les sites i.

Nous allons maintenant voir comment cette modélisation par champ de Markov sur les classes peut être utilisée pour la classification de gènes. Nous disposons d'un graphe et d'une forme de distribution pour la loi des classes. Reste maintenant à introduire les mesures faites sur les objets du graphe, vues comme des réalisations de variables aléatoires.

3.4 Modèles de champ de Markov cachés

Pour en revenir à la fin du paragraphe 3.2, nous voulons relâcher l'hypothèse d'indépendance intrinsèque aux modèles de mélange. Les **champ de Markov cachés** permettent de prendre en compte les dépendances spatiales entre sites voisins dans la distribution de $\mathbf{Z}$.

Une manière de voir un champ de Markov est comme la généralisation d'une chaîne de Markov cachée bilatérale [3] (figure 3.4). La dépendance n'est plus uni-

[3] la valeur dépend d'une fenêtre de valeurs passées et futures à horizon fini : l'ordre de la

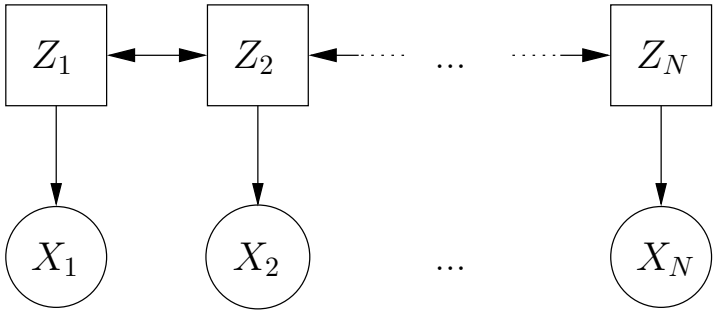


Fig. 3.4: Chaîne de Markov cachée bilatérale (ou non causale)

dimensionnelle (souvent assimilée au temps) mais précisée par la structure spatiale de graphe. La dépendance entre les variables cachées introduite par rapport au modèle indépendant peut être représentées graphiquement (figure 3.5). Les arêtes du graphe (en vert) peuvent être pondérées (voir la partie 3.5) par un sous-ensemble des paramètres définissant $\mathbf{Z}$. On notera que les variables observées sont quant à elles indépendantes conditionnellement aux classes en vertu de la structure imposée au graphe.

Le modèle utilisé

On considèrera la présence de deux grandeurs descriptives du point i : la donnée individuelle x_i et son étiquette de classe z_i. Nous traiterons d'abord le cas où la première est complètement observée et la deuxième cachée. Les classes sont distribuées selon le champ de Markov $\mathbf{Z}$ défini ci-dessus par les équations (3.2) ou (3.4). Les données individuelles $\mathbf{X} = (X_1, \ldots, X_N)$ sont alors indépendantes conditionnellement à $\mathbf{Z}$. Leur loi est paramétrée par θ :

$$P(\mathbf{X}|\mathbf{Z}, \theta) = \prod_{i \in S} P(X_i|Z_i, \theta_{Z_i}).$$

Notons que nous ne considérerons que le cas des distributions normales multivariées (en dimension D) pour $P(X_i|Z_i, \theta_{Z_i})$ dans nos manipulations à la partie 5. La loi marginale d'une observation se décompose alors selon les K classes :

$$P(X_i|\mathbf{\Psi}) = \sum_{k=1}^{k=K} P_G(Z_i = c_k|\Delta) P(X_i|Z_i = c_k, \theta_k).$$

chaîne. Notons qu'une chaîne de Markov unilatérale (ou causale) est bien une chaîne bilatérale du même ordre. L'inverse est en toute généralité faux (voir [96]). La distribution conditionnelle de probabilités aux extrémités d'une chaîne unilatérale doit satisfaire certaines propriétés qui ne sont pas nécessaires pour la chaîne seulement bilatérale. En fait, il faut fixer que deux fois plus de valeurs initiales et finales que l'ordre de la chaîne (un nombre fini donc tout de même) pour avoir une chaîne unilatérale entre ces extrémités fixées.

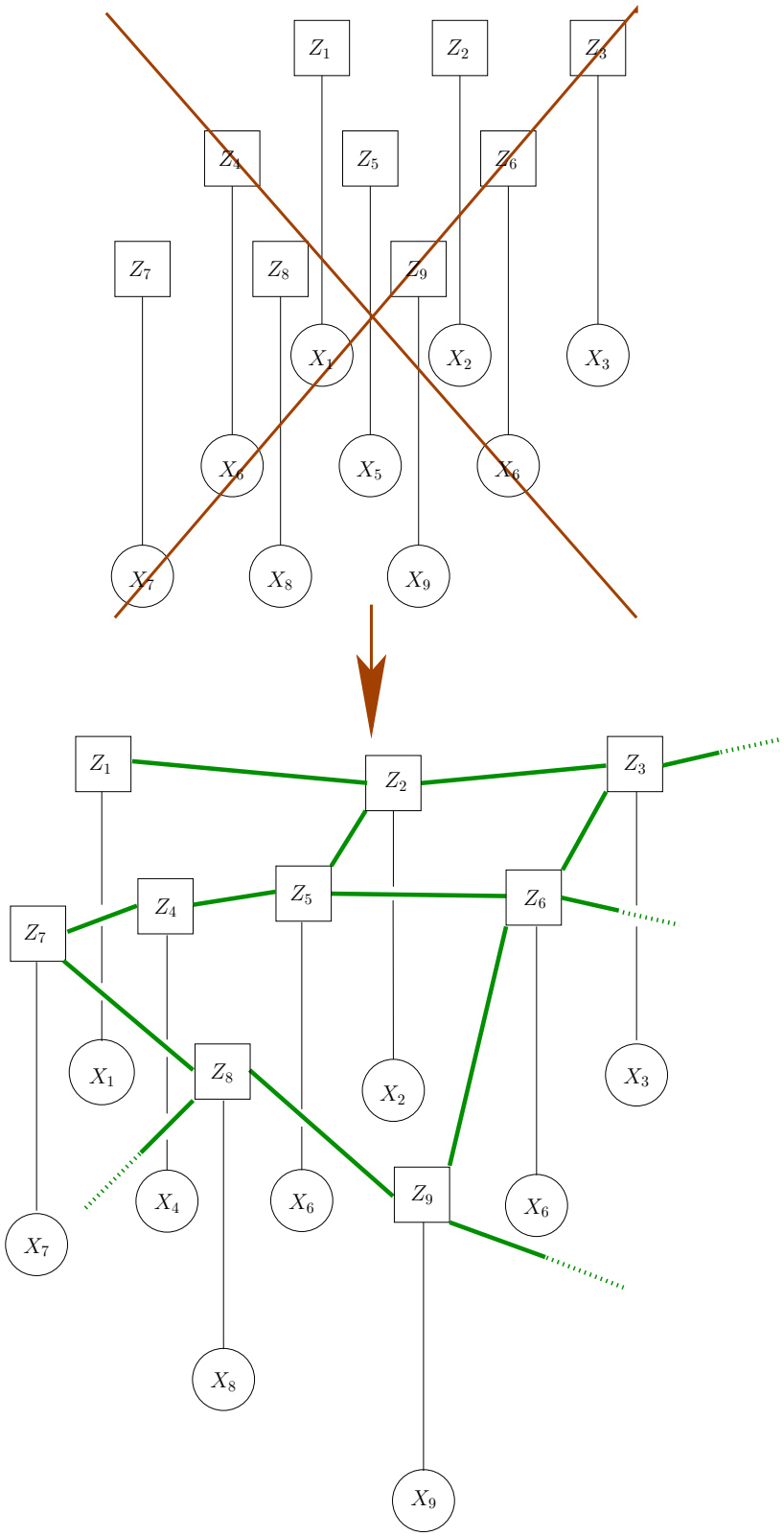


Fig. 3.5: Graphe de champ de Markov caché (en bas) par opposition au modèle de mélange sur le même champ (en haut).

Nous avons alors une formule analogue à celle des modèles de mélange (voir la page 51) sauf que les proportions π_k sont remplacées par les probabilités gibbsiennes $P_G(Z_i = c_k|\Delta)$; (la classe d')une observation n'est plus indépendante des autres (observations) ni du site concerné.

On notera que la probabilité *a posteriori* [4] s'exprime toujours grâce à la formule d'inversion de Bayes :

$$P(\mathbf{Z}|\mathbf{X}, \Psi) = \frac{P(\mathbf{Z}|\Delta)P(\mathbf{X}|\mathbf{Z}, \theta)}{P(\mathbf{X}|\theta)}$$

Le terme au numérateur ne dépend que des données. Donc à une constante près :

$$P(\mathbf{Z}|\mathbf{X}, \Psi) \propto \exp \underbrace{\left\{ -H(\mathbf{Z}|\Delta) + \sum_{i \in S} \log P(X_i|Z_i, \theta) \right\}}_{:= -H(\mathbf{Z}|\mathbf{X}, \Delta)}.$$

D'après le théorème de Hammersley-Clifford, $\mathbf{Z}|\mathbf{X}$ est aussi distribué selon un champ de Markov dont la fonction d'énergie comprend un terme propre au graphe (la fonction d'énergie du champ $\mathbf{Z}$) et un terme dit d'attache aux données : $-\sum_{i \in S} \log P(X_i|Z_i, \theta)$.

On distinguera le champ de Markov marginal $\mathbf{Z}$ et le champ de Markov conditionnel $\mathbf{Z}|\mathbf{X} = \mathbf{x}$. Une difficulté majeure lors de l'utilisation des champs de Markov réside dans le calcul des quantités utilisant des distributions marginales ou conditionnelles qui sont spécifiées par les paramètres. La section suivante 3.5 traitera des approximations envisageables pour approcher les valeurs des paramètres $\Psi = (\theta, \Delta)$.

3.5 Problème de l'estimation des paramètres

À ce stade, nous disposons d'un modèle à deux niveaux. Le premier est celui de la réponse $\mathbf{x}$ dont la loi est donnée conditionnellement au processus latent. Ce processus latent $\mathbf{Z}$ reflète en quelques sortes une hétérogénéité structurée des données qui n'est pas directement observée. La finalité de la procédure est de fournir les étiquettes $\mathbf{z}$ inconnues de façon à permettre une analyse experte ultérieure. On considère ces étiquettes aléatoires, distribuées selon un champ de Markov (caché donc) $\mathbf{Z}$. On a aussi postulé une certaine forme de distribution (paramétrique) de $\mathbf{X}$ conditionnellement à $\mathbf{Z}$. Celle-ci reflète notre conception du bruit, du trouble ou de tout ce qui peut transformer des données parfaites qui correspondrait aux classes théoriques en une observation. Les mécanismes inhérents au processus étudié peuvent aussi être responsables d'inhomogénéités. La

[4] C'est à dire la probabilité de la classification conditionnellement à l'information apportée par les données dans le cadre du modèle.

formule de Bayes permet alors en principe de calculer la distribution *a posteriori* des classes conditionnellement aux données. C'est ainsi qu'il est possible de produire des estimateurs raisonnables pour les étiquettes $\mathbf{z}$ en tenant compte des données $\mathbf{x}$. Nous allons dans la suite essayer d'expliquer l'approche qui permet une estimation simultanée des paramètres de la loi des observations, de la loi de la classification cachée.

Lorsque les paramètres Ψ du modèle sont connus ?

On a vu à la partie 3.1 qu'en adoptant un modèle de champ de Markov caché, on définissait une loi de probabilité $P(\mathbf{X}, \mathbf{Z}|\Psi)$. Cette loi relie données observées $\mathbf{x}$ et classification cachée $\mathbf{z}$. On va alors donner la stratégie de classification. Elle fournit un cadre pour le choix de la valeur de la variable $\mathbf{Z}$ pour laquelle on connaît seulement la distribution conditionnellement aux observations : $P(\mathbf{Z}|\mathbf{X}, \Psi)$.

On définit d'abord le coût d'une décision $\tilde{\mathbf{z}}$ sur $\mathbf{Z}$ par rapport à la vraie valeur cachée $\mathbf{z}$: $l(\tilde{\mathbf{z}}|\mathbf{z})$ (positif et nul si et seulement si $\mathbf{z} = \tilde{\mathbf{z}}$). Pour prendre une décision, on s'appuie sur les données observées $\mathbf{x}$. On s'appuie sur le coût ou risque conditionnel :

$$R(\tilde{\mathbf{z}}|\mathbf{x}, \Psi) = \overbrace{I\!E[l(\tilde{\mathbf{z}}|\mathbf{Z})|\mathbf{X}, \Psi]}^{:=\sum_{\mathbf{Z}} l(\tilde{\mathbf{z}}|\mathbf{Z}) P(\mathbf{Z}|\mathbf{X}, \Psi)} .$$

En effet, la vraie classification étant inconnue, on calcule la valeur du risque conditionnellement à son estimation selon la loi imposée par le modèle. Une fonction de coût simple est celle qui prend la valeur 1 en cas de bonne prédiction et 0 sinon. Ce «coût du $0-1$» conduit à une valeur de risque conditionnel qui n'est autre que la probabilité *a posteriori* de fournir une mauvaise classification $\tilde{\mathbf{z}}$ connaissant les données et les paramètres du système. On choisit donc le $\mathbf{z}$ qui réalise $\arg\min_{\tilde{\mathbf{z}}} P(\mathbf{Z} \neq \tilde{\mathbf{z}}|\mathbf{X} = \mathbf{x}, \Psi)$. C'est exactement celui qui maximise la distribution *a posteriori* : $P(\mathbf{Z} = \mathbf{z}|\mathbf{X} = \mathbf{x}, \Psi)$. On parle alors d'estimation du *Maximum A Posteriori* (MAP).

Cela signifie que l'on choisit la valeur la plus probable de $\mathbf{z}$ conditionnellement aux données. Dans le cas du mélange, cela revient à choisir la classe la plus probable pour chaque observation $i = 1, \ldots, N$ puisque ces dernières sont indépendantes. Dans le cas des champs de Markov cachés, les dépendances entre les N sites ne permettent pas la factorisation en N règles comme pour le mélange. On ne peut pas non plus énumérer toutes les configurations pour trouver celle qui maximise la probabilité *a posteriori*. Stuart et Donald Geman [90] proposent de coupler un échantillonneur de Gibbs et un principe de recuit simulé pour chercher le maximum global de $\mathbf{z} \mapsto P(\mathbf{z}|\mathbf{X} = \mathbf{x}, \Psi)$. L'idée est de simuler non pas la distribution *a posteriori* directement mais une version avec un paramètre supplémentaire de température dont le contrôle permet de maximiser globalement la probabilité d'intérêt en maximisant seulement les probabilités conditionnelles

$P(Z_i = c_k|\mathbf{z}_{S\setminus\{i\}}, \mathbf{x}, \Psi)$. Malgré de bons résultats pratiques, cette méthode souffre de la nécessité de temps de calculs longs. Une alternative peut être un algorithme ICM (*Iterated Conditional Mode*) proposé par Julian Besag [24]. Sa faiblesse réside dans le fait qu'une classification dure est faite à chaque étape, ce qui biaise l'estimation des paramètres.

D'autres fonctions de coût peuvent être utilisées. Par exemple, la fonction de coût quadratique $l'(\tilde{\mathbf{z}}|\mathbf{z}) = ||\tilde{\mathbf{z}} - \mathbf{z}||^2$. Dans le cas des champs de Markov cachés, cela revient à maximiser indépendemment les probabilités marginales *a posteriori* d'appartenance des objets $i = 1, \ldots, N$ aux classes $c_1, \ldots, c_K$. On peut les calculer par méthodes MCMC ou champs moyens. L'intérêt de cette fonction de coût est qu'elle prend en compte le nombre d'erreurs dans une estimation de la classification ce que ne fait pas le MAP qui pénalise identiquement des classifications faisant peu d'erreurs et des classifications très éloignées de la vraie classification. On pourra consulter [61], pp 90–93 pour plus de détails.

Estimation des paramètres à classification fixée

Pour appliquer de telles méthodes de classification, il faudra connaître au préalable les paramètres $\Psi = (\Theta, \Delta)$. Quand les observations $\mathbf{x}$ et la classification $\mathbf{z}$ sont connues ; on est ramené à un problème d'estimation de paramètres. Plus généralement, on veut rendre maximale la **vraisemblance** (ou en d'autres termes la probabilité) d'observer les données $\mathbf{x}$ les paramètres Ψ étant les variables d'intérêt. Formellement on écrit la log-vraisemblance des paramètres :

$$\log L(\Psi) = \log P(\mathbf{x}|\Psi) = \log \sum_{\mathbf{z}} P(\mathbf{x}, \mathbf{z}|\Psi).$$

Vraisemblance complète des paramètres

On a souvent besoin d'écrire la log-vraisemblance complète du modèle de champ de Markov caché pour éviter la sommation sur toutes les configurations du champ caché :

$$\begin{aligned} \log L_c(\Psi) &= \log P(\mathbf{x}, \mathbf{z}|\Psi) \\ &= \log P(\mathbf{x}|\mathbf{z}, \Theta) + \log P_G(\mathbf{z}|\Delta) \\ &= \sum_{i=1}^{N} \log f(x_i|\theta_{z_i}) \\ &\quad - \sum_{c} V_c(\mathbf{z}_c, \Delta) - \log W(\Delta) \end{aligned} \tag{3.6}$$

La partie de la vraisemblance qui correspond aux paramètres de loi des observations conditionnellement aux classes ne pose aucun problème ; il s'agit de

maximiser la vraisemblance des paramètres Θ en se basant sur les observations $\mathbf{x}$ réparties dans les classes $\mathbf{z}$. En revanche la partie de la vraisemblance qui correspond aux paramètres Δ de la loi *a priori* spatiale est problématique ; elle requiert le calcul de la fonction de partition $W(\Delta)$.

Julian Besag [25] proposant de maximiser cette vraisemblance sur des sous-ensembles de sites indépendants entre eux. Il les nomme sous-sensembles de codage (*coding set* en VO) ; il exploite ainsi la nature locale des dépendances du champ de Markov. Chaque ensemble de codage est constitué par des sites non voisins entre eux selon la relation de voisinage ν. Ces ensembles isolent en quelques sorte les objets de l'un d'entre eux : si on connaît les sites de tous ces sous-ensembles sauf un, les variables Z_i de cet ensemble de codage sont mutuellement indépendantes. On montre alors que l'estimation des paramètres spatiaux Δ peut se faire en maximisant un produit de probabilités conditionnelles $P(z_i|\mathbf{z}_{\nu(i)}, \Delta)$ qui ont une forme analytique simple. L'inconvénient de cette technique est la perte d'information en ne se servant que de celle d'un ensemble de codage. Des auteurs suggèrent alors de faire autant d'estimations que d'ensembles de codage et de combiner ces estimations (voir par exemple [147]). Toutefois aucun fondement théorique ne justifie ces approches.

Pour ne pas être confronté au problème du choix ou de la combinaison des différentes estimations issues des différents ensembles de codage, on peut aussi utiliser la pseudo-vraisemblance déjà vue à la section précédente pour remplacer la vraisemblance. Il a été démontré que l'estimateur du maximum de pseudo-vraisemblance est asymptotiquement consistant. Les fonctions $P(z_i|\mathbf{z}_{\nu(i)}, \Delta)$ sont simples à manipuler. Des techniques de type descente de gradient sont utilisées pour approcher les paramètres optimaux : ceux qui maximisent $PL(\mathbf{z})$. D'autres approximations sont envisageables comme une méthode de gradient stochastique (décrite dans [61] pp 97–98).

Nous avons choisi d'associer une méthode de champ moyen à l'algorithme EM.

Principe de l'algorithme EM et structure de données incomplètes

Nous pouvons à présent discuter de l'approche globale que nous avons envisagée. Le contexte sera la structure de données incomplètes : les paramètres du modèle doivent être estimés à partir des données, une partie de ces dernières étant cachées.

L'**algorithme EM** (voir par exemple [65] pour la référence de base et [161] pour un exposé moderne) est un algorithme itératif dont le but est de maximiser la log-vraisemblance en procédant à chaque étape (q) à la maximisation de l'espérance de la log-vraisemblance complète connaissant les observations $\mathbf{x}$ et l'estimation courante (issue de l'étape précédente) des paramètres $\Psi^{(q)}$:

$$\begin{aligned} \Psi \mapsto Q(\Psi|\Psi^{(q)}) &= I\!E_{\Psi^{(q)}}\left[\log P(\mathbf{x},\mathbf{Z}|\Psi)|\mathbf{X}=\mathbf{x}\right] \\ &= \underbrace{\log P(\mathbf{x}|\Psi)}_{=\log L(\Psi)} + \underbrace{I\!E_{\Psi^{(q)}}\left[\log P(\mathbf{Z}|\mathbf{X}=\mathbf{x},\Psi)|\mathbf{X}=\mathbf{x}\right]}_{:=H(\Psi,\Psi^{(q)})} \end{aligned} \tag{3.7}$$

On répète les étapes d'itérations dont chacune peut être décrite comme :

- le calcul de l'espérance conditionnelle $Q(\Psi|\Psi^{(q)})$ et surtout le calcul des probabilités *a posteriori* $P(Z_i = c_k|\mathbf{X}=\mathbf{x},\Psi^{(q)})$ (**étape E**) et
- la mise à jour de $\Psi^{(q)}$ en $\Psi^{(q+1)} := \arg\max_\Psi Q(\Psi|\Psi^{(q)})$ (**étape M**).

On remplace ainsi les informations manquantes par leur espérance. Le calcul des paramètres Ψ n'est quasiment jamais faisable explicitement. Des méthodes de descente de gradient pourraient fournir de bonnes approximations mais des problèmes de convergence apparaissent vite selon la forme assez vite inadaptée de la vraisemblance avec la complexité du modèle [161] (nombre d'individus, de classes, forme du voisinage,...) ou pour un choix inadéquat des paramètres initiaux.

En effet, cet algorithme nécessite une initialisation des paramètres. Ce choix est encore à l'heure actuelle un casse-tête qui n'a probablement pas de réponse générale définitive ([159] pour le cas des modèles de mélange) selon la complexité des données ou de l'espace des paramètres par exemple : [28] propose une bonne stratégie dans le cas des mélanges gaussiens de manière à explorer au mieux les trajectoires possibles de l'algorithme. Nous adopterons une démarche pragmatique en observant la façon dont se comporte l'algorithme pour différentes initialisations aléatoires des paramètres. On se penchera en particulier sur les premières itérations ainsi que sur les valeurs finales obtenues après convergence. Nous expliciterons plus avant cet algorithme dans la section 3.6 suivante dans le cadre des approximations de type champ moyen que nous utiliserons. Un critère d'arrêt doit être décidé. On peut fixer le nombre d'itérations. Ou donner un seuil sous lequel les différences entre deux étapes soit des paramètres du modèle soit de la vraisemblance ne sont pas considérés comme significatifs.

Une bonne propriété de l'algorithme EM est que la fonction $q \mapsto L(\Psi^{(q)})$ est croissante. En effet : $\log L(\mathbf{\Psi}) = Q(\mathbf{\Psi},\mathbf{\Psi}^{(q)}) - H(\mathbf{\Psi},\mathbf{\Psi}^{(q)})$. Or la fonction $\mathbf{\Psi} \mapsto H(\mathbf{\Psi},\mathbf{\Psi}^{(q)})$ atteint son maximum en $\mathbf{\Psi}^{(q)}$ ce qui se montre par inégalité de Jensen (voir [65] ou [161]) :

$$\begin{aligned} H(\mathbf{\Psi}, \mathbf{\Psi}^{(q)}) &- H(\mathbf{\Psi}^{(q)}, \mathbf{\Psi}^{(q)}) \\ &= \mathbb{E}_{\mathbf{\Psi}^{(q)}} \left[\log \frac{P(Z|X, \mathbf{\Psi})}{P(Z|X, \mathbf{\Psi}^{(q)})} | X \right] \\ &\leq \log \mathbb{E}_{\mathbf{\Psi}^{(q)}} \left[\frac{P(Z|X, \mathbf{\Psi})}{P(Z|X, \mathbf{\Psi}^{(q)})} | X \right] \\ &= \log \int P(Z|X, \mathbf{\Psi}).dZ \\ &= \log 1 = 0 \end{aligned}$$

ou tout simplement en écrivant que :

$$H(\mathbf{\Psi}, \mathbf{\Psi}^{(q)}) - H(\mathbf{\Psi}^{(q)}, \mathbf{\Psi}^{(q)}) = \mathbb{E}_{\mathbf{\Psi}^{(q)}} \left[\sum_{z \in \mathcal{Z}} P(z|X, \mathbf{\Psi}^{(q)}) \log \frac{P(z|X, \mathbf{\Psi})}{P(z|X, \mathbf{\Psi}^{(q)})} \mid X \right]$$

et en rappelant que $\log(a) \leq a - 1, \forall\, a \in \mathbb{R}^{+*}$ par concavité de la fonction log. $\mathbf{\Psi}^{(q+1)}$ est choisi de façon a maximiser $\mathbf{\Psi} \mapsto Q(\mathbf{\Psi}, \mathbf{\Psi}^{(q)})$ d'où :

$$\begin{aligned} \log L(\mathbf{\Psi}^{(q+1)}) &= Q(\mathbf{\Psi}^{(q+1)}, \mathbf{\Psi}^{(q)}) - H(\mathbf{\Psi}^{(q+1)}, \mathbf{\Psi}^{(q)}) \\ &\geq Q(\mathbf{\Psi}^{(q)}, \mathbf{\Psi}^{(q)}) - H(\mathbf{\Psi}^{(q)}, \mathbf{\Psi}^{(q)}) = \log L(\mathbf{\Psi}^{(q)}) \end{aligned}$$

Maximiser la vraisemblance ou la fonction Q est donc équivalent. Si $\mathbf{\Psi}^{(q+1)}$ est choisi de façon a maximiser localement (donc en n'étant pas nécessairement le jeu de paramètres conduisant au maximum global), on parle de la famille des algorithmes EM généralisés (*Generalized EM*).

La convergence de l'algorithme EM repose sur le fait que la valeur maximisant localement la log-vraisemblance constitue un point fixe de la fonctionnelle de mise à jour des paramètres à chaque itération. Elle a été étudiée pour des classes de problèmes particuliers ; on pourra consulter [244] ou [245] où l'algorithme EM est interprété comme une descente de gradient sur la vraisemblance. Puis on y opère une projection et selon la forme de celle-ci, on explicite les propriétés qui conditionnent la convergence de l'algorithme.

Pour faire mieux qu'une convergence vers un maximum local, deux solutions sont possibles : recourir à des algorithmes probabilistes de type *Stochastic EM* ou *Simulated Annealing EM* ou bien travailler sur le choix des valeurs initiales des paramètres initiaux ([28]). Le problème d'un algorithme de recuit simulé est qu'on ne dispose plus que d'une convergence en probabilité : des expériences avec les mêmes entrées peuvent donner des résultats différents. Notre stratégie d'initialisation a été de lancer l'algorithme avec différentes classifications de départ tirées

aléatoirement et nous avons comparerons les vraisemblances (corrigées par une pénalisation de la complexité du modèle que l'on expliquera dans la partie 3.8) obtenues selon chaque cas pour garder le plus favorable. Une autre possibilité est d'observer l'évolution des paramètres lors des premières étapes de l'algorithme. Mais cela n'aide que pour des données simulées où les paramètres du modèle sont connus.

Et application aux champs de Markov cachés

Précisons l'écriture de Q dans le cas des champs de Markov caché :

$$Q(\Psi|\Psi^{(q)}) = \overbrace{\sum_{i\in S}\sum_{z_i} P(z_i|\mathbf{x},\Psi^{(q)})\log f(x_i|z_i,\theta_{z_i})}^{:=Q_\theta(\theta|\Psi^{(q)})} - \tag{3.8}$$

$$\underbrace{-\log W(\Delta) + \sum_c \sum_{\mathbf{z}_c} V_c(\mathbf{z}_c|\Delta) P(\mathbf{z}_c|\mathbf{x},\Psi^{(q)})}_{:=Q_\Delta(\Delta|\Psi^{(q)})}\,. \tag{3.9}$$

Le premier terme est indépendant de Δ tandis que les deux suivants ne dépendent pas de θ.

On peut alors procéder séparément pour estimer les valeurs des paramètres qui maximisent Q :

$$\theta^{(q+1)} = \arg\max_\theta Q_\theta(\theta|\Psi^{(q)}),$$

$$\Delta^{(q+1)} = \arg\max_\Delta Q_\Delta(\Delta|\Psi^{(q)}).$$

On est confronté à deux difficultés majeures pour évaluer Q : la fonction de partition $W(\Delta)$ et les probabilités conditionnelles $P_G(z_i|\mathbf{x},\Psi^{(q)})$ et $P_G(\mathbf{z}_c|\mathbf{x},\Psi^{(q)})$ ne peuvent être calculées exactement. Nous présentons donc maintenant l'approche que nous avons utilisée pour surmonter cet obstacle.

3.6 Approche EM et approximations de type champ moyen

Il reste encore une tâche importante à accomplir pour l'utilisation d'un modèle probabiliste tel que celui que nous avons jusqu'alors décrit : l'estimation des probabilités *a posteriori* connaissant les données observées et par la suite de certaines statistiques des variables relativement à cette distribution (comme la moyenne). Le modèle contient pour nous des paramètres déterministes. Nous n'avons pas considéré la généralisation dans un cadre bayésien où une distribution *a priori* est spécifiée pour les paramètres inconnus qui deviennent alors des

variables cachés ([30]). Cette généralisation s'intègre sans problème dans notre formalisme mais nous n'avons pas ressenti le besoin d'explorer cette piste.

De manière classique, l'algorithme EM permet l'évaluation de tels paramètres en calculant la moyenne de la vraisemblance complète des données (3.7). Dans de nombreux cas, cette vraisemblance est incalculable de façon exacte soit que la dimension de l'espace des variables cachées est trop grande soit que la distribution de ces variables cachées a une forme trop complexe. Des approximations stochastiques ou déterministes sont alors envisagées. Des approximations stochastiques comme les techniques MCMC (voir [90] pour la présentation de ce principe, [112, 178] pour des applications en biologie assisté par ordinateur ou autres [147]) permettent l'utilisation du cadre bayésien. Leur avantage est qu'asymptotiquement (si la capacité de calcul est infinie), elle conduise à des estimations exactes. Leurs inconvénients pratiques sont que les temps de calculs peuvent être longs mais surtout qu'il est difficile de vérifier que l'échantillonnage produit bien des réalisations indépendantes de la variable à simuler.

Nous présentons dans la suite une approximation déterministe analytique de la distribution *a posteriori* appelée champ moyen et considérons quelques généralisations présentées par exemple dans [47]. En pratique, nous utiliserons en particulier l'approximation en champ simulé. Elle est basée sur une classe particulière de méthodes dites *variationnelles*. Très rapidement il s'agira d'exprimer une quantité statistique d'intérêt comme une solution d'un problème d'optimisation puis de résoudre une version perturbée du problème. Les examples les plus répandus incluent des algorithmes de *belief propagation* ou *sum-product* ([30, 136, 253] présentent le problème de façon plus générale sous la théorie de la transmission d'un message) et la variété d'algorithmes dits en champ moyen.

Le principe d'une approximation en **champ moyen** correspond à une méthode qui est à l'origine dédiée au calcul de la moyenne d'un champ de Markov. Elle a été étendue pour l'estimation de la distribution du champ ([258]). Elle trouve sa base en mécanique statistique ([49]) où elle a par exemple été utilisée pour l'étude de phénomène de transition de phase dans des matériaux ferro-magnétiques. Plus récemment elle a été grandement employée dans des applications en vision par ordinateur ([147]), en recherche de solution de problèmes issus de théorie des graphes (*e.g.* bipartitionnement de graphes, problème du voyageur de commerce [101]) ou à des méthodes d'approximation particulaire de solutions d'équations aux dérivées partielles ([1]). Ces quelques exemples donnent une idée de la diversité des domaines d'applications du principe du champ moyen sans prétendre nullement être exhaustifs.

Principe du champ moyen

L'idée maîtresse est en considérant un site i de négliger les fluctuations des sites voisins $\nu(i)$ et de les fixer à leur valeur moyenne. Cette approximation est raisonnable pour un système en état d'équilibre. Les champs de Markov suivent

une distribution gibbsienne. Cela permet alors un calcul facilité des espérances et autres statistiques utiles. Le système qui résulte d'une telle approximation se comporte comme un système à variables indépendantes. Les distributions sont alors factorisables ; les calculs deviennent possibles à résoudre.

On fixe Z_j pour tous les $j \in \nu(i)$ à leur valeur moyenne $\mathbb{E}_G[Z_j]$. Notons alors que les variables Z_i ne seront plus à valeurs dans $\{0,1\}^K$ où $Z_{i,k} = 1$ si et seulement si $i \in c_k$ mais plutôt dans $[0,1]^K$ où $z_{i,k}$ sera assimilable à la probabilité d'appartenance de l'objet i à la classe c_k. Cette généralisation ne pose pas de problème ([243]). On notera aussi la probabilité marginale *a posteriori* :

$$t_{ik} = P(Z_i = c_k | \mathbf{x}, \Psi).$$

Une nouvelle fonction d'énergie peut alors être définie par à partir de l'équation (3.5) pour l'objet i :

$$H_i^{mf}(z_i) = H(\mathbf{z})|_{z_j = \mathbb{E}_G[Z_j], j \in S \setminus \{j\}}.$$

Notons que cette définition est locale en chacun des sites. Il lui correspond une nouvelle mesure de probabilité selon :

$$P_i^{mf}(\mathbf{z}) = \frac{\exp\left[-H_i^{mf}(z_i)\right]}{W_i^{mf}} \prod_{j \in S \setminus \{i\}} \mathbb{I}_{Z_j = \mathbb{E}_G[Z_j]}(z_j),$$

le produit exprimant simplement que la mesure se concentre sur l'hyperplan $\{Z_j = \mathbb{E}_G[Z_j], j \in S \setminus \{i\}\}$. La constante $W_i^{mf} = \sum_{z_i} \exp\left[-H_i^{mf}(z_i)\right]$ garantit ici encore que P_i^{mf} soit bien une mesure de probabilité.

Dans la décomposition de l'énergie du champ d'une mesure gibbsienne (3.5), les termes qui font intervenir z_i peuvent être isolés (selon que z_i appartient ou non à l'ensemble complet c dans le graphe G). Cela mène à une décomposition de l'énergie en champ moyen en un terme correspond à une énergie locale en i prenant en compte l'influence des sites «moyens» voisins, $H_i^{mf,loc}(z_i)$ et un terme indépendant de i, $H_i^{mf,ind}(\mathbb{E}_G[\mathbf{Z}_{S\setminus\{i\}}])$. On a la constante de normalisation correspondant à cette énergie locale moyenne : $W_i^{mf,loc} = \sum_{z_i} \exp\left[-H_i^{mf,loc}(z_i)\right]$.

La théorie en champ moyen suggère alors de remplacer la distribution marginale $P_G(z_i) = \frac{\sum_{\mathbf{Z}_{S\setminus\{i\}}} \exp\left[-H(\mathbf{Z})\right]}{W}$ par :

$$\begin{aligned} P_i^{mf}(z_i) &= \frac{\exp\left[-H_i^{mf}(z_i)\right]}{W_i^{mf}} \\ &= \frac{\exp\left[-H_i^{mf,loc}(z_i)\right]}{W_i^{mf,loc}} \end{aligned} \tag{3.10}$$

ce qui se trouve être aussi la probabilité de Z_i conditionnellement à $\mathbf{Z}_{\nu(i)} = I\!E_G[\mathbf{Z}_{\nu(i)}]$.

La distribution en champ moyen qui approxime la distribution jointe gibbsienne $P_G(\mathbf{z})$ est alors donnée par :

$$\begin{aligned} P^{mf}(\mathbf{z}) &= \prod_{i\in S} P_i^{mf}(z_i) \\ &= \frac{\exp\left[-\sum_{i\in S} H_i^{mf,\,loc}(z_i)\right]}{W^{mf}}, \end{aligned} \tag{3.11}$$

où $W^{mf} = \prod_{i\in S} W_i^{mf,\,loc}$.

La différence majeure avec la pseudo-vraisemblance de Besag est que les voisins ne sont plus autorisés à fluctuer. Ainsi fixés à des constantes (leur moyenne), les termes du produit sont bien indépendants et P^{mf} est bien une distribution de probabilité. Il reste à évaluer les moyennes $I\!E_G[Z_i]$. En effet, celles-ci sont inconnues. C'est d'ailleurs un des buts des calculs menés que d'y accéder.

Cohérence des moyennes calculées par le modèle et des moyennes définissant le champ moyen

Il y a en outre une condition de cohérence à vérifier par les valeurs approximées $\overline{z_i}$, $i \in S$: la moyenne calculée en se basant sur l'approximation doit être identique à celle utilisée pour cette approximation. De façon formelle, ceci s'écrit $\forall i \in S$:

$$\overline{z_i} = I\!E_i^{mf}[Z_i] = \frac{\sum_{z_i} z_i \exp\left[-H_i^{mf\,loc}(z_i)\right]}{W_i^{mf\,loc}} := g_i(\overline{\mathbf{z}_{\{i\}\cup\nu(i)}}),$$

puisque le terme $\frac{\sum_{z_i} z_i \exp\left[-H_i^{mf\,loc}(z_i)\right]}{W_i^{mf\,loc}}$ ne dépend que de $\{i\} \cup \nu(i)$. Il convient alors de résoudre cette équation de point fixe (N équations à N inconnues) pour trouver une estimation de $I\!E_G[\mathbf{Z}]$.

Si, par exemple, les fonctions g_i sont contractantes (*i.e.* la différence de deux valeurs prises par g_i en deux points différents est plus petite en norme que la norme de la différence de ces deux points multipliée par une constante < 1) on peut déterminer l'unique solution de façon itérative. Des conditions suffisantes sur la fonction d'énergie $\sum_c V_c(\mathbf{z}_c, \Psi)$ sont données dans [243].

Justification de l'approche en champ moyen

L'approche en champ moyen est justifiée par un **principe variationnel** qui trouve son origine en physique statistique [5]. La formulation de ce principe se fait

[5] Rien de statistique en fait on ne parlera pas d'échantillon ni d'estimation, le terme statistique date ici du XIX^{e} siècle où tout ce qui était aléatoire était qualifié de statistique. Aujour-

en demandant que la valeur d'une quantité typique du système, soit optimale pour la performance effectivement réalisée par le système par rapport à ce qu'elle vaudrait si on imaginait une performance différente. Des contraintes sur le système peuvent être considérées. L'application que nous en ferons est que la forme des distributions sur $\mathbf{Z}$ est trop complexe pour être exprimée explicitement. On construit donc une famille de distributions plus simple et on en cherche le meilleur représentant dans cette famille. Cela nécessite une façon de qualifier la proximité entre distributions ; on a pour cela recours au notions d'entropie ou divergence de Kullback-Leibler (voir plus bas). Plutôt qu'une forme première expliquant un phénomène par une loi locale (*e.g.* lois de Newton ou de Snell-Descartes, équation de Maxwell,...) on adopte une forme globale. Nous voulons utiliser les observations pour inférer une distribution de probabilités. Il y a (en général) plusieurs distributions qui sont compatibles avec les observations et les contraintes. Nous choisissons une méthode qui repose sur un principe d'optimalité pour le choix de la distribution : le maximum d'entropie. L'interprétation de ce principe est qu'on choisit la distribution la plus fidèle aux données tout en conservant un maximum d'incertitude sur l'état du système.

Pour cela on définit la fonction d'énergie libre F du système dont les variations permettent d'accéder à l'évolution d'un système évoluant à température et volume constant. Avec une distribution d'énergie H et une probabilité P toutes deux définies sur l'espaces des configurations possibles d'un système $\mathbf{S}$, on a :

$$F(P) = I\!E[H(\mathbf{Z})] - \mathbf{S}(P),$$

où $\mathbf{S}$ est l'**entropie** de la probabilité P : $\mathbf{S}(P) = -I\!E[\log P(\mathbf{Z})]$. En théorie de l'information, l'entropie est une mesure de la quantité d'information correspondant à la réalisation d'une variable aléatoire suivant une loi de probabilité. Le (second) principe hérité de la thermodynamique [6] impose qu'un système ne passe pas spontanément d'un état vers un autre nettement moins probable. Ainsi l'entropie d'un système isolé reste constante ou augmente. Le système est dit proche de son équilibre lorsque son entropie est proche de son maximum. Si on l'en écarte, se créent alors des mécanismes qui l'y ramènent.

Pourvu que l'on puisse résoudre le problème, la solution prendra alors une forme exponentielle qui est bien analogue à la forme de la distribution désirée (3.4) ([239]).

Si P_G est la distribution gibbsienne associée à H, on en déduit que $F(P) = -\log W + I\!E[\log P(\mathbf{Z})/P_G(\mathbf{Z})]$; on peut aussi voir le problème comme une minimisation de la quantité $I\!E[\log P(\mathbf{Z})/P_G(\mathbf{Z})]$ sur les distributions de probabilité

d'hui on parlerait surement plutôt de stochastique...

[6] Le premier principe stipule que la variation d'énergie interne d'un système est la somme de la chaleur (ou échanges thermiques) et du travail des forces qu'il reçoit de l'extérieur. En fait ce principe donne une contrainte sur les transformations possibles dans les états du système. La chaleur provient de mouvements «désordonnés» liés au changements d'état des composants du système.

P. On minimise une grandeur appelée la divergence [7] de Kullback-Leibler entre P et la vraie distribution de Gibbs P_G. Si $P = P_G$, l'énergie libre minimale vaudrait alors $F(P_G) = -\log W$. On montre aisément que l'approximation en champ moyen est optimale pour le critère de la divergence de Kullback-Leibler pour les systèmes à variables indépendantes ([181]).

En effet, les probabilités se factorisent alors selon :

$$P(\mathbf{z}) = \frac{\exp\left[-H(\mathbf{z})\right]}{W} = \prod_{i\in S} P(z_i) = \prod_{i\in S} \frac{\exp\left[-H_i(z_i)\right]}{W_i} = \frac{\exp\left[-\sum_{i\in S} H_i(z_i)\right]}{\prod_{i\in S} W_i}.$$

Les fonctions d'énergie sont alors de la forme : $H^{fac}(\mathbf{z}) = \sum_{i\in S} z_i^t[V_i + \delta V_i]$, où V_i est le potentiel sur le singleton $\{i\}$ pour l'énergie H. Notons que le terme δV_i contient les informations sur les valeurs prises par tous les voisins de i. Ces voisins sont dans contenus dans au moins un des ensembles complets c telles que $i \in c$.

Pour une telle probabilité gibbsienne associée à une telle fonction d'énergie :

$$F(P^{fac}) = -\log W^{fac} + I\!E_{P^{fac}}[H(\mathbf{Z}) - H^{fac}(\mathbf{z})],$$

En annulant la dérivée de cette énergie libre par rapport δV_i (voir le lemme 11.1 dans [239]), on trouve la condition (le calcul de la constante de normalisation est alors éludé) valable presque partout :

$$H(\mathbf{z}) - \log P^{fac}(\mathbf{z}) + cste = 0.$$

La distribution en champ moyen est bien celle qui minimise l'énergie libre si on garde la contrainte de probabilités qui se factorisent)et fournissent un système de variables indépendantes). On doit toujours satisfaire les conditions de cohérence comme en décrit dans la section 3.6. Rappelons que ceci n'est valable que dans le cas où on se restreint aux probabilités qui se factorisent selon $P^{fac}(\mathbf{z}) = \prod_{i\in S} P_i^{fac}(z_i)$. C'est dans ce cadre restreint que l'approximation en champ moyen est la meilleure au sens de la divergence de Kullback-Leibler.

Si on considère maintenant que la différence $\delta H := H - H^{fac}$ comme une perturbation de l'énergie de la probabilité gibbsienne P_G, dont on peut exprimer la fonction de partition :

$$W = \sum_{\mathbf{z}} \exp\left[-H^{fac}(\mathbf{z}) - \delta H(\mathbf{z})\right] = W^{fac} I\!E_{P^{fac}}\left[\exp\left[-\delta H(\mathbf{Z})\right]\right]. \qquad (3.12)$$

C'est à ce niveau qu'intervient la condition de stationnarité. On suppose que δH est faible comparé aux ordres de grandeurs des énergies considérées. On peut

[7] Notons l'inégalité triangulaire n'est pas vraie : on n'a pas affaire à une distance entre P et P_G. Cependant elle est positive, et nulle si et seulement si $P = P_G$. On pourra consulter [57].

alors procéder au développement de la fonction exponentielle au voisinage de 0 à l'ordre 1 dans l'expression 3.12 ci-avant. On obtient :

$$I\!E_{P^{fac}}\left[\exp\left[-\delta H(\mathbf{Z})\right]\right] \sim \exp\left(-I\!E_{P^{fac}}[\delta H(\mathbf{Z})]\right)$$

et comme $1 + x \leq \exp x, \forall x \in \mathbb{R}$:

$$W^{fac}(1 - I\!E_{P^{fac}}[\delta H(\mathbf{Z})]) \leq \underbrace{W^{fac} \exp\left\{-I\!E_{P^{fac}}[\delta H(\mathbf{Z})]\right\}}_{=\exp\left[-F(P^{fac})\right]} \lesssim W. \tag{3.13}$$

On parle de borne de Gibbs-Bogoliubov-Feynman pour le terme du centre de (3.13). Ce type d'inégalité est une reformulation du second principe de la thermodynamique. Elle est équivalente à une inégalité de Jensen [8].

Maximiser le membre du centre de l'équation (3.13) parmi les distributions ayant une fonction d'énergie de la forme $H^{fac}(\mathbf{z}) = \sum_{i \in S} z_i^t [V_i + \delta V_i]$ pour obtenir une meilleure approximation de la fonction de partition W est équivalent à minimiser l'énergie libre F sur les probabilités qui se factorisent. On a alors la meilleure approximation de W en utilisant P^{mf} par ce qui précède. Nous avons d'ailleurs validé numériquement et utilisé cette approximation dans nos expériences pour le calcul de la vraisemblance (voir les sections 3.8 et 5.1). On pourra aussi consulter [81] pour une comparaison de cette borne avec W^{mf}.

Généralisation : approximation de type champ moyen

L'approximation en champ moyen consiste à négliger les variations de l'environnement d'un objet autour de la valeur centrale qu'est la moyenne. Plus généralement on parlera d'**approximation de type champ moyen** quand la valeur d'un site i ne dépendra pas vraiment des valeurs de chacun des autres sites (ou du moins de ses voisins) mais de constantes (pas forcément la moyenne donc). Cette constante est fixée indépendamment du site i. Nous allons nous servir de cette idée pour le calcul fastidieux de la probabilité jointe $P_G(\mathbf{x}, \mathbf{z}|\Psi)$ dans une procédure de type EM. En particulier on considérera l'algorithme en champ modal et celui en champ simulé. Les constantes seront alors respectivement fixées au mode ou simulées selon la distribution $P_G(\mathbf{z}|\mathbf{x}, \Psi^{(q-1)})$. Pour le choix d'une approximation ou d'une autre, on regardera la section 3.7. Nous nous inspirons ici de l'exposé mené dans [47].

Pour un champ avec interaction jusqu'à l'ordre 2, la fonction d'énergie précisée à l'équation (3.5) est redéfinie à l'aide d'une approximation $\tilde{H}(\mathbf{z}) = \sum_{i \in S} \tilde{H}_i(z_i)$, somme d'énergies locales $H_i(z_i) = V_i(z_i) + \sum_{j \in \nu(i)} V_{ij}(z_i, \tilde{z}_j)$.

[8] Cette inégalité stipule que pour une fonction convexe sur $]a; b[$ et X une variable aléatoire d'espérance finie, à valeurs dans $]a; b[$, alors $f(I\!E[X]) \leq I\!E[f(X)]$ est vrai. Notons qu'elle sert aussi à montrer la positivité de la distance de Kullback-Leibler entre deux probabilités

Par exemple, le champ modal sera semblable au champ moyen à l'exception près que les voisins de l'objet considéré sont fixés à un de leur mode (s'il n'est pas unique). La condition de cohérence devient : $z_i^{mod} = \arg\max_{z'_i} P_G(z'_i|\mathbf{z}_{\nu(i)}^{mod})$ qui se résout de façon itérative.

L'autre solution [9] est l'approximation en champ simulé (*simulated field*). On simule les variables cachées selon la distribution conditionnelle avec un échantilloneur de Gibbs (on fixe la température $T = 1$, [90]). Comme le fait remarquer [181], l'intérêt majeur résidera dans la phase d'estimation des paramètres du champ par EM. En effet, on ne fixe plus ce champ à un valeur fixe (même justifiée) mais on opère une simulation. C'est très certainement cette étape stochastique, contrebalancée par un coût élevé en calcul puisqu'elle demande l'utilisation d'un échantillonneur de Gibbs, qui permet de meilleurs résultats de classification donc probablement une meilleure convergence de EM ([80]).

Décrivons ici la mise en œuvre de l'algorithme EM. On répètera les deux étapes suivantes :

1. Créer une configuration $\tilde{\mathbf{z}}^{(q)}$ à partir des observations $\mathbf{x}$ et de l'estimation courante des paramètres $\Psi^{(q-1)}$. En particulier on aura utilisé durant cette thèse les approximations en $(\forall i \in S)$:
   - . champ moyen : $\tilde{z}_i^{(q+1)} = I\!E_{\tilde{\mathbf{Z}}^{(q)}}[Z_i]$ dont nous avons vu qu'elle était optimale au sens de la divergence de Kullback-Leibler pour les distributions qui se factorisent ;
   - . champ modal : $\tilde{z}_i^{(q+1)} = \arg\max_{z_i} P_{\tilde{\mathbf{Z}}^{(q)}}(z_i|x_i, \Psi^{(q)})$, on a alors convergence de $\tilde{\mathbf{z}}^{(q)}$ vers un maximum local de la distribution gibbsienne. Rien n'assure la convergence vers un maximum global cependant ;
   - . champ simulé : $\tilde{z}_i^{(q)} \sim P_{\tilde{\mathbf{Z}}^{(q)}}(z_i|x_i, \Psi^{(q)})$. Là nous ne disposons pas de preuve théorique, uniquement de meilleurs comportements de l'algorithme sur les données simulées.

   Ainsi pour chaque site i on aura fixé l'état de ses voisins $\mathbf{z}_{\nu(i)}^{(q)}$ à $\tilde{\mathbf{z}}_{\nu(i)}^{(q)}$ et on pourra approcher la distribution marginale théorique $P_G(\mathbf{z}|\Delta)$ par :

$$P_{\tilde{\mathbf{Z}}^{(q)}}(\mathbf{z}|\Delta) = \prod_{i \in S} P_G(z_i|\tilde{\mathbf{z}}_{\nu(i)}^{(q)}, \Delta).$$

   On peut approximer de même la distribution *a posteriori* vu que l'on dispose de la forme des observations conditionnellement aux classes. Il semble plus justifié de l'utiliser parce que les observations sont un indice précieux de la réponse du système au régime latent $\mathbf{Z}$. Si cela est souhaitable, cela est aussi licite parce que nous avons pris la peine de préciser que $\mathbf{Z}|\mathbf{X}$ est aussi un champ de Markov. Cette recommandation d'utiliser le champ

[9] et qui est celle que nous avons retenu pour l'analyse des résultats sur données réelles parce qu'elle fournit les meilleures performances sur données simulées, voir la partie 5

conditionnel plutôt que le champ marginal a été étudiée sur des données simulées seulement. [31] a observé de sensibles améliorations dans le cas de l'emploi du champ conditionnel pour des données de type image.

Notons cependant que la mise à jour est séquentielle sur les sites de 1 à N. Ainsi quand on met à jour le site i à l'étape $(q+1)$, les sites $1, \ldots, i-1$ ont déjà été mis à jour et on utilise donc $\tilde{z}_1^{(q+1)}, \ldots, \tilde{z}_{i-1}^{(q+1)}$ tandis qu'on utilise $\tilde{z}_{i+1}^{(q)}, \ldots, \tilde{z}_N^{(q)}$ pour cette mise à jour (en n'utilisant bien sûr que les sites voisins de i).

2. Appliquer l'algorithme EM pour le modèle de champ de Markov caché défini par la formule ci-dessus et la loi d'observations des données $\mathbf{x}$ conditionnellement aux classes (gaussienne si on veut spécifier la forme que nous avons utilisée dans les expériences du chapitre 5). On part des valeurs $\Psi^{(q)}$ à mettre à jour en $\Psi^{(q+1)}$. La distribution jointe gibbsienne d'origine $P_G(\mathbf{x}, \mathbf{z}|\Psi)$ est remplacée par :

$$\prod_{i \in S} f_i(x_i|z_i, \theta_{z_i}) P_G(z_i|\tilde{\mathbf{z}}_{\nu(i)}^{(q)}, \Delta),$$

qui correspond à une vraisemblance des valeurs observées :

$$\begin{aligned} P_{\tilde{\mathbf{Z}}^{(q)}}(\mathbf{x}|\Psi) &= \sum_{\mathbf{Z}} f(\mathbf{x}|\mathbf{z}, \theta) P_{\tilde{\mathbf{Z}}^{(q)}}(\mathbf{z}|\Delta) \\ &= \prod_{i \in S} \sum_{z_i} f_i(x_i|z_i, \theta_{z_i}) P_G(z_i|\tilde{\mathbf{z}}_{\nu(i)}^{(q)}, \Delta) \\ &= \prod_{i \in S} P_G(x_i|\tilde{\mathbf{z}}_{\nu(i)}^{(q)}, \Psi). \end{aligned} \tag{3.14}$$

Grâce à l'étape 1. de récupération du voisinage, le calcul de la distribution marginale du champ caché est rendu possible. Le calcul de la distribution *a posteriori* est alors aisé puisque nous disposons d'une distribution *a priori* qui se factorise ; à partir des éléments précédents nous pouvons écrire :

$$\begin{aligned} P_{\tilde{\mathbf{Z}}^{(q)}}(\mathbf{z}|\mathbf{x}, \Psi^{(q)}) &= \frac{f(\mathbf{x}|\mathbf{z}, \theta^{(q)}) P_{\tilde{\mathbf{Z}}^{(q)}}(\mathbf{z}|\Delta^{(q)})}{P_{\tilde{\mathbf{Z}}}^{(q)}(\mathbf{x}|\Psi^{(q)})} \\ &= \prod_{i \in S} \left[\frac{f_i(x_i|z_i, \theta^{(q)}) P_G(z_i|\tilde{\mathbf{z}}_{\nu(i)}^{(q)}, \Delta^{(q)})}{\sum_{z_i'} f_i(x_i|z_i', \theta^{(q)}) P_G(z_i'|\tilde{\mathbf{z}}_{\nu(i)}^{(q)}, \Delta^{(q)})} \right] \\ &= \prod_{i \in S} P_G(z_i|x_i, \tilde{\mathbf{z}}_{\nu(i)}^{(q)}, \Psi^{(q)}) \\ &= \prod_{i \in S} P_{\tilde{\mathbf{Z}}^{(q)}}(z_i|x_i, \Psi^{(q)}). \end{aligned} \tag{3.15}$$

Ainsi on pourra facilement estimer $Q_\theta(\theta|\Psi^{(q)})$ et $Q_\Delta(\Delta|\Psi^{(q)})$.
L'étape d'estimation de l'algorithme EM est donc le calcul des

$$t_{ik}^{(q+1)} \approx \tilde{t}_{ik}^{(q+1)} := P_{\tilde{\mathbf{Z}}^{(q)}}(z_i|\mathbf{x}, \Psi^{(q)}).$$

De façon explicite dans le cadre des champs de Markov cachés :

$$\tilde{t}_{ik}^{(q+1)} = \frac{f(x_i|z_i, \theta_k^{(q)}) \exp\left[-\sum_{c,i\in c} V_c(\tilde{\mathbf{z}}^{(q|q+1)}_{c,\tilde{z}_i^{(q|q+1)}=c_k}|\Delta^{(q)})\right]}{\sum_{m=1}^{K} f(x_i|z_i, \theta_m^{(q)}) \exp\left[-\sum_{c,i\in c} V_c(\tilde{\mathbf{z}}^{(q|q+1)}_{c,\tilde{z}_i^{(q|q+1)}=c_m}|\Delta^{(q)})\right]}, \tag{3.16}$$

où la notation $\tilde{\mathbf{z}}^{(q|q+1)}_{c,\tilde{z}_i^{(q|q+1)}=c_k}$ signifie que l'on force la i^{e} composante à valoir c_k et que l'on prend les valeurs mises à jour avant i à leur estimation à l'étape $(q+1)$ et celle qui ne sont pas encore mise à jour à cette étape à leur valeur référée en (q).

La formule de la vraisemblance a la même forme que celle d'un mélange de variables observées indépendantes mais les proportions du mélange dépendent des valeurs constantes données aux voisins du site $\tilde{\mathbf{z}}_{\nu(i)}$. Ce jeu de données est donc différent à chaque itération. C'est une particularité de la prise en compte du réseau.

Il reste alors à mettre à jour les paramètres dans la phase de maximisation de l'algorithme EM selon les formules générales présentées dans l'équation (3.8) et qui seront spécifiées selon la paramétrisation exacte choisie (voir la partie suivante 3.7).

L'utilisation de la pseudo-vraisemblance suivi de simulations de Monte-Carlo est moins consistante parce qu'elle ne vérifie pas forcément la formule de Bayes (3.15). Il ne s'agit pas non plus forcément d'une distribution de probabilités valide.

Finalement, on dispose d'un algorithme qui approxime à chaque étape le champ de Markov caché à l'aide d'un jeu de variables manquantes qui sont supposées indépendantes. L'estimation des paramètres se fait sur cette structure simplifiée par un algorithme EM. Le choix des valeurs des voisins nécessite la résolution d'une équation de point fixe de cohérence ou une simulation du champ.

3.7 Influence des paramètres, du voisinage

Nous allons commencer par étudier la partie spatiale de la mise à jour des paramètres. En effet, elle permet le calcul des probabilités conditionnelles aux données, les t_{ik}, qui serviront pour l'estimation des paramètres de loi de chacun des groupes qui ne pose alors aucun problème. Le calcul des paramètres spatiaux est une particularité de notre modélisation qui apporte une difficulté supplémentaire. Nous avons vu plusieurs solutions et en avons retenu une. Explicitons là ici.

Ces deux parties forment l'étape de maximisation de l'algorithme EM de type champ moyen que nous utilisons.

3.7.1 Distribution a priori

Dans un modèle de champ de Markov caché, la classification non observée $\mathbf{z}$ suit par hypothèse une distribution de Gibbs $P(\mathbf{z}|\Psi)$ de la forme (3.4). On peut décomposer la fonction d'énergie selon les tailles croissantes des ensembles complets :

$$H(\mathbf{z},\Delta) = \sum_{i=1}^{N} V_i(z_i,\Delta) + \sum_{i\sim j} V_{ij}(z_i,z_j,\Delta) + \cdots + \sum_{\{i_1,\ldots i_q\}\in c_q} V_{i_1\cdots i_q}(z_{i_1},\ldots,z_{i_q},\Delta),$$

où la notation $z_i \sim z_j$ signifie que les deux objets i et j sont dans le même ensemble complet de cardinal 2 ; c_q est une clique de taille q qui est la plus grande taille de clique dans le graphe G.

Potentiels sur les singletons et potentiels sur les paires

Dans la plupart des tâches de classifications, on néglige les potentiels pour les cliques de taille supérieure à deux. Cette approximation est héritée de la physique statistique où les forces s'exercent par hypothèse entre deux objets (notamment le modèle de Potts pour le ferromagnétisme). On a donc une expression de la forme :

$$P(\mathbf{z}|\Delta) = \frac{\exp\left[\sum_{i=1}^{N}\alpha_{z_i}(i) + \sum_{i\sim j}\mathbb{B}_{z_i,z_j}(i,j)\right]}{W(\Delta)} \tag{3.17}$$

Les **potentiels sur les singletons** $V_i(z_i,\Delta) = -\alpha_{z_i}(i)$ sont proportionnels à la probabilité *a priori* d'occurrence des classes en chacun des sites en l'absence d'interactions de paires ; rappelons que les champs de Markov cachés sont équivalents au modèle de mélange quand les sites sont indépendants. En mécanique statistique, ils s'interprètent comme l'influence du champ externe. Si $\alpha(i) = 0$, aucune classe n'est «favorisée» au site i. Si $\alpha_{c_k}(i)$ est notablement plus élevé que les autres, (k variant) cela signifie que cette «direction c_k d'un champ externe $\alpha(i)$» est très influente sur «l'orientation z_i» de l'objet i. Inversement, une valeur petite (*i.e.* très négative) de $\alpha_{c_k}(i)$ impose une plus faible probabilité de l'évènement $z_i = c_k$. À k constant, en faisant varier i, on peut voir l'influence de la position dans le réseau sur l'action du champ externe.

Néanmoins on simplifie en général davantage l'expression 3.17 en supposant que ces fonctions sont les mêmes en tous les sites ; le champ externe ne dépend pas de son *locus* d'application. En mécanique statistique, cela signifie que le champ est stationnaire ou uniforme spatialement. La distribution *a priori* ne connaît pas

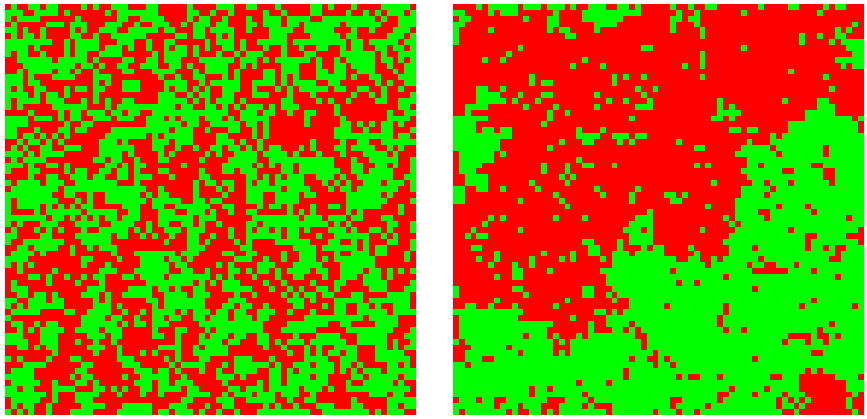
de biais spatial. Les paramètres sur les singletons se résument alors à un vecteur de taille $K : \alpha = (\alpha_{c_1}, \ldots, \alpha_{c_K})$. Dans ce cas on s'autorisera à noter α_k le potentiel associé à tous les objets dans la classe k. Pour résumer ce paragraphe, plus un α_{c_k} est élevé par rapport aux autres, plus il favorise la classe k.

Les **potentiels sur les paires** $V_{ij}(z_i, z_j, \Delta) = -\mathbb{B}_{z_i,z_j}(i,j)$ modélisent la dépendance entre sites voisins. En mécanique statistique, on prend en compte le potentiel de la force d'interaction entre particules voisines. On décompose souvent ce potentiel d'interaction en : $\mathbb{B}_{z_i,z_j}(i,j) = F(d_{ij})\, c(z_i, z_j)$. F est une fonction décroissante d'une certaine distance ou dissimilarité d_{ij} entre les individus i et j. Cette dissimilarité permet de prendre en compte une connaissance experte sur la «position» relative de deux individus i et j. La décroissance de F traduit le fait que l'interaction entre objets est d'autant plus faible que ceux-ci sont éloignés au sens du voisinage défini ν. $c(c_k, c_l)$ est une fonction de $K \times K \rightarrow \mathbb{R}$ qui est un coût algébrique (dont une perte si sa valeur est négative) qui caractérise l'attraction ou la répulsion des 2 classes c_k et c_l quand elles sont voisines. Plus le terme $c(c_k, c_l)$ est grand, plus la configuration est probable. Il peut donc s'interpréter comme un degré de compatibilité entre les classes c_k et c_l. Cette contrainte de forme pour les potentiels de paires n'est pas obligatoire. C'est une forme utile pour l'inclusion de connaissance experte *a priori* sur les interactions. Nous n'avons pas utilisé cette possibilité. Nous pensons qu'elle sera utile pour l'inclusion de degré de confiance sur les arêtes entre deux sites du graphe avec des données comme celles de la base STRING (voir le chapitre 2 sur les données ou celui 5.3 sur les expériences).

Si on fixe le terme F à la valeur 1, on présuppose que les sites voisins sont tous séparés identiquement. Les potentiels ne dépendent alors plus que des classes des sites voisins. Cette hypothèse est en général pertinente sur une image ; c'est le choix que nous adopterons quand aucune information préalable n'est disponible. On notera alors de façon simplifiée $\mathbb{B}_{kl}$ le potentiel d'interaction entre les classes c_k et c_l.

Alors les paramètres d'interactions sont précisés par une matrice $K \times K$. Les éléments de cette matrice peuvent être à leur tour soit fixés en entrée du modèle si une information préalable pertinente est disponible, soit estimés. Sans information supplémentaire, on peut encore simplifier la forme de l'équation (3.17) en prenant $c(c_k, c_l) = b\, \mathbb{I}_{c_k=c_l}$. C'est le modèle de Potts classique. On suppose qu'aucune interaction spécifique n'a lieu entre classes différentes et qu'entre membres d'une même classe, cette interaction est résumée par un unique paramètre b. Éventuellement, on peut avoir un tel paramètre par classe. La matrice d'interaction est alors diagonale. Plus un de ces termes est élevé, plus les voisinages dans une même classe correspondant à ce terme sont favorisés.

La figure 3.6 donne deux exemples de réalisations de modèles de Potts obtenus par simulation sur une image carrée de taille 100×100 à $K = 2$ classes. Pour $\beta = 0,4$, on voit que le caractère spatial commence à jouer un rôle important avec des régions plus homogènes que les petits aggrégats locaux quand $\beta = 0,2$.



$\beta = 0,2$ $\beta = 0,4$

Fig. 3.6: Exemple de simulation de modèle de Potts.

Une sorte de transition de phase est observée.

On peut donner une formule explicite pour la mise à jour des valeurs estimées pour la variable de classe semblable à celle de la section précédente (équation (3.16)) :

$$\tilde{t}_{ik}^{(q+1)} = \frac{f(x_i|z_i,\theta_k^{(q)}) \exp\left[\alpha_k^{(q)} + \sum_{j\sim i}\sum_{l=1}^{K} B_{kl}^{(q)} \tilde{z}_{lj}^{(q|q+1)}\right]}{\sum_{m=1}^{K} f(x_i|z_i,\theta_m^{(q)}) \exp\left[\alpha_m^{(q)} + \sum_{j\sim i}\sum_{l=1}^{K} B_{ml}^{(q)} \tilde{z}_{lj}^{(q|q+1)}\right]}. \tag{3.18}$$

où la notation $\tilde{z}_{lj}^{(q|q+1)}$ signifie que l'on prend $\tilde{z}_{lj}^{(q+1)}$ si $j < i$ (il a déjà été mis à jour) et $\tilde{z}_{lj}^{(q)}$ sinon.

Il faut cependant bien garder à l'esprit que les descriptions des effets des potentiels sur les singletons et sur les paires vont de concert. C'est l'équilibre entre ces potentiels qui donne le comportement du champ de Markov. Remarquons aussi qu'ils sont définis à une constante près. Ainsi les potentiels sur les paires ont pour effet de ralentir l'effet des potentiels sur les singletons sur les différents proportions des différentes classes. Autrement dit les interactions locales des paires régularisent (pour des valeurs positives du coût de voisins d'une même classe) l'effet du champ externe dicté par les potentiels sur les singletons. Dans le cas où on a des valeurs négatives du coût, on ne parlera pas de régularisation mais plutôt d'inversion locale entre voisins face aux effets du champ externe. Dans nos applications, les termes diagonaux évaluent la relative compatibilité entre les classes. Nous n'avons cependant pas d'échelle de comparaison des valeurs des paramètres qui dépendent de toutes les caractéristiques du modèle (N, D, nombre d'arêtes, K, *etc.*).

Les paramètres spatiaux se résument donc à :

$$\Delta = (\alpha, \mathbb{B}),$$

où $\alpha(i) \in \mathbb{R}^K, \forall i \in S$ et $\mathbb{B}(i,j) \in M_K(\mathbb{R}), \forall (i,j) \in S^2$ avec leurs effets relatifs et simplifications possibles discutés ci-avant (les matrices $\mathbb{B}$ sont symétriques dans le cas des graphes non-orientés).

Mise à jour des paramètres spatiaux

Désormais nous ne travaillerons que sur des modèles homogènes c'est à dire que les indices sur les individus dans les paramètres ci-dessus disparaissent. Ils reflètent des interactions sur les classes sans prendre en compte la position intrinsèque au sein du graphe. Il n'y a en effet aucune raison *a priori* pour particulariser des régions du réseau. On a :

$$\begin{aligned} Q_\Delta(\Delta|\Psi^{(q)}) &= \sum_{i\in S}\sum_{k=1}^{K} \tilde{t}_{ik}^{(q+1)}\alpha_k + && (3.19)\\ &\quad \sum_{i\in S}\sum_{j\in\nu(i)}\sum_{k=1}^{K}\sum_{m=1}^{K} \tilde{t}_{ik}^{(q+1)} B_{km}\tilde{z}_{jm}^{(q+1)} && (3.20)\\ &- \sum_{i\in S}\underbrace{\sum_{k=1}^{K}\tilde{t}_{ik}^{(q+1)}}_{=1}\log\sum_{l=1}^{K}\left[\exp\left(\alpha_l + \sum_{j\in\nu(i)}\sum_{m=1}^{K} B_{lm}\tilde{z}_{jm}^{(q+1)}\right)\right] && (3.21) \end{aligned}$$

Les $\tilde{t}_{ik}^{(q+1)}$ sont déterminés en amont de cette étape lors du choix des voisins. En différenciant cette expression, on obtient la condition suffisante $\forall k = 1, \ldots, K$ (voir les calculs à l'annexe B.3) :

$$\tilde{t}_{ik}^{(q+1)} - \frac{\exp\left(\alpha_k + \sum_{j\in\nu(i)}\sum_{l=1}^{K} B_{kl}\tilde{z}_{jl}^{(q+1)}\right)}{\sum_{m=1}^{K}\exp\left(\alpha_m + \sum_{j\in\nu(i)}\sum_{l=1}^{K} B_{ml}\tilde{z}_{jl}^{(q+1)}\right)} = 0, \; ou$$

$$\sum_{j\in\nu(i)} \tilde{z}_{jr}^{(q)} = 0, \text{ (l'effet des voisins se compense comme si le point était isolé)}$$

en approximant comme cela a été proposé P_G par $P_{\tilde{\mathbf{Z}}}$ où $\tilde{\mathbf{z}}$ est le champ constant (moyen, modal ou simulé).

Un technique de descente de gradient permet la mise à jour des paramètres spatiaux.

Nous avons testé le comportement de l'algorithme sur des données simulées de type image (grille régulière). La méthode montre de bonnes performances autant

en termes de classification que pour l'estimation des paramètres (spatiaux ou de loi de classe d'ailleurs) sur des exemples variés. La taille du voisinage a peu d'influence sur ces résultats même s'il semble préférable d'utiliser un voisinage à 8 voisins plutôt qu'à 4. Cependant ces exemples sont (i) sur grille régulière et (ii) la distribution des classes se fait en zones homogènes. Nous aimerions valider l'approche sur des graphes différents et notamment un peu plus réalistes par rapport à ce que nous savons des caractéristiques des réseaux biologiques aujourd'hui en partie révélés. En effet, les graphes correspondants ont des attributs qui diffèrent des graphes réguliers : forme, cœfficient de *clustering* ou topologique, distribution des degrés ou du chemin le plus court entre deux nœuds,...Ainsi il faudrait tester que le bon comportement de l'algorithme reste valable dans des situations variées plus proches de réseaux réalistes (graphes aléatoire de Erdös ou mieux graphe exponentiel [116] pour des interactions protéine-protéine, *small world* [165] pour la définition sociologique,...). Le lecteur désireux de se documenter plus avant sur le sujet de la zoologie des graphes pourra consulter [5, 18, 172] ainsi que pour les problématiques plus sécifiquement biologiques : [4, 116].

3.7.2 Loi des observations

Dans la décomposition de la loi jointe, $P(\mathbf{X}|\mathbf{Z},\theta)$ modélise le processus selon lequel les observations $\mathbf{x}$ sont supposées être générées à partir d'une classification $\mathbf{z}$ donnée. Pour nos expériences, nous nous sommes limités au cas des densités normales multidimensionelles.

Mise à jour des paramètres de la loi des données observées

Il s'agit de maximiser la partie correspondant aux observations de l'espérance complète :

$$Q_\theta(\theta|\Psi^{(q)}) = \sum_{i\in S}\sum_{z_i} P_G(z_i|\mathbf{x},\Psi^{(q)})\log f(x_i|z_i,\theta_{z_i}),$$

où (rappel) la distribution gibbsienne P_G est remplacée par une approximation de type champ moyen $P_{\tilde{\mathbf{Z}}}$. C'est l'étape M de l'algorithme qui mettra à jour les paramètres de la loi des observations selon :

$$\theta^{(q+1)} = \arg\max_\theta \sum_{i\in S}\sum_{z_i} P_{\tilde{\mathbf{Z}}^{(q+1)}}(z_i|\mathbf{x},\Psi^{(Q)})\log f(x_i|z_i,\theta_{z_i}).$$

Dans le cas des densités gaussiennes, on peut détailler les paramètres des classes : les moyennes,

$$\mu_k^{(q+1)} = \frac{1}{\sum_{i\in S} t_{ik}^{(q+1)}}\sum_{i=1}^{N} t_{ik}^{(q+1)} x_i$$

et matrices de covariances pour chacun des *clusters* $c_1, \ldots, c_K$:

$$\Sigma_k^{(q+1)} = \frac{1}{\sum_{i \in S} t_{ik}^{(q+1)}} \sum_{i=1}^{N} t_{ik}^{(q+1)} (x_i - \mu_k^{(q+1)})(x_i - \mu_k^{(q+1)})^t.$$

En effet, on peut écrire :

$$\theta^{(q+1)} = \arg\max_{\theta} \sum_{i=1}^{N} t_{ik}^{(q+1)} \left(-\frac{\log\det \Sigma_k}{2} - \frac{1}{2}(x_i - \mu_k)^t \Sigma_k^{-1} (x_i - \mu_k) \right),$$

et utiliser les formules matricielles (la dernière pour $M \in GL_K(\mathbb{R})$) en les appliquant à $M = \Sigma^{-1}$, la matrice de précision qui est bien symétrique définie positive (on peut vérifier ces calculs composante par composante) :

$$\frac{\partial (x-\mu)^t M (x-\mu)}{\partial \mu} = -(M + M^t)(x - \mu)$$

$$\frac{\partial (x-\mu)^t M (x-\mu)}{\partial M} = (x-\mu)(x-\mu)^t$$

$$\frac{\partial \log\det M}{\partial M} = (M^t)^{-1}.$$

Les applications concernées sont bien dérivables sur les domaines considérés car ce sont des applications linéaires pour les deux premières, le déterminant est une fonction polynômiale des coefficients de la matrice (de degré 1 en chacun des coefficients même) et la fonction log est bien dérivable sur $\mathbb{R}^{+*}$.

Les paramètres μ et Σ de chaque classe s'interprètent comme dans le cas du mélange : la contribution de la donnée x_i de chaque individu i est pondérée par le niveau d'appartenance à la classe c_k : t_{ik}.

3.8 Critère BIC de sélection de modèle

La sélection de modèle est un problème central dans la démarche statistique. Si le modèle est connu, on a un cadre rigoureux pour permettre des inférences pertinentes. Mais dans de nombreux cas, les connaissances *a priori* sur les données sont insuffisantes pour déterminer de façon univoque un modèle sur lequel se baser pour procéder à des estimations ultérieures. C'est pourquoi, des méthodes de **sélection de modèles** se sont rapidement développées : sélection de variables ([134]) ou bien choix du nombre de composantes d'un mélange ([83]), d'un ordre d'auto-régression ([188]) ou d'une chaîne de Markov ([72, 123]). Le choix parmi une liste de modèles du «meilleur» possible est une question difficile. Il est reconnu que les **tests d'hypothèses** dont la décision finale repose sur une P-valeur sont problématiques pour cette tâche ([186]). Par exemple, dès que la taille de

l'échantillon devient importante, l'hypothèse nulle (qui est celle d'un modèle plus simple inclus dans celui de l'hypothèse alternative) est trop facilement rejetée même dans des situations où les données ne semblent soutenir aucune anomalie notoire. Aussi l'interprétation n'est pas la même quand on doit faire un choix entre deux modèles alternatifs ou quand la liste de tous les modèles possibles est grande. Enfin différents modèles peuvent sembler pertinents en regard des données. Néanmoins, ils conduisent à des conclusions éventuellement très différentes en ce qui concerne certaines questions cruciales si on ignore l'incertitude quant aux modèles plausibles.

Une autre méthode de sélection de modèle est la **validation croisée**. Traditionnellement, il s'agit de supprimer un objet au hasard (*leave-one-out*) qui sera utilisé pour la validation tandis que les données restantes seront utilisées pour l'apprentissage des modèles en compétition. On répète cette opération N fois. Mais il peut paraître absurde d'évaluer le modèle avec une seule donnée notamment dans le cadre d'un réseau. Par ailleurs, comme le fait remarquer [72], la taille de nos données rendent cette démarche trop gourmande en temps. On préfèrera alors exclure M données simultanément pour la validation. Cela pose alors un problème d'estimation de paramètres ; on se prive volontairement de certaines données.

On peut par exemple enlever toutes les données d'une classe (d'une partition obtenue aléatoirement ou comme résultat d'un algorithme rapide de type *k-means*), apprendre le modèle sur les données restantes et le tester sur toutes les classes. La moyenne des vraisemblances des données des classes initiales conditionnellement aux paramètres appris sans la classe écartée est un bon critère. Il dépend de la partiton initiale. Comme il n'est pas envisageable de toutes les parcourir, on peut avoir recours à un algorithme de type Monte-Carlo pour répéter cette procédure avec plusieurs partitions. C'est le *Multifold Cross Validation* [259].

Une autre approche est d'apprendre le modèle sur une classe et de calculer la vraisemblance sur toutes les autres données (dans toutes les classes non mises à l'écart) conditionnellement aux paramètres appris sur une classe. Si on répète cette opération sur un nombre raisonnable de partitions, un critère (*Repeated Learning Testing* [46]) est obtenu en moyennant les vraisemblances obtenues comme décrit ci-avant.

Les deux critères vus ci-dessus sélectionnent bien les modèles pour leurs qualités prédictives. Mais ils posent des problèmes puisque les données ne sont plus indépendantes contrairement aux modèles de mélange pour lesquels ils ont été pensés. Une seconde difficulté réside dans l'apprentissage d'un modèle pour lequel des observations ont été (volontairement certes) supprimées. Nous verrons dans la section 4.1 que des adaptations de l'algorithme EM sont alors nécessaires. Nous n'avons pas vraiment réfléchi à une adaptabilité de ces méthodes pour la sélection de modèles qui nous intéressent.

Nous avons choisi d'utiliser le **critère BIC** (*Bayesian Information Criterion*

[207]). D'une part pour sa contruction basée sur des fondements théoriques solides dans le cadre bayésien de l'estimation de paramètres (voir l'annexe B.2.2) : BIC est asymptotiquement (quand la taille de l'échantillon tend vers l'infini) bayésien tout en permettant d'éviter la spécification de *prior* (en fait on considère tous les modèles équiprobables). D'autre part pour son bon comportement lors de nos simulations (voir la partie 5). Son expression est donnée par :

$$BIC(model) = \log P(\mathbf{X}|M) - \frac{\kappa}{2} \log N',$$

où $\log P(\mathbf{X}|M)$ est la log-vraisemblance des données, κ est le nombre de paramètres libres du modèle M (ce nombre est inférieur au nombre de paramètres total du modèle ; des contraintes sur ces paramètres existent comme la somme des probabilités des composantes du mélange qui vaut 1) et N' est la taille des données (attention elle ne vaut pas forcément N le nombre d'objets considérés ; elle vaudra par exemple $N \times D$ pour une matrice de données d'expression telle que nous l'avons présentée).

On pourra aussi se référer à [143] pour une comparaison avec le critère AIC. Le **critère AIC** est en fait une approximation de la divergence de Kullback-Leibler entre la vraie densité responsable de la production des données et les éléments d'une famille de probabilités qui ne contient pas la vraie densité. On voudra retenir le meilleur élément de cette famille c'est à dire le plus proche de la vraie densité du point de vue de la divergence de Kullbac. AIC est donc un critère de la théorie de l'information. [159] note qu'AIC surestime davantage le nombre d e paramètres que BIC. [27] fait remarquer que BIC aussi sélectionne un nombre de classes souvent trop élevé parce qu'il ne prend pas en compte l'objectif de classification. Une solution est de remplacer la vraisemblance par la vraisemblance complète (ou classifiante). BIC semble en revanche bien adapté pour l'estimation de densités : choix du graphe d'interactions (en particulier choix du modèle indépendant ou d'un modèle markovien) choix de la famille de distributions sur les classes ou choix du type de modèle à l'intérieur de la famille gaussienne ([72]). La restauration des états cachés permet en outre de détecter des zones à homogénéité particulière dans le réseau. Notre but est plus de décrire la population étudiée dans une exploration des données (*data minig*). Nous ne cherchons pas à détecter un nombre exact de *clusters* dans les données. Nos prévisions ne souffriront pas d'un nombre de classes déterminé à 6 ou à 7. BIC qui a la propriété d'être consistant est alors un meilleur choix. Aussi [186] a utilisé BIC dans le cadre des graphes non orientés et [81] a donnée une version plus satisfaisante de BIC pour en éviter le calcul en adoptant l'hypothèse d'indépendance des données.

Nous n'avons ici parlé que de la sélection du modèle retenu parmi la liste de ceux testés et plus particluièrement du choix du nombre de classes à retenir. Le problème de la qualité des *clusters* eux-mêmes sera abordé dans la partie 5.3. Nous serons surtout intéressés par la validation biologique des groupes produits par notre algorithme. Pour le problème général de la validation de classification,

on peut se référer à [95] et [44]. Le cas particlier des données post-génomiques est traité dans [97] ou [70].

Nous avons présenté les aspects théoriques du modèle de base qui correspond à notre approche. Avant de passer aux applications à proprement parler, nous étudierons des généralisations qu'il nous a paru nécessaire d'envisager pour mieux faire face à la réalité des données post-génomiques.

4. EXTENSIONS DU MODÈLE

4.1 Observations manquantes ou partielles

Présentation générale

Lorsqu'on est confronté à un problème réel, on doit souvent faire face à des tableaux de données dont certaines valeurs viennent à manquer. Les causes de ces absences sont très variées. Le vide peut être accidentel en cas de défaillance d'un appareillage de mesure. Ou des champs sont «volontairement» renseignées (par les sondés d'un questionnaire par exemple) et d'autres non. Dans d'autres cas, des mesures ne seront pas effectuées volontairement. Elles pourront être jugées peu revélatrices ou trop coûteuses par l'expérimentateur au vu et au su de résultats antérieurs suffisants. Des méthodes adaptées à cette absence de données sont nécessaires sous peine de conduire à des conclusions erronées. De bonnes références générales statistiques sur le sujet sont les ouvrages [197] pour une présentation historique claire du problème et [148] plus récent et traitant le sujet avec moult détails techniques.

Nous supposerons en tout cas qu'il existe des conditions d'observations idéales qui auraient pu permettre l'observation de toutes les valeurs mesurables. Bien entendu nous insistons sur le caractère théorique de ces conditions. Des perturbations engendrent les absences d'observations. Elles peuvent être de nature très différentes selon le champ d'application envisagé : reconnaissance de la parole ([118,190]), analyse de mouvement ([232]), évolution et phylogénie ([242]), météorologie ([111]), nutrition ([55]), écologie ([133,174]), sociologie ([205,213]),... Nous nous pencherons sur le cas des données post-génomiques.

Puces ADN, observations manquantes et remèdes classiques

Dans le cadre pratique des données d'expression de gènes issues de puces ADN, plusieurs raisons conduisent les expériences à ne pas fournir des données complètes ([160]). Par exemple la résolution lors de l'analyse d'image peut être insuffisante, l'image à analyser corrompue ou un problème peut survenir dans la phase d'assignation d'un gène à une mesure de fluoresence. Plus en amont, on peut avoir une poussière ou une éraflure sur la lame elle-même ([34, 231]). Les valeurs absentes peuvent aussi résulter d'une erreur systématique du robot lors du dépot des sondes. En fait on n'arrive pas à dégager les causes et les effets précisément tant le processus expérimental est complexe. [175] rappelle que

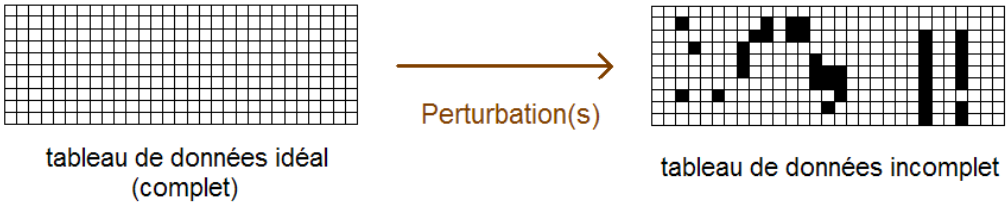


Fig. 4.1: Apparition de valeurs manquantes dans le cas classique de données «rectangulaires» : les lignes représentent les individus observés tandis que les colonnes représentent les différentes variables mesurées.

les données manquantes affectent typiquement 90% des gènes présents dans un expérience de puce ADN.

Des stratégies primaires sont employées pour combler ces vides le plus souvent. Les positions à trous sont repérées manuellement et signalées. On peut les exclure des analyses ultérieures, considérer leur contribution comme neutre : [6] supprime les mesures (*case deletion*) incomplètes pour 80% ou plus des dimensions et remplace les valeurs manquantes par des 0. Une technique à peine plus évoluée, l'imputation par la moyenne est très répandue. Une telle approche est loin d'être optimale parce qu'elle ne prend pas en compte la structure de corrélation sur les données. [175] évalue cette approche en supprimant artificiellement des valeurs dans une matrice de données de gènes et estimant l'erreur commise (par erreur quadratique ou en comptant les différences de partionnement obtenu par une méthode de *clustering*). Souvent même les auteurs ne précisent pas la méthode employée pour compléter les données (ou les supprimer). Pourtant l'impact en matière de classification est reconnu important dans la communauté qui se penche sur le traitement des données manquantes !

Une solution coûteuse et rarement employée pour cause de fonds limités est la répétition d'expériences. Elle pose aussi le problème de l'intégration de mesures répétées. Des modélisations à base de *splines* permettent des ajustements des valeurs manquantes pour les séries temporelles ([17]).

Des méthodes un peu plus sophistiquées sont basées sur une estimation sur les k plus proches voisins ou sur une décomposition en valeurs singulières des profils d'expression ([231]). Plutôt qu'un voisinage basé sur des corrélations, [129] préfère une distance des moindres mais le principe est le même.

Les k plus proches voisins cherchent à compléter la valeur manquante pour une expérience en faisant la moyenne sur des profils où la valeur de cette expérience est présente. Les profils utilisés sont les k plus proches au sens d'une certaine distance. Pragmatique, cette méthode n'a pas vraiment de fondement statistique en ne considérant que les corrélations positives ([210]).

Pour la décomposition en valeurs singulières, toutes les valeurs manquantes sont remplies par moyenne sur la ligne dans une phase préliminaire. Le processus produit alors une suite de profils d'expression orthogonaux dont la combinaison linéaire permet d'approcher la matrice complétée des expressions. Une régression est faite pour les gènes dont au moins une des données manque avec les k gènes les plus significatifs dans la décomposition spectrale de la matrice complétée des expressions (les vecteurs-gènes propres). Les valeurs manquantes sont alors remplacées par la valeur prédite par la régression.

Pour ces deux approches cependant, les auteurs mettent en garde contre une utilisation non raisonnée de leur méthode. Bien qu'elles donnent des résultats d'une plus grande précision que le remplissage par moyenne sur la ligne et soit d'une plus grande robustesse face au pourcentage de données manquantes, les conclusions biologiques doivent être considérées avec précaution. Il faut ici bien garder à l'esprit que les données complétées ne garantissent pas vraiment la réalité du processus biologique.

Nous n'avons pas une connaissance très précise des causes des manques dans nos données. Toutefois, nous en tiendrons compte avec rigueur. Nous essaierons de nous affranchir d'approches heuristiques pour mettre en place des hypothèses probabilistes justifiant le choix de nos traitement.

La technique la plus souvent utilisée est la suppression des individus pour lesquels les observations ne sont pas complètes. On parlera de *case deletion* dans la littérature. Elle se justifie en premier lieu par sa simplicité et le fait que les méthodes pour des données complètes sont applicables. En pratique elle donne de bons résultats quand les données manquent sur un nombre peu élevé d'individus (moins de 5% dans [202]). Aussi, contrairement à l'imputation par la moyenne, cette méthode n'introduit pas de biais dans le tableau de données dans le cas où le mécanisme d'absence est totalement indépendant des données (voir ci-dessous la définition de MCAR).

Cependant, il est courant qu'une proportion notable d'observations soit manquante. Par exemple, si nous agissions de la sorte sur les quelques 4883 gènes qui composent nos données sur la levure (issues de données d'expression de [217] et de données d'interaction de protéines DIP `http://dip.doe-mbi.ucla.edu/`), nous n'aurions par exemple que 373 gènes dont les dimensions de données d'expression sont toutes renseignées (soit $7,6\%$). Ainsi non seulement bon nombre d'individus ne seraient pas classés mais encore les dimensions présentes pour ces individus ne seraient plus prises en compte : ôter les individus incomplets appauvrit grandement l'information disponible. En effet, dans notre exemple, il est fréquent qu'une seule (ou moins d'une dizaine) dimension sur la petite centaine étudiée soit non renseignée.

Il est aussi enviseageable d'ignorer les informations manquantes. On ne calculera les différentes statistiques qu'à l'aide des seules valeurs observées. Même si cette démarche n'est pas tout à fait satisfaisante, elle repose sur un modèle

probabiliste et un critère défini dans la démarche d'estimation du modèle.

Une autre technique classique consiste à remplacer une donnée manquante par une valeur «raisonnable». La littérature parle de *single imputation*. On a présenté ci-dessus l'imputation par des zéros (qui reflète une activité neutre dans les expressions de gènes passées en log) ou par la moyenne. Un raffinement possible est de conditionner par la classe quand on fait du *clustering*. La notion d'imputation par les valeurs des voisins est celle de l'algorithme *KNNimpute* ([231]). On peut imaginer d'autres voisinages que celui de données d'expression similaires. [34] propose une imputation par méthode des moindres carrés (choix de gènes utilisés pour faire une régression en fait en se basant sur d'autres matrices d'expression où les données ne manquent pas). On travaille ensuite avec le tableau complété comme si les valeurs artificielles étaient effectivement observées. Cette démarche est très courante dans l'analyse de puces à ADN. Des petits programmes complètent automatiquement les trous sans que le manipulateur s'inquiète vraiment de la méthode (quand il est conscient que des données manquent !). Cette démarche a de nombreux inconvénients évidents. La moyenne empirique est certes conservée mais la variance se trouve biaisée (diminuée). Aussi on s'interdit de prendre en compte une éventuelle valeur exceptionnelle révélatrice d'un phénomène. Enfin et surtout, rien ne différencie dans le traitement final les valeurs réellement observées de celles inférées. C'est surtout ce point que nous désirons corriger.

Pour pallier l'oubli de la prise en compte de la variabilité des données, une technique de *multiple imputation* ([7]) a été proposée. Les détracteurs de cette technique avancent l'argument que des données sont artificiellement crées par une sorte d'alchimie statistique. [203] précise qu'au contraire on estime l'incertitude.

[204] note que toute anomalie entre modèle d'imputation et d'analyse a pour conséquence une mauvaise estimation des paramètres. Dans ce cadre, les techniques d'imputations bayésiennes multiples ont l'avantage sur la maximisation de la vraisemblance que leurs phases d'imputation et d'analyse sont clairement séparées. Leurs effets respectifs sont donc plus facile à analyser. [226] a utilisé cette technique pour une étude longitudinale de patients atteints de métastases (il y a donc des données censurées : le nombre de patients abandonnant le traitement ou décédant étant important).

Parmi les autres méthodes classiques de traitement de données manquantes, citons : la sélection de variables (dans le cadre de réseaux bayésiens [187]), les méthode de régression ([108] pour une revue sur les aspects théoriques et [262] pour une application en bioinformatique), *etc.* [85] propose d'intégrer des connaissances biologiques dans l'imputation en considérant des ensembles convexes décrivant des annotations.

[135] préfère une solution identifiant les données supplémentaires à effectuer pour apprendre correctement le modèle (en traitant sur le même plan observations manquantes et étiquettes). Mais il se place dans le cadre de l'apprentissage semi-

supervisé.

[127] est une première méthode qui souligne les effets potentiellement néfastes d'imputation préalable pour des analyses ultérieures. Les auteurs proposent donc une correction lors des étapes de l'algorithme de classification. Les données estimées sont progressivement prises en compte avec un poids croissant. Cependant la méthode est déterministe et ne permet pas le calcul de probabilités de confiance pour la classification en utilisant un algorithme de type *k-means*. Vue la diversité des méthodes proposées mais jamais dans un cadre statistique rigoureux pour la prise en compte de ces valeurs absentes, il nous a paru nécessaire de tirer profit des possibilités d'un algorithme de type EM pour traiter la problématique des observations manquantes dans les données de puces ADN. Il reste à préciser les lois des variables sous-jacentes.

Notre proposition : traiter les observations manquantes dans le cadre d'un algorithme EM

Nous préfèrerons voir les valeurs absentes comme des données cachées. On notera Obs_i ou Man_i les indices respectifs dont les données individuelles sont observées ou manquantes pour l'individu i. Remarquons simplement que (Obs_i, Man_i) est une partition de $\{1, \ldots, D\}$. On notera alors $\mathbf{x}^{Obs} := \{x_i^{Obs_i}, i \in S\}$ avec $x_i^{Obs_i} := \{x_{id}, d \in Obs\}$ et $\mathbf{X}^{Man} := \{X_i^{Man_i}, i \in S\}$.

On peut toujours écrire la log-vraisemblance des données complètes pour séparer contribution des classes et contribution spatiale :

$$\begin{aligned} \log L_c(\mathbf{\Psi}) &= \log P(\mathbf{X}, \mathbf{Z}|\mathbf{\Psi}) && (4.1) \\ &= \log P(\mathbf{X}|\mathbf{Z}, \mathbf{\Theta}) + \log P(\mathbf{Z}|\mathbf{\Delta}) && (4.2) \end{aligned}$$

où $\log P(\mathbf{X}|\mathbf{Z}, \mathbf{\Psi})$ est la distribution conditionnelle des observations (gaussienne multivariée) et l'autre composante est le champ de Markov caché défini selon l'équation (3.17).

Il faudra préciser le modèle intégrant le mécanisme d'absence de données R. C'est une matrice de taille $N \times d$ telle que $R_{ij} = 1$ si l'individu i a sa variable j observée et 0 sinon. On supposera que toute variable j est observée pour au moins un individu (*i.e.* $\forall j, \sum_{i=1}^{N} R_{ij} > 0$) et que chaque individu i a au moins une de ses composantes observée (*i.e.* $\forall i, \sum_{j=1}^{d} R_{ij} > 0$). Sinon notre algorithme fonctionnera tout de même en se servant uniquement des informations de voisinage.

L'ensemble des paramétres du mécanisme d'absence est alors noté ρ. On distingue ([197]) :

(a) l'absence complètement aléatoire qui est indépendante de toutes les données : $P(R|\mathbf{X}, \mathbf{Z}, \rho) = P(R|\rho)$ (MCAR pour *Missing Completely At Random*). Ce cas dit juste qu'il y a une distribution du mécanisme d'absence des données mais que celle-ci n'est en aucun cas reliée aux variables étudiées.

(b) l'absence aléatoire qui ne dépend que des données observées : $R \perp \{\mathbf{Z}, \mathbf{X}^{Man}\}$ ou $P(R|\mathbf{X}, \mathbf{Z}, \rho) = P(R|\mathbf{x}^{Obs}, \rho)$ (MAR pour *Missing At Random*). Cette dénomination peut paraître contradictoire puisque le mécanisme d'absence dépend des observations, pas des valeurs non observées. Elle correspond par exemple au cas où des individus seront moins enclins à répondre à des questions d'une enquête (qui seront alors les observations manquantes) s'ils sont d'une certaine catégorie sociale, d'un sexe ou d'un autre ou ont eu un certain niveau d'éducation (qui sont les données observées). Pour un exemple un peu plus technique, songez à un expérimentateur qui prévoit de faire trois mesures d'une même grandeur sur plusieurs sujets. Il pourra décider (pour des raisons de coût ou de temps) de ne pas faire la troisième mesure si la première et la deuxième sont suffisamment proches.

(c) l'absence non aléatoire qui dépend également des données manquantes : $P(R|\mathbf{X}, \mathbf{Z}, \rho) = P(R|\mathbf{x}^{Obs}, \mathbf{X}^{Man}, \rho)$ (NMAR pour *Non Missing At Random*). L'absence peut être le fait de la valeur manquante ($P(R|\mathbf{X}, \mathbf{Z}, \rho) = P(R|\mathbf{X}^{Man}, \rho)$) on parle de données censurées ; par exemple l'observation peut être manquante si elle est trop grande car l'appareil de mesure ne permet plus d'accéder à sa valeur. L'absence non aléatoire peut aussi dépendre de la classe (non observée) de l'objet : $P(R|\mathbf{X}, \mathbf{Z}, \rho) = P(R|\mathbf{Z}, \rho)$. C'est le cas si un certain type d'individus est difficile à observer par exemple. Pour reprendre l'exemple citer dans le cas MAR, manquera-t-on de données sur le poids des individus parce que des individus d'un certain sexe est moins enclin à répondre (cas MAR) ou parce que les individus ayant une surcharge pondérale renseignent l'information plus rarement (cas NMAR) ?

Le modèle NMAR est le plus général et contient le modèle d'absence MAR qui lui même englobe le modèle MCAR.

Dans les deux premiers cas, on saura faire beaucoup de choses grâce à de nombreuses factorisations des probabilités. On pourra alors moyennant quelques transformations appliquer les mêmes outils que ceux vus à la section 3.5. Dans le dernier cas, nous serions confrontés à des difficultés plus difficilement surmontables et il faudrait faire des hypothèses ou disposer d'informations supplémentaires sur le mécanisme d'absence R.

En effet, l'hypothèse d'absence aléatoire MAR permet (pour obtenir les égalités ci-dessous, il faut écrire la sommation sur les variables cachés qui ne sont pas présentes $\mathbf{X}^{Man}$ et/ou $\mathbf{Z}$ pour appliquer les hypothèses MAR ; nous ecrirons explicitement la troisième) :

- la factorisation de la vraisemblance en $\mathbf{x}^{Obs}$ et R :

$$P(\mathbf{x}^{Obs}, R|\mathbf{\Psi}, \rho) = P(R|\mathbf{x}^{Obs}, \rho)\, P(\mathbf{x}^{Obs}|\mathbf{\Psi}).$$

  La séparation des deux jeux de paramètres du modèle des données et du mécanisme d'absence est importante. La maximisation de la vraisemblance

des données obervées (voir la notationL_{Obs} ci-après) permet d'obtenir les paramètres du modèle indépendamment de ceux du mécanisme d'absence (qui ne nous intéresse pas dans ce mémoire).

- la factorisation de la vraisemblance dite «classifiante» :

$$P(\mathbf{x}^{Obs}, R, \mathbf{Z}|\mathbf{\Psi}, \rho) = P(R|\mathbf{x}^{Obs}, \rho)\, P(\mathbf{x}^{Obs}, \mathbf{Z}|\mathbf{\Psi}).$$

Cette égalité permet de faire abstraction de l'hypothèse d'absence aléatoire sous l'hypothèse MAR si son mécanisme ne nous intéresse pas. On pourra se limiter à rechercher la classification $\mathbf{Z}$ et les paramètres $\mathbf{\Psi}$ qui rendent *optimum* la vraisemblance classifiante ci-dessus.

- d'établir l'indépendance de la distribution *a posteriori* des classes du mécanisme d'absence :

$$\begin{aligned} P(\mathbf{Z}|\mathbf{x}^{Obs}, R, \mathbf{\Psi}, \rho) &= \frac{P(\mathbf{Z}, \mathbf{x}^{Obs}, R, \mathbf{\Psi}, \rho)}{\sum \mathbf{Z}' P(\mathbf{Z}', \mathbf{x}^{Obs}, R, \mathbf{\Psi}, \rho)} \\ &= \frac{\int P(\mathbf{Z}, \mathbf{x}^{Obs}, \mathbf{X}^{Man}, R, \mathbf{\Psi}, \rho) d\mathbf{X}^{man}}{\sum \mathbf{Z}' \int P(\mathbf{Z}', \mathbf{x}^{Obs}, \mathbf{X}^{Man}, R, \mathbf{\Psi}, \rho) d\mathbf{X}^{Man}} \\ &= \frac{P(R|\mathbf{x}^{Obs}, \rho) \int P(\mathbf{Z}, \mathbf{x}^{Obs}, \mathbf{X}^{Man}, \mathbf{\Psi}) d\mathbf{X}^{man}}{P(R|\mathbf{x}^{Obs}, \rho) \sum \mathbf{Z}' \int P(\mathbf{Z}', \mathbf{x}^{Obs}, \mathbf{X}^{Man}, R, \mathbf{\Psi}, \rho) d\mathbf{X}^{Man}} \\ &= \frac{P(\mathbf{Z}, \mathbf{x}^{Obs}, \Psi)}{P(\mathbf{x}^{Obs}, \Psi)} = P(\mathbf{Z}|\mathbf{x}^{Obs}, \mathbf{\Psi}). \end{aligned}$$

On peut donc assigner les objets aux classes par le principe du MAP par exemple sans se soucier de la structure d'absence de certaines données, toujours sous l'hypothèse MAR.

On se référera utilement à [205]. Notamment pour l'applicabilité de l'hypothèse MAR qui est le meilleur compromis entre faisabilité des calculs et réalisme. Ainsi lorsque l'expérimentateur omet volontairement certaines mesures de variables après les avoir faites sur certains individus, l'hypothèse peut être considérée valide. Aussi cette méthode donne des résultats satisfaisants dans certains cas où l'absence est due aux valeurs manquantes. L'argument peut être essentiellement présenté ainsi : si les valeurs observées $\mathbf{x}^{Obs}$ sont assez informatives pour prédire celles qui manquent $\mathbf{X}^{Man}$ et $\mathbf{Z}$, alors très certainement, connaissant $\mathbf{x}^{Obs}$, on doit avoir une faible dépendance entre R et $\mathbf{X}^{Man}$ et $\mathbf{Z}$. Donc l'équation du point (b) est presque vérifiée et l'hypothèse MAR est raisonnable. [100] donne un aperçu des difficultés rencontrées lorsque un mauvais choix d'hypothèse est fait entre les modèles MAR ou MCAR. [208] précise des conditions aussi peu restrictives que possibles pour ignorer le mécanisme d'absence des données dans un modèle décomposable et pour mettre à jour la vraisemblance marginale du modèle. [206] évalue différentes méthodes d'imputation sur des données d'expression de gènes selon le mécanisme d'absence des données ; selon les auteurs, 5%

de données manquantes NMAR est équivalent à une proportion entre 10 et 30% de données MAR. Ils s'intéressent aussi à l'influence de la méthode d'estimation pour la détection de gènes différentiellement exprimés.

Dans toute la suite de ce mémoire, nous nous placerons dans l'hypothèse que le mécanisme qui a généré l'absence des données est MAR.

La log-vraisemblance des données observées change d'expression par rapport au cas où les données $\mathbf{x}$ sont complètes. Elle devient :

$$L_{Obs}(\mathbf{\Psi}) := \log P(\mathbf{x}^{Obs}|\mathbf{\Psi})$$

.

Si nous formons le rapport :

$$k(\mathbf{X}, \mathbf{Z}|\mathbf{x}^{Obs}, \mathbf{\Psi}) := L_c(\mathbf{\Psi})/L_{Obs}(\mathbf{\Psi})$$

qui est la densité conditionnelle du couple $(\mathbf{X}, \mathbf{Z})$ à $\mathbf{x}^{Obs}$ fixées. On a la même équation qu'en 3.7 où $P(\mathbf{Z}|\mathbf{X}, \mathbf{\Psi})$ est remplacé par $k(\mathbf{X}, \mathbf{Z}|\mathbf{x}^{Obs}, \mathbf{\Psi})$ et $L(\mathbf{\Psi})$ par $L_{Obs}(\mathbf{\Psi})$.

En intégrant les deux côtés de cette égalité par rapport à la distribution conditionnelle de $(\mathbf{X}, \mathbf{Z})$ à $\mathbf{x}^{Obs}$ donné avec un jeu de paramètres $\mathbf{\Psi} = \mathbf{\Psi}^{(q)}$ pour calculer les intégrales, on a :

$$\begin{aligned} \log L_{Obs}(\mathbf{\Psi}) &= \mathbb{E}_{\mathbf{\Psi}^{(q)}}\left[\log L_c(\mathbf{\Psi})|\mathbf{x}^{Obs}\right] \\ &\quad \underbrace{- \mathbb{E}_{\mathbf{\Psi}^{(q)}}\left[\log k(\mathbf{X}, \mathbf{Z}|\mathbf{x}^{Obs}, \mathbf{\Psi})|\mathbf{x}^{Obs}\right]}_{:=H^*(\mathbf{\Psi}, \mathbf{\Psi}^{(q)})} \\ &= Q(\mathbf{\Psi}, \mathbf{\Psi}^{(q)}) - H^*(\mathbf{\Psi}, \mathbf{\Psi}^{(q)}) \end{aligned} \tag{4.3}$$

Et on montre de même qu'au paragraphe de la page 3.5 que la fonctionnelle $\mathbf{\Psi} \mapsto H^*(\mathbf{\Psi}, \mathbf{\Psi}^{(q)})$ atteint son maximum en $\mathbf{\Psi}^{(q)}$.

Ainsi maximiser la log-vraisemblance des données observées revient à maximiser l'espérance pour un jeu de paramètre donné $\mathbf{\Psi}^{(q)}$ de la log-vraisemblance complète. L'utilisation d'une telle notation nous est suggérée pour la faire intervenir à l'étape $(q+1)$ d'un algorithme de type EM qui maximise la (log-)vraisemblance. Il reste à exprimer :

$$\begin{aligned} Q(\mathbf{\Psi}, \mathbf{\Psi}^{(q)}) := & \quad \mathbb{E}_{\mathbf{\Psi}^{(q)}}\left[\log P(\mathbf{X}, \mathbf{Z}|\mathbf{\Psi})|\mathbf{x}^{Obs}\right] \\ = & \underbrace{\mathbb{E}_{\mathbf{\Psi}^{(q)}}\left[\log P(\mathbf{x}^{Obs}, \mathbf{X}^{Man}|\mathbf{Z}, \mathbf{\Theta})|\mathbf{x}^{Obs}\right]}_{:=Q_\Theta(\mathbf{\Theta}|\mathbf{\Psi}^{(q)})} \\ & + \underbrace{\mathbb{E}_{\mathbf{\Psi}^{(q)}}\left[\log P(\mathbf{Z}|\mathbf{\Delta})|\mathbf{x}^{Obs}\right]}_{:=Q_\Delta(\mathbf{\Delta}|\mathbf{\Psi}^{(q)})} \end{aligned} \tag{4.4}$$

Cette décomposition permet de réestimer les paramètres des classes indépendamment des paramètres de voisinage comme dans le cas des données complètes (voir le chapitre 3).

On exprimera d'une part avec l'approximation de type champ moyen :

$$\begin{aligned} Q_\Delta(\mathbf{\Delta}|\mathbf{\Psi}^{(q)}) &\simeq \sum_{i\in S} \mathbb{E}_{\mathbf{\Psi}^{(q)}}\left[\log P(Z_i|\mathbf{\Delta}, \tilde{z}_{\nu(i)})|x_i^{Obs_i}\right] \\ &= \sum_{\{c_k\}_{k=1,\ldots,K}} \sum_{i\in S} P_{\mathbf{\Psi}^{(q)}}(Z_i = c_k|x_i^{Obs_i}) \log P(Z_i = c_k|\mathbf{\Delta}, \tilde{z}_{\nu(i)}) \\ &\simeq \sum_{\{c_k\}_{k=1,\ldots,K}} \sum_{i\in S} \tilde{t}_{ik}^{(q)} \log P(Z_i = c_k|\mathbf{\Delta}, \tilde{z}_{\nu(i)}) \end{aligned} \tag{4.5}$$

En posant $P_{\mathbf{\Psi}^{(q)}}(Z_i = c_k|X_{Obs}) = t_{i,k}^{(q)}$, $t_{i,k}$ sera la probabilité d'appartenance *a posteriori* de l'objet i à la classe k. On estimera son approximation de type champ moyen $\tilde{t}_{i,k}^{(q)} := P_{\mathbf{\Psi}^{(q)}}(Z_i = c_k|x_i^{Obs_i}, \tilde{z}_i)$.

D'autre part, avec l'hypothèse d'indépendance des observations conditionnellement aux classes :

$$Q_\Theta(\mathbf{\Theta}|\mathbf{\Psi}^{(q)}) = \sum_{i\in S} \mathbb{E}_{\mathbf{\Psi}^{(q)}}\left[\log P_{\mathbf{\Theta}}(x_i^{Obs_i}, X_i^{Man_i}|Z_i)|x_i^{Obs_i}\right]$$

$$= \sum_{\{c_k\}_{k=1,\ldots,K}} \sum_{i\in S} \int P_{\mathbf{\Psi}^{(q)}}(X_i^{Man_i}, Z_i = c_k|x_i^{Obs_i}) \log P_{\mathbf{\Theta}}(x_i^{Obs_i}, X_i^{Man_i}|Z_i = c_k) dX_i^{Man_i}$$

Or :

$$\begin{aligned} P_{\mathbf{\Psi}^{(q)}}(X_i^{Man_i}, Z_i = c_k|x_i^{Obs_i}) &= P_{\mathbf{\Psi}^{(q)}}(Z_i = c_k|x_i^{Obs_i}) P_{\mathbf{\Theta}^{(q)}}(X_i^{Man_i}|x_i^{Obs_i}, Z_i = c_k) \\ &\simeq \tilde{t}_{ik}^{(q)} P_{\Theta_k^{(q)}}(X_i^{Man_i}|x_i^{Obs_i}) \end{aligned}$$

Donc en poursuivant les calculs ci-avant :

$$\begin{aligned} Q_\Theta(\mathbf{\Theta}|\mathbf{\Psi}^{(q)}) \simeq \sum_{\{c_k\}_{k=1\ldots K}} \sum_{i\in S} \tilde{t}_{i,k}^{(q)} \\ \int P_{\Theta_k^{(q)}}(X_i^{Man_i}|x_i^{Obs_i}) \log P_{\Theta_k}(x_i^{Obs_i}, X_i^{Man_i}) dX_i^{Man_i} \end{aligned} \tag{4.6}$$

Ainsi en notant :

- $\Sigma_k^{Obs_i}$ la matrice $\{(\Sigma_k)_{st}, (s,t) \in Obs_i^2)\}$ (notation $\Sigma_k^{Man_i}$ définie de façon similaire),
- $\Sigma_k^{Obs_i, Man_i}$ la matrice $\{(\Sigma_k)_{st}, s \in Obs_i, t \in Man_i)\}$ (notation $\Sigma_k^{Man_i, Obs_i}$ définie de façon similaire) et

- $\mu_k^{Obs_i}$ le vecteur $\{(\mu_k)_s, s \in Obs_i\}$ (notation $\mu_k^{Man_i}$ définie de façon similaire).

On a les lois gaussiennes :

$$P(x_i^{Obs_i}|\Theta_k) \sim \mathcal{N}(x_i^{Obs_i}|\mu_k^{Obs_i}, \Sigma_k^{Obs_i}) \quad (4.7)$$
$$P(x_i^{Man_i}|x_i^{Obs_i}, \Theta_k) \sim \mathcal{N}(x_i^{Man_i}|\eta_{i,k}, \Gamma_{i,k}) \quad (4.8)$$

où :

$$\eta_{i,k} = \mu_k^{Man_i} + \Sigma_k^{Man_i,Obs_i}(\Sigma_k^{Obs_i})^{-1}(x_i^{Obs_i} - \mu_k^{Obs_i}) \quad (4.9)$$
$$\Gamma_{i,k} = \Sigma_k^{Man_i} - \Sigma_k^{Man_i,Obs_i}(\Sigma_k^{Obs_i})^{-1}\Sigma_k^{Obs_i,Man_i} \quad (4.10)$$

L'algorithme (dénommé *SFmiss* dans le cas du champ simulé) prend alors la forme suivante à l'itération (q) (on pourra consulter l'annexe B.4 pour certains calculs qui ne sont pas détaillés ici) :

Étape NR Détermination de $(\tilde{z}_i^{(q)})_{i\in S}$ à partir de $\mathbf{x}^{Obs}$ et de l'estimation courante des paramètres $\mathbf{\Psi}^{(q-1)}$. Dans le cas de l'algorithme en champ simulé, on simulera la configuration des classes. Ces valeurs pour les Z_i serviront à remplacer la distribution markovienne incalculable $P_G(\mathbf{Z})$ par une distribution factorisée $\prod_{i\in S} P_G(z_i|\tilde{z}_{\nu(i)}^{(q)})$.

Étape E Il faut calculer :

$$\tilde{t}_{ik}^{(q)} = P_G(Z_i = c_k|x_i^{Obs_i}, \mathbf{\Psi}^{(q-1)}, \tilde{z}_{\nu(i)}^{(q)}) \quad (4.11)$$

Notons que le cas $Obs_i = \emptyset$ n'est pas formellement écarté. On conditionnera alors seulement en se servant des informations sur les voisins du réseaux. C'est le mieux que puisse faire cet algorithme dans le cas où aucune mesure individuelle n'est disponible.

Étape M Il s'agit de mettre à jour les paramètres de chacune des K classes :

$$\theta_k^{(q)} = \arg\max_{\theta_k} \sum_{i\in S} \tilde{t}_{i,k}^{(q)} \int P(X_i^{Man_i}|x_i^{Obs_i}, \theta_k^{(q-1)}) \log P(x_i^{Obs_i}, X_i^{Man_i}|\theta_k) dX_i^{Man_i}.$$

Ce qui mène à (en notant $\otimes$ la multiplication élément par élément des valeurs contenues dans un vecteur ou une matrice) :

$$\mu_k^{(q+1)} = \frac{\sum_i t_{ik}^{(q)}[r_i \otimes x_i + (1 - r_i) \otimes \eta_{ik}^{(q)}]}{\sum_i t_{ik}^{(q)}},$$

pour la moyenne de la classe k, en convenant que $0 \times a = 0$ pour une valeur a absente et :

$$\Sigma_k^{(q)} = \frac{\sum_i t_{ik}^{(q)} S_{ik}^{(q)}}{\sum_i t_{ik}^{(q)}},$$

avec

$$\begin{aligned} S_{ik}^{(q)} &= r_i.r_i^t \otimes (x_i - \mu_k^{(q)}).(x_i - \mu_k^{(q)})^t \\ &+ r_i.(1-r_i)^t \otimes (x_i - \mu_k^{(q)}).(\eta_{ik}^{(q)} - \mu_k^{(q)})^t \\ &+ (1-r_i).r_i^t \otimes (\eta_{ik}^{(q)} - \mu_k^{(q)}).(x_i - \mu_k^{(q)})^t \\ &+ (1-r_i).(1-r_i)^t \otimes (\eta_{ik}^{(q)} - \mu_k^{(q)}).(\eta_{ik}^{(q)} - \mu_k^{(q)})^t + \Gamma_{i,k}^{(q)}. \end{aligned}$$

pour la matrice de covariance.

On voit ici qu'il ne suffit pas de remplacer les données manquantes $(X_i)_{i \in Man}$ par la moyenne estimée à l'itération précédente. Un terme d'écart normalisé des données effectivement observées à leur moyenne se rajoute dans l'expression de η_{ik}. Aussi un terme supplémentaire explique que la meilleure (au sens de la vraisemblance) estimation de la variance soit plus élevée que dans le cas où on fait de l'imputation par une moyenne (conditionnée ou non).

La mise à jour des paramètres spatiaux de la distribution markovienne P_G reste quant à elle inchangée par rapport au cas des observations complètes (les calculs sont données dans l'annexe B.3).

Maintenant que nous avons spécifié l'apprentissage du modèle de champs de Markov cachés, nous en présenterons la validation expérimentale à la section 5.3.2. Avant cela, nous continuons avec une autre particularité de nos données : leur grande dimension et les difficultés engendrées.

4.2 Réduction de dimension

La grande dimension des données modernes est une de leurs caractéristiques. Par exemple, la résolution des images prises par des appareils photos numériques aujourd'hui courants a été multipliée par un facteur 100 en une trentaine d'années et l'augmentation a été spectaculaire ces dernières années. C'est aussi le cas des mesures faites par des appareils de technologie récente (laser, mesures dans le cadre de météorologie ou de photos satellites de planètes,...et données à haut débit post-génomiques). Une fois réglés les problèmes liés au stockage et à l'accès, cette grande dimension des données peut poser des problèmes aux méthodes qui se proposent de les analyser. Nous envisagerons naturellement intégrer des solutions existantes pour ce type d'approches dans notre modélisation. Nous voudrons tirer parti des économies en temps de calcul (qui peuvent tout simplement rendre les calculs faisables) et de méthodes adaptées qui ont apporté une amélioration lors du traitement des données en grande dimension.

4.2.1 Techniques classiques

Problématique des données de grandes dimensions

[20] fut le premier à parler du **fléau de la dimension** (*dimensionality curse*). Il faisait référence aux difficultés pour visiter une grille régulière en dimension 10 pour trouver un *optimum* d'une fonction défini sur cet espace discret. D'autres auteurs parleront de phénomène d'**espace vide**. Pour des dimensions à partir de l'ordre de 10, notre intuition n'est plus capable de bien appréhender la distribution des données. Par exemple, en dimension 10, un point de la boule unité a presque deux chances sur trois d'être «coincé» entre les sphères de rayons $0,9$ et 1. Plus généralement, en grandes dimensions, une proportion importante de données va se retrouver dans des espaces de dimensions inférieurs.

Un modèle de mélange complet requiert un nombre de paramètres de l'ordre de la dimension des données au carré. Ceci est principalement dû aux matrices de covariances qu'il est en outre nécessaire d'inverser. En plus des temps de calculs fastidieux, se pose la question de la précision des estimations et donc de la qualité de la classification obtenue en regard des données disponibles.

Notons tout de même que cette «surdimensionalité» des données face à leur nombre de degrés de liberté intrinsèque peut avoir une utilité pratique dans un objectif de classification ; il semble que les classifieurs aient des performances meilleures, toutes choses étant par ailleurs égales, pour des données de grande dimension. Les SVM exploitent cet effet en augmentant artificiellement la dimension des données pour améliorer les performances de discrimination. Les classes sont alors mieux séparées.

Propositions pour aborder les grandes dimensions

Une première solution est de dégager les espaces dans lesquels «vivent» réellement les données. C'est à dire que l'on opérera une **réduction de dimension** pour etudier seulement les dimensions significatives en vue de capter les tendances, les variations importantes des données. C'est une solution naturelle que de prendre ce problème à la source. Si des points se distribuent sur une variété de dimension inférieure à celle de l'espace original, il faut identifier cette variété pour réduire l'espace d'analyse. Certaines techniques sélectionnent les variables qui rassemblent une partie de l'information aussi grande que possible. Se pose le problème de choisir un sous-espace parmi tous les sous-espaces possibles quand la dimension est grande. D'autres techniques extraient des caractéristiques des variables de départ. La technique la plus répandue est certainement l'ACP qui cherche un sous-espace affine tel que l'inertie des données par rapport à leur projection sur le sous-espace soit la plus petite possible ([117]). Il existe des généralisations non linéaires de l'ACP. Les inconvénients de la réduction de dimension sont que cette manipulation préalable à la classification ne tient pas compte de l'objectif de création de groupes. Une perte d'informations a pu être enregistrée.

Il est par exemple fréquent d'avoir du mal à observer des classes dans les projections sur les premiers axes d'une ACP par exemple. Ou au contraire d'observer des *clusters* relativement différents selon différentes projections. Il semble alors dangereux de modifier un jeu de données de façon définitive avant la classification. La perte de données est irréversible.

Une extension possible est de chercher ces sous-espaces par des heuristiques (on cherche soit des dimensions séparant les groupes, soit directement des groupes dans des sous-espaces de l'espace de départ) ou modélisant le fait que les données sont incluses dans un sous-espace (on peut par exemple faire un modèle de mélange dont les composantes n'ont pas la même dimension [230]).

Lorsque la dimension des données est trop forte par rapport à la taille de l'échantillon disponible, nous sommes souvent confrontés à un mauvais conditionnement voire des singularités des matrices de covariances. Une solution est de **régulariser** ces matrices pour les rendre inversibles. Bien sûr alors, plus rien n'assure la qualité des estimations des paramètres même si le problème calculatoire peut être résolu. Ce biais sera d'autant plus fort que la dimension sera élevée par rapport à la quantité de données. Par ailleurs, si les classes sont effectivement de dimension plus faible que celle originale des données, les termes de la matrice de covariance qui correspondent à ces dimensions peu informatives ne décriront plus les données. Si les matrices de covariance sont mal conditionnées, c'est très certainement en raison de cette occupation très partielle de l'espace initial des données.

Une toute autre approche agira sur le modèle en le contraignant à une simplicité pour autoriser une estimation des paramètres sans trop de problèmes. On aura recours à une certaine **parcimonie** vis-à-vis du nombre de paramètres. On trouvera par exemple une description des modèles de mélange gaussiens utiles en classification dans [16] ou [48]. Nous avons adopté des modèles à matrices de covariance diagonale pour nos expériences. Cependant il faut rester conscient que cette dernière solution connaît aussi des limitess. Les modèles les plus performants sont soit les modèles les plus simples, soit les plus complexes ([48]). Les premiers induisent des hypothèses très restrictives sur les données tandis que les seconds ne permettent pas une limitation suffisante du nombre de paramètres alors que c'est leur principale motivation.

Après toutes ces méthodes pour surmonter certaines caractéristiques gênantes des données en grande dimension, nous présenterons le modèle de [38] qui intègre les points forts de ces approches en cherchant à limiter les problèmes associés.

4.2.2 Réduction de dimension par classes

Une paramétrisation du modèle est proposée. Elle permet de ne pas modifier les données préalablement à leur analyse. Par ailleurs, elle limite le nombre de paramètres en proposant des contraintes plus ou moins fortes. Enfin, en modélisant les données dans leur dimension complète, elle espère tirer parti des performances

des classifieurs en hautes dimensions. La complexité du modèle est contrôlée par les dimensions des sous-espaces des classes, et non par la dimension de l'espace d'origine.

Avec les mêmes notations que les nôtres, [38] adopte une reparamétrisation basée sur la décomposition spectrale des matrices de covariance de [16]. Pour tout $k \in \{1, \ldots, K\}$, on peut écrire $\Sigma_k = Q_k \Delta_k Q_k^t$, où Q_k est la matrice orthogonale de taille $D \times D$ contenant les vecteurs propres de Σ_k et Δ_k est une matrice diagonale avec les valeurs propres sur sa diagonale.

La paramétrisation proposée est de la forme :

$$\Delta_k = \begin{pmatrix} a_{k,1} & & 0 & & & & \\ & \ddots & & & & 0 & \\ 0 & & a_{k,d_k} & & & & \\ & & & b_k & & & 0 \\ & & & & \ddots & & \\ & 0 & & & & \ddots & \\ & & & 0 & & & b_k \end{pmatrix}$$

où les $(a_{k,l})_{l=1,\ldots,d_k} \geq b_k$ sont les d_k plus grandes valeurs propres de la matrice de covariance de la classe c_k et b_k remplace les $p - d_k$ plus petites. A la matrice Q_k corespond une matrice $\tilde{Q}_k$ composée des d_k premières colonnes de Q_k complétée par des 0. La figure 4.2 illustre la paramétrisation proposée dans le cas de deux classes. La dimension originale des données est $D = 3$ et dans les deux cas, $d_1 = d_2 = 2$.

Le modèle complet est noté $[a_{k,l} b_k Q_k d_k]$. Cette paramétrisation permet de contrôler les classes du modèle. Le paramètre d_k représente la dimension la plus adaptée à la modélisation de la classe c_k par une densité gaussienne. Une méthode intitulée *scree-test* de Cattell est utilisée pour déterminer cette dimension. Très brièvement, il s'agit de détecter la dimension qui correspond à l'arrêt de l'éboulis des valeurs propres de la matrice de covariance qui stagnent et sont faibles. On a alors capter l'essentiel de la variance des données dans le modèle. La matrice orthogonale Q_k contrôle l'orientation de la classe par rapport aux D axes originaux. Les $(a_{k,1}, \ldots, a_{k,d_k})$ contrôlent la forme de la distribution gaussienne dans l'espace de dimension d_k. Le paramètre b_k modélise le poids de la variance résiduelle dans l'espace orthogonal à celui défini par Q_k et qui modélise les données de c_k.

Ce modèle a prouvé son efficacité pour le traitement de données images (voir les publications associées à la thèse [38]) aussi bien en reconnaissance d'objets que pour des reconnaissances de textures. Nous n'avons effectué que des tests préliminaires avec nos données. Les résultats furent décevants. Probablement, un nouvel essai serait nécessaire ; l'algorithme n'était pas encore dans une forme définitive et donc peut-être pas encore complètement fiable et optimal, la collaboration avec Juliette Blanchet au sein de l'équipe Mistis n'était alors qu'à ses

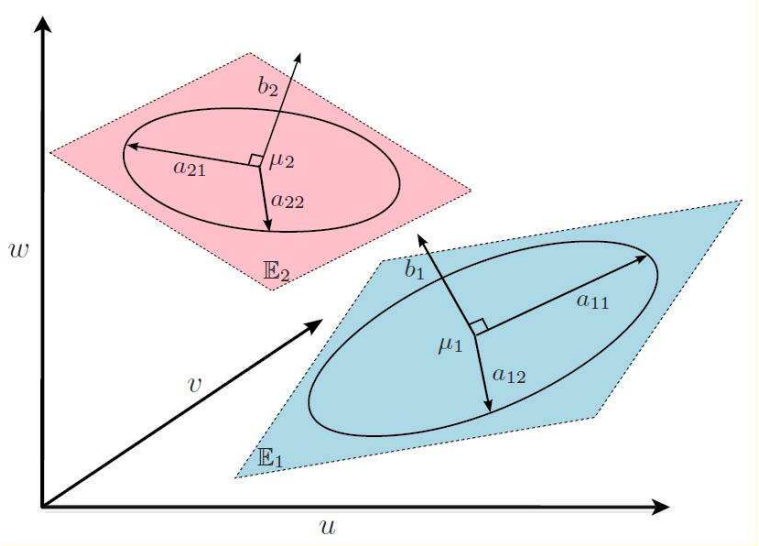


Fig. 4.2: Illustration du modèle $[a_{k,l}b_kQ_kd_k]$ de [38] en dimension $D = 3$ pour deux classes modélisées par deux distributions gaussiennes dans des sous-espaces de dimension 2 (les plans E_1 et E_2).

balbutiements. Cependant, nous avons observé que certaines classes étaient «vidées», tandis que ne s'opérait aucune réduction de dimension de façon naturelle sur d'autres. Une première idée est que les données que nous avons testées sont de nature temporelle. Ainsi il existe des dépendances évidentes entre les dimensions surtout entre les dimensions proches. Il est alors difficile de saisir comment les composantes principales contiennent cette information. Hors c'est précisément la tactique adoptée par la modélisation que nous venons de voir. Aussi, nous pensons que l'estimation empirique de d_k est un point problématique. Malgré ces débuts difficiles, il nous semble que ce travail a soulevé des points importants. Quitte à revoir la reparamétrisation en question, nous pensons que la spécificité d'un modèle pour des données de grande dimension telles que celles face auxquelles nous serons toujours plus confrontées en biologie moléculaire peut être très profitable.

Après avoir passé en revue des extensions pour la prise en compte d'observations individuelles manquantes et pour les techniques spécifiques des données de grandes dimensions, nous allons présenter la validation de notre modèle *via* les expériences qui ont été menées.

5. APPLICATIONS AUX DONNÉES

NOUS PRÉSENTONS dans ce chapitre les résultats des expériences mises en place.

5.1 Données simulées

L'avantage des données simulées est que l'on maîtrise entièrement les données et leur classe d'appartenance. Mais on connait aussi le mécanisme qui a produit les données. C'est un avantage supplémentaire pour savoir si on est dans le cadre des hypothèses du modèle ou si on s'en éloigne. On peut ainsi commencer à juger de la stabilité de la méthode vis-à-vis de données qui ne suivent pas forcément les hypothèses du modèle. Notamment l'hypothèse de distribution gaussienne des données dans chacune des classes est très discutable dans de nombreuses applications. Aussi les données simulées fournissent des répliques en grand nombre à faible coût.

Cependant nous ne prétendons pas que nos données reflètent peu ou prou des caractéristiques de données réelles. Nous avons seulement essayé de les choisir cohérentes et diversifiées. En fait, la modélisation de données d'expression de gènes par exemple est à ce jour un sujet à part entière bien loin d'être clos. De nombreux modèles ont été proposés : [164, 257].

Résultats sur données simulées avec un réseau de type image

Dans un premier travail ([82]), nous nous sommes inspirés des méthodes de simulation de [254]. Les données ainsi obtenues modélisent un comportement cyclique (sinusoïdal) ou une augmentation/diminution (linéaire) de l'activité au cours du temps de données de type expression de gènes.

Nous avons créé 5 jeux de données selon ce modèle. Chaque jeu est composé d'un ensemble de $N = 1536$ objets (appelons les gènes) dans $D = 20$ dimensions (ou conditions expérimentales). Ces gènes sont rangés dans $K = 6$ classes de taille identique (256 gènes par classe).

Le gène i de la classe c_k dans la condition j est contraint de prendre la forme :

$$x_{ij} = \sin(2\pi j/10 - \pi k/4) + \epsilon_{ij}, \text{ pour } k = 1, \ldots, 4, \tag{5.1}$$

et

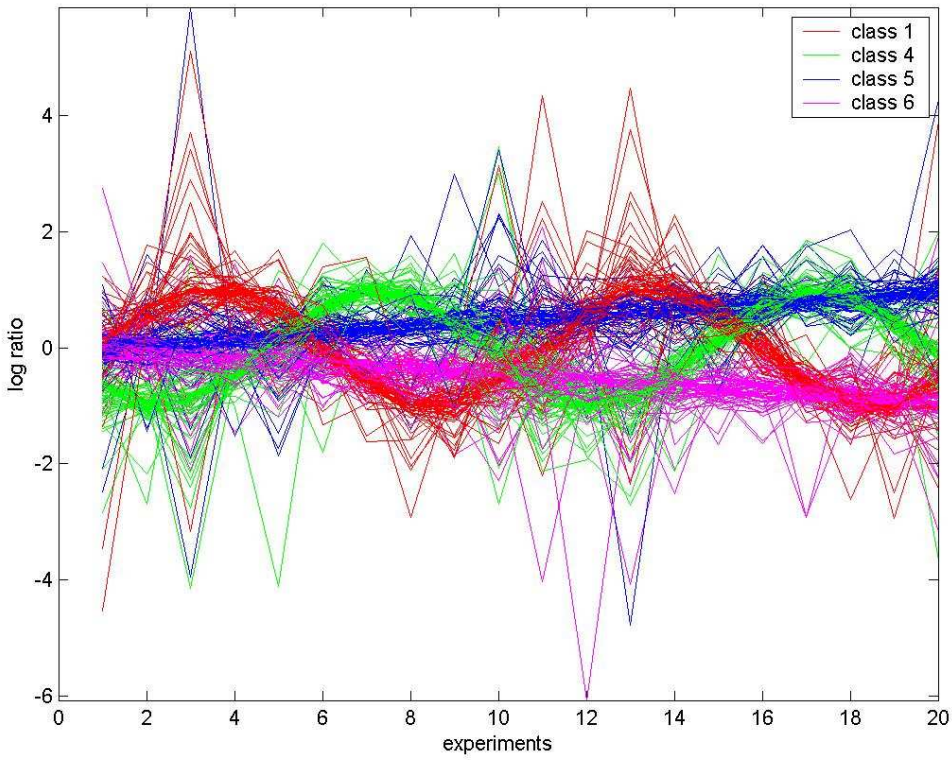


Fig. 5.1: Illustration de données synthétiques comme dans [82]

$$x_{ij} = j/20 + \epsilon_{ij}, \text{ pour } k = 5 \qquad (5.2)$$
$$x_{ij} = -j/20 + \epsilon_{ij}, \text{ pour } k = 6.$$

Les ϵ_{ij} sont générés selon des bruits gaussiens de moyenne 0 et d'écart-type 6 fois ceux pris sur des données réelles décrits dans [110]. Un tel bruit rend ces données non triviales et tente de modéliser la variabilité qui peut désynchroniser un gène. Les données sinusoïdales sont déphasées selon la condition expérimentale (pour rendre compte des instants successifs de mesures, nous aurons ici deux cycles complets observés) et la classe d'appartenance (pour rendre compte de régulation dans des phases différentes du cycle cellulaire par exemple comme dans [51] ou [217]). Une illustration de deux classes sinusoïdales et des deux classes linéaires pour un des cinq jeux de données est fourni à la figure 5.1.

Nous utiliserons conjointement à ces données individuelles simulées un réseau d'interactions. Il n'y a pas à notre connaissance de méthode bien établie pour simuler des classes sur un réseau d'interactions. Par exemple [63] donne une revue

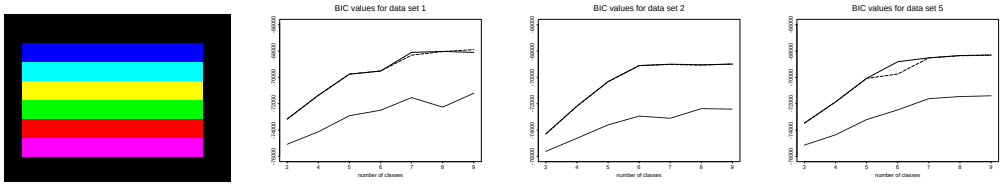


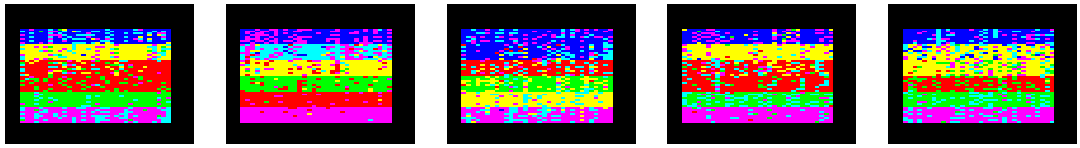
Fig. 5.2: Classification de référence et courbes BIC pour trois jeux de données lorsque K varie de 3 à 9. La ligne pleine est celle correspondant à l'algorithme en champ simulé tandis que la ligne pointillée est celle de l'algorithme EM standard et la ligne interrompue celle de l'algorithme en champ moyen.

des méthodes qui traitent des réseaux géniques de régulation. Nous ne pouvons pas vraiment identifier des comportements de classes sur ce type de réseau parce que le nombre de gènes analysables simultanément par ces méthodes dépasse rarement la dizaine.

À titre d'illustration, nous avons considéré que les données décrites par les équations (5.1) et 5.3 voient leurs classes réparties sur six bandes d'une grille régulière de type image. La taille de l'image est 48×32. Chaque bande est donc de taille 8×32 et est représentée par une couleur (voir la figure 5.2 tout à gauche). Nous avons pris en compte les huit voisins sur la grille régulière (Nord, Nord-Est, Est, Sud-Est, Sud, Sud-Ouest, Ouest et Nord-Ouest). Un tel réseau n'a évidemment aucune interprétation biologique. Il est très certainement trop homogène pour prétendre être réaliste. Nous sommes conscients de cette écueil. Mais la classification est aisément visualisable et il est facile de calculer le gain dans la prise en compte du réseau par notre méthode.

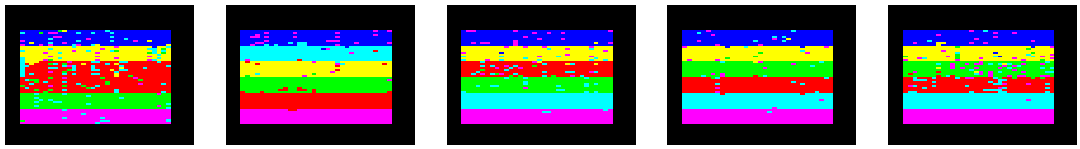
Nous avons comparé l'algorithme EM standard (classifications à la figure 5.3 pour $K = 6$) qui considère les sites comme indépendants au sein d'un modèle de mélange et les procédures de type EM (classifications de l'algorithme en champ simulé aux figures 5.4 pour $K = 6$ et 5.5 pour $K = 7$) que nous proposons. Des valeurs typiques de l'indice BIC sont reportées à la figure 5.2 (de gauche à droite pour les jeux de données 1, 2 et 5).

Les procédures de type champ moyen montrent des valeurs de BIC supérieures à l'algorithme EM standard. Le critère sélectionne le bon nombre de classes à l'exception des jeux de données 1 et 4 où $K = 7$ est sélectionné. Cependant cela est cohérent avec les classifications obtenues à la figure 5.5. Il semble que tout en étant plus performant (voir la table 5.1), l'algorithme SF confonde deux bandes. Ce qui est toujours le cas pour l'algorithme EM standard sauf pour le 2^{e} jeu de données. Les classifications en $K = 7$ groupes sont visuellement meilleures pour les jeux de données 1 et 4 comme les valeurs de BIC le suggèrent. L'interprétation que nous pouvons en donner est que dans les cas très bruités, on peut s'autoriser à prendre en compte une classe supplémentaire. Cette dernière n'a pas vraiment de signification autre que de rassembler des mesures exceptionnelles ou trop ambiguës.



(a) (b) (c) (d) (e)

Fig. 5.3: Algorithme EM standard : classification en 6 groupes pour les cinq jeux de données.

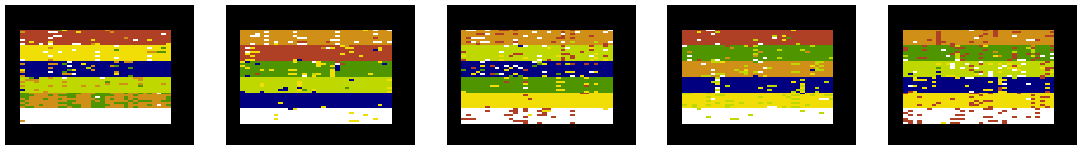


(a) (b) (c) (d) (e)

Fig. 5.4: Algorithme en champ simulé : classification en 6 groupes pour les cinq jeux de données.

Les algorithmes en champ simulé et en champ moyen (voir la partie 3.6) se comportent de façon très proche sur cet exemple sauf pour le 5^{e} jeu de données où l'algorithme en champ simulé sélectionne $K = 6$ classes et obtient de meilleures performances de classification. Nous avons eu l'occasion d'observer ce type de comportement sur de nombreux exemples variés. C'est pourquoi nous avons porté notre choix sur l'algorithme en champ simulé pour analyser les données réelles.

Le tableau 5.1 montre la proportion de gènes correctement groupés pour les algorithmes EM standard et SF. Tous les jeux de données montre une amélioration notable entre les algorithmes EM standard et en champ simulé. Le tableau 5.2 donne la matrice de confusion entre la classification de référence (en lignes) et la classification pour $K = 6$ classes sélectionnées par le critère BIC pour l'algoritme SF (colonnes). Les termes sur la diagonale sont les proportions de gènes bien classés ensembles (les indices des classes sont arbitraires). Les autres termes comptabilisent les gènes classés ensembles à tort. On voit bien comme à l'image la plus à droite de la figure 5.4 que la classe 6 est entièrement retrouvée. En revanche des éléments $(10, 5\%)$ de la classe 3 sont inclus à tort dans cette classe SF



(a) (b) (c) (d) (e)

Fig. 5.5: Algorithme en champ simulé : classification en 7 groupes pour les cinq jeux de données.

data sets	1	2	3	4	5
EM	64,8	79,2	63,3	68,1	64,3
Simulated field	77,5	95,8	93,6	78,6	91,3

Tab. 5.1: Taux de bonne classification en pourcentages pour les cinq jeux de données simulées ($K = 6$).

Taux de bonne classification globale : 91, 3%						
Class	1	2	3	4	5	6
1	94,1	1,2	0	0	0,8	3,9
2	1,2	89,1	3,1	0	2,0	4,7
3	0	1,2	80,9	1,6	5,9	10,5
4	0	0	1,6	84,0	12,1	2,3
5	0	0	0	0	99,6	0,4
6	0	0	0	0	0	100

Tab. 5.2: Matrice de confusion pour la classification issue de SF sur le 5^{e} jeu de données correspondant à la figure 5.4, image la plus à droite.

6. De même on peut aussi remarquer que des éléments de la classe 4 sont inclus à tort dans la classe SF 5 qui est par ailleurs presque parfaitement retrouvée. Ces deux tendances sont bien visibles sur l'image qui représente la classification obtenue.

Le gain lors de la prise en compte de l'information de réseau (ou de dépendances) apparaît clairement sur ces données simulées. De meilleures performances de classification sont observées. Le critère BIC ou son approximation dans le cadre de champ de Markov apparaît comme un critère satisfaisant pour la sélection du nombre de classes aussi bien que celle du modèle le plus performant. Il reste cohérent avec la visualisation que nous avons proposée. Ces premières conclusions nous guiderons pour l'analyse des données réelles à la partie 5.3.

Résultats sur données simulées incomplètes et prise en compte du réseau

Nous présentons ici des expériences faites sur données simulées correspondant à nos travaux dans [32] et [33].

Nous avons d'abord mené des expériences pour comparer quelques approches existantes pour la prise en compte des valeurs manquantes dans un jeu de données. Les méthodes suivantes ont été testées :

- EMmiss (voir partie 4.1) où le réseau n'est pas pris en compte,
- SFmiss (*idem*) qui est l'algorithme que nous avons développé,
- ZERO+SF : remplacement des valeurs manquantes par des zéros avant d'utiliser l'algorithme en champ simulé sur les données complétées,

- MEAN+SF : imputation préalable des valeurs manquantes par la moyenne de la colonne (c'est à dire de la dimension qui représente une condition expérimentale) puis SF classique.
- CMEAN+SF (moyenne conditionnelle) : semblable à SFmiss mais on biaise l'estimation de la variabilité en remplaçant les valeurs manquantes par leur moyenne η_{ik} sans se soucier de la variabilité des observations plutôt que la procédure décrite dans la section 4.1.
- UMEAN+SF (moyenne non conditionnelle) : comme CMEAN mais en remplaçant la valeur manquante par la moyenne des éléments de la classe sur la colonne $\mu_k^{Man_i}$.
- KNN+SF : imputation préalable par les $K(= 15)$ plus proches voisins au lieu de toutes les valeurs dans (U)MEAN. Des poids peuvent être définis, voir [231].

On considère deux expériences ici en faisant varier le modèle pour voir comment les méthodes ci-avant réagissent face à des modèles de voisinage, des mécanismes d'absence de données et des dimensions variées [1] :

1. Un modèle de Potts à deux classes en dimension $D = 10$ est simulé avec paramètre $\beta = 0.4$. Les sites sont répartis comme des pixels sur une image de taille 100×100. On considère les 8 plus proches voisins. Nous avons aussi testé la prise en compte du voisinage à 4 voisins. C'est un résultat général à toutes les données simulées que les résultats sont alors un peu moins bons. Les classes ont pour distributions respectives $\mathcal{N}(0, 2Id_{10})$ et $\mathcal{N}(u_1, Id_{10})$, où $u_1 = (1, 1, ..., 1)$ et Id_{10} est la matrice identité en dimension 10. Des valeurs sont supprimées en proportion ρ qui varie de 0 à 1 dans un cadre MCAR. Les taux globaux de bonne classification sont donnés à la figure 5.6 (A).
2. Une image «damier» à 4 couleurs de taille 128×128 est bruitée par $\mathcal{N}(0, \Sigma)$ où $\Sigma_{k,l} = 0,5$ si $k = l$ et $0,2$ sinon. On s'est ici placé en dimension $D = 4$ et les moyennes des quatre classes avant bruit sont $(i-1, i-1, i-1, i-1)$ pour la classe $c_i, i \in \{1, 2, 3, 4\}$. On considère ici aussi les 8 plus proches voisins. Les valeurs sont censurées (processus NMAR donc). Une fraction ρ des valeurs est enlevée : les $\rho/2$ plus petites et $\rho/2$ plus grandes. Les taux globaux de bonne classification sont donnés à la figure 5.6 (B) avec des visualisations des classifications pour $\rho = 0$, 30, 60 et 70%. Le modèle SFmiss estime le β de cette image à $2,14$ (rappelons qu'une telle image n'est pas du tout la réalisation d'un modèle de Potts) ; cette valeur ne peut pas en toute rigueur être comparée directement aux $\beta = 0,4$ du modèle de Potts ci-avant parce

[1] Nous n'avons pas réellement abordé le cas de la grande dimension. Nous nous sommes limité à des modèles en dimension 10 et encore en considérant le cas de matrices de covariance diagonales. Cette limitation est la même que dans le cas des modèles de mélange. On pourra consulter [48] à ce sujet ou [38] pour une approche que nous aimerions inclure dans notre modèle.

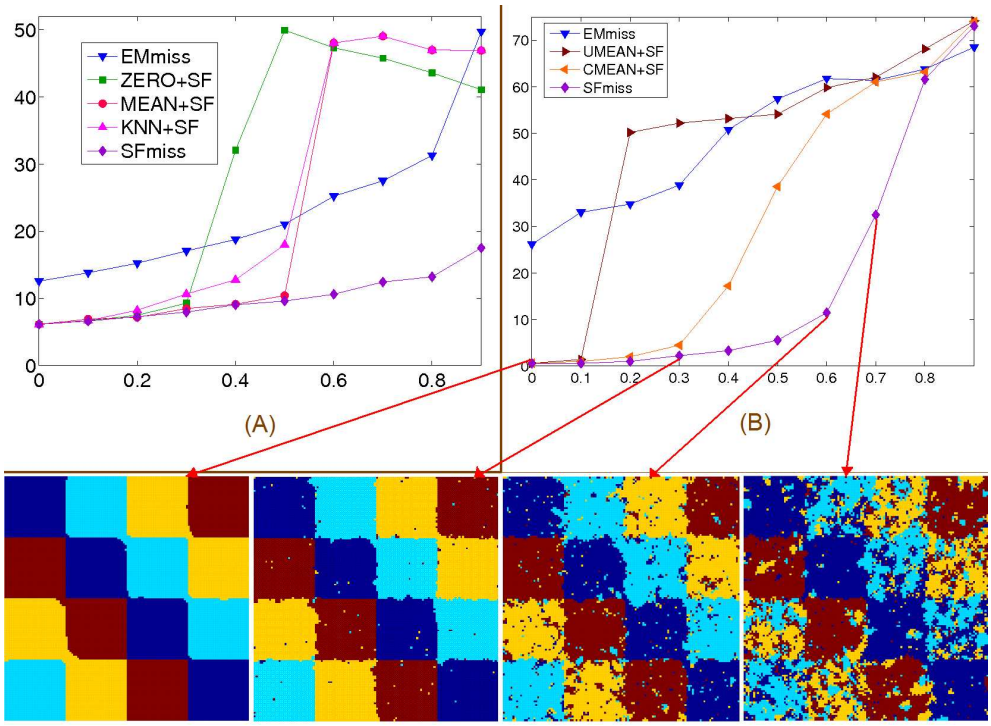


Fig. 5.6: (A) Simulation sur un modèle de Potts : graphe du pourcentage de pixels mal classés en fonction de la fraction d'observations manquantes dans le cas MCAR ; (B) même graphe pour le jeu de données synthétiques image à 4 classes dans le cas d'observations censurées (cas NMAR) et visualisation de certains résultats de classifications.

que le nombre d'objets n'est pas le même : 128×128 ici contre 100×100. Cependant l'ordre de grandeur est comparable. Notons que le bruit dicté Σ fait que les données sont également réparties au sens de la distance de Mahanalobis par rapport au cas précédent avec $D = 10$ et $\Sigma = 2Id_{10}$. En une dimension, cela correspondrait à une variance $\sigma^2 = 0,25$ (voir les expériences sur modèle de Potts unidimensionnel ci-après). La distance avec les moyennes choisies est de l'ordre de 2 soit des gaussiennes moyennement séparées (donc bien séparées quand $D = 1$ et $\sigma^2 = 0,1$ et mal séparées quand $D = 1$ et $\sigma^2 = 1$).

La figure 5.6 donne donc les performances des différentes méthodes. On voit tout d'abord nettement que dans les deux cas, SFmiss a de meilleures performances que tous les autres algorithmes surtout pour des valeurs de ρ élevées. Aussi la prise en compte du voisinage semble cruciale. Même si la qualité de l'algorithme que nous proposons contrebalance cet effet quand de plus en plus de données viennent à manquer. Alors que l'imputation préalable par la moyenne de la colonne semblait acceptable jusqu'à $\rho = 0,5$ dans la simulation du modèle de Potts, cette méthode s'effondre dès que $\rho > 0,1$ dans le cas de l'image à quatre

couleurs. L'imputation préalable par les K (=15) plus proches voisins (KNN) est un peu meilleure mais ses performances restent moins bonnes que l'imputation par la moyenne calculée sur la colonne (MEAN). Cette dernière technique a des performances comparables à SFmiss jusqu'à $\rho = 0,5$ sur le modèle de Potts et $\rho = 0,2$ sur l'image-damier.

La dimension semble avoir peu d'influence même si des dimensions pas trop petites permettent d'avoir des vraies observations pour un individu. Le cas extrême $D = 1$ sera traité ci-dessous ; dans ce cas, une valeur manquante est synonyme de toute la donnée x_i manquante ! Cela explique sûrement la robustesse de SFmiss dans le modèle de Potts dans le cas MCAR même quand $\rho = 90\%$! Notons enfin que SFmiss se comporte bien même si modèle n'est pas le bon (une image n'est pas la réalisation d'un modèle de Potts !) et si les données sont NMAR alors que notre modélisation fait l'hypothèse qu'on est MAR. Le taux de bonne classification atteint encore 88% alors que la proportion d'observations manquantes est de $\rho = 60\%$. Pour des valeurs de ρ plus fortes, la méthode commence alors à connaître ses limites.

Après avoir regardé les méthodes existantes pour traiter les valeurs absentes, et avoir validé l'algorithme SFmiss, nous allons regardé les effets du modèle sur l'estimation des paramètres et sur les performances en terme de classification.

Ici nous avons affaire à un champ de Potts à deux classes simulées sur une grille régulière de taille 100×100. Les données sont unidimensionnelles et les distributions des classes sont des gaussiennes $\mathcal{N}(0, \sigma^2)$ et $\mathcal{N}(1, \sigma^2)$. On considèrera des valeurs de σ de $0,1$ (bruit faible), $0,25$ et 1 (cas très bruité). On simulera trois modèles différents pour $\beta = 0,2$, $0,3$ et $0,4$. La visualisation de ces jeux de données est donnée à la figure 5.7. Notons qu'il n'y a pas d'égalité du nombre d'objets dans chacune des deux classes. Entre $0,2$ et $0,4$, on observe une transition du regroupement géographique des classes en îlots ou agrégats d'abord puis en régions presque homogènes. On se place donc dans des situations où le spatial commence à avoir de l'importance sans être très fort comme dans une image comme en 5.6 (B).

La figure 5.8 donne les distributions originales qui sont adjointes au modèle de Potts selon le bruit envisagé et le paramètre β. En les comparant aux figures 5.9, 5.13 et 5.15, on peut mieux comprendre les manipulations effectuées pour simuler un mécanisme d'absence de certaines observations. La table 5.3 montre les estimations correspondantes des paramètres et les taux de classification obtenus quand aucune observation $\mathbf{x}$ n'est manquante. On peut voir que les paramètres sont très bien estimés. Il apparaît toutefois que les modèles où le caractère spatial est peu marqué sont plus durs à capter avec un modèle par champ de Markov caché. Cependant ce modèle fonctionne toujours mieux que le modèle indépendant. Aussi un niveau de bruit élevé rend la reconnaissance des classes plus difficiles.

Nous avons considéré les absences de données suivantes que nous notons pour ce jeu synthétique :

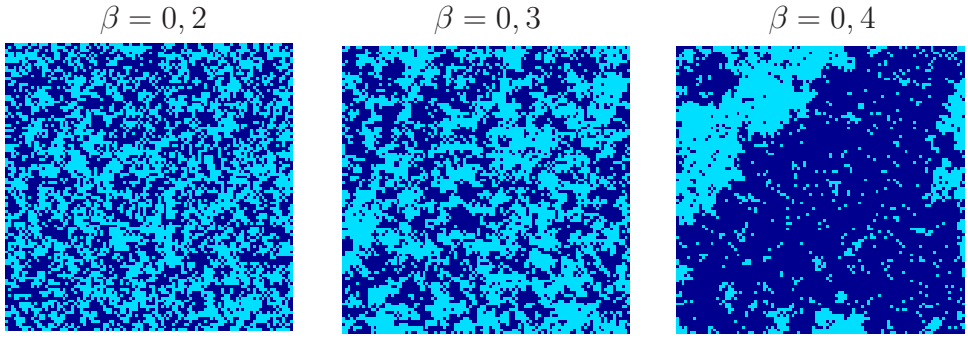


Fig. 5.7: Modèle de Potts simulé par échantillonneur de Gibbs pour différentes valeurs de β

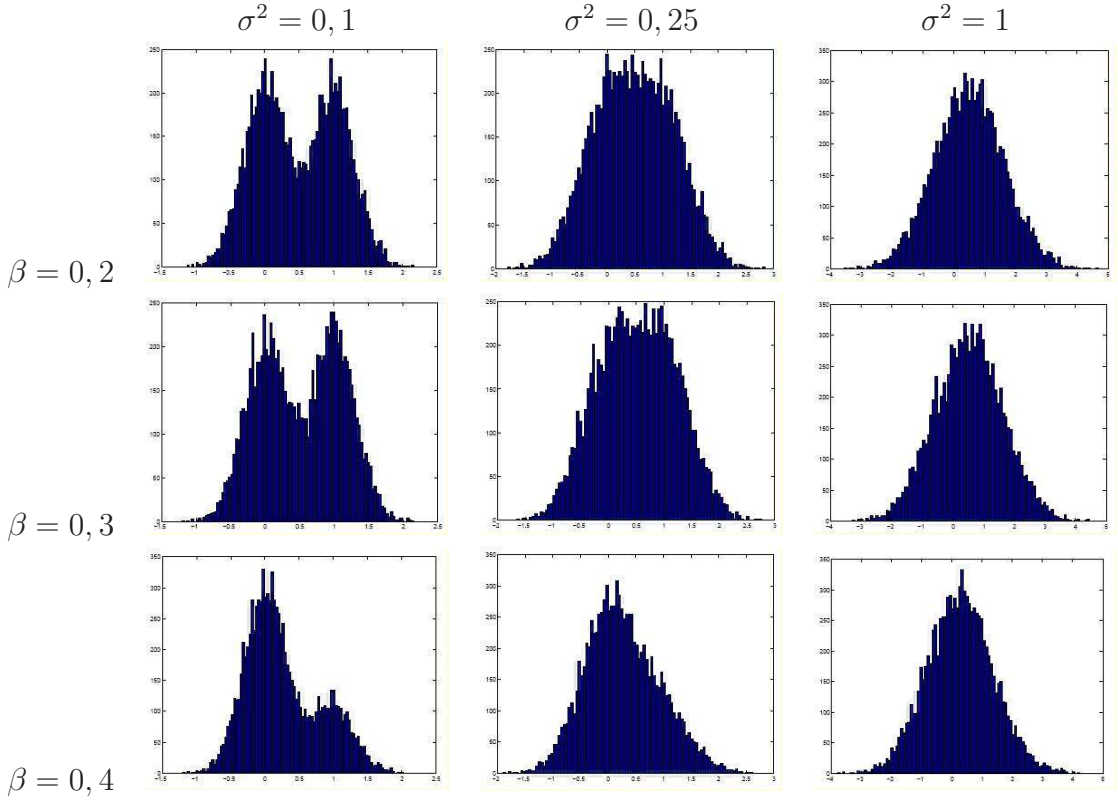


Fig. 5.8: Histogrammes des données simulées sur le champ du modèle de Potts de la figure 5.7. La distribution est normale de moyenne 0 ou 1 selon la classe et d'écart-type variant de $0,1$ à 1.

β	μ_1	σ_1^2	μ_2	σ_2^2	erreur
0,21	-0,00	0,10	1,00	0,10	5,11
0,30	0,00	0,10	1,00	0,10	4,60
0,40	-0,00	0,10	0,99	0,10	2,80
0,22	-0,01	0,25	0,99	0,26	14,40
0,29	0.00	0,25	0,99	0,25	12,56
0,40	0,00	0,25	0,99	0,25	7,11
0,24	0,01	1,03	1,00	1,00	28,42
0,29	-0,00	1,01	0,99	1,00	26,84
0,39	-0,00	1,00	0,99	1,00	12,22

Tab. 5.3: Paramètres estimés et taux d'erreur de classification sur les données complètes.

- cas «mcar» $P(R_i = 0|\mathbf{x}^{Obs}, \mathbf{x}^{Man}, \mathbf{z}, \rho) = \rho$; on tire $100 \times \rho\%$ des données de manière aléatoire et indépendante et on les supprime. Les distributions correspondantes sont données dans la figure 5.9 pour le cas $\sigma^2 = 0,1$; les distributions dans le cas $\sigma^2 = 1$ ne sont pas représentées. La quantité totale d'observations diminue mais pas les caractéristiques de la distribution des données qui restent très proches. La figure 5.10 reporte les estimations des paramètres et le taux de classification obtenu pour les trois valeurs de β dans le cas $\sigma^2 = 0,1$. Ces estimations des paramètres sont données en fonction de la proportion de données manquantes ρ. La figure 5.12 donne les mêmes visualisations quand $\sigma^2 = 1$.

  La figure 5.11 fournit des visualisations dans le cas $\beta = 0,4$ des classifications obtenues pour tous les mécanismes d'absence (un par colonne) de données considérés. Les cas $\beta = 0,2$ et $0,3$ sont beaucoup moins parlant visuellement.

  Dans le cas où $\sigma = 0,1$, les estimations sont excellentes même pour des taux d'absence élevé. La classification est aussi très bien retrouvée ; on est à 22% d'erreur alors que 50% des données sont manquantes dans le cas le moins favorable où $\beta = 0,2$. Alors que le résultat de classification dépend de manière visible de β, cela ne semble pas être le cas pour l'estimation des paramètres de classes. Le cas $\rho = 90\%$ est mauvais mais s'accompagne conjointement d'une mauvaise estimation de β. Le modèle a un comportement très bon comportement jusqu'à $\rho = 0,5$ et qui reste très acceptable jusqu'à $\rho = 0,8$.

  Le cas «mcar» lorsque $\sigma^2 = 1$ est comme nous l'attendions plus mitigé. Notons que c'est une situation très bruitée puisque l'écart-type est alors aussi important que la différence du comportement moyen des deux classes. Les taux de classification pour $\beta = 0,2$ restent très bons mais ceux de $\beta = 0,3$ ou $0,4$ sont fortement déteriorés. On a un taux d'erreur de 30% pour $\beta = 0,3$ et de 35% pour $\beta = 0,4$ losque $\rho = 0,5$. Rappelons ici qu'un

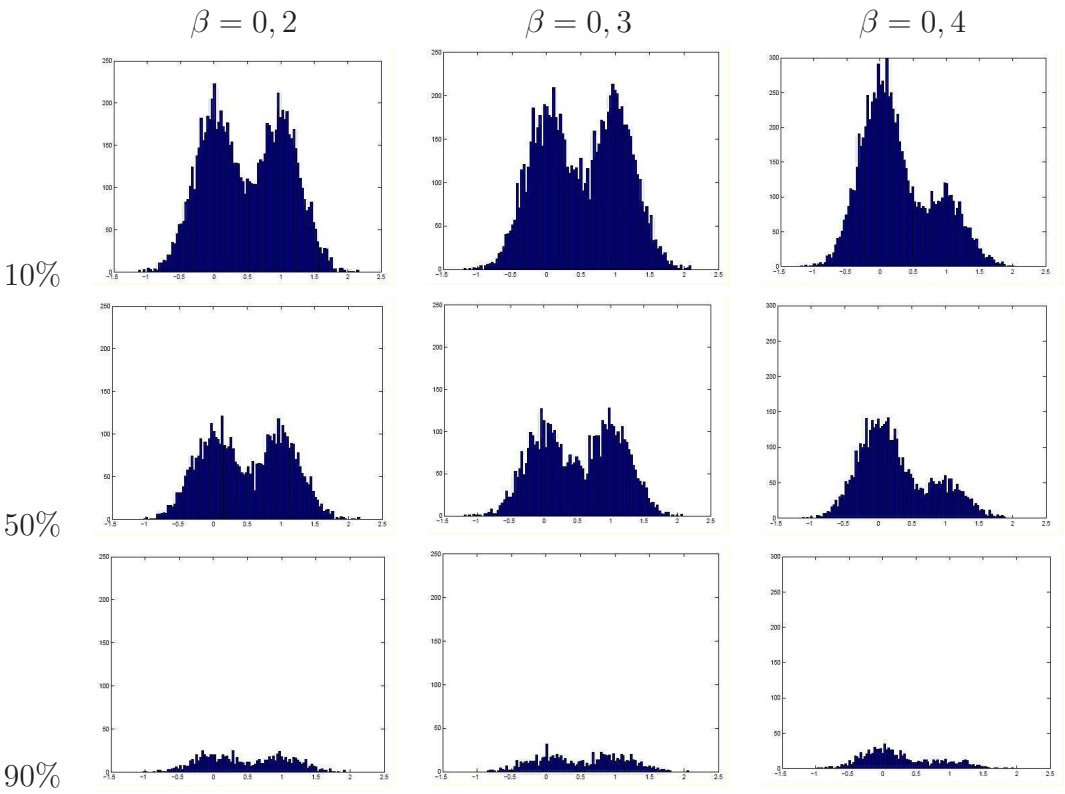


Fig. 5.9: Distribution des données simulées pour différentes proportions de valeurs manquantes dans le cas MCAR ($\sigma^2 = 0,1$).

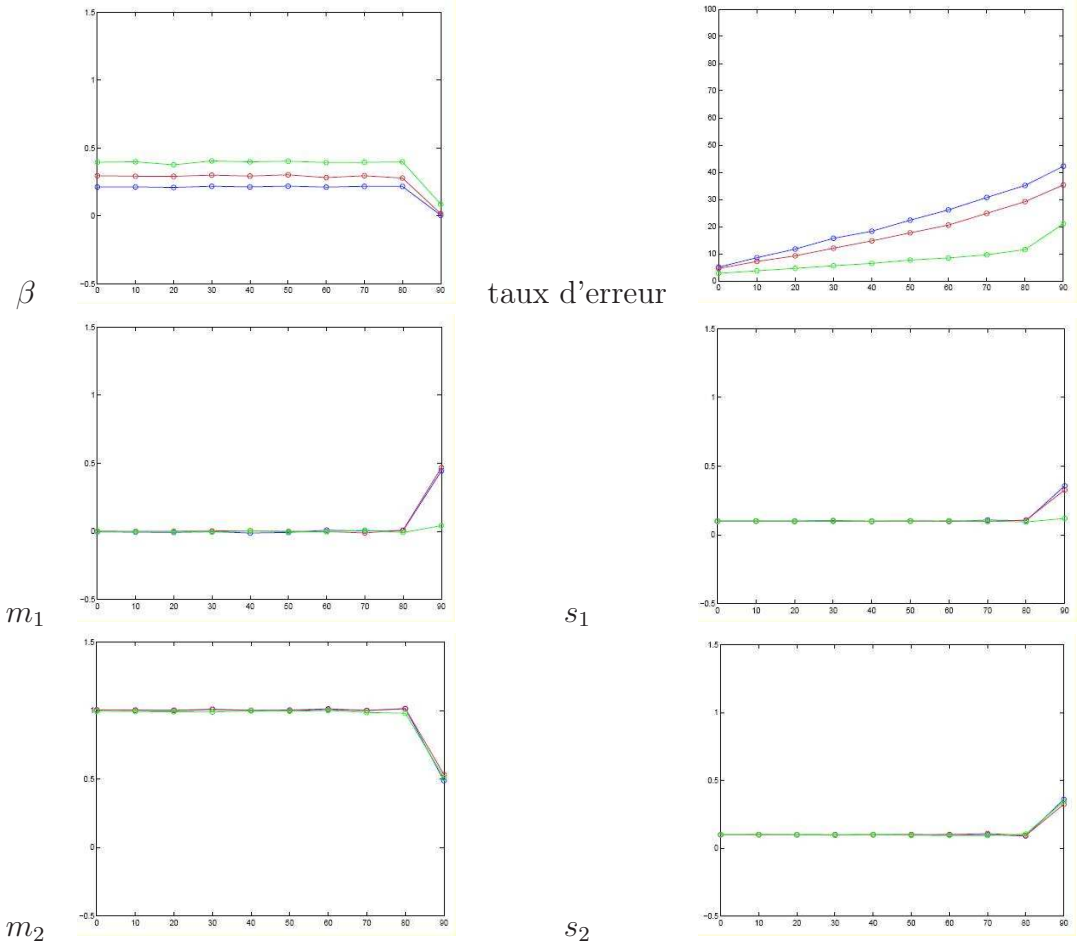


Fig. 5.10: Estimation des paramètres du modèle et taux de classification pour les trois simulations de 5.7 : $\beta = 0,2$ (bleu), $0,3$ (rouge) et $0,4$ (vert) dans le cas où les données manquantes sont MCAR ($\sigma^2 = 0,1$).

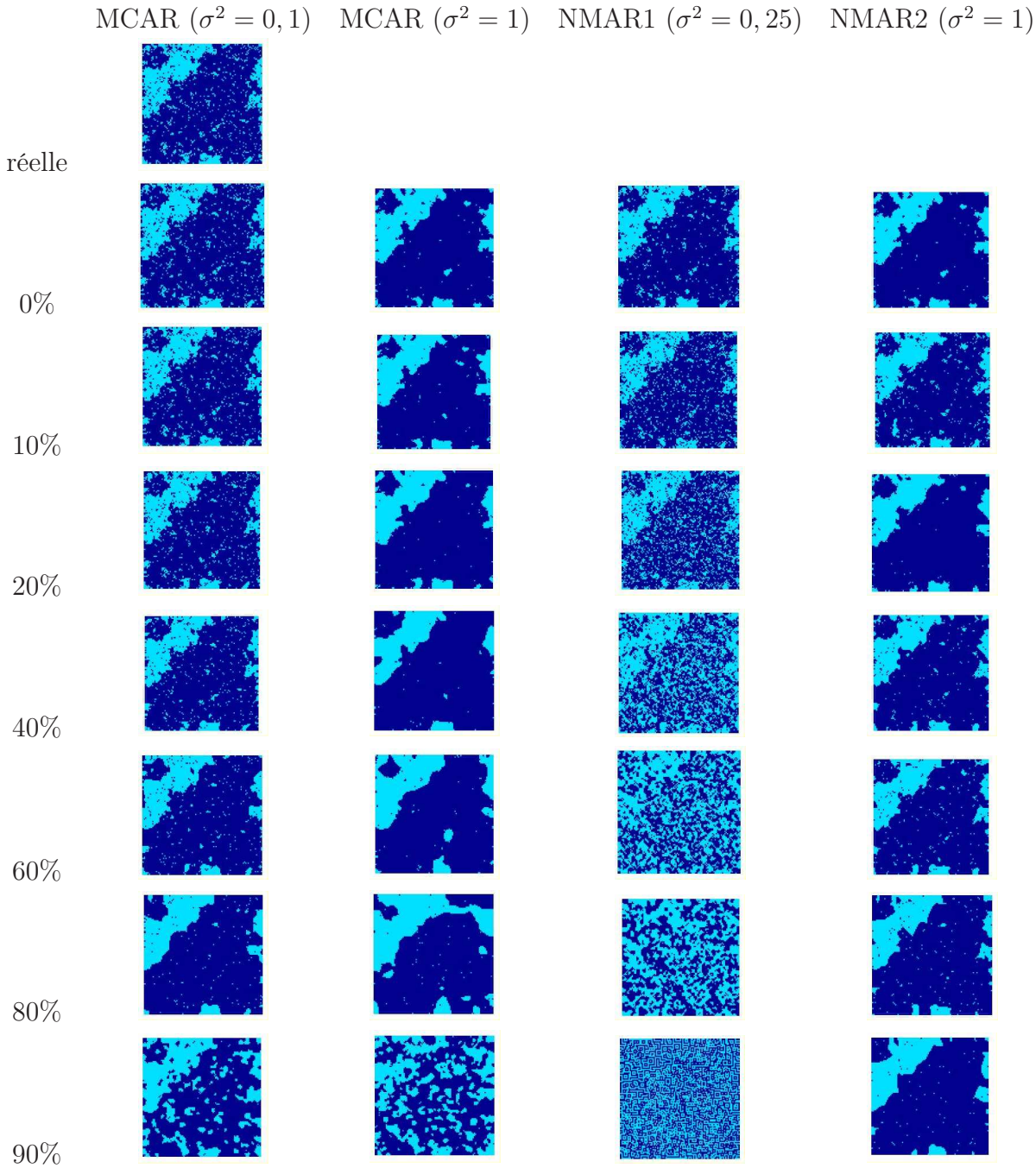


Fig. 5.11: Visualisation de la classification obtenue pour différentes proportions de valeurs manquantes selon divers mécanismes et bruits (les quatre cas choisis pour la visualisation sont indiqués sur la première ligne). La première colonne précise le pourcentage d'observations manquantes (réelle signifie que c'est l'image originale simulée à retrouver). On observe bien une dégradation de plus en plus importante des résultats de classification avec l'augmentation de la quantité d'observations manquantes à des degrés moindres selon le mécanisme d'absence et le bruit considérés.

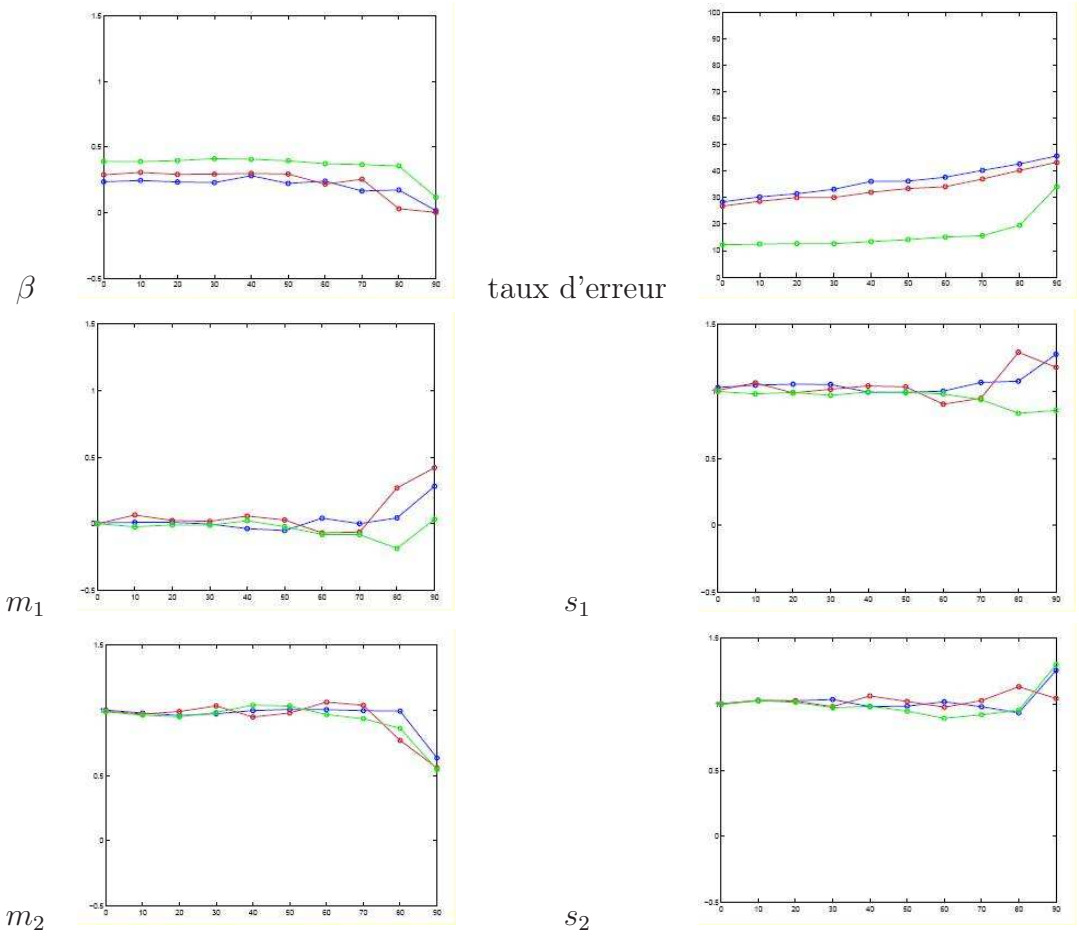


Fig. 5.12: Estimation des paramètres du modèle et taux de classification pour les trois simulations de 5.7 : $\beta = 0,2$ (bleu), $0,3$ (rouge) et $0,4$ (vert) dans le cas où les données manquantes sont MCAR ($\sigma^2 = 1$).

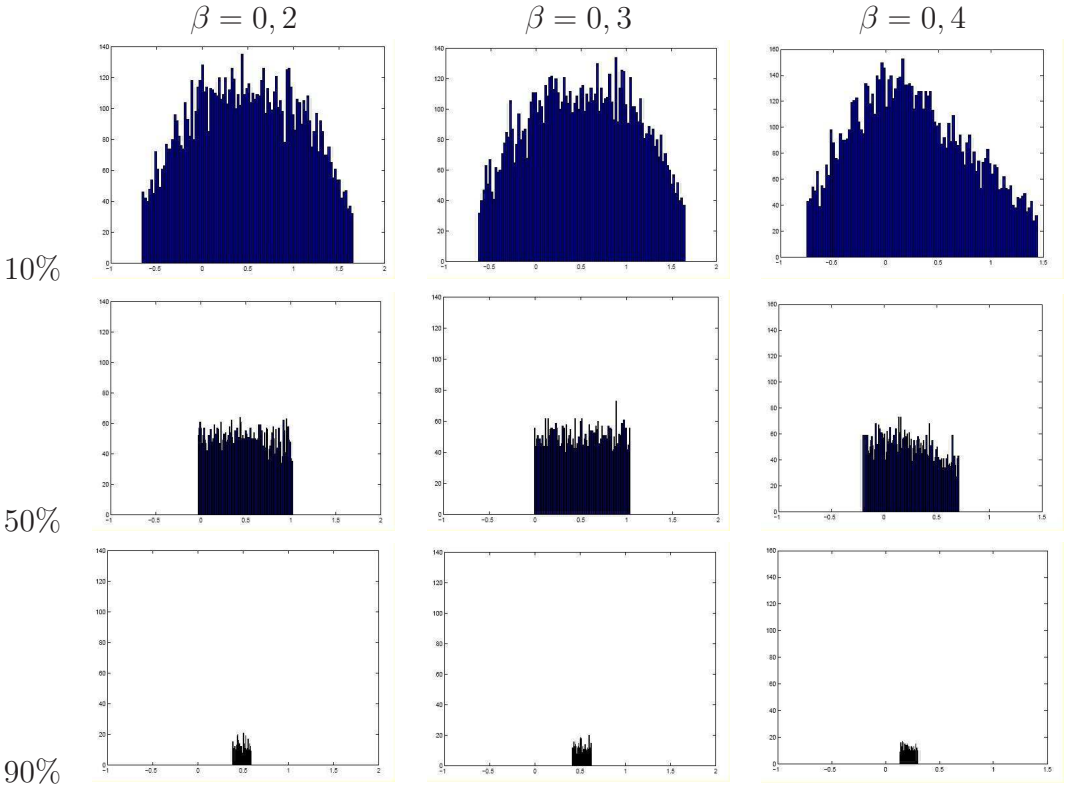


Fig. 5.13: Distribution des données simulées pour différentes proportions de valeurs manquantes dans le cas NMAR1 ($\sigma^2 = 0,25$).

classifieur aléatoire naïf (souvent cité comme le pire qu'on puisse faire dans le littérature) devrait obtenir des scores proches de 50% d'erreur dans le cas où il assigne les objets aux classes c_1 et c_2 en jouant à pile ou face avec une pièce équilibrée... Les estimations des paramètres sont acceptables jusqu'à un niveau de $\rho = 0,7$ mais elles semblent nettement moins stables à la perte de données. On voit aussi visuellement que le modèle réagit en régularisant fortement la classification en se servant du voisinage : dans les cas très bruités, le manque de données provoque des difficultés à notre modèle qui fait alors de plus en plus confiance au réseau.

- cas «nmar1» : on censure les données respectivement en-dessous et au-dessus des quantiles empiriques d'ordre $\rho/2$ et d'ordre $1 - \rho/2$. Formellement, $P(R_i = 0|\mathbf{x}^{Obs}, \mathbf{x}^{Man}, \mathbf{z}, \rho) = 1$ si $x_i < q_{\rho/2}$ ou $x_i > q_{1-\rho/2}$ et 0 sinon. $100 \times \rho/2\%$ des données les plus petites en valeur manquent et de même $100 \times \rho/2\%$ des données les plus grandes manquent. Nous nous sommes ici placés dans un cas assez bruité (en tout cas vis-à-vis du mécanisme de censure très exigeant) : $\sigma^2 = 0,25$.

  On voit à la figure 5.13 que la forme des distributions est assez rapide-

ment modifiée en profondeur quand la proportion de valeurs manquantes augmente et pas selon la quantité totale de données disponibles. Cela a nécessairement un effet sur l'estimation des paramètres que l'on peut voir à la figure 5.14. Au delà de $\rho = 10\%$, ces estimations sont très faussées. Pourtant les taux de classification restent acceptables tant que $\rho \leq 30\%$ quel que soit β. On classe alors correctement 70% des sites. Au delà, la difficulté à retrouver l'image simulée initialement est bien visible (*cf.* figure 5.11). Cela s'explique par le fait que même si les paramètres sont mal estimés dans l'absolu, les distributions des deux classes sont bien séparées (dans le sens de la distance de Mahanalobis). Nécessairement on estime $\mu_1 = 0$ plus fort et $\mu_2 = 1$ plus petit que leur vraie valeur puisqu'on a censuré les valeurs petites et grandes. De même on a ainsi réduit la variabilité totale des données donc l'estimation de σ_1 et σ_2 est aussi biaisée vers le bas. On voit quand même que notre modèle arrive à se sortir de situations qui lui sont aussi défavorables tant que les données ne manquent pas en trop grand nombre. On touche cependant la limite où le modèle peut être appliqué avec confiance. Il conviendrait de modéliser le mécanisme d'absence des observations. Mais c'est un sujet difficile que nous n'aborderons pas...dans cette thèse.

- cas «nmar2» : on censure les deux classes différemment :
  $P(R_i = 0|\mathbf{x}^{Obs}, \mathbf{x}^{Man}, \mathbf{z}, \rho) = 1$ si $z_i = c_1$ et $(x_i < q^{c_1}_{\rho/2}$ ou $x_i > q^{c_1}_{1-\rho/2})$ ou si $z_i = c_2$ et $(x_i < q^{c_2}_{\rho/2}$ ou $x_i > q^{c_2}_{1-\rho/2})$ en notant q^{c_j} le quantile de la classe c_j. Ainsi $100 \times \rho/2\%$ des observations les plus petites et les plus grandes de chacune des classes manquent. On modélise ici en quelque sorte un appareil de mesure qui ne censurerait pas de la même façon des observations selon qu'elles sont issues d'une classe ou d'une autre. On pourrait penser comme exemple d'application au cas où une même mesure de fluoresence aurait à être effectuée sur deux rayonnements de couleurs différentes. Les limites de détection de l'appareil pourraient alors dépendre de la longueur d'onde. On se place dans un cas très bruité : $\sigma^2 = 1$.

  Certes les distributions sont modifiées mais les données restantes permettent de bien différencier les objets dans une classe ou dans l'autre bien que le bruit soit élevé. Cela explique certainement que les paramètres soient bien estimés (à l'exception des variances forcément diminuées par les procédures de censure) même pour de fortes valeurs de ρ. Les classifications restent acceptables et même très bonnes pour $\beta = 0,4$ tant que $\rho \leq 70\%$ (ceci est aussi visible à la figure 5.11 dans le cas favorable $\beta = 0,4$ où le caractère spatial commence à être marqué). Il semble même que des valeurs moyennes de ρ soit plus favorables en aidant à discerner les deux pics de la distribution qui sont majoritairement constitués de mesures d'objets d'une même classe.

Ces études sur données simulées nous ont permis en premier lieu de valider l'algorithme SF(miss) plus performant que les autres approches employées traditionnellement. Le critère BIC ou son approximation est un bon indicateur pour le

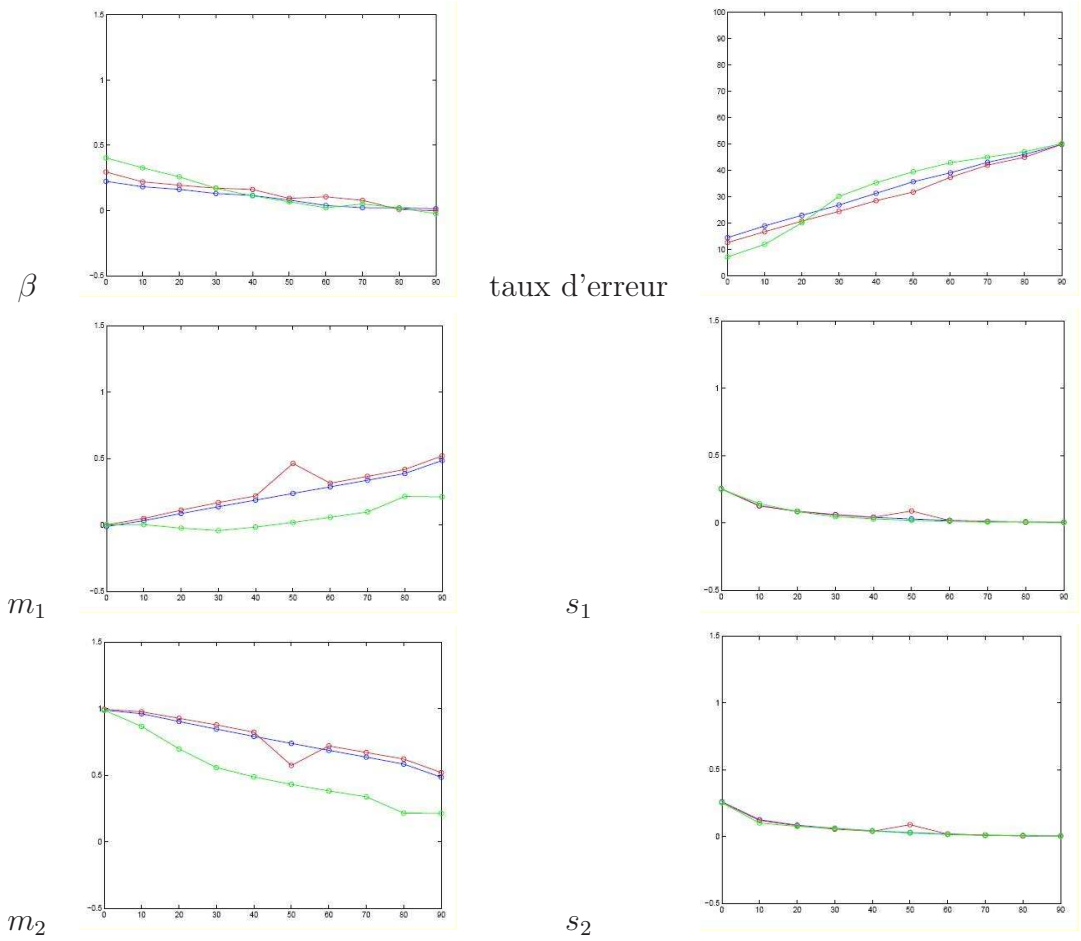


Fig. 5.14: Estimation des paramètres du modèle et taux de classification pour les trois simulations de 5.7 : $\beta = 0,2$ (bleu), $0,3$ (rouge) et $0,4$ (vert) dans le cas où la totalité des données manquantes sont NMAR censurées à gauche et à droite ($\sigma^2 = 0,25$).

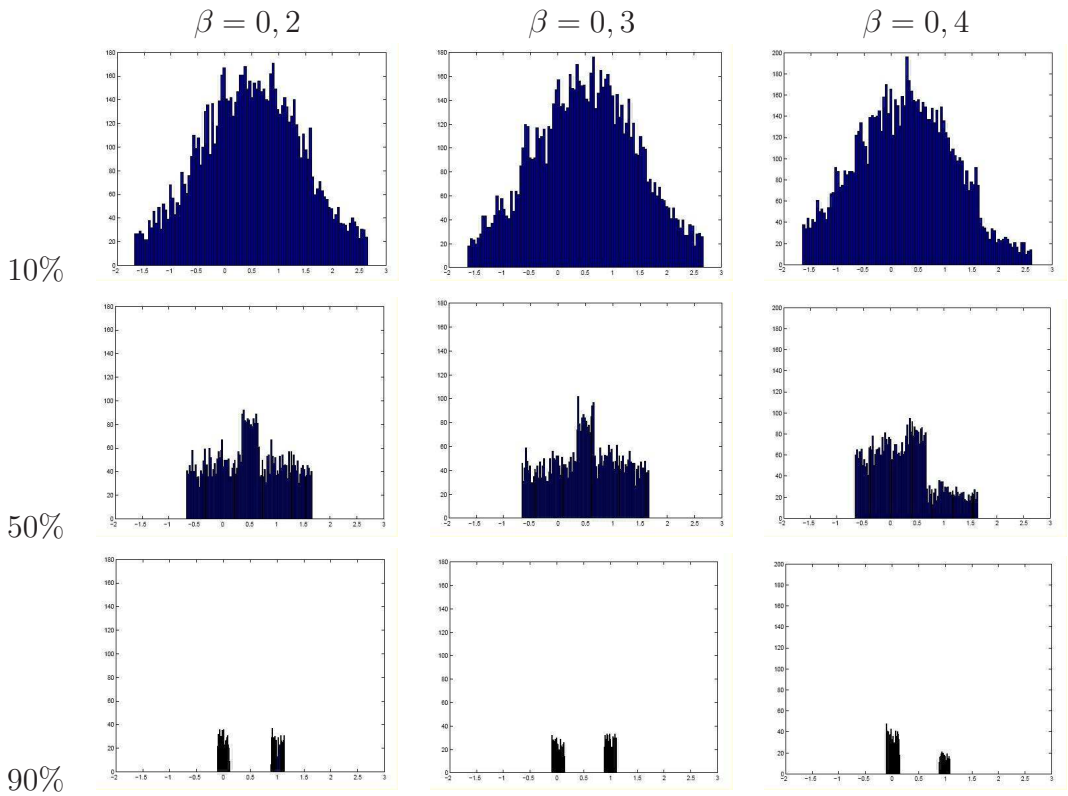


Fig. 5.15: Distribution des données simulées pour différentes proportions de valeurs manquantes dans le cas NMAR2 ($\sigma^2 = 1$).

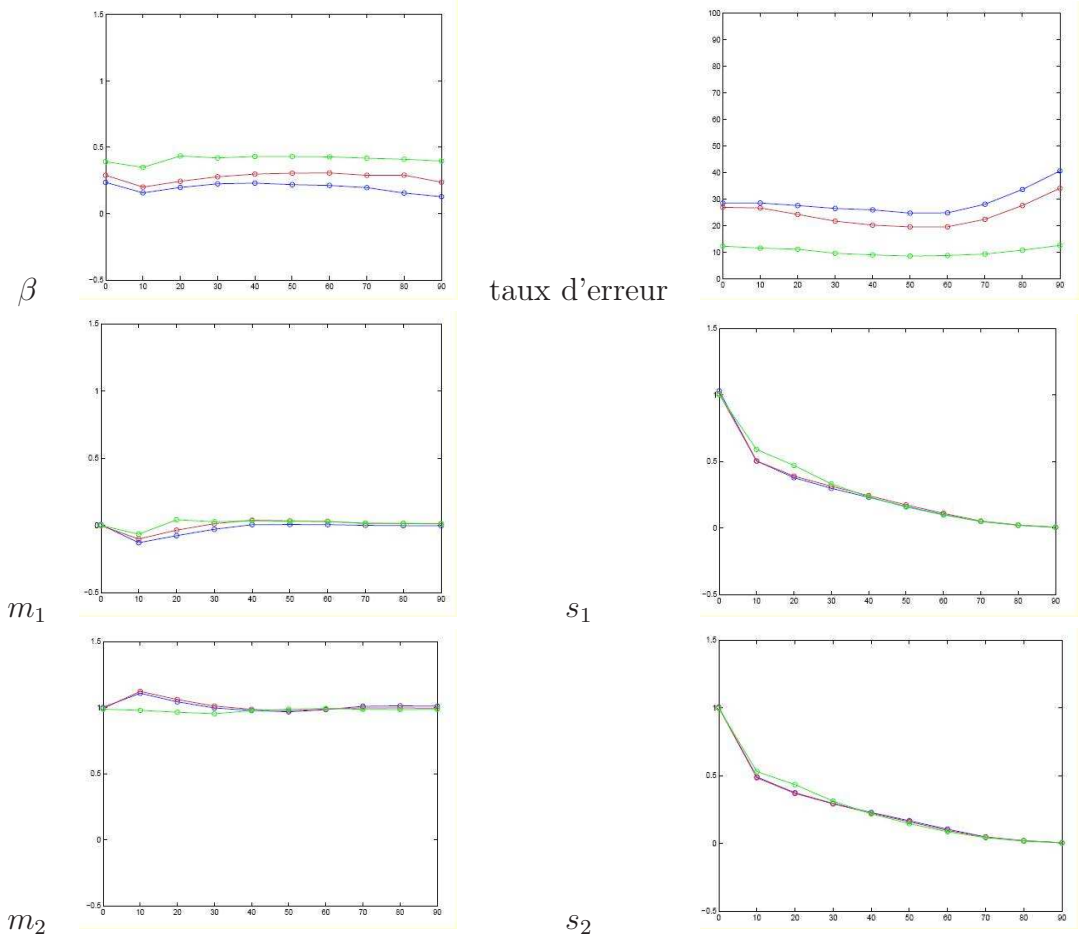


Fig. 5.16: Estimation des paramètres du modèle et taux de classification pour les trois simulations de 5.7 : $\beta = 0,2$ (bleu), $0,3$ (rouge) et $0,4$ (vert) dans le cas où la totalité des données manquantes sont NMAR censurées à gauche et à droite ($\sigma^2 = 1$).

choix de modèle, en particulier pour la sélection d'un nombre de classes K approprié. Enfin nous avons vu que dans des situations variées, avec des mécanismes de génération des données qui peuvent s'écarter des hypothèses de notre modèle, les performances restent bonnes : estimation des paramètres, taux d'erreur de classification.

5.2 Préparation des données réelles

Bien que notre méthode ne soit pas spécifique d'un organisme ni même d'un type de données particulier parmi celles décrites aux parties 2.1 et 2.2, nous avons dans le cadre de cette thèse fait porté le thème sur l'analyse conjointe de données d'expression avec intégration d'une information sur le réseaux des gènes pour la levure. Nous avons expliqué ces choix au chapitre 1. Vouloir analyser des données (post-)génomiques est une chose. Rendre cette envie concrète en est une autre, c'est une des leçons que nous avons retenu lors du déroulement de cette thèse.

Les données ne sont pas sous une forme directement utilisable. Nous avons du leur faire subir des pré-traitements et opérer de nombreuses vérifications. Le bénéfice a été grand en s'attaquant à cette difficulté : nous nous sommes formés aux bases de données biologiques. Et l'exploration de ces bases de données a aussi été l'occasion de comprendre petit à petit leur contenu ; nous prétendons avoir fait un peu de Biologie. Et pas sans le savoir comme aurait dit monsieur Jourdain ? !

Au début de cette thèse, un problème qui se posa fut de savoir gérer la taille des fichiers issus des techniques. Encore récemment, nous plantâmes plusieurs fois un PC récent en parcourant un fichier d'interactions entre protéines produit par la base de données STRING (`http://string.embl.de/`). Il faisait 2 Go pour la version gratuite...Nos résultats encourageants pourraient intéresser quelque informaticien disposant d'une grille de calculs.

Construction des réseaux

Nous avons déjà empiété sur ce paragraphe dans la partie page 36. Nous y avons exposé comment construire un réseau d'interactions entre gènes déduits du métabolisme d'une espèce. Ce travail a été exploité dans [82] et [235]. Rappelons-en les étapes essentielles avec les particularités auxquelles nous avons été confronté.

Deux réactions du fichier `reaction_main.lst` (téléchargeable sur l'accès FTP de la base KEGG) de la base LIGAND de KEGG sont voisines si elles ont un ou des composés communs. Les lignes de ce fichier se présentent sous la forme :

index-réaction : composant(s)-réactif(s) <=> composant(s)-produit(s)

sans que réactifs ou produits puissent être distingués ; les quantités stœchiométriques sont données. Nous avons donc créé une liste des composants pour chacune

des réactions. Puis en comparant ces listes, dès que l'intersection est non vide et que ces composés sont jugés pertinents, une arête est établie entre les deux réactions correspondantes.

Puis nous avons considéré le fichier listant toutes les réactions : `reaction`. Il a ses entrées sous la forme :

ENTRY	index-réaction Reaction
NAME	nom chimique de la transformation
DEFINITION	réaction écrites avec les noms courants
EQUATION	réaction écrite avec les index des composés
COMMENT	des commentaires
RPAIR	liste de paires de réactifs
PATHWAY	liste des voies métaboliques où cette réaction est rencontrée
ENZYME	la ou les enzymes catalysant la réaction
ORTHOLOGY	liste de gènes orthologues correspondant à une enzyme
///	

Il a donc fallu pour chaque enzyme dans le fichier `enzyme` dont le format est comparable à celui de `reaction`, lister toutes les réactions qu'elle catalysait. Deux enzymes sont alors reliées si et seulement si elles sont en correspondance avec deux réactions voisines à l'étape précédente.

Le passage du réseau d'enzymes au réseau de gènes se fait exactement de la même manière en repérant les gènes qui sont ceux de la levure (SCE).

La difficulté n'est pas tant dans le principe mais dans la manipulation effective de tous ces fichiers. Il faut tout d'abord connaître leur structure exacte pour ne pas commettre d'erreur. Ensuite il nous aura fallu programmer quelques scripts pour transformer les fichiers ci-avant en données de format compatible avec le logiciel développé dans l'équipe. Pour cela, nous avons majoritairement utilisé le logiciel (*G*)*awk* qui permet de parcourir un fichier en filtrant ligne par ligne.

Nous avons aussi utilisé dans [33] la base de données STRING (`http://string.embl.de/`) pour des interactions provenant de divers sources considérés comme des preuves. Chaque source donne un score à chaque interaction. La combinaison de ces scores donne un niveau de confiance global pour l'interaction d'une paire de gènes. Il s'agissait pour ces données de convertir le format des interactions entre les gènes pour qu'il soit accepté par notre programme.

Maintenant que nous avons décrit la façon dont nous avons récupéré et transformé les informations biologiques à notre disposition, nous allons présenter les résultats et les interprétations des expériences qui ont été entreprises.

5.3 Données réelles

Nous allons décrire ici deux manipulations majeures menées sur des données issues de la levure de boulanger (*Saccharomyces cerevisæ*). La première expérience a pour but de montrer le gain dans la prise en compte d'un réseau d'interactions

(métaboliques ici [120]) entre les gènes sur des données d'expression pendant la sporulation ([52]). La deuxième manipulation évaluera la performance de notre approche en traitant explicitement des observations de puces ADN de cycle cellulaire ([217]) manquantes avec la prise en compte d'un réseau d'interactions entre gènes obtenu dans la base de données STRING ([237]).

5.3.1 Prise en compte du réseau dans le modèle statistique complet

La classification de gènes est une technique exploratoire qui vise, par le regroupement de gènes en *clusters* homogènes à refléter des structures des données peu visibles *a priori*. L'information ainsi synthétisée permet de vérifier des connaissances préalables ou d'en inférer de nouvelles grâce aux connaissances partielles sur les éléments des groupes. On voudra adopter une approche intégrative. Plutôt que d'étudier séparément les acteurs d'un système complexe comme un organisme vivant, on voudra obtenir une vue globale. En effet, il semble en général vain dans un premier temps de chercher à décortiquer l'effet d'un gène précis. On cherchera plutôt à connaître ses interacteurs, les processus dans lesquels il joue un rôle. Les données post-génomiques sont à même de révéler ces relations. Les réseaux biologiques connectent alors les éléments du vivant et nous renseignent sur la complexité du système.

À titre d'illustration, nous nous sommes intéressés à l'étude d'un phénomène bien particulier : la sporulation (la gamétogénèse qui consiste en une méiose superposée à la formation de spores) chez la levure (organisme largement étudié et aux mécanismes bien documentés) en l'éclairant sous l'angle de son métabolisme. Quatre sous-ensembles distincts de gènes qui jouent un rôle dans ce processus étaient utilisés pour caractériser ce phénomène : précoces, moyens, moyens-tardifs et tardifs. [52] a montré que ce découpage était sous-optimal et lui préfère une description en 7 schémas d'expression temporels. Durant le déroulement de cette expérience, les changements de la concentration en ARNm furent mesurés pour les quelques 6118 ORF à sept instants après synchronisation (les cellules de levure furent transfèrées dans un milieu aux ressources azotées limitées qui induit la sporulation) : 0h, $0,5$h, 2h, 5h, 7h, 9h et $11,5$h. Trois points additionnels furent mesurés après qu'un facteur essentiel de la transcription connu pour être activé en fin de méiose soit réprimé ; la sporulation n'a alors pas lieu. Nous disposons donc pour ce jeu de données de profils de dimension $D = 10$ pour capter l'activité des gènes pendant la sporulation.

En ce qui concerne le réseau d'interaction, nous avons utilisé la base de données *Reaction* de KEGG comme décrit dans la partie 5.2. Le graphe résultant est composé de 635 gènes (seulement) qui sont connus et que l'on peut identifier automatiquement dans le réseau métabolique. Nous nous sommes donc restreints à l'analyse de ces 635 gènes pour voir l'importance jouée par la prise en compte du réseau. Le réseau compte 7111 arêtes et 92% des paires sont connectées. La figure 5.17 donne quelques caractéristiques du réseau que nous considérons. Celles-ci

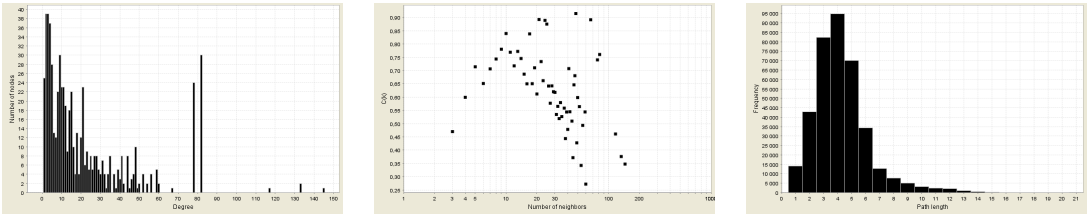
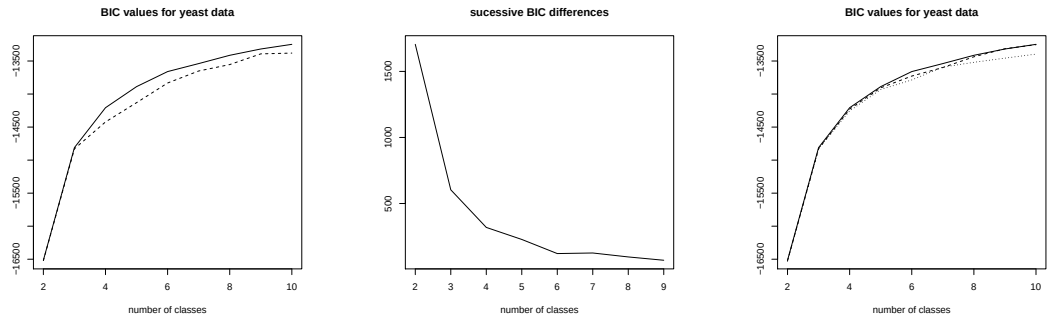


Fig. 5.17: À gauche : la distribution des degrés des nœuds du réseau. Au centre : les cœfficients de *clustering* moyen en fonction du nombre de voisins. À droite : distribution du chemin le plus court entre deux nœuds.

permettent de dire que le graphe que nous étudions n'a pas les propriétés théoriques de réseaux «pilotes» tels que les graphes exponentiels (la distribution des degrés ne suit visiblement pas une loi exponentielle) ou les graphes *small world* (le chemin le plus court entre deux points du graphe est parfois...un peu long). On pourra consulter [4, 5, 18, 116, 164, 172] pour des études de caractéristiques de graphes *aléatoires*, *small-world*, *scale-free* (ou exponentiels) ou *réguliers* et/ou celles des réseaux biologiques.

Le nombre correct de classes est inconnu. Nous avons donc calculé les valeurs de BIC pour $K = 2$ à $K = 10$ pour plusieurs modèles. Les courbes correspondantes à la figure 5.18 (a) ne montrent pas de maximum clair ni de coude (critère adopté par [27]). Nous avons considéré comme une valeur raisonnable de K, la valeur après laquelle deux différences successives de BIC n'étaient plus significatives (voir figure 5.18). Nous avons donc retenu $K = 6$; les analyses à suivre sont effectuées pour cette valeur du nombre de classes. Les résultats complets et toutes les données sont disponibles sur le site de matériel supplémentaire de l'article [235] (`http://mistis.inrialpes.fr/people/forbes/transparentia/supplementary.html`). Notons cependant que ce nombre est cohérent avec l'article qui porte sur les données [52] qui attendait entre 5 et 7 groupes se dégageant des données et [62] qui a choisi de travailler avec 7 *clusters*.

Nous avons ensuite comparé plus en détail les classifications issues des algorithmes EM standard et SF. La figure 5.19 donne une visualisation de la matrice de confusion entre ces deux classifications. Tandis que certains groupes restent très semblables voire identiques dans le cas des *clusters* SF et EM 5, il est parfois difficile d'identifier une correspondance. Par exemple le *cluster* EM 2 se coupe en deux *cluster* SF 1 and 6 ou bien le *cluster* SF 2 prend des éléments dans les *clusters* EM 3 et 6. En outre, le tableau 5.4 donne quelques caractéristiques descriptives parfois fort différentes des groupes SF. Leur connectivité interne varie beaucoup de même que la topologie des réseaux. Par exemple, malgré une très bonne connectivité interne, le groupe SF 2 a un diamètre de valeur forte pour un réseau à 86 nœuds : 12. Ses paires sont aussi très connectées (66%) comparées à celle d'un groupe comparable (groupe SF 4). Les tailles des groupes sont aussi



(a) (b) (c)

Fig. 5.18: Valeurs BIC pour le jeu de données de la levure avev K variant de 2 à 10. **(a)** Comparaison de SF et de EM. Ligne continue : algorithme en champ simulé, ligne interrompue : algorithm EM standard ; **(b)** Différences entre deux valeurs successives de BIC pour l'algorithme en champ simulé ; **(c)** comparaison de plusieurs modèles de champ de Markov pour l'algorithme SF. Ligne continue : matice d'interaction diagonale $B = b \times I_K$,Ligne interrompue : matrice B diagonale et ligne pointillée : matrice d'interactions B pleine.

k	1	2	3	4	5	6
taille du *cluster*	180	86	52	123	34	160
nombre d'arêtes internes	537	195	75	254	8	903
% paires connectées	14	66	17	14	1	16
diametre du *cluster*	8	12	5	8	2	8

Tab. 5.4: Quelques caractéristiques des *clusters* SF.

très différentes.

Évaluer la qualité intrinsèque d'une classification en l'absence de classification digne de confiance est une tâche difficile ; il n'existe pas de critère universel. En conséquence, nous avons choisi d'illustrer le principal avantage dans l'utilisation de notre méthode en pointant des caractéristiques biologiques spécifiques qui apparaissaient dans ses résultats. Un outil de visualisation de la classification dans son ensemble fait cruellement défaut. Aussi bien pour représenter les données que pour résumer l'information biologique contenue dans les données. De manière idéale, nous voulons vérifier que les *clusters* obtenus par notre approche ont une signification biologique renforcée par rapport à une méthode comparable n'utilisant pas l'information de réseau : le modèle de mélange. En revanche nous n'avons comparé notre classification ni à une classification obtenue par une utilisation sous-optimale du réseau (par exemple en modifiant les classes une fois la classification terminée avec de l'information *a priori* comme dans [216]) ni à une classification obtenue en intégrant différentes sources d'information dans un traitement des données que nous jugeons aussi sous-optimal (comme par exemple

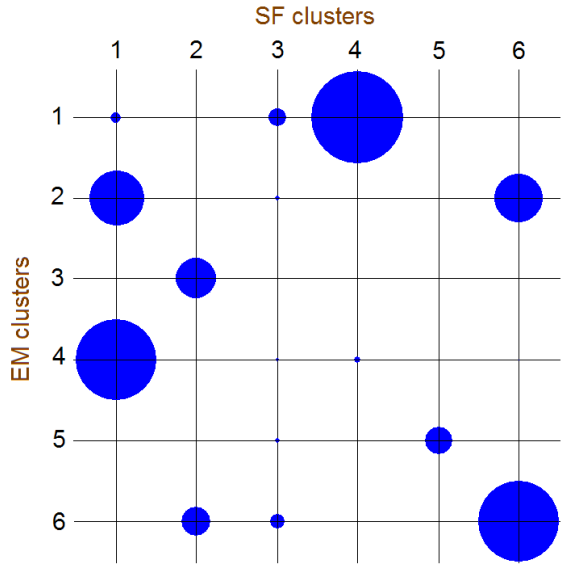


Fig. 5.19: Visualisation de la matrice de confusion entre les classifications EM et SF ; le rayon du cercle est proportionnel au cardinal de l'intersection des *clusters*. Les choix des étiquettes n'ont aucune autre signification particulière que d'avoir été attribuées par les algorithmes de classification.

la transformation de tous les types de données en distances dans [98] pour ensuite combiner toutes ces distances puis d'appliquer un *clustering* hiérarchique par lien moyen). En effet, nous n'avons pas pu disposer des outils et des méthodes développés par les auteurs de ces méthodes malgré des demandes faites dans ce sens.

[138] propose une méthode pour détecter les voies métaboliques qui jouent un rôle significatif en regard des données d'expression. Par exemple, ils ont analysé les données de [52] que nous utilisons ici. Trois fonctions de score sont décrites pour caractériser les voies métaboliques au niveau transcriptionnel, de corégulation ou d'effet cascade. De grands scores d'activité (*Activity Scores*) récompensent des voies métaboliques qui contiennent de nombreux gènes sur- ou sous-exprimés (par rapport à des seuils). Cela signifie que de nombreux gènes de cette voie métabolique voient leur expression modifiée du fait de la condition expérimentale. Le score de corégulation (*Corregulation Score*) est d'autant plus élevé pour une voie métabolique que les gènes qu'elle contient ont des profils d'expression similaires. On peut donc interpréter l'activité de ces gènes comme concertée par un mécanisme régulateur. Enfin le score d'effet cascade (*Cascade Score*), prend en compte les gènes qui n'ont pas de déviance prononcée par rapport à la condition initiale et par rapport à la structure et à l'ordre des réactions dans la voie métabolique. En particulier, ils sont utiles pour mettre en évidence qu'une réaction en chaîne

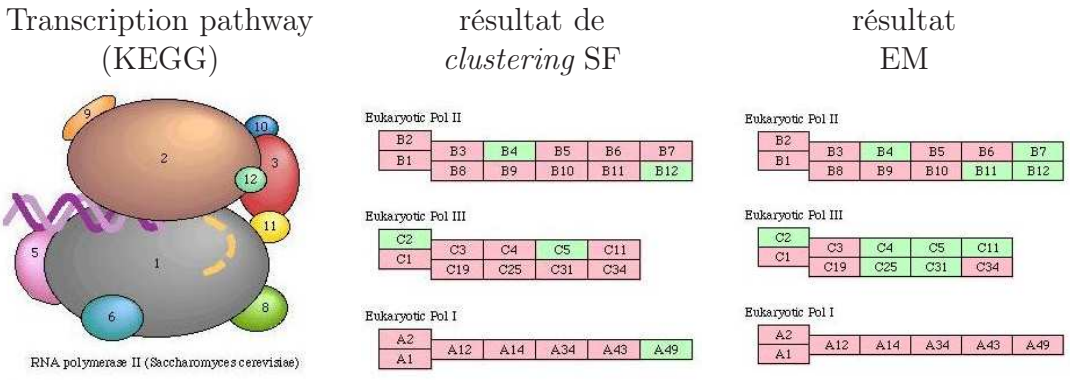


Fig. 5.20: Schéma d'ARN polymérase de la voie métabolique «transcription» donnée dans KEGG. Les colonnes du milieu et de droite donnent les résultats obtenus en regroupant deux *clusters* selon qu'on utilise respectivement les algorithmes en champ simulé et EM standard. Les protéines colorées en rose sont celles incluses dans ces *clusters* tandis que celles en vert sont soit dans un autre *cluster* soit non présentes dans le jeu de données. L'approche en champ simulé est plus performante qu'EM pour grouper cette classe fonctionnelle de gènes. Aucun gène ne correspond à B12 dans notre jeu de données bien qu'il existe au moins un gène de la levure associé à cette protéine.

particulière est activée ou éteinte dans une voie métabolique.

Par exemple, les voies métaboliques des transcriptions et traductions ont de hauts scores d'activité. L'algorithme SF capte bien ces éléments : 16 des 28 gènes de la voie de la transcription dans notre jeu de données sont regroupés dans un seul *cluster* contre 11 au mieux pour EM standard. Si on autorise le groupement de deux clusters, on retrouve 24 gènes regroupés (presque tous) contre 19 pour EM (voir la figure 5.20).

Dans le même esprit, nous pouvons nous intéresser au métabolisme des vitamines qui obtient un bon score d'effet cascade pour notre jeu de données. SF rassemble 26 de ces gènes dans un même *cluster* contre seulement 19 pour EM. En regroupant deux *clusters*, on monte à 44 des 70 gènes dans le jeu de données rassemblés par SF pour 35 avec EM. Un dernier exemple est la voie métabolique de phosphorilation oxydante qui obtient un bon score de corégulation dans [138] pour les données de [52]. Notre méthode trouve 24 gènes dans le *cluster* SF 6 tandis qu'EM en groupe au plus 16 parmi les 52 impliqués dans notre jeu de données dans cette voie. En outre, tous ces gènes détectés sont régulés *up* au deuxième instant de mesure ; ils sont spécifiques de la synthèse de l'ATP. D'autres voies métaboliques comme la glycolyse sont bien captées par l'algorithme en champ simulé ; 24 gènes de la glycolyse sont regroupés contre 19 sur 44 par EM. De plus, les gènes manquants ne gènent pas l'ouverture de certaines voies de synthèse complètes (voir la figure 5.21). Il s'agit du *cluster* SF 2 impliqué dans les

mécanismes des carbohydrates (voir la tableau 5.5). Notre méthode groupe ainsi les gènes $YLR153C$ (EC 6.2.1.1) et $YPL061W$ (EC 1.2.1.3) avec les gènes de la glycolyse malgré leur dissimilarité d'expression avec le comportement global des gènes de cette voie (sur-expression rapide des gènes de la conversion du glucose 6-phosphate en pyruvate ou inversement) en ayant une augmentation douce de leur expression. Tous ces résultats soutiennent la meilleure sensibilité de l'algorithme en champ simulé par rapport au modèle de mélange standard. Des gènes fonctionnellement liés sont ainsi bien groupés même si la seule donnée de l'expression d'ARNm n'était pas un indice suffisant.

Pour continuer la validation des résultats, nous avons considéré l'analyse ontologique (par vocabulaire contrôlé et classé en hiérarchies) des classes en termes GO (*Gene Ontology* `http://www.geneontology.org/`). GO est représenté comme un graphe orienté avec les classes fonctionnelles (ou attributs moléculaires des gènes) aux nœuds et où une arête orientée signifie qu'une classe (en aval) est incluse dans une autre (en amont). Les termes situés aux feuilles de la hiérarchie GO (précis donc) sont associés à des gènes des organismes pour lesquels l'information est disponible. Nous nous sommes servis des 1935 termes disponibles au moment de notre étude. Parmi ces termes, 1016 sont classés dans la branche «processus biologiques» (*biological process*). Deux autres branches principales existent : «fonction moléculaire» et «localisation cellulaire». La première décrit dans quelles transformations un gène est impliqué au niveau de la cellule. L'autre situe le produit de gènes dans la cellule. Les processus biologiques décrivent comment un gène contribue à un objectif biologique. De nombreuses autres classifications fonctionnelles existent (*FunCat* [198] du MIPS, les KOG [223]). GO semble être une des plus utilisées pour la validation de classification de gènes par rapport à l'enrichissement ou l'appauvrissement de certaines catégories fonctionnelles : [13, 36, 40, 87, 126, 191, 193]. Plus on observe une sur-représentation de termes GO trouvés dans le jeu de données dans les groupes issus d'une classification, plus celle-ci est sensible. Plus les groupes de cette classification isolent des termes, plus la méthode est sélective.

Nous avons testé la sur- et la sous-représentation des termes GO. Les P-valeurs ont été calculées avec le FDR (*False Discovery Rate*) qui est le contrôle de la proportion de rejets erronés parmi tous les rejets de l'hypothèse nulle (pas de sur-/sous-représentation donc distribution hypergéométrique des catégories). Ainsi le problème du grand nombre simultané de tests est pris en compte. En outre, nous avons ici choisi d'autoriser des dépendances arbitraires entre les catégories GO testées. La méthode est décrite dans [23]. Le tableau 5.5 résume quelques résultats intéressants de sur-représentation de termes GO dans les *cluster* SF avec les P-valeurs correspondant. Il donne aussi la meilleure P-valeur obtenue parmi les composants du mélange dans le cadre du *clustering* par EM standard. Cette façon de comparer ne peut pas avantager SF. En général, des termes sur-représentés dans un *cluster* sont sous-représentés dans les autres. Mais les P-valeurs ne sont pas souvent très significatives.

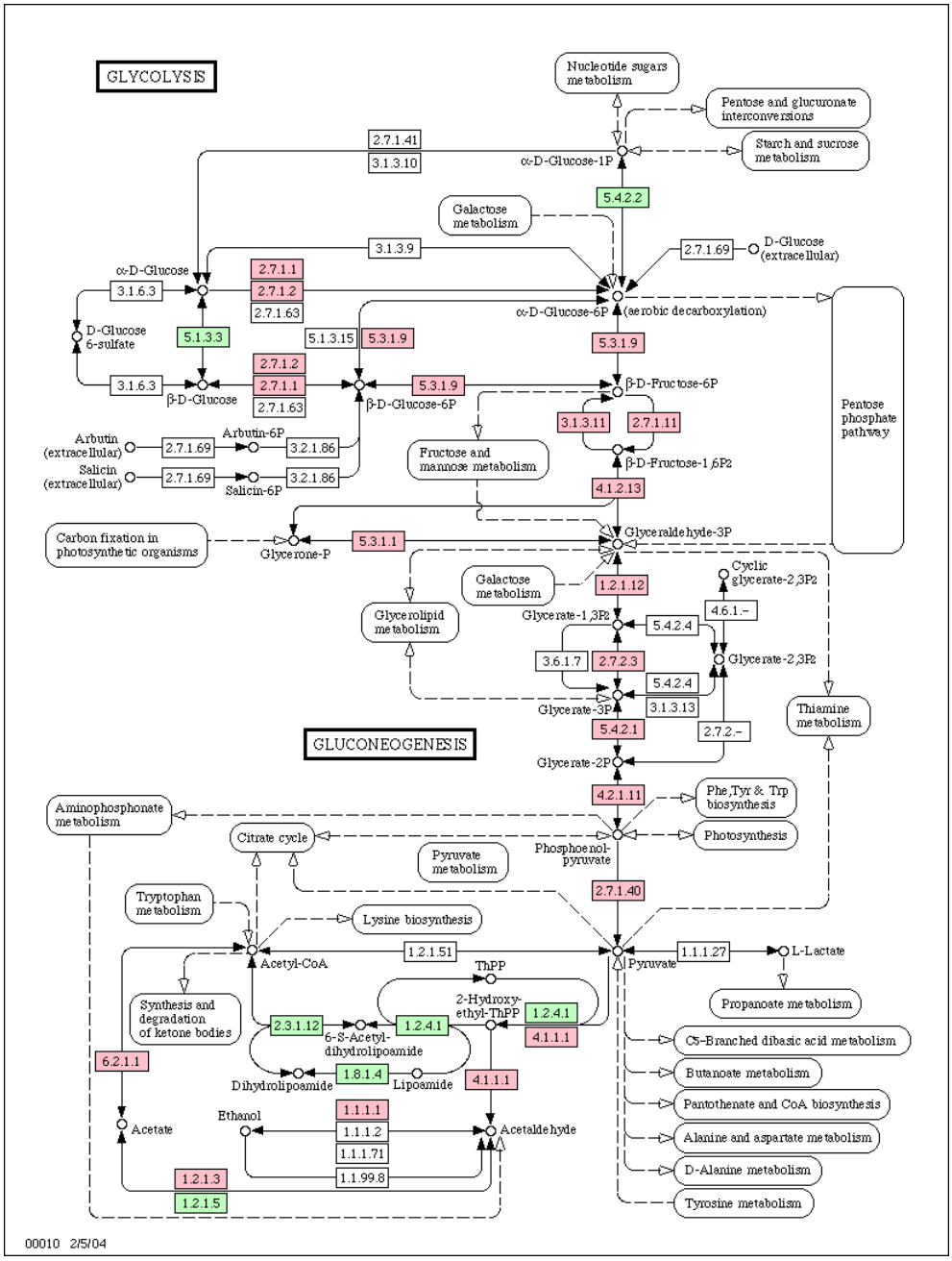


Fig. 5.21: Voie métabolique de la glycolyse : les EC colorés sont dans notre jeu de données ; les roses sont dans un même cluster SF tandis que les verts sont répartis dans les autres.

Cluster SF	Termes GO	P-valeurs SF	P-Valeurs EM standard
1	GO :0008652 : amino-acid biosynth.	**1.1E-2**	0.193
2	GO :0006006 : glucose metabolism	**1.2E-7**	8.7E-7
	GO :0006090 : pyruvate metabolism	5.9E-5	**8.7E-7**
	GO :0006144 : purine base metabolism	**2.2E-2**	0.259
	GO :0015980 : energy dev. by oxid...	**1.8E-2**	3.3E-2
3	GO :0006259 : DNA metabolism	**4.1E-2**	1
	GO :0006261 : DNA-dep. DNA replic.	**4.1E-2**	0.193
	GO :0006271 : DNA strand elong.	**4.1E-2**	0.208
4	no significant GO term	N.A.	N.A.
5	GO :0030437 : sporulation	**4E-2**	0.26
6	GO :0006360 : transcr. from RNA pol.	**1.6E-2**	2.5E-2
	GO :0006164 : purine nucleo biosynt.	**2.0E-2**	6.1E-2

Tab. 5.5: Analyse ontologique des *clusters* issus SF et EM : les catégories GO sur-représentés et les P-valeurs correspondantes sont données.

L'analyse ontologique est cohérente avec les observations précédentes sur les voies métaboliques. En outre, cela soutient toujours la conclusion que l'approche en champ simulé fournit une meilleure classification (plus spécifique) du point de vue de l'interprétation biologique. Seuls les termes GO avec un nombre raisonnable (de l'ordre de la dizaine) de représentants ont été reportés dans le tableau 5.5. Les catégories GO avec de trop nombreux gènes sont peu informatives sur la capacité de la méthode à distinguer des processus spécifiques dans une classe. À l'inverse, des catégories avec un trop petit nombre de termes (typiquement de 2 à 4) ne fournissent à notre sens pas une réelle validation pratique.

À titre d'exemple, les gènes classés dans la catégorie $GO : 0008652$: biosynthèse des acides aminés et dans le *cluster* SF 1 montre une P-valeur significative de $1,1E-2$ alors qu'elle ne l'est pas du tout pour EM $(0,193)$. Le *cluster* SF 5 est encore plus pertinent. Il contient la plupart des gènes spécifiques de la sporulation ($GO : 0030437$) repertorié dans [52]. La conclusion du test est que ce *cluster* est bien spécifique de la fonction en question avec une P-valeur de 4%. En comparaison, la meilleure P-valeur parmi les groupes EM est de $0,26$ ce qui ne mène pas à la conclusion que ce terme est sur-représenté avec des seuils acceptables usuels. Cela peut sembler surprenant puisque ces gènes ne sont reliés par aucune des associations (co-expression, localisation sur le chromosome, fusion, expérience de paillasse, co-occurrence) envisagées dans la base de données STRING [237] sauf le lien par fouille de textes qui retrouve évidemment le fait que ces gènes sont co-cités dans [52]. Ainsi il semble que l'algorithme SF sache capter de l'information de réseau même quand une partie est manquante : 32 des 34 gènes du *cluster* 5 jouent un rôle dans la sporulation d'après le papier [52] sans inclure la classe des gènes «métaboliques» (à induction rapide) qui sont presque tous dans un autre

(grand) *cluster*. Cela permet d'une part d'émettre une hypothèse à vérifier pour un biologiste en ce qui concerne les deux gènes du *cluster* 5 qui n'étaient pas précedemment connus comme jouant un rôle dans la sporulation. D'autre part, si les gènes de la classe temporelle «métabolique» sont séparés, c'est fort probablement qu'ils ont une régulation toute différente.

Ces résultats préliminaires sont intéressants par plusieurs aspects. L'algorithme en champ simulé produit des groupes de gènes biologiquement satisfaisants et meilleurs que ceux du modèle de mélange qui ne se base que sur les données d'expression. En comparaison des méthodes usuelles pour l'incorporation de données d'interactions post-génomiques, il a l'avantage d'être un modèle statistique bien fondé qui ne demande pas le choix *a priori* d'une distance. Il permet aussi de répondre à la question du choix du modèle (grâce à un critère comme BIC). Il fournit aussi une classification qui se base sur des probabilités d'appartenance des gènes à une classe, pas une classification dure (d'habitude plus biaisée). Enfin il donne un cadre adéquat pour le traitement des observations individuelles manquantes. Nous allons dans la partie suivante 5.3.2 présenter notre analyse d'un autre jeu de données réelles contenant des «trous» dans la matrice d'expression.

5.3.2 Modèle prenant en compte les observations manquantes

Les données ont été préparées de façon similaire à celle décrite à la partie 5.2. L'étude de [217] qui s'intéresse aux expressions de gènes lors du cycle cellulaire chez la levure a été utilisée. Le jeu de données initiales contient plusieurs profils d'expression de cultures de levures synchronisées par différentes manipulations : elutriation, arrêt par α-facteur ou par allèle particulier aux gènes CDC15 ou CDC28 rendant la levure sensible aux différences de température. Il inclut aussi les données par induction Clb2 et Cln3. Le jeu de données total contient 6179 gènes, chacun d'eux renseigné par un profil d'expression de dimension $D' = 77$. Ce jeu de données contient environ 5% des observations manquantes sans que ce fait soit réellement documenté. Nous avons essayé de l'analyser entièrement mais le fait que la prise en compte des grandes dimensions ne soit pas encore totalement opérationnelle dans le modèle n'a pas permis d'avancer dans les résultats. Les modèles diagonaux ne sont pas suffisants pour capter une information aussi riche. Aussi il a été observé que la méthode de réduction de dimension que nous envisageons ne se prête pas très bien aux séries temporelles.

Dans la suite, nous nous sommes donc concentrés sur l'expérience CDC28 initialement menée par [51] et utilisée par [217] pour une analyse de concert avec leurs propres données. Les levures furent synchronisées après un arrêt en phase G1 du cycle cellulaire ; la température fut montée à $37°C$ et le cycle cellulaire redémarra lorsque la température revient à $25°C$. $D = 17$ mesures furent collectées toutes les dix minutes en laissant donc ainsi assez de temps à presque deux cycles cellulaires de se dérouler. Nous avons choisi cette expérience parce que la synchronie y fut contrôlée avec un grand soin et pour simplifier la présentation

des résultats par rapport au jeu de données complet sur lequel nous avons aussi effectué l'analyse. Dans ce sous-ensemble, il y a aussi environ 5% des observations des niveaux d'ARNm manquants.

Beaucoup de gènes impliqués dans le cycle cellulaire ne s'expriment différentiellement qu'une fois pendant le cycle. Les processus dans lesquels ils jouent un rôle incluent la synthèse, la réplication, l'entretien, *etc.* de l'ADN, la «germination» (développement n'est pas assez précis), l'alimentation et la mitose par exemple. De plus, une bonne partie de ces gènes contrôlent (par boucle de rétroaction) le cycle cellulaire mais la liste de ceux dont la transcription est nécessaire n'est pas arrêtée. [217] fournit une liste de 800 gènes identifiés comme régulés «par» le cycle cellulaire grâce à une analyse de Fourier de leur profil d'expression. Ils affichent cinq groupes correspondant à différentes phases du cycle cellulaire en regard des instants où un pic d'expression est observé. Cette classification a été obtenue par classification hiérarchique ascendante et est reconnue comme un bon standard pour l'évaluation de résultats de *clustering*. Nous nous sommes donc concentrés sur l'analyse de ces 800 gènes.

Parmi les nombreux réseaux biologiques disponibles, nous avons ici porté notre choix sur le réseau complet issu de la base de données STRING version 7 (`http://string.embl.de`, [237]). Il s'agit d'une base de données d'interactions déterminées ou prédites entre protéines. La partie consacrée à la levure contient 401 948 interactions sur 5611 gènes. Une même paire de gènes peut refléter des interactions issues de sources différentes. Nous n'avons pas filtré les interactions selon leur niveau de confidence. L'intersection entre les 800 gènes appartenant à une classe temporelle de [217] et les gènes ayant une information de réseau dans STRING se solde par un ensemble de $N = 612$ gènes aux nœuds d'un graphe à 3530 arêtes. Le graphe entier est visible dans la partie gauche de la figure 5.24. Nous avons utilisé Cytoscape (`http://www.cytoscape.org/`) comme outil de visualisation du réseau biologique. Ici aussi, nous ne présenterons pas tous les résultats mais certains passages significatifs. La totalité est disponible sur le site de matériel supplémentaire de l'article [33] à l'adresse `http://mistis.inrialpes.fr/people/blanchet/supbibe.html`.

Nous commençons par déterminer la classification qui reflète le mieux la structure des données au sens du critère BIC. Ses valeurs sont calculées pour différents algorithmes et tracées à la figure 5.22. Un maximum est visible pour $K = 9$ et le modèle SFmiss est préféré parmi tous ceux testés. On voit qu'il y a un gain net dans la prise en compte du réseau (algorithmes basés sur SF) par rapport au modèle de mélange généralisé pour la prise en compte des valeurs manquantes dans les profils d'expression (EM miss). Aussi notre approche obtient de meilleurs valeurs de l'indice BIC que les approches avec imputation préalable par les k plus proches voisins (*KNNimpute* de [231]) ou par modèle auto-regressif d'ordre 1 pourtant bien adapté à une série temporelle. Passons maintenant à l'analyse proprement dite du contenu des *clusters* SFmiss en les comparant aux autres

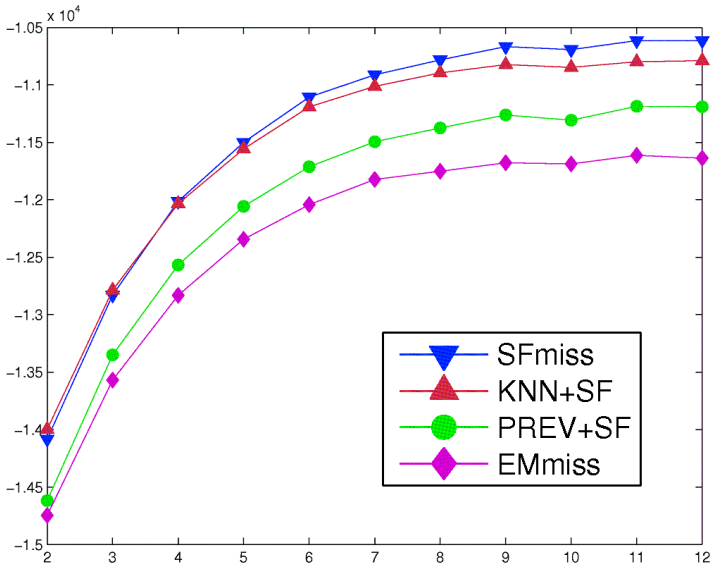


Fig. 5.22: Valeurs BIC en fonction du label de classe pour les algorithmes SFmiss, KNN+SF, PREV+SF (algorithme SF avec imputation préalable par un modèle auto-régressif d'ordre 1) et EMmiss.

classifications lorsque cela est nécessaire. Nous essaierons ici encore de dégager des caractéristiques biologiques intéressantes qui démontrent l'amélioration de l'interprétabilité des résultats grâce à notre approche. La figure 5.23 donne une visualisation qui permet de juger des différences entre les classifications issues de SFmiss et EMmiss.

Une première tendance générale est que SFmiss regroupe davantage les gènes qui ont tendance à interagir que les autres algorithmes. Cela est illustré dans la partie de droite (zoom) de la figure 5.24. Cela est cohérent avec l'estimation à $\beta = 0,41$ du paramètre spatial : le réseau est bien pris en compte.

La fiabilité des *clusters* peut être évaluée à l'aide de l'analyse ontologique des catégories de *Gene Ontology* (GO, `http://www.geneontology.org`). On voudra connaître le contenu des *clusters* et savoir s'ils permettent de distinguer des catégories fonctionnelles. *GOstat* (`http://gostat.wehi.edu.au/`) nous a servi pour l'analyse ontologique des termes GO contenus dans les groupes des classifications. Nous avons considéré les catégories des trois branches principales : processus biologiques, localisation cellulaire et fonction moléculaire parce que les termes de ces trois branches apportent des visions complémentaires du processus biologique que nous étudions : le cycle cellulaire chez la levure. Plus un terme présent dans le jeu de données a de représentants dans un même *cluster*, plus la méthode de clustering est sensible. Plus les 9 *clusters* isolent les différentes catégories GO, plus la méthode sera spécifique.

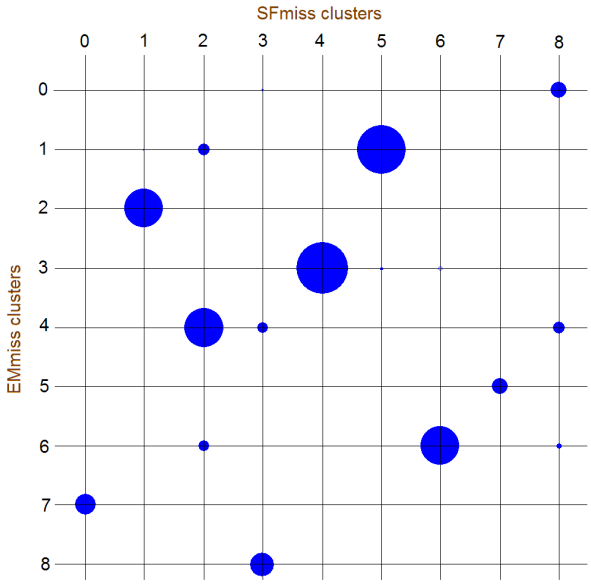


Fig. 5.23: Visualisation de la matrice de confusion entre les classifications EMmiss et SFmiss pour $K = 9$; le rayon du cercle est proportionnel au cardinal de l'intersection des *clusters*. Les choix des étiquettes n'ont aucune autre signification particulière que d'avoir été attribués par les algorithmes de classification.

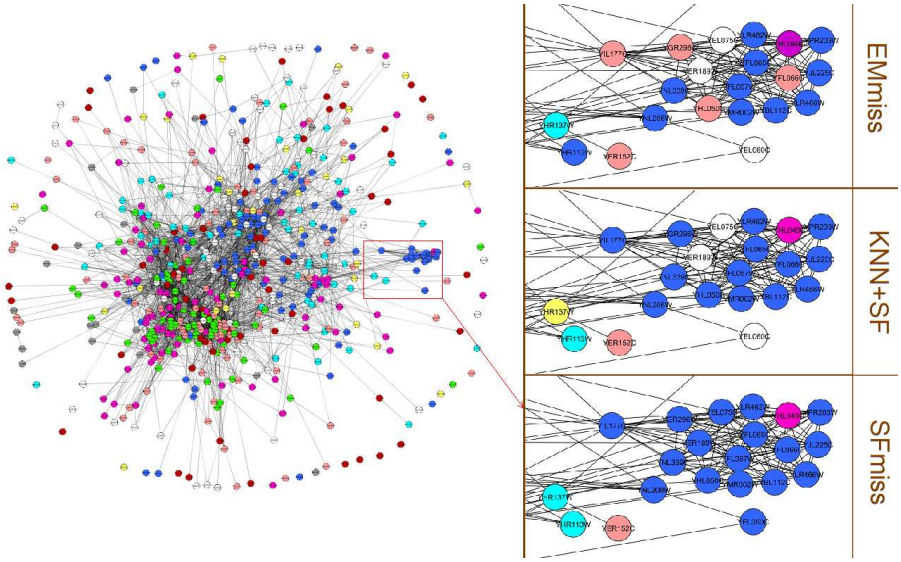


Fig. 5.24: À gauche : Réseau complet avec les nœuds colorés selon les labels SFmiss. À droite : zooms sur la partie encadrée pour les différents algorithmes.

Le tableau 5.6 présente les résultats qui nous permettent de conclure que SFmiss est légèrement plus sensible (il détecte mieux des *clusters a priori*) que les autres algorithmes étudiés. Nous parlerons de sa spécificité (capacité à séparer des gènes qui n'ont pas de point commun connu évident) un peu plus bas (analyse des classes temporelles fournies par [217] par exemple). Comme toutes les catégories présentes dans le jeu de données sont testés pour une sur- ou sous-représentation, plusieurs centaines voire quelques milliers de tests sont effectués en parallèle. Parmi les corrections à notre disposition pour tenir compte de la multiplicité des tests et des dépendances entre les catégories GO, nous avons ici retenu le FDR de [22] pour le calcul des P-valeurs. Une hypothèse de dépendance de régression positive entre les catégories GO. Le fait qu'elles vérifient effectivement cette hypothèse est une question ouverte. Ce problème peut être résolu en intégrant des dépendances générales entre les catégories testées dans le calcul des P-valeurs ([23]). Mais cette procédure montre une puissance de test nettement moindre ; c'est le prix à payer pour la prise en compte de dépendances arbitraires !

Hormis pour de rares exceptions, le modèle SFmiss montre de meilleurs performances que les autres algorithmes. Des gènes annotés similairement sont mieux groupés par SFmiss. Le cas du *cluster* $k = 8$ est une situation particulièrement dévaforable où le *cluster* le plus semblable contient 43 éléments alors que le *cluster* SFmiss 8 en contient 43. Ainsi une légère amélioration dans le regroupement d'objets à fonction similaire n'implique pas nécessairement une plus petite P-valeur. Nous pouvons détailler l'exemple des gènes $YDL105W$, $YER111C$, $YKR077W$, $YJL196C$, $YLR212C$ et $YNL082W$ qui sont regroupés dans le *cluster* 1 SFmiss et dans aucun *cluster* EMmiss. Tous jouent un rôle dans les processus du cycle cellulaire : réparation du complexe de fuseau mitotique ($YLR212C$), transition G1/S du cycle mitotique ($YER111C$) par exemple. $YKR077W$ est quant à lui annoté comme activateur putatif de transcription. Notre méthode suggère que cette information est cohérente et que ce gène joue sûrement un rôle-clé dans la régulation du cycle cellulaire ou en tant que gène régulé par ce processus. Ainsi notre méthode ne sert pas qu'à résumer des connaissances biologiques déjà établies mais donne des indications sur les fonctions possibles de gènes en tant que composants d'un système complexe : un organisme et ses interactions. Nous pouvons aussi illustrer ces conclusions de résultats de *clustering* en comparant les résultats dans le cas de l'imputation préalable par l'algorithme *KNNimpute* de [231]. Les gènes $YBL002W$, $YGL093W$ et $YPL269W$ sont tous les trois dans le *cluster* 4 SFmiss et ne sont ensembles dans aucun *cluster* KNN+SF. Leurs annotations respectives (assemblage de la chromatine, nécessaire à la localisation précise en vue du désappariement du chromosome du coté nucléaire du fuseau et nécessaire pour l'orientation-polarisation des microtubules chez la levure) font bien sens par rapport à celle déjà vue pour le *cluster* entier (tableau 5.6) et confirme une description fonctionnelle du *cluster*. Les nombreuses références biologiques pour ce qui prend ici seulement l'allure d'affirmations sont disponibles dans la bibliogaphie complémentaire de [33].

cluster SFmiss	termes GO & p-valeurs correspondantes	meilleures p-valeurs parmi les *clusters* EMmiss	meilleures p-valeurs parmi les *clusters* KNN+SF
k=3	GO :0006732, coenzyme met. process		
	$\mathbf{1.1\,10^{-2}}$	> 0.1	> 0.1
k=4	GO :0005819, spindle		
	$\mathbf{4.6\,10^{-9}}$	$6.7\,10^{-7}$	$2.0\,10^{-6}$
	GO :0006790, sulf. met. process		
	$\mathbf{1.1\,10^{-4}}$	$2.4\,10^{-4}$	$8.7\,10^{-4}$
	GO :0000278, mitotic cell cycle		
	$\mathbf{2.2\,10^{-3}}$	$7.7\,10^{-3}$	> 0.1
	GO :0030472, mit. spin. org. & biogen. in nucleus		
	$\mathbf{5.2\,10^{-3}}$	$8.8\,10^{-3}$	$2.0\,10^{-2}$
k=5	GO :0006974, resp. to DNA dam. stim.		
	$\mathbf{1.8\,10^{-3}}$	$3.0\,10^{-3}$	$8.0\,10^{-3}$
	GO :0000724, dbl-str. bk rep. via hom. comb.		
	$\mathbf{1.9\,10^{-2}}$	$2.7\,10^{-2}$	$4.6\,10^{-2}$
	GO :0000030, mannosyltransf. act.		
	$\mathbf{1.1\,10^{-2}}$	$1.2\,10^{-2}$	$2.7\,10^{-2}$
k=8	GO :0042555, MCM cplx		
	$\mathbf{3.4\,10^{-4}}$	$8.3\,10^{-4}$	$4.0\,10^{-4}$
	GO :0008026, ATP-dep. helicase act.		
	$5.5\,10^{-4}$	$1.3\,10^{-3}$	$\mathbf{4.5\,10^{-4}}$
	GO :0006268, DNA unwind. replic.		
	$2.8\,10^{-3}$	$6.7\,10^{-3}$	$\mathbf{1.1\,10^{-3}}$
	GO :0042623, ATPase act. coupl.		
	$\mathbf{4.4\,10^{-3}}$	$1.5\,10^{-2}$	$4.3\,10^{-2}$

Tab. 5.6: Quelques termes GO représentatifs des *clusters* obtenus par les modèles testés.

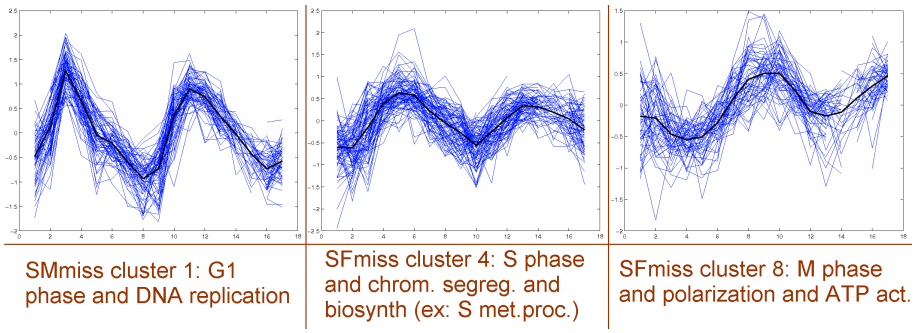


Fig. 5.25: Exemples de profils d'expression moyens de trois *cluster* SFmiss.

	$k=0$	$k=1$	$k=2$	$k=3$	$k=4$	$k=5$	$k=6$	$k=7$	$k=8$
M/G1	0	2	28	27	0	2	1	11	13
G1	0	69	34	14	20	84	10	9	0
S	1	0	7	1	29	5	13	4	0
G2	2	0	14	3	38	0	31	0	0
G2/M	33	0	18	16	1	0	20	6	46

Tab. 5.7: Classification croisée des *clusters* SFmiss comparés avec les classes temporelles du cycle cellulaire de la levure identifées par [217] : G1, S, S/G2, G2/M et M/G1

Un autre aspect plaisant des résultats du modèle SFmiss est que les *clusters* sont bien interprétables en terme des classes temporelles identifiées par [217] : G1, S, S/G2, G2/M et M/G1. Ainsi le *cluster* 0 est presque entièrement inclus dans la classe G2/M et les *clusters* 1 et 5 sont dans G1. Le *cluster* 8 est plus amplement concerné par des gènes actifs en phase M. Ces remarques sont soutenues à la fois par les profils moyens des groupes (quelques uns sont représentés dans la figure 5.25) et la classification croisée entre classes temporelles et classification SFmiss (tableau 5.7). D'autres *cluster* sont plus difficilement interprétables parce que diffus sur les classes temporelles. Même si par exemple le profil moyen du *cluster* 4 bien que peu marqué semble bien cyclique et que la plupart de ses gènes semblent centrés autour de la phase S. Toutes ces propositions de classifications faites avec l'appui des profils moyens sont en accord avec les profils moyens de [51], figure 4 (C, D) qui sont données pour des gènes déjà annotés comme jouant un rôle précis dans les phases du cycle cellulaire.

Enfin un dernier aspect intéressant de SFmiss que nous présenterons est sa grande stabilité vis-à-vis de la diminution du nombre d'observations disponibles. La figure 5.26 illustre ceci. Les données manquantes supplémentaires ont été générées sous l'hypothèse MCAR. Même si SFmiss est meilleur, tous les algorithmes ont des mauvaises performances dans le cas NMAR que nous avons mis en place. Nous avons censuré les données et l'analyse des données d'expression devient rapidement très difficile même quand $10-15\%$ des rapports d'intensité les plus significatifs viennent à manquer ! La classification de référence pour évaluer les différences est celle de l'algorithme testé quand aucune valeur n'est manquante sinon celles initialement absentes du jeu de données. Hormis SFmiss, tous les algorithmes ont un comportement hautement instable dès que la proportion supplémentaire de données manquantes dépasse les 4%. Au dessus de 7%, nous n'avons plus été capable de trouver une concordance entre les classifications et celle de départ pour EMmiss, KNN+SF et PREV+SF. Ceci suggère que ces algorithmes ont un comportement très indésirable dès que le jeu de données a «seulement» environ 9% de valeurs absentes. À l'opposé, SFmiss montre une bonne stabilité par rapport à sa classification de référence lorsque la proportion totale de valeurs

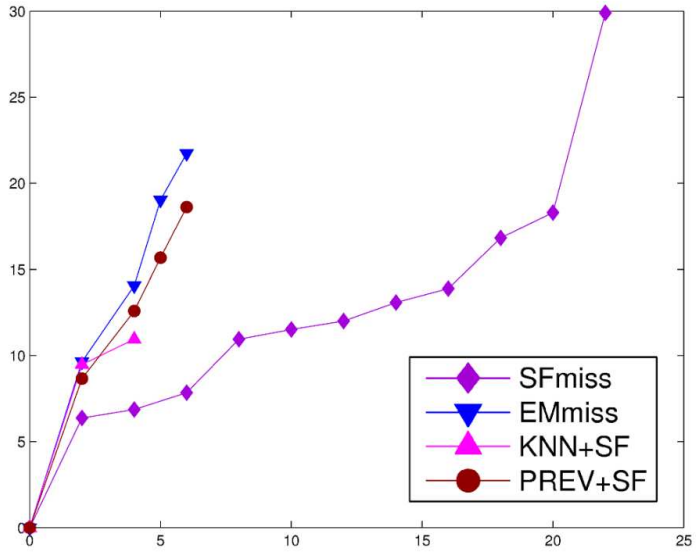


Fig. 5.26: Pourcentages de gènes mal classés par les différents algorithmes en fonction du pourcentage de valeurs manquantes ajoutées (aux $\sim$ 5% d'observations manquantes inhérentes au jeu de données).

manquantes augmente significativement. Lorsque cette proportion atteint 25%, sa stabilité est alors seulement mise à défaut. Cette valeur permet d'envisager le traitement de nombreux jeux de données vues les proportions usuelles de données manquantes. Notons toutefois que cet algorithme ne serait probablement pas applicable dans le cas où on souhaiterait par exemple analyser deux échantillons et que le recouvrement des mesures faites entre ces deux échantillons est faible. Cela demanderait en tout cas des expériences supplémentaires.

6. CONCLUSION

Bilan

Nous avons présenté dans ce mémoire l'essentiel de notre réflexion sur l'analyse des données post-génomiques. Notre état de l'art a en effet mis en évidence que le traitement actuel de ces données ne connaissait pas de consensus. Des techniques très variées sont considérées dans ce domaine. Nous avons eu pour objectif d'utiliser des hypothèses raisonnables quant aux connaissances biologiques sur les données. En outre, nous avons présenté une approche avec un outil unique pour intégrer des données de différentes sources et produire une classification des gènes. Cet outil s'appuie sur les champs de Markov cachés qui permettent une modélisation probabiliste rigoureuse et flexible des phénomènes.

Un autre travail important de cette thèse a été l'analyse des résultats de la classification. Nous pensons notamment avoir montré les avantages de notre modèle sur ceux utilisés communément.

Nous rappelons toutefois que l'approche que nous préconisons n'est pas spécifique de données issues de la biologie assistée par ordinateur. Par exemple, les champs de Markov cachés se sont révélés très performants pour l'analyse d'images et sont encore très utilisés dans ce domaine. Mais notre contribution a été de réfléchir à la présentation du modèle pour une utilisation biostatistique, à la mise en œuvre des expériences et à la validation de leurs résultats.

Perspectives

Nous ne prétendons pas avoir répondu à tous les points délicats de la problématique biologique qui nous a intéressé. De nombreuses pistes peuvent être encore explorées et demanderaient des travaux ultérieurs dont certains sont en cours.

Nous souhaiterions par exemple proposer l'utilisation de notre méthode à des experts biologistes désireux d'analyser des données issues d'expériences en cours. En effet, de tels experts seraient à même de juger de la qualité des résultats de classification ; notre jeune expérience dans les phénomènes biologiques mis en jeu bride une analyse réellement fructueuse sur de vraies données qui ne seraient pas issues d'organismes pilotes déjà bien étudiés. Pour ce faire, des développements préalables de la méthode devront éventuellement être envisagés.

Par exemple, il faudrait regarder plus finement toutes les spécifications et pos-

sibilités de paramétrisations offertes par notre modèle. Nous n'avons appliqué le modèle qu'avec des matrices de covariance diagonales pour la distribution gaussienne. Bien entendu l'utilisation de matrices pleines théoriquement plus riche est rapidement rendue impossible avec l'augmentation de la dimension des données ; leur estimation est un obstacle numérique. Nous avons testé le modèle de [38] qui s'adapte aux dimensions intrinsèques des classes. Mais les résultats n'ont pas été convaincants, probablement en raison du caractère fortement corrélé des données entre les différentes conditions expérimentales. Par ailleurs, nous avons testé le modèle avec des matrices d'interaction entre les classes $\mathbb{B}$ de la forme $b \times Id$, diagonales ou pleines. Dans l'état courant de nos expériences, les valeurs du critère BIC que nous avons obtenues ne permettent pas de distinguer ces différents modèles. Les expériences présentées dans ce mémoire ne considèrent qu'un paramètre qui résume l'intensité de l'interaction de voisinage. La nature de $\mathbb{B}$ (diagonale ou pleine) reflète les influences relatives des différentes classes à travers le réseau (attraction, répulsion). Mais encore une fois, nous manquons d'expertise biologique pour interpréter de tels effets. Nous mentionnerons aussi la possibilité offerte par notre modèle pour pondérer les arêtes du réseau. De tels poids pourraient rendre compte d'une distance *a priori* entre les nœuds du graphe ou d'une mesure de confiance quant à l'interaction entre deux nœuds.

En ce qui concerne la validation du nombre de classes à considérer (problème difficile en classification), nous aimerions aussi comparer les choix qui ont été guidés par l'utilisation du critère BIC avec d'autres approches. L'utilisation de BIC est apparue naturellement avec la participation dans les projets IS2 puis MISTIS. Ainsi, [228] propose l'utilisation d'une *gap statistic* pour détecter K. Il s'agit de comparer une statistique du résultat de la classification avec l'espérance de cette statistique sous une hypothèse nulle qui exprime l'absence de structure dans la classification. Mais des difficultés théoriques sont rencontrées pour des données de grandes dimensions et le choix du modèle nul de distribution neutre des données n'est pas simple. D'autres méthodes considèrent une stabilité de la classification face à de petites perturbation. [255] construit une *Figure Of Merit* (FOM) en enlevant une à une les conditions expérimentales et mesurant l'écart par rapport à la classification avec toutes les données. [62] considère différents indices pour le même problème et compare des algorithmes de classification variés. [21] considère aussi l'addition de bruit. Enfin certaines mesures comme l'indice de Dunn [71] ou la silhouette [124] sont fréquemment utilisées. Mais ils nécessitent la définition d'une distance entre les gènes et nous n'avons pas de distance naturelle ([93] donne un bon aperçu des méthodes dans le cas où une distance existe).

Pour une extension plus réaliste du modèle, nous avons mené un important travail de reflexion sur la mise en place d'un modèle en classes empiétantes. En effet, l'hypothèse de base qu'un objet appartient à une classe unique est très rigide. Le fait de considérer une classification floue [26, 64] ne suffit pas. Il existe de nombreuses applications où il faut considérer qu'un objet possède des ca-

ractéristiques de plusieurs classes : classification de documents selon leur genre littéraire («Pour faire le portrait d'un oiseau de Jacques PRÉVERT ne court-il pas le risque de finir classé parmi les recettes de cuisines ?) ou de genre de films, analyse d'images, *marketing*, évolution (*cf.* orchidées hybrides naturelles), mais aussi biologie moléculaire. Dans ce dernier domaine, le premier algorithme empiétant a été proposé dès 1991 dans [152]. Même si leur problématique est éloignée de la nôtre, ce travail souligne qu'une classification traditionnelle peut être trop restrictive du point de vue des interprétations ultérieures. Une organisation des processus cellulaires dont l'activité s'adapte aux circonstances environnementales apparaît comme une hypothèse très raisonnable qui explique bien les phénomènes observés notamment au niveau de la régulation de la transcription pour assurer l'expression des gènes lorsque leur action est nécessaire. Il faut alors un cadre adéquat par exemple pour identifier les gènes-clefs qui participent à plusieurs mécanismes. Un gène peut coder pour une protéine intervenant dans plusieurs voies métaboliques ([209]). [86] utilise un algorithme de type *fuzzy c-means* pour arriver à une classification recouvrante sous forme d'un diagramme de Veine qui résume les activités potentielles de nombreux gènes révélées par co-expression. [141] propose un algorithme de superposition de couches de processus pour l'analyse de données d'expression de gènes. Nous préfèrerons nous baser sur le travail récent de [19] qui peut être vu comme une généralisation des modèles de mélange [15]. Des travaux récents avec le problème de la lecture de clichés d'IRM du cerveau ([233, 260]) sont aussi instructifs. Un effet de *volume partiel* explique que les pixels de l'image observée sont souvent trop grossiers pour bien contraster entre les différents tissus (inobservables directement donc). L'image observée n'est pas produite par des processus observables aux pixels de l'image mais il faut considérer une superposition de processus théoriques «purs». Pour résumer, nous voudrons modéliser (i) le fait que les gènes participent dans un ou plusieurs processus, (ii) qu'une condition expérimentale invoque la mise en route d'un ou plusieurs processus selon différentes intensités et (iii) que les mesures effectuées sur chacun des gènes sont dues à la contribution combinée de différents processus. Si l'écriture du modèle généralisant [15] est terminée, nous n'avons pas d'algorithme pour une estimation des paramètres du modèle. La méthode n'est donc pas opérationnelle à l'heure actuelle. Une autre piste pourrait être l'utilisation de *fuzzy Markov random fields* [201] (outil dont le développement est actif en télédétection) qui considèrent des classes pures et des classes continues, combinaisons convexes des classes pures pour modéliser l'imprécision des étiquettes de classes cachées.

Un autre point que nous aimerions étudier est le comportement de l'algorithme sur des graphes non réguliers. En effet, toutes nos expériences «contrôlées» ont été faites sur des grilles régulières. Ce type de réseau n'est très certainement pas représentatif d'une quelconque réalité biologique. Il a été très utile pour évaluer notre méthode en terrain connu où la notion de voisinage avait une interprétation

simple. Cependant, il faudra avoir une idée de la répartition des gènes selon leur rôles respectifs dans l'organisme au sein d'un réseau réel ou du moins plausible ([4, 116]). [142] propose une adaptation de la forme des voisinages aux valeurs des étiquettes dans l'algorithme qui sert à l'estimation d'un modèle de champs de Markox cachés pour la segmentation d'images.

Enfin, de façon similaire au traitement des observations manquantes, il nous a été récemment suggéré d'étudier la stabilité de notre modèle face à des absences d'arêtes dans le réseau. Encore une fois une telle manipulation ne demandera pas des efforts insurmontables dans le cadre des données de type image mais il sera difficile d'évaluer sa précision sur des réseaux biologiques réels. Nos connaissances à leur sujet ne sont pas entièrement fiables ; nous ne savons pas quelle proportion de ces réseaux est actuellement disponible dans les ressources des bases de données actuelles. Cependant cette piste de recherche intéresse au plus haut point les biologistes.

BIBLIOGRAPHIE

[1] François Bolley. *Applications du transport optimal à des problèmes de limites de champ moyen.* PhD thesis, École Normale Supérieure de Lyon, décembre 2005.

[2] Jean-François Boulicaut and Olivier Gandrillon, editors. *Informatique pour l'analyse du transcriptome.* Hermès, Paris, juillet 2004.

[3] Gene Ontology Consortium. The gene ontology (GO) database and informatics resource. *Nucleic Acids Research*, 32(Database Issue) :D258–D261, janvier 2004.

[4] Réka Albert. Scale-free networks in cell biology. *Journal of Cell Science*, 118(21) :4947–4957, novembre 2005.

[5] Réka Albert and Albert-László Barabási. Statistical mechanics of complex networks. *Review of Modern Physics*, 74(1) :47–97, janvier 2002.

[6] Ash A. Alizadeh, Michael B. Eisen, R. Eric Davis, Chi Ma, Izidore S. Lossis, Andreas Rosenwald, Jennifer Boldrick, Hajeer Sabet, Truc Tran, Xin Yu, John I. Powell, Liming Yang, Gerald E. Marti, Troy Moore, James Hudson Jr, Lisheng Lu, David B. Lewis, Robert Tibshirani, Gavin Sherlock, Wing C. Chan, Timothy C. Greiner, Dennis D. Weisenburger, James O. Armitage, Roger Warnke, Ronald Levy, Wyndham Wilson, Michael R. Grever, John C. Byrd, David Botstein, Patrick O. Brown, and Louis M. Staudt. Distinct types of diffuse large b-cell lymphoma identified by gene expression profiling. *Nature*, 403(6769) :503–511, février 2000.

[7] Paul D. Allison. Multiple imputation for missing data : a cautionary tale. *Sociological Methods and Research*, 28(3) :301–309, février 2000.

[8] Stephen F. Altschul, Thomas L. Madden, Alejandro A. Schäffer, Jinghui Zhang, Zheng Zhang, Webb Miller, and David J. Lipman. Gapped blast and psi-blast : a new generation of protein database search programs. *Nucleic Acids Research*, 25(17) :3389–3402, septembre 1997.

[9] Siv G.E. Andersson and Kimmo Eriksson. *Comparative Genomics*, chapter Dynamics of gene order structures and genomic architectures, pages 267–280. Kluwers Academic Publishers, octobre 2000.

[10] Miguel A. Andrade and Peer Bork. Automated extraction of information

in molecular biology. *Federation of European Biochimical Societies Letters*, 476 :12–17, juin 2000.

[11] Mircea Andrecut and Stuart A. Kauffman. Mean-field model of genetic regulatory networks. *New Journal of Physics*, 8(148), août 2006.

[12] Barry C. Arnold and S. James Press. Compatible conditional distributions. *Journal of the American Statistical Association*, 84(405) :152–156, mars 1989.

[13] Francisco Azuaje, Fatima Al-Shahrour, and Joaquin Dopazo. *Bioinformatics and Drug Discovery*, volume 316 of *Methods in Molecular Biology*, chapter 5–Ontology-driven approaches to analyzing data in functional genomics, pages 67–86. Humana Press, octobre 2005.

[14] Amos Bairoch, Rolf Apweiler, Cathy H. Wu, Winona C. Barker, Brigitte Boeckmann, Serenella Ferro, Elisabeth Gasteiger, Hongzhan Huang, Rodrigo Lopez, Michele Magrane Maria J. Martin, Darren A. Natale, Claire O'Donovan, Nicole Redaschi, and Lai-Su L. Yeh. The universal protein resource (uniprot). *Nucleic Acids Research*, 33(Database Issue) :D154–D159, janvier 2005.

[15] Arindam Banerjee, Chase Krumpelman, Joydeep Gosh, Sugato Basu, and Raymond J. Mooney. Model-based overlapping clustering. In *Proceedings of the Eleventh ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, pages 532–537, Chicago, août 2005.

[16] Jeffrey D. Banfield and Adrian E. Raftery. Model-based gaussian and non-gaussian clustering. *Biometrics*, 49(3) :803–821, septembre 1993.

[17] Ziv Bar-Joseph, Georg Gerber, Itamar Simon, David K. Gifford, and Tommi S. Jaakkola.

[18] Andrew D. Barbour and Resine Geinert. Small worlds. *Random Structures and Algortihms*, 19(1) :54–74, août 2001.

[19] Alexis Battle, Eran Segal, and Daphne Koller. Probabilistic discovery of overlapping cellular processes and their regulation. In Philip E. Bourne and Dan Gusfield, editors, *Proceedings of the 8th Annual International Conference on Computational Molecular Biology*, pages 167–176. ACM, mars 2004.

[20] Richard Bellman. *Dynamic programming*. Princeton University Press, juin 1957.

[21] Asa Ben-Hur abd Andre Eliseeff and Isabelle Guyon. A stability based method for discovering structure in clustered data. In *Proceedings of the Pacific Symposium on Biocomputing 7*, pages 6–17, Kauai, janvier 2002.

[22] Yoav Benjamini and Yosef Hochberg. Controlling the false dicovery rate – a practical and powerful approach to multiple testing. *Journal of the Royal Statistical Society B*, 57(1) :289–300, février 1995.

[23] Yoav Benjamini and Daniel Yekutieli. The control of the false discovery rate in multiple testing under dependency. *Annals of Statistics*, 29(4) :1165–1188, août 2001.

[24] Julian Besag. On the statistical analysis of dirty pictures (with discussion). *Journal of the Royal Statistical Society, Series B*, 48(3) :259–302, mai 1974.

[25] Julian Besag. Spatial interaction and the statistical analysis of lattice systems (with discussions). *Journal of the Royal Statistical Society, Series B*, 36(2) :192–236, mars 1974.

[26] James C. Bezdek, James M. Keller, Raghu Krishnapuram, and Nikhil R. Pal. *Fuzzy models and algorithms for pattern recognition and image processing*, volume 4 of *The Handbooks of Fuzzy Sets*. Kluwer Academic Publishers, Boston, août 1999.

[27] Christophe Biernacki. Précision sur les données et coude de la vraisemblance pour trouver le nombre de classes dans un mélange. *Revue de Statistique Appliquée*, 47(1) :47–62, 1999.

[28] Christophe Biernacki. Initializing em using the properties of its trajectories in gaussian mixtures. *Statistics and Computing*, 14(3) :267–279, août 2004.

[29] Adrian P. Bird. CpG islands as gene markers in the vertebrate nucleus. *Trends in Genetics*, 3(12) :342–374, décembre 1987.

[30] Christopher M. Bishop. *Pattern recognition and machine learning*. Springer, août 2006.

[31] Juliette Blanchet, Florence Forbes, and Cordelia Schmidt. Markov random fields for textures recognition with local invariant regions and their geometric relationships. In *British Machine Vision Conference*, septembre 2005.

[32] Juliette Blanchet, Florence Forbes, and Matthieu Vignes. Clustering with incomplete dependent data. In *Actes des* 39^{es} *Journées De Statistiques*, Angers, juin 2007.

[33] Juliette Blanchet and Matthieu Vignes. Combined expression data with missing values and interaction network analysis : a markovian integrated approach. In *Proceedings of the 7th IEEE Symposium on BioInformatics and BioEngineering*, Boston, octobre 2007.

[34] Trond Hellem Bø, Bjarte Dysvik, and Inge Jonassen. Lsimpute : accurate estimation of missing values in microarray data with least square methods. *Nucleic Acids Research*, 32(3) :e34, février 2004.

[35] Joseph Bockhorst, Yu Qiu, Jeremy Glasner, Mingzhu Liu, Frederick Blattner, and Mark Craven. Predicting bacterial transcription units using sequence and expression data. *Bioinformatics*, 19(Suppl.1) :i34–i43, juillet 2003.

[36] Nadia Bolshakova, Francisco Azuaje, and Pádraig Cunningham. An integrated tool for microarray data clustering and cluster validity assessment. *Bioinformatics*, 21(4) :451–455, février 2005.

[37] Peter Bork, Berend Snel, Gerrit Lehmann, Mikita Suyama, Thomas Dandekar, Warren Lathe III, and Martijn Huynen. *Comparative Genomics*, chapter Comparative genome analysis : exploiting the context of genes to infer evolution and predict fonction, pages 281–294. Kluwers Academic Publishers, octobre 2000.

[38] Charles Bouveyron. *Modélisation et classification des données de grande dimensions - application à l'analyse d'images.* PhD thesis, Université Joseph Fourier Grenoble I, septembre 2006.

[39] Frédéric Boyer, Anne Morgat, Laurent Labarre, Joël Pothier, and Alain Viari. Syntons, metabolons and interactons : an exact graph-theoretical approach for exploring neighbourhood between genomic and functional data. *Bioinformatics*, 21(23) :4209–4215, décembre 2005.

[40] Elizabeth I. Boyle, Shuai Weng, Jeremy Gollub, Heng Jin, David Botstein, J. Michael Cherry, and Gavin Sherlock. Go : :termfinder–open source software for accessing gene ontology information and finding significantly enriched gene ontology terms associated with a list of genes. *Bioinformatics*, 20(18) :3710–3715, décembre 2004.

[41] Jerome V. Braun and Hans-Georg Müller. Statistical methods for dna sequence segmentation. *Statistical Science*, 13(2) :142–162, mai 1998.

[42] Michael P.S. Brown, William Noble Grundy, David Lin, Nello Cristianini, Charles Walsh Sugnet, Terrence S. Furey, Manuel Ares Jr., and David Haussler. Knowledge-based analysis of microarray gene expression data by using support vector machines. *Proceedings of the National Academy of Sciences*, 97(1) :262–267, janvier 2000.

[43] Christine Brun, Jérôme Wojcik, Alain Guénoche, and Bernard Jacq. Etude bioinformatique des réseaux d'interactions : Prodistin, une nouvelle méthode de classification fonctionnelle des protéines. In *Actes JOBIM 2002, INRIA*, pages 171–182, juin 2002.

[44] Marcel Brun, Chao Sima, Jianping Hua, James Lowey, Brent Carroll, Edward Suh, and Edward R. Dougherty. Model-based evaluation of clustering validation measures. *Pattern Recognition*, 40(3) :807–824, mars 2007.

[45] Alan T. Bull, Alan C. Ward, and Michael Goodfellow. Search and discovery strategies for biotechnology : the paradigm shift. *Microbiology and Molecular Biology Reviews*, 2000.

[46] Prabir Burman. A comparative study of ordinary cross-validation, v-fold cross-validation and the repeated learning-testing methods. *Biometrika*, 76(3) :503–514, septembre 1989.

[47] Gilles CELEUX, Florence FORBES, and Nathalie PEYRARD. EM procedures using mean-field like approximations for markov-model based image segmentation. *Pattern recognition*, 36 :131–144, janvier 2003.

[48] Gilles CELEUX and Gérard GOVAERT. Gaussian parsimonious clustering models. *Pattern Recognition*, 28(5) :781–793, mai 1995.

[49] David CHANDLER. *Introduction to Modern Statistical Mechanics*. Oxford University Press, septembre 1987.

[50] Guoan CHEN, Tarek G. GHARIB, Chiang-Ching HUANG, Jeremy M.G. TAYLOR, David E. MISEK, Sharon L.R. KARDIA, Thomas J. GIORDANO, Mark D. IANNETTONI, Mark B. ORRINGER, Samir M. HANASH, and David G. BEER. Discordant protein and mRNA expression in lung adenocarcinomas. *Molecular and Cellular Proteomics*, 1(4) :304–313, avril 2002.

[51] Raymond J. CHO, Michael J. CAMPBELL, Elizabeth A. WINZELER, Lars STEINMETZ, Andrew CONWAY, Lisa WODICKA, Tyra G. WOLFSBERG, Andrei E. GABRIELIAN, David LANDSMAN, David J. LOCKHART, and Ronald W. DAVIS. A genome-wide transcriptional analysis of the mitotic cell cycle. *Molecular Cell*, 2 :65–73, juillet 1998.

[52] Shelley CHU, Joe DERISI, Michael EISEN, Jon MULHOLLAND, David BOTSTEIN, Patrick O. BROWN, and Ira HERSKOWITZ. The transcriptional program of sporulatin in budding yeast. *Science*, 282 :699–705, octobre 1998.

[53] David CLAYTON and John KALDOR. Empirical bayes estimate of age-standardized relative risks for use disease mapping. *Biometrics*, 3(43) :671–681, septembre 1987.

[54] Robert G. COWELL, A.Philip DAWID, Steffen L. LAURITZEN, and David J. SPIEGELHALTER. *Probabilistic networks and expert systems*. Springer, juillet 1999.

[55] Iogen S. COWIN and Pauline M. EMMET. The effect of missing data in the supplements to McCance and Widdowson's food tables on calculated nutrient intakes. *European Journal of Clinical Nutrition*, 53(11) :891–894, novembre 1999.

[56] Mark CRAVEN, David PAGE, Jude SHAVLIK, Joseph BOCKHORST, and Jeremy GLASNER. A probabilistic learning approach to whole-genome operon prediction. In *Proceedings of the 8th International Conference on Intellingent Systems for Molecular Biology*, juillet 2000.

[57] Didier DACUNHA-CASTELLE and Marie DUFLO. *Probabilités et statistiques - 1.Problèmes à temps fixe*. Masson, 1982.

[58] Antoine DANCHIN. *La barque de Delphes*. Editions Odile Jacob, avril 1998.

[59] Antoine DANCHIN. From protein sequence to function. *Current Opinion in Structural Biology*, 9(3) :363–367, juin 1999.

[60] Thomas Dandekar, Berend Snel, Martijn Huynen, and Peer Bork. Conservation of gene order : a finger print of proteins that physically interact. *Trends in Biochemical Sciences*, 23(9) :324–328, septembre 1998.

[61] Mô Van Dang. *Classification de données spatiales : modèles probabilistes et critères de partitionnement*. PhD thesis, Université technologique de Compiègne, décembre 1998.

[62] Susmitta Datta and Somnah Datta. Comparisons and validation of statistical clustering techniques for microarray gene expression data. *Bioinformatics*, 19(4) :459–466, mars 2003.

[63] Hidde De Jong. Modeling and simulation of genetic regulatory systems : a literature review. *Journal of Computational Biology*, 9(1) :67–103, janvier 2002.

[64] Doulaye Dembélé and Philippe Kastner. Fuzzy c-means method for clustering microarray data. *Bioinformatics*, 19(8) :973–980, mai 2003.

[65] Arthur P. Dempster, Nan M. Laird, and Donald B. Rubin. Maximum likelihood from incomplete data via the em algorithm. *Journal of the Royal Statistical Society*, 39(1) :1–38, février 1977.

[66] Joseph L. DeRisi, Vishwanath R. Iyer, and Patrick O. Brown. Exploring the metabolic and genetic control of gene expression on a genomic scale. *Science*, octobre 1997.

[67] Damien Devos and Alfonso Valencia. Practical limits of function prediction. *Proteins*, 41(1) :98–107, octobre 2000.

[68] Patrik D'haeseleer. How does gene expression clustering work ? *Nature Biotechnology*, 23(12) :1499–1501, décembre 2005.

[69] Inderjit S. Dhillon, Edward M. Marcotte, and Usman Roshan. Diametrical clustering for identifying anti-correlated gene clusters. *Bioinformatics*, 19(13) :1612–1619, septembre 2003.

[70] Edward R. Dougherty and Ulisses Braga-Neto. Epistemology of computational biology : mathematical models and experimental prediction as the basis of their validity. *Biological Systems*, 14(1) :65–90, mars 2006.

[71] Joseph C. Dunn. Well separated clusters and optimal fuzzy partitions. *Journal of Cybernetics*, 4 :95–104, 1974.

[72] Jean-Baptiste Durand. *Modèles à structure cachée : inférence, sélection de modèles et applications*. PhD thesis, Université Joseph Fourier, Grenoble I, janvier 2003.

[73] Richard Durbin, Sean R. Eddy, Anders Krogh, and Graeme Mitchison. *Biological sequence analysis : probabilistic models of proteins and nucleic acids*. Cambridge University Press, mai 1998.

[74] Michael B. Eisen, Paul T. Spellman, Patrick O. Brown, and David Botstein. Cluster analysis and display of genome-wide expression patterns. *Proceedings of the National Academy of Sciences*, 95 :14863–14868, décembre 1998.

[75] David Eisenberg, Edward M. Marcotte, Ioannis Xenarios, and Todd O. Yeates. Protein function in the post-genomic era. *Nature*, 405 :823–826, juin 2000.

[76] Anton J. Enright, Ioannis Iliopoulos, Nikos C. Kyrpides, and Christos A. Ouzounis. Protein interaction maps for complete genomes based on gene fusion events. *Nature*, 402(6757) :86–90, novembre 1999.

[77] Maria D. Ermolaeva, Owen White, and Steven L. Salzberg. Prediction of operons in microbial genomes. *Nucleic Acids Research*, 29(5) :1216–1221, mars 2001.

[78] Walter N. Fitch. Homology, a personal view on some of the problems. *Trends in Genetics*, 16(5) :227–231, mai 2000.

[79] Hans Föllmer. Random economies with many interacting systems. *Journal of Mathematical Economics*, 1(1) :51–62, 1974.

[80] Florence Forbes and Gersende Fort. Combining Monte Carlo and mean-field-like methods for inference in hidden Markov random fields. *IEEE Transactions on Image Processing*, 16(3) :824–837, mars 2007.

[81] Florence Forbes and Nathalie Peyrard. Hidden markov random field model selection criteria based on mean field-like approximations. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 25(9) :1089–1101, septembre 2003.

[82] Florence Forbes and Matthieu Vignes. Champs de Markov cachés et fusion de données individuelles et pairées pour l'identification de groupes de gènes. In *Actes des* 6^{es} *Journées Ouvertes Biologie, Informatique et Mathématiques*, Lyon, juillet 2005.

[83] Chris Fraley and Adrian E. Raftery. How many clusters ? Which clustering method ? Answers via model-based cluster analysis. *The Computer Journal*, 41(8) :578–588, novembre 1998.

[84] Michael Y. Galperin and Eugene V. Koonin. Who's your neighbor ? new computational approaches for functional genomics. *Nature Biotechnology*, 18 :609–613, juin 2000.

[85] Xiangchao Gan, Alan Wee-Chung Liew, and Hong Yan. Microarray missing data imputation based on a set theoretic framework and biological knowledge. *Nucleic Acids Research*, 34(5) :1608–1619, mars 2006.

[86] Audrey P. Gasch and Michael B. Eisen. Exploring the conditional coregulation of yeast gene expression through fuzzy k-means clustering. *Genome Biology*, 3(11) :0059.1–0059.22, octobre 2002.

[87] Irit Gat-Viks, Roded Sharan, and Ron Shamir. Scoring clustering solutions by theur biological relevance. *Bioinformatics*, 19(18) :2381–2389, décembre 2003.

[88] Irit Gat-Viks, Amos Tanay, Daniela Raijman, and Ron Shamir. A probabilistic methodology for integrating knowledge and experiments on biological networks. *Journal of Computational Biology*, 13(2) :165–181, mars 2006.

[89] Andrew Gelman and Terry P. Speed. Characterizing a joint probability distribution by conditionals. *Journal of the Royal Statistical Society, Series B*, 55(1) :185–188, mars 1993.

[90] Stuart Geman and Donald Geman. Stochastic relaxation, gibbs distribution and the bayesian restoration of images. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 6(6) :721–741, novembre 1984.

[91] Debashis Ghosh and Arul M. Chinnaiyan. Mixture modelling of gene expression data from microarray experiments. *Bioinformatics*, 18(2) :275–286, février 2002.

[92] Francis D. Gibbons and Frederick P. Roth. Judjing the quality of gene expression-based clustering methods using gene annotation. *Genome Research*, 12(10) :1574–1581, octobre 2002.

[93] Alain Guénoche and Henri Garreta. Representation and evaluation of partitions. In Krzysztof Jajuga, Andrzej Sokolowski, and Hans-Herman Bock, editors, *Proceedings of the Conference of the International Federation of Classification Societies*, pages 131–138. Springer, juillet 2002.

[94] Steven P. Gygi, Yvan Rochon, B. Robert Franza, and Ruedi Aebersold. Correlation between protein and mRNA abundance in yeast. *Molecular and Cellular Biology*, 19(3) :1720–1730, mars 1999.

[95] Maria Halkidi, Yannis Batistakis, and Michalis Vazirgiannis. On clustering validation techniques. *Journal of Intelligent Information Systems*, 17(2/3) :107–145, décembre 2001.

[96] Haluk Haluk and Patrick A. Kelly. Discrete-index markov-type random processes. *Proceedings of the IEEE*, 77(10) :1485–1510, octobre 1989.

[97] Julia Handl, Joshua Knowles, and Douglas B. Kell. Computational cluster validation in the post-genomic data analysis. *Bioinformatics*, 21(15) :3201–3212, août 2005.

[98] Daniel Hanisch, Alexander Zien, Ralf Zimmer, and Thomas Lengauer. Co-clustering of biological networks and gene expression. *Bioinformatics*, 18(Suppl.1) :S145–S154, juillet 2002.

[99] Erev Hartuv and Ron Shamir. A clustering algorithm based on graph connectivity. *Information Processing Letters*, 76(4–6) :175–181, décembre 2000.

[100] Daniel F. Heitjan and Srabashi Basu. Distinguishing "missing at random" and "missing completely at random". *American Statistician*, 50(3) :207–213, août 1996.

[101] John Hertz Anders Krogh Richard G. Palmer. *Introduction to the Theory of Neural Computation*. Perseus Books Group, janvier 1991.

[102] Ralf Herwig, Albert J. Poustka, Christine Müller, Christof Bull, Hans Lehrach, and John O'Brien. Large-scale clustering of cDNA-fingerprinting data. *Genome Research*, 9(11) :1093–1105, novembre 1999.

[103] Laurie J. Heyer, Semyon Kruglyak, and Shibu Yooseph. Exploring expression data : identification and analysis of coexpressed genes. *Genome Research*, 9(11) :1106–1115, novembre 1999.

[104] Masami Yokota Hirai, Mitsuru Yano, Dayan B. Goodenowe, Shigehiko Kanaya, Tomoko Kimura, Motoko Awazuhara, Masanori Arita, Toru Fujiwara, and Kazuki Saito. Integration of transcriptomics and metabolomics for understanding of global responses to nutritional stresses in arabidopsis thaliana. *Proceedings of the National Academy of Science*, 101(27) :10205–10210, juillet 2004.

[105] Elaine Holmes and Henrik Antti. Chemometric contributions to the evolution of metabolomics : mathematical solutions to characterising and interpreting complex biological nmr spectra. *The Analyst*, 127(12) :1549–1557, décembre 2002.

[106] Gen Hori, Masati Inoue, Shin ichi Nishimura, and Hiroyuki Nakahara. Blind gene classification based on ica of microarray data. In *Proceedings of the 3rd International Conference on independent component analysis and blind signal separation*, pages 332–336, San Diego, décembre 2001.

[107] David Horn and Inon Axel. Novel clustering algorithm for microarray expression data in truncated svd space. *Bioinformatics*, 19(9) :1110–1115, juin 2003.

[108] Nicholas J. Horton and Ken P. Kleinman. Much ado about nothing : a comparison of missing data methods and software to fit incomplete data regression models. *The American Statistician*, 61(1) :79–90, février 2007.

[109] Mia Hubert and Sanne Engelen. Robust PCA and classification in biosciences. *Bioinformatics*, 20(11) :1728–1736, juillet 2004.

[110] Timothy R. Hughes, Matthew J. Marton, Allan R. Jones, Christopher J. Roberts, Rollan Stoughton, Christopher D. Armour, Holly A. Bennett, Ernest Coffey, Hongyue Dai, Yudong D. He, Matthew J. Kidd, Amy M. King, Michael R. Meyer, David Slade, Pek Y. Lum, Sergey B. Stepaniants, Daniel D. Shoemaker, Daniel Gachotte, Kalpana Chakraburtty, Julian Simon, Martin Bard, and Stephen H. Friend. Functional discovery via a compendium of expression profiles. *Cell*, 102 :109–126, juillet 2000.

[111] Dafeng Hui, Shiqiang Wan, Bo Su, Gabriel Katul, Russel Monson, and Yiqi Luo. Gap-filling missing data in eddy covariance measurements using multiple imputation (MI) for annual estimations. *Agricultural and Forest Meteorology*, 121(1–2) :93–111, janvier 2004.

[112] Dirk Husmeier. Sensitivity and specificity of infering genetic regulatory interactions from microarray experiments with dynamic bayesian networks. *Bioinformatics*, 19(17) :2271–2282, novembre 2003.

[113] Martijn Huynen, Berend Snel, Warren Lathe III, and Peer Bork. Exploitation of gene context. *Current Opinion in Structural Biology*, 10(3) :366–370, juin 2000.

[114] Jacques Istas. *Introduction aux modélisations mathématiques pour les sciences du vivant*. Springer, juin 2000.

[115] Takeshi Itoh, Keiko Takemoto, Hirotada Mori, and Takashi Gojobori. Evolutionnary instability of operon structures disclosed by sequence comparisons of complete microbial genomes. *Molecular Biology and Evolution*, 16(3) :332–346, mars 1999.

[116] Hawoong Jeong, Sean P. Mason, Albert-László Barabási, and Zoltan N. Oltvai. Lethality and centrality in protein network. *Nature*, 411 :41–42, mai 2001.

[117] Ian T. Jolliffe. *Principal Component Analysis*. series in statistics. Springer, New-York, mai 1986.

[118] Ljubomir Josifovski, Martin Cooke, Phil Green, and Ascension Vizinho. State based imputation of missing data for robust speech recognition and speech enhancement. In *Proceedings of the 6th European Conference on Speech Communication and Technology*, pages 2837–2840, Budapest, septembre 1999.

[119] Maeds Kærn, Timothy C. Elston, William J. Blake, and James J. Collins. Stochasticity in gene expression : from theories to phenotypes. *Nature Reviews Genetics*, 6(6) :451–464, juin 2005.

[120] Minoru Kanehisa, Susumu Goto, Masahiro Hattori, Kiyoko F. Aoki-Kinoshita, Masumi Itoh, Shuichi Kawashima, Toshiaki Katayama, Michihiro Araki, and Mika Hirakawa. From genomics to chemical genomics : new developments in KEGG. *Nucleic Acids Research*, 34(Database issue) :D354–D357, janvier 2006.

[121] Samuel Karlin and Jan Mrázek. Predicted highly expresssed genes of diverse prokaryotic genomes. *Journal of Bacteriology*, 182(18) :5238–5250, septembre 2000.

[122] Samuel Karlin, Jan Mrázek, Allan Campbell, and Dale Kaiser. Characterizations of highly expressed genes of four fast-growing bacteria. *Journal of Bacteriology*, 183(17) :5025–5040, septembre 2001.

[123] Richard KATZ. On some criteria for estimating the order of a Markov chain. *Technometrics*, 23(3) :243–249, août 1981.

[124] Leonard KAUFMAN and Peter J. ROUSSEEUW. *Finding groups in data : an introduction to cluster analysis*. Probability and mathematical statistics. Wiley–Interscience, mars 1990.

[125] Douglas B. KELL, Robert M. DARBY, and John DRAPER. Genomic computing. explanatory analysis of plant expression profiling data using machine learning. *Plant Physiology*, 126 :943–951, juillet 2001.

[126] Purvesh KHATRI and Sorin DRAGHICI. Ontological analysis of gene expression data : current tools, limitations and open problems. *Bioinformatics*, 21(18) :3587–3595, septembre 2005.

[127] Dae-Won KIM, Ki-Young LEE, Kwang H. LEE, and Doheon LEE. Towards clustering of incomplete microarray data without the use of imputation. *Bioinformatics*, 23(1) :107–113, janvier 2007.

[128] Dae-Won KIM, Kwang H. LEE, and Doheon LEE. Detecting clusters of different geometrical shapes in microarray gene expression data. *Bioinformatics*, 21(9) :1927–1934, mai 2005.

[129] Hyynsoo KIM, Gene H. GOLUB, and Haesun PARK. Missing value estimation for DNA microarray gene expression data : local least squares imputation. *Bioinformatics*, 21(2) :187–198, janvier 2005.

[130] Ross KINDERMANN and J. Laurie SNELL. *Markov random fields and their applications*. American Mathematical Society, Providence, Rhode Island, 1980.

[131] Scott KIRKPATRICK, C. Daniel GELATT, and Mario P. VECCHI. Optimization by simulated annealing. *Science*, 220(4598) :671–680, mai 1983.

[132] Marc W. KIRSCHNER. The meaning of system biology. *Cell*, 121(4) :503–504, mai 2005.

[133] Astrid KODRIC-BROWN and James H. BROWN. Incomplete data sets in community ecology and biogeography : a cautionary tale. *Ecological Applications*, 3(4) :736–742, novembre 1993.

[134] Balaji KRISHNAPURAM, Lawrence CARIN, and Alexander HARTEMINK. *Kernel methods in computational biology*, chapter Gene expression analysis : joint feature selection and classifier design, pages 299–318. MIT Press, août 2004.

[135] Balaji KRISHNAPURAM, David WILLIAMS, Ya XUE, Lawrence CARIN, Mário A.T. FIGUEIREDO, and Alexander J. HARTEMINK. Active learning of features and labels. In Stefan RÜPING and Tobias SCHEFFER, editors, *Proceedings of the International Conference on Machine Learning 2005 Workshop on learning with multiple views*, pages 43–50, août 2005.

[136] Frank Kschischang, Brendan J. Frey, and Hans-Andrea Loeliger. Factor graphs and the sum product algorithm. *IEEE Transactions on Information Theory*, 47(2) :498–519, février 2001.

[137] Hans Künsch, Stuart Geman, and Athanasios Kehagias. Hidden markov random fields. *The Annals of Applied Probability*, 5(3) :577–602, août 1995.

[138] Manish P. Kurkehar, Sudeshna Adak, Suchit Jhunjhunwala, and Karthik Raghupathy. Genome-wide pathway analysis and visualization using gene expression data. In *Proceedings of the Pacific Symposium on Biocomputing 7*, pages 462–473, Kauai, janvier 2002.

[139] Gert R.G. Lanckriet, Tijl De Bie, Nello Critiani, Michael I. Jordan, and William Stafford Noble. A statistical framework for genomic data fusion. *Bioinformatics*, 20(16) :2626–2635, novembre 2004.

[140] Steffen Lauritzen. *Graphical models*. Oxford Statistical Science Series. Oxford University Press, juin 1996.

[141] Laura Lazzeroni and Art B. Owen. Plaid models for gene expression data. *Statistica Sinica*, 12(1) :61–86, janvier 2002.

[142] Sylvie Le Hégarat-Mascle, Abdelaziz Kallel, and Xavier Descombes. Ant colony optimization for image regularization based on a nonstationary Markov modeling. *IEEE Transactions on Image Processing*, 2007.

[143] Émilie Lebarbier and Tristan Mary-Huard. Une introduction au critère BIC : fondements théoriques et interprétation. *Journal de la Société Française de Statistiques*, 147(1) :39–57, 2006.

[144] Irene Lee and Anthony J. Berdis. *Enzymes and their inhibitors : drug development*, chapter 3 - Kinetics. CRC, juillet 2004.

[145] Arthur M. Lesk. *Bioinformatics*. Oxford University Press, 2nd edition edition, mai 2006.

[146] Benjamin Lewin. *GENES VIII*. Pearson Prenctice Hall, janvier 2004.

[147] Stan Z. Li. *Markov random field modeling in image analysis*. Springer, juillet 2001.

[148] Roderick J.A. Little and Donald B. Rubin. *Statistical analysis with missing data*. Probability and Statistics. Wiley, second edition, septembre 2002.

[149] Alexander V. Lukashin and Mark Borodovsky. GeneMark.hmm : new solutions for gene finding. *Nucleic Acids Research*, 26(4) :1107, février 1998.

[150] Alexander V. Lukashin and Rainer Fuchs. Analysis of temporal gene expression profiles : clustering by simulated annealing and determining the optimal number of clusters. *Bioinformatics*, 17(5) :405–414, mai 2001.

[151] Jiong Ma, Allan Campbell, and Samuel Karlin. Correlations between shine-dalgarno sequences and gene features such as predicted expression

levels and operon structures. *Journal of Bacteriology*, 184(20) :5733–5745, octobre 2002.

[152] Giorgio MANCINI and Alessandro VALBONESI. Mcs/sel/bas program – an overlapping clustering method with examples from mating type interactions of ciliated protozoa. *Computer Applications in the Biosciences*, 7(3) :365–371, juillet 1991.

[153] Edward M. MARCOTTE, Matteo PELLEGRINI, Ho-Leung NG, Danny W. RICE, Todd O. YEATES, and David EISENBERG. Detecting protein function and protein-protein interactions from genome sequence. *Science*, 285 :751–753, juillet 1999.

[154] Edward M. MARCOTTE, Matteo PELLEGRINI, Michael J. THOMPSON, Todd O. YEATES, and David EISENBERG. A combined algorithm for genome-wide prediction of protein function. *Nature*, 402 :83–86, novembre 1999.

[155] Antoine MARIN. *Développement d'une méthode bioinformatique de reconnaissance de repliements des protéines et application aux séquences «orphelines» issues du séquençage de B.Subtilis*. PhD thesis, Université Paris 7, janvier 2002.

[156] Olivier MARTIN. *Approches statistiques pour l'analyse de données de puces à ADN*. PhD thesis, Université Joseph Fourier, Grenoble I, décembre 2002.

[157] Tristan MARY-HUARD, Franck PICARD, and Stéphane ROBIN. *Mathematical and computational methods in Biology*, chapter Introduction to statistical methods for microarray data analysis. Hermann, Paris, mars 2006.

[158] Catherine MATHÉ, Marie-France SAGOT, Thomas SCHIEX, and Pierre ROUZÉ. Current methods of gene prediction, their strenghts and weaknesses. *Nucleic Acids Research*, 30(19) :4103–4117, octobre 2002.

[159] Geoffrey MCLACHLAN and David PEEL. *Finite mixture models*. Probability and Statistics. John Wiley & Sons, New-York, octobre 2000.

[160] Geoffrey J. MCLACHLAN, Kim-Anh DO, and Christophe AMBROISE. *Analyzing microarray gene expression data*. Probability and Statistics. Wiley, août 2004.

[161] Geoffrey J. MCLACHLAN and Thriyambakam KRISHNAN. *The EM algorithm and extensions*. Wiley, novembre 1996.

[162] Claudine MÉDIGUE, Thierry ROUXEL, Ph. VIGIER, Alain HÉNAUT, and Antoine DANCHIN. Evidence for horizontal gene transfer in escherichia coli speciation. *Journal of Molecular Biology*, 222(4) :851–856, décembre 1991.

[163] Pedro MENDES. Emerging bioinformatics for the metabolome. *Briefings in Bioinformatics*, 3(2) :134–145, juin 2002.

[164] Pedro MENDES, Wei SHA, and Keying YE. Artificial gene networks for objective comparison of analysis algorithms. *Bioinformatics*, 19(Suppl.2) :ii122–ii129, septembre 2003.

[165] Stanley Milgram. The small world problem. *Psychology Today*, 1(1) :60–67, mai 1967.

[166] Ian Milne, Franck Wright, Glenn Rowe, David F. Marshall, Dirk Husmeier, and Gràinne McGuire. TOPALi : software for automatic identification of recombinant sequences within DNA multiple alignments. *Bioinformatics*, 20(11) :1806–1807, juillet 2004.

[167] Tom Misteli. Spatial positioning : a new dimension in genome function. *Cell*, 119 :153–156, octobre 2004.

[168] Michel Morange. *Histoire de la biologie moléculaire*. La Découverte, avril 1994.

[169] Gabriel Moreno-Hagelsieb and Julio Collado-Vides. A powerful nonhomololy method for the prediction of operons in prokaryotes. *Bioinformatics*, 18(Suppl.1) :S329–S336, juillet 2002.

[170] John Moussouris. Gibbs and Markov random systems with constraints. *Journal of Statistical Physics*, 10(1) :11–33, janvier 1974.

[171] Jan Mrázek, Devaki Bhaya, Arthur R. Grossman, and Samuel Karlin. Highly expressed and alien genes of the *synechocystis* genome. *Nucleic Acids Research*, 29(7) :1590–1601, avril 2001.

[172] Mark Newman. The structure and function of complex networks. *SIAM Review*, 45(2) :167–256, mai 2003.

[173] Patrick Nitschké, Pascale Guerdoux-Jamet, Hélène Chiapello, Gaël Faroux, Corinne Hénaut, Alain Hénaut, and Antoine Danchin. Indigo : a world-wide-web review of genomes and gene functions. *Federation of European Microbiological Societies Microbiology Reviews*, 22(4) :207–227, octobre 1998.

[174] R. Bob O'Hara, Elja Arjas, Hannu Toivonen, and Ilkka Hanski. Bayesian analysis of metapopulation data. *Ecology*, 83(9) :2408–2415, septembre 2002.

[175] Ming Ouyang, William J. Welsh, and Panos Georgopoulos. Gaussian mixture clustering and imputation of microarray data. *Bioinformatics*, 20(6) :917–923, avril 2004.

[176] Roos Overbeek, Michael Fonstein, Mark D'Souza, Gordon D. Pusch, and Natalia Maltsev. The use of genes clusters to infer functional coupling. *Proceedings of the National Academy of Sciences*, 96(6) :2896–2901, mars 1999.

[177] Paul Pavlidis and William Noble Grundy. Combinig microarray expression data and phylogenetic profiles to learn gene functional categories using support vector machines. Technical Report CUCS-011-00, Columbia University Computer Science Departement, avril 2000.

[178] Dana Pe'er, Aviv Regev, Gal Elidan, and Nir Friedman. Inferring subnetworks from perturbed expression profiles. *Bioinformatics*, 17(Suppl.1) :S215–S224, juin 2001.

[179] Matteo Pellegrini, Edward M. Marcotte, Michael J. Thompson, David Eisenberg, and Todd O. Yeates. Assigning protein functions by comparative genome analysis : Protein phylogenetic profiles. *Proceedings of the National Academy of Sciences*, 96(8) :4285–4288, avril 1999.

[180] Guy Perrière, Jean R. Lobry, and Jean Thioulouse. Correspondence discriminant analysis : a multivariate method for comparing classes of protein and nucleic acid sequences. *Computer Applications of Biosciences*, décembre 1996.

[181] Nathalie Peyrard. *Approximations de type champ moyen des modèles de champ de Markov pour la segmentation de données spatiales*. PhD thesis, Université Joseph Fourier, Grenoble I, octobre 2001.

[182] Julie Peyre. *Analyse statistique des données issues des biopuces ADN*. PhD thesis, Université Joseph Fourier, Grenoble I, septembre 2005.

[183] Jie Qin, Darrin P. Lewis, and William Stafford Noble. Kernel hierarchical gene clustering from micoarray expression data. *Bioinformatics*, 19(16) :2097–2104, novembre 2003.

[184] John Quackenbush. Computational analysis of microarray data. *Nature Reviews Genetics*, 2(6) :418–427, juin 2001.

[185] John Quackenbush. Microarray data normalization and transformation. *Nature Genetics*, 32(Suppl.2) :496–501, décembre 2002.

[186] Adrian Raftery. *Social Methodology*, volume 25, chapter 4 - Bayesian model selection in social research, pages 111–163. Blackwell, Cambridge, MA, 1995.

[187] Marco Ramoni and Paola Sebastiani. Robust learning with missing data. *Machine Learning*, 45(2) :147–170, novembre 2001.

[188] Marco F. Ramoni, Paola Sebastiani, and Isaac S. Kohane. Cluster analysis of gene expression dynamics. *Proceedings of the National Academy of Sciences*, 99(14) :9121–9126, juillet 2002.

[189] Franck Rapaport, Andrei Zinovyev, Marie Dutreix, Emmanuel Barillot, and Jean-Philippe Vert. Classification of microarray data using gene networks. *BMC Bioinformatics*, 8(35), février 2007.

[190] Philippe Renevey. *Speech recognition in noisy conditions uding mssing feature approach*. PhD thesis, École Polytechnique Fédéale de Lausanne, décembre 2000.

[191] Isabelle Rivals, Léon Personnaz, Lieng Taing, and Marie-Claude Potier. Enrichment or depletion of a GO category within a class of genes : which test ? *Bioinformatics*, 23(4) :401–407, février 2007.

[192] Stéphane Robin, Sophie Schbath, and Vincent Vandewalle. Statistical tests to compare motif count exceptionalities. *BMC Bioinformatics*, 8(84), mars 2007.

[193] Mark A. Robinson, Jörg Grigull, Naveed Mohammad, and Tomothy R. Hughes. Funspec : a web-based cluster interpreter for yeast. *BMC Bioinformatics*, 2(35), novembre 2002.

[194] Eduardo Rocha. *Analyse exploratoire des génomes bactériens*. PhD thesis, Université de Versailles Saint-Quentin-en-Yvelines, avril 2000.

[195] Eduardo P.C. Rocha, Pascale Guerdoux-Jamet, Ivan Moszer, Alain Viari, and Antoine Danchin. Implication of gene distribution in the bacterial chromosome for the bacterial cell factory. *Journal of Biotechnology*, 78(3) :209–219, mars 2000.

[196] Eduardo P.C. Rocha, Agnieszka Sekowska, and Antoine Danchin. Sulphur islands in the *escherichia coli* genome : markers of the cell's architecture ? *Federation of European Biochemical Societies Letters*, 476(1) :8–11, juin 2000.

[197] Donald B. Rubin. Inference and missing data. *Biometrika*, 63(3) :581–592, décembre 1976.

[198] Andreas Ruepp, Alfred Zollner, Dieter Maier, Kaj Albermann, Jean Hani, Martin Mokrejs, Igor Tetko, Ulrich Güldener, Gertrud Mannhaupt, Martin Münsterkötter, and H. Werner Mewes. The funcat, a functional annotation scheme for systematic classification of proteins from whole genomes. *Nucleic Acids Research*, 32(18) :5539–5545, octobre 2004.

[199] Heladia Salgado, Gabriel Moreno-Hagelsieb, Temple F. Smith, and Julio Collado-Vides. Operons in *escherichia coli* : genomic analyses and predictions. *Proceedings of the National Academy of Sciences*, 97(12) :6652–6657, juin 2000.

[200] Steven L. Salzberg, Arthur L. Delcher, Simon Kasif, and Owen White. Microbial gene identification using interpolated markov models. *Nucleic Acids Research*, 26(2) :544–548, janvier 1998.

[201] Fabien Salzenstein and Christophe Collet. Fuzzy Markov random fields versus chains for multispectral image segmentation. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 28(11) :1753–1767, novembre 2006.

[202] Joe L. Schafer. *Analysis of incomplete multivariate data*. Chapman & Hall, london edition, août 1997.

[203] Joseph Schafer. Multiple imputation : a primer. *Statistical Methods in Medical Research*, 8(1) :3–15, février 1999.

[204] Joseph Schafer. Multiple imputation in multivariate problems when the imputation and analysis models differ. *Statistica Neerlandica*, 57(1) :19–35, février 2003.

[205] Joseph L. SCHAFER and John W. GRAHAM. Missing data : our view of the state of the art. *Psychological Methods*, 7(2) :147–177, juin 2002.

[206] Ida SCHELL, Magne ALDRIN, Ingrid K. GLAD, Ragnhild SØRUM, and Heidi LYNG. The influcence of missing value imputation on detection of differentially expressed genes from microarray data. *Bioinformatics*, 21(23) :4272–4279, décembre 2005.

[207] Gideon SCHWARZ. Estimating the dimension of a model. *The Annals of Statistics*, 6(2) :131–134, avril 1978.

[208] Paola SEBASTIANI and Marco RAMONI. Bayesian selection of decomposable models with incomplete data. *Journal of the American Statistical Association*, 96(456) :1375–1386, décembre 2001.

[209] Eran SEGAL, Alexis BATTLE, and Daphne KOLLER. Decomposing gene expression into cellular processes. In *Proceedings of the Pacific Symposium on Biocomputing 8*, pages 89–100, Lihue, janvier 2003.

[210] Muhammad Shoaib SEHGAL, Iqbal GONDAL, and Laurence S. DOOLEY. Collateral missing value imputation : a new robust missing value estimation algorithm for microarray data. *Bioinformatics*, 21(10) :2417–2423, mai 2005.

[211] Roded SHARAN, Adi MARON-KATZ, and Ron SHAMIR. Click and expander : a system for clustering and visualizing gene expression data. *Bioinformatics*, 19(14) :1787–1799, septembre 2003.

[212] Paul M. SHARP and Wen-Hsiung LI. The codon adaptative index–a measure of directional synonymous. *Nucleic Acids Research*, 15(3) :1281–1295, février 1987.

[213] Sandip SINHARAY, Hal S. STERN, and Daniel RUSSEL. The use of multiple imputation for the analysis of missing data. *Psychological Methods*, 6(4) :317–329, décembre 2001.

[214] V. Anne SMITH and Erich D. JARVIS a nd Alexander J. HARTEMINK. Evaluating functional network inference using simulations of complex biological systems. *Bioinformatics*, 18(Suppl.1) :S216–S224, juillet 2002.

[215] Gordon K. SMYTH, Yee Hwa YANG, and Terry SPEED. *Functional genomics : methods and protocols*, volume 224 of *Methods for molecular biology*, chapter Statistical issues in cDNA microarray data analysis, pages 111–136. Humana Press, mars 2003.

[216] Florian SOHLER, Daniel HANISCH, and Ralf ZIMMER. New methods for joint analysis of biological networks and expression data. *Bioinformatics*, 20(10) :1517–1521, juillet 2004.

[217] Paul T. SPELLMAN, Gavin SHERLOCK, Michael Q. ZHANG, Vishwanath R. IYER, Kirk ANDERS, Michael B. EISEN, Patrick O. BROWN, David BOTSTEIN, and Bruce FUTCHER. Comprehensive identification of cell cycle-

regulated genes of the yeast *Saccharomyces cerevisiae* by microarray hydridization. *Molecular Biology of the Cell*, 9(12) :3273–3297, décembre 1998.

[218] Benjamin J. Stapley and Gerry Benoit. Biobibliometrics : information retrieval and visualisation from co-occurrences of gene names in medline abstracts. In *Proceedings of the Pacific Symposium on Biocomputing 5*, pages 526–537, Oahu, janvier 2000.

[219] Ralf Steuer, Jürgen Kurths, Carstens O. Daub, Janko Weise, and Joachim Selbig. The mutual information : detecting and evaluating dependencies between variables. *Bioinformatics*, 18(Suppl.2) :S231–S240, octobre 2002.

[220] Javier Tamames. Evolution of gene order conservation in prokaryotes. *Genome Biology*, 2(6) :0020.1–0020.11, juin 2001.

[221] Pablo Tamayo, Donna Slonim, Jill Mesirov, Qing Zhu, Sutisak Kitareewan, Ethan Dmitrovsky, Eric S. Lander, and Todd R. Golub. Interpreting patterns of gene expression with self-organizing maps : Methods and application to hematopoietic differentiation. *Proceedings of the National Academy of Sciences*, 96(6) :2907–2912, mars 1999.

[222] Amos Tanay, Roded Sharan, Martin Kupiec, and Ron Shamir. Revealing modularity and organization in the yeast molecular network by integrated analysis of highly heterogeneous genomewide data. *Proceedings of the National Academy of Sciences*, 101(9) :2981–2986, mars 2004.

[223] Roman L. Tatusov, Natalie D. Fedorova, John D. Jackson, Aviva R. Jacobs, Boris Kiryutin, Eugene V. Koonin, Dmitri M. Krylov, Raja Mazumder, Sergei L. Mekhedov, Anastasia N. Nikolskayaand B.Sridhar Rao, Sergei Smirnov, Alexander V. Sverdlov, Sona Vasudevan, Yuri I. Wolf, Jodie J. Yin, and Darren A. Natale. The COG database : an updated version includes eukaryotes. *BMC Bioinformatics*, 4(1) :41, septembre 2003.

[224] Roman L. Tatusov, Eugene V. Koonin, and David J. Lipman. A genomic perspective on protein families. *Science*, 278 :631–637, octobre 1997.

[225] Saeed Tavazoie, Jason D. Hughes, Michael J. Campbell, Raymond J. Cho, and George M. Church. Systematic determination of genetic network architecture. *Nature Genetics*, 22(3) :281–285, juillet 1999.

[226] Herbert Thijs, Geert Molenberghs, Bart Michiels, Geert Verbeke, and Desmond Curran. Strategies to fit pattern-mixture models. *Biostatistics*, 3(2) :245–265, juin 2002.

[227] Julie D. Thompson, Desmond G. Higgins, and Toby J. Gibson. CLUSTAL W : improving the sensitivity of progressive multiple sequence alignment through sequence weighting, position-specific gap penalties and weight matrix choice. *Nucleic Acids Research*, 22(22) :4673–4680, novembre 1994.

[228] Robert TIBSHIRANI, Guenther WALTHER, and Trevor HASTIE. Estimating the number of clusters in a data set via the gap statistic. *Journal of the Royal Society B*, 63(2) :411–423, avril 2001.

[229] Luke TIERNEY and Joseph B. KADANE. Accurate approximations for posterior moments and marginal densities. *Journal of the American Statistical Association*, 81(393) :82–86, mars 1986.

[230] Michael E. TIPPING and Christopher M. BISHOP. Mixtures of probabilistic principal component analysers. *Neural Computation*, 11(2) :443–482, février 1999.

[231] Olga TROYANSKAYA, Michael CANTOR, GavinSherlock, Pat BROWN, Trevor HASTIE, Robert TIBSHIRANI, David BOTSTEIN, and Russ B. ALTMAN. Missing values estimation methods for dna microaarays. *Bioinformatics*, 17(6) :520–525, juin 2001.

[232] Peter VAMPLEW and Anthony ADAMS. Recognition and anticipation of hands motions using a recurrent neural network. In *Proceedings of the IEEE International Conference on Neural Networks*, Perth, Australia, novembre 1995.

[233] Koen VAN LEEMPUT, Frederik MAES, Dirk VANDERMEULEN, and Paul SUETENS. A unifying framework for partial volume segmentation of brain mr images. *IEEE Transactions on Medical Imaging*, 22(1) :105–119, janvier 2003.

[234] Jean-Philippe VERT and Minoru KANEHISA. Graph-driven features extraction from microarray data using diffusion kernels and kernel cca. In Suzanna BECKER, Sebastian THRUN, and Klaus OBERMAYER, editors, *Advances in Neural Information Processing System 15*, pages 1425–1432. MIT Press, 2003.

[235] Matthieu VIGNES and Florence FORBES. Gene clustering via integrated markov models combining individual and pairwise features. *IEEE/ACM Transactions on Computational Biology and Bioninformatics*, à paraître, 2007.

[236] Christan VON MERING, Rolan KRAUSE, Berend SNEL, Michael CORNELL, Stephen G. OLIVIER, Stanley FIELDS, and Peer BORK. Comparative assesment of large-scale data sets of protein-protein interactions. *Nature*, 417(6887) :399–403, mai 2002.

[237] Christian VON MERING, Lars J. JENSEN, Michael KUHN, Samuel CHAFFRON, Tobias DOERKS, Beate KRÜGER, Berend SNEL, and Peer BORK. String 7–recent developments in the integration and prediction of protein interactions. *Nucleic Acids Research*, 35(Database issue) :D358–D362, janvier 2007.

[238] Martin J. WAINWRIGHT and Michael I. JORDAN. Graphical models, exponential families, and variational inference. Technical report, Department of Statistics, University of California, Berkeley, 2003.

[239] Martin J. WAINWRIGHT and Michael I. JORDAN. *New directions in statistical signal processing – From systems to brain*, chapter Chapter 11 – A variational principle for graphical models, pages 155–202. Neural Information Processing. MIT Press, août 2006.

[240] Wolfgang WEIDLICH. The statistical description of polarization phenomena in society. *British Journal of Mathematical and Statistical Psychology*, 24(2) :251–266, novembre 1971.

[241] Joe WHITTAKER. *Graphical models in applied multivariate statistics*. Wiley Series in Probability & Statistics. Wiley, New-York, avril 1990.

[242] John J. WIENS. Incomplete taxa, incomplete characters and phylogenetic accuracy : what is the missing data problem ? *Journal of Vertebrate Paleontology*, 23(2) :297–310, juin 2003.

[243] Chi-hsin WU and Peter C. DOERSCHUK. Cluster expansions for the deterministic computation of bayesian estimators based on Markov random fields. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 17(3) :275–293, mars 1995.

[244] Chien-Fu Jeff WU. On the convergence properties of the em algorithm. *Annals of Statistics*, 11(1) :95–103, mars 1983.

[245] Lei XU and Michael I. JORDAN. On convergence properties of the em algorithm for gaussian mixtures. *Neural computation*, 8(1) :129–151, janvier 1996.

[246] Ying XU, Victor OLMAN, and Dong XU. Clustering gene expression data using a graph-theoretic approach : an application of minimum spanning trees. *Bioinformatics*, 18(4) :536–545, avril 2002.

[247] Tetsushi YADA, Mitsuteru NAKAO, Yasushi TOTOKI, and Kenta NAKAI. Modeling and predicting transcriptionnal units of*escherichia coli* genes using hidden markov models. *Bioinformatics*, 15(12) :987–993, décembre 1999.

[248] Yoshihiro YAMANISHI, Jean-Philippe VERT, and Minoru KANEHISA. Protein network inference from multiple genomic data : a supervised approach. *Bioinformatics*, 2004.

[249] Yoshihiro YAMANISHI, Jean-Philippe VERT, and Minoru KANEHISA. Supervised enzyme network inference from the integration of genomic data and chemical information. *Bioinformatics*, 21(Suppl.1) :i468–i477, juin 2005.

[250] Yoshihiro YAMANISHI, Jean-Philippe VERT, Akihiro NAKAYA, and Minoru KANEHISA. Extraction of correlated gene clusters from multiple genomic

data by generalized kernel canonical correlation analysis. *Bioinformatics*, 19(Suppl.1) :i323–i330, juillet 2003.

[251] Itai YANAI, Adnan DERTI, and Charles DELISI. *Genomic technologies : present and future*, chapter Chapter 12 : Beyond sequence similarity, or sequence analysis in the age of the genome. Caister Academic Press, 2002.

[252] Itai YANAI, Yuri I. WOLF, and Eugene V. KOONIN. Evolution of gene fusions : horizontal transfer versus independant events. *Genome Biology*, 3(5) :0024.1–0024.13, avril 2002.

[253] Jonathan M. YEDIDIA, William T. FREEMAN, and Yair WEISS. *Exploring artificial intelligence in the new millenium*, chapter Chapter 8 – Understanding belief propagation and its generalizations, pages 239–270. Morgan Kaufmann, août 2002.

[254] Ka Yee YEUNG, Chris FRALEY, Alejandro MURUA, Adrian RAFTERY, and Larry RUZZO. Model-based clustering and data transformations for gene expression data. *Bioinformatics*, 17(10) :977–987, octobre 2001.

[255] Ka Yee YEUNG, David R. HAYNOR, and Walter L. RUZZO. Validating clustering for gene expression data. *Bioinformatics*, 17(4) :309–318, avril 2001.

[256] Ka Yee YEUNG, Mario MEDVEDOVIC, and Roger E. BUMGARNER. Clustering gene-expression data with repeated measurements. *Genome Biology*, 4(5) :R34, avril 2003.

[257] Ka Yee YEUNG and Larry RUZZO. Principal component analysis for clustering gene expression data. *Bioinformatics*, 17(9) :763–774, septembre 2001.

[258] Jun ZHANG. The mean field theory in em procedure for markov random fields. *IEEE Transactions on Signal Processing*, 40(10) :2570–2583, octobre 1992.

[259] Ping ZHANG. Model selection via multifold cross validation. *Annals of Statistics*, 21(1) :299–313, mars 1993.

[260] Yongyue ZHANG, Michael BRADY, and Stephen SMITH. Segmentation of brain mr images through a hidden markov random field model and the expectation-maximisation algorithm. *IEEE Transactions on Medical Imaging*, 20(1) :45–57, janvier 2001.

[261] Yu ZHENG, Joseph D. SZUSTAKOWSKI, Lance FORTNOW, Richard J. ROBERTS, and Simon KASIF. Computational identification of operons in microbial genomes. *Genome Research*, 12(8) :1221–1230, août 2002.

[262] Xiaobo ZHOU, Xiaodong WANG, and Edward R. DOUGHERTY. Missing-value estimation using linear and non-linear regression with bayesian gene selection. *Bioinformatics*, 19(17) :2302–2307, novembre 2003.

[263] Dongxiao ZHU, Alfred O. HERO, Hong CHENG, and Ritu KHANNA. Network constrained clustering for gene microarray data. *Bioinformatics*, (21) :4014–4020, novembre 2005.

[264] Heng Zhu, Metin Bilgin, Rhonda Bangham, David Hall, Antonio Casamayor, Paul Bertone, Ning Lan, Ronald Jansen, Scott Bidlingmaier, Thomas Houfek, Tom Mitchell, Perry Miller, Ralph A. Dean, Mark Gerstein, and Michael Snyder. Global analysis of protein activities using proteome chips. *Science*, 293(5537) :2101–2105, septembre 2001.

[265] Jianhua Zhu and Zhiping Weng. Fast : a novel protein structure alignment algorithm. *Proteins*, 58(3) :618–27, février 2005.

ANNEXE

A. BIOLOGIE

A.1 Une petite histoire de biologie moléculaire à la bioinformatique

Orphée
Regardez cette troupe infecte
Aux mille pattes, aux cent yeux :
Rotifères, cirons, insectes
Et microbes plus merveilleux
Que les sept merveilles du monde
Et le palais de Rosemonde !

Guillaume Apollinaire, *le Bestiaire ou cortège d'Orphée*, 1911.

Le but de cette section n'est pas de faire un exposé de connaissances biologiques en soi mais de comprendre au mieux les phénomènes et le vocabulaire des mécanismes sur lesquels nous nous sommes penché et que nous avons jugés pertinents de prendre en compte en vue de de l'analyse de données post-génomiques. Notamment, nous ne prétendons en aucun cas refaire la preuve de nombreux résultats de génétiques anciens ou récents qu'ils fassent l'unanimité (ou presque) ou soient controversés. Nous nous contentons de préciser le cadre de travail que nous adoptons. On prendra aussi bien garde aux affirmations à caractère trop général. La biologie est truffée de contre-exemples.

Sur la piste de l'ADN

On a observé très tôt la conservation de caractères phénotypiques lors de croisement de plantes ou d'animaux. On a d'abord admis qu'il existait une transmission d'informations chez les êtres vivants d'une génération à l'autre. Se posent alors les questions de savoir comment cette information est stockée et par quels mécanismes elle est transmise.

La cellule est l'unité élémentaire du monde vivant. Ceci est connu depuis leur observation microscopique par l'anglais Robert Hooke mais surtout la compréhension de cette observation par le biologiste allemand Theodor Schwann en 1839. Un organisme peut être composé d'une cellule (**procaryote**) ou de plusieurs (**eucaryote**). Une cellule possède une **membrane** lipidique qui l'isole du monde extérieur. De manière simplifiée, son destin est de croître et de se diviser.

Ce qui nécessite déjà une grande diversité dans ses composants. La cellule est composée à 70% d'eau. Les 30% restants sont principalement constitués par : des acides nucléiques (ADN et ARN), des glucides (source d'énergie et constituant de certaines parois), des lipides (acides gras surtout utilisés pour la compartimentalisation) et des protéines. Ces grandes classes possèdent précurseurs et dérivés et il faut y ajouter plusieurs centaines de types de molécules (vitamines,...).

La plupart de ces composés sont essentiels à la vie telle que nous la connaissons. La connaissance du vivant passe par la compréhension du rôle et des interactions de ces molécules ([58]).

Les protéines représentent à elles seules plus de la moitié de la masse sèche. Elles couvrent une très grande partie des fonctions moléculaires et possèdent des implications au niveau cellulaire telles que la réplication de l'ADN, la régulation et l'expression des gènes (voir plus bas), la plupart des fonctions enzymatiques, l'élaboration de la structure cellulaire et le maintien de l'équilibre entre les milieux extra- et intra-cellulaires. Elles ont en outre un large panel de fonctions au niveau des organismes : reconnaissance d'un corps étranger pour le système immunitaire, communication hormonale à longue distance, contraction musculaire ou encore communication entre synapses du système nerveux. Cela découle de deux propriétés fondamentales : reconnaître des molécules et catalyser des réactions chimiques. Une telle diversité est rendue possible grâce au nombre gigantesque de combinaisons dans la suite des acides aminés qui composent les protéines (travaux de l'américain Linus Pauling entre 1939 et 1941) et conditionnent leur structure. En effet, il y a vingt acides aminés principaux et une séquence polypeptidique classique se compose d'une dizaine à quelques milliers d'acides aminés. Cependant toutes les combinaisons ne sont pas explorées. Tout d'abord, la nature n'en a pas eu le temps. Pensons qu'avec une petite séquence de 60 résidus, on arrive à $20^{60} = 10^{78}$ combinaisons possibles, soit *grosso modo* le nombre d'atomes dans la partie visible de l'univers ! Aujourd'hui, on connaît moins d'un million ($= 10^6$) de séquences polypeptidiques. Aussi, pour observer une protéine, il faut que celle-ci ait une configuration stable. Sinon sa fonctionnalité ne peut pas être sélectionnée par évolution naturelle. Souvent les repliements de la chaîne polypeptidique (théoriquement infinis) sont en nombre restreint voire unique. Aussi de nombreuses conformations ou structures spatiales se ressemblent localement ou globalement. On est confronté à une apparente contradiction entre le nombre limité de repliements possibles et la grande variété des fonctions des protéines. Cela s'explique en grande partie par la stabilité des structures qui laisse une grande liberté à la variabilité de la séquence. C'est la clef de la richesse des fonctions observées et de leur versatilité ([155]).

Lors de la division cellulaire, à la base des mécanismes de la vie, on doit avoir une conservation du potentiel fonctionnel. Cette caractéristique découle justement de l'information génétique. Le suisse Friedrich Miescher en 1869 mit en évidence une nouvelle molécule intrigante qu'il nomma nucléine. Cette dé-

nomination se substitua à celle d'acide désoxyribonucléique quand on découvrit qu'il s'agissait d'un acide contenant un sucre, le désoxyribose. Les travaux de MIESCHER permirent en outre de démontrer que l'acide nucléique contenait du phosphore contrairement à la plupart des molécules biologiques alors connues. Peu de chercheurs s'intéressèrent à cette nouvelle molécule car on ne comprenait pas du tout quel pouvait être son rôle biologique. À la fin des années 20, Phoebus LEVINE put établir qu'outre le phosphore et le désoxyribose, elle contient quatre éléments de construction différents appelés **bases azotées** : l'adénine, la guanine, la thymine et la cytosine. Mais sa structure exacte ne put être élucidée.

Les recherches reprirent à la suite d'une série de travaux démontrant son rôle dans l'hérédité. En 1932, un microbiologiste anglais, Frederick GRIFFITH, à la recherche d'un vaccin contre la pneumonie, démontra que des pneumocoques (les microbes responsables de cette maladie) tués par la chaleur pouvaient transmettre certains de leurs caractères, notamment leur virulence, à des souches de pneumocoques ne provoquant pas la pneumonie. L'acquisition de cette caractéristique devenait héréditaire. Cette découverte était tellement incroyable que GRIFFITH attendit quatre ans avant de publier ses résultats. Seul le transfert d'une substance chimique entre des bactéries mortes et des bactéries vivantes pouvait expliquer cette transformation. Ce fut Oswald AVERY qui, en 1944, réussit à isoler la substance responsable, l'**ADN** ou Acide DesoxyriboNucléique qui est contenu dans les chromosomes. Tous les biologistes étaient alors convaincus que la transmission des caractères héréditaires d'une génération à l'autre dépendait des protéines depuis que le médecin anglais Archibald GARROD avait fait le lien entre les maladies héréditaires et les protéines en 1909. Aussi, les résultats d'AVERY représentaient-ils une véritable révolution conceptuelle et ils suscitèrent une vague de travaux (et de nombreux remous) sans équivalent dans l'histoire de la biologie : on avait découvert la molécule responsable à la fois du fonctionnement et des caractéristiques propres de chaque espèce. En somme, c'était la molécule-clef de la vie.

Les plus grands noms de la recherche s'attelèrent alors à la tâche difficile consistant à établir la structure exacte de cette molécule afin de comprendre comment elle pouvait assurer ses fonctions. Ce travail devait donner bien du fil à retordre aux chercheurs. Il fut rendu possible par le développement de techniques nouvelles d'investigation. C'est ainsi qu'en 1949 Erwin CHARGAFF et James N. DAVIDSON, en utilisant la chromatographie sur papier, purent montrer que tous les ADN étudiés avaient un point commun : il y a toujours autant de thymine que d'adénine et autant de guanine que de cytosine. Par ailleurs, une nouvelle méthode d'analyse, la cristallographie par diffraction des rayons X ayant montré sa puissance pour déterminer la structure complexe des protéines, elle fut appliquée à l'étude de l'ADN par Maurice WILKINS et Rosalind FRANKLIN au King's College de Londres et par Linus PAULING et Max PERUTZ aux États-Unis. Toutefois, les structures proposées alors présentaient toujours quelques défauts les rendant incompatibles avec les données expérimentales. Le grand mérite de James WAT-

son et Francis Crick est d'avoir réussi la synthèse des informations fournies par les différentes techniques d'analyse et de les avoir concrétisées en réalisant un modèle moléculaire de l'ADN. Ce modèle, d'abord incorrect, fut progressivement amélioré jusqu'à ce qu'il colle sans erreur avec l'ensemble des résultats publiés. C'est ainsi que Crick put pousser son cri de victoire un jour de mars 1953 au *pub* l'*Eagle* à proximité de Cambridge : «Nous avons trouvé le secret de la vie !». C'est ce qui valu neuf ans plus tard le prix Nobel de médecine à Crick, Watson et Wilkins.

Il a été démontré que l'information héréditaire avait un support physique. Et la forme de ce support est aussi connue, ce qui est essentiel : en biologie, souvent forme et structure sont intimement liées. C'est plus que vrai pour l'ADN qui permit de comprendre les bases moléculaires de l'hérédité. En tant que matériel génétique d'un être vivant il doit :

(i) détenir l'information propre à son espèce. L'ordre d'enchaînement des nucléotides constitue un message ; il gouverne l'assemblage des protéines. Ce sont les protéines synthétisées par un être vivant qui en détermine le phénotype.

(ii) avoir la capacité de transmettre l'information qu'il contient d'une génération à la suivante avec le moins d'erreurs possibles. Lors de la division cellulaire, les deux cellules résultantes doivent intégralement contenir le programme génétique initial pour conserver le plan d'élaboration de l'organisme. Justement la structure de l'ADN s'y prête à merveille. Crick et Watson l'avaient présenti. Aussi le brin complémentaire assure une stabilité essentielle à la molécule d'ADN. Mais il fallut attendre les travaux de Matthew Meselson et Franklin Stahl en 1958 pour prouver que le modèle de réplication de l'ADN est semi-conservatif (un brin sur deux).

(iii) autoriser des modifications raisonnables de l'information. L'ADN n'est pas d'une stabilité à toutes épreuves. Un nucléotide peut être remplacé par un autre. Le message s'en trouve alors modifié conformément au code génétique. Ces mutations [1] sont à la base de l'évolution ET de la survie des espèces. Des bases peut aussi être effacées ou insérées.

(iv) pouvoir être lu par la cellule. Contenir l'information n'est pas tout : il faut qu'elle soit utilisable, c'est à dire traduit en une forme fonctionnelle et agissante : les protéines.

Au sujet des points (i) et (iv) ci-dessus l'ADN est l'équivalent du disque dur d'un ordinateur. Un disque dur qui contient les programmes des protéines, c'est à dire l'intégralité de leur plan d'élaboration. La taille du disque nécessaire au

[1] On prendra bien garde au fait que la mutation n'a pas de finalité *per se* en vue de l'acquisition d'une nouvelle fonction qui confèrerait un avantage sélectif. Les nouvelles capacités qui découlent des mutations se découvrent lorsqu'elles sont confrontées à l'environnement puis la sélection intervient en boucle avec l'amplification.

stockage des bases qui composent l'ADN est surprenamment peu élevée : quelques 800 Mo [2] suffiront pour l'homme (qui contient environ trois milliards de paires de bases). En revanche, 5 Go seront nécessaires pour le blé (qui a un génome d'environ 16.10^9 pb) et quelques dizaines de Mo pour les bactéries (dont les génomes sont de l'ordre du million de pb). À titre de comparaison, un disque dur standard actuel contient de l'ordre de 100 Go ; les écrits de William SHAKESPEARE occuperaient 8 Mo... Le point (ii) dira que ce disque sait se transmettre d'une génération à l'autre ou plutôt que l'information qu'il contient est transmise. Le point (iii) concède que des modifications de cette information sont possibles (et indispensables en fait).

On distingue les cellules procaryotes qui ne possèdent qu'une seule molécule d'ADN, le plus souvent circulaire, des organismes eucaryotes. Dans ces derniers, presque tout l'ADN est contenu dans le noyau. Chaque molécule d'ADN s'enroule autour d'histones pour former un nucléosome. Cet ensemble de nucléosome s'enroule à son tour pour former un chromosome. Le nombre de chromosomes dépend de l'espèce.

À ce niveau, on peut se demander comment un molécule biologique peut contenir de l'information *via* le simple enchaînement des nucléotides qui compose l'ADN. Cette séquence renseigne la synthèse des protéines un peu à la façon d'un code de correspondance simple entre un alphabet à quatre lettres (les nucléotides de l'ADN) et un autre à vingt lettres (les acides aminés des protéines) : on parle de code génétique qui associe triplet de nucléotides sur un brin d'ADN (ou **codon**) et acides aminés. On prendra garde au fait qu'un triplet de nucléotides n'est pas un acide aminé. De même le mot «crapaud» tel que tu le vois ici imprimé cher lecteur n'est pas un crapaud mais une représentation dans le langage écrit d'un certaine langue (le français en l'occurrence). René MAGRITTE a peint une pipe qui n'en est pas une. C'est la trahison des images !

Le **code génétique** (figure A.1) est quasiment universel. Il est dégénéré puisqu'il code pour vingt acides aminés à l'aide de $64 = 4^3$ triplets de nucléotides. On ne sait pas expliquer pourquoi ce code a été choisi par la nature et pas un autre.

Dans un premier temps, on verra le **gène** comme l'entité, le segment sur le brin d'ADN qui contient l'information nécessaire à la synthèse d'un produit spécifique qui conduira vers la protéine. Cette définition est à la fois peu précise et pas tout à fait exacte comme nous le verrons plus loin. On pourra aussi le situer comme une unité de base de l'héritage biologique d'une personne provenant de ses parents. Géographiquement on l'identifiera en un *locus* spécifique d'un chromosome. On pourra aussi le caractériser selon l'action qu'il induit dans l'organisme. Ce qui peut d'ailleurs n'avoir de sens que pour un groupe de gènes et dépendre du niveau auquel on se place.

[2] Un octet contient huit *bit* qui sont l'unité de référence de l'information binaire (de type 0 ou 1). Pour les Il faut donc quatre *bit* soit un demi octet pour coder l'information d'une paire de bases (pb). Attention, octet se dit *byte* en anglais !

Base 1 du codon	Base 2 U	Base 2 C	Base 2 A	Base 2 G	Base 3
U	phénylalanine phénylalanine leucine leucine	sérine	tyrosine tyrosine STOP STOP ⋆	cystéine cystéine STOP ⋆ tryptophane	U C A G
C	leucine	proline	histidine histidine glutamine glutamine	arginine	U C A G
A	isoleucine isoleucine isoleucine ⋆ méthionine	thréonine	asparagine asparagine lysine lysine	sérine sérine arginine arginine	U C A G
G	valine	alanine	ac. aspartique ac. aspartique glutamate glutamate	glycine	U C A G

⋆ donne d'autres résultats chez certains micro-organismes

Fig. A.1: Code génétique de la quasi totalité des formes de vie terrestre connues.

Au tour des protéines

On a une correspondance très restreinte (univoque la plupart du temps) entre une protéine vue comme une molécule dont la forme tri-dimensionnelle après repliements explique essentiellement son activité et la chaîne linéaire articulée de ses acides aminés.

Le processus de **transcription** procaryote commence avec la fixation d'un **ARN polymérase** (ARNp) sur le **promoteur** (courte séquence de bases azotées qui précéde la zone à transcrire). Une petite protéine additionnelle, le facteur σ se lie à la polymérase pour le stabiliser. Puis l'ARNp ouvre la double hélice de l'ADN pour former un œil dans lequel les bases de l'**ARN messager** (ARNm) pourront s'assembler de façon complémentaire du brin d'ADN lu. Après cette phase d'initiation, on assiste à l'élongation : l'ARNp se déplace le long de l'ADN. Il l'utilise comme un patron pour fabriquer une molécule d'ARNm. Remarquons que la base thymine de l'ADN est remplacée par une base uracile dans la séquence d'un ARN. Après le dernier gène à transcrire, l'ARNp rencontre le terminateur qui le détache de l'ADN et stoppe le processus, illustré à la figure A.2.

On vient d'assister au premier mécanisme du **dogme central de la biologie moléculaire**. Chez un organisme eucaryote, la différence majeure est que les gènes peuvent alterner zones codantes (**introns**) et zones non codantes (**exons**). Seuls les introns sont transcrits en ARNm dans une unité de transcription (zone entre un promoteur et un terminateur). Certains d'entre eux peuvent être transcrits selon les besoins de l'organisme ; on parle d'épissage alternatif.

On trouve aussi sur la molécule ADN des **séquences régulatrices**. Elles régulent l'activité d'un ou plusieurs gènes qui peuvent être transcrits de manière

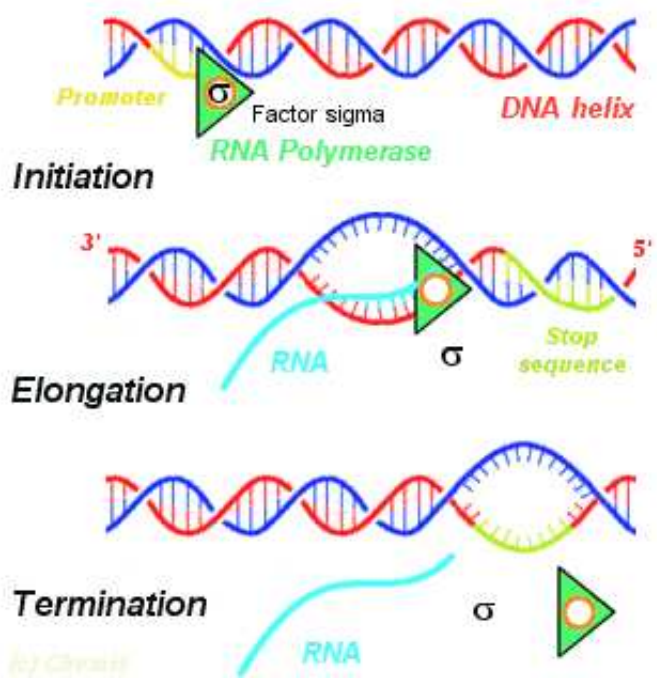


Fig. A.2: Les étapes de la transcription.

concertée. C'est à dire qu'elles contrôlent l'initiation de la transcription en la favorisant ou la réprimant. Les mécanismes peuvent être directs ou non et font souvent réponse à l'environnement cellulaire. Leur identification est non triviale. De manière générale, la recherche de sens de séquences non-codantes (c'est à dire qui ne code pas pour des gènes mais qui peuvent avoir une importance primordiale) est plus difficile que pour du codant. Souvent, l'enchaînement de leurs bases sur la séquence est remarquable mais le petit nombre de telles séquences identifiées est-il représentatif ?

Les **puces à ADN** (voir aussi la section 2.1.3) exploitent ces propriétés de liaisons spécifiques ou hybridation d'un simple brin d'ADN à un brin complémentaire d'ARNm ou d'ADNc (pour ADN cible). Une forte relation est attendue entre co-expression et co-régulation des gènes. Elle n'est pas forcément directe. Par exemple, les données d'expression peuvent servir à détecter la co-régulation comme mentionné dans [251] pour la levure. Il faut savoir qu'on se réfère uniquement à la régulation à l'initiation de la transcription.

Pour qu'un gène soit exprimé, plusieurs conditions doivent être remplies. Tout d'abord il faut un signal pour initier sa transcription. Un ligand qui s'attache à un récepteur va accomplir cette tâche. Le signal active une protéine appelée **facteurs de transcription** (*Transcription Factor, TF*). Ce facteur lie généralement de façon simultanée l'ADN, l'ARNp et d'autres agents nécessaires. Tous ces acteurs peuvent être régulés par des altérations structurales irréversibles : phosphory-

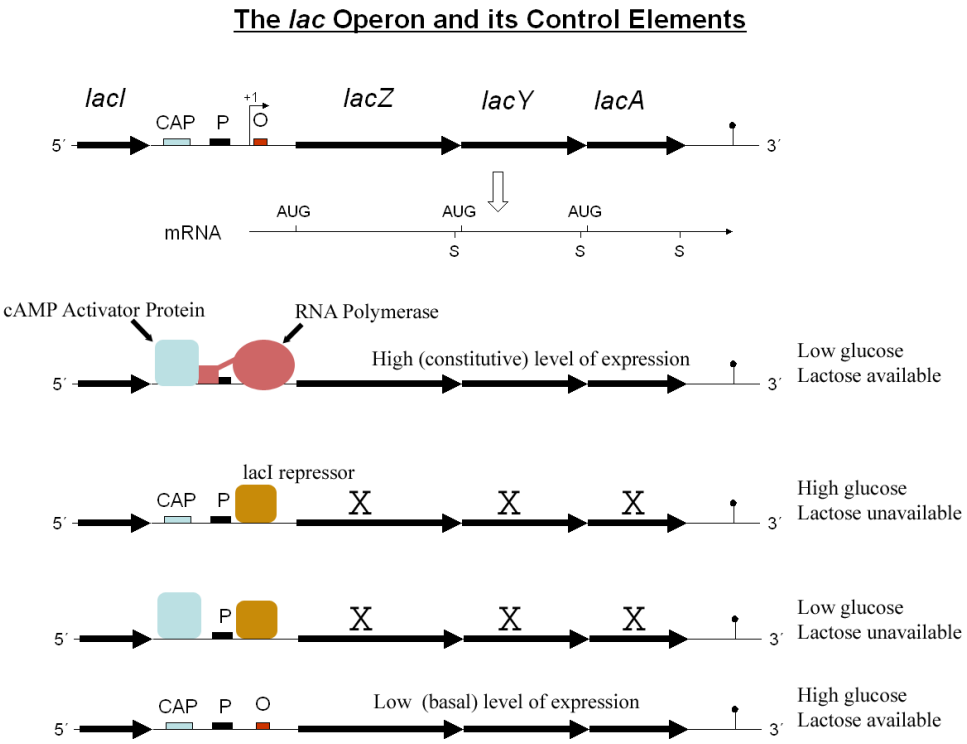


Fig. A.3: L'opéron *Lac* ou la combinaison d'«interrupteurs».

lation ou protéolyse. La transcription commence au promoteur. Des séquences ADN loin de ce point peuvent aider à cet assemblage de la transcription.

L'exemple de l'opéron *Lac* (figure A.3) est instructif pour voir un exemple de régulation selon les configurations environnementales. Il a été mis en évidence chez *Escherichia Coli* par François JACOB et Jacques MONOD de l'Institut Pasteur. Ceci participa à l'obtention de leur prix Nobel de médecine en 1965.

On trouve les gènes structurels : *LacZ* une enzyme intracellulaire qui casse le lactose en glucose et galactose, *LacY*, une protéine de transport qui se lie sur la membrane et permet le pompage du lactose dans la cellule et *LacA*, une enzyme qui transfert un groupement acétyle utile aux deux protéines précédentes. On notera qu'ici, on confond allègrement gènes et protéines pour lesquelles ils codent. En amont on trouve le site de liaison CAP et le site opérateur O qui entourent le promoteur P.

Si la cellule ne contient pas ou peu de lactose, on trouve un élément de régulation appelé le répresseur *lac*. En s'associant à l'opérateur, elle prévient toute transcription car elle couvre une partie du promoteur ; l'ARNp n'a pas la place pour agir. Les protéines de l'opéron ne sont pas nécessaires. En présence de lactose, un de ses isomères se lie au répresseur, l'inactive et laisse le site de l'opérateur libre.

En l'absence de glucose, on trouve beaucoup de $cAMP$ [3]. Ces derniers se lient au *Catabolite Activator Protein* et l'activent. Ce dernier peut alors se lier au site CAP, courbe l'ADN et l'ARNp peut alors s'attacher au promoteur et commencer la transcription de l'opéron. En présence de glucose, le niveau de $cAMP$ est bas et le CAP n'est pas activé ; il ne peut donc se lier au site CAP ; l'ARNp s'associe difficilement au promoteur : les enzymes sont produites en faibles quantités. Le lactose n'est pas la source de carbone préférée.

En résumé, expression et régulation sont fortement interdépendantes de part :

- la stabilité des ARNm,
- la configuration des promoteurs,
- l'initiation et la terminaison de la transcription ou de la traduction,
- l'usage du codon qui peut servir à caractériser le niveau d'expression des gènes ([121]) ou détecter des transferts horizontaux ([171]),
- les repliements des protéines et leur localisation dans la cellule.

Nous n'avons pas mis en évidence toutes ces influences mutuelles. On pourra consulter [146].

La seconde étape du dogme central est l'interprétation des ARNm en protéines : la **traduction**. Une sous-unité d'un **ribosome** [4] se fixe sur l'extrémité de l'ARNm. Les **ARN de transfert** (ARNt) transportent les acides aminés jusqu'au ribosome en reconnaissant lesquels doivent être assemblés grâce à un anti-codon. Le code génétique (figure A.1) est alors mis à contribution. Le ribosome se translate (on parle de translocation) triplet par triplet de nucléotides sur l'ARNm. Une liaison peptidique se forme entre les acides aminés successifs et les ARNt sont libérés. Tout cela est schématisé à la figure A.4. Il faut bien garder en mémoire que ces mécanismes font grandement usage de l'aspect tridimensionnel des molécules mises en jeu par leur imbrication.

Quand le codon STOP est rencontré, comme aucun ARNt ne lui correspond, le site où arrive les acides aminés reste vide, la translocation s'interrompt et la protéine finale est relâchée.

Les protéines sont les acteurs réels de l'activité d'un organisme à différents niveaux. Les technologies à haut débit actuel sur puces mesurent le plus souvent des données relatives aux ARN messagers, intermédiaires de la production des protéines depuis le mat ériel génétique hérité, principalement pour des raisons techniques. Pourtant le lien entre ces deux indicateurs est loin d'être direct ; en tout cas nettement moins simple qu'une relation de proportionnalité ([50, 94]).

Après le lien entre gène et protéine, la génomique fonctionnelle voudrait arriver à une étude systématique des génomes pour connaître nombre et position

[3] *cyclic Adenisine MonoPhosphate*, dérivé de l'ATP utile dans de nombreux processus biologiques. L'ATP est la «monnaie intracellulaire» des échanges d'énergie.

[4] un ribosome peut être perçu comme la machinerie cellulaire «tête de lecture» de la traduction constitué d'ARN ribosomaux (ARNr)

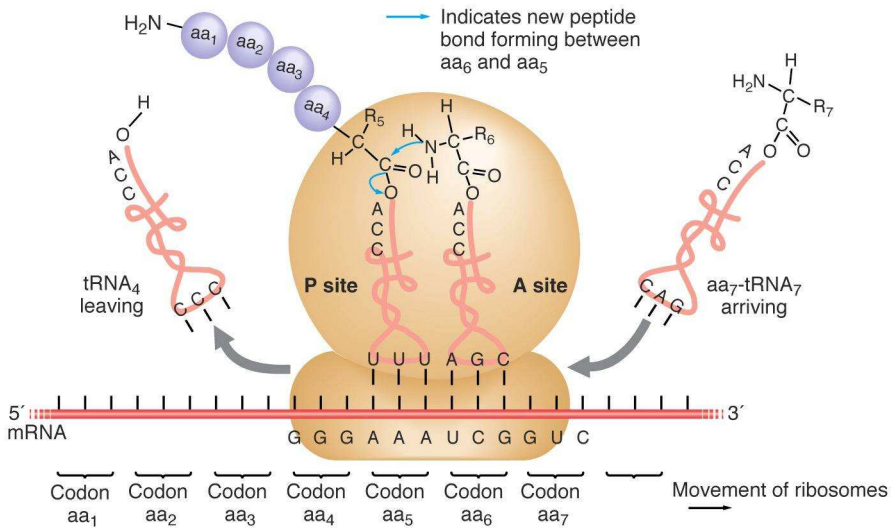


Fig. A.4: Organisation de la traduction d'un ARNm en protéine autour du ribosome.

des gènes, leur séquence, leurs produits,...Le **séquençage** consiste à délimiter et répertorier des segments d'ADN dont on voudrait connaître la fonction. En effet, depuis la découverte de la technique de la *Polymerase Chain Reaction* (PCR) en 1983 par Kary B. Mullis, il est plus facile d'obtenir la séquence d'une protéine à partir de la séquence ADN qui code pour cette protéine qu'à partir de cette protéine elle-même. Cela n'a pas toujours été la cas en biologie moléculaire. Isolement et séquençage des protéines sont actuellement plus long et couteux que ceux de molécules d'ADN. Le pendant de cette rapidité est que l'absence d'expérimentation sur la protéine ne produit plus d'information biochimique comme sa localisation cellulaire ou sa fonction.

On utilise actuellement deux techniques physiques pour la détermination expérimentale de la structure des protéines : la diffraction des rayons X sur des cristaux de protéines et la Résonance Magnétique Nucléaire (RMN). On connait alors en sortie les positions dans l'espace de tous les atomes qui composent la protéine. La *Protein Data Bank* (PDB, `http://www.rcsb.org/pdb/`) répertorie et stocke aujourd'hui environ 45 000 structures.

Les repliements des protéines sont possibles grâce aux rotations possibles autour des liaisons simples formées par le carbone α et l'azote d'une part (angle ϕ) et le carbone α et la carbone carbonyle d'autre part (angle ψ). La liaison peptidique qui possède un caractère double partiel forme un plan (matérialisé en gris clair sur la figure A.5) comprenant les atomes d'azote (en bleu), de carbone (en noir) carbonyle, d'hydrogène (en blanc) et d'oxygène (en rouge). Les résidus latéraux (en violet) déterminent le type de l'acide aminé.

On distingue généralement quatre niveaux de description de la structure des protéines :

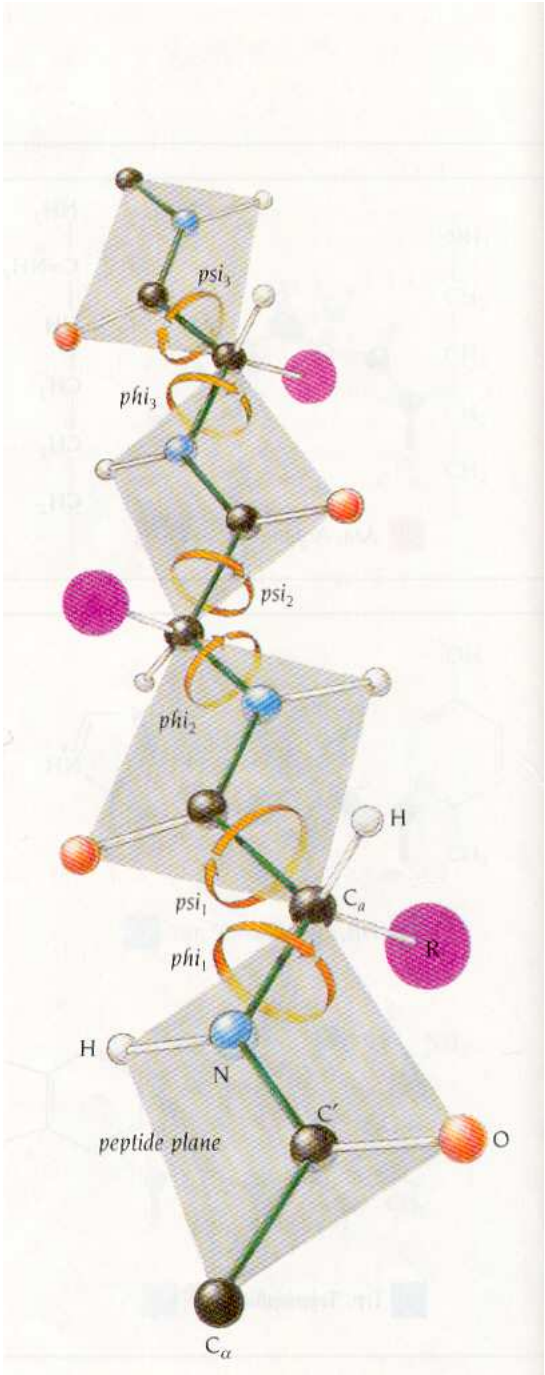


Fig. A.5: Squelette d'une protéine : la chaîne polypeptidique

- la structure primaire qui est la suite des acides aminés le long de la séquence de la protéine. Elle est connue dès le séquençage de la protéine.
- la structure secondaire qui représente les motifs de la chaîne des polypeptides.
- . la structure tertiaire qui est l'agencement des motifs de la structure secondaire ; elle est entièrement déterminée par les couples d'angles dièdres ϕ et ψ sus-cités.
- la structure quaternaire qui est l'assemblage de plusieurs chaînes polypeptidiques à l'aide de liaisons non covalentes (faibles) ou covalentes (fortes) comme les ponts disulfures.

Le chemin de la chaîne polypeptidique est contraint par l'encombrement des atomes. Le seul facteur qui varie est le résidu latéral qui se fixe sur le carbone α. En raison de cela, toutes les valeurs de (ϕ, ψ) ne sont pas admissibles. En plus de ces contraintes, la chaîne se replie en motifs périodiques pour presque toutes les protéines. Ces motifs continus, d'une longueur de l'ordre de la dizaine de résidus sont les éléments de la structure secondaire. On distingue principalement les hélices α et les feuillets β. Environ la moitié des résidus est placée au sein d'une telle structure. Ils sont séparés par des zones apériodiques appelées boucles (ou *coil*).

Ces éléments de structures secondaires s'assemblent à leur tour de façon non arbitraire et selon certaines règles. Ils forment alors des domaines structuraux qui reflètent la nature modulaire des protéines. Cette tendance peut être expliquée par des bouleversements de l'ADN comme la duplication de gènes, les transferts,...Ces modifications complètes ou partielles autorisent alors les différentes sortes d'assemblages de protéines. On a vu qu'on pouvait déterminer structures secondaire et tertiaire à l'aide d'expérimentations physiques. C'est aujourd'hui un défi bioinformatique difficile que d'arriver à prédire la structure secondaire et la structure tertiaire des protéines ([155]).

La structure quaternaire concerne les protéines dont l'activité n'est pas portée par une seule chaîne. L'amalgame des différentes chaînes forme un oligomère. Il s'agit en général de protéines agissant à l'intérieur de la cellule. Celles qui sont sécrétées c'est à dire exportées hors de la cellule sont plutôt à chaîne unique : monomérique. Pour être connue, la structure quaternaire demande à l'heure actuelle des expérimentations complémentaires.

Bien sûr il faut garder à l'esprit la grande flexibilité des protéines. Sa conformation native est souple plutôt que rigide. Cette souplesse lui permet de s'associer et de se dissocier d'autres molécules selon le type d'action envisagée.

En route vers la fonction

La fonction d'une protéine est le résultat de l'agencement spatial des résidus qui permettra un type d'interaction précis avec une autre molécule. Une recon-

naissance, une catalyse, un changement de conformation ou une combinaison de ces trois facteurs engendre une fonction. D'ailleurs une protéine peut porter plusieurs fonctions.

Comment définir alors la fonction d'une protéine ?

De nombreuses méthodes expérimentales sont disponibles. On retiendra juste qu'elles apportent des informations sur les protéines à plusieurs niveaux. On peut avoir des pistes moléculaires comme la connaissance des molécules qui interviennent dans une réaction enzymatique. On peut obtenir une information cellulaire comme la localisation d'une protéine ou la voie métabolique à laquelle elle appartient. Le renseignement peut aussi être phénotypique comme la couleur observée sur tout ou partie d'un organsime et due à la présence d'une protéine. Prenons l'exemple de l'hémoglobine contenue dans les hématies. Sa fonction est-elle la coloration du sang en rouge, le transport de l'oxygène du poumon aux muscles, la fixation de l'oxygène ? Connaître sa fonction de transport n'explicite pas le moyen de le faire. La coloration rouge une fois oxygènée ne nous dit rien quant à ce transport. Or ces fonctions sont des conséquences des association/dissociation avec l'oxygène mais vues à des niveaux différents.

[43] distingue donc bien différents niveaux de fonctions moléculaire (ou biochimique ou local : kinase, phosphatase,...) puis cellulaire (transport de molécules, communication inter-cellules, contrôle du cycle cellulaire,...). On verra ensuite apparaître la fonction au niveau d'une voie, d'un processus (*e.g.*, synthèse d'un sucre, d'un acide aminé). Le niveau suivant peut être à son tour scindé selon le type d'organisme avec, en continuant du plus petit au plus grand : la fonction au sein d'un tissu/organe (différentiation cellulaire), d'un organisme entier (pathologie, trait morphologique) et d'une population d'organismes (comme le concept de *fitness*). Il existe d'autres classifications envisageables et un obstacle majeur est qu'il n'y a, à l'heure actuelle, aucune norme internationale en la matière qui s'est imposée. Néanmoins, des initiatives telles que *Gene Ontology* ([3], `http://www.geneontology.org/`), *UniProt* ([14], `http://www.expasy.uniprot.org/`) ou encore *Clusters of Orthologous Genes* ([223, 224], `http://www.ncbi.nlm.nih.gov/COG/`) tentent de combler cette lacune en rassemblant et structurant le plus possible d'informations. *GO* a été crée dans le but de définir les rôles des gènes et protéines à l'aide d'un vocabulaire structuré. *UniProt* est né du rassemblement des informations contenues dans *Swiss-Prot*, *TrEMBL* et *PIR*. C'est une ressource internationale reconnue de séquences protéiques et de connaissances fonctionnelles associées. Les *COG* classent hiérarchiquement des groupes d'orthologues (voir A.2). On a d'abord quatre grandes classes : processus de stockage et traitement de l'information, processus cellulaire, métabolisme et la dernière classe regroupe les éléments peu ou mal caractérisés. Puis chaque classe est divisée en sous-classes contenant chacune certains groupes d'orthologues identifiés chacun par un numéro. Il existe par ailleurs des classifications qui se fondent sur d'autres critères. Par exemple les domaines protéiques avec *InterPro*.

Le terme de fonction pour une protéine a donc un précision toute relative pour le moment. Sa définition nous renseigne sur son niveau de description. En fait il faudra prendre garde aux critères que nous utilisons pour savoir quelle est l'information (fonctionnelle ou autre) qu'il pourrait être pertinent de transférer après un résultat de *clustering* par exemple.

La bioinformatique participe à ce niveau à l'analyse des génomes. Son premier axe d'intervention est conforme au rôle traditionnel de l'informatique : stockage, organisation et gestion de la quantité considérable de données brutes produites par les projets de séquençages. Le second axe majeur concerne l'analyse biologique des données. Il est très large : mise à bout des morceaux de séquences ADN séquencés, détection de parties codantes (les gènes) ou de sites d'intérêt sur la séquence par exemple. En marge de notre travail, on mentionera un autre grand pan de la bioinformatique : l'analyse phylogénétique permettant de retracer l'evolution des organismes.

En aval on trouve le champ dans lequel nous intervenons : la génomique fonctionnelle ou post-génomique. Le but idéal est de déterminer quels gènes agissent dans quel environnement pour avoir la réponse d'un tissu cellulaire. Bref de comprendre les mécanismes de fonctionnement d'un organisme. Nous passerons par un analyse *in silico* ([58]) des organismes. Traditionnelement, l'analyse fonctionnelle reposait sur des expériences biochimiques et biophysiques. On voulait caractériser au maximum les protéines c'est à dire connaître leur structure, localisation cellulaire, fonction, liste d'interacteurs, régulation d'activité ou d'expression. Cependant les techniques pour arriver expérimentalement à ces caractéristiques sont longues et coûteuses et il n'est pas possible de traiter toutes les séquences en suivant la cadence des projets de séquençage. Seules les analyses biochimiques sont des preuves expérimentales de la détermination fonctionnelle du protéome. Mais il est possible d'obtenir tout ou partie de ces informations à l'aide de modèles mathématiques en passant par une mise en œuvre informatique. On veut éviter la goulot d'étranglement ([194]) en aval du séquençage. Il existe heureusement des rétroactions entre ces deux points de vue : les expérimentations fournissent des hypothèses à la construction de modèle tandis que l'analyse fonctionnelle *in silico* ciblera les expériences permettant d'améliorer les connaissances.

Désormais, des études simultanées sur de très nombreux gènes d'un organisme peuvent être menées grâce aux avancées biotechnologiques :

- analyse du transcriptome *via* l'ARN messager qui mesure les différences et les variations du niveau d'expression des gènes,
- analyse du protéome qui identifie les protéines présentes dans la cellule,
- génomique structurelle qui détermine la structure spatiale des protéines.

Et d'autres encore. La notion même de fonction d'un gène est problématique quand on dit qu'on les groupes et qu'ils ont bien une «fonction biologique commune». On a distingué plusieurs niveaux pour la description de cette fonction.

Une telle hiérarchisation est utile au traitement informatique et à l'analyse à grande échelle telle que nous l'entreprenons. En outre, on peut parler d'action plutôt que d'utiliser le terme polémique de fonction au sein d'un tissu/organe, d'un organisme entier ou d'une population d'organismes. Mais nous ne considérerons que les deux premiers niveaux de [43] : description locale et au niveau des interactions cellulaires (niveau des voies métaboliques ou des interactions entre protéines).

En connaissant une liste de gènes -pas forcément tous mais ceux présentant des caractéristiques communes par exemples, voir la sectionclassif- d'un organisme, on se demande :

- quelle est leur fonction c'est à dire leur rôle dans les processus cellulaires,
- comment il sont régulés, comment ils (ou leurs produits plutôt) interagissent et au sein de quels réseaux,
- quel est leur niveau d'expression et les changements possibles selon le type cellulaire ou les conditions par exemple.

Pour ce faire un outil intéressant est la puce ADN. On aura à l'esprit les problèmes de normalisation de données ([156]. Les profils d'expression ont été beaucoup étudiés pour identifier ou prédire les caractéristiques fonctionnelles ([36, 182]).

Notes sur certains organismes

Pour en savoir un peu plus sur les petites bestioles que nous avons cotoyés...(nous n'avons finalement pas travaillé sur les bactéries ; mais cela ne saurait attendre !)

- *Saccharomyces cerevisiæ* : c'est levure utilisée dans l'élaboration du pain, du vin, de la bière (fermentation haute). Ces procédés sont connus depuis la préhistoire. Toutefois la compréhension biologique des mécanismes mis en jeu date de Louis Pasteur au XIX^{e} siècle. Elles servent aussi à la production d'antibiotiques. Elle peut produire l'énergie nécessaire à sa survie et sa reproduction de deux manières différentes selon le milieu ambiant : par respiration ou fermentation alcoolique du glucose. C'est un champignon unicellulaire (voir la figure A.6). C'est un organisme eucaryote (premier séquencé en 1996) très utilisé comme organisme modèle en biologie.

- *Escherichia coli* : bactérie gram-négative [5] présente dans une flore intestinale humaine normale, elle sert notamment à la production de vitamine K bien utile pour digérer. Des brins pathogènes sont à la source d'infection urinaires, pneumonie, méningite, tourista par exemple. Attention aux traitements antibiotiques parce qu'elle libère des toxines en mourant ;

[5] une bactérie gram-positive se distingue d'une bactérie gram-négative par la présence d'une épaisse paroi externe et l'absence de menbrane externe. On les distingue expérimentalement par le test de Gram qui est l'absorption d'un colorant violet dans la protéine et l'observation de son lieu de fixation.

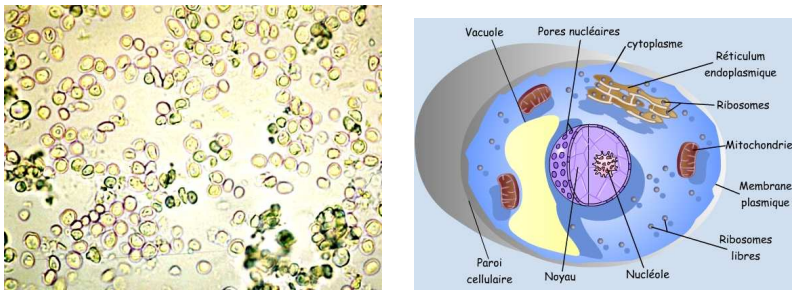


Fig. A.6: Gauche : levure de bière, sous-produit lavé, tamisé, puis press é et desséché et droite : schéma d'une levure

- *Bacillus subtilis* : bactérie gram-positive non pathogène de la soie. Elle sécrète de nombreuses enzymes utiles dans l'industrie. Elle vit dans le sol. L'intérêt que l'on porte à cette bactérie date de 1958 quand le microbiologiste John Spizizen réussit à effecter un transfert d'ADN dans cette bactérie [6]. Ceci permit de comprendre structure, organisation et mécanismes de régulation des gènes de *B.subtilis*. A l'heure actuelle, *E.Coli* (et ses proches parents) et *B.Subtilis* sont les bactéries pilotes du monde vivant les plus étudiées.

- *Hæmophilus influenzae* : bactérie qui fait partie de la flore normale du nez et du pharynx. Mais elle peut provoquer des maladies graves chez certains patients présentant des faiblesses immunologiques. Elle est la cause principale de méningite chez les enfants agés de cinq mois à cinq ans.

- *Helicobacter pylori* : bactérie de l'estomac humain vivant dans le mucus. Une personne sur deux en est porteuse, souvent sans conséquence ; elle peut provoquer des ulcères.

- *Mycoplasma genitalium* : bactérie ; une des 1resséquencées à petit génome. Parasite de nombreux hôtes (humains, animaux, plantes et même cellules élevées en culture). En plus de leur intérêt comme pathogène, cette bactérie a un grand intérêt du fait qu'on pense qu'elle représente une sort de forme de vie minimale.

- *Synechocystis sp.* : microbe, cyanobactérie aquatique et photosynthétique. L'intérêt des cyanobactéries est qu'elles portent une quantité importante de gènes analogue de ceux des plantes participant à la photosynthèse. Tout en ayant un génome nettement plus maniable.

- *Vibrio cholerae* : bactérie microbe d'eau fraîche.

A.2 Homologie

En biologie de l'évolution, l'homologie se réfère d'abord à toute similarité entres caractères hérité d'une génération antfieure. Tandis qu'on a observé des

[6] on appelle «compétente» une bactérie qui permet d'intégrer par transformation de l'ADN exogène.

homologies dans la structure anatomique des animaux pendant longtemps, on compare en génétique homologie de séquences ADN ou protéines. Le terme homologie trouve son origine dans le terme grec $o\mu o\lambda o\gamma\varepsilon\iota\nu$ qui signifie "accord, convention".

Pour adopter au premier abord une vue simplificatrice, cette notion sert à mesurer la ressemblance entre une séquence test et celles stockées dans des bases de données. Des algorithmes de comparaison de séquences sont utilisés. Ils sont de trois types : programmation dynamique, heuristique ou apprentissage machine. Tous font appel à un **alignement** de séquences. L'alignement est un processus qui compare deux ou davantage de séquences afin d'obtenir la meilleure correspondance entre les éléments de la séquence.

On mentionnera principalement les programmes *Fasta* (alignement local [265]) et *Blast* (heuristique [8]) et qui ne sont que des tests qui donnent une décision et un seuil (ou score) associé pour dire si deux séquences sont homologues. Leur utilisation dans le milieu bioinformatique est très répandue. Il existe bien entendu aujourd'hui des programmes nettement plus performants. *Clustal W* [227] est un algorithme dédié à l'alignement multiple «progressif». On peut aussi citer Topali pour des raisons professionnelles récentes qui devient une plateforme d'échange entre plusieurs programmes bioinformatiques [166].

De façon plus biologique, la théorie de l'évolution postule que les caractères observables des organismes dérivent d'ancêtres communs. Losque deux caractères se ressemblent et si cette analogie est significative -et c'est le cœur du problème-, il est légitime d'inférer que ces deux caractères ont été hérités d'un ancêtre commun.

On définit plusieurs concepts d'**homologie** entre des gènes selon la nature du phénomène de laquelle découle la ressemblance observée :

- orthologue : se dit de deux gènes présents sur deux organismesdifférents qui sont connectés par descente évolutiveverticale. Un phénomène de *spéciation* a eu lieu (une espèce connaît une évolution divergente en deux espèces). Les deux copies des gènes dans ces espèces sont dits othologues. La meilleure preuve que duex gènes similaires soient orthologues est que dans une expérience, le fait de remplacer le gène d'une espèce par la copie d'un mutant d'une autre espèce ne perturbe pas les mécanismes en jeu. En général, les gènes/protéines ont des structures et/ou fonctions très proches. La figure A.7 différencie bien cette notion qui apparaît aux divergence de l'arbre de gauche de celle à venir dans le point suivant et qui intervient lors de la spéciation indiquée par une flêche.
- paralogue : gène dupliqué (le même gène se trouve dans deux positions du génome de la même espèce ; leur séquence peut évoluer différemment) au sein d'une espèce. Les «copies» d'un même gène peuvent acquérir des fonctions bien différentes. En effet, une seule copie dont les mutations seront conditionnées par contrainte fonctionnelle est finalement nécessaire. .

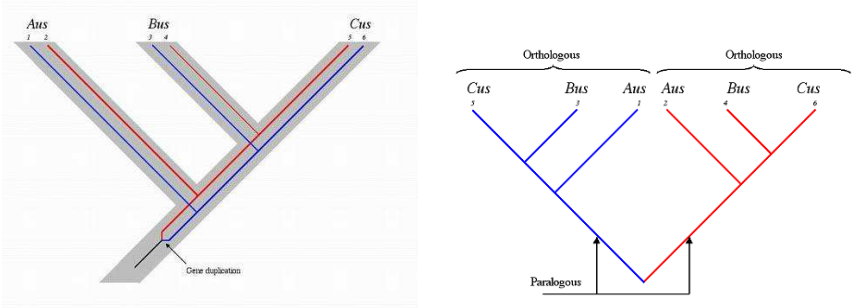


Fig. A.7: Homologie entre gènes ; orthologues et paralogues

L'exemple le plus connu est la classe des gènes de l'hémoglobine. Bien qu'ils soient tous impliqués dans le transport d'oxygène, l'hémoglobine fœtale a une bien meilleure affinité au dioxygène que l'hémoglobine adulte. Aussi des paralogues ne proviennent pas forcément d'une même espèce ; par exemple, les gènes de l'hémoglobine humains sont paralogues de ceux de la myoglobine du chimpanzé. C'est plsu généralement un probème bioinformatique que de savoir sire si deux gènes à séquences semblables partagent une fonction similaire ou non.

- ohnologue : ils s'agit de gènes paralogue résulant d'une duplication globale du génome.
- xénologue : ce type repose sur un évènement de transfert horizontal. La fonction change principalement si l'environement des deux organismes diffère beaucoup.

On prendra bien garde à distinguer le type d'homologie : fonctionnelle ou structurelle par exemple. Et ces deux types ne sont pas équivalents. Des protéines homologues fonctionnelles descendront d'un ancêtre commun et présenteront donc une homologie de structure (qui est mieux conservée que la fonction). En effet, on écarte l'évolution convergente qui ne ferait apparaître que des analogies de phénotypes, notion moins forte que l'homologie. En revanche il existe des homologues structuraux à fonctions différentes.

De manière générale, la fonction est la caractérisitique la mieux conservée d'une protéine. La pression sélective sera d'autant plus forte que la fonction a un rôle essentiel. Puis vient la structure et ensuite la séquence. On observe des séquences à repliements très comparables avec seulement 20% de résidus conservés.

Les méthodes qui mettent en évidence une homologie entre deux protéines sont surtout les comparaisons de séquences et les comparaisons de structures. Les premières ont l'avantage de nécessiter moins d'information (la structure primaire seulement) et d'être plus rapides. Les secondes sont plus sensibles et utilisent le fait que les protéines adoptent des repliements particuliers. En pratique, ces approches sont complémentaires ([155]).

L'homologie est aussi utilisée par Marcotte et coll. ([153]). Ils utilisent la nature modulaire des protéines et leur tendance à fusionner. Ils en concluent que les protéines mises en jeu ont de fortes chances d'avoir le même type de fonction cellulaire et/ou d'interagir.

Problème : l'orthologie n'est pas une relation d'équivalence ; un gène dans une ligné phylogénétique peut correspondre à toute une famille de paralogues dans une autre lignée.

Construction des COG (*Clusters of Orthologous Genes*) : (0) on réalise toutes les comparaisons 2 à 2 de séquences de protéines. On considère (1) un groupe d'au moins 3 orthologues d'espèces éloignées (des gènes d'espèces proches sont au préalables fusionnés) qui sont plus similaires entre elles qu'elles le sont de toutes les autres protéines de leur génome [7] sont favorisées pour être dans une même famille d'orthologues. Ceci même si le niveau de ressemblance est plutôt bas (pour prendre en compte les gènes évoluant lentement et rapidement). (1') si les plus similaires sont dans le même génome : paralogues que l'on regroupe. (2) avec les triangles ainsi formés à l'étape (1), il s'agit de les regrouper s'ils possèdent une arête commune. (3) analyse au cas par cas pour enlever les faux positifs et casser les protéines multi-domaines (on recommence avec ces domaines. (3') certains COG grands peuvent être cassés en 2. (+) 75 % des gènes bactériens et archæ (parmi 66 génomes) sont dans 4873 COG. 54 % des gènes de 7 eukaryotes ds 4852 KOG ([223]). Ainsi en principe des connaissances très poussées sur environ 4500 gènes devraient suffire pour comprendre les parties assez conservées entre génomes. (3) souvent des orthologues d'un même COG voient leurs annotations varier dans les bases de données publiques. Il y a donc une nécessité de clarifier les annotations.

Aussi si deux séquences voient l'enchaînement de leur gène mélangé, quelle est la probabilité P d'avoir un meilleur score d'alignement ? Ou quelle est la signification statistique d'un score d'alignement ? [179] répond en disant qu'on peut fabriquer $n*m$ nombre de séquences différentes si on a n protéines dans un organisme et m sur l'autre. Si on considère les séquences aléatoires et équiprobables, on a une probabilité-seuil de $\frac{1}{n*m}$.

Le lecteur désireux de voir un exposé des problèmes variés liés à la notion d'homologie en biologie pourra consulter [78]. On y aborde notamment des notions que nous avons volontairement écarté pour simplifier l'exposé : évolution convergente ou parallèle, le problème des répétitions dans les séquence ADN, problème des recombinaisons, *et.* Et le fameux problème des ailes d'oiseaux et de chauves-souris si semblables et pourtant...

[7] cela implique que la construction des COG ne dépend pas d'un seuil de similarité

A.3 Autres types de données

Données protéomiques

Dans le chapitre 2 il a été question de mesures lues sur la séquence ou de mesures de puces ADN. Ces dernières relèvent la quantité en ARN messager, un intermédiaire dans la production des protéines, véritables acteurs du monde vivant.

Obtenir directement une quantification des protéines présentes dans une culture cellulaire ou un tissu est bien évidemment tentant. On voudra aussi accéder à leur localisation dans les compartiments cellulaires voire leurs modifications post-transcriptionnelles. Elles sont un meilleur reflet du fonctionnement cellulaire. C'est une partie appelée **protéomique**. Elle n'est pas complètement redondante des puces ADN : une protéine encodée par un gène et synthétisée peut ensuite être modifiée et sont activité légèrement altérée. Un organisme peut posséder jusqu'à deux millions de protéines pour «seulement» plusieurs milliers de gènes. indexproteomique@protéomique

Les profils protéiniques représentent un grand pas en avant. De riches développement thérapeutiques pourraient en découler. Notamment en prenant en compte l'aspect des protéines vues comme des biomarqueurs de l'état d'une cellule. Mais sa mise au point est autrement plus complexe malgré des techniques innovantes récentes ou empruntées à d'autres disciplines et peaufinées vers son objectif [8]. Pour simplifier, on considèrera qu'il faudra extraire les protéines du matériel biologique présent dans la culture cellulaire, dans le tissu, l'organisme ou le liquide biologique. Suit une seconde étape de séparation des protéines en fonction de leurs caractéristiques physiques, chimiques ou leur affinité à un ligand : électrophorèse et chromatographie sont les deux techniques les plus répandues. Enfin il conviendra d'identifier les protéines. La spéctrométrie de masse en tandem permet de comparer les indices produits par un échantillon à ceux déjà référencés (technique du *fingerprinting*).

En tout état de cause, nous n'avons pas trouvé de données disponibles sur des phénomènes aussi variés que pour les données de puces ADN. Le coût de fabrication de puces à protéines à haut débit n'y est pas étranger ; il est nettement plus élevé que celui des puces ADN. Aussi comme il a été vu à la partie 2.2, ce type de puces est très utile pour obtenir des données d'interactions, les protéines ayant une forte propension à former des complexes avec d'autres acteurs du monde vivant. Cependant, l'obtention de données d'interaction tolère actuellement des données plus bruitées donc moins coûteuses. Donc des expériences réalisables en grand nombre. On pourra regarder [264].

Cependant il est bon de noter que de nombreuses raisons plaident en faveur de

[8] *e.g.* spéctrométrie de masse provenant de la physique et de l'analyse chimique pour identifier les protéines, quantifier leur niveau d'expression, localiser les peptides dans un tissu ou chercher des biomarqueurs spécifiques d'une pathologie

l'utilisation du protéome. Estimer les taux de protéines est un problème crucial pour avoir une image complète d'un processus biologique. Comprendre leurs localisations et leurs mouvements peut aussi aider à comprendre l'activité biologique qu'elles accomplissent. Les différentes étapes de la maturation d'une protéine sont également un bon indicateur de la machinerie cellulaire. Nous terminerons ce court paragraphe en rappelant que l'achèvement du projet de séquençage du génome humain soulève de nombreuses questions. Par exemple la faible quantité de gènes qu'il contient par rapport à la diversité des protéines comme acteurs centraux des processus biologiques. Mais nous devrons patienter encore un peu pour nous attaquer à de tels problèmes.

Données métabolomiques

Traditionnellement les protéines sont identifiées au sein d'un mélange complexe par spectrométrie de masse. C'est une technique physique qui caractérise des molécules par une mesure combinée de leur masse et de leur charge électrique. Les molécules à analyser doivent souvent être «cassées» pour observer leurs «empreintes» caractéristiques : l'ensemble des sous-éléments obtenus en brisant la molécule. On l'utilise majoritairement pour analyser les composants d'un complexe protéique, le séquençage de protéines et l'analyse de modifications post-traductionnelles, la substitution de séquence.

Une autre branche qui se sert de telles techniques est la **métabolomique** : l'étude systématique des traces biochimiques laissées par un processus cellulaire, c'est à dire les métabolites. Un métabolite est un intermédiaire ou un produit du métabolisme. Il désigne usuellement de petites molécules impliquées dans le développement ou la reproduction des organismes (métabolites primaires, *e.g.* glucose) ou dans des fonctions écologiques (métabolites secondaires, *e.g.* antibiotiques, nicotine ou pigments). On voudra identifier et quantifier ces métabolites. Ils sont les produits finaux indirects de l'expression des gènes. Les profils métaboliques donnent une vision instantannée de l'état physiologique de la cellule. Avant d'identifier les métabolites par spéctrométrie de masse ou par Raisonnance Magnétique Nucl'eaire [9], il faudra les séparer. Essentiellement on utilise alors la chromatographie (en phase liquide ou gazeuse) ou un principe de capillarité. Les applications vont de la toxicologie, aux génomiques nutritionnelle et fonctionnelle.

Les méthodes d'analyse de la métabolomique sont très proches de celles utilisées pour la transcriptomique et la protéomique en raison de la dimension ([105]) et de la structure des données aussi bien que des propriétés des molécules biologiques étudiées. Des adaptations sont nécessaires notamment selon l'objectif de

[9] Très rapidement, la RMN consiste à mettre les molécules dans un champ magnétique qui, en oscillant très légèrement fait vibrer les liaisons internes. La vibration de ces liaisons provoque l'émission de signaux qu'il est possible de détecter. L'ensemble de ces signaux, sinon caractéristique de la molécule en donne une forte présomption de présence en livrant presque son squelette chimique.

l'analyse : classification, identification de molécules ou inférence de mécanismes de régulation ([163]).

Un des défis de la biologie des systèmes va être d'intégrer transcriptome, protéome et métabolome (voir par exemple [104]). On devrait alors avoir une vision bien plus complète des organismes vivants et de leur fonctionnement microscopique en ayant une vue sous plusieurs angles de l'échantillon biologique au niveau de sa composition. Il semble alors évident que les différentes mesures présentées dans cette partie bien que relatives à des entités différentes sont très fortement interdépendantes. La partie 2.2 traite des données sur les interactions qui structurent le fonctionnement d'un organisme. Ces données apporteront une vision complémentaire des données individuelles mentionnées dans cette section, sortes de clichés instantanés de la cellule à différentes résolutions.

B. ASPECTS TECHNIQUES

B.1 Apparition de la statistique jusqu'a son application à la biologie

Au commencement, les statistiques n'étaient pas une affaire de mathématiciens. Plutôt une affaire d'État -comme le suggère son étymologie latine *statisticus* : relatif à l'État- ou de politique qui concernait l'évaluation du nombre d'habitants et surtout celle de la perception des taxes par exemple.

La nécessité de données sur une population et ses conditions de vie remonte au moins à l'antiquité. On mentionne souvent les exemples chinois ou égyptien pour recenser les productions agricoles ou légiférer sur le devoir du «citoyen» de fournir des renseignements exacts.

Cependant la génèse formelle du concept moderne de statistiques peut être attribuée à trois écoles différentes au XVII^{e} siècle :

- allemande (en Prusse), où est tout d'abord introduit le terme *Statistik* et où seront développées quelques notions dans un souci principalement académique,
- anglaise où l'école des arithméticiens politiques (William Petty, John Graunt, Edmund Halley) développe principalement des outils propres au système d'assurances (tableaux de mortalité). On y met aussi en évidence certains phénomènes qui dépassent le cadre des statistiques descriptives,
- française avec Jean-Baptiste Colbert et Vauban qui exécutent moult relevés relatifs à la population et des ressources du pays.

Les XVIII^{e} et XIX^{e} siècle sont une ère de développement des techniques de collecte de données. On voit déjà apparaître les différents sens du mot statistiques : ensemble de données, activité qui consiste en leur collecte, leur traitement et leur interprétation.

Ce n'est qu'en 1853 que se réunit le premier congrès international de statistiques à Bruxelles à l'initiative d'Adolphe Quételet. Ce dernier fut certainement un des premiers à concevoir que la statistique pouvait être fondée sur un calcul probabiliste.

Le calcul des probabilités se développait jusqu'alors de manière florissante, mais dans un mouvement de pensée relativement indépendant des statistiques. Les Blaise Pascal, Jean et Jacques Bernoulli, Abraham de Moivre, Pierre

de FERMAT, *etc.* au XVII^{e} siècle ou Thomas BAYES, Richard PRICE, Pierre Simon LAPLACE, Jean-Antoine Nicolas de Caritat de CONDORCET, *etc.* au XVIII^{e} avaient des motivations nettement différentes. Cet essor eut pour principale motivation les résultats de jeux de hasard. Le souci était davantage de comprendre de beaux problèmes que de s'intérsser d'un point de vue pragmatique à une grosse masse de données. Le formalisme récent des probabilités repose sur les axiomes introduits par Andreï Nikoloevich KOLMOGOROV au début du XX^{e} siècle.

Les notions d'erreur et d'espérance sont débattues et une théorie bien formalisée de l'inférence voit le jour. Tests de Student ou du χ^2 sont développés. Les travaux de Ronald Aylmer FISHER en 1930 en agronomie sont à la base de la féconde analyse de variance et d'une théorie générale de plans d'expériences. Les statistiques se développent en parallèle de nombreuses disciplines. Elles suivent alors des développements particuliers : économétrie, physique statistique, gestion, sciences humaines,...Et biologie avec les travaux de Francis GALTON sur l'utilisation de la régression linéaire dans une étude de l'hérédité de certains caractères chez les petits pois puis l'étude de mesures de corrélation avec Karl PEARSON. Ce rapprochement est consommé en 1901 avec la naissance de la revue *Biometrika*. La biostatistique, plutôt qu'une discipline à part entière, rassemblera un ensemble de techniques statistiques dédiées à l'étude de problématiques biologiques.

Mais alors les probabilités ou les statistiques en biologie : pour quoi faire ? A en croire bon nombre (pas tous !!) d'étudiants de Licence de Biologie, le cours de statistiques relève surtout d'un calvaire imposé qu'il est souhaitable d'oublier au plus vite. La présentation qui leur est faite n'est probablement pas toujours adaptée. Il est aussi indéniable qu'il faut surmonter chez beaucop un *a priori* sur la difficulté à comprendre les concepts de base. Essayons d'abord de cerner les rôles des probabilistes/statisticiens.

Un probabiliste est-il quelqu'un d'astucieux qui sait donner les chances de tirer un carré d'as au poker ou le temps d'attente probable à une queue de guichet ? Un statisticien est-il celui qui est capable de déterminer si le tabac favorise l'apparition d'un cancer ou de quel nombre de voix sera crédité tel candidat aux prochaines élections ([57]) ? Le lien entre ces propositions est qu'elles font intervenir un jugement sur le hasard. Un modèle abstrait est construit avec un cadre mathématique fixé. Il permettra de répondre à ces questions. La différence entre les deux personnages est que le statisticien est un «travailleur de terrain». Si on lui demande si tel médicament peut être utilisé il doit participer à une décision la meilleure (ou la moins mauvaise) possible grâce à des outils le plus souvent probabilistes.

Nous avons mentionné le terme **modèle**. Les dés semblent irrémédiablement jetés. C'est l'étape préalable au traitement mathématique. Qu'est-ce alors qu'un «bon modèle» ? Son but sera de mettre en évidence des phénomènes importants et généraux. Il devra être simple pour avoir valeur de paradigme ([114]). Dans la mesure du possible, il devra rendre compte du comportement sinon quantitatif du moins qualitatif du système étudié. Le modélisateur ne devra pas craindre la perte

de réalisme inhérente à cette pratique. Elle est d'autant plus inévitable au sein des systèmes biologiques complexes. Les mécanismes ou même la simple description de la plupart de ces systèmes ne sont souvent que partiellement connus. Des modèles fournissent néanmoins une information riche pour valider voire préciser les processus mis en jeu.

Nous avons vu un aperçu historique de l'arrivée nécessaire d'outils statistiques généraux ou parfois spécifiques dans des problèmes issus de la biologie. Il aura alors fallu définir la problématique qui nous concerne...Mais c'est justement le sujet de la partie principale de ce mémoire !

B.2 Autour du Bayesian Information Criterion

B.2.2

Le but des quelques pages qui suivent est de fournir une justification mathématique pour légitimer le choix du critère BIC [207] pour la sélection de modèle. Ainsi un lecteur biologique pourra comprendre pourquoi ce critère est particulièrement adapté au travail d'analyse de données tel que celui présenté dans cette thèse. Il lui faudra cependant accepter un peu de formalisme mathématique même si la volonté a été ici de mettre l'accent sur l'interprétation.

B.2.1 Tests d'hypothèses bayésien

Le but de l'estimation bayésienne est de fournir un niveau de certitude sur les paramètres inconnus d'un modèle. L'hypothèse de base est que ceux-ci sont la réalisation d'une variable aléatoire.

On écrit ici D l'ensemble des données de taille N et $\theta = (\theta_1, \ldots, \theta_d)$ l'ensemble des paramètres. Antérieurement à toute observation de données, nos connaissances et incertitudes sur les paramètres sont reflétées dans la **distribution** *a priori* : $P(\theta)$. Pour un modèle, on peut définir la vraisemblance $P(D|\theta)$ qui est la probabilité d'observer les données D connaissant la valeur des paramètres. La formule de Bayes autorise le calcul de la **distribution** *a posteriori* des paramètres *theta* sachant que les données D sont effectivement observées :

$$P(\theta|D) = \frac{P(D|\theta)P(\theta)}{P(D)},$$

où le dénominateur $P(D)$ est l'intégrale sur numérateur pour toutes les valeurs de paramètres possibles : $P(D) = \int_{\theta'} P(D|\theta')P(\theta')$ (loi des probabilités totales).

Au vue du but poursuivi -l'estimation la meilleure possible des paramètres θ-, la probabilité *a posteriori* n'a d'intérêt que pour comparer ses valeurs pour différents jeux de paramètres. La constante multiplicative $P(D)$ indépendante de θ n'a donc pas besoin d'être calculée. On trouve souvent dans la littérature une formule du type :

$$P(\theta|D) \propto P(D|\theta)P(\theta)$$

qui signifie que tout l'information nécessaire à des inférences sur θ est contenue dans le membre de droite de l'expression ci-dessus. Reste à savoir comment utiliser au mieux cette information ! Si on s'intéresse à un paramètre particulier (*e.g.* θ_1), il faudra intégrer l'expression sur les autres paramètres $(\theta_2, \ldots, \theta_d)$. Des valeurs utiles pour un paramètres pourront par exemple être le mode postérieur : θ_1 qui maximise $P(\theta_1|D)$, l'intervalle de confiance bayésien à 95% défini par les quantiles de niveau $0,025$ et $0,975$, le paramètre moyen et l'écart-moyen à cette valeur ([186]).

L'inférence bayésienne a été sujet à controverses parce qu'elle fait appel à la distribution *a priori* des paramètres. Cette dernière est la plupart du temps déterminée de façon subjective. Cependant, pour des échantillons de grande taille, ce choix a peu d'impacts. Par exemple, sa contribution à la moyenne et à la variance *a posteriori* est de l'ordre de $0(1/N)$. Toujours quand N est grand, le mode postérieur est très proche de l'estimateur de maximum de vraisemblance (EMV) et les intervalles de confiance bayésiens sont très proches des intervalles classiques. Asymptotiquement, pour des modèles réguliers [1], la distribution *a posteriori* est normale multivariée de moyenne l'EMV et de matrice de covariance l'inverse de la matrice d'information de Fisher (observée ou mieux son espérance). Ainsi des estimations par des méthodes d'EMV et bayésienne, les réponses données pour des modèles réguliers et avec des grands jeux de données sont quasiment identiques.

Supposons désormais que l'on cherche à comparer deux modèles M_1 et M_2 avec des jeux de paramètres $\theta^{[1]}$ et $\theta^{[2]}$ respectivement. Le théorème de Bayes donne la probabilité *a posteriori* que le modèle M_1 est correct conditionnelement à D et au fait que le vrai modèle est M_1 ou M_2 :

$$P(M_1|D) = \frac{P(D|M_1)\,P(M_1)}{P(D|M_1)\,P(M_1) + P(D|M_2)\,P(M_2)}.$$

On a une expression similaire pour $P(M_2|D)$ et par construction : $P(M_1|D) + P(M_2|D) = 1$.

$P(D|M_i)$ est obtenu par intégration sur $\theta^{[i]}$, qui représentera ici les paramètres du modèle M_i :

[1] un modèle regulier est tel que l'EMV est asymptotiquement normal de moyenne la valeur exacte et de matrice de covariance l'inverse de l'espérance de la matrice d'information de Fisher. Un exemple de modèle non régulier simple est celui où les données sont indépendantes, uniformément ditribuées entre 0 et θ (inconnu) ; l'EMV est alors la plus grande observation de θ et n'a pas la distribution attendue.

$$\begin{aligned} P(D|M_i) &= \int P(D|\theta^{[i]}, M_i)\, P(\theta^{[i]}|M_i)\, d\theta^{[i]} \\ &= \int (vraisemblance \times proba_{a\ priori}) d\theta^{[i]} \end{aligned}$$

Cette probabilité intégrée du modèle M_i permet de définir l'avantage *a posteriori* du modèle M_2 sur le modèle M_1 comme le rapport :

$$\frac{P(M_2|D)}{P(M_1|D)} = \underbrace{\frac{P(D|M_2)}{P(D|M_1)}}_{:=B_{21},\ \textbf{facteur de Bayes}\ de\ M_2\ contre\ M_1.} \underbrace{\frac{P(M_2)}{P(M_1)}}_{avantage\ antérieur.}$$

Dans le cas où aucun modèle n'est favorisé *a priori*, si $B_{21} > 1$, les données favorisent M_2 par rapport à M_1 et l'inverse quand $B_{21} < 1$. Ce facteur permet aussi différentes échelles de décisions qualitatives pour classer les modèles.

Le calcul de la vraisemblance intégrée est ainsi nécessaire. Le problème est que l'intégrale multiple peut être incalculable de manière exacte. C'est là qu'interviennent des approximations analytiques ou numériques telles que BIC. En fait nous verrons que cette approximation correspond plutôt à un bon choix de l'avantage antérieur, raisonnable dans un but pragmatique. Le logarithme de la vraisemblance intégrée peut aussi être interprété comme le score prédit du modèle. Le facteur de Bayes, en plus d'être un avantage quantitatif d'un modèle sur l'autre, désignera le modèle qui donnera les meilleures prédictions à partir des données sans supposer que M_1 ou M_2 est le vrai modèle.

B.2.2 L'approximation BIC

On donnera une approximation (étape un peu technique) de la vraisemblance intégrée qui conduit au critère BIC dont nous verrons combien il peut se montrer utile pour l'estimation de modèles. Pour simplifier l'écriture, on renote la vraisemblance intégrée :

$$P(D) = \int \underbrace{P(D|\theta)\, P(\theta)}_{:=\exp g(\theta)} d\theta.$$

Aussi on considèrera que D est un ensemble de N observations indépendantes et identiquement distribuées, $x_1, \cdots, x_N$ qui sont chacune un vecteur. Les résultats restent valables sous des hypothèses moins restrictives (par exemple, pour la plupart des modèles de relevés historiques concernant un év'enement où les données peuvent manquer, ne pas être indépendantes ou de même loi).

On fait un développement de Taylor de g autour de $\hat{\theta}$ qui maximise g (le mode postérieur donc) :

$$g(\theta) \;=\; g(\hat{\theta}) + (\theta-\hat{\theta})^t\, g'(\hat{\theta}) + \frac{(\theta-\hat{\theta})^t}{2} g''(\hat{\theta})(\theta-\hat{\theta}) + o(||\theta-\hat{\theta}||^2),$$

où $g'(\theta) = (\frac{\partial g(\theta)}{\partial \theta_1}, \cdots, \frac{\partial g(\theta)}{\partial \theta_d})^t$ est le vecteur des dérivées partielles premières et $g''(\theta)$ est la matrice hessienne de de g ; son élément en position (i,j) est : $\frac{\partial^2 g(\theta)}{\partial \theta_i \partial \theta_j}$. Aussi $g'(\hat{\theta}) = 0$ parce que c'est un extremum. Donc :

$$g(\theta) \approx g(\hat{\theta}) + \frac{(\theta-\hat{\theta})^t}{2} g''(\hat{\theta})(\theta-\hat{\theta}),$$

pourvu que θ soit suffisamment proche de $\hat{\theta}$. Cette condition est bien réalisée quand N est grand parce que la vraisemblance est concentrée sur son maximum et décroît rapidement quand θ s'écarte de $\hat{\theta}$ donc seules les valeurs proches de $\hat{\theta}$ vont contribuer significativement à l'intégrale définissant $P(D)$. Il s'ensuit :

$$\begin{aligned} P(D) &= \int \exp g(\theta)\, d\theta \\ &\approx \exp g(\hat{\theta}) \int \exp \frac{(\theta-\hat{\theta})^t}{2} g''(\hat{\theta})(\theta-\hat{\theta}) d\theta \\ &\approx \exp g(\hat{\theta}) (2\pi)^{d/2} |A|^{-1/2}, \end{aligned}$$

en reconnaissant que l'intégrande de la 2e ligne est proportionnel à une densité normale multidimensionnelle, avec d paramètres du modèles et $A = -g''(\hat{\theta})$. Il s'agit d'une méthode de Laplace pour les intégrales et l'erreur commise est en $O(1/N)$. Pour toutes les preuves techniques et une formalisation de ces arguments, on pourra consulter [229].

Ainsi :

$$\log P(D) = \log P(D|\hat{\theta}) + \log P(\hat{\theta}) + d/2 \log 2\pi - 1/2 \log |A| + O(1/N).$$

Et pour de grands échantillons, $\hat{\theta} \approx \bar{\theta}$ où $\bar{\theta}$ est l'EMV et $A \approx N.I$, avec I la matrice d'information de Fisher d'élément (i,j) : $-I\!E[\frac{\partial^2 \log P(y_1|\theta)}{\partial \theta_i \partial \theta_j}_{|\theta=\bar{\theta}}]$. Ainsi, $|A| \approx N^d |I|$. Ces deux approximations introduisent une erreur en $O(N^{-1/2})$ et l'équation précédente devient :

$$\begin{aligned} \log P(D) &= \log P(D|\hat{\theta}) + \log P(\hat{\theta}) + d/2 \log 2\pi - d/2 \log N \\ &\quad -1/2 \log |I| + O(N^{-1/2}) \\ &= \log P(D|\hat{\theta}) - d/2 \log N + O(1) \end{aligned}$$

Ceci signifie que la log-vraisemblance intégrée vaut la log-vraisemblance maximisée au terme correctif $-d/2 \log N$ près. L'erreur commise ne disparaît pas même pour des échantillons infinis. L'ordre des deux premiers termes : $O(N)$ et $O(\log N)$ indique que ceux-ci vont alors dominer. Cependant, une erreur en $O(1)$ peut faire penser que l'approximation est un peu grossière. En tout cas elle est systématique.

Empiriquement, il a été observé que l'expression ci-avant est plus précise que ne le ferait penser le terme en $O(1)$. En fait l'erreur semble être bien plus faible pour un choix adéquat des probabilités *a priori*. Par exemple, si $P(\theta)$ est une loi normale multidimensionnelle de moyenne $\bar{\theta}$ et de matrice de covariance I, grossièrement on peut dire que la distribution initiale contient la même quantité d'information qu'une simple observation. Cela semble raisonnable si on considère la situation fréquente où l'information *a priori* n'est pas très importante. Quantitativement, l'erreur commise est en $O(N^{-1/2})$. Cela peut servir pour approcher le facteur de Bayes d'un modèle par rapport à un autre par exemple.

Quand on compare plusieurs modèles, l'approximation menée ci-dessus sur le facteur de Bayes montre que celui qui doit être retenu est celui qui a la plus grande valeur de BIC (qui est cependant bien toujours négative) puisque :

$$\begin{aligned} 2 \log B_{21} &\approx 2(\log P(D|\bar{\theta^{[2]}}, M_2) - \log P(D|\bar{\theta^{[1]}}, M_1)) \\ &\quad -(d_2 - d_1) \log N + O(N^{-1/2}) \\ &= BIC_2 - BIC_1 + O(N^{-1/2}) \end{aligned}$$

Si on se trouve en présence de K modèles au lieu de deux, il peut être utile de les comparer chacun soit à un modèle nul M_0 soit à un modèle saturé M_S.

Plus généralement, deux modèles M_i et M_j peuvent être comparés par leur différence de valeurs BIC.

B.2.3 Interprétation de BIC

Je reprendrai ici l'exposé ds [143] qui explique clairement ce que l'approximation BIC implique dans le choix de ce qu'on présente comme un bon modèle parmi K qui diffère par leur nombre de groupes. Ils sont donc emboîtés.

Dans un premier temps, les auteurs montrent la consistance de BIC vis-à-vis de la divergence Kullback-Leibler (KL). Cette divergence entre deux densités f et g est définie par :

$$d_{KL}(f, g) = \int \log \left(\frac{f(x)}{g(x)} \right) f(x) dx$$

On peut alors noter par abus la divergence entre le modèle M_i pour $i \in \{1, \ldots, K\}$ et f, la densité à estimer :

$$d_{KL}(f, M_i) = \inf_{\theta^{[i]}} d_{KL}(f, P(.|\theta^{[i]}, M_i))$$

La fonction $i \mapsto d_{KL}(f, M_i)$ est donc décroissante. Notons M_t le modèle à partir duquel cette distance ne diminue plus. Du point de vue de la divergence de KL, M_t doit être préféré à tous les modèles plus simples $(M_1, \ldots, M_{t-1})$ parce qu'il est plus proche de f. Il doit aussi être favorisé par rapport à $M_{t+1}, \ldots, M_K$ par ce qu'il est plus parcimonieux (sans être plus proche de f). On nomme M_t le «quasi-vrai modèle».

Il s'agit alors de former la différence $BIC_t - BIC_i$. On différencie les cas (a) $i < t$ et (b) $i > t$.

(a) si $i < t$, on peut écrire :

$$\begin{aligned} BIC_t - BIC_i &= 2\log P(D|\bar{\theta}^{[t]}, M_t) - 2\log P(D|\bar{\theta}^{[i]}, M_i) + \\ &\quad (K_i - K_t)\log N \\ &= 2N\left[\frac{1}{N}\sum_{j=1}^{N}\log P(x_j|\bar{\theta}^{[t]}, M_t) - \sum_{j=1}^{N}\log P(x_j|\bar{\theta}^{[i]}, M_i)\right] + \\ &\quad (K_i - K_t)\log N \\ &= 2N\left[\frac{1}{N}\sum_{j=1}^{N}\log\left(\frac{f(x_j)}{P(x_j|\bar{\theta}^{[i]}, M_i)}\right) - \right. \\ &\quad \left. \frac{1}{N}\sum_{j=1}^{N}\log\left(\frac{f(x_j)}{P(x_j|\bar{\theta}^{[t]}, M_t)}\right)\right] + (K_i - K_t)\log N \end{aligned}$$

Les deux sommes de la dernière ligne convergent en probabilité respectivement vers $d_{KL}(f, M_i)$ et $d_{KL}(f, M_t)$. Cela suffit à expliquer le comportement asymptotique quand $n \to \infty$, le premier terme de l'ordre de $O(N)$ dominant sur le deuxième en $O(\log N)$. La divergence de KL de M_t par rapport à f est plus petite que celle de M_i par rapport à f. Donc le cœfficient du terme d'ordre N est positif et on a bien $BIC_t - BIC_i > 0$ pour N assez grand.

(b) si $i > t$, on reconnaît que le terme $-(\log P(D|\bar{\theta^{[t]}}, M_t) - \log P(D|\bar{\theta^{[i]}}, M_i))$ s'apparente à un rapport de vraisemblance pour deux modèles, l'un contenant le second. Sous l'hypothèse nulle, cette statistique suit une loi du Chi 2 à $K_i - K_t$ degrés de liberté. On a donc :

$$BIC_t - BIC_i \approx -2\chi^2_{K_i - K_t} + (K_i - K_t)\log N$$

C'est ici le second teme en $O(\log N)$ qui domine et tend vers l'infini quand N tend vers l'infini. Les modèles surajustés $M_{t+1}, \ldots, M_K$ sont eux aussi disqualifiés par le critère BIC.

Cette propriété de convergence vers le quasi-vrai modèle est appelée consistance de BIC vis-à vis de la dimension.

On peut aussi interpréter la probabilité *a posteriori* du modèle :

$$P(M_i|D) = \frac{P(D|M_i)P(M_i)}{\sum_{j=1}^{K} P(D|M_j)P(M_j)}$$

Plaçons nous dans le cas où aucun modèle n'est préféré *a priori*, c'est à dire que $P(M_i) = 1/K$, $\forall i$. Sans cette hypothèse, ce qui suit reste valable dès lors que $n \to \infty$. Les résultats relatifs à l'approximation BIC mènent à l'approximation $P(D|M_i) \propto \exp -1/2(\overbrace{BIC_{max} - BIC_i}^{:=\Delta BIC_i})$. Par la formule des probabilités totales, on en déduit que :

$$P(M_i|D) \approx \frac{\exp -1/2(\Delta BIC_i)}{\sum_{j=1}^{K} \exp -1/2(\Delta BIC_j)}.$$

Cette probabilité tend d'ailleurs vers 1 pour le quasi-vrai modèle tandis qu'elle tend vers 0 pour tout autre.

Cela implique-t-il que le modèle qui réellement engendré les données doive faire partie de la liste des modèles testées ? Cette question a toujours eu des réponses plus ou moins obscures ou alors trop restrictives dans la littérature.

Nous venons de voir qu'il n'en était aucunement besoin pour obtenir la consistance vers le quasi-vrai modèle. Sauf si on est intŕessé par la convergence vers le vrai modèle, auquel cas on identifie celui-ci comme le quasi-vrai modèle.

Si on regarde les applications, supposer l'appartenance d'un modèle réel (pourvu que celui-ci existe) semble même utopique. Seules des simulations peuvent par essence être des phénomènes simples et bien décrits. Dans les champs d'applications que nous pouvons mentionner dans cette thèse, les modèles sont nécessairement imparfaits. En revanche, cela n'entraîne aucune conséquence malheureuse dans le cas de la comparaison de modèles entre eux. Cependant, le quasi-vrai modèle est le meilleur au sens de la divergence de KL parmi la liste étudiée. Il peut donc être relativement éloigné ; la consistance de BIC ne garantit pas la qualité du modèle sélectionné. Celle-ci découle plutôt de la construction qui est proposée par une expertise sur les données.

Finalement l'approximation BIC fournit en général de bons résultats pourvu que les données et les modèles mis en jeu soient raisonnables. Rappelons le caractère asymptotique par rapport à la taille des données de sa construction. Ce critère est assez simple et manipulable pour autoriser la sélection de modèles tout en équilibrant une meilleure adaptation du modèle et un nombre pas trop élevé de paramètres.

B.3 Mise à jour des paramètres spatiaux pour le modèle de champ de Markov caché du chapitre 3

Dans toute cette thèse, nous avons noté :

$$\frac{\partial toto(M)}{\partial M} = (\frac{\partial toto(M)}{\partial M_{ij}})_{(i,j)\in r\times s}, \tag{B.1}$$

si $M \in M_{r,s}(\mathbb{R})$ et $toto : \mathbb{R}^r,\mathbb{R}^s \longrightarrow E$ est une fonction différentiable pour deux entiers positifs r et s. Ce cas inclue celui de la dérivation par rapport à un vecteur en prenant $s = 1$.

Rappelons qu'il s'agit de maximiser :

$$Q_\Delta(\mathbf{\Delta}|\mathbf{\Psi}^{(q)}) = \mathbb{E}_{\mathbf{\Psi}^{(q)}}\left[\log P(\mathbf{Z}|\mathbf{\Delta})|\mathbf{x}^{Obs}\right].$$

Avec les équations (3.19), (3.20) et (3.21), en dérivant par rapport à α_k, il vient :

- $\frac{\partial(3.19)}{\partial\alpha_k} = \sum_{i\in S}\tilde{t}_{ik}^{(q)}$,
- $\frac{\partial(3.20)}{\partial\alpha_k} = 0$ et
- $\frac{\partial(3.21)}{\partial\alpha_k} = -\sum_{i\in S}\frac{\exp(\alpha_k+\sum_{j\in\nu(i)}\sum_{l=1}^K B_{kl}\tilde{z}_{jl}^{(q+1)})}{\sum_{m=1}^K\exp(\alpha_m+\sum_{j\in\nu(i)}\sum_{l=1}^K B_{ml}\tilde{z}_{jl}^{(q+1)})}$.

On en déduit donc la formule à vérifier par les α_k et $\mathbb{B}$ annoncées dans la section 3.7.1.

Si on dérive par rapport à B_{pr}, on déduit :

- $\frac{\partial(3.19)}{\partial B_{pr}} = 0$,
- $\frac{\partial(3.20)}{\partial B_{pr}} = \frac{\partial}{\partial B_{pr}}\left(\sum_{i\in S}\sum_{j\in\nu(i)}\sum_{k=1}^K\sum_{m=1}^K\tilde{t}_{ik}^{(q+1)}B_{km}\tilde{z}_{jm}^{(q+1)}\right)$ qui vaut
  $\sum_{i\in S}\sum_{j\in\nu(i)}(\tilde{t}_{ip}^{(q)}\tilde{z}_{jr}^{(q)} + \tilde{t}_{ir}^{(q)}\tilde{z}_{jp}^{(q)})$ si $p \neq r$
  et $\sum_{i\in S}\sum_{j\in\nu(i)}\tilde{t}_{ip}^{(q)}\tilde{z}_{jp}^{(q)}$ si $p = r$.
- Nous détaillons $\frac{\partial(3.21)}{\partial B_{pr}}$ ci dessous.

$$\begin{aligned}
\frac{\partial(3.21)}{\partial B_{pr}} &= -\sum_{i\in S}\frac{\partial}{\partial B_{pr}}\log\sum_{l=1}^K\left[\exp(\alpha_l + \sum_{j\in\nu(i)}\sum_{m=1}^K B_{lm}\tilde{z}_{jm}^{(q+1)})\right] \\
&= -\sum_{i\in S}\frac{\frac{\partial}{\partial B_{pr}}\sum_{l=1}^K\exp(\alpha_l + \sum_{j\in\nu(i)}\sum_{m=1}^K B_{lm}\tilde{z}_{jm}^{(q+1)})}{\sum_{l=1}^K\exp(\alpha_l + \sum_{j\in\nu(i)}\sum_{m=1}^K B_{lm}\tilde{z}_{jm}^{(q+1)})} \\
&= \begin{cases} -\sum_{i\in S}\frac{(\sum_{j\in\nu(i)}\tilde{z}_{jp}^{(q)})U_r+(\sum_{j\in\nu(i)}\tilde{z}_{jr}^{(q)})U_p}{\sum_{l=1}^K\exp(\alpha_l+\sum_{j\in\nu(i)}\sum_{m=1}^K B_{lm}\tilde{z}_{jm}^{(q+1)})} & \text{si } p\neq r, \\ -\sum_{i\in S}\frac{(\sum_{j\in\nu(i)}\tilde{z}_{jp}^{(q)})\exp(\alpha_p+\sum_{j\in\nu(i)}\sum_{m=1}^K B_{pm}\tilde{z}_{jm}^{(q+1)})}{\sum_{l=1}^K\exp(\alpha_l+\sum_{j\in\nu(i)}\sum_{m=1}^K B_{lm}\tilde{z}_{jm}^{(q+1)})} & \text{si } p=r, \end{cases}
\end{aligned}$$

où $U_{r/p} = \exp(\alpha_{r/p} + \sum_{j\in\nu(i)} \sum_{m=1}^{K} B_{r/p\,m} \tilde{z}_{jm}^{(q+1)})$. En combinant les trois dérivées de (3.19), (3.20) et (3.21) par rapport à B_{pr}, on trouve :

- si $p = r$:

$$\frac{\partial Q_\Delta}{\partial B_{pp}} = \sum_{i\in S} (\sum_{j\in\nu(i)} \tilde{z}_{jp}^{(q)}) \left(\tilde{t}_{ip}^{(q)} - \frac{\exp(\alpha_p + \sum_{j\in\nu(i)} \sum_{m=1}^{K} B_{pm} \tilde{z}_{jm}^{(q+1)})}{\sum_{l=1}^{K} \exp(\alpha_l + \sum_{j\in\nu(i)} \sum_{m=1}^{K} B_{lm} \tilde{z}_{jm}^{(q+1)})} \right)$$

- si $p \neq r$:

$$\begin{aligned} \frac{\partial Q_\Delta}{\partial B_{pp}} &= \sum_{i\in S} (\sum_{j\in\nu(i)} \tilde{z}_{jp}^{(q)}) \left(\tilde{t}_{ir}^{(q)} - \frac{\exp(\alpha_r + \sum_{j\in\nu(i)} \sum_{m=1}^{K} B_{rm} \tilde{z}_{jm}^{(q+1)})}{\sum_{l=1}^{K} \exp(\alpha_l + \sum_{j\in\nu(i)} \sum_{m=1}^{K} B_{lm} \tilde{z}_{jm}^{(q+1)})} \right) \\ &+ \sum_{i\in S} (\sum_{j\in\nu(i)} \tilde{z}_{jr}^{(q)}) \left(\tilde{t}_{ip}^{(q)} - \frac{\exp(\alpha_p + \sum_{j\in\nu(i)} \sum_{m=1}^{K} B_{pm} \tilde{z}_{jm}^{(q+1)})}{\sum_{l=1}^{K} \exp(\alpha_l + \sum_{j\in\nu(i)} \sum_{m=1}^{K} B_{lm} \tilde{z}_{jm}^{(q+1)})} \right) \end{aligned}$$

On en déduit de même les formules annoncées dans la section 3.7.1 à vérifier par les α et $\mathbb{B}$.

B.4 Mise à jour des paramètres dans le cas de données manquantes

On rappelle que pour des distributions normales :

$$P(x_i|\theta_k) = P(x_i^{Obs_i}, x_i^{Man_i}|\theta_k) = \frac{1}{2}(Cste - \log|\Sigma_k| - (x_i - \mu_k)^t)\Sigma_k^{-1}(x_i - \mu_k).$$

B.4.1 Mise à jour des centres (moyennes) des classes :

On va se servir de la formule (pour Σ une matrice carrée inversible et μ un vecteur de dimension le nombre de lignes de Σ) :

$$\frac{\partial (x-\mu)^t \Sigma^{-1} (x-\mu)}{\partial \mu} = -(\Sigma^{-1} + (\Sigma^{-1})^t).(x-\mu);$$

Et cette formule se simplifie puisque Σ est une matrice de covariance donc symétrique : $\Sigma^{-1} + (\Sigma^{-1})^t = 2\Sigma^{-1}$.

En réordonnant les indices si nécessaire pour avoir une écriture plus simplifiée des expressions :

$$\frac{\partial \log P(X_i|\theta_k)}{\partial \mu_k} = \Sigma_k^{-1} \begin{pmatrix} x_i^{Obs_i} - \mu_k^{Obs_i} \\ \nu_{i,k}^{(q)} - \mu_k^{Man_i} \end{pmatrix}$$

Finalement :

$$\frac{\partial Q_\theta(\theta|\mathbf{\Psi}^{(q)})}{\partial \mu_k} = \sum_i \sum_k t_{i,k}^{(q)} \Sigma_k^{-1} \begin{pmatrix} x_i^{Obs_i} - \mu_k^{Obs_i} \\ \nu_{i,k}^{(q)} - \mu_k^{Man_i} \end{pmatrix}$$

Et cette expression s'annule si et seulement si : (en prenant la convention $0 \times a = 0$ si $a \in Man_i$) :

$$\mu_k^{(q+1)} = \frac{\sum_i t_{i,k}^{(q)} (r_i.x_i + (1 - r_i).\eta_{i,k}^{(q)})}{\sum_i t_{i,k}^{(q)}},$$

qui est bien le condition annoncée dans la partie 4.1.

B.4.2 Mise à jour de la matrice de co-variance

On établit que (rappel : Σ_k est symétrique en tant que matrice de covariance) :

$$\frac{\partial \log |\Sigma_k|}{\partial \Sigma_k^{-1}} = -\frac{\partial \log |\Sigma_k^{-1}|}{\partial \Sigma_k^{-1}} = -(\Sigma_k)^t = -\Sigma_k$$

$$\frac{\partial (X_i - \mu_k)^t \Sigma_k^{-1} (X_i - \mu_k)}{\partial \Sigma_k^{-1}} = (X_i - \mu_k)(X_i - \mu_k)^t,$$

donc : $\frac{\partial \log P(X_i|\theta_k)}{\partial \Sigma_k^{-1}} = \frac{1}{2}\left[\Sigma_k - (X_i - \mu_k)(X_i - \mu_k)^t\right]$.

En ordonnant les indices de façon à avoir les valeurs observées Obs_i pour l'individu i avant les valeurs manquantes Man_i :

$$\frac{\partial}{\partial \Sigma_k^{-1}} \int P(X_i^{Man_i}|x_i^{Obs_i}, \theta_k^{(q)}) \log P(x_i^{Obs_i}, X_i^{Man_i}|\theta_k) dX_i^{Man_i} =$$

$$\frac{1}{2}\left[\Sigma_k - \begin{pmatrix} (x_i^{Obs_i} - \mu_k^{Obs_i})(x_i^{Obs_i} - \mu_k^{Obs_i})^t & (x_i^{Obs_i} - \mu_k^{Obs_i})(\eta_{i,k}^{(q)} - \mu_k^{Man_i})^t \\ (\eta_{i,k}^{(q)} - \mu_k^{Man_i})(x_i^{Obs_i} - \mu_k^{Obs_i})^t & (\eta_{i,k}^{(q)} - \mu_k^{Man_i})(\eta_{i,k}^{(q)} - \mu_k^{Man_i})^t + \Gamma_{i,s}^{(q)} \end{pmatrix}\right].$$

Il s'ensuit :

$$2\frac{\partial Q_\theta(\theta|\mathbf{\Psi}^{(q)})}{\partial \Sigma_k} = \sum_{i \in S} \sum_k t_{i,k}^{(q)}.$$

$$\left[\Sigma_k - \begin{pmatrix} (x_i^{Obs_i} - \mu_k^{Obs_i})(x_i^{Obs_i} - \mu_k^{Obs_i})^t & (x_i^{Obs_i} - \mu_k^{Obs_i})(\eta_{i,k}^{(q)} - \mu_k^{Man_i})^t \\ (\eta_{i,k}^{(q)} - \mu_k^{Man_i})(x_i^{Obs_i} - \mu_k^{Obs_i})^t & (\eta_{i,k}^{(q)} - \mu_k^{Man_i})(\eta_{i,k}^{(q)} - \mu_k^{Man_i})^t + \Gamma_{i,s}^{(q)} \end{pmatrix}\right]$$

Finalement on a la condition :

$$\frac{\partial Q_\theta(\theta|\mathbf{\Psi}^{(q)})}{\partial \Sigma_k} = 0 \Leftrightarrow \Sigma_k^{(q+1)} = \frac{\sum_i t_{i,k}^{(q)} S_{i,k}^{(q+1)}}{\sum_i t_{i,k}^{(q)}}$$

avec la notation pour tout $(i,k) \in S \times [1,...,K]$ et $(s,t) \in [1,...,D]$:

$$\begin{aligned}(S_{i,k})_{st} &= r_i.r_i^t \otimes (x_i) - \mu_k^{(q)}).(x_i - \mu_k^{(q)})^t \\ &+ r_i(1-r_i)^t \otimes (x_i - \mu_k^{(q)}).(\eta_{i,k}^{(q)} - \mu_k^{(q)})^t \\ &+ (1-r_i).r_i^t \otimes (\eta_{i,k}^{(q)} - \mu_k^{(q)}).(x_i - \mu_k^{(q)})^t \\ &+ (1-r_i).(1-r_i)^t \otimes (\eta_{i,k}^{(q)} - \mu_k^{(q)}).(\eta_{i,k}^{(q)} - \mu_k^{(q)})^t + \Gamma_{i,k}^{(q)}\end{aligned}$$

Ce qui est bien la forme annoncée à la fin de la partie 4.1

C. PUBLICATIONS ASSOCIÉES À LA THÈSE

C.1 Classification de gènes via des modèles de Markov qui intègrent données individuelles et données d'interactions

Gene clustering via integrated Markov models combining individual and pairwise features

Matthieu Vignes and Florence Forbes

***Abstract*— Clustering of genes into groups sharing common characteristics is a useful exploratory technique for a number of subsequent computational analysis. A wide range of clustering algorithms have been proposed in particular to analyze gene expression data, but most of them consider genes as independent entities or include relevant information on gene interactions in a sub-optimal way.**

We propose a probabilistic model that has the advantage to account for individual data (*eg.* expression) and pairwise data (*eg.* interaction information coming from biological networks) simultaneously. Our model is based on hidden Markov random field models in which parametric probability distributions account for the distribution of individual data. Data on pairs, possibly reflecting distance or similarity measures between genes, are then included through a graph where the nodes represent the genes and the edges are weighted according to the available interaction information. As a probabilistic model, this model has many interesting theoretical features. Also, preliminary experiments on simulated and real data show promising results and points out the gain in using such an approach.

Availability: The software used in this work is written in C++ and is available with other supplementary material at `http://mistis.inrialpes.fr/people/forbes/transparentia/supplementary.html`.

***Index Terms*— Markov random fields, model-based clustering, metabolic networks, gene expression.**

I. Introduction

As an increasing amount of post-genomic data is available, there is a great need to develop methodologies to analyze and to use theinformation contained in this data. A major challenge in bioinformatics is to reveal interactions between components of living organisms and discover the corresponding networks responsible for their biological complexity. In this framework, clustering of genes into groups sharing common characteristics is a useful exploratory technique. It is frequently used as the basis for further computational analysis. As an example, the function of a gene can be predicted according to known functions of other genes from the same cluster. With the introduction of DNA microarray technology, researchers are now able to measure the expression levels of thousands of genes simultaneously at various time points of the biological process or under various experimental conditions. As data accumulate, the tendency to investigate general regulatory mechanisms by clustering genes from their expression profiles increases. A wide range of clustering algorithms have been proposed to analyze gene expression data. Various methods have been applied such as hierarchical clustering [9], self-organizing maps [21], k-means algorithms [23], and more recently Support Vector Machines methods [5] or graph analysis by bi-clustering [22]. More generally, approaches fall mainly in two categories. Some focus on individual data and assume that they are independent. Typically, [26] use a statistically based model which does not incorporate possible relationships between genes. Others try to integrate several sources of data, setting for instance, expression data into a Bayesian graphical model framework [12], combining expression data with phylogenetic profiles [18], or defining distances between genes combining different data types [16]. Typically, the procedure in the work of [11] uses information on pairs of genes in the form of networks or graphs and combines it with distances computed from individual expression profiles. This requires transforming individual information into distances or similarity measures and does not directly use individual data associated to genes in the networks, loosing some potentially interesting information in the process. Kernel-based approaches to data fusion ([15], [25], [24]) also consist of representing various data sets via kernel functions which define generalized similarity relationships. Also, sequential procedures that cluster first individual data alone and incorporate additional information only after the clusters are determined are necessarily suboptimal.

It appears that models able to integrate simultaneously information on individuals (without reducing it to pairwise information) and pairwise relationships in the same procedure have not yet been proposed. The novelty of our work is to propose a model-based approach, as opposed to the distance-based approaches mentioned above, to take into account simultaneously data from individual genes, *ie.* data that make sense and exist for each genes, and data from pairs of genes reflecting for instance some distance or some similarity measure defined on the genes, possibly using some recent kernel-based approaches. To our knowledge, the only similar attempts have been proposed in [19]. However, the formulation of their probabilistic model does not fully exploit gene dependencies. It is written to account for gene interaction but one of the assumptions made is only valid under gene independence. In addition, no estimation procedure is proposed to estimate the model parameters and they then need to be fixed to arbitrary values. We propose an integrated Markov model, meaning by that a specific instance and usage of a Hidden Markov model. Parametric probability distributions account for the distribution of individual data while data on pairs are included through a graph where the nodes represent the genes and the edges are weighted according to the available interaction information (*eg.* distances or similarity measures between genes). As regards parameter estimation and classification step, we consider recent procedures based on the EM algorithm and *mean field*-like approximations [7]. Such procedures were shown to be more efficient in many ways than standard Gibbs samplers or Markov Chain Monte Carlo (MCMC) techniques traditionally used in computer vision.

Manuscript received July 24, 2006; revised February 22, 2007.

M. Vignes is with BioSS at the Scottish Crop Research Institute, Invergowrie, Dundee, DD2 5DA, Scotland, UK.
E-mail: matthieu@bioss.ac.uk

F. Forbes is head of team Mistis, INRIA Rhône-Alpes, ZIRST, 655, avenue de l'Europe, Montbonnot, 38334 Saint Ismier Cedex, France.
E-mail: florence.forbes@inrialpes.fr

This model and the EM classification framework (Section II) have many interesting features. As a probabilistic model, it leads to various possible statistical criteria to select automatically the number of clusters and it provides confidence measures such as posterior probabilities that an object (*eg.* a gene) is assigned to a class. It is flexible in that various pairwise relationship information and features on individual data can be easily incorporated possibly with different weights. Its generalization to include missing data, that often occur when dealing with expression data, is straightforward and its extension to overlapping clustering methods, to deal with more realistic situations where genes can belong to many groups at the same time, can also be considered. Although such a model is relevant in various other applications, we specify in Section III the type of data used in this work. Experiments on simulated data are reported and results on real data are then shown in Section IV. A discussion section ends the paper.

II. Integrated Markov models

The basic assumption is that measures (*e.g.* expression profiles) corresponding to each objects are random variables with a specific probability distribution in each class. A standard way to represent class-specific density functions is to approximate them as Gaussian distributions whose parameters depend on the class. In the work of [26], a Gaussian mixture model is assumed which corresponds to Gaussian class-specific distributions but also to genes independence. This is not fully satisfying since strong neighbourhood relationships between genes sharing common functions can exist. To overcome this limitation, we propose to improve on the Gaussian mixture model by assuming that the distribution of the observed features is that of a Hidden Markov Random Field (HMRF) with K components and appropriate parametrization. To define such a model, one needs to specify a neighbourhood structure indicating which genes are statistically linked but this structure is not necessarily related to the clusters. Dependent genes may be in different classes and genes from the same class may be independent.

A. Hidden Markov fields for biological networks

Let n be the number of genes to be clustered and $x_1, \ldots, x_n$ denote the individual data observed for the genes numbered by $\{1, \ldots, n\}$. The observed data are usually multi-dimensional vectors, *e.g.* expression profiles. For $i = 1, \ldots, n$, we model the probability of observing x_i as

$$P(x_i|\mathbf{\Psi}) = \sum_{k=1}^{K} P(Z_i = c_k|\beta)\ f(x_i|\theta_k),$$

where $f(x_i|\theta_k)$ denotes the multivariate Gaussian distribution with parameters θ_k namely the mean μ_k and covariance matrix Σ_k. Notation Z_i denotes the random variable representing the class of gene i. Z_i can take values in $\{c_k, k = 1 \ldots K\}$ denoting the K possible classes. More specifically, it is convenient to consider c_k as a K-dimensional indicator vector with all components being 0 except the k^{th} which is 1. Note that we assume in this section that K is fixed but this can be generalized (see Section II-B). Notation β denotes additional parameters defining the distribution of the Z_i's and $\mathbf{\Psi}$ denotes the whole model parameters *i.e.* $\mathbf{\Psi} = (\theta_k, \beta, k = 1 \ldots K)$. As an example, the model used by [26] for $P(x_i|\mathbf{\Psi})$ is an *independent* Gaussian mixture model and corresponds, in our framework, to assume that the Z_i's are independent variables. Our approach differs in that our aim is to account for dependencies. This requires the definition of neighbourhood relationships between genes. We will think of a set of genes as a graph with edges emanating from each gene to other genes within its neighbourhood. We will illustrate in Section III how such a graph can be built from biological network data. The dependencies between neighbouring genes are then modelled by further assuming that the joint distribution of $Z_1, \ldots, Z_n$ is a discrete Markov Random Field on this specific graph. Denoting $\mathbf{z} = (z_1, \ldots, z_n)$ specified values of the Z_i's, we define

$$P(\mathbf{z}|\beta) \quad = \quad W(\beta)^{-1} \exp(-H(\mathbf{z}, \beta)) \tag{1}$$

where $W(\beta)$ is a normalizing constant and H is a function assumed to be of the following form (we restrict to pair-wise interactions), $H(\mathbf{z}, \beta) = \sum_{i=1}^{n} V_i(z_i, \beta) + \sum_{\substack{i,j \\ i \sim j}} V_{ij}(z_i, z_j, \beta)$, where the V_i's and V_{ij}'s are respectively functions referred to as singleton and pair-wise potentials. We write $i \sim j$ when genes i and j are neighbours on the graph, so that the second sum above is over neighbouring genes. Parameters β consist of two sets $\beta = (\alpha, \mathbb{B})$ where α and $\mathbb{B}$ are defined as follows. We consider pair-wise potentials V_{ij} that depend on z_i and z_j but also possibly on i and j. Since the z_i's can only take a finite number of values, for each i and j, we can define a $K \times K$ matrix $\mathbb{B}_{ij} = (\mathbb{B}_{ij}(k,l))_{1 \leq k,l \leq K}$ and write without lost of generality $V_{ij}(z_i, z_j, \beta) = -\mathbb{B}_{ij}(k,l)$ if $z_i = c_k$ and $z_j = c_l$. Using the indicator vector notation and denoting z_i^t the transpose of vector z_i, it is equivalent to write $V_{ij}(z_i, z_j, \beta) = -z_i^t \mathbb{B}_{ij} z_j$. This latter notation has the advantage to make sense also when the vectors are arbitrary and not necessarily indicators. This will be useful when describing the algorithms of Section II-C. Similarly we consider singleton potentials V_i that may depend on z_i and on i, so that denoting by α_i a K-dimensional vector, we can write $V_i(z_i, \beta) = -\alpha_i(k)$ if $z_i = c_k$, where $\alpha_i(k)$ is the k^{th} component of α_i, or equivalently $V_i(z_i, \beta) = -z_i^t \alpha_i$. This vector α_i acts as weights for the different values of z_i. When α_i is zero, no class is favored, *i.e.* for a given gene i, if no information on the neighbouring genes is available, then all classes have the same probability. If in addition, for all i and j, $\mathbb{B}_{ij} = b \times I_K$ where b is a real scalar and I_K is the $K \times K$ identity matrix, parameters β reduce to a single scalar interaction parameter b and we get the Potts model traditionally used for image segmentation. Note that this model is probably the more appropriate for classifying genes since it tends to favor neighbours that are in the same class. However, cases where the $\mathbb{B}_{ij}$'s are far from $b \times I_K$ could be useful in situations where neighbouring genes are likely to be in different classes. Also, when distance or similarity data, $(d_{ij})_{i,j=1,\ldots,n}$, between genes are available, $\mathbb{B}_{ij}(k,l)$ can be decomposed as $\mathbb{B}_{ij}(k,l) = F(d_{ij})\, c(k,l)$ where F is a non increasing function of $\mathbb{R}^+$ and $c(k,l)$ corresponds to a gain (or a loss depending on its sign) of assigning genes i and j respectively to class c_k and c_l. This is part of the flexibility and modelling capabilities of the model. However, without specific information, we can choose $c(k,l) = b$ if $k = l$ and $c(k,l) = 0$ otherwise. In this case, parameter b can be interpreted as a strength of interaction between neighbours. The higher b the more weight is given to the interaction graph. If b is set to 0, only the individual features are taken into account, reducing our model to traditional existing approaches. In practice, these parameters can be tuned

according to expert or *a priori* knowledge or they can be estimated from the data. In the first case, our software can deal with the most general parametrization, namely $\beta = (\alpha_i, \mathbb{B}_{ij})$. In the latter case, the part to be estimated is usually assumed independent of the genes indices i and j, so that in what follows we will reduce α and $\mathbb{B}$ respectively to a single vector and a single matrix. Note that in Section IV, the model is further reduced to α equal to 0 and $\mathbb{B}$ equal to $b \times I_K$ (see comments in this section).

Meanwhile, to keep a general presentation, the observed data is then represented by an HMRF defined by parameters $\mathbf{\Psi}$ being $\mathbf{\Psi} = (\{\mu_k, \Sigma_k\}_{k=1,\ldots,K}, \alpha, \mathbb{B})$.

B. Selecting the number of classes

Choosing the probabilistic model that best accounts for the observed data is an important first step for the quality of the subsequent estimation and classification stages. In statistical problems, a commonly used selection criterion is the Bayesian Information Criterion (BIC) of [20]. The BIC is computed given the data $\mathbf{x}$ and a model M with parameters $\mathbf{\Psi}$. It is defined by:

$$BIC(M) = 2 \log P(\mathbf{x} \mid \mathbf{\Psi}^{ml}) - d \log n \ ,$$

where $\mathbf{\Psi}^{ml}$ is the maximum likelihood estimate of $\mathbf{\Psi}$, $\mathbf{\Psi}^{ml} = \arg\max_{\mathbf{\Psi}} P(\mathbf{x} \mid \mathbf{\Psi}, M)$, d is the number of free parameters in model M and n is the number of observations. BIC allows comparison of models with differing parametrizations and/or differing number of classes. Many other approaches can be found in the literature on model selection but BIC has become quite popular due to its simplicity and its good results. In this study, we will consider the Markov model (α, $\mathbb{B}$) as fixed. More specifically, the experiments reported in Section IV correspond to the simplest model with $\alpha = 0$ and $\mathbb{B} = b \times I_K$. More important in practice is the choice of K and of the covariance model (Σ_k's). For multivariate Gaussian class-specific distributions, there exists a number of different choices for the Σ_k's. See [1] and [6] for a description of these forms and their meaning. The simplest models are those for which the Σ_k's are diagonal matrices. Our choice of K and Σ_k's then can be based on BIC. However, for HMRFs, its exact computation is not tractable due to the dependence structure induced by the Markov model. A possibility is then to compute BIC for independent mixture models, forgetting any spatial information. Not to loose such information, we rather choose to use the mean field like approximations of BIC proposed by [10] (see Section II-C for additional details). As regards covariance matrices, we restrict to diagonal models in most cases or consider an original reduction dimension techniques [4]. In the context of the present work however, we did not observe significant improvement over the simple diagonal models for the data (only 10 dimensional) we consider in Section IV.

C. Classifying genes

Our aim is to classify each gene in one of the K classes. To do so we consider a Maximum Posterior Marginal (MPM) principle consisting in assigning gene i to class c_k that maximizes $P(Z_i = c_k | \mathbf{x}, \mathbf{\Psi})$. Such maximizations depend on $\mathbf{\Psi}$ which is usually unknown, or partly unknown when prior knowledge can be incorporated, and has to be estimated. The parameters to be estimated are the parameters defining the Gaussian distributions namely the μ_k and Σ_k for $k = 1, \ldots K$ and the parameters defining the interaction model, namely the $\alpha(k)$ for $k = 1, \ldots K$ and the $K \times K$ dimensional matrix $\mathbb{B}$. The EM algorithm is a commonly usedalgorithm for parameter estimation in problems with missing data (here the class assignments). For independent mixture models, the independence assumption leads to an easy implementation of the algorithm. For HMRFs, due to the dependence structure, the exact EM is not tractable and approximations are required. In this paper, we use some of the approximations presented in [7]. These approximations are based on the mean field principle which consists in replacing the intractable Markov distributions by factorized ones for which the exact EM can be carried out. This allows to take the Markovian structure into account while preserving the good features of EM. [7] generalized the mean field principle and introduced different factorized models resulting in different procedures. Note that in practice, these algorithms have to be extended to incorporate the estimation of matrix $\mathbb{B}$ and to include irregular neighbourhood structure coming from biological networks and not from regular pixel grids like in [7].

Briefly, these algorithms can be presented as follow. They are based on the EM algorithm which is an iterative algorithm aiming at maximizing the log-likelihood (for the observed variables $\mathbf{x}$) of the model under consideration by maximizing at each iteration the expectation of the complete log-likelihood (for the observed and hidden variables $\mathbf{x}$ and $\mathbf{z}$) knowing the data and a current estimate of the model parameters. When the model is an Hidden Markov Model with parameters $\mathbf{\Psi}$, there are two difficulties in evaluating this expectation. Both the normalizing constant $W(\beta)$ in (1) and the conditional probabilities $P(z_i \mid \mathbf{x}, \mathbf{\Psi})$ and $P(z_i, z_j \mid \mathbf{x}, \mathbf{\Psi})$ for j in the neighbourhood $N(i)$ of i, cannot be computed exactly. Informally, the mean field approach consists in approximating the intractable probabilities by neglecting fluctuations from the mean in the neighbourhood of each gene i. More generally, we talk about mean field-like approximations when the value for gene i does not depend on the value for other genes which are all set to constants (not necessarily to the means) independently of the value for gene i. These constant values denoted by $\tilde{z}_1, \ldots, \tilde{z}_n$ are not arbitrary but satisfy some appropriate consistency conditions (see [7]). Let $z_{N(i)}$ denote the set of variables $\{z_j, j \in N(i)\}$ associated to the set $N(i)$ of neighbours of i. It follows that $P(z_i \mid \mathbf{x}, \mathbf{\Psi})$ is approximated by

$$\begin{aligned} P(z_i \mid \mathbf{x}, \tilde{z}_{N(i)}, \mathbf{\Psi}) &\propto f(x_i | z_i^t \theta).P(z_i | \tilde{z}_{N(i)}, \beta) \\ &\propto f(x_i | z_i^t \theta). \\ &\quad \exp[z_i^t (\alpha + \mathbb{B} \sum_{j \in N(i)} \tilde{z}_j)], \end{aligned}$$

where θ denotes the vector $(\theta_1, \ldots, \theta_K)$. The normalizing constant is not specified but its computation is not an issue. Then, for all $j \in N(i)$, $P(z_i, z_j \mid \mathbf{x}, \mathbf{\Psi})$ is approximated by $P(z_i \mid \mathbf{x}, \tilde{z}_{N(i)}, \mathbf{\Psi})\ P(z_j \mid \mathbf{x}, \tilde{z}_{N(j)}, \mathbf{\Psi})$. Both approximations are easy to compute. Using such approximations leads to algorithms which in their general form consist in repeating two steps. At iteration q,

(1) Create from the data $\mathbf{x}$ and some current parameter estimates $\mathbf{\Psi}^{(q-1)}$ a configuration $\tilde{z}_1^{(q)}, \ldots \tilde{z}_n^{(q)}$, *i.e.* values for the Z_i's. Replace the Markov distribution $P(\mathbf{z}|\beta)$ of (1) by the factorized distribution $\prod_{i=1}^{n} P(z_i | \tilde{z}_{N(i)}^{(q)}, \beta)$. It follows that the joint distribution $P(\mathbf{x}, \mathbf{z} | \mathbf{\Psi})$ can also be approximated by a factorized

distribution:

$$\prod_{i=1}^{n} f(x_i|z_i^t\theta)P(z_i|\tilde{z}_{N(i)}^{(q)},\beta)$$

and the two problems encountered when considering the EM algorithm with the exact joint distribution disappear. The second step is therefore,

(2) Apply the EM algorithm for this factorized model with starting values $\mathbf{\Psi}^{(q-1)}$, to get updated estimates $\mathbf{\Psi}^{(q)}$ of the parameters.

In particular the *mean field* and *simulated field* algorithms consist in two different ways of performing step **(1)**. The *mean field* algorithm consists in updating the $\tilde{z}_i^{(q)}$'s by setting, for all $i = 1,\ldots,n$, $\tilde{z}_i^{(q)}$ to the mean of distribution $P(z_i \mid \mathbf{x}, \tilde{z}_{N(i)}^{(q)}, \mathbf{\Psi}^{(q-1)})$. Note that as z_i is an indicator vector, the mean value $\tilde{z}_i^{(q)}$ is a vector made of the respective probabilities to be in each of the K classes. In the *simulated field* algorithm, $\tilde{z}_i^{(q)}$ is simulated from $P(zi \mid \mathbf{x}, \tilde{z}_{N(i)}^{(q)}, \mathbf{\Psi}^{(q-1)})$. Note also that to save additional notation, the updating described above is synchronous while we actually implemented a sequential updating of the $\tilde{z}_i^{(q)}$'s: each node i is updated in turn using the new values of the other nodes as soon as they become available rather than waiting until all nodes have been updated. Intuitively, the stochastic feature of the *simulated algorithm* enables to avoid convergence towards a saddle point or a local minimum, dependence of the converging state towards initialization or slow rate of convergence. These well-documented pitfalls of the EM algorithm are sorted out in the spirit of Stochastic EM (SEM [27]). However both algorithm cannot be really compared since SEM is not Markovian: it doesn't account for dependencies between observations and is well suited to deal with mixture models.

In practice, at step **(2)**, performing one EM iteration is usually enough. Then, the HMRF estimation provides us with estimations for the means and covariance matrices of the K Gaussian distributions, namely μ_k and Σ_k for $k = 1,\ldots K$, but also for the hidden field parameters, matrix $\mathbb{B}$ and vector α. It follows easily approximations of the $P(Z_i = c_k|\mathbf{x},\mathbf{\Psi})$ required to classify each genes using the MPM principle.

In this work, we mainly consider the so-called *simulated field* (SF) algorithm for its better performance in practice.

III. From biological networks to interaction graphs

Many kinds of biological networks are freely available. They contain a lot of information that should not be ignored to provide optimal clustering but the quality and the access to that information is far from being uniform. As an example, biological networks are not all related to the same objects. They may contain links between genes, gene products, proteins complexes or families, *etc.* and the links may stand for experimentally based or assumed relationships. Our goal is to build a graph with objects which are individually subject to other measurements, as genes are to microarrays. There is no universal way to build such a graph but we give an illustration in this section. We choose to focus on gene expression data and metabolic networks like those given in the Reaction (part of Ligand) KEGG database (`http://www.genome.ad.jp/kegg/reaction/`). A mapping between genes and objects in the network must then be derived. Chemical reactions of interest are those which are assigned one or several EC numbers corresponding to enzymes that may catalyze them. Toeach EC number are associated one or more genes.

A first stage consists in building a graph whose nodes are enzymes. An edge exists between two enzymes if and only if they catalyze two reactions that share at least a common chemical compound either as substrate or product. The interpretation is that an edge stands for the possibility that two reactions follow each other in metabolic pathways. However, all the links between reactions cannot be considered. In particular those which involve compounds that are very common (*eg.* water, *etc.*) are usually not relevant to the biological interpretation and may hide or bias the biological information. Two possibilities are either to use the main compounds (according to KEGG database) or to remove compounds which would link too many reactions (above a given threshold). We choose the first solution for the restricted database has the advantage of being produced by experts who manually removed somewhat irrelevant compounds such as water, carbon dioxide, *etc.* In addition, weights $(d_{ij})_{i,j=1,\ldots,n}$ can be assigned to edges. They may reflect in a quantitative way enzymes proximity or thermodynamical properties. Such information is not available yet but would be easily dealt with in our model. As an illustration, we consider the *Saccharomyces cerevisiae* genome and derive a graph on genes with the network of chemical reactions given in the database. Figure 1 gives an example based on the compounds acting in the `Citrate cycle` with only part of the reactions represented for clarity (Figure 1 (A)). For example, reactions $R00479$, $R01900$ and $R00709$ all share compound `Isocitrate`. They are therefore neighbours and so are the enzymes catalyzing each of these reactions (Figure 1 (B)). Two enzymes catalyzing the same reaction are neighbours as well (*eg.* $EC\,2.3.3.8$ and $EC\,2.3.3.1$). Reversibility is allowed. Also a common enzyme may catalyze different reactions, *eg.* $EC1.1.1.42$ is active in reactions $R01899$, $R00268$ and $R00267$. It must then be linked to any enzyme catalyzing reactions sharing a compound with the later.

A second stage in building the final graph is to go from enzymes to genes. Two cases have to be considered. In the first one a gene maps to several enzymatic functions while in the second one several genes map to a single enzymatic EC number. A way to deal with both cases is to consider couples of objects $(gene, EC)$ and connect them in the graph as soon as their second components are connected. In the first case, enzymes already correspond to different nodes. These nodes only need to be fused keeping neighbourhood relationships. The measure about the gene is then assigned to the resulting node. The second case is illustrated in the transition from graph (B) to graph (C) of Figure 1. Links (see graph (B)) between enzymes correspond to solid lines while each set of associated genes corresponds to dotted lines. A node is added for each of the gene corresponding to the same enzymatic function. New nodes are then linked to keep the same relationships than that existing between enzymes. In our example (Figure 1 (C)), $EC\,4.2.1.3$ splits into genes $YJL200C$ and $YLR304C$. Note that information related to $EC\,2.3.3.8$ is lost because no known yeast gene is assigned to that enzyme. Besides an obvious limitation of our graph construction is that it ignores genes not related to EC numbers. Many of them (*eg.* regulators) can be responsible for relevant interactions. A more complete (and less automatic) graph construction would have required additional expert knowledge not available in this study. Note that beyond the biological relevance, the size of the graph is not a problem. The model can deal with large numbers of genes, edges and experiments. In different contexts, experiments

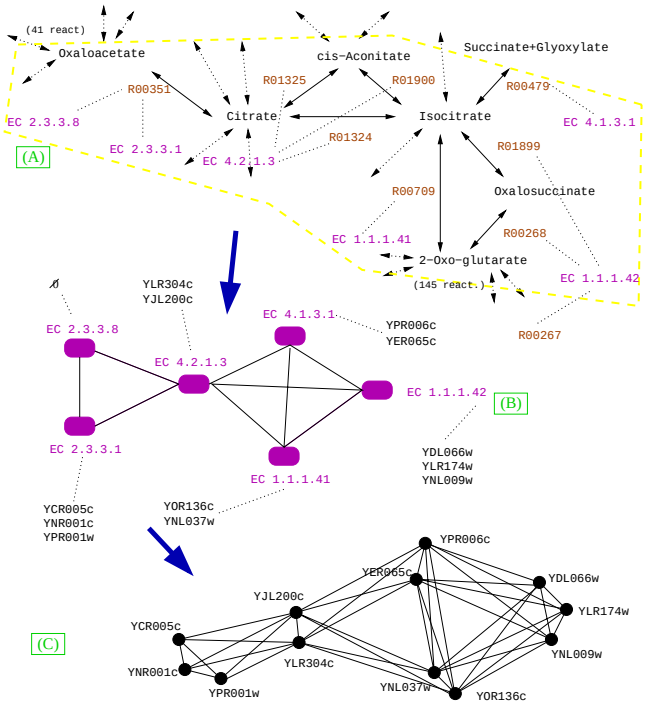


Fig. 1. From the graph of chemical metabolic reactions (A) to the gene interaction network (C) *via* the enzyme network (B). For clarity, only edges between reactions in the metabolic subgraph (A) are represented.

were carried out with the equivalent of thousands of genes and edges and up to 300 experimental conditions (dimension of the data) using diagonal covariance matrices or dimension reduction techniques of [4].

IV. Results

As mentioned earlier, the experiments reported in this section correspond to the simplest Markov model with $\alpha = 0$ and $\mathbb{B} = b \times I_K$. In particular for the yeast data, more complex models, when estimated, seem to be penalized by their larger number of parameters (see Figure 5 (b) for an illustration).

A. Synthetic data sets

We first assess our method performance on synthetic data for which the classes are known. Modelling gene expression data sets is an ongoing effort by many researchers and there is no well-established model to represent gene expression data yet. The simulation method we use is based on a proposal by [26]. It aims at simulating cyclic data, *ie.* cyclic behavior of genes over different time points. We create five data sets following the same model. Each set is made of 1536 genes for which we simulate 20 experiments. These genes come from 6 classes equal in size (256 genes per class) corresponding to different behavior over the time course. Let x_{ij} be the simulated expression level of gene i under experiment j in the data set. We first consider the following periodic behaviors (before adding noise). When the gene class is $z_i = c_k$ with $k = \{1, \ldots, 4\}$, we set

$$y_{ij} = sin(2\pi j/10 - \pi k/4) \quad \text{for } j = 1, \ldots 20.$$

When $k = 5$ and $k = 6$, we consider the linear behaviors $y_{ij} = j/20$ et $y_{ij} = -j/20$. Noise is then added,

$$x_{ij} = y_{ij} + \epsilon_{ij} \quad \text{for } i = 1, \ldots 1536 \text{ and } j = 1, \ldots, 20,$$

where the ϵ_{ij}'s are generated according to the normal distribution with mean 0 and standard deviation σ_{ij}. The σ_{ij}'s are drawn, randomly from standard deviations observed on the real data described by [13]. We further increase the noise by multiplying the ϵ_{ij}'s by 6 (the corresponding standard deviation is then $6 * \sigma_{ij}$). We refer to [17] and the web site `http://expression.washington.edu/publications/kayee/medvedovic_bioinf2003/` for a graphical illustration of such data (see also supplementary material).

As regards network data, we are not aware of any well established simulation methods. For a simple illustration, we consider the genes as the nodes of a 48×32 regular grid with neighbourhoods made of the 8 nearest neighbours. The 6 classes are then chosen as shown in Figure 2 (left-hand image) where each color is associated to a class. Although such a network has no biological interpretation, the classification quality is easy to assess by non expert users and it provides a clear visual illustration of the gain in taking into account network relationships. We compare the standard EM algorithm, which assumes genes independence and the EM-like procedures we propose. BIC is computed in both cases for $K = 3$ to $K = 9$. Typical curves are shown in Figure 2. EM-like procedures show higher BIC values than standard EM. The criterion selects the right number of classes except for data sets 1 and 4 for which 7 classes are preferred. However, this is consistent with the obtained classifications shown in Figure 3. For standard EM, 2 bands are wrongly merged except for data set 2. For the *simulated field* algorithm, the bands are correctly recovered except for sets 1 and 4 . In these latter cases, the 7-group classifications are visually better for data sets 1 and 4 (bottom row of Figure 3) as suggested by BIC values. The interpretation is that in these very noisy cases it may be worth considering an extra class with no specific meaning but that gathers outliers or too ambiguous measures. *Simulated field* and *mean field* algorithms perform similarly except for data set 5. In this case the *simulated field* algorithm selects 6 classes and gives a better classification. In the following developments, we will only refer to the *simulated field* algorithm.

Table I shows the global recognition rates (proportions of well-classified genes) obtained with the EM and *simulated field* algorithms for each data sets, while Table II shows the confusion matrix obtained for set 5. Rows correspond to the true classes while columns correspond to the obtained classes. The diagonal terms are the proportions of well classified genes in each class. The other terms are proportions of badly classified genes. All data sets show similar improvements when comparing EM to the *simulated field* algorithm.

On such synthetic data, the gain in taking into account network information or dependencies between genes through their labels appears clearly with improved recognition rates. BIC or its approximation in our Markov field setting, also appears to be satisfying criterion as regards the selection of the number of classes. It selects a number of classes which is consistent with the visual quality of the corresponding classification. These first conclusions will guide, in the next section, our analysis of the experiments on real data for which no ground-truth is available.

B. Saccharomyces cerevisiae (yeast) data

Although, our approach is valid for any organism provided individual data and network information is available, we focus

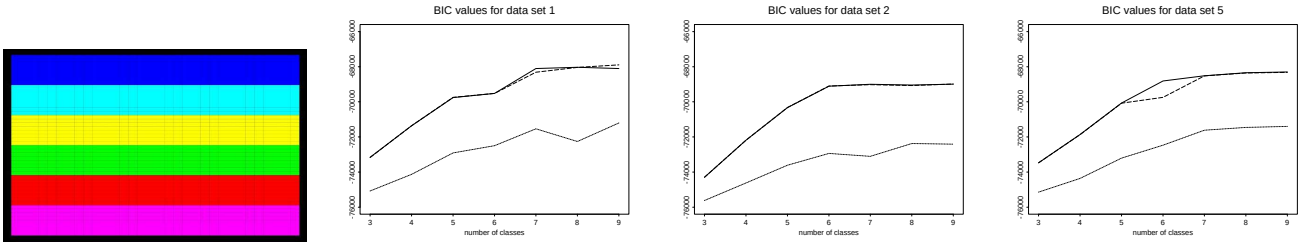


Fig. 2. Reference classification and BIC values for 3 data sets when K varies from 3 to 9. Solid line: Simulated field algorithm, Dotted line: EM algorithm, Dashed line: Mean field algorithm.

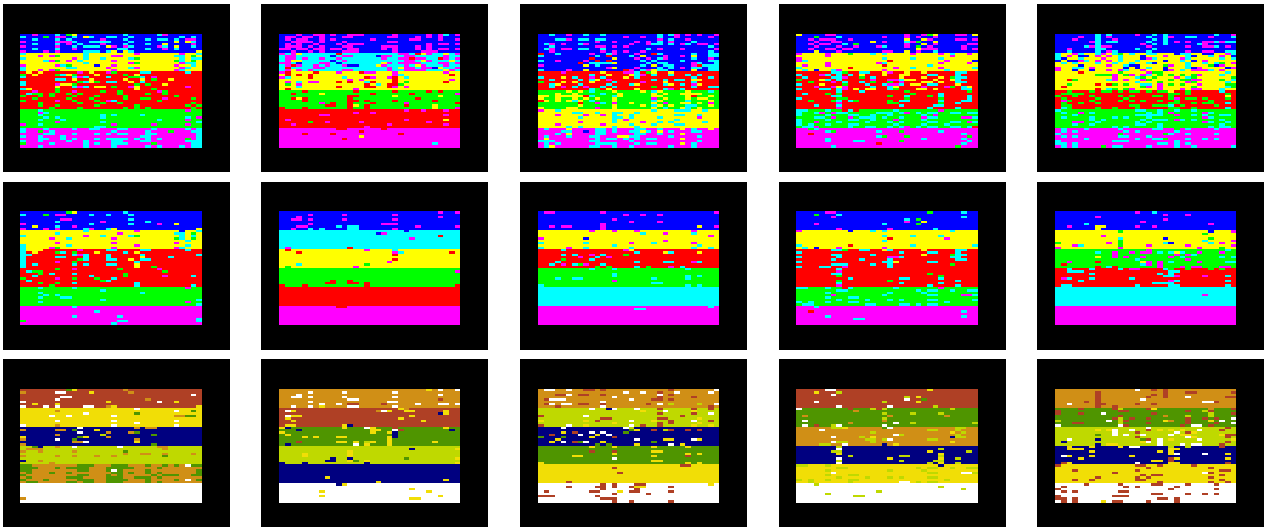


Fig. 3. Top and middle rows: 6 color classifications for 5 synthetic data sets using standard EM algorithm assuming independence (top row) and simulated field algorithm (middle row). Bottom row: 7 color classifications using the Simulated field algorithm. Note that the colors are arbitrarily assigned and may not match.

TABLE I
RECOGNITION RATES IN % FOR SYNTHETIC DATA ($K = 6$).

data sets	1	2	3	4	5
EM	64.8	79.2	63.3	68.1	64.3
Simulated field	77.5	95.8	93.6	78.6	91.3

TABLE II
CONFUSION MATRIX FOR FIGURE 3 MIDDLE RIGHT IMAGE.

global recognition rate= 91.3 %						
Class	1	2	3	4	5	6
1	94.1	1.2	0	0	0.8	3.9
2	1.2	89.1	3.1	0	2.0	4.7
3	0	1.2	80.9	1.6	5.9	10.5
4	0	0	1.6	84.0	12.1	2.3
5	0	0	0	0	99.6	0.4
6	0	0	0	0	0	100

on data related to *Saccharomyces cerevisiae* which is a widely studied organism with well established information and data on its mechanisms. The expression data we use are described by [8] and correspond to the developmental program of sporulation (gametogenesis in yeast). It consists of meiosis overlapped by spore formation. Sporulation can be characterized in terms of four distinct sets of genes which play different sequential roles according to their transcriptional activation during the process: early, middle, mid-late and late. The study proved this characterization to be suboptimal and a seven expression patterns description was preferred. Changes in the concentrations of the mRNA transcripts from each gene were measured at seven successive intervals after synchronisation; yeast cells were transferred to a nitrogen-limited medium that induces sporulation. The samples were taken at times (0h, 0.5h, 2h, 5h, 7h, 9h, 11.5h) based on the independently monitored expression pattern of known early, middle, mid-late, and late genes. Three additional points were measured when an essential transcription factor known to be activated at the end of the meotic prophase is missing; cells are then non-sporulating. The measures we use are related to these specific times. This leads to 10 dimensional profiles that should capture essential activity behavior of yeast genes during sporulation. As regards network data, we use the KEGG Reaction database as described in Section III. The resulting graph consists of 635 genes (amongst the 6118 ORFs expression measurements available, only 635 are present in the metabolic network). Since our aim is mainly to assess the benefit in adding network information, we then restrict to these 635 genes. The networks has 7111 edges and 92% of pairs are connected. Figure 4 gives a summary of characteristics of the network. In particular, the graph on the left reports that a significant number of edges are very connected : 245 nodes have more than 20 neighbours and 78 of them have 50 or more neighbours. The most connected nodes ($YER005W$) was found to have 145 interactors.

In this case, the appropriate number of classes is unknown. We compute BIC values for $K = 2$ to $K = 10$. The corresponding

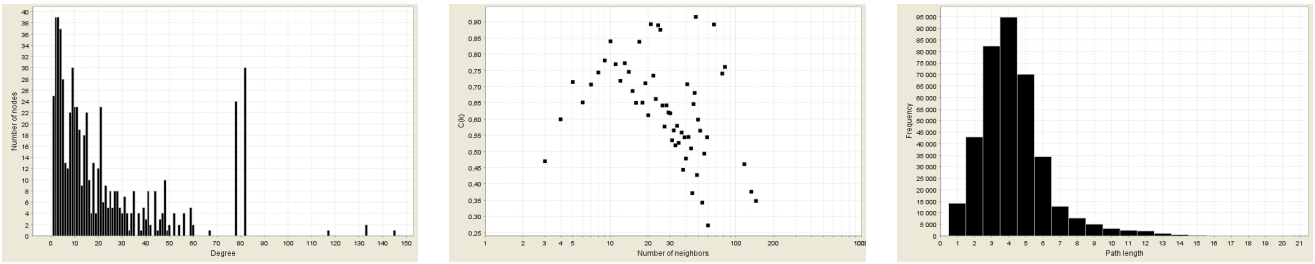


Fig. 4. Left: distribution of the node degrees of the network under study. Center: average clustering coefficients of nodes vs their number of neighbours. Right: distribution of the shortest path between nodes.

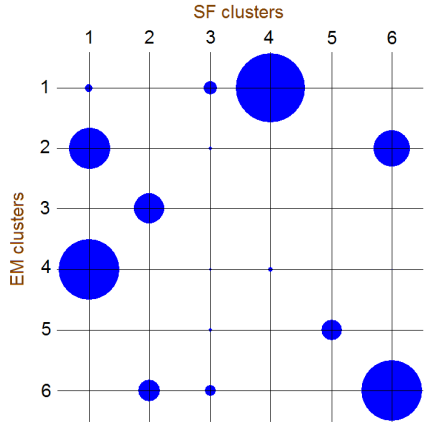


Fig. 6. Visualization of the confusion matrix of EM and SF classifications. Circle radius are proportional to cluster intersection cardinals. Note that SF and EM cluster labels have no special meaning and were arbitrarily and independently assigned as outputs of both clustering algorithm.

TABLE III
SOME CHARACTERISTICS OF SF CLUSTERS.

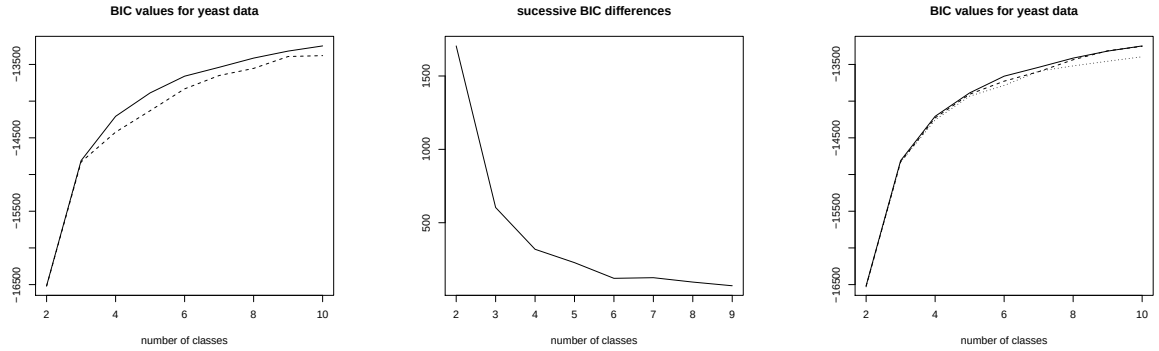
k	1	2	3	4	5	6
Nodes	180	86	52	123	34	160
Internal edges	537	195	75	254	8	903
% of connected pairs	14	66	17	14	1	16
Diameter	8	12	5	8	2	8

curve (Figure 5 (a)) does not show a clear maximum. We then consider as a reasonable choice of K, the value after which the difference in two successive BIC does not increase significantly anymore. This leads to selecting $K = 6$ as the number of classes (Figure 5 (b)). We then compare into more details classifications obtained with standard EM and with the *simulated field* algorithm for parameter $K = 6$. A visualization of the confusion matrix between SF and EM classifications is provided on Figure 6. While some clusters remain very similar (even identical in the cas of clusters SF and EM 5), it is sometimes difficult to identify correspondance; EM cluster 2 splits into SF clusters 1 and 6 and SF cluster 2 content comes from EM clusters 3 and 6. Moreover, characteristics for SF clusters are reported in Table III. All results are made available on the supplementary website.

To assess the quality of such classifications is not an easy task since there is no universal criteria to measure the relative performance of the algorithms. We therefore illustrate the gain in using our approach on the following specific features chosen for their relevance with regards to the data under consideration. Note that presenting the resulting clustering as a whole is not possible due to the size of the graph. An appropriate visualization tool is missing to provide a global biologically meaningful idea of the clusters. However, the clusters are available in separate files on our website.

Ideally, we would like to check whether our approach results in clusters better related to real biological networks. Since this experiment is based on a graph that accounts for dependencies that are expected to be strongly related to pathway information, we assess the quality and relevance of the various clusters by comparing them to groups of genes in the same metabolic pathway or in related pathways. [14] propose a method to detect significant pathways associated to the [8] expression data set we are using. They describe three scoring functions to characterize pathways at the transcriptional level based on gene expression, coregulation and cascade effect. Their pathway scores show relevance towards the biological background. This work provides an interesting tool to evaluate the performance of gene expression clustering techniques. High *Activity Scores* are awarded to pathways that exhibit many genes expressed above a given threshold or under another threshold in the case of repression effects. *Coregulation Scores* are higher for pathways in which genes show greater similarity in their expression patterns. *Cascade Scores* account for genes that do not show huge deviation from the reference time point and for the structure and ordering of the reactions in the pathway. In particular, they are useful to find out in which pathway a reaction chain is active or shut down for the particular experiment under study. For example, Transcription/Translation pathways are given a high *Activity Score* in results of [14]. This is well captured by our *simulated field* algorithm which gathers 16 out of the 28 genes involved in Transcription mechanisms in cluster 6. In comparison, standard EM succeeds in gathering 11 of these genes at best. When considering two clusters, these numbers raise respectively to 24 genes for our approach against 19 for the independent gene case (see Figure 7). Note that due to the restriction of our data set, we have no gene corresponding to B12 in our data although it corresponds to some yeast gene. Similar results hold for Translation involved genes.

We can also refer to the Vitamins metabolism that is given a high *Cascade Score* by [14]. The *simulated field* algorithm gathers 26 genes in the same cluster while EM recovers 19 at best. If two clusters are merged, these numbers respectively raise to 44 and 35 genes out of the 70 involved in the Vitamins Metabolism. Another pathway that is reported to be related to sporulation is the Oxydative Phosphorilation pathway that has a high *Coregulation Score* in [14]. Our method finds 24 genes in cluster 6 while EM groups at most 16 out of the 52 genes involved. The detected genes are up-regulated at the second time point and



(a) (b) (c)

Fig. 5. BIC values for yeast data when K varies from 2 to 10. **(a)** comparing Simulated field and EM algorithms. Solid line: Simulated field algorithm, Dashed line: EM algorithm; **(b)** Differences in two successive BIC for the Simulated field algorithm; **(c)** comparing various Markov models for the Simulated field algorithm. Solid line: $B = b \times I_K$ model, Dashed line: diagonal B model, Dotted line: full B model.

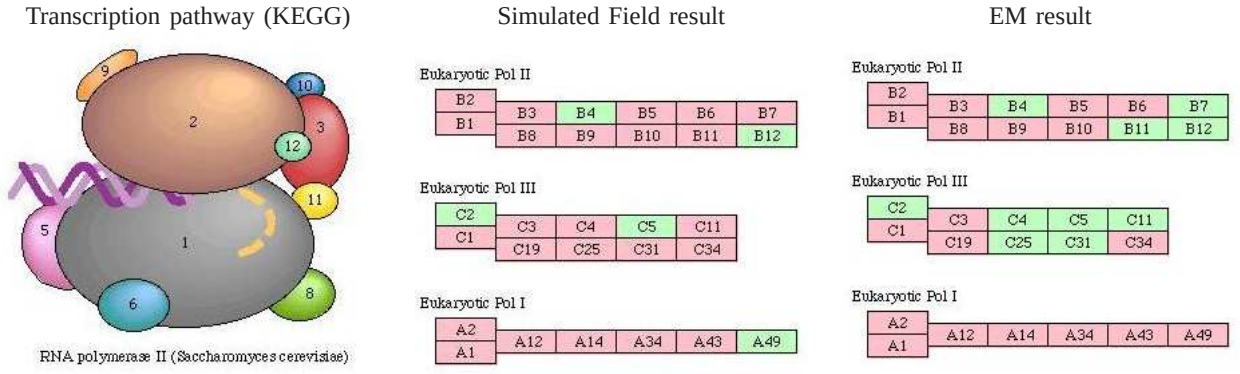


Fig. 7. RNA polymerase Transcription pathway as taken from KEGG. The middle and right columns show the results obtained by merging two clusters, respectively using the *simulated field* algorithm (middle) and the EM algorithm (right). Pink colored proteins correspond to genes that are included in the two merged clusters while green ones correspond either to genes that do not belong to the clusters or to genes that were not included in the analysis. The *simulated field* algorithm (missing proteins: B4, B12, C2, C5, A49) outperforms EM (missing proteins: B4, B7, B11, B12, C2, C4, C5, C11, C25, C31) in grouping together this class of genes.

are specific to ATP synthesis. The analysis shows that cluster 6 is related to Energy metabolism (*eg.* Oxydative Phosphorilation) as well as metabolisms that deal with Transcription (*eg.* RNA polymerase), Translation (*eg.* Aminoacyl-tRNA synthetase) and Vitamins. Other pathways can be more fully recovered using our approach and the additional graph information. As an illustration, for the glycolysis pathway, 24 genes belong to the same *simulated field* cluster while EM groups 19 out of the 44 in our data set. These results show the inclination of our method to a better sensitivity to group genes involved in meaningful identified metabolism subunits than standard EM. Figure 8 shows genes assigned to *simulated field* cluster 2 that are involved in Glycolysis. This cluster is mainly related to Carbohydrate metabolism (see Table IV).

Our method has the ability to group genes with a coordinated activity during glycolysis despite some expression dissimilarities. This is the case for $YLR153C$ ($EC6.2.1.1$) and $YPL061W$ ($EC1.2.1.3$) which have a slowly increasing expression while genes in the main way converting glucose 6-phosphate into pyruvate (or conversely) are immediately over-expressed. As a matter of fact the two former genes are not assigned to the same cluster as the others when using standard EM. The glycolysis example suggests that, as expected, our method outperforms traditional clustering methods in grouping functionally related genes into clusters even if their expression pattern is not a sufficient clue.

To further assess the gain in using network information, we also consider an ontological analysis approach to help with the biological interpretation of the results. We used the 1935 GO terms available -out of them is a subset of 1016 terms involved in *biological process*)- at the time of the study, from the Gene Ontology (`http://www.geneontology.org/`) database. The full list and additional information is made available on our website (`http://mistis.inrialpes.fr/people/forbes/transparentia/supplementary.html`). Two series of statistical tests are driven. The null hypothesis always being that a GO term is not over/under-represented and the alternative being that a GO term is over- (first series) or under-represented (second series). P-Values are computed with False Discovery Rate (expected proportion of erroneous rejections among all rejections) corrections, which addresses the multiple testing issue. Moreover, arbitrary dependencies between groups are taken into account (see [3]). The power of this correction is lower than in [2] (null hypothesis are more easily accepted even if wrong) that is valid under positive regression dependencies. It is not sure wheither the GO terms hierarchy fulfills this statement. Eventually, we will miss some over-represented GO categories but we can be confident about identified ones. This is the price to pay to account for multiple testing under arbitrary

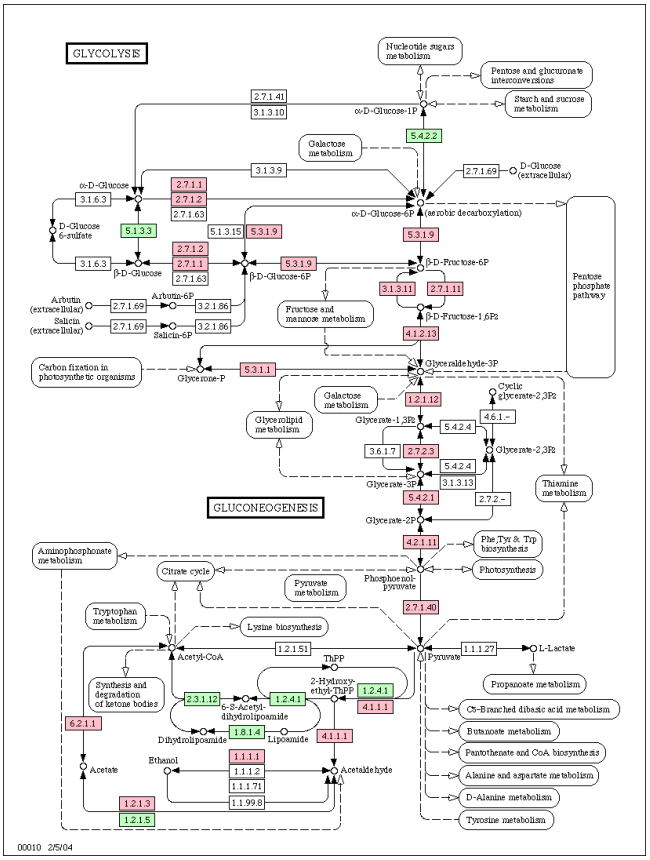


Fig. 8. Glycolysis pathway: colored EC numbers are in our data set. Pink ones belong to the same *simulated field* cluster while green ones (numbers 5.4.2.2, 5.1.3.3, 2.3.1.12, 1.2.4.1, 1.8.1.4, 1.2.1.5) do not.

dependencies. The analysis is summarized in Table IV which shows respective P-values for over-represented GO terms in the clusters found by the *simulated field* and standard EM algorithms. Note that since clustering results for *simulated field* and standard EM algorithms differ, the cluster numbering corresponds to the *simulated field* algorithm. The last column of the table shows the best corresponding P-Values computed among all the EM clusters. Although this does favor EM, the results show that the *simulated field* approach still performs better. Under-represented GO-terms are not listed for sake of brevity and because most of over-represented GO terms in one cluster show under-representation in the other clusters and this with significant P-values.

The ontological analysis is consistent with the previous observations on pathways. In addition, it suggests that our method tends to produce clusterings with more specificity than traditional EM in the sense that GO terms that are significantly over-expressed in one cluster are significantly under-expressed in the other clusters. This is usually true but to a much lesser extent for clusters found by EM. The *simulated field* algorithm provides highly specific clusters. To exhibit such a property we considered GO terms with a reasonnable number of components (ten or so). GO terms gathering are not specific to a distinctive class of genes. They do not give a satisfactory evidence that our method can distinguish between genes specific to some processes. GO terms with a too low number of genes cannot lead to a real validation of the method neither.

The genes classified under GO term $GO : 0008652$: amino-

TABLE IV

ONTOLOGICAL ANALYSIS: OVER-REPRESENTED GO CATEGORIES RELATED TO THE DIFFERENT CLUSTERS AND CORRESPONDING P-VALUES. SIMULATED FIELD (RESP. STANDARD EM) P-VALUES ARE COMPUTED FOR SIMULATED FIELD (RESP. STANDARD EM) CLUSTERS.

Simul. field Cluster	GO terms	Simul. field P-values	Standard EM P-values
1	GO:0008652: amino-acid biosynth.	**1.1E-2**	0.193
2	GO:0006006: glucose metabolism	**1.2E-7**	8.7E-7
	GO:0006090: pyruvate metabolism	5.9E-5	**8.7E-7**
	GO:0006144: purine base metabolism	**2.2E-2**	0.259
	GO:0015980 : energy dev. by oxid...	**1.8E-2**	3.3E-2
3	GO:0006259: DNA metabolism	**4.1E-2**	1
	GO:0006261: DNA-dep. DNA replic.	**4.1E-2**	0.193
	GO:0006271: DNA strand elong.	**4.1E-2**	0.208
4	no significant GO term	N.A.	N.A.
5	GO:0030437: sporulation	**4E-2**	0.26
6	GO:0006360: transcr. from RNA pol.	**1.6E-2**	2.5E-2
	GO:0006164: purine nucleo biosynt.	**2.0E-2**	6.1E-2

acid biosynthesis (see Table IV) are a first example. The P-Value (1.1%) shows that the *simulated field* cluster (1) is highly specific towards this function whereas the corresponding standard EM cluster isn't (P-value equal to 0.193). *Simulated field* cluster 5 is an even more relevant example. It contains most of the sporulation specific genes ($GO : 0030437$) listed in [8] (available on the paper website or with our data in supplementary material). The test conclusion is that this cluster is specific towards the invoked function with a P-Value of 4%. For comparison, the best results among standard EM clusters is 0.26 which does not lead to the conclusion that this term is over-represented. Note that this is somewhat surprising since these genes are apparently not linked by any of the association types provided in the STRING database (`http://string.embl.de`). We looked for links related to databases, co-expression, physical location on the chromosome, fusion, experiments, co-occurrence in different genomes. But only a text-mining link was detected, certainly due to the fact that many of the genes are referenced in the [8] paper. According to this paper, 32 among 34 genes in cluster 5 take a significant part in the temporal program of yeast sporulation. This cluster does not include however the class of *metabolic* genes (quickly induced) that are mainly recovered in another much bigger cluster. A possible interpretation is that these latter genes have a quite different regulator.

V. DISCUSSION AND CONCLUSION

Our aim was to show that Hidden Markov models could be introduced to incorporate various types of information about biological objects (*eg.* genes) and in particular to account for interactions between these objects (through biological networks for instance). We focused on the task of classifying genes from their expression profiles and from metabolic pathways data as an illustration. The introduction of Markov models in this context is new. They provide parametric models where the parameters have a natural interpretation. Some of them (the α_k's) can be related to class proportions while others (matrix $\mathbb{B}$) to pair-wise interactions (see Section II-A). In our method, parameters are estimated but tuning is also possible, for instance, to incorporate *a priori* knowledge regarding class proportions or strength of interactions to put more weight on network data. Other clustering methods are much less readable in that sense.

Preliminary results are promising. Experiments on simulated data show that our approach can improve significantly classifica-

tion rates. They also suggest that criteria based on BIC could be used to guide the choice of the number of classes. Additional experiments on real data (yeast) point out further interesting features of our approach. The *simulated field* algorithm leads to biologically more plausible and more fully identified clusters. When compared to clustering methods based on gene expression only (*eg.* EM clustering), it has the advantage to produce clusters associated to pathways with possible coordinated change in gene expression. When compared with methods incorporating network data, it has the advantage to consist in a statistically well founded approach which does not require to choose a distance or a kernel function and allows further statistical analysis regarding additional issues such as model selection. It is also part of the *soft* clustering methods that provide membership probabilities instead of *hard* (usually more biased) classifications.

Future work would be to investigate this general methodology in other contexts, with applications in proteomics, using genes or proteins as central concepts through a variety of information sources such as sequences, structures, expression patterns, position in networks, *etc*. Additional experiments could also be useful to challenge our model on incomplete interaction data. The difficult passage here is to determine the full pair interaction network. In the synthetic data case, the 8 nearest neighbourhood setting is chosen because it is widely used in image analysis. But it doesn't claim to be the network that accounts at best for interactions in an image. In the real biological dataset, we are certainly faced to a graph only summarizing a fraction of the real hypothetical gene interaction network of *Saccharomyces cerevisæ*. Biological interaction networks are known to be incomplete and the reliability of the interactions vary a lot. Hence it makes it difficult to assess the proportion of information initially present in the network. A possibility would be to take into account a wider range of pair dependencies and to assign weights according to prior knowledge on the reliability of the interaction. Our model was designed to deal with such a refinement. This kind of information is not available yet on the metabolic network we considered. Alternately, it might be worth building an interaction network using a database like STRING (`http://string.embl.de/`) which offers confidence level for gene interactions from various sources (physical chromosome neighbourhood, two-hybrid, literature link,...). Before that, more specific analysis would be useful as regards the generalization to missing data that often occur in biological studies. Our mean field-like framework allows such a generalization. Also, in a variety of applications, overlapping clustering, wherein some items are allowed to be members of two or more discovered clusters, is more appropriate. Methods have been proposed that would be worth investigating in the context of genetic data analysis.

ACKNOWLEDGMENT

The authors would like to thank Frédéric Boyer and Juliette Blanchet for their help with the data and the experiments. We are also grateful to Alain Viari and Éric Coissac for fruitful discussions. Lastly we would like to thank anonymous reviewers for their detailled comments and suggestions on our work.

REFERENCES

[1] J. D. Banfield and A. E. Raftery, "Model-based Gaussian and non Gaussian Clustering", *Biometrics*, vol. 49, no. 3, pp. 803–821, Sept. 1993.

[2] Y. Benjamini and Y. Hochberg, "Controlling the false dicovery rate - a practical and powerful approach to multiple testing", *Journal of the Royal Statistical Society B*, vol. 57, no.1, pp. 289–300, Feb. 1995.

[3] Y. Benjamini and D. Yekutieli, " The control of the false discovery rate in multiple testing under dependency," *Annals of Statistics*, vol. 29, no. 4, pp. 1165–1188, Aug. 2001.

[4] C. Bouveyron, S. Girard and C. Schmid, "Class specific subspace discriminant analysis for high dimensional data," In *Lect. Notes Comp. Sci., Springer*, no. 3940, pp139–150, 2006.

[5] M.P.S. Brown, W.N. Grundy, D. Lin, N. Cristianini, C.W. Sugnet, T.S. Furey, M. Ares Jr. and D. Haussler, "Knowledge-based analysis of microarray gene expression data by using support vector machines," *Proc. Nat. Acad. Sci.*, vol 97, no.1, pp.262–267, Jan 2000.

[6] G. Celeux and G. Govaert, "Gaussian Parsimonious clustering models", *J. of Pat. Rec. Soc.*, 28, pp. 781–793, 1995.

[7] G. Celeux, F. Forbes and N. Peyrard, "EM procedures using mean-field like approximations for Markov-model based image segmentatio," *Pat. rec.*, vol. 36, no. 1, pp. 131–144, Jan 2003.

[8] S. Chu, J.L. DeRisi, M.B. Eisen, J. Mulholland, D. Botstein, P.O. Brown and I. Herskowitz, " The transcriptional program of sporulation in budding yeast, " *Science*, vol. 282, pp. 699–705, Oct 1998.

[9] M.B. Eisen, P.T. Spellman, P.O. Brown and D. Botstein, "Cluster analysis and display of genome-wide expression patterns," *Proc. Nat. Acad. Sci.*, vol. 95, pp.14863–14868, Dec 1998.

[10] F. Forbes and N. Peyrard, "Hidden Markov random field model selection criteria based on mean field-like approximations," *IEEE Trans. PAMI*, vol. 25, no. 9, pp. 1089–1101, Sep 2003.

[11] D. Hanisch, A. Zien, R. Zimmer and T. Lengauer, "Co-clustering of biological networks and gene expression," *Bioinformatics*, vol. 18 no. Suppl.1, pp. S145–S154, Jul. 2002.

[12] A.J. Hartemink, D.K. Gifford, T.S. Jaakkola and R.A. Young, "Using graphical models and genomic espression data to statistically validate models of genetic regulatory networks," *Proc. Pacific Symp. Biocomputing 7*, pp. 422–433, Jan. 2002.

[13] T.R. Hugues, M.J. Marton, A.R. Jones, C.J. Roberts, R Stoughton, C.D. Armour, H.A. Bennett, E. Coffey, H. Dai, Y.D. He, M.J. Kidd, A.M. King, M.R. Meyer, D. Slade, P.Y. Lum, S.B. Stepaniants, D.D. Shoemaker, D. Gachotte, K. Chakraburtty, J. Simon, M. Bard and S.H. Friend, "Functional discovery via a compendium of expression profiles," *Cell*, vol. 102, pp. 109–126, Jul. 2000.

[14] M.P. Kurhekar, S. Adak, S. Jhunjhunwala and K. Raghupathy, "Genome-wide pathway analysis and visualization using gene expression data," *Proc. Pacific Symp. Biocomputing 7*, pp. 462–473, Jan 2002.

[15] G. Lanckriet, T. De Bie, N. Christianini, M. I. Jordan and W. Noble, "A statistical framework for genomic data fusion," *Biofinformatics*, vol. 20, no. 16,pp. 2626–2635, Nov. 2004.

[16] E.M. Marcotte, M. Pellegrini, M.J. Thompson, T.O. Yeates and D. Eisemberg, "A combined algorithm for genome-wide prediction of protein function," *Nature*, vol. 402, pp. 83–86, Nov. 1999.

[17] M. Medvedovic, K.Y. Yeung and R.E. Bumgarner, "Bayesian mixture model based clustering of replicated microarray data," *Bioinformatics*, vol. 20, no. 8, pp. 763–774, Apr. 2004.

[18] P. Pavlidis, J. Weston, J. Cai and W. N. Grundy, "Gene functional classification from heterogeneous data," *Proc. Fifth Annual Int. Conf. Comp. Biol.*,, pp. 249–255, Apr. 2001.

[19] E. Segal, H. Wang and D. Koller, "Discovering molecular pathways from protein interaction and gene expression data," *Bioinformatics*, vol. 19, no. Suppl 1, pp. i264–i272, Jul. 2003.

[20] G. Schwarz, "Estimating the dimension of a model," *Ann. Stat.*, vol. 6, no.2, pp. 131–134, Apr. 1978.

[21] P. Tamayo, D. Slonim, J. Mesirov, Q. Zhu, S. Kitareewan, E. Dmitrovsky, E.S. Lander, and T.R. Golub, "Interpreting patterns of gene expression with self-organizing maps : methods and application to hematopoietic differentiation," *Proc. Nat. Acad. Sci.*, vol. 96, no. 6, pp. 2907–2912, Mar. 1999.

[22] A. Tanay, R. Sharan, M. Kupiec and R. Shamir, "Revealing modularity and organization in the yeast molecular network by integrated analysis of highly heterogeneous genomewide data," *Proc. Nat. Acad. Sci.*, vol. 101, no. 9, pp. 2981–2986, Mar. 2004.

[23] S. Tavazoie, J. D. Hughes, M.J. Campbell, R.J. Cho and G.M. Church, "Systematic determination of genetic network architecture," *Nature Genetics*, vol. 22, no.3, pp. 281–285., Jul. 1999.

[24] J.-P. Vert and M. Kanehisa, "Graph-driven features extraction from microarray data using diffusion kernels and kernal CCA," *Adv. Neural Inf. Proc. Sys. 15*, pp. 1425–1432, 2003.

[25] Y. Yamanishi, J.-P. Vert, A. Nakaya and M. Kanehisa, "Extraction of correlated gene clusters from multiple genomic data by generalized kernel canonical correlation analysis," *Bioinformatics*, vol. 19, no. Suppl 1, pp. i323–i330, Jul. 2003.

[26] K.Y. Yeung, C. Fraley, A. Murua, A. Raftery and L. Ruzzo, "Model-based clustering and data transformations for gene expression data," *Bioinformatics*, vol. 17, no. 10, pp. 977–987, Oct. 2001.

[27] G. Celeux and J. Diebolt, "The SEM algorithm: a probabilistic teacher algorithm derived from the EM algorithm for the mixture problem," *Computational Statistics Quaterly*, vol. 2, pp. 73–82, 1985.

Matthieu Vignes Matthieu Vignes graduated from University Claude Bernard, Lyon, France in 2002. He is now finishing a PhD in Statistics at University Joseph Fourier, Grenoble as part of team Mistis (INRIA Rhône-Alpes). He was recently appointed by BioSS (Biomathematics & Statistics Scotland). His research interests include Graphical models, Missing data, Overlapping clustering and Statistical applications to system biology and compositional analysis of crop plants.

Florence Forbes Florence Forbes received the PhD degree in applied probability in 1996, from University Joseph Fourier, Grenoble, France. She is a research scientist at the Institut National de Recherche en Informatique et Automatique (INRIA). She joined IS2 team at INRIA Rhone-Alpes in 1998 and is now leading research team Mistis. Her research activities include Bayesian image analysis, Mixture models, Markov processes, Markov random fields, and hidden structure models.

C.2 Analyse combinée de données d'expression avec des valeurs manquantes et d'un réseau d'interactions : une approche markovienne

Combined expression data with missing values and gene interaction network analysis: a Markovian integrated approach

Juliette Blanchet
INRIA Rhône-Alpes
655, av. de l'Europe
38330 Saint Ismier Cedex, France
Email: juliette.blanchet@inrialpes.fr

Matthieu Vignes
BioSS at Scottish Crop Research Institute
Invergowrie, Dundee
DD2 5DA, Scotland, UK
Email: matthieu@bioss.ac.uk

***Abstract*—DNA microarray technologies provide means for monitoring in the order of tens of thousands of gene expression levels quantitatively and simultaneously. However data generated in these experiments can be noisy and have missing values. When it is not ignored, the last issue has been solved by imputing the expression matrix in order to keep going with traditional analysis method. Although it was a first useful step, it is not recommended to use value imputation to deal with missing data. Moreover, appropriate tools are needed to cope with noisy background in expression levels and to take into account a dependency structure among genes under study. Alternative approaches have been proposed but to our knowledge none of them has the ability to fulfil all these features. We therefore propose a clustering algorithm that explicitely accounts for dependencies within a biological network and for missing value mechanism to analyze microarray data. In experiments on synthetic and real biological data, our method demonstrates enhanced results over existing approaches.**

I. Introduction

A major challenge addressed to biostatistics is to reveal complex interactions between components of living systems. A common and efficient way of representing these interactions is to build a network responsible for the observed biological complexity. As the amount of high-throughout data continuously increases, there is a need for methodologies to extract information contained in it.

In this framework, organizing produced data into meaningful structures is highly desirable as a first step in unsupervised exploration of the large number of genes involved in complex biological systems. As an example, many clustering algorithms have been proposed over the last decade to decipher the message contained in microarray data (the interested reader can see a comprehensive overview of gene expression data clustering methods in [1]).

Ideally, microarray data would reflect the activity level of all messengers that control the build-up of proteins in the cell. The complexity of microarray experiments induce quite noisy data. The use of statistical models can overcome the absence of uncertainty for the estimated expression level of each gene. [2] proposed a Gaussian mixture model to deal with microarray analysis. However a key assumption of mixture models is that they consider gene measurements to be independent. Hence we proposed in a previous work ([3]) an extension of this approach to account for individual features (*e.g.* microarray data) and dependencies between genes in a united fashion thanks to Hidden Markov Random Field. As a probabilistic model, it leads to possible statistical criteria to select the number of clusters, an often difficult issue in clustering processes.

All the works mentioned above use a full matrix of expression data as an input. An unfortunate feature of microarray experiments is that they often produce multiple missing values. Most of the time, these missing entries appear because of various experimental issues: dust or scratches on the slide, corrupted images, difficulties in measuring fluorescence intensity, systematic error of the robot that drops the probes, problem with accurate gene spotting on the array,... As a consequence, certain items of the expression matrix are missing.

Here, missing data refer to values that are missing due to bad measurement or corrupted observations but that should be otherwise available in ideal experimental conditions. We distinguish such missing values from the usual clustering paradigm often referred to as a missing data problem. The goal is to label each individual or site in agreement with the data. In this case the labels (*e.g.* the biological processes bringing out the observations) are unknown and are intrinsically missing, hence referred to as missing data too.

A prevalent practice is to remove genes and/or arrays from the analysis to end up with a matrix without missing value. However it can provoke the loss of important informations. Up to 90% of genes (rows) or experimental conditions (columns) can be affected ([4]). The mechanism is named data deletion. It can also conceal interactions by deleting nodes in the considered interacting network. Another widespread practice to circumvent these absences is to replace them by zeros or by unconditional means. Such a naive filling-in strategy is a particulary case of single imputation. It is known to cause spurious estimations of summary statistics. As a consequence, clustering results can be misleading ([5]).

Several more sophisticated methods have been proposed

since the pioneering work of [6]. Most of them propose single imputation methods to transform the data matrix into a full matrix as needed by subsequent classical statistical analysis (see [4], [6]–[8],...). More recently, promising approaches make use of multiple imputation ([9]) or iterative alternate blended clustering and missing values estimation ([1]). Although System Biology nowadays advocates the consitent integration of available data to reveal the mechanisms of living organisms, none of these approaches take into account relationships, hence dependencies, between genes like protein-protein interactions or metabolomic networks.

We propose to tackle these issues in a unique statistical framework. We take advantage of many features of the probabilistic aspect of the model. In a previous work ([3]), we mentioned the ability of a straightforward extension of the model to deal with missing values. It is now implemented and we prove it to be successful at dealing with different absence patterns either on simulated or real biological data sets. We emphasize that our model can be useful in a great range of applications for clustering entities of interest (such as genes, proteins, metabolites in post-genomics studies). It requires individual possibly incomplete measurements taken on these entities related by a relevant interaction network. Hence our method is neither organism- nor data-specific. We can imagine as well the interest of the method presented here in a wide variety of fields where missing data is a common feature: social sciences, computer vision, remote sensing, speach recognition and of course biological systems.

We organized the present work as follows: the model will be presented in Section II with the EM estimation procedure and classification framework. We already mentioned the advantage of deriving a statistical criteria from the computed likelihood of the data. Furthemore, it provides *a posteriori* probabilities of entities (*e.g.* genes) to be allocated to classes given the data. This can be seen as a confidence measure of assignment. Another flexibility of our approach is that features on individuals and various pairwise relationships can easily be incorporated in a unique statistical model. Last but not least, the framework used is appropriate to deal with missing data and the algorithm is a straightforward extension. We specify the data we use in the present work and show the kind of results that can be obtained. Experiments on synthetic image-like data are reported in Section III while results on real yeast cell-cycle data combined with a network of interacting proteins are presented in Section IV. For sake of brievety, we chose some highlights of our results to reveal the performance of our method. We strongly encourage interested readers to refer to our supplementary material website at http://mistis.inrialpes.fr/people/blanchet/supbibe.html for the latest version of the software, enlarged and colour figures, full results and analysis of experiments on yeast as text files, *etc*.

II. Clustering with incomplete data

In this work, we assume that values are Missing At Random (MAR case [5]). It means that missingness can only be attributed to observed data not to missing data (*e.g.* the fact that a data is missing cannot depend on its label). If the absence does not depend on processes giving birth to the data at all, missing values are Missing Completely At Random (MCAR case).

MAR hypothesis allows a separate estimation of the missingness process on one hand and of the data distribution on the other hand. In fact, the likelihood can be factorized in two terms. The first one does not depend on the process generating the data and the second one does not depend on the missingness process [5]. In real applications, the MAR assumption might not be true as regards the phenomenon generating missing values. Just think of censorship issues due to machine limits of detection or data missing in polls because surveyees choose not to answer to peculiar items. Data are then said to be Not Missing At Random (NMAR case). Methods based on MAR assumption can however produce satisfactory results if observed values contain enough information to predict missing values ([5] and references therein).

We present in this section a general algorithm for clustering incomplete dependent data. Dependencies between *sites* (pixels in section III, genes in section IV) are defined through a network (or a neighbourhood structure). Let us denote $\mathcal{S}$ the set of N sites and $\mathbf{x} = \{x_i \in \mathbb{R}^D\}$ the $N \times D$ matrix of observations. Some entries of this matrix can be misssing. For each $i \in \mathcal{S}$, we will write $o_i \subset [\![1, D]\!]$ the index set corresponding to the observed values x_{id} and m_i the complementary set of missing values ($o_i \cup m_i = [\![1, D]\!]$). We shall denote $x_i^{o_i} = \{x_{id}, d \in o_i\}$, $x_i^{m_i} = \{x_{id}, d \in m_i\}$, $\mathbf{x}^o = \{x_i^{o_i}, i \in \mathcal{S}\}$ and $\mathbf{x}^m = \{x_i^{m_i}, i \in \mathcal{S}\}$. We address the issue of clustering *i.e.* distinguishing meaningful groups in a dataset. In other words, each site $i \in \mathcal{S}$ is assigned one of the K labels $z_i \in [\![1, K]\!]$. Dependencies between neighbouring observations are modeled by further assuming that the joint distribution of $\mathbf{Z} = \{Z_i, i \in \mathcal{S}\}$ is a discrete Markov Random Field (MRF):

$$P_G(\mathbf{z}) = W^{-1} \exp(-\textstyle\sum_{c \in C} V_c(\mathbf{z}_c))$$

where C denotes the set of cliques, V_c is a potential function for a specific clique c and W is a normalizing constant. We eventually assume that data are independent conditionnally on classes, this is to say that $P(\mathbf{x}|\mathbf{z}) = \prod_{i \in \mathcal{S}} P(x_i|z_i)$. Under this assumptions, $(\mathbf{X}, \mathbf{Z})$ is said to be an *Hidden Markov Random Field*. Note that the *a posteriori* probability distribution $P_G(\mathbf{Z}|\mathbf{x})$ of classes $\mathbf{Z}$ conditionnaly on the observations $\mathbf{X} = \mathbf{x}$ is also Markovian. Let us then denote θ_k the parameter of $P(x_i|Z_i = k)$ (for example mean vector and covariance matrix for a Gaussian distribution) and ϕ the parameters of P_G. Parameters $\theta_1, \cdots, \theta_K$ and ϕ of the model have to be estimated.

The Expectation-Maximisation (EM) algorithm [10] is a commonly used algorithm for parameters estimation in problems with missing data. Here, value absence refer to observations that are missing and to the labels. The algorithm aims at maximizing the log-likelihood for the observed variables $\mathbf{x}^o$ of the model. At each iteration, it maximizes the expectation of the complete log-likelihood, for the observed $\mathbf{x}^o$ and missing $\mathbf{x}^m$, $\mathbf{z}$ data given the observed data and a current estimate of

the model parameters. More specifically, with the full set of parameters for the model $\Psi = (\theta_1, \cdots, \theta_K, \phi)$, at iteration (q), a current estimate $\Psi^{(q-1)}$ is available and the algorithm maximises the function Q defined as:

$$Q(\Psi|\Psi^{(q-1)}) \equiv \mathbb{E}[\log P(\mathbf{x}^o, \mathbf{X}^m, \mathbf{Z}|\Psi)|\mathbf{x}^o, \Psi^{(q-1)}]$$

When data are independent, $P(\mathbf{x})$ is an independent mixture model and EM is tractable [5]. When the model is an hidden Markov model, due to the dependence structure, there are two major difficulties in evaluating this expectation. Both the normalizing constant W and the conditional probability $P(\mathbf{z}|\mathbf{x})$ cannot be computed exactly; approximations are required to make the algorithm tractable. In this paper, we propose to use a mean field-like approximation of the Markovian *a posterori* distribution $P_G(\mathbf{z}|\mathbf{x})$ as proposed in [11] in the framework of complete data clustering. Informally, mean field-like approximations consist in neglecting fluctuations in the neighbourhood N_i of each location i by setting his neighbours $j \in N_i$ to fixed values $\tilde{z}_j$ (to means for example). The intractable Markovian distributions $P_G(\mathbf{z})$ and $P_G(\mathbf{z}|\mathbf{x})$ are then respectively approximated by $\prod_{i\in\mathcal{S}} P_G(z_i|\tilde{z}_{N_i})$ and $\prod_{i\in\mathcal{S}} P_G(z_i|\tilde{z}_{N_i}, x_i)$.

Using such approximations leads to algorithms which in their general form consist in repeating two steps. In what follows, we will focus on the *Simulated Field* approximation for it performance (see [11]). At iteration (q),

- *[NR]* For each site $i \in \mathcal{S}$, simulate from the observed data $x_i^{o_i}$ and some current parameter estimates $\Psi^{(q-1)}$ a configuration $\tilde{z}_i^{(q)}$ *i.e.* values for the Z_i's. Replace the Markov distribution $P_G(\mathbf{z})$ by the factorized distribution $\prod_{i\in\mathcal{S}} P_G(z_i|\tilde{z}_{N_i}^{(q)})$.
- *[EM]* Apply one step of the EM algorithm for this factorized model with starting values $\Psi^{(q-1)}$, to get updated estimates $\Psi^{(q)}$ of the parameters.

Note that such an algorithm allows to take into account the Markovian structure of the data while reducing to a factorized distribution on which EM is tractable. When observations are incomplete, the differences with the complete data case are in the E-step which becomes: $\forall i \in \mathcal{S}$, $\forall k \in [\![1, K]\!]$, compute

$$\tilde{t}_{ik}^{(q)} = \begin{cases} P_G(Z_i = k|x_i^{o_i}, \tilde{z}_{N_i}^{(q)}) & \text{if } |o_i| \neq 0 \\ P_G(Z_i = k|\tilde{z}_{N_i}^{(q)}) & \text{otherwise} \end{cases}$$

and in the updating of the parameters $(\theta_k)_{k\in[\![1,K]\!]}$ in the M-step: for all $k \in [\![1, K]\!]$,

$$\theta_k^{(q)} = \arg\max_{\theta_k} \sum_{i\in\mathcal{S}} \tilde{t}_{ik}^{(q)} \int P(x_i^{m_i}|x_i^{o_i}, \theta_k^{(q-1)}) \log P(x_i^{o_i}, x_i^{m_i}|\theta_k)\, dx_i^{m_i}.$$

The updating of the parameters ϕ of the Markovian distribution $P_G(\mathbf{z})$ remains unchanged and can easily be perfomed with a gradient descent:

$$\phi^{(q)} = \arg\max_{\phi} \sum_{i\in\mathcal{S}} \sum_k \tilde{t}_{ik}^{(q)} \log P_G(Z_i = k|\tilde{z}_{N_i}^{(q)}, \phi)$$

If in addition, we assume $P(.|\theta_k)$ to be a Gaussian distribution $\mathcal{N}(\mu_k, \Sigma_k)$, the updating of parameter $\theta_k = (\mu_k, \Sigma_k)$ in the M-step can be further specified (see [5]). Denote $\Sigma_k^{o_i o_i} = \{(\Sigma_k)_{st}, s \in o_i, t \in o_i\}$, $\Sigma_k^{o_i m_i} = \{(\Sigma_k)_{st}, s \in o_i, t \in m_i\} = (\Sigma_k^{m_i o_i})^T$ and $\Sigma_k^{m_i m_i} = \{(\Sigma_k)_{st}, s \in m_i, t \in m_i\}$. Then,

$$\begin{aligned} P(x_i^{o_i}|\theta_k) &= \mathcal{N}(x_i^{o_i}|\mu_k^{o_i}, \Sigma_k^{o_i o_i}) \\ P(x_i^{m_i}|x_i^{o_i}, \theta_k) &= \mathcal{N}(x_i^{m_i}|\eta_{ik}, \Gamma_{ik}) \end{aligned}$$

with,

$$\begin{aligned} \eta_{ik} &= \mu_k^{m_i} + \Sigma_k^{m_i o_i}(\Sigma_k^{o_i o_i})^{-1}(x_i^{o_i} - \mu_k^{o_i}) \\ \Gamma_{ik} &= \Sigma_k^{m_i m_i} - \Sigma_k^{m_i o_i}(\Sigma_k^{o_i o_i})^{-1}\Sigma_k^{o_i m_i} \end{aligned}$$

and, at iteration (q) the component $s \in [\![1, D]\!]$ of μ_k is updated as:

$$(\mu_k^s)^{(q)} = \frac{\sum_i \tilde{t}_{ik}^{(q)}(r_i^s x_i^s + (1 - r_i^s)\eta_{ik}^{s\,(q)})}{\sum_i \tilde{t}_{ik}^{(q)}} \quad (1)$$

with $r_i^s = 1$ if variable x_i^s is observed, 0 otherwise. Note that equation (1) consists in replacing the missing variable x_i^s by the mean $(\eta_{ik}^s)^{(q)}$ of the distribution $P(X_i^{m_i}|x_i^{o_i}, \theta_k^{(q)})$.

Similarly, the component $s, t \in [\![1, D]\!]$ of the covariance matrix Σ_k is updated as:

$$\Sigma_k^{st\,(q)} = \frac{\sum_i \tilde{t}_{ik}^{(q)}(S_{ik}^{st})^{(q)}}{\sum_i \tilde{t}_{ik}^{(q)}}$$

with for all $i \in S$, $k \in [\![1, K]\!]$, $s, t \in [\![1, D]\!]$,

$$\begin{aligned} &(S_{ik}^{st})^{(q)} \\ &= r_i^s r_i^t (x_i^s - \mu_k^{s\,(q)})(x_i^t - \mu_k^{t\,(q)}) \\ &+ r_i^s (1 - r_i^t)(x_i^s - \mu_k^{s\,(q)})(\eta_{ik}^{t\,(q)} - \mu_k^{t\,(q)}) \\ &+ (1 - r_i^s) r_i^t (\eta_{ik}^{s\,(q)} - \mu_k^{s\,(q)})(x_i^t - \mu_k^{t\,(q)})) \\ &+ (1 - r_i^s)(1 - r_i^t)((\eta_{ik}^{s\,(q)} - \mu_k^{s\,(q)})(\eta_{ik}^{t\,(q)} - \mu_k^{t\,(q)}) + \Gamma_{ik}^{st\,(q)}) \end{aligned}$$

Note that, because of the $\Gamma_{ik}^{st\,(q)}$ term in the last factor, this is not equivalent to replacing a missing variable x_i^s by the mean $(\eta_{ik}^s)^{(q)}$. This is consistent with the remark that mean imputation technique lowers the estimated variance ([5]).

This procedure will be refered in what follows as "*SFmiss algorithm*" standing for **S**imulated **F**ield approximation with **miss**ing values.

When Ψ is estimated, sites can be clustered using the MAP or MPM rule and mean field-like approximation. Such a rule consists in classing a site $i \in \mathcal{S}$ in the most probable class conditionally on observed values $x_i^{o_i}$:

$$z_i^{MAP} = \arg\max_k P_G(Z_i = k|x_i^{o_i}) = \arg\max_k \tilde{t}_{ik}$$

The experiments reported on the two next sections correspond to the Potts model with parameter $\beta \in \mathbb{R}^+$:

$$P_G(\mathbf{z}) = W^{-1} \exp(\beta \sum_{i\sim j} 1_{z_i = z_j}) = W^{-1} \exp(\beta n(\mathbf{z})) \quad (2)$$

where $i \sim j$ means that sites i and j are neighbours and $n(\mathbf{z})$ is the number of homogenous pairs (neighbours in the same class). The higher the parameter β, the smoother the

realization and the more importance is given to the network. Note that our model could be used with more general potential functions. Nevertheless, in particular for the yeast data, more complex models, seem to be penalized by their larger number of parameters (data not shown).

III. Validation on synthetic data

We first assess the performance of our method on image-like (*i.e.* the network is a regular grid) data for which real classes are known. It allows us to compute the accuracy of the method. We performed many experiments with different noises and missing value mechanisms. For sake of brievety, we illustrate here the performance of our method on two examples. Other experiments can be found in the supplementary material. We used 8 nearest neighbours as a neighbourhood setting. 4 nearest neighbours give comparable still slighlty deteriorated results. The correct number of clusters can be successfully determined by the BIC ([12]) criterion (see [13] or [3] for an effective use in the complete date case).

In a first experiment, we simulated a 2 class Potts model with parameter $\beta = 0.4$ (see equation 2). We considered data in $D = 10$ dimensions. Cluster means are set to $\mu_k = k \times (1, ..., 1)^T$ for $k = 1, 2$. The image is corrupted by a diagonal Gaussian noise $\mathcal{N}(0, 2\mathbb{I}_{10})$. This noise is equivalent to a standard deviation of 0.5 for one-dimensionnal Gaussian distribution in terms of Mahanalobis distance. It corresponds to Gaussian distributions that are not quite well separated. We then artificially removed randomly a certain proportion of data (MCAR case). Figure 1 (A) shows the percentage of missclassified pixels (sites) versus the percentage of missing data. 5 methods are compared: our SFmiss algorithm described in Section II, the corresponding EM algorithm for an independent mixture model (EMmiss), and the Simulated Field algorithm (SF) of [11] with prior imputation performed by K-Nearest neighbours (KNN+SF, *cf.* [6]), with filling in zeros (ZERO+SF) or column mean (MEAN+SF). Note that experiments with different β values were performed and gave similar results (data not shown).

We also considered a 4 classes synthetic image. Cluster means are chosen as $\mu_k = k \times (1, ..., 1)^T$ for $k = 1, ..., 4$ but here $D = 4$. We then corrupted the image by adding a non-diagonal Gaussian noise $\mathcal{N}(0, \Sigma)$ with $\Sigma_{dd'} = 0.5$ if $d = d'$ and 0.2 otherwise. This perturbation induce Mahanalobis distances between Gaussian distribution similar to those in the Potts model above. We then artificially removed a certain proportion of the highest and the lowest data (left and right censorship, NMAR case). Figure 1 (B) shows the percentage of missclassified pixels (sites) versus the percentage of missing data. Visualization of clustering output pictures are provided below the accuracy graph. 4 methods are compared: the SFmiss, EMmiss algorithms, and the SF algorithm with built-in unconditional (UMEAN+SF) or conditional (CMEAN+SF)class-dependent mean imputation. Various other experiments can be found in the supplementary material. According to all these experiments, it appears that the SFmiss algorithm performs better than imputation methods. It

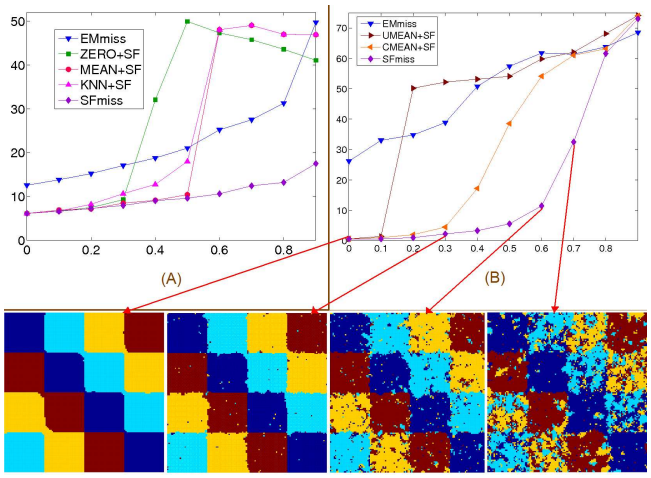


Figure 1. (A) Simulation of a Potts model: percentage of missclassified pixels vs. percentage of missing data (MCAR case), (B) 4 class synthetic image: idem but censored data (NMAR case) and visual results.

provides a way of modeling uncertainty over the value of the observation that is missing. Furthermore, taking into account dependencies (through the use of Markovian models) improves significantly the results. It can also be noted that our algorithm performs well even when the correct underlying model is not set as an hypothesis: the synthetic image is not a realization of a Potts model and censored data are NMAR! Censored data case seem to be more difficult than MCAR case but SFmiss provides reasonably good classifications for highs percentage of missing values (up to 60%).

IV. Experiments on yeast cell-cycle data

A. Individual Data

We use the study of [14] on *Saccharomyces cerevisiae* that focuses on the identification of cell-cycle regulated genes. The data set contains expression profiles from yeast cultures synchronized by different methods: α-factor arrest, arrests of $cdc15$ and $cdc28$ temperature sensitive mutants respectively and elutriation. Further it includes the data for the $Clb2$ and $Cln3$ induction experiments. The full data set consists of a 77 dimensional vector for each of the 6179 genes. It must be noted that the initial dataset has 5% overall missing entries.

In the following, we will focus on the $cdc28$ experiment initially performed by [15] ([14] used this data along with their own for their analysis). Yeast were synchronized by stopping them in late G1 phase of the cell cycle: the temperature was raised to $37°C$ and the cell cycle started again when the temperature was set to $25°C$. 17 time points were collected every 10 minutes so nearly two cell cycle occured. We chose this experiment because the synchrony was carefully controlled and because it is simpler to present the results. However the analysis was also performed on the full data set; results are not showed here for sake of briefness (see supporting material).

Many genes involved in cell cycle manifest themselves only once per cycle. Processes in which they are known to play

a key role include DNA (synthesis, replication, maintenance, *etc.*) processes, budding, nutrition or mitosis for example. Additionally, many of these genes control the cell cycle itself but it is not certain which transcription among them is required. [14] provide a list of 800 genes that are identified as cell cycle regulated thanks to a Fourier analysis of their expression pattern. Moreover, they assign a phase of the cell cycle to each of the genes thanks to a hierarchical clustering of the expression profiles. They exhibit five clusters of different phases of the cell cycle with according to peaks timing in the expression profiles of the 800 above genes. These temporal classes are recognized to be valuable as gold standard for the evaluation of clustering results. Hence we analyzed the set of genes contained within these 800 genes and that have interaction data, which we will now present.

B. Interaction Network

Many biological networks are (freely) available. They contain a lot of information that should not be ignored to provide optimal clustering. We aim at building a graph with biological entities (genes) at each nodes. These entities are subject to other individual measurements: transcript level measures in our study. An edge will stand for a confirmed link between two entities. It may relate a wide variety of evidences: interactions between genes, gene products, complexes of proteins, families, metabolic pathways,...Their format and reliability vary a lot.

As regards network data, we use the release 7 of STRING (http://string.embl.de/), a consistent database of known and predicted protein-protein interactions. It gathers information from a wide variety of different sources: genomic context, literature knowledge, physical interactions, *etc.* The current version contains $401\,948$ curated interactions [1] for 5611 genes of *Saccharomyces cerevisiæ*.

The intersection between the set of genes identified as cell-cycle regulated and those contained in the STRING database is considered. The resulting graph consists of 612 nodes (genes) and 3530 edges accounting for one or more interaction(s) treated equally (no weight on edges, see V). Just remind that each node has a 17-dimensional expression profile, with some dimensions possibly missing.

C. Cluster number selection

In this study the appropriate number of clusters is unknown. We computed Bayes Information Criterion (BIC, a penalized likelihood that accounts for the complexity of the model and the size of the data) [12] values with parameter K from 2 to 12 (*cf* Figure 2) for different models. Actually it is an approximation in our MRF setting but the approximation we used based on mean-field principle was shown efficient ([13]). A maximum is clearly visible for $K = 9$. Moreover the BIC doesn't increase with K above 9. It indicates that there is no gain in considering more clusters given the increased number of parameters. Hence $K = 9$ is the most likely number of clusters. Another comment is that the use of the network imply a clear benefit when taking the network into account (SF based algorithms) rather than ignoring it (EM algorithm). Actually, SF's curves are located EMmiss one. It means that SF based algorithms are a better compromise between gain in likelihood and model complexity. Last but not least the comparison between different ways of dealing with missing data shows that our integrated algorithm performs better than prior imputation ([6]'s KNNimpute or an autoregressive model of order 1). As a conclusion, the full SFmiss model is prefered among others. We will now analyze these results more precisely looking at the content of clusters to evaluate the clustering results.

[1] Two or more interactions may occur between the same couple of genes. This is just because differents kinds of interactions are considered.

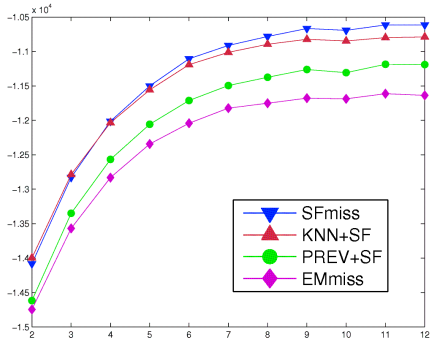


Figure 2. BIC values vs total number K of clusters for the SFmiss, KNN+SF, PREV+SF (SF algorithm with prior imputation thanks to a AR(t-1) model) and EMmiss algorithms.

D. Results/Validation

Assessing the quality of clusters produced by unsupervised algorithms is not an easy task. There is no consensus criterion to rely on. We therefore illustrate the gain in using our approach on some specific biological features of interest as regards the data under study. Also note that the presentation of the clustering results as a whole is very difficult to read. We used Cytoscape (http://www.cytoscape.org/) and GOstat (http://gostat.wehi.edu.au/) as tools to visualize and give a biologically meaningful overview of the clusters. The files of the full results are available for download on the paper website.

We first check whether the output clusters of our model are well-suited to summarize biological knowledge with respects to other algorithms. The full graph with colors assigned to nodes according to the clustering output of the SFmiss model is presented in the left part of Figure 3. A general trend is that the SFmiss algorithm gathers genes that are known to interact better than other algorithms. This is made visible on the zooms provided on the right side of Figure 3 (note that some reorganizations were made for improved readibility). This is consistent with the estimated spatial parameter β of our model: β is estimated to 0.41, which means that the neighbourhood plays a significant role.

The clusters reliability can be quantified using GO (Gene Ontology, http://www.geneontology.org/) terms representativity. GO defines an ontology: a structured and controled vocabulary to describe molecular attributes across organisms.

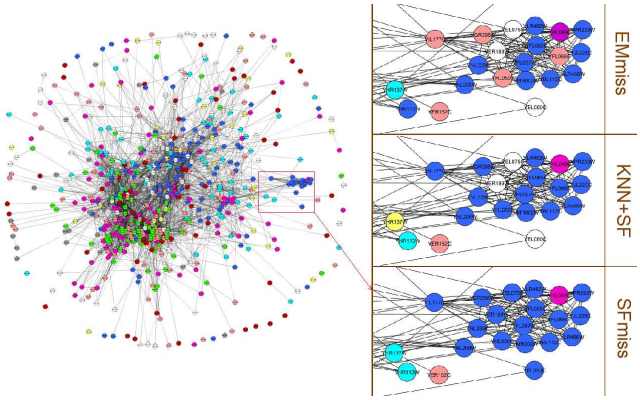


Figure 3. Left: entire graph with nodes colored according to SFmiss model clusters assignement. Right: zooms on the framed area of the left graph for different algorithms.

SFmiss cluster	GO terms & corresponding p-values	best p-values among EMmiss clust.	best p-values among KNN+SF clust.
k=3	GO:0006732, coenzyme met. process		
	$\mathbf{1.1\,10^{-2}}$	> 0.1	> 0.1
k=4	GO:0005819, spindle		
	$\mathbf{4.6\,10^{-9}}$	$6.7\,10^{-7}$	$2.0\,10^{-6}$
	GO:0006790, sulf. met. process		
	$\mathbf{1.1\,10^{-4}}$	$2.4\,10^{-4}$	$8.7\,10^{-4}$
	GO:0000278, mitotic cell cycle		
	$\mathbf{2.2\,10^{-3}}$	$7.7\,10^{-3}$	> 0.1
	GO:0030472, mit. spin. org. & biogen. in nucleus		
	$\mathbf{5.2\,10^{-3}}$	$8.8\,10^{-3}$	$2.0\,10^{-2}$
k=5	GO:0006974, resp. to DNA dam. stim.		
	$\mathbf{1.8\,10^{-3}}$	$3.0\,10^{-3}$	$8.0\,10^{-3}$
	GO:0000724, dbl-str. bk rep. via hom. comb.		
	$\mathbf{1.9\,10^{-2}}$	$2.7\,10^{-2}$	$4.6\,10^{-2}$
	GO:0000030, mannosyltransf. act.		
	$\mathbf{1.1\,10^{-2}}$	$1.2\,10^{-2}$	$2.7\,10^{-2}$
k=8	GO:0042555, MCM cplx		
	$\mathbf{3.4\,10^{-4}}$	$8.3\,10^{-4}$	$4.0\,10^{-4}$
	GO:0008026, ATP-dep. helicase act.		
	$5.5\,10^{-4}$	$1.3\,10^{-3}$	$\mathbf{4.5\,10^{-4}}$
	GO:0006268, DNA unwind. replic.		
	$2.8\,10^{-3}$	$6.7\,10^{-3}$	$\mathbf{1.1\,10^{-3}}$
	GO:0042623, ATPase act. coupl.		
	$\mathbf{4.4\,10^{-3}}$	$1.5\,10^{-2}$	$4.3\,10^{-2}$

Table I
SOME REPRESENTATIVE GO TERMS ANALYSIS OF CLUSTERS OBTAINED BY TESTED MODELS.

It comprises three different branches often referred to as taxonomy: molecular function –what a gene product does at the molecular level–, biological process –contribution of the gene product to a biological objective– and cellular component –location of the gene product in the cell. It is represented as a direct graph with vocabulary classes at nodes and oriented edge standing for hierarchical relations "is a" between nodes descriptors. We considered terms taken from these three branches because we noticed that genes can be associated with terms in different branches that give complementary informations on the biological process under study: the yeast cell cycle.

The more terms present in the data set are shared by genes in the same cluster, the more sensitive the method is. The more the 9 clusters isolate different parts of GO, the more specific the method is.

We present in Table I a summary of the results that allow us to conclude that the SFmiss algorithm is slightly more sensitive and specific than other cited methods. For each GO category, a test is performed to determine whether the category is over-represented in each cluster. Under-representation can be tested as well but under-representation analysis is not presented here for brievety reasons. One waits an over-represented category in one cluster to be under-represented in others for the method to be as specific as possible.

Because all GO labels present in the data set are tested, several hundred of tests are performed at a time. Thus usual individual testing error I (false positive) control is not adapted: if fixed at 5%, it would imply more than ten or so categories declared over-represented by error (in that case, 5 categories out of 100 are expected to be identified as over-represented just by chance). Hence multiple testing control is necessary. One of the most used and simple correction is the FWER (Family-Wise Error Rate) which is defined as the probability of making at least one type I error. When performing the Bonferoni control at a level of 5%, you are 95% sure (it means that if you do 100 analysis, the statement to follow is true 95 times on average) that the over-represented identified categories are not false positives. It is usually assumed rather conservative but it has been reported that the Bonferoni correction can be quite liberal especially when used with tests that show mutual dependencies ([16]). And indeed GO categories are not mutually independent for an obvious reason: their stacking.

An alternative to the FWER correction is the FDR (False Discovery Rate) namely the expected proportion of false positives among the positively identified tests. Generally, this type of correction is more appropriate for our purpose, since we would typically rather have more power (less false negatives) at the cost of a few more false positives.

P-values in the Table I are computed with the FDR correction of [17] which provides strong control over the FDR under positive regression dependency of the null hypotheses. It is not sure whether the GO hierarchy fulfills this positive regression dependency requirement. Nevertheless, this correction is widely used and has proven its efficiency. Alternatives include the [18] procedure, which controls the FDR under arbitrary dependency. However this latter one exhibits severely decreased power compared to the former correction; this is the price to pay for allowing arbitrary dependencies.

Apart from few exceptions, the SFmiss model performs better at grouping genes with similar annota-

tions than other algorithms [2]. For a detailled example, genes $YDL105W$, $YER111C$, $YKR077W$, $YJL196C$, $YLR212C$ and $YNL082W$ are found in cluster 1 produced by the SFmiss algorithm and are not present in the corresponding EMmiss cluster. All of them are brought into play during cell cycle processes: mitotic spindle complex repair ($YLR212C$) or G1/S transition of the mitotic cycle ($YER111C$) for example. $YKR077W$ is annotated as a putative transcription activator. Our method suggests that this annotation is fully coherent and that this gene plays a key role either as a cell cycle regulator or as a regulated gene of the process. So it does not only summarize known biological knowledge but can give directions for putative functional guess on components of living organisms based on the clustering results. We can also illustrate the advantage of accounting for missing values in a united fashion as compared to prior filling-in with [6]'s KNNimpute. Genes $YBL002W$, $YGL093W$ and $YPL269W$ belong to SFmiss cluster 4 and not to the corresponding KNN+SF cluster. Their annotations are making sense when compared with those of their cluster and confirm a possible functional description of the cluster: chromatin assembly, required for accurate chromosome segragation localized to the nuclear side of the spindle pole body and required for cytoplasmic microtubule orientation in yeast (polarization) respectively. All references for these statements can be found on the supplementary material website.

Another nice feature of the outputs of the SFmiss model is that we were able to interpret clusters when compared to temporal classes of the cell cycle identified by [14]: G1, S, S/G2, G2/M and M/G1. So cluster 0 is almost entirely included in [14]'s G2/M group and cluster 1 is in G1 just like cluster 5. Cluster 2 include genes regulated in late G2, M and early G1 phases (quite broad, certainly a reason why no specific function is highlighted in this cluster). Cluster 3 is similar but with a more coondensed expression pattern and with earlier start. Cluster 8 is focused on M-regulated genes. Cluster 4 shows its temporal peak in S phase. Lastly, cluster 6 has many genes from early G2 to M.

These interpretations are corroborated if we investigate the expression profiles for each meaningful clusters. Examples of such additional evidences are given in Figure 4. These profiles are very similar to those obtained by [15]'s Figure 4 (C, D) for annotated genes.

Last but not least we would like to present another major advantage of our approach: it responds with a much greater level of stability than any other method when the number of observed data decreases. This is illustrated in Figure 5. Additional missing values were generated under MCAR. All algorithms perform poorly in the NMAR case (data not shown) on real data simply because we generated NMAR data thanks to censorship. Hence it makes it very difficult to analyze gene expression data when most significant ratios are very likely

[2]the case of cluster $k = 8$ in Table I is a particularly adverse situation where the cluster found to be the most similar has 43 genes whereas the SFmiss cluster has 61 genes. Consequently, a slight enhancement in grouping obejcts with similar function does not necessarily imply a better p-value!

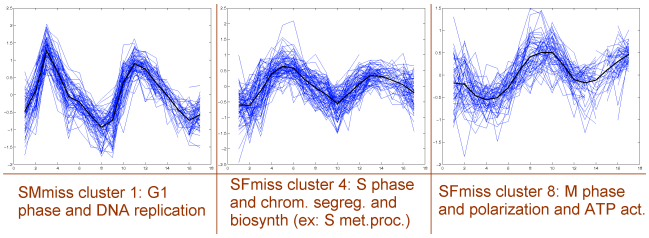


Figure 4. Examples of expression profiles for three SFmiss clusters.

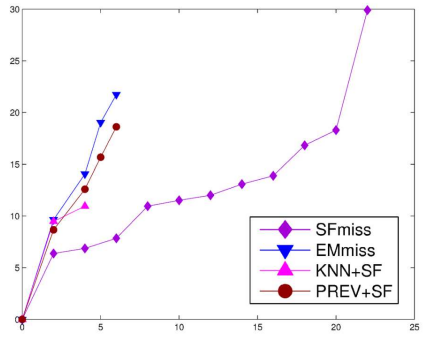


Figure 5. Percentage of error for different algorithms vs. percentage of added (to the inherent $\sim 5\%$ in the dataset) missing value.

to be removed! The reference clustering for comparison was the one resulting from the considered algorithm when no additional data was removed. Apart from SFmiss, all algorithms show dramatical unstability when the rate of added missing value increases above 4%. Above 7%, we were not able to give classification error because clustering results were too far from the initial one for EMmiss, KNN+SF and PREV+SF. This suggests that these algorithms have a very unpleasant behaviour when they are facing datasets even with "as few as" nearly 9% of total missing data in the favourable MCAR case. On the contrary, the SFmiss algorithm shows a very good stability towards the rate of missing value. Its performance impairs significantly above 25% of overall missing data which is quite acceptable as regards real situation encountered.

V. Conclusions and Future Work

Missing data can bring many difficulties in data analysis simply because most methods procedures were not designed for them. This is particularly true in the context of the integration of post-genomic data. Data absence is usually a nuisance, not the focus of inquiry. We presented a comprehensive integrated statistical tool for modelling individual measurements that have a network-dependant structure. To this occasion, we overcame raised conceptual and computational challenges. We demonstrated the performance of our method both on synthetic and real biological datasets.

These encouraging results lead on further study. We plan to analyze a dataset with all yeast genome under the supervision of a biological expert. We restricted our analysis to genes with prior knowledge for validation purpose. Another prospect would be to take into account the missingness mechanism

to improve performances in NMAR generated data. A final plan is to account for missing edges as we did for missing individual measurements. However this sounds to us like a very difficult task: we would need to consider every possible interactions between nodes, a computationally unbearable solution. A possible alternative would be to consider confidence levels for interactions; gene interaction data are known to be incomplete. Their reliability vary a lot when reported by two-hybrid screening for example. This latter technique is often criticized because it identifies many false positive and negative pairs of interacting proteins. Integrating such information would be highly beneficial. This feature is already available with our algorithm but we don't have substancial data to analyze in scope.

ACKNOWLEDGMENT

The authors would like to thank Florence Forbes and Jim McNicol for fruitful discussions and critical reading of the manuscript. We are also grateful to the anonymous reviewers for their comments on the present work.

REFERENCES

[1] D.-W. KIM, K.-Y. LEE, K. H. LEE, and D. LEE, "Towards clustering of incomplete microarray data without the use of imputation," *Bioinformatics*, vol. 23, no. 1, pp. 107–113, 2007.

[2] K. Y. YEUNG, C. FRALEY, A. MURUA, A. RAFTERY, and L. RUZZO, "Model-based clustering and data transformations for gene expression data," *Bioinformatics*, vol. 17, no. 10, pp. 977–987, 2001.

[3] M. VIGNES and F. FORBES, "Gene clustering via integrated Markov models combining individual and pairwise features," *IEEE/ACM Transactions on Computational Biology and Bioninformatics*, vol. to appear, 2007.

[4] T. H. Bø, B. DYSVIK, and I. JONASSEN, "LSimpute: accurate estimation of missing values in microarray data with least square methods," *Nucleic Acids Research*, vol. 32, no. 3, p. e34, 2004.

[5] R. J. LITTLE and D. B. RUBIN, *Statistical analysis with missing data*, 2nd ed., ser. Probability and Statistics. Wiley, 2002.

[6] O. TROYANSKAYA, M. CANTOR, GavinSherlock, P. BROWN, T. HASTIE, R. TIBSHIRANI, D. BOTSTEIN, and R. B. ALTMAN, "Missing values estimation methods for DNA microarrays," *Bioinformatics*, vol. 17, no. 6, pp. 520–525, 2001.

[7] S. OBA, M. aki SATO, I. TAKEMASA, M. MONDEN, K. ichi MATSUBARA, and S. ISHII, "A bayesian missing value estimation method for gene expression profile data," *Bioinformatics*, vol. 19, no. 16, pp. 2088–2096, 2003.

[8] M. OUYANG, W. J. WELSH, and P. GEORGOPOULOS, "Gaussian mixture clustering and imputation of microarray data," *Bioinformatics*, vol. 20, no. 6, pp. 917–923, 2004.

[9] M. S. SEHGAL, I. GONDAL, and L. S. DOOLEY, "Collateral missing value imputation: a new robust missing value estimation algorithm for microarray data," *Bioinformatics*, vol. 21, no. 10, pp. 2417–2423, 2005.

[10] A. P. DEMPSTER, N. M. LAIRD, and D. B. RUBIN, "Maximum likelihood from incomplete data via the EM algorithm," *Journal of the Royal Statistical Society*, vol. 39, no. 1, pp. 1–38, 1977.

[11] G. CELEUX, F. FORBES, and N. PEYRARD, "EM procedures using mean-field like approximations for Markov-model based image segmentation," *Pattern recognition*, vol. 36, pp. 131–144, 2003.

[12] G. SCHWARZ, "Estimating the dimension of a model," *The Annals of Statistics*, vol. 6, no. 2, pp. 131–134, 1978.

[13] F. FORBES and N. PEYRARD, "Hidden markov random field model selection criteria based on mean field-like approximations," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 25, no. 9, pp. 1089–1101, 2003.

[14] P. T. SPELLMAN, G. SHERLOCK, M. Q. ZHANG, V. R. IYER, K. ANDERS, M. B. EISEN, P. O. BROWN, D. BOTSTEIN, and B. FUTCHER, "Comprehensive identification of cell cycle-regulated genes of the yeast *Saccharomyces cerevisiae* by microarray hydridization," *Molecular Biology of the Cell*, vol. 9, no. 12, pp. 3273–3297, 1998.

[15] R. J. CHO, M. J. CAMPBELL, E. A. WINZELER, L. STEINMETZ, A. CONWAY, L. WODICKA, T. G. WOLFSBERG, A. E. GABRIELIAN, D. LANDSMAN, D. J. LOCKHART, and R. W. DAVIS, "A genome-wide transcriptional analysis of the mitotic cell cycle," *Molecular Cell*, vol. 2, pp. 65–73, 1998.

[16] E. I. BOYLE, S. WENG, J. GOLLUB, H. JIN, D. BOTSTEIN, J. M. CHERRY, and G. SHERLOCK, "GO::TermFinder–open source software for accessing Gene Ontology information and finding significantly enriched Gene Ontology terms associated with a list of genes," *Bioinformatics*, vol. 20, no. 18, pp. 3710–3715.

[17] Y. BENJAMINI and Y. HOCHBERG, "Controlling the false dicovery rate – a practical and powerful approach to multiple testing," *Journal of the Royal Statistical Society B*, vol. 57, no. 1, pp. 289–300, 1995.

[18] Y. BENJAMINI and D. YEKUTIELI, "The control of the false discovery rate in multiple testing under dependency," *Annals of Statistics*, vol. 29, no. 4, pp. 1165–1188, 2001.

• *Résumé* :

Les recherches que nous présentons dans ce mémoire s'inscrivent dans le cadre de l'intégration statistique de données post-génomiques hétérogènes. La classification non supervisée de gènes vise à regrouper en ensembles significatifs les gènes d'un organisme, vu comme un système complexe, conformément aux données expérimentales afin de dégager des actions concertées de ces gènes dans les mécanismes biologiques mis en jeu.

Nous basons notre approche sur des modèles probabilistes graphiques. Plus spécifiquement, nous utilisons l'outil de champs de Markov cachés qui permet la prise en compte simultanée de données propres à chacun des gènes grâce a des distributions de probabilités et de données traduisant un réseau d'interaction au sein de l'organisme à l'aide d'un graphe non-orienté entre les gènes.

Apres avoir présenté la problématique et le contexte biologique, nous décrivons le modèle utilisé ainsi que les stratégies algorithmiques d'estimation des paramètres (*i.e.* approximations de type champ moyen). Puis nous nous intéresserons à deux particularités des données auxquelles nous avons été confrontés et qui amènent des développements du modèle utilisé, notamment la prise en compte de l'absence de certaines observations et la haute dimensionnalité de celles-ci. Enfin nous présenterons des expériences sur données simulées ainsi que sur données réelles sur la levure qui évaluent le gain apporté par notre travail. Notamment, nous avons voulu mettre l'accent sur des interprétations biologiques plausibles des résultats obtenus.

• *Abstract* :

The research work presented in this dissertation is on keeping with the statistical integration of post-genomics data of heterogeneous kinds. Gene clustering aims at gathering genes of a living organism –modeled as a complex system– in meaningful groups according to experimental data to decipher the roles of the genes acting within biological mechanisms under study.

We based our approach on probabilistic graphical models. More specifically, we used Hidden Markov Random Fields (HMRF) that allow us to simultaneously account for gene-individual features thanks to probability distributions and network data that translate our knowledge on existing interactions between these genes through a non-oriented graph.

Once the biological issues tackled are set, we describe the model we used as well as algorithmic strategies to deal with parameter estimation (namely mean field-like approximations). We then examine two specificities of the data we were faced to : the missing observation problem and the high dimensionality of this data. They lead to refinements of the model under consideration. Lastly, we present our experiments both on simulated and real Yeast data to assess the gain in using our method. In particular, our goal was to stress biologically plausible interpretations of our results.