Agrégation de classifieurs et d'experts pour la recherche d'homologues chez les cytokines à quatre hélices alpha

Nicolas Beaume

UNIVERSITE DE NANTES

FACULTE DE MEDECINE

Agrégation de classifieurs et d'experts pour la recherche d'homologues chez les cytokines à quatre hélices alpha

THESE DE DOCTORAT

Ecole Doctorale CHIMIE-BIOLOGIE
Discipline SCIENCES DE LA VIE ET DE LA SANTE
Spécialité BIOINFORMATIQUE

présentée et soutenue publiquement par

NICOLAS BEAUME

le 25 juin 2008, devant le jury ci-dessous

Président	: M. Jean-Jacques Schott, CR, UMR 915, Nantes
Rapporteur	: M. Alain Guénoche, DR, IML, Marseille
Rapporteur	: M. Hugues Gascan, DR, INSERM U564, Angers
Examinateur	: M. Younès Bennani, Professeur, Paris XIII, Paris
Directeur de thèse	: M. Yannick Jacques, DR, INSERM U892, Nantes
Co-encadrant	: M. Gérard Ramstein, MC, EPUN, Nantes

UNIVERSITE DE NANTES

FACULTE DE MEDECINE

Agrégation de classifieurs et d'experts pour la recherche d'homologues chez les cytokines à quatre hélices alpha

THESE DE DOCTORAT

Ecole Doctorale CHIMIE-BIOLOGIE
Discipline SCIENCES DE LA VIE ET DE LA SANTE
Spécialité BIOINFORMATIQUE

présentée et soutenue publiquement par

NICOLAS BEAUME

le 25 juin 2008, devant le jury ci-dessous

Président	: M. Jean-Jacques Schott, CR, UMR 915, Nantes
Rapporteur	: M. Alain Guénoche, DR, IML, Marseille
Rapporteur	: M. Hugues Gascan, DR, INSERM U564, Angers
Examinateur	: M. Younès Bennani, Professeur, Paris XIII, Paris
Directeur de thèse	: M. Yannick Jacques, DR, INSERM U892, Nantes
Co-encadrant	: M. Gérard Ramstein, MC, EPUN, Nantes

A Flo, qui a supporté un de ces illuminés d'apprentis-chercheurs pendant la pire moitié d'une thèse.
A Claire, qui m'a fait faire mes premiers pas dans le monde de la bioinformatique.

Seul l'impossible mérite réflexion

Deszö Kosztolányl

Remerciements

Les remerciements sont sûrement une des parties les plus agréables à écrire, d'une part parce que c'est l'une des dernières à être rédigées mais plus important parce que c'est un des (rares) espaces de liberté d'une thèse, où l'on peut librement s'adresser à tous ceux que notre chemin a croisés pendant ces quelques années.

Profitant de cet espace,
Je vais m'accorder le double plaisir
De dire avec grâce
A ceux qui m'ont aidé à parcourir
Ce chemin bordé de précipices
Ma gratitude éternelle
Tout en succombant à l'un de mes vices.
Comme Cyrano se battait en duel,
C'est en rimes et en vers
Que j'exprimerai mes émotions les plus sincères.

Il me faut tout d'abord remercier
Yannick et Gérard, qui m'ont dirigé,
M'ont abreuvé de conseils et d'idées
Et ont éclairé de la flamme de leur expérience
Ce tronçon de mon chemin en science.
Merci à tous les deux pour vos suggestions avisées
Et de votre patience.

A tous ceux qui à Nantes, m'ont accueilli
Collègues de travail et sûrement amis
Pour les sérieuses discussions
Et quelques moments de détente
Les repas que nous partagions
Et les idées passionnantes,
Qui ont rendu ces années un peu plus brêves.
Merci en particulier à Julien, mon frère de thèse.

Au delà de mes voisins de bureau,
J'adresse mes remerciements à tous ceux
Qui m'ont soutenu, et ce n'est pas peu,
Dans la vie hors de mon bureau.
Flo, première parmi ceux-là,
Fut, est et sera
Une compagne stimulante
Drôle et patiente
Qui supporte avec humour
L'intellectuel que je serai toujours
Parfois perdu dans ses chimères.
Merci, ma tanguera, d'être si entière
Et de m'avoir soutenu,
Malgré le sentier ardu.

A ma famille, parents, frères et presque belle-soeurs,
vous qui, durant toutes ces années d'étude,
N'avez jamais manqué d'encouragements et de sollicitude,
Je voulais vous dire merci du fond du coeur.
Chaque pensée de vous, et il y en eu pléthore,
M'est, à travers le sang, parvenue,
Et fut une raison de plus

De poursuivre l'effort.

Je n'oublierai pas mes amis,
De Nantes, Toulouse, Londres ou Paris
Vieux compagnons indéfectibles,
Partenair(e)s de tango ou de karaté.
Un remerciement tout particulier,
A Claire pour les PBIB,
Mais aussi pour ses conseils,
Nos long échanges par emails,
Et sa patience infinie,
En relisant ce manuscrit,
Relevant avec justesse,
La moindre de ses faiblesses.

Mille excuses à tous ceux que je n'ai pas pu citer,
Que tous ces collègues, parents ou amis ne se croient pas oubliés,
De tous ceux qui pendant ces trois ans et demi,
M'ont aidé, encouragé ou simplement suivi,
Je garderai un vif souvenir,
Et de ces moments partagés, un grand plaisir.

A toutes et à tous, mes remerciements les plus sincères

Résumé

Cette thèse, à l'interface entre la biologie et l'informatique, s'intégre dans le champ de l'extraction de connaissances appliqué aux données biologiques. Je me suis intéressé à une famille de protéines, les cytokines à quatre hélices alpha, connues pour leurs implications dans la réponse immunitaire et l'inflammation. L'objectif de ce travail est la mise au point d'une méthode de détection d'homologues inconnus, qui pourraient se révéler, entre autres, des cibles ou des agents thérapeutiques intéressants. Un travail précédent ayant démontré que les Séparateurs à Vastes Marges (SVM) sont la technique la plus efficace pour rechercher les homologues éloignés, j'ai comparé plusieurs classifieurs utilisant cette stratégie, sur la base de leur capacité à classer les cytokines à quatre hélices alpha parmi un ensemble de contre-exemples. Pour affiner les résultats de ces classifieurs, j'ai proposé d'ajouter, sous la forme d'experts automatiques, des connaissances spécifiques à la famille étudiée. Bien qu'elles ne soient pas nécessairement discriminantes par elles-mêmes, ces connaissances, combinées aux classifieurs, améliorent sensiblement les capacités de discrimination du système. Enfin, afin de maximiser l'efficacité de l'association classifieurs-experts, j'ai comparé différentes méthodes d'agrégation et sélectionné la plus adaptée à la classification des cytokines à quatre hélices alpha. Je propose à l'issue de ce travail une méthode performante, utilisant aussi bien des techniques génériques que des outils adaptés au problème. Elle a été optimisée pour la recherche d'homologues de cytokines à quatre hélices alpha humaines, mais s'avère facilement généralisable à d'autres familles de protéines.

Abstract

This thesis, taking place at the interface between Biology and Computer Science, is confronted with the problem of the extraction of knowledge from biological data. I was working on a particular gene family : the four helix cytokines. The cytokines are involved in numberous physiological process, including immune response, making it an important therapeutique target. Althougth one of the largest gene family of the human genome, previous observations suggest that all four helix cytokines have not been identified. The major purpose of this work was to obtain classification models from our cytokines families set of interest to detected still unknown members in the human genome. A previous work in our team demonstrate that Support Vector Machine (SVM) was the best strategy for homologs research and lead to the developement of a SVM classifier which achieve good results at classifying cytokines. The first part of my work was to add SVM classifiers from the literature, to evaluate all those classifiers and to complete the developement of a tool which handles the whole discovery process, from data storage to identification of putative members. It comprises five classifiers especially designed for biological sequence. During the second part of my work, i designed automatical exeperts which deal with information like proteine structure, length of the sequence or other biological which can't be include by the way of SVM classifiers. Those experts can be add to the classifiers via aggregation to achieve better ranking of the candidates. The last part of my work consisted in evaluating methods to optimize aggregation of classifiers and experts into an unique ranking of candidates. A large spectrum of aggregation methods have been tested and some of them achieve better results than the best classifer alone. This methode can easily be adapted to another gene family of interest.

Table des matières

Abréviations et Notations

Cytokines

BCRF : Bam HI C fragment rightward Reading Frame
BSF-3 : B cell Stimulating Factor 3
CLC : Cardiotrophique Like Cytokine
(G/GM-)CSF : (Granulocytes/Granulocyte Macrophage-)Colony Stimulating Factor
CT-1 : Cardiotrophine 1
CNTF : Ciliary Neutrophic Factor
EPO : Erythropoiétine
FLT-3 : Fms-Like Tyrosine kinase receptor-3 gp130 : Glycoprotéine 130
IFN : Interféron
IL- : Interleukine
LIF : Leukemia Inhibitory Factor
NNT-1 : Novel Neutrophin 1
NGF : Nerve Growth Factor
OSM : Oncostatine M
SCF : Stem Cell Factor
TGF : Transforming Growth Factor
TNF : Tumor Necrosis Factor
TSLP : Tymic Stromal Lymphopoietic Cytokine

Les récepteurs de cytokines sont généralement nommé à partir du nom de la cytokine auquel on ajoute un "R" *e.g.* EPOR pour le récepteur de l'EPO.

Biologie

ADN : Acide DeoxyriboNucléique
Alk-1 : Orphan activin receptor-Like Kinase 1
AP-1 : Activator Protein-1
ARN : Acide RiboNucléique
ATF2 : alcohol acetyltransferase II
bZip : leucine-zipper
CMH : Complexe Majeur d'Histocompatibilité
FLIP : Fas-associated death-domain-Like IL-1beta-converting enzyme-Inhibitory Protein
FNIII : fibronectin type III
GAS : Gamma-Activated Sequences
HCP : Hematopolietic Cell Phosphatase
Ig : Immunoglobuline
ISRE : Interferon-Stimulated Response Elements
JAK : Janus Associated Kinase
kDa : KiloDalton
LB : Lymphocyte B
LT : Lymphocyte T
MAPK : Mitogen-Activated Protein Kinase
MEF-2C : Myocyte Enhancer Factor-2C
NFκB : Nuclear Factor κB
NK : Natural Killer
PI-3 kinase : PhosphoInositide 3-kinase
PLC : Phospholipase C
RMN : Résonance Magnétique Nucléaire
SBE : STAT Binding Element
SIDA : Syndrome Immunitaire Deficient Acquis
SH2 : Src-Homology 2
STAT : Signal Transducer and Activator of Transcription
TCR : T Cell Receptor

Th : Lymphocyte T helper
Treg : Lymphocyte T régulateur
Tr1 : Lymphocyte T régulateur de type 1
VIH : Virus de l'Immunodéficience Humaine

Bioinformatique

API : Application Programming Interface
AUC : Area Under (ROC) Curve
BLAST : Basic Local Alignment Search Tool
dVC : Dimension de Vapnik-Chervonenkis
E-MM : expert basé sur la masse moléculaire de la protéine
E-PI : Expert basé sur le point isoélectrique de la protéine
E-SS : expert basé sur la structure secondaire de la protéine
E-T : Expert basé sur la taille de la séquence
FFF : Fuzzy Functional Forms
HMM : Hidden Markov Models
LA kernel : Local Alignment kernel
max : méthode du maximum
min : méthode du minimum
PDB : Protein Data Bank
PSI-BLAST : Position Specific Iterative BLAST
RBF : Radial Basis Function
ROC : Receiver Operating Characteristic curve
SCOP : Structural Classification Of Proteins
SOV : Segment OVerlapping
SVM : Support Vector Machine / Séparateur à Vaste Marge
SW score : score de Smith & Watermann

Introduction

Cette thèse a pour principal objectif la mise en place d'une méthode de recherche de nouvelles cytokines à quatre hélices α, une famille de protéines connues pour leur rôle dans la réponse immunitaire, dans le génome humain. Cette méthode a également pour ambition d'être généralisable à d'autres familles de protéines.
Des travaux précédents ont démontré l'efficacité des SVM pour les problèmes de recherche d'homologues, entre autre chez les cytokines à quatre hélices α. En m'appuyant sur ces travaux, j'ai analysé les performances de plusieurs classifieurs basés sur cette méthode.
Afin d'améliorer les performances de ces classifieurs, j'ai proposé l'utilisation de connaissances spécifiques à la famille, au travers d'outils appelés "experts". J'ai ainsi décris et évalué quatre experts dédiés aux cytokines à quatre hélices α.
Enfin, j'ai cherché à combiner ces classifieurs et ces experts de manière optimale à l'aide de méthodes d'agrégation. Après évaluation de ces méthodes, j'ai retenues deux d'entre elles, dont une particulièrement, comme les efficaces pour combiner classifieurs et experts.
Dans cette introduction, je décrirai le contexte de ce travail, les problèmes posés ainsi que l'intérêt dont ils relèvent. Je présenterai brièvement les solutions proposées sous forme des contributions que j'ai apporté à ce champ d'étude, avant d'expliquer l'organisation de ce manuscrit.

Contexte

Cette thèse se focalise sur un groupe de protéines appelées "cytokines" qui sont connues entre autres pour leurs rôles dans la réponse immunitaire. Découvertes dans les années 60 et particulièrement étudiées à partir des années 70, les cytokines demeurent un vaste champ de recherches à l'heure actuelle. Ces protéines sont associées à des mécanismes cruciaux tels que l'inflammation et à de nombreuses pathologies comme le psoriasis, la maladie de Crohn ou certaines formes de cancers, ce qui explique l'intérêt qu'elles ont suscité et suscitent toujours.
Il s'est rapidement avéré que les cytokines étaient une famille complexe à étudier du fait des diverses fonctions et des interactions multiples entre chacun de ses membres. S'il est acquis que les cytokines induisent des signaux cellulaires à courte et moyenne distance, le signal délivré par chaque cytokine est, quant à lui, spécifique de celle-ci et de son environnement. Alors que dans les années 80 les cytokines étaient classées selon leur tendance pro-ou anti-inflammatoire, on les considère de nos jours comme des protéines avec un vaste spectre d'action et s'insérant au sein de réseaux de régulation qui déterminent leurs multiples fonctions. Cette vision bien plus complexe explique l'effort porté par de nombreuses équipes de recherche pour élucider le fonctionnement précis de cette famille.

Pour transmettre un signal, les cytokines interagissent avec des récepteurs membranaires à la surface de la cellule cible. Ces récepteurs activent des voies de signalisation qui vont conduire à la régulation de certains gènes, induisant un changement d'état, généralement la prolifération, la différenciation ou l'apoptose de la cellule. Il existe un vaste jeu de récepteurs aux cytokines, chacun d'entre eux étant capable de fixer plusieurs ligands et d'activer différentes voies de signalisation. Cette variété de combinaisons est une des principales explications du large spectre d'action des cytokines et de leurs redondances fonctionnelles mais il n'en demeure pas moins qu'un certain nombre de mécanismes demeurent inexpliqués, du fait de la com-

plexité du réseau d'interaction. Certaines observations pointent des "trous" dans ce réseau d'interaction, suggérant l'existence d'acteurs encore inconnus. Le fait que de nouvelles cytokines soient régulièrement découvertes et que l'une des dernières découvertes présente un mécanisme d'action inconnu jusqu'alors, indique clairement que les membres de cette famille ne sont pas identifiés de manière exhaustive. La découverte de nouvelles cytokines est un enjeu important, à plusieurs titres. D'un point de vue purement scientifique, ces découvertes apportent des données supplémentaires pour mieux comprendre la famille et son fonctionnement, en permettant de lever des zones d'ombre dans les connaissances actuelles ou en identifiant de nouveaux mécanismes. D'un point de vue médical, la découverte de nouvelles cytokines permet d'étendre le spectre des cibles et agents thérapeutiques potentiels contre les pathologies où ces protéines sont impliquées.

Depuis une vingtaine d'années, de nombreuses équipes se sont attelées à cette recherche, ce qui a mené à la découverte d'un grand nombre de ses membres. Toutefois, ces travaux basés sur des algorithmes simples de recherche de similarité de séquences, n'ont fait ressortir à chaque fois qu'une ou deux nouvelles cytokines et aucune méthode n'a permis de mettre en évidence de façon exhaustive l'ensemble des membres de cette famille. Cela est probablement dû à la nature très hétérogène des membres de cette famille. En effet, un simple alignement de séquences montre que les cytokines ont une faible similarité entre elles. L'histoire de l'étude des cytokines montre que l'identification de nouveaux membres à grandement modifié la perception qu'on en avait, entre autre parce que ces nouveaux membres n'avaient pas la même nature ou la même origine que les cytokines connues jusqu'à maintenant. Ainsi la découverte de membres qui n'étaient pas produits par des lymphocytes, contrairement à ce qui était connu jusqu'alors, à provoqué le changement de nom de la famille de "lymphokine" en "cytokine". Il semblerait donc que les cytokines ne forment pas un ensemble homogène mais plutôt un agglomérat de sous-familles parta-

geant des propriétés fonctionnelles. Ces sous-familles peuvent être définies à l'aide de critères d'homologie, particulièrement la structure protéique, et constituent donc des familles de protéines, au sens évolutif du terme.

Formulation du problème et solutions envisagées

Ainsi que je viens de le présenter, les cytokines forment une vaste famille de protéines, aux multiples intérêts, tant thérapeutiques que purement intellectuels. Ces protéines forment un réseau d'interaction complexe dont le fonctionnement reste encore à élucider. De nombreuses observations tendent à montrer que tous les membres de ce réseau n'ont pas été encore identifiés, ce qui nuit à la compréhension de certains mécanismes. La découverte de ces éléments inconnus est donc cruciale.
La principale difficulté qui se présente est la grande hétérogénéité de cette famille, dont les critères d'appartenance ne sont pas toujours liés à des considérations évolutives. Ceci peut expliquer que les méthodes de recherche d'homologues existantes ne soient pas parvenues à en identifier tous les membres.

Le problème qui se dégage de ces observations est donc : comment mettre en évidence l'ensemble des homologues de la familles des cytokines ?
Un élément de réponse est fourni par le fait qu'il existe parmi les cytokines des sous-familles qui sont, elles, des familles de protéines au sens évolutif du terme. Une recherche sous-famille par sous-famille paraît donc être la stratégie optimale pour mettre en évidence l'ensemble des homologues de cytokines. Malheureusement, ces sous-familles elles-mêmes présentent une certaine hétérogénéité dans les séquences, reflétée par une faible similarité en leur sein et un nouvel échec des méthodes de recherche d'homologues existantes. Toutefois, cette approche permet la décomposition du problème en sous-problèmes que l'on peut qualifier de "mieux posé" puisqu'on se place dans le cadre de véritables familles de protéines. Dans ce optique, il est possible de développer des outils adaptés au cas d'une famille de

protéines hétérogènes.

Dans cette thèse, j'ai choisi de m'intéresser à la sous-famille de cytokines à quatre hélices α, qui présente une remarquable similarité de structure secondaire et qui est la plus grande des sous-familles de cytokines, puisqu'elle représente plus d'un tiers de cette famille.
L'objectif de cette thèse étant de rechercher les membres inconnus d'une manière exhaustive, l'utilisation de méthodes bioinformatiques semble s'imposer. Il existe dans la littérature un vaste panel de méthodes de recherche d'homologues, dont l'intérêt dépend du problème posé. Dans le cadre d'un travail précédent, plusieurs méthodes génériques ont été évaluées et l'une d'entre elles, la méthode des Séparateurs à Vastes Marges (SVM) s'est avérée la plus performante dans le cas qui nous concerne. Cette méthode, issue du champ de la classification supervisée, permet, à partir d'un jeu d'exemples et de contre-exemples d'objets à reconnaître, de créer un modèle de reconnaissance, pouvant être appliqué par la suite à des objets inconnus. Les SVM possèdent entre autres une bonne capacité de généralisation du modèle, ce qui est explique son efficacité à reconnaître des objets hétérogènes comme les cytokines à quatre hélices α. Comme toutes les méthodes de classification, les SVM passent par une phase d'extraction de caractéristiques jugées intéressantes par le décideur, le choix de ces caractéristiques étant évidement un point important pour les performances finales de l'outil. On trouve dans la littérature plusieurs classifieurs (*i.e.* des méthodes d'extraction de caractéristiques couplées à une technique de classification) à base de SVM et spécialisés pour les séquences biologiques. La diversité de ces classifieurs nécessite une évaluation de chacun d'entre eux pour déterminer les plus efficaces relativement au problème posé, mais ouvre également la voie à une utilisation combinée de ces classifieurs.

La bioinformatique est un champ théorique qui a essentiellement pour objectif de proposer des modèles généraux applicables à des données biolo-

giques. La capacité de traitement relativement rapide des outils informatiques, associé à un coût réduit, en fait une méthode de "première ligne", qui permet de réduire le champ d'investigation des méthodes biologiques, plus longues et plus coûteuses. Un outil informatique a donc pour vocation d'écarter au maximum les données erronées pour ne laisser à la validation biologique que les données les plus sûres. Le traitement bioinformatique tend donc de plus en plus à ajouter aux traitements directement issus de l'informatique, un post-traitement biologique. Ici se pose un dilemme entre automatisation ou expertise manuelle par un annotateur humain. L'expertise automatique a l'indéniable avantage de traiter un grand nombre de données en un court laps de temps, mais s'avère moins précise que l'annotation manuelle, qui est plus laborieuse. Selon le problème posé, le recours à l'une où à l'autre des options, ainsi que le choix des critères d'expertise a un impact sur le type de résultat.

Motivation de la thèse

Cette thèse s'articule autour de deux thèmes de recherche, d'une part la mise en évidence d'homologues de membres connus d'une famille de protéines et d'autre part l'étude d'une famille de protéines précise : la famille des cytokines à quatre hélices α. Cette partie a pour objectif d'expliciter les raisons qui ont conduit à la mise en place de ce projet et à certains choix que j'ai effectués.

L'intérêt premier de ce travail est bien sûr centré sur les cytokines. Ainsi que je l'ai dit ci-dessus, les membres de cette famille sont impliqués, comme agents ou comme causes, dans de nombreuses pathologies dont certains fléaux du siècle dernier et de celui-ci, tels que le cancer ou le SIDA. La découverte de nouvelles cytokines est donc porteuse d'un espoir thérapeutique important. L'apport d'une meilleure compréhension des mécanismes de fonctionnement de cette famille, permet d'envisager à terme une meilleure compréhension de son implication dans certaines pathologies, en particu-

lier leur dysfonctionnement. L'autre espoir est de découvrir des cytokines pouvant avoir un effet bénéfique dans certaines pathologies.

D'un point de vue plus académique, ce projet a pour ambition de proposer une nouvelle méthode permettant de rechercher des homologues de protéines connues. Cette tâche, parfois difficile à mener, est un sujet fertile dans la littérature et la diversité des approches montre qu'il est loin d'avoir été épuisé. Bien que ce travail soit très clairement orienté vers la recherche de cytokines à quatre hélices α, une volonté de généralisation y sera toujours présente.

En tant que sujet mutidisciplinaire, cette thèse a pour ambition de combiner le savoir du biologiste avec des méthodes de pointes en informatique. Cet esprit d'interdisciplinarité est resté présent dans l'ensemble de ce travail, y compris la rédaction de ce rapport, et jusque dans mes choix méthodologiques. Ainsi une partie de la méthode que je propose se base sur les Séparateurs à Vastes Marges (SVM), une des meilleures techniques de classification en informatique, pour traiter les objets biologiques. Cette partie exploite pleinement cette méthode sans véritablement s'intéresser à la signification biologique des objets qu'elle manipule. Au contraire, la deuxième partie a pour vocation de modéliser une expertise de séquences candidates. L'un des objectifs de cette partie de mon travail était de redonner un sens biologique à la classification des candidats, en exploitant des caractéristiques très spécifiques de la famille de protéines étudiée. Ces deux approches sont finalement combinées afin de faire ressortir les candidats les plus intéressants, qui pourront être minutieusement étudiés. Cette démarche s'inscrit donc pleinement dans l'idée de pluridisciplinarité de la bioinformatique.

Contributions

Ainsi que je l'ai indiqué ci-dessus, un des objectifs de ce sujet est de proposer une nouvelle méthode de recherche d'homologues, utilisant un maximum de connaissances de la famille étudiée. Il s'agit là du principal apport de cette thèse. Plus précisément, ce travail apporte l'idée originale d'ajouter à des classifieurs SVM, possédant une capacité de discrimination propre, des experts basés sur des caractéristiques peu discriminantes mais dont l'ajout apporte une information supplémentaire. Dans cette thèse, je propose une définition conceptuelle de ces experts, ainsi qu'un mécanisme de fonctionnement, basé sur des probabilités bayesiennes. Cette notion a été appliquée aux cytokines à quatre hélices α, mais la définition très générale que j'en donne permet de le réappliquer, éventuellement sous une autre forme, à d'autres familles de protéines.

Le deuxième apport de cette thèse, très pratique cette fois, est la description d'un outil complet de détection d'homologues de cytokines, comprenant l'utilisation, proposée lors d'une thèse précédente, de classifieurs SVM, d'experts basés sur certaines caractéristiques de cette famille ainsi qu'un ensemble de méthodes d'agrégation pour associer ces deux composantes. Bien que l'application ne soit pas encore accessible à un utilisateur non-informaticien, les principaux composants de cette plateforme sont opérationnels grâce au travail de développement effectué pendant les travaux qui ont précédé les miens et auxquels quelques personnes et moi-même avons ajouté les récents développements.

Plan de la thèse

Cette thèse est à l'interface entre la biologie et l'informatique, dans le champ maintenant bien établit de la bioinformatique. J'ai voulu lui donner une approche réellement multi-disciplinaire en l'organisant par thèmes et non en utilisant une structure classique "état de l'art/matériel et méthodes-

/résultats/conclusion". Chaque chapitre contient les éléments de bibliographie et de méthodologie qui seront utiles au thème abordé ainsi que les résultats des méthodes qui y sont mises en oeuvre.

Le chapitre I aborde les cytokines en général et propose une synthèse des connaissances actuelles sur cette famille de protéines. Un bref historique montrera d'abord à quel point il est difficile de définir les cytokines en terme de relations entre ces protéines.
J'essaierai dans un second temps de proposer une définition des cytokines avant de présenter leur mécanisme de fonctionnement moléculaire, les récepteurs qu'elles utilisent pour remplir leur fonction et qui sont essentiels dans la compréhension des cytokines et les pathologies auxquelles elles sont associées, justifiant ainsi l'intérêt qu'on leur porte d'un point de vue médical.
Je discuterai ensuite du problème de la classification des cytokines, en m'appuyant sur les éléments précédemment présentés, avant d'entrer dans le détail de ma famille d'intérêt : les cytokines à quatre hélices α.
Dans cette partie, je présenterai plus particulièrement les caractéristiques de cette sous-famille, qui seront utilisées par la suite dans le chapitre IV.

Le chapitre II s'intéresse aux méthodes automatiques de recherche d'homologues. Après un état de l'art de ces méthodes, je détaillerai plus précisément les Séparateurs à Vastes Marges (SVM), une méthode de classification supervisée donnant d'excellents résultats pour cette tâche. Je décrirai ensuite plusieurs classifieurs spécialement conçus pour la recherche d'homologues avant de les évaluer.

Dans le chapitre III, je décris une autre forme d'outils que j'appelle "experts". Ces experts représentent le jugement d'un annotateur utilisant des critères biologiques pour déterminer si une protéine appartient ou non à une famille déterminée. Cette démarche ajoute une dimension biologique

aux résultats des classifieurs proposés dans le chapitre précédent et permettent d'intégrer des informations supplémentaires, pas nécessairement discriminantes à elles seules. Dans ce chapitre, je commence par définir ce qu'on appelle un expert avant de présenter quelques critères utilisables dans le cas de la recherche d'homologues chez les cytokines. J'expose ensuite concrètement comment concevoir un expert avant d'évaluer quatre d'entre eux, développés à partir des critères présentés auparavant.

Le chapitre IV présente des techniques dites d'agrégation afin de combiner efficacement classifieurs et experts pour améliorer la détection d'homologues. Après un état de l'art sur ce type de méthodes, je présente certaines d'entre elles, qui me paraissent intéressantes pour combiner experts et classifieurs. Ces méthodes sont ensuite évaluées afin de déterminer celle(s) qui présente(nt) les meilleures performances dans la recherche d'homologues de cyokines.

Ce rapport se termine sur un chapitre de conclusion qui récapitule l'ensemble des résultats obtenus et discute des perspectives de ce travail.

Chapitre 1

Cytokines

1.1 Présentation des cytokines

1.1.1 Historique

Pour bien comprendre les cytokines et la façon dont ces molécules ont été regroupées dans la même famille, il convient de rappeler comment elles ont été découvertes. [52]

Si l'ère des cytokines peut être située entre les années 70 et 90, l'aventure des cytokines démarre dès le début du vingtième sciècle. En 1906 Carnot et Deflandre montrent que le serum de lapins anémiques induit, chez des lapins normaux, une forte augmentation de la production de globules rouges. Les auteurs concluent que le serum contient un facteur stimulant cette production. Ce facteur est identifié cinquante ans plus tard comme l'érythropoiétine. En 1957 Isaacs et Lidenmann proposent l'existence d'une activité biologique nommée "interféron", accordant aux cellules une résistance à l'infection virale. Plusieurs autres "activités biologiques" de ce type sont indentifiés dans les années suivantes. La demonstration du fait que les lymphocytes sont à l'origine de ces activité biologiques, donne naissance en 1969 au le terme "lymphokine", pour désigner les molécules à l'origines de ces "activités". Ce terme est remplacé en 1974 par le terme "cytokine" (Cohen *et al*). Plusieurs cytokines ont déjà été découvertes parallèlement dans différentes équipes utilisant des nomenclatures de noms

différentes. Pour remettre de l'ordre dans cette jungle appelations, le terme "interleukine" est proposé par la communauté en 1979. Enfin le terme "chémokine", pour distinguer de petites molécules ayant des propriétés chémoattractantes appartenant à la famille des cytokines, fait son apparition en 1992.

Entre temps de nombreuses cytokines sont découvertes ou identifiées comme telles.

Le tableau 1.1 récapitule les dates de découverte des principales cytokines. Il est à noter que la prééminence de découverte suggérée par les noms ne

nom	date de découverte		nom	date de découverte
EPO	1906		CSF	1990
interféron γ	1957 (Isaac et Lidenmann)		IL-13	1993 (Minty *et al*
IL-5	1970 (Basten et Beeson)		IL-12	1994 (Wolf, Sieburth et Sypek)
IL-1	1972 (Gery et Waksman)		IL-15	1994 (Carson *et al*)
IL-2	1976 (Morgan, Ruscetti et Gallo)		IL-24	1995
IL-6	1980 (Weissman *et al*		leptine	1996
IL-4	1982 (Isakson *et al*)		NNT-1	1999
IL-3	1983 (Ihle *et al*)		IL-21	2000
G-CSF	1985 (Welte *et al*)		IL-22	2000
OSM	1986		IL-23	2000
interféron α	1986		IL-26	2000
interféron β	1986		IL-20	2001
LIF	1987 (Gearing *et al*)		IL-30	2002
IL-7	1988		IL-31	2004
IL-10	1989		IL-32	2004
IL-9	1990		IL-33	2005
IL-11	1990 (Bennet *et al*)		Zsig81	2007 (West et Tannheimer)

TAB. 1.1 – date de découverte des principales cytokines

reflète pas toujours la réalité historique. Cela peut s'expliquer par un écart entre la découverte réelle et sa date de publication officielle, comme c'est le cas pour IL-3 et IL-4, mais cela peut être également dû à un changement de nom postérieur à la découverte, comme par exemple IL-26, initialement baptisée AK155. A ce sujet, la découverte d'IL-32 en 2004 est également intéressante. IL-32 a été identifiée comme cytokine à cette date mais le gène avait été décrit sous le nom "NK4" une douzaine d'années auparavant comme un transcrit existant chez les NK activées par IL-2 (Kim *et*

al, 2005).

.

La première constatation que l'on peut dégager de cet historique est que les cytokines ont été regroupées sur la base d'une activité biologique observée et non de critères d'homologie. Dès lors, les cytokines ne sont pas toutes nécéssairement apparentées sur le plan évolutif.
La seconde est que depuis les années 70, de nombreuses cytokines ont été découvertes et que de nouveaux membres de cette famille sont mis en évidence quasiment chaque année depuis 1990. Ainsi que je l'expliquerai plus loin, les cytokines se fixent à des récepteurs membranaires. Boulay *et al* [6] ont considéré qu'il devait rester deux membres dans la famille des cytokines à quatre hélices α à découvrir. Cette affirmation s'appuyait sur le fait qu'il existait deux récepteurs orphelins associés à cette famille, c'est à dire des fixant aucune cytokine connue. L'idée était que si chaque cytokine possède un et un seul récepteur spécifique et que chaque récepteur ne fixe qu'une cytokine, alors l'existence de deux récepteurs orphelins prédit celle de deux cytokines non découvertes. Comme je l'expliquerai dans la partie sur les récepteurs, le paradigme 1 récepteur = 1 ligand, ne s'applique pas pour les cytokines [44, **?**], il pourrait donc exister des cytokines inconnues se fixant à des récepteurs déjà connu pour fixer d'autres membres de la famille. La découverte de Zsig81, dernière cytokine à quatre hélice α en date [58], complexifie encore le problème. En effet, cette cytokine s'est avérée capable de se fixer non à un récepteur mais à une autre cytokine. Ce phénomène n'avait encore jamais observé dans cette famille et ouvre la voie vers beaucoup de nouvelles possibilités car il est probable qu'il existe d'autres cytokines ayant cette propriété. Toutes ces observations suggèrent que la famille des cytokines pourrait encore s'aggrandir et justifient pleinement des travaux sur cette problématique.

1.1.2 Description des cytokines

Décrire les cytokines est un exercice délicat compte tenu de la diversité de ces protéines et de leurs actions. Il est important de garder à l'esprit qu'il existe toujours des contre-exemples aux généralités que j'énoncerai ici. Les sources principales que j'ai utilisé pour cette description sont : [52], [40]

Définition

La définition la plus générale que l'on puisse en donner est que les cytokines sont des protéines solubles de faible poids moléculaire (15 à 30 KDa), ayant pour principal rôle d'assurer la transmission de signaux de prolifération, différenciation ou de mort cellulaire.

Production, cibles et modes d'action

Les cytokines sont le plus souvent sécrétés par les cellules productrices dans le milieu extracellulaire pour agir sur leurs cibles. Elles sont produites par différentes lignées cellulaires, principalement des agents de la réponse immunitaire et secrétées dans l'ensemble du système périphérique ainsi que dans certains organes tels que le coeur, le foie ou le cerveau. Les cytokines ont de multiples cellules cibles sur lesquelle elles peuvent agir de manière endocrines (action sur des cellules distantes via le système circulatoire), paracrines (action sur les cellules avoisinantes), juxtacrines (action impliquant un contacte entre la cellule sécrétrice et la cellule cible) voire même autocrines. D'une manière générale, on considère que les cytokines pourvoient à la transmission de signaux à courte distance. Cette constatation est renforcée par la faible demi-vie des cytokines dans le milieux extracellulaire, qui ne leur permet pas une action sur le long terme.

Fonctions principales

Les cytokines sont principalement connues pour leurs actions dans le système immunitaire, ce pourquoi elles ont d'abords été étudiées. D'une

manière générale, elles sont impliquées dans la transmission de signaux de prolifération, différenciation ou mort cellulaire. Les cytokines ont rarement une activité catalytique et on les retrouve la plupart du temps dans l'espace inter-cellulaire, bien qu'il en existe des formes membranaires. La principale caractéristique des cytokines est leur pléiotropie, c'est à dire qu'une même cytokine assure plusieurs fonctions différentes. A ceci s'ajoute une certaine redondance entre les cytokines. En effet plusieurs cytokines peuvent avoir la même action sur un type cellulaire. On observe également des relation additives, synergiques ou antagonistes.
Ces caractéristiques rendent la déterminations du ou des rôles d'une cytokine difficile à appréhender. Pour comprendre l'action d'une cytokine, il faut la remettre dans son environnement "cytokinique" [23], cellulaire et temporel. De fait, il est actuellement admit qu'une cytokine n'est plus considérée seule mais dans son contexte à savoir le type de cellules cible, le type de réaction mise en jeu (médiation cellulaire, inflammation, réaction non-immunitaire ...), le stade de cette réaction et surtout les autres cytokines dans l'environnement.

Régulations

La complexité de ce réseau d'interactions, la synergie, l'additivité, l'antagonisme et la redondance des cytokines entre elles et la pléiotropie de chacune implique une régulation très fine de l'action de ces protéines. Les principaux mécanismes de régulation sont au niveau transcriptionnels. La transcription des cytokines est activée ou réprimée par des cascades de signalisations, souvent contrôlées par d'autres cytokines (*cf.* "Effet des cytokines sur la régulation de gènes" 1.2.4). Ce phénomène explique en partie les relations synergiques, additives ou antagonistes qu'ont les cytokines entre elles. On peut également citer d'autres mécanismes de régulation, tels que la faible demi-vie évoquée plus haut de la protéine mature, qui contribue à une action très ciblée.

Résumé

Cette partie a permis de mettre en évidence les principales caractéristiques des cytokines, leur rôle de transmition d'information, leur implication dans la prolifération, différenciation ou mort cellulaire, leur pleiotropisme, redondance et synergisme/antagonisme, expliquant la particularité de cette famille, ses rôles et la complexité d'étude qu'elle représente.

1.2 Mécanisme d'action des cytokines

Cette partie a pour objectif de présenter le schéma d'action des cytokines. Le rôle principal des cytokines étant en lien avec le système immunitaire, c'est dans ce cadre que j'illustrerai leurs mécanismes d'action.

1.2.1 Cellules sécrétrices et cellules cibles

Les cytokines interviennent, entre autre, lors de la différenciation des Th0 (lymphocytes T helper précurseurs) en Th1 (lymphocytes T helper impliqués dans la réponse à médiation cellulaire) ou en Th2 (lymphocytes T helper impliqués dans la réponse à médiation humorale) [24]. La sécrétion de cytokines telles que IFNγ, Il-2, IL-3, IL-4, orientera la différenciation des Th0 en Th1, alors que la sécretion d'autres cytokines comme IL-5, IL-10 ou TGFβ l'orientera vers une différentiation en Th2. Récement de nouveaux type de cellules T ont été mis en évidence : les TH17, Treg, TH1-like et Tr1. La différenciation des TH0 en l'une de ces lymphocytes est schématiquement présentée dans la figure 1.1. L'ensemble de ces cellules T différenciées vont elles-même secréter des cytokines qui leur permettront de différencier les lymphocytes en d'autres types cellulaires spécialisés tels que les NK pour la réponse à médiation cellulaire ou les macrophages pour la réponse à médiation humorale. Le tableau 1.2, bien que loin d'être exhaustif, récapitule quelques fonctions des principales cytokines intervenant dans la réponse immunitaire.

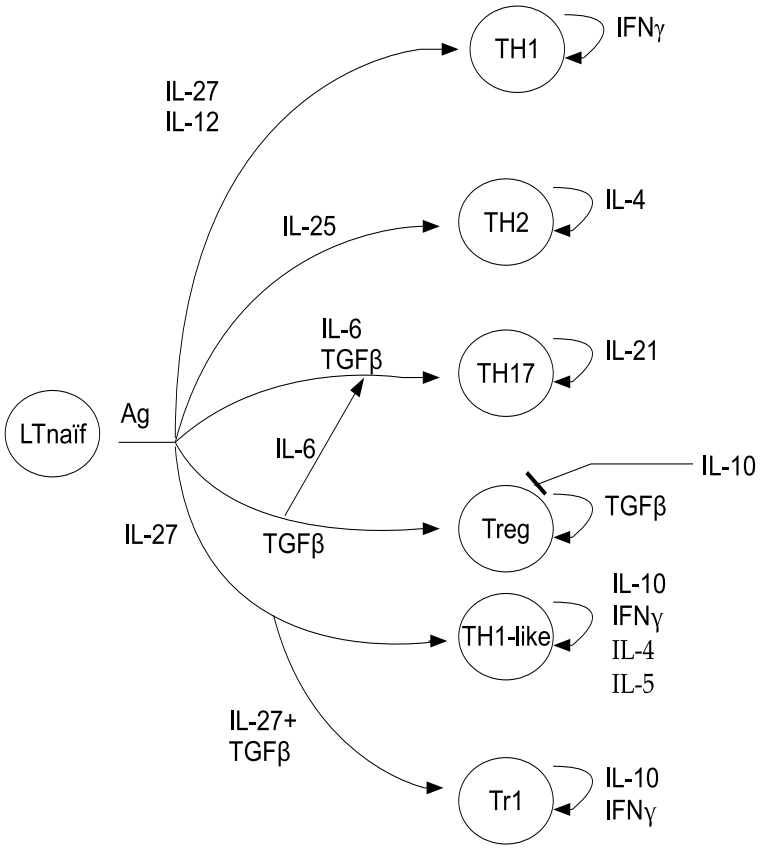

FIG. 1.1 – Différenciation des lymphocytes T précurseurs selon la présence de cytokines dans le milieu

Cytokines	cellules sécrétrices	cellules cibles	fonctions
IL-2	Th1	LT et LB activés, NK	croissance, prolifération, activation
IL-3	Th NK	Stem cell mastocyte	croissance et différenciation croissance et relargage d'histamine
IL-4	Th2 Th2 Th2	LB activé Macrophage LT	prolifération, différenciation expression des molécules du CMH prolifération
IL-5	TH2	LB activé	proliferation, différenciation
IL-6	monocytes macrophages TH2 cellules stromales	LB activés cellules plasmatiques stem cell divers	différenciation en cellules plasmatiques sécrétion d'anticorps différenciation réponse inflammatoire
IL-7	stroma de la moelle osseuse stroma thymique	stem cell stem cell	diférenciation des progéniteurs LB et LT diférenciation des progéniteurs LB et LT
IL-8	macrophage cellules endotheliales	stem cell stem cell	chimiotactisme chimiotactisme
IL-10	Th2 Th2	macrophage LB	inhibition de la production de cytokines activation
IL-12	macrophage macrophage	cellules Tc activées cellules NK	différenciation activation
IFN α	leukocyte	divers	inhibition de la production virale, expression des molécules du CMH
IFN β	fibroblasts	divers	inhibition de la production virale, expression des molécules du CMH
IFNγ	Th1, cellule NK Th1, cellules NK Th1, cellules NK	macrophage TH2 divers	elimination des pathogènes inhibition de la prolifération inhibition de la réplication virale
TGF β	LT, monocytes LT, monocytes LT, monocytes LT, monocytes	monocyte, macrophages macrophage activés LB activées divers	chimiotactisme synthèse IL-1 synthèse IgA inhibition de la prolifération
TNFα	macrophage, mastocytes, NK macrophage, mastocytes, NK	macrophage cellule tumorales	production de cytokines mort cellulaire
TNFβ	Th1 Th1	phagocytes cellule tumorales	phagocytose mort cellulaire

TAB. 1.2 – principales fonction des cytokines dites "immunitaires"

1.2.2 Mode d'action au niveau moléculaire

Le rôle des cytokines est de transmettre un signal de prolifératon/différenciation/mort cellulaire mais, comme je l'ai précisé dans la partie précédente, les cytokines n'ont généralement pas d'activité enzymatique intrinsèque [52]. Cette transmition s'effectue donc par fixation à des récepteurs spécifiques, présents à la surface des cellules cibles. Ce mécanisme de fixation à un récepteur explique plusieurs propriétés des cytokines [23] :

- le mode d'action endocrine, paracrine, autocrine et juxtacrine
- la pléiotropie
- la redondance

Concernant les modes d'actions des cytokines, l'expression de récepteurs à la surface de cellules proches de la cellule sécrétrice, voire sur la cellule elle-même, permet de comprendre les actions paracrines et autocrine des cytokines. Le mode d'action juxtacrine s'opère par le fait qu'une cellule 1 peut fixer la cytokine, sans être activée, pour la présenter à une cellule 2 voisine, cette dernière possèdant des récepteurs dont les capacités de fixation sont moindres. Les cellules 1 et 2 coopèrent ainsi pour activer la cellule 2.

On peut expliquer la pléiotropie par le fait que plusieurs types cellulaires expriment des récepteurs à la même cytokine [45]. La diffèrence de fonction d'une même cytokine sur ces différentes cibles provient de la structure même des récepteurs. Comme il le sera précisé un peu plus bas, les récepteurs des cytokines possèdent deux domaines distincts : un domaine chargé de la fixation de la cytokine et un domaine chargé de la transduction du signal à l'intérieur de la cellule. Cette dernière partie peut activer différentes cascades de signalisation, produisant différents effets à la fixation de la cytokine. Ceci peut également être observé à l'echelle d'une même cellule ayant plusieurs types de récepteurs à une même cytokine, chaque récepteur ayant une affinité différente. Partant de ce principe, la redondance des cytokines peut être expliquée par le fait que plusieurs récepteurs fixant des cytokines différentes activent la même cascade de signalisation.

1.2.3 Les cascades de signalisation

La conséquence principale de la fixation d'une cytokine sur un récepteur est la transduction d'un signal au travers d'une cascade de signalisation. Les cytokines sont ainsi capable d'activer plusieurs voies de transduction, abondamment décrites dans la littérature. La plupart de ces processus de cascades de signalisations débutent soit par la phosphorylation du récepteur lui même lors de la dimérisation induite par la fixation de la cytokine, soit par l'activation des kinases de type JAK, décrites ci-dessous [40].

Les JAKs

Les JAK (Janus Associated Kinase) forment une famille de quatre membres JAK1, JAK2, JAK3 et Tyk2. La partie N-terminal des différentes chaines constituant un récepteur de cytokines se trouve à l'intérieur de la cellule (*cf.* section "récepteurs des cytokines"), reliés aux JAK par l'intermédiaire de motif de type Box (Box1 et Box2). La partie C-terminale des JAK comporte trois domaines : les domaines JH1 et JH2 impliqués dans l'activité tyrosine-kinase et un domaine de type SH2. Les domaines de type SH2 sont présents dans de nombreuses protéines impliquées dans des cascades de signalisation et possèdent la capacité de lier des tyrosines phosphorylées.
lors de la fixation d'une cytokine par un récepteur, les différentes chaines réceptrices se rapprochent, permettant la dimérisation des JAK qui y sont attachées. Ces dernières s'activent par trans-phosphorylation. Ceci permet un changement de conformation chez les JAK les rendant capable de phosphoryler les tyrosines d'autres protéines chargées de transmettre à leur tour le signal. Cette transduction peut passer par l'intermédiaire de plusieurs voies de signalisation, les deux principales étant la voie STAT et la voie des MAPkinases.

La voie STAT

La famille STAT (Signal Transducers and Activators of Transcription) comprend 7 facteurs de transcription (STAT1, STAT2, STAT3, STAT4, STAT5a, STAT5b et STAT6) impliqués dans la régulation de différents gènes. A l'état déphosphorylés et monomériques, les STAT sont localisés à proximité de la membrane, près des chaines intracellulaires des récepteurs des cytokines. Les KAK activent STAT par phosphorylation de leurs tyrosines. Une fois phosphorylés, les STAT se dimérisent *via* leurs domaines SH2 pour former des homodimères ou des hétérodimères. Cette étape de phosphorylation et dimérisation leur permet également d'être adressées au noyau.

La voie MAP Kinase

La voie MAP (Mitogen Activated Proteins) Kinase est une des voies de signalisation cellulaire les mieux connues et les plus étudiées. Elle permet de conduire un signal de régulation génétique depuis la membrane cellulaire jusqu'au noyau par des séries de phosphorylations/déphosphorylations successives, activé par des kinases particulières. Ce mécanisme ne sera pas détaillé ici. Le terme "voie des MAP Kinase" désigne en réalité un ensemble de voies de signalisation basées sur le même principe mais dont les acteurs sont plus ou moins spécifiques à chaque voie. D'une manière générale, il y a quatres niveaux dans une voie MAP Kinase. Le premier niveau est formé par les MKKK (MAP Kinase Kinase Kinase) qui sont des protéines proches de la partie intracellulaire de la membrane et qui sont phosphorylées par des kinases couplées aux récepteurs. Cette phosphorylation passe par l'intermédiaire de petites protéines G. Les MKK (MAP Kinase Kinase), phosphorylées par les MKKK constituent le deuxième niveau. Le troisième niveau est représenté par les MK (MAP Kinase), phosphorylées par les MKK . Les MK vont à leur tour phosphoryler les facteurs de transcriptions, qui constituent le dernier niveau. Des protéines dites "scaffold" (échaffaudage) sont suseptibles d'intervenir entre ces niveaux pour faciliter

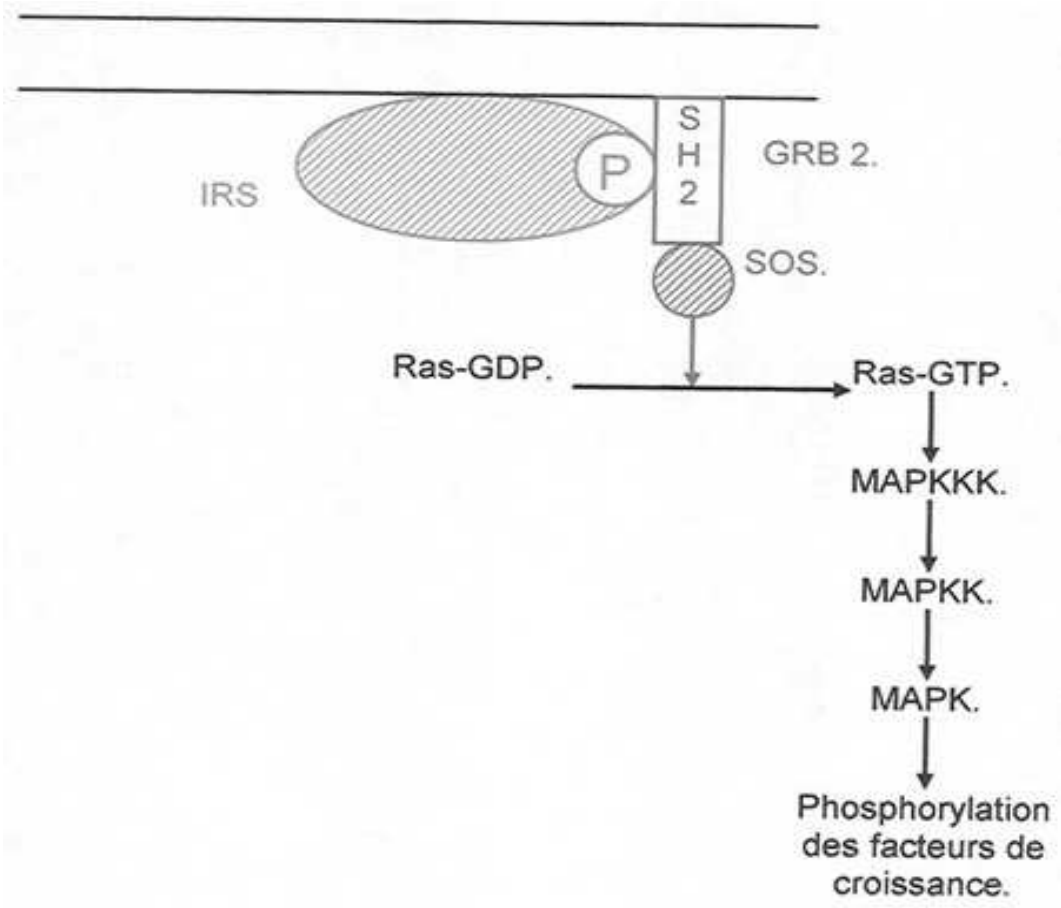

FIG. 1.2 – Exemple de cascade de MAP Kinases

le passage du signal. Sous forme phosphorylés, les facteurs de transcription sont transloqués dans le noyau pour réguler leur(s) gène(s) cible(s). La complexité de cette architecture de transmission permet, entre autre, une régulation très fine de ces cascades. Dans le cas d'un signal induit par des cytokines, Les MKKK potentielles seraient des protéines des familles MEKK et MLK, induites par des GTPase de la famille Rho, les MKK sont principalement MKK3, MKK6, MKK4 et MKK7. Les deux principales MK sont p38 et JNK qui phosphorylent des facteurs de transcription tels que ATF2, Alk-1, MEF-2C et c-Jun, appartenant à la famille de AP-1. la figure 1.2 illustre une cascade de MAP kinases.

Autres voies de signalisations

D'autres voies de signalisation peuvent être utilisé par les cytokines. C'est entre autre le cas de la voie PI 3-kinase (PhosphoInositide 3-kinase), de la voie PLC-γ (Phospholipase C γ) et de la voie HCP (Hematopolietic Cell Phosphatase). Ces voies fonctionnent souvent sur le même principe que les voies décritent ci-dessous à savoir phosphorylation par les JAK, ou par le récepteur lui même afin d'enclancher une cascade de phosphorylation/déphosphorylation visant à transloquer des facteurs de transcription dans le noyau.

1.2.4 Effets des cytokines sur la régulation de gènes

Comme on vient de le voir, les cytokines permettent la translocation de facteur de transcription dans le noyau, ces derniers agissant directement sur les gènes cibles.
Les STATs possèdent dans leur région N-terminales un domaine bZip (leucine-zipper) de fixation à l'ADN spécifique à certains promoteurs comportant des séquences de type SBE (STAT Binding Element). Il existe plusieurs exemples de SBE, les GAS (Gamma-Activated Sequences) et les ISRE (Interferon-Stimulated Response Elements) en étant les plus connues. Il faut noter que la phosporylation des STAT induit leur homo- ou hétéro-dimérisation. La combinaison de monomères de STAT détermine quels promoteurs et donc quels gènes vont être affectés. En C-terminal, les dimères STAT possèdent un domaine d'activation de la transcription, qui varie pour chaque STAT. Cette diversité, tant du point de vue de la reconnaissance des sites de fixation du facteur de transcription que des domaines d'activation de transcription permet aux STAT une grande versatilité d'actions et de cibles.

c-JUN ou AFT2, induits par la voie des MAP kinases, possèdent également des site de fixation à l'ADN de type bZip ainsi que des sites d'activation de la transcription. [54] les AP-1 ont tendance à se dimériser pour se fixer

à des séquences promotrices consensus $TGAGTCA$
Des homodimères ou des hétérodimères peuvent être formés, ces combinaisons fixant des variations de la séquence consensus avec des affinités différentes. Le type de régulation (activation ou répression) dépend également de la combinaison de dimères et de l'environnement génétique de la séquence fixée par ce dimère, permettant une grande finesse dans la régulation des gènes cibles. Par ailleur les facteurs de transcription eux même sont régulés par leur propre action.

1.2.5 Résumé

Dans cette partie, j'ai détaillé les mécanismes d'action moléculaire des cytokines. Leur action passe par la fixation à un récepteur membranaire qui déclanche différentes cascades de signalisation, menant à la régulation transcriptionnelle d'un grand nombre de gènes de la cellule cible. Ce mécanisme permet d'expliquer les caractéristiques de pléiotropie, redondance, additivité, synergie ou antagonisme des cytokines.
L'interaction de la cytokine avec son récepteur est l'élément déclancheur de la cascade. Le récepteur assure également la spécificité d'action. L'importance de ces récepteurs dans le mode d'action des cytokines sera détaillé dans la partie suivante.

1.3 Récepteurs des cytokines

Cette section présente les récepteurs des cytokines. Je commencerai par décrire les récepteurs d'un point de vue général avant d'entrer dans les détails de leur classification et de leur mode de fonctionnement.

1.3.1 Description générale

Ces dernières années, les récepteurs de cytokines ont été l'objet d'un intéret croissant, du fait de leur importance dans le déclanchement des casades de signalisation médiées par les cytokines. Si les cytokines présentent

peu de similarité de séquences entre elles, ce n'est pas le cas de leurs récepteurs. Ces derniers peuvent être divisés en plusieurs sous-familles, extrêmement homogènes. Les différentes sous-familles de récepteurs seront détaillées plus loin mais on peut d'ore et déjà dégager quelques caractéristiques communes.

Organisation des domaines

Les récepteurs à cytokines sont des protéines membranaires constituées de plusieurs modules. Le cas le plus général est une partie extracellulaire assignée à la liaison avec le ligand et une partie intracellulaire ayant pour rôle la transmission du signal. La partie extracellulaire est elle-même constituée de plusieurs domaines. Ainsi, par exemple TNFR (Tumor Necrosis Factor Receptor) est constitué de trois régions riches en cystéines, impliquées dans l'interaction avec le ligand et de trois domaines membranaires de type FNIII (fibronectin type III) [40]. Cette structure est retrouvée dans plusieurs récepteurs de cytokines. La partie intracellulaire est également constituée de plusieurs domaines. On trouve des domaines d'interaction protéine-protéine et, souvent, des domaines ayant une activité enzymatique, majoritairement kinase/phosphatase. Ces multiples domaines permettent d'activer différentes voies de signalisations décrites précédemment.

Mode d'action

Inactifs en l'absence de cytokines, les récepteurs se dimérisent, voire trimérisent, lors de la liaison au ligand. Le plus souvent, on observe cette oligomérisation au moment de la fixation du ligand mais dans certains cas comme celui de l'EPOR, le récepteur peut être dimérisé en l'absence de ligand. Dans ce cas la fixation de ligand induit un changement de comformation dans le dimère qui active le récepteur et permet la propagation du signal.

Cette oligomérisation s'explique par le mode de fixation du ligand sur le récepteur. En effet, trois épitopes de contact sur le ligand, nommés site I, site II et site III, sont requis pour fixer les chaînes réceptrices. Il a été démontré que ces épitopes ne changent presque pas d'une cytokine à l'autre [44], si le récepteur est capable de fixer plusieurs cytokines. Ceci implique que c'est la topologie du récepteur après fixation qui est le mécanisme d'activation de ce dernier et donc que, quel que soit le ligand, des topologies similaires doivent être obtenues.
Un récepteur de cytokine peut être un homodimère ou un hétérodimère de chaînes réceptrices. Une même chaine réceptrice peut donc être associée à plusieurs chaines effectrices différentes pour constituer différents récepteurs, n'ayant pas nécessairement le même ligand et/ou n'activant pas les mêmes cascades de signalisation. En effet, dans un hétérodimère, les deux chaines peuvent avoir des fonctions différentes, l'une ayant pour rôle d'augmenter l'affinité du récepteur pour un ligand particulier (jouant ainsi sur la spécificité du récepteur) et l'autre étant reliée à une cascade de signalisation précise, influant donc sur la fonction de la cytokine fixée. Par exemple, gp130 peut être associé à IL-6Rα pour former un récepteur à l'IL-6 ou s'associer avec IL-11Rα pour former un rcépteur à l'IL-11. IL-6Rα et IL-11Rα sont les chaines spécifiques à IL-6 et IL-11 respectivement alors que gp130 est une chaine liée à des cascades de signalisation intracellulaire. Les différentes associations des récepteurs dit de type I sont présentées das le tableau 1.3 [45] .

On constate qu'un même récepteur peut fixer plusieurs cytokines (par exemple IL-4Rα+IL-13Rα1 peut fixer IL-4 et IL-13) et qu'une cytokine peut se fixer à plusieurs récepteurs (par exemple IL-4 peut se lier à IL-4Rα+IL-13Rα1 et à γc+IL-4Rα. Ceci permet d'expliquer la pleiotropie et la redondance des cytokines. La redondance provient du fait que plusieurs cytokines peuvent se lier à un même récepteur, déclanchant ainsi la même cascade de signalisation. Cela est également possible si deux cytokines utilisent la même sous-unité transductrice du signal. La pléiotropie

Sous-unitée partagée	autres composants	ligand
βc	IL-3Rα	IL-3
	GM-CSFRα	GM-CSF
	IL-5Rα	IL-5
gp130	IL-6Rα	IL-6
	IL-11Rα	IL-11
	IL-27R	IL-27
	CNTFRα et LIFRβ	CNTF, NNT-1, BSF-3, CLC/CLF
	LIFRβ	LIF, OSM, CT1
	OSMR	OSM
IL12-Rβ1	IL12-Rβ2	IL-12
	IL-23R	IL-23
γc	IL-2Rβ	IL-2 (affinité intermédiaire), IL-15
γc	IL-2Rβ et IL-2Rα	IL-2 (affinité forte)
	IL-4Rα	IL-4
	IL-7Rα	IL-7
	IL-9Rα	IL-9
	IL-15Rα et IL2-Rβ	IL-15
	IL-21Rα	IL-21
IL-4Rα	IL-13Rα1	IL-4 et IL-13
TSLPR	IL-7Rα	TSLP
OSMR	GPL	IL-31

TAB. 1.3 – Composition des principaux récepteurs des cytokines dites de type I

vient du fait qu'une cytokine se fixe à plusieurs récepteurs. Ces récepteurs utilisent la même sous-unité pour fixer la cytokine mais diffèrent par leurs sous-unités transductrices. Une cytokine peut ainsi activer plusieurs voies de signalisation différentes.

Après avoir présenté les caractéristiques générales des récepteurs de cytokines, je vais présenter plus précisément les différentes famille de récepteurs.

1.3.2 Les familles de récepteurs

Les récepteurs de cytokines se divisent en cinq grandes familles : les récepteurs de type I, de type II, de type III, les récepteurs apparentés à la superfamille des immunoglobulines et les récepteurs à chémokines. Je vais décrire ces différentes familles en me focalisant sur les récepteurs de type I et de type II, qui concernent plus spécifiquement les cytokines à quatre hélices α, étudiées de cette thèse.

récepteurs de type I

Les récepteurs de type I ont pour principaux ligands IL-2, IL-3, IL-4, IL-5, IL-6, IL-7, IL-9, IL-11, IL-12, EPO, GM-CSF, G-CSF, LIF, CNTF et TPO [21]. Ils se composent de plusieurs domaines, intra et extra-cellulaires. Leur domaine extracellulaire, d'environ 200 acides aminés, se caractérise par la présence d'un motif ($WSXWS$) en C-terminal (proche du domaine transmembranaire) et quatre cystéines conservées en N-terminal (noté C4). Le motif ($WSXWS$) s'est avéré indispensable pour la fixation du ligand, quant aux quatre cystéines, elles interviennent dans le repliment du domaine. Du fait de leur importante similarité de séquences, les récepteurs de type I ont été principalement détecté en utilisant des techniques basées sur le BLAST. 34 récepteurs de type I ont été ainsi identifiés [6] et 25 d'entre eux possèdent exactement les motifs C4 et ($WSXWS$). Un autre point commun est la structure génomique de ces récepteurs. Leur domaine extracellulaire est codé par quatre exons et le domaine intracellulaire par

deux : l'un codant pour le site d'attachement à JAK et l'autre codant le reste du domaine. Les récepteurs de type I n'ont pas d'activité kinase instrinsèque mais sont capable de lier des protéines de type JAK/STAT pour déclancher une cascade de phosphorylation. On constate que l'utilisation des JAK/STAT n'est pas aléatoire. Boulay *et al* ont reconstruit la phylogénie des récepteurs de type I et l'ont comparée à celle des cytokines de type I. On y distingue clairement cinq sous-groupes, dont trois fixent les cytokines à chaines longues et deux fixent les cytokines à chaines courtes (*cf.* 1.5.5), or chaque sous-groupe a une utilisation des JAK/STAT différente et quasi homogène à l'intérieur du groupe.
Cette famille de récepteurs, tant par le nombre de cytokines qui en dépendent, que par ces caractéristiques très homogènes, est *de facto* la famille la plus étudiée.

récepteurs de type II

Moins bien étudiés que les récepteurs de type I, les récepteurs de type II n'en sont pas moins la deuxième plus importante famille de récepteurs de cytokines. Leurs ligands principaux sont IL-10, IL-19, IL-20, IL-22, IL-24, IFN-α, IFN-β et IFN-γ. Apparentés aux récepteurs de type I, ils forment un groupe relativement homogène et terme de séquence et de structure. Les récepteurs de type II possèdent dans leur domaine extracellulaire un ou deux tandems de domaines fibropectine III, probablement issus de duplications. Ils sont également caractérisés par deux doublets de cystéines en C-terminal des domaines fibropectine III. Les récepteurs de type II peuvent être également séparés selon leur utilisation des différentes voie de signalisation [48].

récepteurs de type III

Les récepteurs de type III sont un petit groupe conportant les deux récepteurs au TNF ainsi que le récepteur au NGF et certains antigènes de surface. Cette famille, peu étudiée dans la littérature, a été distinguée

par la présences de quatre domaines extracellulaire riches en cystéines, où la position de ces dernières semble extrémement bien conservée dans la famille [40].

récepteurs apparentés à la superfamille des immunoglobulines

Ces récepteurs possèdent trois à cinq domaines immunoglobine-like dans leur partie extracellulaire. Ces domaines sont utilisés pour fixer le ligant. Les principaux ligands de ces récepteurs de cette famille sont IL-1 ainsi que des facteurs de croissance tels que M-CSF et SCF. Dans leur partie intracellulaire ces récepteurs peuvent posséder ou non une activité tyrosine kinase. L'absence ou la présence de cette activité détermine les voies de signalisation que le récepteur peut activer [40].

récepteurs au chemokines

Les chemokines forment un groupe à part des cytokines, découvert récement. Leurs récepteurs sont extrémement différents des autres récepteurs de cytokines et rentre plutôt dans la catégorie des protéines à sept domaines transmembranaires. L'activation de la cascade cellulaire se fait par un couplage du récepteur à des protéines G assurant une transduction β-adrénergique du signal [40].

1.4 Cytokines et pathologies

Du fait de leur rôle dans le système immunitaire, les cytokines sont impliquées, par leurs actions ou par leur dérèglement, dans de nombreuses pathologies. J'aborderai quatre grandes pathologies : l'inflammation chronique, les maladies auto-immunes en général, le cancer et l'infection par le VIH [56].

1.4.1 L'inflammation

L'inflammation est une réaction de défense immunitaire stéréotypée du corps à une agression : infection, brûlure, allergie ...Elle n'est pas spécifique au type d'agression mais représente plutôt une réponse primaire en attendant que le système immunitaire identifie le type d'agression et mette en place une réponse plus adapatée. Les cytokines sont connues pour leurs rôles dans la régulation de ce processus, il s'agit d'ailleurs d'une des premières fonctions pour lesquelles elles ont été étudiées. Traditionnellement, on répartit les cytokines en pro-inflamatoires et anti-inflamatoires selon qu'elles activent ou inhibent l'inflamation. Le tableau 1.4 récapitule les principales cytokines impliquées dans l'inflammation.

Le rôle inflammatoire ou anti-inflammatoire de chaque cytokine reste tou-

Cytokines proinflammatoires	cytokines anti-inflammatoires
IL-6	IL-10
IL-1	IL-4
IL-8	IL-11
IL-12	IL-13
IL-15	TGFβ
TNFα	IFNγ

TAB. 1.4 – quelques cytokines pro- et anti-inflammatoires

tefois dépendant du contexte, le tableau 1.4 ne fait que donner leur tendance inflammatoire/anti-inflammatoire.

Chaque cytokine agit sur l'inflammation d'une façon différente et à différents stades, mais d'une manière générale, les cytokines pro-inflammatoires stimulent le recrutement, la différentiation en effecteurs de l'inflammation tels que les monocytes, macrophages, monocytes, lymphocytes et leur prolifération alors que les cytokines anti-inflammatoires s'y opposent et activent le processus de mort cellulaire de ces effecteurs. Le fonctionnement général de l'inflammation, du point de vue des cytokines, est le suivant : lors des premières étapes de l'inflammation, quelque cytokines pro-inflammatoires sont sécrétées (TNFα, IL-1, IFNγ). Elles stimulent les effecteurs de l'inflammation mais également la synthèse d'autre cytokines

pro-inflammatoires, dont la leur. Avec la progression de l'inflammation, le contexte cytokinique change, amenant la synthèse de cytokines anti-inflammatoires "primaire" (IL-10, IL-4) qui ont pour rôle de stopper l'inflammation en inhibant la prolifération des effecteurs et en déclanchant leur apoptose. Elles agissent également sur les cytokines pro-inflammatoires, qu'elle inhibent, et stimulent la synthèse de cytokines anti-inflammatoires, dont leur propre auto-activation.

Je viens de décrire le processus "normal" de l'inflammation, mais un déréglement de ce phénomène peut donner lieu à diffrentes pathologies, particulièrement dans le cadre d'inflammations chroniques, qui créent un contexte favorable à l'apparition de maladies auto-immunes (cf ci-dessous). Les inflammations chroniques résultent le plus souvent d'un dérèglement de la régulation des cytokines, du fait d'une sur-stimulation par les cytokines inflammatoires, ou d'une sous-inhibition par les cytokines anti-inflammatoires. Ces inflammations chroniques peuvent être également dûe à un facteur pathogène que le système immunitaire ne parvient pas à éliminer complétement et qui stimule en continu l'inflammation.
Plus rarement, la réaction inflammatoire peut être défectueuse du fait d'une sous-stimulation par cytokines inflammatrices, ou d'une sur-inhibition par les cytokines anti-inflammatoires.

1.4.2 Les maladie auto-immunes

Le dérèglement d'une ou plusieurs cytokines peut facilement mener à des maladies auto-immunes, c'est à dire des maladies où le système immunitaire réagit contre les cellules du soi. D'une manière générale, les cytokines qui activent la différenciation des lymphocytes dans la voie Th1 sont plutôt pro-inflammatoires et génératrices de maladies auto-immunes alors que les cytokines impliquées dans la différenciation des lymphocytes dans la voie Th2 sont plutôt anti-inflammatoires, particulièrement si elles ont une action inhibitrice de la différenciation vers la voie Th1, et s'opposent à l'appa-

rition de maladies auto-immunes [24]. TNFβ, IL-2 et IFN$\alpha/\beta/\gamma$ semblent être les cytokines les plus impliquées dans l'émergence de mécanismes d'autoimmunité. Cela s'explique par le fait que ces cytokines sont parmi les premières dans la cascade pro-inflammatoire dont elles contrôlent le déclanchement. Toutefois, ce paradigme est à considérer avec prudence car les cytokines pro-inflammatoires peuvent également avoir une action immuno-supprésives (*i.e. activatrices des cellules T régulatrices*), principalement lorsqu'elle agissent pendant une longue durée sur leurs cellules cibles.
L'IL-2 est considérée comme une pièce maitresse de la régulation des cytokines pro-inflammatoires et donc de l'apparition de maladies auto-immunes. Une déficience d'IL-2 est souvent notée dans le cas de patients atteints de l'une de ces maladies. Cela peut s'expliquer par le fait qu'IL-2 peut déclancher l'apoptose de lymphocytes par l'activation de la transcription de plusieurs gènes impliqués dans ce mécanisme (BCL-2, Fas ...) et par la répression de la transcription de protéines s'opposant à l'apoptose tel que la protéine FLIP.

Ainsi qu'il l'a été dit plus haut, les cytokines peuvent avoir des action opposées. C'est également vrai pour leur propention à favoriser ou non l'apparition de réactions auto-immunes. Pour illustrer ce phénomène, voici deux exemples de cytokines ayant des activités plutôt enclintes à favoriser ce genre de pathologies mais possèdant également une activité immuno-suppréssive.
Le TNF possède des propriétés pro-inflammatoires qui en fait une cible thérapeutique pour le traitement des maladies auto-immunes. L'inhibition du TNF a été utilisée de manière probante pour traiter la maladie de Crohn ou l'artrite rhumatoïde. Toutefois, le TNF a des effets immuno-suppresseurs multiples comme l'inhibition de la signalisation par le TCR, responsable de la reconnaissance des complexes CMH/peptide chez les cellules T, le déclanchement de l'apoptose chez ces mêmes cellules T ou encore l'induc-

tion d'autre cytokines (IL-10 et TGFβ) promouvant la différenciation des lymphocytes dans la voie Th2.
Les IFNα/β ont également cette dualité pro-inflammation/immuno-supprésion. D'une part les IFNs sont impliqués dans la différenciation des lymphocytes dans la voie Th1. D'autre part, les IFNs activent la voie Fas, impliquée dans l'apoptose. Ils activent également des cytokines immuno-suppresives telle que l'IL-10.

Ce mécanisme de fonctionnement général des maladies auto-immunes, en plus d'expliquer le mode d'apparition de ces pathologies, est une bonne illustration de la dualité des fonctions des cytokines.
Je présente par la suite deux pathologies pour lesquelles une origine immunitaire est fortement suspectée, afin d'illustrer le fonctionnement des maladies auto-immunes.

La sclérose en plaque

Cette maladie auto-immune est dûe à la démylinisation dans la substance blanche de l'encéphale et de la moelle, entrainant une dimminution de l'influx nerveux. Aux stades les plus graves de la maladie, le patient peut être atteint de handicaps moteurs dus à une rupture de cet influx. Cette démylinisation est provoquée par une sur-inflammation dans le système nerveux. La sclérose en plaque est une maladie dont la prévalance (proportion de la population touchée) et l'incidence(proportion de nouveaux cas par an) est supérieure dans les zones tempérées par rapport aux zones tropicales. La sclérose en plaque est également corrélée à des facteurs génétiques. Des études montrent que la probabilité de contracter la sclérose en plaque est plus importante dans une famille où des cas de cette maladie ont été observés. Elles montrent que certains allèles du HLA sont plus fréquemment présents chez les malades. De même certains allèles de IL-7RA et IL-2RA (codant respectivement pour une sous-unité des récepteurs à l'IL-7 et à l'IL-2) sont fréquents chez les patients atteints de sclérose en plaque. Les

causes de cette maladie ne sont pas encore clairement élucidées. On suppose qu'elle est déclanchée de manière enssentiellement auto-immune. La cause pourrait en être une sur-stimulation de l'inflammation ou une faiblesse de la reconnaissance des cellules du soi par le système immunitaire. Certain auteurs suggèrent également un embalement de l'inflammation lors d'une réponse immunitaire normale. L'augmentation de l'occurence de la maladie au fur et à mesure que l'on s'éloigne de l'équateur suggère que l'exposition au Soleil a un rôle protecteur contre la maladie. Il a été proposé que la diminution d'exposition au Soleil pouvais réduire la production de vitamine D, qui joue un rôle de protection sur la myeline.
La sclérose en plaque est donc probablement une maladie multifactorielle, d'origine inflammatoire. Le rôle des cytokines dans l'inflammation en fait des cibles thérapeutiques de première importance. Des cytokines anti-inflammatoires, particulièrement l'interféron β sont utilisés comme traitement, avec une certaine efficacité.

Le psoriasis

Le psoriasis est une affection dermatologique non contagieuse qui se caractérise par l'apparition de plaques à la surface de la peau. Selon la gravité de la pathalogie, ces plaques peuvent être localisées (cuir chevelu, visage, ongles, articulations, parties génitales) ou couvrir une importante partie du corps. Le psoriasis peut se manifester chez les hommes aussi bien que chez les femmes et à tout age à partir de 15 ans. Sa prévalance est d'environ 3% dans les pays occidentaux. D'un point de vue cellulaire, Le psoriasis se présente comme un renouvellement trop rapide des cellules de la peau. On observe dans les zones où se manifeste le psoriasis, une hyperplasie c'est à dire l'absence de cellules granulaires, une différenciation incomplète des keratinocytes, et une accumulation de cellules T [20]. Il existe deux types de keratinocyte, les $CD29^{+}K1/K10^{-}$, qui ont un cycle cellulaire lent et donc un taux de réplication faible et les $CD29^{+}K1/K10^{+}$ qui ont un taux de réplication rapide. Les $CD29^{+}K1/K10^{-}$ sont majoritaires

dans les couches supérieures de l'épiderme alors que les $CD29^+K1/K10^+$ se retrouvent plutôt dans les couches inférieures et servent de "cellules de transferts" Dans le cas du psoriasis, on observe que les kératinocytes de type $CD29^+K1/K10^-$ ont un cycle de prolifération plus rapide que la normale.

L'origine du psoriasis est encore mal connue. Les deux principales hypothèses avancées sont un dérèglement naturel de la prolifération cellulaire dans le derme et l'épiderme, ou une sur-activation du système immunitaire qui provoque une inflammation dans les zones touchées, cette inflammation augmentant la production de cellule de la peau, comme dans le processus normal d'inflammation. Le psoriasis est une maladie multi-factorielle avec des composantes génétiques, environnementales (stress, consomation d'alcool ou de tabac, exposition au Soleil...)

Une forte concentration d'IL-2 et d'interferon γ a été observée dans les zones touchées par le psoriasis. Ajouté au fait que certaines cytokines anti-inflammatoires sont capable de stimuler les kératinocytes $CD29^+K1/K10^-$ pour accelérer leur cycle cellulaire, cette observation privilégie nettement la piste inflammatoire, et une implication des cytokines dans cette pathologie. Les rôles de plusieurs cytokines, entre autre le TNF α, IL-22 et l'OSM, sont particulièrement examinés. Un traitement à base d'inhibiteur du TNF α est classiquement proposé aux patients atteints de psoriasis.

1.4.3 Le Cancer

Du fait même des fonctions prolifératives et/ou apoptotiques des cytokines, il est évident qu'elles aient un lien avec différentes formes de cancer. Actuellement, on considère qu'une inflammation continue et/ou sur-activée puisse créer un micro-environnement favorable au développement de cellules malignes [36]. De plus, nombres de cancers sont associés à une infection virale, les cytokines intervenant dans la réponse immunitaire, on peut voir ainsi une deuxième forme d'interaction possible. J'expliciterai principalement le lien entre inflamation et cancer, qui semble le mieux étudié

et le plus fréquent. En effet, l'infection virale peut induire des tumeurs de différentes façons mais il semble que dans de nombreux cas, c'est l'inflammation chronique déclanchée par une infection mal éliminée qui permet l'apparition de tumeurs, ramenant au cas d'un cancer d'origine inflammatoire. Je parlerai également de "cancer" en général sans entrer dans les différentes formes de cancers qui existent, avec leurs spécificités.

Distinctions entre cytokines pro- et anti-tumorales

Le rôle des cytokines en général et de chaque cytokine en particulier sur le développement de tumeurs peut être variable, à cause de leur pléiotropie et de leur synergie/additivité/antagonisme. On peut grossièrement considérer que les cytokines pro-inflammatoires sont plutôt favorables à l'initiation et au développement de tumeurs alors que les cytokines anti-inflammatoires sont plutôt inhibitrices de ces phénomènes. L'un des éléments clé du dévellopement tumorale est le contrôle de la voie de signalisation NF-κB. Cette voie stimule des gènes ayant des effets pro-tumoraux ou inhibe des gènes anti-tumoraux :

- activation/inhibition de gènes anti-apoptotiques/pro-apoptotiques,
- activation du cycle cellulaire, favorisant la progression de la tumeur,
- augmentation de la transition épithélio-mesenchymal, qui favorise l'invasion de la tumeur,
- maintient de l'inflammation qui entretient le micro-environnement, favorable au développement de nouvelles tumeurs.

Les cytokines ayant tendance à activer la voie NF-κB seront donc plutôt pro-tumorales alors que celles reprimant cette voie seront plutôt anti-tumorales. Cette partie n'ayant pas la prétention de faire une étude exhaustive de l'implication des cytokines dans les cancers, je présenterai les principales cytokines connues pour avoir un rôle direct dans le cancer, mais il convient de ne pas oublier que les cytokines forment un réseau complexe et que leurs implications sont sûrement plus subtiles qu'il n'y parait.

Les cytokines pro-tumorales

Les cytokines pro-tumorales sont principalement IL-1, TNFα, IL-6 et IL-17. La présence de hautes concentrations de ces cytokines dans l'environnement tumoral ainsi que la régression des tumeurs lors de la neutralisation de ces cytokines dans des modèles murins démontrent clairement leur action pro-tumorale.

TNFα joue un rôle dans l'apparition de tumeurs en activant directement la voie NF-κB, avec comme cibles principales des molécules anti-appototiques, et en stimulant la production de molécules génotoxiques, pouvant endommager l'ADN et causer des mutations favorables à l'acquisition de la malignité. TNFα est également capable de promouvoir la progression tumorale à cause de son rôle suppresseur sur la cytotoxicité des macrophages et la réponse médié par les cellules T. Le rôle pro-tumorale de TNFα a été entre autre démontré dans des cancers de la peau, du foie et du colon.

IL-6 active principalement deux voies de signalisation dans la cellule : les voies STAT1 et STAT3. STAT3 a des effets prolifératifs sur les tumeurs alors que STAT1 est plutôt décrite comme ayant des effets inhibiteurs sur leur croissance, mais, curieusement, il semble qu'IL-6 soit très majoritairement pro-tumorale. En effet, il a été observé que la plupart des gènes activés par l'action d'IL-6 sont des gènes favorisant la progression du cycle cellulaire et suppresseur de l'apoptose. Contrairement à TNFα, le rôle d'IL-6 semble se limiter à la prolifération des cellules malignes. Une suractivation d'IL-6 est associée à de nombreux cancers tels que le cancer du sein, du colon ou le lymphome d'Hodgkin. Il n'est toutefois pas exclu qu'IL-6 puisse jouer, dans certaines conditions, un rôle anti-tumoral via son activation de STAT1

IL-17 est également une cytokine plutôt pro-tumorale mais ce rôle semble moins direct que dans le cas de TNFα et IL-6. Impliquée dans l'activation de la voie NF-κ B, IL-17 est plus étudiée pour son action pro-inflammatoire, au début de cette réaction. IL-17 a entre autre un rôle important dans l'induction d'autres facteurs inflammatoirs tels que TNFα et IL-6. Son rôle

pro-tumoral est moins bien tranché car IL-17 c'est également montrée capable d'une activité anti-tumorale dans certains modèle murins, mais ce rôle reste encore à démontrer.

Les cytokines anti-tumorales

A contrario, des facteurs cités précédemment, certaines cytokines ont une activité plutôt anti-tumorales. Il s'agit principalement d'IL-10, IL-12, IFNγ et TGFβ. Dans la même idée que pour les cytokines pro-tummorales, l'apparition accrue de tumeurs après inhibition de ces cytokines ou leur utilisation avec succès dans des protocoles anti-tumoraux confirment cette classification.

Bien que pro-inflammatoire, IL-12 se présente comme une cytokine clairement anti-tummorale dans plusieurs modèles de cancers chez la souris. IL-12 inhibe la cancerogénèse et induit une régression tumorale via son action de contrôle sur l'immunité de type Th1.

L'IFNγ, par la stimulation de réactions toxiques contre les cellules tumorales et son activité anti-angiogénique, est également un cytokine anti-tumorale importante.

IL-10 étant un puissant agent anti-inflammatoire, son action anti-tummorale n'a rien de surprenant. L'action d'IL-10 s'explique par son rôle d'inhibiteur directe de la voie NF-κB et de la production de cytokines pro-tumorales. Ainsi IL-10 agit sur la prolifération de cellules tumorales et leur progression. IL-10 pourrait également activer des processus immunitaires anti-tumoraux. Toutefois IL-10 pourrait également être capable de promouvoir la prolifération et la survie des cellules tumorales via son rôle activateur sur la voie STAT3. Ainsi, bien que considérée comme majoritairement anti-tumorale, IL-10 aurait une action bien plus complexe qu'il n'y parait.

TGF-β est une cytokine ant-inflamatoire possèdant également une activité immuno-suppressive. Elle a donc un effet anti-tumorale préventif, du fait

de sa capacité à limiter la création d'un micro-environnement favorable au développement tumoral. TGF-β est également un contre-poids d'IL-6 s'opposant ainsi à la croissance tumorale. Toutefois, à l'instar d'IL-10, TGF-β semble être capable de promouvoir l'invasivité des tumeurs par un mécanisme angiogenique, tout en inhibant des cellules de l'immunité anti-tumorale. Bien que la tendance de TGF-β soit plutôt anti-tumorale, elle peut donc basculer en cytokine pro-tumorale, en fonction du type de cancer et du stade de la tumeur.

Résumé

Ces observations montrent une forte implication des cytokines dans les cancers et en font soit des cibles soit des agents thérapeutiques puissants, et ce malgré des effets secondaires possible à cause de leurs multiples fonctions. Il est intéressant de constater à quel point la dualité entre les cytokines est importante, certaines pouvant être qualifiées de pro-tumorales alors que d'autre seront plutôt anti-tumorales. Ces exemples révelent toutefois que cette discrimination n'est pas si simple, puisqu'une même cytokine peut avoir des effets à la fois pro- et anti-tumoraux. Cette balance d'effets est d'autant plus complexe qu'elle n'est pas la même pour chaque cytokine. Ainsi, IL-6 active de voie de signalisation l'une favorisant la croissance des tumeurs l'autre l'inhibant, ce sont donc deux rôles diamétralement opposés. L'activité pro-tumorale peut s'expliquer par une prépondérance de l'activation de STAT3 sur STAT1. *A contrario*, TGF-β a un rôle préventif sur l'apparition de tumeurs mais peut également, à d'autres stades de la maladie devenir pro-tumorale en augmentant les capacités invasives de la tumeur. Les deux fonctions sont ici différentes et décalées dans le temps. On peut également noter que les cytokines anti-tumorales semblent pouvoir basculer plus facilement en pro-tumorales que le contraire.

1.4.4 L'infection par le VIH

Il est aisé de prévoir qu'avec leur rôle dans l'immunité, les cytokines ont un lien avec l'infection par le VIH. Dès 1997 [56] on a pu établir qu'elles pouvaient accélérer ou contenir cette maladie.
Certaines cytokines, s'avèrent être des activateurs du VIH par augmentation du niveau de rétrotranscription du virus (GM-CSF, TNF-α, TNF-β). Ces cytokines induisent une voie de signalisation se terminant par la translocation de NFκB, un facteur de transcription pour lequel les promoteurs de certains gènes du VIH possède un site de fixation. D'autres cytokines, ayant une fonction redondante avec celles-ci ou activant d'autre voies de signalisation, augmente le niveau de transcription de certains gènes du VIH ou permettent la multiplication des cellules infectées. D'une manière générale, les cytokines dites pro-inflamatoires sont considérées, avec quelques exceptions, comme plutôt activatrices du VIH.
Il existe également des cytokines considérées comme plutôt inhibitrices du VIH, IFNα/β et IL-13 étant celles pour lesquelles cette propriété est la plus évidente. IFNα et β semblent pouvoir agir sur la rétro-transcription du virus ou sur la transcription du virus intégré dans le génome. IL-13 est un suppresseur de réplication pour le VIH, mais cette action semble limitée à cetains types de cellule comme les monocytes issus de macrophages. Certaines cytokines anti-inflammatoires sont également capables d'inhiber l'expression du VIH dans certaines lignées, par l'augmentation de la sécretion d'IL-1ra, un antagoniste de l'IL-1, qui semble impliqué dans l'expression du VIH.
Certaines cytokines ont des effets variables sur le VIH, tantôt activatrices, tantôt inhibitrices. Ainsi IFNγ et TGFβ sont impliqués à la fois dans l'augmentation et la diminution de la répication du VIH dans les monocytes issus de macrophage, selon que ces molécules sont ajoutées avant ou après l'infection. IL-2 est un facteur de croissance important et, par conséquent, considérée comme une cytokine inductrice du VIH, mais des études ont montré que ce rôle n'était pas aussi clair.

Ces quelques exemples montrent la complexité de la relation entre les cytokine et l'infection par le VIH, ouvrant de nombreuses pistes de recherche pour combattre cette maladie.*De facto* des cytokines, ou des inhibiteurs de certaines cytokines ont été intégrés dans des protocoles thérapeutiques.

1.4.5 Résumé

On constate que les cytokines sont impliquées dans de nombreuses pathologies, comme cause de dérèglement ou comme remède potentiel. Ce rôle est souvent lié à une action sur l'inflammation. Cette dernière étant une arme à double tranchant capable d'éliminer une agression extérieure mais aussi de créer un environnement favorable à l'installation de pathologies telles que les maladies auto-immunes ou le cancer. Cette très forte implication dans des pathologies explique en partie l'intérêt qu'elles sucitent. Découvir de nouveaux membres dans cette famille, permettrait de mettre en évidence de nouvelles cibles thérapeutiques potentielles et ouvrirait la voie vers des traitements plus fins et plus efficaces.
On peut également avoir un aperçu du niveau de complexité de leurs actions et interactions. Cette partie illustre la nature duale des cytokines.

1.5 Classification des cytokines

Je présente dans cette partie différentes classifications des cytokines. Ces classifications sont employées par des communautés ayant des points de vue et des objectifs différents, du praticien au phylogéniste, en passant par le biologiste moléculaire. J'ai mis en évidence, dans les parties précédentes, la complexité et le niveau d'imbrication des cytokines entre elles, ce qui explique l'existence de différentes classifications, constituées au fil du temps et des découvertes.
Cette partie a pour objectif de présenter la classification qui sera utilisée tout au long de ce travail.

1.5.1 Classification nominale

C'est la première tentative de classification des cytokines. Comme je l'ai présenté dans l'historique, les premières "activités biologiques" découvertes furent appelées "interférons" qui devint ensuite un type de cytokine. Partant du principe qu'elles étaient produites par des lymphocytes, ces activités biologiques furent ensuite baptisées lymphokines. Elles finirent par être dénommées "cytokines". Puis on distinga les interleukines, les Colony Stimulating Factors(CSF), les Tumor Necrosis Factors (TNF), les Tranforming Growth Factors (TGF) et les chémokines.
Cette classification pose des difficultés d'utilisation car il est difficile de comprendre la signification de chaque niveau. Interleukines, interférons et chémokines représentent trois types de cytokines. Si les chémokines semblent se différencier des autres par leur propriétés chimiotractantes et leur faible poids moléculaire, l'avancée de la connaissance des cytokines montre que les distinctions entre les autres types de cytokines, basée sur les cellules cibles et productrices, perd de son sens.
Cette classification historique est donc basée sur les caractéristiques primaires (effet global, cellules sécrétrices) des cytokines. Elle très imparfaite car liée aux contingences de découverte des cytokines, elle n'en reste pas moins parfois utilisée.

1.5.2 Pro- ou anti-inflammatoire

J'ai présenté précédement le rôle des cytokines dans l'inflammation et les maladies auto-immunes. On y a constaté que la dualité pro/anti-infllamatoire des cytokines était corrélée à l'action des cytokines dans de nombreuse pathologies. Ce critère est fréquemment utilisé pour classer les cytokines. Le tableau 1.4 présente cette classification.
Cette classification a trois inconvénients de taille. Le premier est évident : les cytokines n'ayant pas de rôle dans l'inflammation n'y figure pas. Le second est dû au fait que l'on utilise un critère fonctionnel alors que les cytokines sont connues pour leur pléiotropie. En effet, rien ne justifie de

se baser sur la tendance à stimuler ou à inhiber l'inflammation alors que les cytokines présentent toutes un nombre important de fonctions. Quand bien même la plupart des cytokines auraient un rôle dans l'inflammation, permettant ainsi de toutes les classer selon ce critère, il n'en demeure pas moins que le rôle pro ou anti-inflammatoire de chaque cytokine n'est pas clairment déterminé. J'ai présenté plusieurs exemples dans la partie 1.4 de cytokines ayant les deux types d'activités, soit simultanément soit décalées dans le temps.
Si cette classification reste globalement parlante, elle n'est pas satisfaisante pour classer les cytokines en dehors du contextede l'inflammation.

1.5.3 Classification selon le type de réponse

Toujours basé sur les caractéristiques les plus apparentes des cytokines, cette classification propose d'améliorer la première en se basant sur le type de réponse immunitaire dans laquelle intervienent les cytokines. Ces dernières classées en quatre catégories :

- cytokines intervenant dans la réponse immunitaire (la majorité des interleukines, IFNγ, TNFα et TNFβ,
- les cytokines anti-virales (IFNα, IFNβ, IFN γ [1], et IL16),
- les cytokines impliquées dans l'inflammation et la fibrose (certaines cytokines pro- et anti-inflammatoires (IL-1, IL-6, TNF, IL-10 ...) et des cytokines impliquées dans la fibrose (TGFβ ...)),
- cytokines de l'hématopoïèse (les CSF, SCF, IL-3, IL-5, IL7),
- les chimiokines.

Cette classification a le mérite d'être plus consistante que la classification nominale, plus exhaustive que la classification pro/anti-inflammatoire et d'utiliser des critères cohérants du point de vue des connaissances actuelles. Elle présente toutefois deux inconvénients. Le premier est la redondance de certaines cytokines qui appartiennent à plusieurs catégories. Le second, comme pour la classification basée sur la tendance inflammatoire, est l'uti-

[1] A noter l'existece d'une classification des interférons basée sur leur thermosensibilité

lisation d'un unique critère fonctionnel, malgré la pléiotropie des cytokines. Cette classification reste l'une des classifications fonctionnelles plus cohérantes que l'on puisse actuellement proposer.

1.5.4 Classification selon les récepteurs

A contrario des classifications précédentes qui se basaient sur des caractéristiques fonctionnelles des cytokines, la classification selon les récepteurs utilise une caractéristique moléculaire : le type de récepteur auquel elles se fixent. Ainsi qu'il l'a été mentionné dans la partie 1.3, ces récepteurs se répartissent en cinq groupes :

- les récepteurs de type I (fixant IL-2, IL-3, IL-4, IL-5, IL-6, IL-7, IL-9, IL-11, IL-12, EPO, GM-CSF, G-CSF, LIF, CNTF et TPO)
- les récepteurs de type II (fixant IL-10, IL-19, IL-20, IL-22, IL-24, IFN-α, IFN-β et IFN-γ)
- les récepteurs de type III (fixant les TNF)
- les récepteurs apparentés au immunoglobulines (fixant entre autre IL-1, M-CSF, Mast/SCF)
- les récepteurs de chimiokines.

Les récepteurs étant très facile à classer par similarité, cette classification présente l'intéret de donner une base phylogénique à la classification des cytokines ainsi que de s'affranchir d'un critère fonctionnel, limitatif du fait de la pléiotropie des cytokines. Boulay *et al* [6] ont présenté des arguments suggérant fortement une co-évolution cytokines-récepteurs qui reforce cette classification. Elle présente toutefois un désavantage dû au fait qu'une cytokine peut fixer plusieurs récepteurs, risquant de créer une certaine redondance dans la classification des cytokines. Plus intrigant encore, la découverte de Zsig81 suggère que deux cytokines puissent se fixer l'une à l'autre. Il n'a pour l'instant pas été démontré que Zsig81 soit capable de fixer un récepteur connu. Ce nouveau type de processus pourrait donc rendre caduque une classification par fixation à un récepteur.

1.5.5 Classification selon de la structure protéique

Les progrès de la crystallographie et de la bioinformatique permettent de disposer de structures protéiques, au moins prédites, pour chaque nouvelle séquence découverte. Les structures des cytokines est donc globalement connue. Ce critère est l'un des plus pertinants pour classer des protéines en familles puisqu'on sait que la structure protéique est en général bien conservée dans l'évolution. C'est particulièrement le cas dans la famille des cytokines où plusieurs types de structures secondaires se dégagent très nettement. On distingue trois types de structures :

- les cytokines à quatre hélices α
- les cytokines majoritairement en tonneaux β
- les cytokines ayant une structure en hélice α et feuillet β

Cette classification est employée par SCOP. Il est possible de distinguer, dans chaque groupe, des sous-familles, possédant des particularitées qui leur sont propres, ainsi qu'il le sera précisé pour les cytokines à quatre hélices α, rendant cette classification relativement souple. Elle présente l'immense avantage de classer les cytokines selon un critère d'homologie, s'affranchissant de considérations dûes à la pléiotropie ou à la multiplicité des cellules productrices/cibles et indépendant de la fixation aux récepteur. Ce critère me parait être le plus pertinent pour organiser un groupe de protéines. Cette classification sera donc celle employée tout au long de cette thèse.

1.6 Caractéristiques des cytokines à quatre hélices α

Cette partie a pour objectif de présenter les principales caractéristiques de ma famille d'intéret : les cytokines à quatre hélices α. Le vocable "cytokine" y fera désormais référence, et non à la famille entière, par soucis de commodité. Elle a également pour but de réunir des critères permettant de reconnaitre une cytokine à quatre hélices, critères qui pourront être utilisés comme experts, ainsi qu'il le sera expliqué dans le chapitre 3.

1.6.1 Structure protéique

La structure protéique est le premier critère permettant de distinguer cette famille de gènes. Les cytokines à quatre hélices α ont, comme leur nom l'indique, une structure composée de 4 hélices α, notées A, B, C et D depuis le N-terminal, reliées par des coudes ou des "random coil". On observe quelques résidus impliqués dans la formation de feuillets β. Les hélices sont orientées en "up-up-down-down"[25], c'est à dire que le haut de l'hélice A est relié au haut de l'hélice B qui est relié au bas de l'hélice C qui est relié au bas de l'hélice D. A noter que cette configuration est spécifique des cytokines et n'existe dans aucune autre famille de protéines à l'heure actuelle [21]. Les hélices peuvent être longues ou courtes, ce qui permet de subdiviser cette famille en deux sous-familles. Certaines cytokines à hélices longues peuvent possèder une cinquième hélice un peu plus courte, impliquée dans la fixation avec des chaînes receptrices particulières. Cette cinquième hélice permet de distinguer une troisième sous-famille de cytokines à 4 hélices α. Ces trois sous-familles, les cytokines à hélices longues (ou IL-6), cytokines à chaines courtes (ou IL-2) et à 5 hélices (ou IL-10/interféron) utilisent chacune des récepteurs de façon préférencielle et donc sont plus ou moins liées à certaines fonctions. Le tableau 1.5 présente les principaux membres de chaque sous-famille de cytokines.

La structure protéique étant une propriété bien conservée dans une famille, cette classification peut être considérée comme solide, d'autant qu'elle est appuyée par des données phylogéniques.

1.6.2 Séquences

La séquence protéique est le premier élément par lequel on tente d'estimer l'homologie entre deux protéines et donc par lequel on définit une famille de protéines. Cette estimation passe par le calcul d'un score de similarité entre les deux homologues putatifs. A noter qu'il faut éviter la confusion fréquente entre similarité et homologie. La similarité désigne le degré de ressemblance entre deux séquences alors que l'homologie signifie

Sous-famille IL-6	sous-famille IL-2	sous-famille IL-10
IL-6	IL-2	IL-10
G-CSF	EPO	IFNα
LIF	CSF2	IFNβ
somatotropine	IL-3	IFNγ
CNTF	IL-4	IL-17F
lactogene	IL-5	IL-19
leptine	IL-13	IL-20
oncostatine M	IL-15	IL-22
IL-12	FLT3	IL-24
IL-23p19	IL-7	BCRF
IL-11	TSLP	IL-26
prolactine	IL-9	IL-27
cardiotrophine	TPO	IL28A
NNT-1	IL-21	IL-28B
IL27p28	SCF	IL29

TAB. 1.5 – les trois sous-familles de cytokines à 4 hélices α

que les séquences sont apparentées sur le plan phylogénique. On suspecte souvent une homologie quand la similarité est importante mais deux homologue peuvent avoir un faible taux de similarité (c'est entre autre le cas chez les cytokines), et deux séquences assez similaires peuvent ne pas être homologues.

Pour comparer des séquences, on utilise des méthodes exactes telles que l'algorithme de Smith et Watermann ou des heuristiques telles que BLAST et ses dérivés ou FASTA, pour les recherches dans de larges bases de données. Dans le cas des protéines à quatre hélices α, la similarité entre la plupart des membre est relativement faible, bien que plus élevée que parmi les cytokines en général. Les scores de similarité varient entre 97% pour le couple IL-28A/IL-28B et 15,48% pour le couple IL-11/IFNγ, avec une tendance plus marquée vers des scores de l'ordre de 20%. IL-28A/IL28B étant probablement, avec IL-29, des copies récentes d'un même gène ancestral, leur similarité est assez exceptionnelle du point de vue des cytokines.

1.6.3 Structure des gènes

Les cytokines à 4 hélices α présentent une structure génomique bien conservée. Leurs gènes possèdent en général quatre exons principaux, séparés par trois introns généralement en phase 0 [4] et ce malgré une similarité de séquence très faible. Cette constatation peut s'expliquer par une évolution relativement récente de la famille (depuis son apparition chez les vertébrés et son expension chez les mamifères).L'évolution rapide mais sur une courte durée d'un gène permet plus facilement de procéder à des mutations qui altèrent principalement la fonction qu'à des mutations modifiant la structure du gène et/ou de la protéique.

1.6.4 Localisation dans le gènome

Les cytokines à quatre hélices α sont disséminées sur plusieurs chromosomes mais on a tendance à les retrouver en clusters. Le tableau 1.6 indique la localisation de chacune dans le gènome.

chromosome	gène et localisation
1	IL-10 (1q31-32), IL-19 (1q32.2), IL-20 (1q32), IL24 (1q32)
3	Trombopoiétine (3q27), IL-12A (3p12-q13.2)
4	IL-2 (4q26-27), IL-15 (4q31), IL-21 (4q26-27)
5	csf2 (5q31.1), TSLP (5q22.1), IL-3 (5q31.1), IL-4 (5q31.1), IL-5 (5q31.1), IL-9 (5q31.1) IL-13 (5q31)
6	prolactine (6p22.2-21.3), IL-17F (6p12)
7	EPO (7q22), leptine (7q31.3), IL-6 (7p21)
8	IL-7 (8q12-13)
9	IFNα (9p22), IFNβ (9p21)
11	CNTF (11q12)
12	IFNγ (12q14), kit-L (12q22), IL-22 (12q15), IL-23p19 (12q13.3), IL-26 (12q15)
16	cardiotrophine(16p11.2-11.1), IL-27 (16p11.2)
17	lactogène (17q24.2), GH1 (17q24.2), csf3 (17q11.2-12)
19	FLT3L (19q13.3), IL-11 (19q13.3-13.4), IL-28A (19q13.13), IL28B (19q13.13), IL-29 (19q13.13)
22	LIF (22q12.2), oncostatine (22q12.2)

TAB. 1.6 – localisation chromosomique des cytokines à quatre hélices α

On constate que la plupart des cytokines sont regroupées en clusters

dans le génome. Les clusters sur les chromosomes 5, 12 et 19 sont les plus importants.

1.6.5 caractéristiques physico-chimiques

Les caractéristiques physico-chimiques, bien que n'ayant pas de rôle direct pour la fonction, peuvent être relativement conservés dans une famille de protéines. Ceci s'explique par le fait qu'elles sont indirectement liées à la composition de la séquence, elle même liée à la structure. Ces mesures présentent également l'avantage d'être faciles à utiliser, entre autre pour une expertise.

Taille

Si dans certaines famille de protéines, la taille des différents membres peut être ammenée à varier grandement, ce n'est pas le cas des cytokines à quatre hélices α. Leurs tailles sont comprises entre 132 acides aminés et 352 acides aminés. La taille moyenne se situe à 192,35 acides aminés. A noter que la trombopoiétine (352 acides aminés) se démarque nettement des autre cytokines de cette sous-famille car la deuxième cytokine la plus grande comprend 273 acides aminés et que sans la trombopoiétine, la moyenne des cytokines serait à 188,7 acides aminés. Cette relative homogénéité peut s'expliquer par l'action des cytokines au travers de récepteurs. Compte tenu du fait que certaines chaines receptrices fixent plusieurs cytokines, il est envisageable que les cytokines maintiennent une certaine homogénéité de taille pour répondre à des contraintes stériques dans le site de fixation du récepteur.

Masse moléculaire

La masse moléculaire est une grandeur physique définie comme le rapport entre la masse d'une molécule et l'unité de masse des atomes (UMA équivalente à $1/12^{eme}$ de la masse d'un atome de carbone 12, soit environ 1 Dalton).

Les masses moléculaires des cytokines sont comprises entre 14,32 kDa et 37,82 kDa avec une moyenne de 21,63 kDa. Cette relative homogénéité s'explique de manière assez similaire à la taille.

Point isoélectrique

Le point isoelectrique est le pH où une protéine est à son état zwitterionique c'est à dire où sa charge nette est nulle. Plus que la taille, cette caractéristique dépend de la composition de la séquence en acides aminés, chargés ou non.
Les cytokines ont des points isoelectriques compris entre 5,04 et 11,39 avec une moyenne de 8,04 et un écart-type faible, la variabilité de point isoélectrique étant dûe à quelques cytokines particulières. Cette relative homogénéité s'explique encore une fois par la fixation des cytokines à leurs récepteurs, un récepteur devant parfois reconnaitre plusieurs cytokines différentes. Bien que le mode de fixation de chaque cytokines à son (ses) récepteur(s) ne soit pas toujours connu, on sait qu'un même récepteur utilise des sites de fixation similaires pour ses différents ligands [21] et que ce sont ces derniers qui sont adaptés à leur fixation sur le récepteur. Partant de ce principe, les interactions ligand-récepteur seront de même nature pour l'ensemble des cytokines. Les interactions électrostatiques étant une des principales formes d'interaction protéine-protéine, le partage par plusieurs cytokine de mêmes sites de fixation suggère une conservation au moins locale de la charge chez les cytokines, ce qui n'est probablement pas sans conséquences sur le point isoélectrique.

1.7 Conclusion

Ce chapitre m'a permis de décrire l'objet d'étude de cette thèse à savoir les cytokines et plus particulièrement les cytokines à quatre hélices α. J'ai commencé par résumer la façon dont les cytokines avaient été découvertes et j'ai dégagé des arguments suggérant que tous les membres de cette famille n'avaient pas nécessairement été mis en évidence.

J'ai ensuite tenté de définir ce qu'était une cytokine et j'ai précisé quelques caractéristiques de ces protéines comme leur mode d'action, les fonctions principales pour lesquelles elles sont connues et quelques éléments concernant leur régulation. Cette présentation très générale nous a permis de mettre en évidence l'extraordinaire complexité du réseau fonctionnel des cytokines. Ce réseau existe du fait de quatre caractéristiques importantes :

- la pléiotropie,
- la redondance,
- l'additivité,
- la synergie,
- l'antagonisme.

Ce qui m'a permit d'introduire le concept d'environnement cytokinique.

J'ai ensuite détaillé les mécanismes moléculaires de fonctionnement des cytokines. Nous y avons vu que les cytokines se fixaient à des récepteurs à la surface des cellules cibles, ce qui induisait des cascades de signalisation qui aboutissaient à la régulation de gènes impliqués dans certaines fonctions (prolifération, différenciation, mort cellulaire ...). Ce mode d'action explique en grande partie les principales caractéristiques des cytokines et m'a permis d'introduire la notion de récepteur à cytokines que j'ai approfondie dans la partie suivante.
J'ai décris la structure de ces récepteurs en deux domaines : une chaine réceptrice chargée de fixer la cytokine, associée à une chaine effectrice, chargée de transmettre le signal lors de la fixation du ligand. Ce mode de fonctionnement permet d'expliquer la pléiotropie des cytokines.

J'ai présenté les différentes pathologies dans lesquelles les cytokines étaient suceptibles d'intervenir. Depuis les maladies auto-immunes, inflammations chroniques, jusqu'au cancer et à l'infection par le VIH, nous avons vu que le spectre de ces pathologies était assez vaste mais impliquait le plus souvent une hyper-inflammation dû à un déséquilibre entre cytokines pro- et

anti-inflammatoires. Cette partie m'a permi de montrer l'importance des cytokines sur le plan médical, notament dans les pathologies comportant des composantes immunitaires. Cette partie justifie l'intéret d'étudier ces protéines et laisse entrevoir quelques possibilités qu'apporterait la découverte de nouveau membres de cette famille.

Cette thèse s'appuyant sur des techniques de classification (qui seront abordées dans le chapitre suivant), j'ai présenté les classifications existantes des cytokines. J'ai montré que les classifications usuelles des cytokines avaient des fondements discutables du point de vue phylogénique et pouvaient être mises en difficulté par certaines exceptions. La seule classification solide que l'on puisse retenir est la classification basée sur les structures protéiques, qui servira de référence pendant toute cette thèse.

Je me suis ensuite intéressé de plus près à la famille de cytokine qui fait l'objet de cette thèse : les cytokines à quatre hélices α. J'ai décris les caractéristiques qui les différenciaient des protéines n'étant pas des cytokines mais aussi des autres familles de cytokines. La principale caractéristique de cette famille est, sans surprise, sa structure protéique à quatre hélices α en conformation "'up-up-down-down"'. Ces caractéristiques ont été présentées dans le but d'être utilisées par la suite pour différencier de potentielles nouvelles cytokines parmi les candidats obtenus après classification.

Chapitre 2

Classification supervisée

L'objectif de ce chapitre est de décrire la stratégie et les méthodes employées pour rechercher des homologues de cytokines dans des bases de données. Après un état de l'art de ces méthodes, j'expliquerai la stratégie de recherche des homologues, avant de décrire les outils utilisés. Enfin, une section sera consacrée au processus d'apprentissage de ces outils et à l'évaluation de leur efficacité vis à vis du problème posé.

2.1 État de l'art des méthodes de recherches d'homologues

Cette partie présente les méthodes les plus courantes dans le domaine de la recherche d'homologues. L'objectif est de recenser les différentes stratégies existantes et de dégager les avantages et inconvénients de chacune afin de choisir la (les) plus adaptée(s).

La recherche d'homologues au sein d'une famille est une problématique classique en bioinformatique car c'est de l'identification de ces homologues que proviennent la plupart des connaissances sur les séquences protéiques [57]. Pour rappel, on définit comme homologues deux séquences (de gènes ou de protéines) ayant un ancêtre commun. Nombre de méthodes ont été développées à partir de stratégies très différentes. Ces méthodes de recherche d'homologues peuvent être classées en quatre grandes catégories :

- méthodes basées sur l'analyse de séquences,
- méthodes basées sur les structures protéiques,
- méthodes hybrides,
- méthodes d'apprentissage.

Les méthodes basées sur l'analyse des séquences sont historiquement les premières à avoir vu le jour. Elles sont employées aujourd'hui principalement pour des recherches d'homologues proches dans des tâches où le temps imparti à cette recherche est court (*e.g.* la prédiction de structure par enfilement (threading)).
Théoriquement, les méthodes basées sur les structures protéiques sont plus efficaces que les méthodes basées sur l'analyse des séquences car la structure protéique est mieux conservée que la séquence au cours de l'évolution. En pratique elles sont plus rarement employées car pour une protéine, la structure est plus rarement connue. Les structures étant difficiles à obtenir, les possibilités d'exploration restent limitées aux protéines, peu nombreuses, qui ont été cristallisées. A titre de comparaison, en décembre 2007, la PDB (Protein Data Bank, la principale base de données de structures protéiques déterminées expérimentalement) comportait 44 360 entrées et GenBank (une des trois principales bases de données de séquences biologiques) 80 388 382 entrées. Il y a également des biais dans la représentativité des structures, par exemple le rapport protéines solubles/protéines membranaires dans les banques de données ne reflète pas la situation dans la nature car les protéines membranaires sont plus difficiles à cristalliser. La solution la plus souvent employée pour contourner ces problèmes est de prédire les structures de protéines à partir de leurs séquences mais cette prédiction est loin d'être aussi précise que la détermination par cristallographie ou RMN.
Quelques stratégies hybrides se basant à la fois sur les séquences et sur les structures ont été mises au point pour tirer parti des avantages de ces deux types de données simultanément.
Enfin, les méthodes d'apprentissage représentent une classe à part, capable

d'utiliser aussi bien la séquence, la structure que d'autres données sur la protéine. Ces méthodes se basent sur l'utilisation des membres connus de la famille, ainsi que de contre-exemples, pour "apprendre" à détecter un nouvel homologue.
La littérature sur le domaine étant bien trop vaste pour être décrite de manière exhaustive, je me contenterai de citer quelques exemples pour chaque type de méthodes.

2.1.1 Méthodes basées sur l'analyse de séquences

Ces méthodes exploitent les caractéristiques de la séquence nucléique ou protéique pour rechercher des homologues dans une base de données de séquences. Il existe plusieurs façons d'identifier des homologues à partir des séquences, depuis une simple mesure de similarité à des méthodes plus complexes impliquant des graphes.

Similarité

Les méthodes basées sur le pourcentage de similarité sont très efficaces pour détecter les homologues proches mais cette efficacité diminue grandement pour des homologues distants. Les outils de type BLAST (et plus récemment PSI-BLAST) [3] sont largement utilisés dans cette famille de méthodes. On sait toutefois que deux protéines peuvent être homologues et avoir un faible pourcentage de similarité, soit parce que les protéines ont beaucoup divergé en terme de fonctions et de structures, donc de séquences, soit parce que la structure a été bien conservée lors de l'évolution mais pas la séquence. Pour détecter des homologues distants (faible similarité de séquences) des méthodes plus élaborées ont été mises en place. Par exemple, on peut utiliser les similarités, non plus directement mais en passant par un homologue intermédiaire. Ainsi, l'une des méthodes les plus récentes [29] utilise PSI-BLAST pour faire des alignements multiples de la protéine connue, afin d'obtenir les intermédiaires (sélectionnés selon des critères de score) qui vont eux-mêmes être utilisés pour trouver les homo-

logues distants par similarité. D'autres approches permettent d'améliorer encore ces stratégies *e.g.* en cherchant à reconstruire phylogéniquement des séquences ancestrales de la famille d'intérêt [8]

Motifs

Les méthodes basées sur les motifs sont une extension des méthodes de similarité. La notion de motifs représente de courtes séquences plus ou moins bien conservées dans une famille de gènes ou de protéines. Ces motifs sont parfois directement liés à une caractéristique fonctionnelle ou structurale de la famille. A titre d'illustration, on peut évoquer le motif ($WSXWS$) des récepteurs aux cytokines de type I, décrit dans le premier chapitre. 25 des 34 membres de cette famille possèdent ce motif ce qui en fait un bon moyen d'identifier des homologues.
L'identification de motifs spécifiques se fait généralement par alignement des séquences des membres connus de la famille et identification, souvent visuelle, de zones conservées, avec des dégénérescences possibles, permises par l'utilisation de matrices de substitutions. De ce point de vue, les séquences protéiques offrent une plus grande variété de motifs que les séquences nucléiques, de part la diversité de leur alphabet mais aussi par la possibilité de substitutions plus importantes. Les séquences nucléiques offrent tout de même la possibilité d'utiliser des motifs, particulièrement dans des zones non traduites telles que le promoteur ou, dans une moindre mesure, les UTR (UnTranslated Regions, région non traduite de l'ARNm). De nombreuses bases de données de motifs sont disponibles, les principales étant PROSITE pour les motifs protéiques et TRANSFAC qui est une base de facteurs de transcription incluant leurs motifs de fixation sur le promoteur.
Une fois le(s) motif(s) spécifique(s) défini(s), il suffit d'appliquer des algorithmes de "pattern matching" sur les séquences d'intérêt. Ces motifs sont également utilisés pour améliorer les performances des algorithmes classiques d'alignement pour la recherche d'homologues [22].

Cette stratégie est assez séduisante car elle s'intéresse à des considérations évolutives et semble relativement simple à mettre en oeuvre. Ce n'est toutefois pas toujours le cas du fait de l'absence de motifs détectables dans certaines familles, entre autre celles des cytokines qui sont particulièrement difficiles à aligner du fait de leur faible similarité. Certaines solutions ont été proposées comme celle de Mikolajczak [41] qui permet d'identifier un grand nombre de motifs de manière automatique en autorisant une dégénérescence importante basée sur les propriétés physico-chimiques des acides aminés. Bien que peu adéquate pour une recherche de motifs en elle-même, cette possibilité peut être exploitée par des méthodes d'apprentissage.

Graphes

Couramment utilisés dans plusieurs domaines informatiques, les graphes ont connu plusieurs applications en bioinformatique. Les graphes sont un outil mathématique formalisé dans les années 60 et utilisé pour décrire des ensembles, flux, réseaux .... Schématiquement, un graphe est un ensemble d'objets, appelés sommets, reliés par des arêtes. Un graphe possède des propriétés comme le nombre de sommets, d'arêtes, leur orientation, leur type (parallèles, arc ...) ou leur coloration, qui sont utilisées pour décrire l'ensemble d'objets à représenter. Cet outil s'est avéré extrêmement efficace dans un grand nombre d'applications réelles. Concernant la recherche d'homologues distants, on peut citer le logiciel TRIBES [15], qui convertit les scores de similarité d'un ensemble de protéines en une matrice et effectue une marche au hasard sur le graphe représentant cette matrice pour retrouver les groupes d'homologues, et CLUSTER-C [42] qui s'appuie sur la recherche de cliques. Dans ces deux cas, la démarche diffère de la plupart des autres méthodes de recherche d'homologues. Ces méthodes cherchent à former des groupes les plus cohérents possibles dans l'ensemble de la base de séquences, sans contrainte sur la famille étudiée alors que, classiquement, on s'appuie sur les membres connus d'une famille pour en trouver

les homologues.

2.1.2 Méthodes basées sur les structures protéiques

La plupart de ces méthodes essayent de définir une structure canonique pour la famille d'intérêt puis d'aligner les structures connues en donnant un score à chaque alignement. Les alignements dont le score est supérieur à un seuil (calibré généralement à partir des membres connus de la famille) sont retenus comme candidats potentiels [17]. Par exemple, la technique dite des FFF (Fuzzy Functional Forms), affinée par Cammer *et al* [9], recherche un profil structural des sites actifs de la famille, ce qui permet d'élargir le champ de recherche (puisqu'un site actif a des chances d'être plus conservé que le reste de la structure) ou au contraire de se restreindre à une sous-classe précise de la famille. Le défaut de cette méthode est qu'elle ne détecte que les homologues ayant conservé la même fonction et dont le site actif est correctement défini.
Ces méthodes sont généralement plus efficaces pour rechercher des homologues distants du fait de la conservation des structures dans l'évolution. La limitation du nombre de structures protéiques oblige toutefois à prédire les structures des protéines pour des applications à l'échelle génomique, ce qui tend à rendre ces méthodes moins performantes.

2.1.3 Méthodes hybrides

Ces méthodes combinent l'utilisation des structures et celle des séquences pour détecter des homologues. Elles tentent de tirer parti des avantages de chaque type d'entrée : la quantité de données disponibles pour les séquences et la fiabilité de la relation d'homologie pour les structures. Cette définition exclut les nombreuses méthodes utilisant les structures mais qui s'appuient sur des séquences pour prédire cette structure, car elles n'utilisent les séquences que pour obtenir une information structurale. Par exemple, certaines méthodes hybrides se basent sur une recherche de similarité de séquences et une superposition de structure [57] [18]. Ces

méthodes calculent un score de probabilité d'homologie qui a deux composantes : l'une structurale et l'autre de similarité de séquence. Si le score est supérieur à un seuil, les deux séquences sont déclarées homologues. On voit bien ici que, si la fonction de score est bien conçue, ces méthodes seront aussi efficaces pour les homologues proches (par le côté similarité de séquences) que pour les homologues distants ayant conservé une similarité de structure, pour peu que les structures prédites soient correctes, ce qui demeure leur point faible.

2.1.4 Apprentissage de reconnaissance d'homologues

Les méthodes d'apprentissage qui nous concernent sont issues de la problématique de la classification. On divise traditionnellement ces méthodes en "supervisées", c'est à dire s'appuyant sur un jeu de données labellisées, et "non-supervisées", c'est à dire s'appuyant sur un jeu de données dont les individus sont de classes inconnues. Dans le cadre de la recherche d'homologues où les classes sont prédéfinies (membre ou non-membre de la famille étudiée), les méthodes non-supervisées présentent moins d'intérêt car elles s'appliquent aux cas où les labels sont inconnus et où l'objectif est de constituer des groupes à partir des données, comme par exemple la clusterisation de résultats d'hybridation de puces à ADN. Je ne m'intéresserai donc ici qu'aux méthodes supervisées.
Il existe plusieurs façons de faire de l'apprentissage mais toutes s'appuient sur l'idée que, partant d'un jeu de séquences, dont l'appartenance ou non à la famille étudiée est connue, il est possible de mettre en place un système capable de reconnaître de nouveaux membres de la famille à partir de caractéristiques particulières. Ces caractéristiques peuvent aussi bien être la composition en acides aminés, la présence de sous-séquences de taille fixée ou de motifs dégénérés, que la structure de la protéine. De ce point de vue, ces méthodes ne sont pas spécifiquement basées sur la séquence, encore que ce soit le type de données le plus couramment utilisé, ni sur aucun autre type de données. En ce sens, c'est la famille de méthodes la plus

flexible. Le jeu de séquences connues, appelé jeu d'apprentissage, est utilisé par le programme pour générer un modèle de discrimination des séquences qui sera ensuite utilisé pour prédire l'appartenance ou non à la famille d'intérêt d'une séquence inconnue. La nature de ce modèle varie selon la méthode d'apprentissage. Les grandes forces des méthodes d'apprentissage sont l'utilisation des membres connus pour rechercher des homologues et la variabilité des caractéristiques exploitables, leur conférant une grande souplesse. Ces méthodes se sont montrées bien plus efficaces que toutes autres pour la recherche d'homologues distants.
Les deux méthodes d'apprentissage les plus récentes et les plus utilisées sont les chaînes de Markov cachées (HMM) [31] [16] et les Séparateurs à Vastes Marges (SVM, [28] [33] [26]). Ces derniers sont considérés comme ayant les meilleures performances dans le cadre de la recherche d'homologues.

2.1.5 Résumé

Dans cette partie, j'ai présenté les grandes méthodes de recherche d'homologues en fonction du type de données utilisé. J'ai montré que les méthodes basées sur les séquences étaient couramment utilisées et qu'elles permettaient de mettre facilement en évidence les homologues proches mais qu'elles s'avéraient moins performantes pour les homologues distants. Une approche de type graphe peut s'avérer sensiblement meilleure mais n'assure pas un regroupement systématique des homologues entre eux. D'un autre côté, les méthodes basées sur la structure protéique, théoriquement plus fiables, présentent un écueil de taille : le faible nombre de structures connues oblige à avoir recours à la prédiction qui fait perdre beaucoup de l'efficacité à ces méthodes. L'utilisation de méthodes hybrides est envisageable mais un quatrième type de stratégies, basé sur des méthodes d'apprentissage, présente des performances supérieures. Ces méthodes d'apprentissage, en plus de permettre l'utilisation de la séquence et/ou d'autres caractéristiques répondent exactement à la question posée : connaissant un certains nombre de membres de la famille étudiée, quels sont leurs homologues dans le

jeu de séquences considéré ? A l'heure actuelle, ces méthodes surclassent les autres pour la recherche d'homologues. Parmi elles, les Séparateurs à Vastes Marges (SVM) s'avèrent les plus efficaces, c'est pourquoi je les ai choisis comme technique de recherche de nouveaux membres de la famille des cytokines.

2.2 Les SVM

Dans cette partie, je décris le principe de fonctionnement des SVM. Cette thèse ayant pour objet d'utiliser et non de perfectionner cette technique, je m'en tiendrai à une description intuitive avec quelques éléments de formalisme. Le lecteur intéressé par des précisions sur les SVM est invité à se référer à des ouvrages spécialisés et aux thèses de Markowetz [39] et J. Micklojzak [40].
Je vais commencer par expliquer l'idée générale des SVM avant de présenter brièvement la théorie mathématique sur laquelle ils s'appuient.

2.2.1 Philosophie des SVM

Les Support Vector Machines également appelés en français Séparateurs à Vastes Marges (SVM) ont été décrits par Vapnik [55]. Il s'agit d'une technique d'apprentissage supervisé qui a fait ses preuves dans plusieurs domaines y compris la bioinformatique ([53], [43], [60]).
Comme signalé ci-dessus, les SVM appartiennent au champ des méthodes d'apprentissages dites supervisées, *i.e.* utilisant un jeu d'apprentissage où la classe de chaque individu est connue. Le principe de base de cette technique consiste à représenter les données par des vecteurs et à projeter ces vecteurs dans un espace à grande dimension. On détermine si un objet appartient à la classe d'intérêt en regardant la position de son vecteur par rapport à une frontière délimitant cette classe. L'objectif de l'apprentissage est de déterminer cette frontière. Un hyperplan de séparation est déterminé en fonction de la position des vecteurs exemples par rapport aux

vecteurs contre-exemples. La force des SVM consiste à maximiser la distance entre l'hyperplan et les exemples/contres-exemples les plus proches (appelés "vecteurs supports"). Ceci a pour conséquences de laisser une marge entre ces vecteurs supports et l'hyperplan, permettant de classer des données inconnues qui seraient plus difficiles que celles du jeu d'apprentissage (*i.e.* plus proches de l'hyperplan). Cette stratégie s'appuie sur des considérations statistiques et un ensemble de justifications mathématiques décrites ci-après.

2.2.2 Minimisation de risques

Dans le cadre de la classification discriminante de données, on utilise une fonction de décision f telle que $f : \chi \rightarrow \{-1; +1\}$ associe à chaque donnée de χ une classe (-1 pour un contre-exemple, +1 pour un exemple. On définit donc un risque dit risque réel (ou encore structural) de mauvaise classification qui se calcule de la manière suivante :

$$R(f) = \int_{\chi \times U} L[f(x), u] dP(x, u)$$

où x représente la donnée à classer, $f(x)$ sa classe prédite et u la classe réelle.

Le risque R dépend donc de la fonction de coût L et de la distribution de probabilités $P(x, u)$. Cette distribution est inconnue et ne peut pas être calculée. On peut approcher le risque réel par le risque empirique à l'aide du jeu d'apprentissage.

Le risque empirique permet d'estimer le nombre d'erreurs de classification dans le jeu d'apprentissage. La meilleure fonction de décision que l'on puisse trouver est donc celle qui minimise le risque empirique. Il est évident que plus le jeu d'apprentissage est représentatif de l'ensemble des données, plus le risque empirique s'approche du risque réel. En pratique cela revient à essayer d'avoir le plus large jeu d'apprentissage possible.

Deux problèmes se posent : si la fonction de décision est trop simple les

risques de mauvaises classifications du jeu d'apprentissage et donc la valeur du risque empirique seront trop importants ; mais si la fonction de décision cherche, par sa complexité, à trop bien représenter le jeu d'apprentissage, il y a risque de "surapprentissage" (apprentissage "par coeur") et cette fonction classera mal des données inconnues, autrement dit, le risque empirique ne sera pas forcément assimilable au risque réel. Il est donc nécessaire d'assurer un compromis entre efficacité de classification du jeu d'apprentissage et généralisation de la classification.
C'est dans ce but qu'a été définie la dimension de Vapnik-Chervonenkis (dVC). Cet indicateur a pour fonction de contrôler la complexité de la fonction de décision. Cette dimension permet de borner le risque réel en fonction du risque empirique, en d'autres termes d'optimiser le risque réel en fonction des performances d'apprentissage et de la complexité de la fonction de décision. Cette notion peut être définie de la façon suivante :

$$R(f) = R_{emp}(f) + D(N, dVC, \epsilon)$$

Où $R(f)$ désigne le risque réel, $R_{emp}(f)$ le risque empirique, D une fonction de décision, N, la taille du jeu d'apprentissage et ϵ l'écart autorisé. L'objectif est de définir une fonction D optimale c'est à dire dont la dVC est suffisante pour obtenir un bon taux de bonne classification, sans être trop grande, pour éviter le surapprentissage.
Le problème du risque de mauvaise classification ayant été défini, il est maintenant possible de chercher une marge de séparation entre les exemples et les contre-exemples.

2.2.3 Données linéairement séparables

Le cas le plus simple que l'on puisse rencontrer est celui de données linéairement séparables. J'expliquerai un peu plus loin comment l'on peut généraliser les concepts présentés ci-après à des données non-linéairement séparables.
Il est possible de tracer plusieurs hyperplans pour séparer deux groupes

de données, la question qui se pose est de savoir quel est le meilleur de ces hyperplans. Comme nous l'avons vu précédemment, la réponse à cette question est : "l'hyperplan qui sépare au mieux les données tout en permettant la meilleure généralisation possible". En d'autres termes, il faut donc que l'hyperplan soit le plus éloigné possible des points "limites" de chaque groupe. Ainsi dans la figure 2.1, l'hyperplan de gauche est le plus optimal des deux exemples.

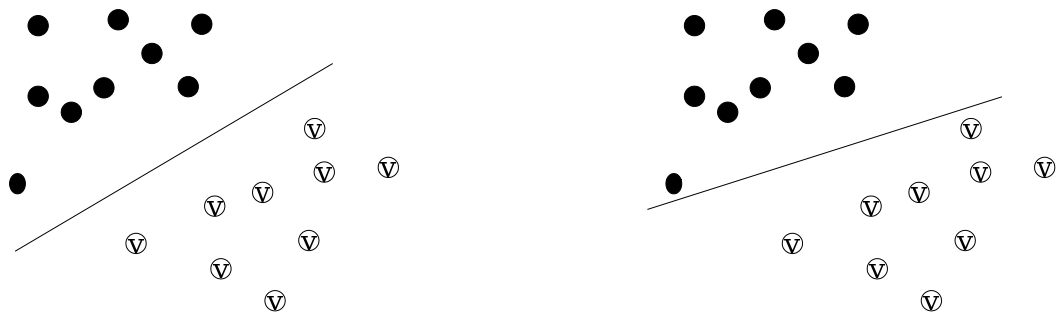

FIG. 2.1 – Deux hyperplans possibles pour discriminer deux groupes de données (les points noirs et les points contenant un "v").

Je vais maintenant expliquer comment calculer cet hyperplan optimal.
Soit x un vecteur de label y représentant un objet ($y = 1$ si l'objet est de la classe cherchée et $y = -1$ sinon). L'équation générale d'un hyperplan est : $w.x + b = 0$ où w représente le vecteur normal à l'hyperplan, x un vecteur de l'hyperplan et b la distance minimale de l'hyperplan à l'origine (*i.e.* le biais). Les données sont classées selon leur position par rapport à cet hyperplan, vérifiant donc les conditions suivantes :

$$\begin{cases} w.x + b \geq 1 \quad si\, y = +1 \\ ou \\ w.x + b \leq 1 \quad si\, y = -1 \end{cases}$$

On appelle vecteurs supports les vecteurs qui satisfont :

$$\begin{cases} w.x + b = 1 & si\, y = +1 \\ ou & \\ w.x + b = 1 & si\, y = -1 \end{cases}$$

Ces vecteurs représentent les "points limites" de chaque groupe et les plus proches de l'hyperplan. Seuls ces vecteurs support sont pris en compte pour la détermination de ce dernier, comme le montre la figure 2.2.

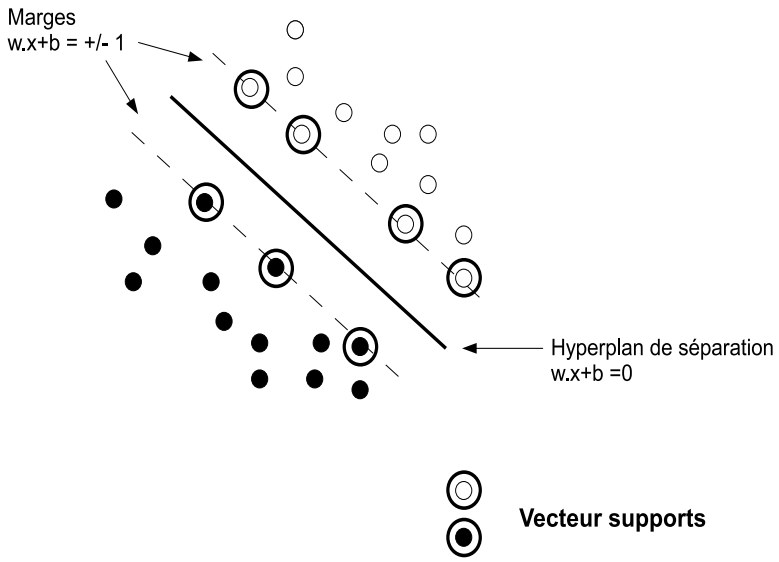

FIG. 2.2 – Représentation des vecteurs supports

Leur distance à cet hyperplan vaut :

$$\begin{cases} \frac{1}{\|w_0\|} & \text{si } y = +1 \\ \frac{-1}{\|w_0\|} & \text{si } y = -1 \end{cases}$$

La marge de séparation, c'est à dire la distance entre les deux classes est alors $\frac{2}{\|w_0\|}$. L'objectif étant de maximiser cette marge, on va chercher à minimiser w_0. Ceci revient à un problème d'optimisation primal qui peut être reformulé de la manière suivante : minimiser la fonction de coût Φ,

où $\Phi(w) = \frac{1}{2}||w||^2$. Une résolution sous contrainte [51] en utilisant les lagrangiens permet de montrer que les vecteurs supports sont les vecteurs ayant un multiplicateur de Lagrange strictement positif et que ce sont les seuls points nécessaires pour déterminer l'hyperplan.

2.2.4 Données non linéairement séparables

La détermination d'un hyperplan optimal est un problème soluble pour des données linéairement séparables mais cette hypothèse est rarement vérifiée en pratique. Dans le cas de données non linéairement séparables, il n'existe pas de méthode déterministe pour résoudre à la fois le problème de maximisation de la marge et celui du positionnement de la frontière. On doit donc faire un compromis entre la capacité de généralisation du modèle et la minimisation des erreurs de classification. Pour résoudre ce problème, il faut, d'après le théorème de Cover [12] rentrer dans un espace de représentation à grande dimension en utilisant une transformation non linéaire pour rendre le problème linéairement séparable. Le problème d'optimisation devient un problème dual, ce qui revient donc à maximiser :

$$\sum_i \alpha_i - \frac{1}{2} \sum_i \sum_j \alpha_i \alpha_j y_i y_j (x_i . x_j)$$

α représentant les poids du problème d'optimisation. Pour résoudre ce problème, il faut faire appel à des fonctions dites noyaux ("kernel" en anglais, qui est une dénomination souvent retrouvée, même dans les articles francophones), permettant de projeter les données dans un espace de dimension supérieure. Le choix de la fonction noyau conditionne l'efficacité du SVM. Une fonction noyau qui projetterait les données dans un espace de trop faible dimension ne permettrait pas de revenir à un cas linéairement séparable alors qu'une fonction qui les projetterait dans un espace trop grand risquerait de mener à du surapprentissage [43]. Plusieurs fonctions noyaux "classiques" existent :

– linéaire : $K(x, x') = x.x'$

- polynomiale : $K(x, x') = (\gamma.x.x' + coef0)^{degree}$
- Radial Basis Function (RBF) : $K(x, x') = e^{-\gamma||x-x'||^2}$
- sigmoïdale : $K(x, x') = tanh(\gamma.x.x' + coef0)$

avec x et x' désignant deux vecteurs. A noter qu'il existe d'autres fonctions noyaux, développées pour s'adapter à des problèmes particuliers.

Algorithme SVM

On peut résumer les étapes d'une classification par SVM de la manière suivante :

1. apprentissage
   - vectorisation des objets du jeu d'apprentissage
   - projection dans l'espace à grande dimension *via* une fonction noyau présélectionnée
   - détermination de l'hyperplan de séparation
2. classification
   - vectorisation des objets à classer
   - projection dans l'espace à grande dimension *via* la fonction noyau
   - détermination de la classe de l'objet en fonction de sa position par rapport à l'hyperplan de séparation

2.2.5 Probabilité d'appartenance à une classe

Une caractéristique peu usitée des SVM est la distance à la marge. En effet, en sortie du classifieur, on reçoit classiquement la classe de l'objet vectorisé en entrée. Cette réponse est de type booléen, ne donnant aucune indication sur la confiance que l'on peut y accorder. Plusieurs outils de classifications récents disposent d'un moyen d'estimer cette confiance. Intuitivement, plus la distance qui sépare un vecteur de la marge est importante et plus il est probable que l'objet appartienne à la classe qui lui a été assignée. De récents développements des algorithmes de SVM [10] permettent ainsi de calculer une probabilité d'appartenance à la classe assignée. Grossièrement, l'obtention de cette probabilité s'effectue en procédant

à une série de validations croisées ayant pour but de calculer la probabilité bayésienne d'appartenance à la classe, connaissant la distance à la marge. Cette mesure d'un taux de confiance de la prédiction est plus intéressante, dans notre problème de recherche d'homologues, qu'une assignation binaire car elle présente l'avantage de permettre un classement des séquences selon leur probabilité d'appartenance à la classe des cytokines.

2.3 Méthodes de vectorisation pour la biologie

J'ai expliqué dans la partie précédente comment fonctionnaient les SVM. Ces derniers représentent les données sous forme de vecteurs algébriques qui sont projetés dans un espace vectoriel afin de prédire leur position par rapport à un hyperplan frontière. Ceci implique donc que les données (des séquences biologiques, des structures ...) soient transformées en vecteurs. Cette partie présente différentes méthodes dites de vectorisation appliquées aux objets biologiques. Je commencerai par expliquer d'une manière générale le principe de vectorisation avant de présenter les cinq principales méthodes décrites dans la littérature, méthodes que j'utiliserai dans la suite de cette thèse.

2.3.1 Classifieurs

J'appelle "classifieur" un outil de classification disposant d'une méthode de vectorisation et d'une méthode de classification basée sur ce vecteur. Les classifieurs que je présente ci-après ont la même méthode de classification, à savoir les SVM décrits précédemment, je présenterai donc uniquement les méthodes de vectorisation par lesquelles ils diffèrent.

2.3.2 Spectrum kernel

Spectrum kernel [33] utilise comme composantes de vecteur la composition en sous-séquences de la séquence traitée. L'idée consiste à fixer une taille k de sous-séquence et à regarder la composition de la séquence en

sous-séquences de cette taille. On obtiendra donc des vecteurs de taille fixe, égale à 20^k (*i.e.* le nombre de sous-séquences possibles de taille k pour une séquence protéique, 20 étant le nombre d'acides aminés constituant, à quelques exception près, les protéines). Les composantes du vecteur pourront prendre les valeurs 0 ou 1 ; 1 si la sous-séquence est présente dans la séquence vectorisée, 0 sinon. Ces composantes représentent les sous-séquences possibles. Pour obtenir le vecteur, on balaye la séquence avec une fenêtre glissante de taille k et pour chaque sous-séquence rencontrée, on inscrit 1 dans la coordonnée correspondante.
Par exemple pour une fenêtre de taille 3, voici la vectorisation obtenue pour la séquence AACYYY :

indice	0	1	...	20^3
sous-séquences	AAA	AAC	MGT	YYY
valeur	0	1	0	1

Le principal avantage de Spectrum kernel est sa rapidité d'exécution (dû à la simplicité de sa méthode de vectorisation) qui en fait une technique très adaptée à la classification d'une grande quantité de séquences, comme c'est le cas dans cette étude.

2.3.3 Mismatch kernel

Ce classifieur [34] est un raffinement de Spectrum kernel. L'idée est toujours de regarder la composition en sous-séquences de la séquence à vectoriser mais en permettant une variation par rapport aux sous-séquences possibles. Cette variation se présente sous la forme d'un certain nombre de mésappariements (mismatch) entre la sous-séquence observée et les sous-séquences théoriques. Ainsi si on autorise un mésappariement, les séquences AAC, AAD, ..., AAY, ACA, ..., AYA, CAA, ..., YAA seront reconnues comme autant de variations de AAA. Cette variabilité a été introduite pour tenir compte de la variabilité biologique des séquences et rendre ainsi le classifieur plus souple. En pratique on obtiendra des vecteurs similaires

à ceux de Spectrum kernel à savoir de taille k (nombre de sous-séquences possibles) et booléens (1 si la sous-séquence ou une de ces variations est présente, 0 sinon). Les vecteurs auront toutefois tendance à avoir plus de composantes à 1 que ceux de Spectrum kernel car même des sous-séquences absentes de la séquence pourront être représentées par des variations (*e.g* AAA avec deux variations est représenté aussi bien par AAA, AA*, A*A, *AA, A**, *A*, **A, où "*" signifie "n'importe quel acide aminé"). Cet élargissement du nombre de sous-séquences améliore la performance du classifieur par rapport à Spectrum kernel mais son temps d'exécution est augmenté du fait de la plus grande complexité de vectorisation et de l'augmentation du temps de traitement de chaque vecteur.

2.3.4 Pairwise

Ce classifieur [35] utilise des scores de recherche de similarité pour vectoriser les séquences. L'idée est tout simplement d'aligner la séquence à vectoriser contre des séquences positives et négatives et de mettre dans le vecteur les scores d'alignements. On obtiendra donc un vecteur de taille n, n étant le nombre de séquences de référence utilisées, et dont les composantes seront des e-values. L'utilisation de scores de similarité, qui avaient été écartés, peut sembler de prime abord peu efficace, toutefois, l'intéret est que l'on travaille ici sur non pas un score mais un profil de scores, qui est donc un peu plus sensible qu'un score isolé. *De facto*, il est probable qu'une cytokine inconnue soit plus similaire à au moins quelques membres connus qu'à des contre-exemples. Ceci justifie donc l'emploi de ce classifieur d'autant qu'en pratique il donne d'excellents résultats, supérieurs à Spectrum kernel ou Mismatch kernel dans la littérature. Son principal défaut est sa lenteur à d'exécution car le calcul du score de Smith & Watermann (SW score) est de complexité quadratique et répété sur l'ensemble du jeu de référence.

2.3.5 PairwiseBlast

Ce classifieur est une variante de Pairwise, qui utilise BLAST à la place de l'algorithme de Smith et Watermann pour aligner le candidat contre les séquences du jeu d'apprentissage. Cette stratégie accélère grandement la vectorisation, du simple fait que BLAST est une heuristique de complexité moindre que l'algorithme de Smith et Watermann.

2.3.6 LA kernel (Local Alignments kernel)

LA kernel [50] est un raffinement de Pairwise et PairwiseBlast, toujours basé sur le calcul d'un score d'alignement. Saigo *et al* [50] démontrent que le SW score n'est pas une fonction noyau théoriquement valide pour les SVM. La principale raison est que le SW score ne calcule que le score de l'alignement optimal entre deux séquences alors qu'une fonction noyau valide devrait sommer les contributions de tous les alignements possibles pour obtenir le score. De plus le SW score est basé sur le logarithme du score de similarité, ce qui ne conserve pas une des propriétés mathématiques des fonctions noyaux à savoir que la matrice les représentant soit définie et positive. Dans cette optique LA kernel est une correction de Pairwise qui prend en compte toutes les contributions des alignements possibles entre la séquence à vectoriser et les séquences de la famille. En dehors de cela ce classifieur fonctionne de la même façon que Pairwise. En pratique Saigo *et al* démontrent que LA kernel surclasse les autres classifieurs dans la recherche d'homologues distants, en utilisant la base SCOP [11]. Le principal inconvénient de ce classifieur est son temps de calcul considérablement supérieur à tous les autres classifieurs y compris Pairwise.

2.3.7 Résumé sur les SVM

Dans les deux parties précédentes, j'ai décrit les SVM et des méthodes de transformations d'objets biologiques en vecteurs mathématiques utilisables pour les SVM. Les SVM s'avèrent une excellente méthode de classification.

D'une part, elle fait appel à un fond mathématique solide, d'autre part le fait de prendre en compte uniquement les exemples les plus difficiles à classer permet de maximiser les capacités de généralisation. Cette technique est applicable aux objets biologiques grâce aux différents travaux sur les méthodes de vectorisation que j'ai présentées. Leurs multiples avantages justifient leur application à des problèmes complexes tels que la recherche d'homologues distants chez les cytokines.

2.4 Mise en oeuvre des SVM sur la classification des cytokines

Cette partie décrit les données que j'ai utilisé pour l'apprentissage et les tests de performance des classifieurs. Elle présente également le score ROC ainsi que la technique de validation croisée, tous deux employés pour l'évaluation de ces classifieurs.

2.4.1 Logiciels

Il existe d'excellents logiciels libres implémentant la méthode des SVM. Parmi eux, j'ai choisi d'utiliser LibSVM [10] qui présente plusieurs avantages. Outre son implémentation éprouvée et ses mises à jour régulières, ce logiciel offre des fonctionnalitées comme le calcul de la probabilité d'appartenance à la classe d'intérêt ou un test par validation croisée basé sur le taux de bonne classification. De plus, il existe une version de LibSVM en JAVA, langage qui fut utilisé, avant ce travail, pour le développement de certains outils et que j'ai moi-même repris.

Les méthodes de vectorisation présentées ci-dessus ont été développées au laboratoire à partir des articles de la littérature.

2.4.2 Données

Les SVM étant des méthodes de classification supervisée, il est nécessaire de disposer de séquences labellisées pour procéder à l'apprentissage. Le jeu d'apprentissage contient à la fois des exemples de cytokines (labellisées +1) et des contre-exemples de cytokines (protéines connues pour ne pas appartenir à cette famille, labellisées −1). Cette considération vaut également pour l'évaluation où le label des séquences doit être connu de l'expérimentateur pour pouvoir analyser les performances de l'outil. Pour constituer le jeu de séquence positives, j'ai utilisé les 45 séquences, appartenant aux trois sous-familles de cytokines à quatre hélices α : la sous-famille IL-6, la sous-famille IL-2, la sous-famille IL-10/IFN. Le tableau 2.1 présente ces différents membres avec leur numéro d'accession SwissProt.

famille IL-6	ID SwissProt	famille IL-2	ID SwissProt	famille IL-10	ID SwissProt
IL-6	P05231	IL-2	P01585	IL-10	P22301
G-CSF	P09919	EPO	P01588	IFNα	P01563
LIF	P15018	CSF2	P04141	IFNβ	P01574
somatotropine	P01241	IL-3	P08700	IFNγ	P01579
CNTF	P26441	IL-4	P05112	IL-17F	Q96PD4
lactogen	P01243	IL-5	P05113	IL-19	Q9UHD0
leptine	P41159	IL-13	P35225	IL-20	Q9NYY1
oncostatine M	P13725	IL-15	P40933	IL-22	Q9GZX6
IL-12	P29459	FLT3	P49771	IL-24	Q13007
IL-23p19	Q9H2A5	IL-7	P13232	BCRF	P03180
IL-11	P20809	TSLP	Q969D9	IL-26	Q9NPH9
prolactine	P01236	IL-9	P15248	IL-27	Q8TAD2
cardiotrophine	Q16619	TPO	P40225	IL28A	Q8IZJ0
NNT-1	?	IL-21	?	IL-28B	Q8IZI9
IL27p28	Q8NEV9	SCF	P21583	IL29	Q8IU54

TAB. 2.1 – les trois sous-familles de cytokines à 4 hélices α

Pour constituer les jeux de contre-exemples, j'ai utilisé 6493 séquences non-cytokines, tirées de la base SCOP [11]. SCOP est une base de données de séquences protéiques, classées par type de structure. Les séquences présentes dans cette base sont uniquement des séquences dont la structure est connue par des moyens expérimentaux, ce qui inclut l'ensemble des structures de la PDB. La base elle-même est hiérarchisée en trois niveaux de repliements, du plus général au plus particulier.

Ces niveaux sont :

1. repliement global : Les protéines de même repliement global ont une structure secondaire similaire dans son ensemble (globalement en hélices α, globalement en feuillets β, ou globalement constitué ces deux types d'éléments). Les protéines de même repliement global ne sont pas nécessairement apparentées.
2. super-famille : Ce sont des protéines ayant une faible similarité mais dont les caractéristiques structurales suggèrent une origine commune. A noter que les cytokines, toutes sous-familles confondues ne forment pas, du point de vue de SCOP, une super-famille car elles ont des repliements globaux différents (il existe des sous-familles dont le repliement est en feuillet β).
3. famille : Ce niveau, le plus précis, permet de regrouper des protéines clairement apparentées, selon les indices structuraux. La plupart des séquences d'une même famille présentent une bonne similarité entre elles, mais ce pas une règle stricte. Ainsi, les cytokines à quatre hélices α forment une famille au sens de SCOP, bien que certains membres aient une faible similarité entre eux.

Cette structure de la base SCOP facilite grandement l'extraction de contre-exemples de cytokines à quatre hélices α. En récupérant toutes les séquences n'appartenant pas au repliement de type *4-helical cytokines*, on obtient un jeu de contre-exemples représentatifs des repliements non-cytokine.

Je dispose donc comme données connues de 45 séquences de cytokines à quatre hélices α et de 6493 contre-exemples.

2.4.3 Le score *ROC*

Classiquement, on utilise des mesures telles que le taux de bonne classification (nombre de séquences pour lesquelles le classifieur renvoie un label correct rapporté au nombre total de séquences) pour déterminer l'efficacité d'un classifieur. Cette mesure implique de considérer le classifieur comme

retournant une réponse binaire : "appartient" ou "n'appartient pas" à la classe d'intérêt. Ainsi que je l'ai dit précédemment, les SVM peuvent renvoyer une probabilité d'appartenance à la classe d'intérêt, qui est une mesure plus fine qu'un simple résultat booléen. Dans ce cas, il est nécessaire de fixer un seuil de risque à partir duquel on considère qu'une séquence appartient à cette classe d'intérêt pour calculer un taux de bonne classification. Outre le problème fondamental de la détermination de ce seuil, qui dépend de la stringence que l'on veut donner au classifieur, on constate qu'il y a une perte importante de finesse de l'information, puisque cette technique nous ramènerai à un classifieur binaire.

Il est possible d'éviter cette perte en considérant les résultats du classifieur sous un autre angle : celui d'un classement. Si l'on décide de classer les résultats selon leur probabilité d'appartenance à la classe d'intérêt, un classifieur idéal devrait classer tous les positifs avant les négatifs. Cette approche comporte deux principaux avantages. Le premier est l'affranchissement par rapport à une valeur seuil. On conserve ici l'information de probabilité d'appartenance à la classe qui permet de choisir en priorité les séquences les plus intéressantes, sans *a priori* sur la valeur de leur probabilité, pour laquelle il est parfois difficile de donner une interprétation. Le second avantage est plus proprement lié à ce travail. Considérant que ces outils seront amenés à être appliqués à un grand ensemble de séquences, on peut s'attendre à ce qu'un outil, aussi performant soit-il, produise un grand nombre de faux positifs, c'est à dire de candidats n'étant pas des cytokines. Dans cette optique, classer les séquences selon un score permet de définir un ordre rationnel dans lequel les candidats doivent être examinés. C'est dans ce contexte de classement que le score *ROC* (Receiver Operating Characteristic curve [63]) prend tout son sens. Cette méthode permet de quantifier le bon ordonnancement du classement, à savoir dans quel mesure les positifs sont classés avant les négatifs. Le score *ROC* peut être vu comme la probabilité, lorsqu'on tire un positif et un négatif dans le classement, que le positif soit classé avant le négatif. Mathématiquement, on

peut le définir comme :

$$ROC = \frac{\Sigma_{i=1}^{n^+}\Sigma_{j=1}^{n^-}\mathbf{1}\pi_{f(x_i^+)>f(x_j^-)}}{n^+n^-}$$

où $f(.)$ représente la fonction de score utilisée pour le classement (*i.e.* les probabilités d'appartenance à la classe des cytokines renvoyées par les classifieurs), x^+ (x^-) le jeu des positifs (négatifs), n^+ (n^-) le nombre de positifs (négatifs) et 1_π est une fonction renvoyant 1 si le prédicat π est vrai et 0 sinon. Le score ROC peut être représenté graphiquement comme l'aire sous la courbe ROC (AUC : Area Under Curve) du nombre de positifs rencontrés dans le classement contre le nombre de négatifs. Un exemple de courbe ROC est présenté dans la figure 2.3. Concrètement, pour chaque rang du classement, un point de la courbe ROC représente le nombre de négatifs rencontrés avant lui en abscisses et le nombre positifs rencontrés avant lui en ordonnées. On calcule ensuite l'intégrale de cette courbe. Plus cette intégrale est proche de 1 et plus le classement à tendance à avoir les positifs en premiers et les négatifs ensuite. Une intégrale à 0 signifie que la méthode inverse positifs et négatifs mais qu'elle les discrimine parfaitement. Une intégrale de 0.5, indique un classement aléatoire, ce qui est le pire résultat pour un classifieur.

Il est possible d'affiner encore la mesure de ROC en ne tolérant qu'un certain nombre de négatifs, au-delà desquels les candidats ne seront plus considérés. L'AUC est calculée de la même manière mais le calcul s'arrête quand on a rencontré un nombre déterminé de négatifs. En pratique, il est possible, à partir d'un même classement, de calculer n'importe quel ROC_x où x est le nombre de négatifs tolérés. Le ROC_x présente comme intérêt de ne s'occuper que du haut du classement, il est donc employé dans des cas, comme celui de ce travail, où le classement est de grande taille et où on sait que l'on sera limité dans le nombre de candidats étudiés par la capacité de traitement de l'expert humain. Dans cette situation, le ROC_x prend tout son sens puisqu'il évalue plus précisément la partie du classement que l'on choisira de traiter.

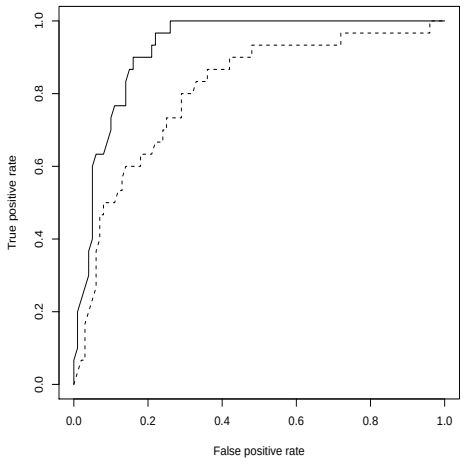

FIG. 2.3 – Exemples de courbes *ROC*. La courbe en traits pleins (C1) et la courbe en pointillés (C2) représentent de courbes *ROC*. Le score *ROC* de C1 s'approche plus de l'axe des ordonnés, indiquant que dans les premiers rangs de son classement, les positifs sont plus nombreux que dans les premiers rangs du classement de C2. De plus C1 forme un plateau à partir de 30% de faux positifs, indiquant qu'à partir d'un certain rang, le classement n'a plus que de négatifs. Au contraire C2 n'atteint jamais de plateau ce qui signifie qu'il reste des positifs dans les derniers rangs du classement. L'aire sous la courbe de C1 est visiblement supérieure à celle de C2, ce qui signifie que le classement C1 est mieux ordonné que le classement C2.

Le calcul de l'AUC de la courbe *ROC* est une méthode d'évaluation plus intéressante que le simple calcul du taux de bonne classification et elle sera employée pendant toute cette thèse. Par commodité, j'appellerai tout simplement "score *ROC*" le calcul de l'AUC de la courbe *ROC*.

2.4.4 Validation croisée

La validation croisée est un moyen couramment employé dans la littérature pour évaluer l'efficacité d'une méthode de classification. Le principe général de la validation croisée consiste à partitionner un jeu de données labellisé en apprentissage et test, de façon à ce que les séquences employées en test soient différentes de celles du jeu d'apprentissage. Les cas particuliers de classification (*i.e.* des séquences testées trop proches de celles utilisées lors de l'apprentissage, ce qui mènerait à une évaluation trop optimiste, ou au contraire trop éloignées, ce qui mènerait à une évaluation trop pessimiste) sont évité par répétition de cette méthode n fois.
Le fonctionnement de validation croisée est le suivant : Le jeu de données

va être partitionné aléatoirement en n lots (n est choisi selon la taille du jeu de données), un de ces lots est mis de côté pendant que les $n-1$ autres servent à l'apprentissage proprement dit. Une fois l'apprentissage terminé, on classe le lot laissé à part pour une classification afin d'estimer l'efficacité du classifieur. Ce processus est répété n fois, en laissant à chaque fois un lot différent de côté. Les n lots sont de cette façon tous testés. La mesure de performance n'est pas intrinsèque à la validation croisée. Classiquement on procède à un simple calcul du taux de bonne classification mais d'autres méthodes, telles que le calcul du score ROC peuvent lui être substitués. Dans ce travail, j'ai utilisé le ROC comme méthode d'évaluation proprement dite, pour les raisons précisées ci-dessus.
Il existe une version extrême de validation croisée qui ne laisse qu'un seul exemple (dans mon cas une seule séquence) du jeu d'apprentissage pour tester l'efficacité du classifieur, cette étape étant répétée autant de fois qu'il y a d'exemples dans le lot d'apprentissage (le tirage est sans remise c'est à dire que chaque exemple servira à tester une fois). Cette méthode, appelée $leave-one-out$, a un caractère déterministe, puisque tous les exemples servent $n-1$ fois d'apprentissage et elle donne donc une réponse exacte. Elle est toutefois entachée d'un biais lorsque les exemples se ressemblent beaucoup, puisqu'ayant plusieurs copies similaires dans le jeu d'apprentissage, ces exemples sont faciles à classer positivement. Cette stratégie d'évaluation ne peut donc être utilisée qu'en dernier recours, quand le nombre d'exemples est trop petit pour être divisé en lots de taille significative.

2.4.5 Taux de corrélation de Kendall

Ce taux de corrélation, qui sera nommé taux de Kendall dans la suite, permet de calculer la similarité entre deux classements ordonnés *i.e.* le nombre d'association rang-objet identique entre deux classements. Les objets sont des items des deux classements, supposés uniques. La formule de

individus	A	B	C	D
rang selon le classement 1 (référence)	1	2	3	4
rang selon le classement 2	2	3	1	4

TAB. 2.2 – exemple de rangement selon un classement de référence

ce taux de corrélation est :

$$\tau = \frac{4P}{n(n-1)} - 1$$

où P représente le nombre de paires concordantes entre les deux classements et n le nombre total d'objets dans les deux classements. Pour que le taux de Kendall ait un sens, les deux classements doivent être de même taille et comporter les même items.
P est calculé de la manière suivante : un des deux classements est choisi arbitrairement comme classement de référence, les individus sont rangés selon ce classement. Puis leurs rangs selon le second classement sont reportés comme il suit : soit deux classement 1 et 2 :

1. A, B, C, D
2. C, A, B, D

Pour calculer P, on part du premier individu (A) et on compte le nombre d'individus après lui selon le classement non-référence (classement 2) ; il y en a deux (B et D) dans cet exemple. On répète cette opération pour les autres individus selon l'ordre du classement référence (classement 1 *i.e.* B,C,D) et on somme l'ensemble des résultats obtenus (ici 1 pour B, 1 pour C et 0 pour D), ce qui donne $P = 2+1+1+0 = 4$). Le taux de Kendall de ces deux classements est donc égal à $\frac{4\times4}{4} - 1$ soit environ 0.333, indiquant qu'ils sont faiblement sont faiblement corrélés.
Le taux de Kendall varie entre -1 et 1. Un taux de Kendall de -1 implique que les deux classements sont parfaitement inversés, un taux de Kendall de 0 implique que les deux classements n'ont aucune paire concordante et un taux de Kendall de 1 indique qu'il s'agit de deux fois le même classement.

2.5 Apprentissage

Dans cette partie, je vais plus spécifiquement parler de l'apprentissage des classifieurs. La phase d'apprentissage est une étape clé de la construction d'un classifieur. Cette étape est sous-divisée en deux phases que sont :

– le choix de la fonction noyau,

– la création du modèle.

Je vais décrire ici ces deux phases.

Choix de la fonction noyau

La fonction noyau RBF étant généralement la plus performante, elle est particulièrement recommandée comme fonction par défaut [10]. Son utilisation impliquant des cycles de validations croisées pour optimiser ses paramètres, elle est toutefois plus coûteuse en temps de calculs que la fonction linéaire, qui ne comprend pas de paramètres à optimiser. Pour comparer les performances de ces deux fonctions noyau, j'ai procédé à 100 classifications pour chacune, en utilisant un jeu d'apprentissage de 15 cytokines et 100 contre-exemples et un jeu de test de 30 cytokines et 6000 contre-exemples. Le score *ROC* de chaque itération a été mémorisé et j'ai déterminé le score *ROC* moyen pour chaque fonction noyau. J'ai utilisé le test de Student pour comparer les distributions de *ROC*. Le tableau 2.3 présente les scores *ROC* moyens pour le classifieur Spectrum kernel. Des résultats comparables ont été obtenus pour les quatre autres classifieurs.

Les deux fonctions donnent des résultats en moyenne comparables, im-

fonction noyau linéaire	fonction noyau RBF
0.9751	0.9721

TAB. 2.3 – score *ROC* moyen de Spectrum kernel, utilisant les fonctions noyaux linéaire et RBF

pression confirmée par le test de Student, pour lequel l'hypothèse H0 est retenue à plus de 99%.

Le gain obtenu par rapport à l'utilisation d'une fonction noyau linéaire

étant faible, j'emploierai préférentiellement cette dernière pour les analyses de performances des classifieurs.

2.5.1 Génération du modèle de classification

Une fois la fonction noyau choisie et ses paramètres optimisés, un modèle peut être obtenu. Par modèle, on entend une définition, dans un espace vectoriel, de l'hyperplan de séparation entre les classes représentées dans le jeu d'apprentissage (généralement deux, bien que les dernières implémentations des SVM supportent une classification multi-classes). Concrètement, on définit, par le biais de l'apprentissage, la frontière entre cytokines et non-cytokines. Cette frontière dépend des caractéristiques du classifieur (fonction noyau et ses paramètres) mais aussi et surtout du jeu d'apprentissage. Pour déterminer les conditions optimales d'apprentissage, j'ai évalué l'efficacité des classifieurs en fonction du nombre de contre-exemples en apprentissage et du type de contre-exemples (issus uniquement du génome humain ou de tous les génomes connus). J'évoque également la possibilité d'utiliser des orthologues (homologues issus d'autres génomes) de cytokines pour enrichir le jeu d'apprentissage.

Nombre de contre-exemples en apprentissage

Le choix du nombre de contre-exemples est également un paramètre important de l'apprentissage. Intuitivement, plus il y a de contre-exemples (et d'exemples) différents et plus le classifieur pourra trouver un hyperplan optimal pour discriminer les deux classes. Il est toutefois intéressant d'évaluer le nombre de données nécessaires et suffisantes en apprentissage pour obtenir les meilleures performances de classification. Un nombre important de données dans le jeu d'apprentissage peut considérablement augmenter le temps de calcul de cette étape. Pour des classifieurs de grande complexité algorithmique comme Pairwise ou LA kernel, qui utilisent également le jeu d'apprentissage lors de la vectorisation d'une séquence inconnue, un large jeu d'apprentissage peut rendre ce temps de calcul rédhibitoire. Il convient

donc de choisir judicieusement la taille et la composition du jeu d'apprentissage. Dans le problème qui nous intéresse, le nombre total d'exemples est suffisamment faible pour permettre d'en employer la plus grande partie possible lors d'un apprentissage. Ce n'est en revanche pas le cas des contre-exemples.

Il est généralement conseillé, pour l'apprentissage d'un SVM d'utiliser un nombre égal d'exemples et de contre-exemples. Cette pratique revient à présenter au classifieur une répartition qui n'existe pas dans la réalité (où le nombre de contre-exemples est beaucoup plus important que le nombre d'exemples).
Afin de déterminer un nombre de contre-exemples idéal, j'ai étudié la variation du *ROC* en fonction du nombre de contre-exemples dans le jeu d'apprentissage. J'ai utilisé trois jeux de données de départs comportant respectivement 18 exemples/18 contre-exemples, 27 exemples/27 contre-exemples, 36 exemples/36 contre-exemples et j'ai augmenté à 100 puis par paliers de 100 jusqu'à 1000 le nombre de contre-exemples dans le jeu d'apprentissage. Chaque apprentissage et classification a été réitéré 50 fois et la moyenne de la différence de *ROC* entre la classification initiale (*e.g.*18 exemples/18 contre-exemples) et la classification courante a été calculée. La figure 2.4 présente les résultats de classification avec le classifieur PairwiseBlast, des résultats similaires ont été obtenu avec les quatre autres classifieurs.

Les résultats montrent une quasi-saturation dès que l'on dépasse 100 contre-exemples, et ce quelque soit le nombre d'exemples dans le jeu de données. Le nombre de contre-exemples d'un jeu d'apprentissage peut donc être limité à 100 sans encourir une baisse de performances de classification.

Choix des contre-exemples

Concernant le choix des contre-exemples, la question est plus délicate à résoudre. L'idéal serait de présenter aux classifieurs les contre-exemples

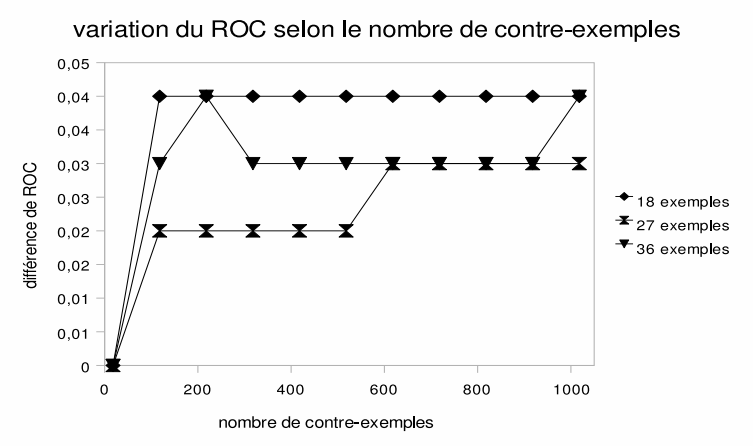

FIG. 2.4 – Variation selon le nombre de contre-exemples pour Spectrum kernel

les plus représentatifs des protéines non-cytokines et les plus proches de ces dernières (*i.e.* les plus difficiles à classer) possibles. SCOP offre une possibilité intéressante de part sa structure. En effet, au niveau "repliement global", les séquences sont classées en sept grands types de repliements, indexés de A à G. La solution qui donne la plus grande diversité de contre-exemples est donc un tirage uniformément réparti d'un nombre identique de contre-exemples dans chacun de ces repliements globaux. La solution qui consiste à ne retenir que les séquences humaines de SCOP est séduisante mais elle limite le nombre de contre-exemples possibles à 1074 et ne permet pas de disposer de représentants pour chaque grand type de repliement, or l'absence de certaines de ces catégories parmi les protéines humaines de SCOP ne signifie pas que ces repliements n'existent pas parmi les protéines humaines *in vivo*. En intégrant toutes les séquences de SCOP, on dispose d'un ensemble de 6493 séquences, représentant plusieurs fois chacun des grands types de repliements. Cet argument m'a incité à inclure, pour le jeu de contre-exemples, des séquences de SCOP qui ne sont pas issues du génome humain.

Choix des exemples

Ainsi que je l'ai signalé auparavant, le nombre de cytokines à quatre hélices α connues est de 45. Le nombre d'exemples des jeux d'appren-

tissage est donc limité à 45 au maximum, ce qui peut paraître relativement faible. Une stratégie classique en recherche d'homologues, quand on dispose de peu de données sur la famille d'intérêt, est d'utiliser des orthologues, *i.e.* des membres de la même famille présents dans d'autres génomes, comme source d'information complémentaire. Autant les cytokines à quatre hélices α forment une famille hétérogène, autant les orthologues sont bien conservés chez des espèces voisines de l'homme (souris, rat, chien ...), et partagent une grande similarité. Pour évaluer la possibilité d'utiliser des orthologues dans des jeux d'apprentissage ou de test, j'ai testé les performances des classifieurs avec des jeux de données contenant des orthologues présents dans plusieurs espèces. J'ai tiré au hasard un jeu d'apprentissage et un jeu de test pour lesquels j'ai effectué un apprentissage et une classification. J'ai ensuite ajouté au jeu d'apprentissage les orthologues connus des cytokines présentes dans ce jeu. Ces orthologues proviennent des génomes de la souris, du rat, du chien, du boeuf et du cheval. J'ai procédé à un nouvel apprentissage, avec le jeu d'apprentissage précédent, complété par les orthologues, et à une nouvelle classification sur le jeu de test précédent en utilisant le nouveau modèle ainsi généré. La figure 2.5 résume ce processus. J'ai comparé le classement obtenu avec celui résultant de la classification du même jeu à partir d'un apprentissage sans orthologues.

J'ai calculé la différence de score *ROC* entre les deux classements ainsi que leur taux de Kendall. Ce processus a été répété 100 fois et la moyenne des écarts de scores *ROC* entre les deux classement a été calculée, ainsi que la moyenne des taux de Kendall.
Je présente dans le tableau 2.4 les résultats sur le classifieur Spectrum kernel, les autres classifieurs présentant des écart de *ROC* et des taux de Kendall similaires.

	ecart de *ROC*	taux de Kendall
moyenne	0.006	0.67
ecart type	0.009	0.036

TAB. 2.4 – Comparaison de classements obtenus par apprentissage avec ou sans orthologues

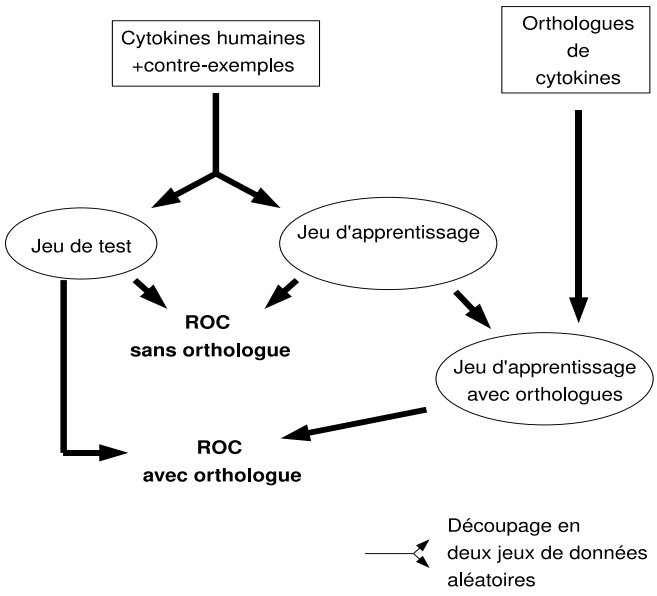

FIG. 2.5 – Création des jeux de données pour évaluer l'intérêt de l'ajout d'orthologues au jeu d'apprentissage

Ces résultats montrent que le gain moyen de score *ROC* par l'ajout d'orthologues est relativement faible, suggérant que ces derniers n'apportent rien au classement global. Le taux de Kendall donne une valeur de 0.67 entre les classements obtenus avec et sans orthologues. Cette valeur assez élevée indique une forte corrélation entre les deux classements et donc peu de différences entre eux.
Je conclus donc que les orthologues apportent peu d'information supplémentaire pour améliorer le classement, pour un coût important en temps de calcul lors de l'apprentissage. Je m'en tiendrai donc aux 45 cytokines humaines comme jeu d'exemple.

2.5.2 Résumé

Cette partie a permis de mettre en évidence les conditions optimales de l'apprentissage. Les fonctions noyaux linéaire et radiale ont des performances comparables pour classer les cytokines connues.
J'emploierai donc la fonction linéaire car cette dernière est plus rapide que la fonction radiale qui, de plus, nécessite d'optimiser ses paramètres.
Pour procéder à un apprentissage optimal, j'ai montré qu'à partir de 100

contre-exemples dans le jeu d'apprentissage, on obtenait une saturation des performances de classification.
L'utilisation de contre-exemples issus d'autres génomes que le génome humain ne semblent pas diminuer les performances de classification mais donne accès à une plus grande diversité de contre-exemples. En revanche, l'enrichissement du jeu d'apprentissage par des orthologues de cytokines n'améliore pas sensiblement les performances de classification.

2.6 Test de performances des classifieurs

Cette partie présente l'évaluation des classifieurs sur la classification des cytokines à quatre hélices α. L'objectif est de déterminer si ces classifieurs sont capables de reconnaître les membres de cette famille auxquels ils n'ont jamais été confrontés auparavant. Je commencerai par expliquer la façon dont j'ai procédé pour effectuer cette évaluation avant de présenter les résultats eux-mêmes ainsi que leur interprétation.

2.6.1 Méthode de création des jeux de données

En partant du jeu de données global de 45 cytokines et 6493 contre-exemples, j'ai aléatoirement partitionné ce jeu en un jeu d'apprentissage (E1) de 15 cytokines et 100 contre-exemples et un jeu de test (E2) de 30 cytokines et de 6393 contre-exemples. La figure 2.6 présente cette répartition. Un apprentissage a été effectué sur le jeu d'apprentissage pour chaque classifieur, en utilisant la fonction noyau linéaire comme indiqué ci-dessus. Pour évaluer les cinq classifieurs que j'ai présenté auparavant, j'ai utilisé les scores ROC et ROC_{200}, décrit ci-avant comme indices de performance. Ce seuil de 200 négatifs autorisé a été choisi pour représenter le nombre total de candidats qu'un biologiste serait capable d'analyser en prenant les candidats par ordre croissant de rang. Compte-tenu du fait qu'il y a 30 cytokines dans le jeu de test, cela signifie que l'on s'intéresse au mieux (*i.e.* si toutes les cytokines sont classées au moins avant le 200^{ieme} négatif) aux

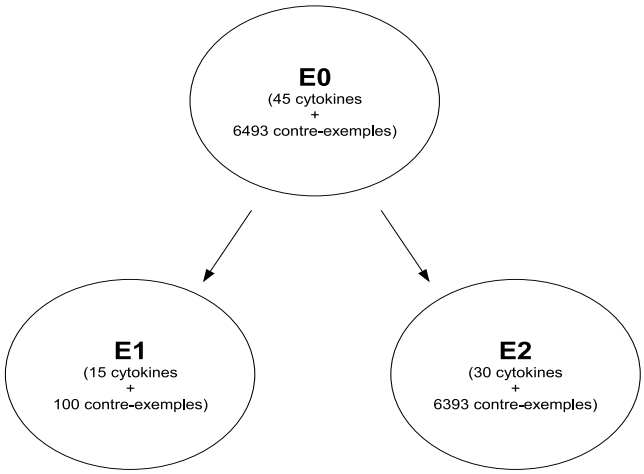

FIG. 2.6 – Répartition des données en jeu d'apprentissage (E1) et jeu de test (E2)

230 premiers candidats. L'ensemble du processus a été réitéré 100 fois afin de constituer 100 fois deux jeux de données et la moyenne de chaque indice de performance a été calculée.

2.6.2 Scores *ROC*

Les résultats des cinq classifieurs sont présentés dans la figure 2.7

Pour chaque mesure, l'écart-type de la distribution de score a été calculé et il est toujours inférieur à 10^{-2}.

Ces résultats montrent que quatre classifieurs parmi les cinq (Spectrum kernel, Mismatch kernel, Pairwise et LA kernel) ont un score *ROC* supérieur à 0.94, le score *ROC* de PairwiseBlast est significativement inférieur (0.85). Lakernel, Mismatch kernel et Spectrum kernel présentent des performances comparables, compte-tenu des écarts-types des distributions de score de ces trois classifieurs. Pairwise présentent des performances légèrement inférieures. Pour chaque classifieur, le score ROC_{200} est nettement mois élevé que le score *ROC* et aucun ROC_{200} n'excède 0.82. LA kernel et Mismatch kernel demeurent les classifieurs les plus performants en ROC_{200}, avec des scores

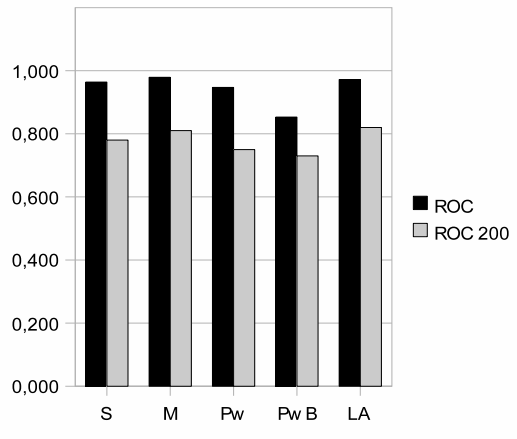

FIG. 2.7 – Comparaison des cinq classifieurs par *ROC* S = Spectrum kernel, M = Mismatch kernel, Pw = Pairwise, PwB = PairwiseBlast, LA = LA kernel

comparables (0.82 et 0.81). Spectrum kernel s'avère légèrement moins performant (0.78) mais surclasse encore Pairwise qui obtient un score ROC_{200} similaire à celui de PairwiseBlast (0.75 et 0.73 respectivement). Pour ce dernier la diminution de performance entre le score *ROC* et le score ROC_{200} semble moins importante (-12%) que pour Spectrum kernel, Mismatch kernel et Pairwise (-21%, -17% et -19%). LA kernel subit une baisse de performance d'environ 15%.

2.6.3 Correlations entre les classifieurs

Ainsi que je l'ai indiqué ci-dessus, le taux de Kendall est utilisé pour mesurer la corrélation entre deux classements. Cette mesure a toutefois une limite importante en ce sens qu'elle ne peut être appliqué sur des classements de grande taille car au-delà d'une certaine limite le taux de Kendall tend rapidement vers 0 du fait des quelques distorsions locales, inévitables dans ce cas. Pour palier à ce problème, j'ai calculé les taux de Kendall de l'ordonnancement des cytokines les unes par rapport aux autres dans les classements des classifieurs. En effet, cette comparaison n'engage que 30 individus (*cf.* "Méthode de création des jeux de données"), ce qui est

une taille raisonnable pour calculer un taux de Kendall. Pour ce faire, j'ai extrais les cytokines des classements totaux en conservant leur rang relatifs les unes par rapport aux autres (*i.e.* si une cytokine est mieux classé qu'une autre dans le classement total, elle le restera dans le classement des cytokines entre elles). J'ai ensuite calculé le taux de Kendall sur ce nouveau classement. Le tableau 2.5 présente les taux de Kendall entre chaque classifieur. On observe que les corrélations entre les différents classifieurs

	S	M	P	PB	LA
S	/	0.712 (0.07)	0.346 (0.103)	0.227 (0.139)	0.352 (0.125)
M		/	0.379 (0.104)	0.241 (0.134)	0.391 (0.134)
P			/	0.491 (0.117)	0.541 (0.098)
PB				/	0.524 (0.127)
LA					/

TAB. 2.5 – Taux de corrélation des classifieurs entre eux. S = Spectrum kernel, M = Mismatch kernel, P = Pairwise PB = PairwiseBlast, LA = LA kernel. L'écart-type des taux de Kendall est précisé entre parenthèses

oscillent entre 0.227 et 0.712, avec des écart-types relativement importants. La corrélation entre Spectrum kernel et Mismatch kernel est la plus importante observée. On remarque que les trois classifieurs basés sur la similarité de séquences (Pairwise, PairwiseBlast et LA kernel) possède également des corrélations élevées entre eux, bien qu'inférieures à la corrélation entre Spectrum kernel et Mismatch kernel.

2.6.4 Discussion

Les résultats de classification montrent clairement qu'au niveau du classement total, les classifieurs possède de bonnes capacités de classification, particulièrement dans le cas de Spectrum kernel, Mismatch kernel et LA kernel. Toutefois ces bonnes performances sont probablement dues à la présence d'un grand nombre de contre-exemples triviaux, donnant une fin de classement très bien classée. Les résultats de score ROC_{200} montrent en effet que la zone de tête du classement est moins bien classée que le classement dans son ensemble. Ceci pose un problème important pour la

suite du processus car cela signifie que parmi les têtes de classement, on risque de trouver un certain nombre de négatifs dont l'analyse nécessaire à les invalider représente un coût important pour le biologiste.
Parmi les classifieurs, LA kernel et Mismatch kernel sont également les classifieurs les plus performants pour la tête du classement. Ce résultat est conforme aux observations de la littérature, où ces deux classifieurs sont considérés comme les plus performants dans leurs catégorie respectives (classifieurs basés sur la composition des séquences et classifieurs basés sur la similarité de séquences), toutefois Mismatch kernel est ici aussi performant que LA kernel, ce qui est un résultat inattendu. On peut supposer que cela est dû au fait que l'information de composition de séquences est aussi efficace à discriminer les cytokines que la similarité, ce qui n'est pas le cas en général.
Spectrum kernel et Pairwise s'avèrent également beaucoup moins efficace sur le classement de tête, indiquant qu'ils doivent globalement bien classer les cytokines mais qu'ils placent un certain nombre de contre-exemples en tête de classement. PairwiseBlast montre une plus faible diminution de performance entre le classement global et le classement de tête. Ceci pourrait signifier qu'il classe relativement mal certaines cytokines mais que les cytokines bien classées sont très proches de la tête de classement.

Les test de corrélation montrent que cette dernière est très forte entre Spectrum kernel et Mismatch kernel et relativement importante entre les classifieurs basés sur la similarité de séquences. La corrélation devient plus faible entre classifieurs basés sur la composition de la séquence et classifieurs basés sur la similarité de séquences, indiquant que ces deux types d'information sont complémentaires. On observe également que toutes les corrélations sont positives, ce qui signifie que ce sont globalement les mêmes cytokines qui sont bien classées par l'ensemble des classifieurs. PairwiseBlast est le classifieur le moins corrélé aux autres, ce qui semble s'accorder avec ses performances moindres.

On peut résumer l'ensemble de ces observations par les trois points suivants :

1. Les classifieurs présentent de bonnes capacités de discrimination globales mais sont beaucoup moins efficaces pour classer la tête de classement. Cette dernière étant l'élément réellement pris en compte par le biologiste, il est important d'améliorer leurs performances sur ce point.
2. LA kernel et Mismatch kernel se présentent comme les meilleurs classifieurs, Spectrum kernel et Pairwise ont également des performances intéressantes mais sont moins stables concernant le passage à la tête de classement. PairwiseBlast est le plus faible des classifieurs mais le plus stable pour la tête de classement.
3. Les tests de corrélation montrent que les classifieurs présente une relative corrélation entre eux, particulièrement si ils ont le même type de méthode de vectorisation, mais ils ne sont pas redondants.

Ces constatations me conduisent à chercher une méthode pour améliorer les performances des classifieurs, particulièrement concernant le classement de tête. Malgré des différences d'efficacité, je n'estime pas judicieux de supprimer des classifieurs, d'une part parce que les corrélations ne sont pas élevées au point de suggérer une redondance, d'autre part parce qu'elles ne sont calculées que sur les cytokines entre-elles, sans préjuger de la possibilité qu'un contre-exemple difficile pour un classifieur ne soit pas facile à éliminer par un classifieur même plus faible.

2.7 Conclusion

J'ai passé en revue dans ce chapitre plusieurs méthodes de recherches d'homologues employées en bioinformatique. J'ai présenté des méthodes basées sur la comparaison de séquences, de structures protéiques, des méthodes hybrides et des techniques d'apprentissage. Parmi ces méthodes, les tech-

niques d'apprentissage sont apparues comme intéressantes, et plus particulièrement les Séparateurs à Vastes Marges (SVM).
J'ai ensuite décrit le fonctionnement des SVM, mettant en évidence qu'en plus de présenter des taux de bonnes classifications élevés, ces méthodes sont conçues pour maximiser leur capacité de généralisation lors de l'apprentissage. Ces éléments expliquent leurs performances. Nous avons vu que, pour un même type d'objet à classer, différents classifieurs pouvaient être développés, en conservant les SVM comme méthode de classification. Cette variété de classifieurs est possible en modifiant le passage de l'objet à classer à un vecteur numérique, utilisé comme entrée par le SVM. Cette phase de vectorisation dépend de l'objet. Des classifieurs spécifiques des séquences biologiques ont été développés par différents auteurs. J'ai présenté cinq classifieurs classiques de la littérature, basés, avec quelques variantes, sur la composition de la séquence en n-uplets (Spectrum kernel, Mismatch kernel) et sur le calcul du score de similarité entre la séquence à vectoriser et le jeu d'apprentissage (Pairwise, PairwiseBlast, LA kernel).
Après avoir discuté de la composition du jeu d'apprentissage et de la fonction noyau à utilisé, j'ai présenté les performances de ces classifieurs sur la classification des cytokines à quatre hélices α. Les cytokines connues et un vaste jeu de contre-exemples, tirés de la base de données SCOP ont été découpés en jeu d'apprentissage et jeu de test. Travaillant sur des classements plutôt que sur une classification binaire, j'ai employé le critère *ROC* pour évaluer les performances. Les classifieurs se sont avérés globalement capables de classer correctement les cytokines, ce qui signifie que les critères choisis (composition des séquences et similarité) sont pertinents pour les cytokines. Toutefois, mes résultats montrent clairement qu'un certain nombre de contre-exemples obtiennent des rangs équivalents aux cytokines. En considérant un nombre de candidats raisonnablement analysable par un biologiste, on constate que le nombre de négatifs bien classés est relativement important, ce qui pose un problème lors de la sélection des candidats destiné à être étudiés expérimentalement.

J'ai également montré qu'il existait une certaine corrélation entre les classifieurs, particulièrement ceux utilisant des caractéristiques similaires, pour le classement de cytokines.
Bien qu'un des classifieurs présente des performances plus faibles que les autres, il n'apparait pas judicieux de l'écarter du processus de classification car on ignore s'il n'apporte pas une information intéressante.

Les classifieurs sont basés sur des critères très généraux, applicables à n'importe quelle famille de protéines et n'utilisent pas l'ensemble des connaissances spécifiques à la famille étudiée. ces critères sont pertinents dans le cas des cytokines, il est toutefois possible d'envisager l'ajout de critères plus spécifiques afin d'améliorer encore les performances d'agrégation, principalement en ce qui concerne le classement de tête. Cette hypothèse sera examinée dans le chapitre suivant.

Chapitre 3

Expertises biologiques automatisées

Ce chapitre décrit un concept original dans le champ de la recherche d'homologues par des méthodes bioinformatiques : la notion d'experts. Je commencerai par expliquer l'intérêt de cette notion dans un processus automatisé avant de la définir plus précisément. Je décrirai ensuite plusieurs experts que j'ai mis au point dans le cadre de la recherche de cytokines dans le génome humain avant de présenter les performances de ces experts seuls, puis associés aux classifieurs SVM.

3.1 Motivation d'une expertise automatisée

J'ai décris dans le chapitre précédent des classifieurs SVM utilisés pour la recherche d'homologues. Le résultat de ces classifieurs est un classement de candidats selon leur probabilité d'appartenance à la classe des cytokines à quatre hélices α. Les classifieurs SVM utilisent certaines caractéristiques des cytokines comme la composition en acides aminés de leurs séquences (Spectrum kernel et Mismatch kernel) ou des scores de similarité entre les cytokines et la séquence candidate (Pairwise, PairwiseBlast, LA kernel).

Malgré les performances intéressantes de ces classifieurs, j'ai pu observer qu'ils plaçaient en tête de classement un grand nombre de contre-exemples. Partant du principe qu'un biologiste ne s'intéressera qu'aux n séquences les mieux classées, il sera confronté à ces contre-exemples et souhaitera les

éliminer. Ce filtrage des candidats en tête de classement, que j'appellerai "expertise", s'effectue à l'aide de caractéristiques spécifiques à la famille et non utilisées par les classifieurs. Ainsi, l'analyse des candidats par un expert humain selon des critères qui lui sont intelligibles est une phase nécessaire de la recherche de nouveaux membres d'une famille de protéines. Les analyses biologiques étant longues et coûteuses, il est indispensable que chaque candidat soit évalué par de multiples critères avant d'y soumettre des candidats.

Dans ce processus, chaque candidat est observé du point de vue de plusieurs critères jugés pertinents. Ces critères sont généralement utilisés comme un "faisceau de présomptions" afin de déclarer si le candidat peut être soumis ou non à des tests biologiques. Certains critères peuvent être rédhibitoires ou au contraire qualifier d'office le candidat. Ainsi un candidat présentant une structure secondaire à quatre hélices α en up-up-down-down serait immédiatement considéré pour des analyses biologiques alors qu'un candidat présentant une structure protéique en feuillets β en serait écarté. Des exemples de critères peuvent être la recherche d'un alignement, manuellement vérifié par l'annotateur, du candidat avec des membres de la famille, la prédiction de sa structure protéique, l'utilisation d'indications sur le candidat telles que des annotations fonctionnelles déjà connues, le type cellulaire où le gène est exprimé, la localisation cellulaire de la protéine ...La conclusion de soumettre ou non un candidat à des analyses biologiques est placé entre les mains de l'annotateur qui s'appuie sur ces indices biologiques.

Ce processus d'expertise manuelle comporte plusieurs limitations. La première, évidente, est qu'un annotateur humain est limité dans le nombre de candidats qu'il peut analyser. Pour illustrer ceci, j'indiquerai simplement que l'analyse des 100 premiers candidats d'un classement effectué par Spectrum kernel sur environ 3 330 000 transcrits humains de la base Unigene [46], m'a demandé un mois de travail. Ceci montre l'ampleur de la tâche de l'annotateur et la quantité forcément restreinte de candidats

qui peuvent être considérés en un temps raisonnable. Ceci d'autant que j'ai précédemment présenté des classifieurs SVM dont un des intérêts était le traitement d'un grand nombre de séquences. La deuxième limitation tient au principe du "faisceau de présomption". L'annotateur n'utilise pas, du fait de la nature des données, un seuil de décision, par ailleurs délicat à déterminer, mais une intuition générale au vu des données. Cette approche intuitive, quoique fort utile dans nombre de problématiques, peut être améliorée par des systèmes d'aide à la décision.
Ces constatations m'ont amené à proposer un système d'expertise automatique ayant pour objectif de faciliter ce travail fastidieux et améliorer les performances des classifieurs, principalement pour la tête de classement.

3.2 Critères d'expertise

Cette partie présente les caractéristiques générales d'un expert automatique et propose quelques exemples d'experts adaptés à la reconnaissance de cytokines putatives.

3.2.1 Choisir des critères d'expertises

La première qualité d'un expert est l'apport d'informations biologiques inexploitées jusqu'ici. Alors que les classifieurs exploitent des caractéristiques particulières, un expert doit enrichir le champ des critères employés pour apporter une diversité d'information, en dépit même de la capacité discriminatrice de celle-ci. D'autre part, l'expertise humaine étant un processus coûteux en temps, il est nécessaire de remplacer celle-ci par une expertise automatique. En outre, l'expert apporte également une information directement interprétable par un biologiste. La notion d'expert que je propose est donc basée sur ces trois impératifs :

- ajout d'information n'ayant pu être intégrée par les classifieurs,
- automatisation possible de la tâche,
- interprétabilité de l'information en termes biologiques.

La première propriété d'un expert est donc l'utilisation d'une information non-exploitée par un classifieur, parce qu'elle n'est pas jugée suffisamment discriminante. Un exemple de ce genre d'information est la taille de la séquence. Dans la partie 1.6, j'ai indiqué que les cytokines à quatre hélices α avaient des tailles de séquences comprises entre 132 et 352 acides aminés, le critère taille de séquence permet donc d'éliminer un candidat ayant une taille significativement différente de celles des cytokines mais pas d'affirmer qu'un candidat ayant une taille comparable à celles des cytokines appartient à cette famille. Cette information peut être exploitée par un expert car ce dernier n'a pas pour vocation d'avoir une grande puissance discriminante mais doit s'ajouter à d'autres experts et classifieurs pour améliorer la classification globale. L'intérêt de l'expert sera alors de légèrement favoriser les cytokines, ou de défavoriser les contre-exemples, de façon à ce qu'elles gagnent suffisamment de rangs pour être mieux classées ces derniers. Par exemple, un expert basé sur la taille diminuera le score global (classifieur+expert) d'un contre-exemple de taille importante alors qu'un expert basé sur la structure protéique augmentera ce même score pour un candidat ayant une structure similaire à celles des cytokines.

La deuxième propriété la plus importante d'un bon critère est son automatisation. *In fine*, c'est la propriété la plus contraignante. Cette propriété exclut entre autre l'utilisation d'outils qui impliquent une interaction avec l'utilisateur lors du traitement. De même, des outils requièrant des calculs longs pour chaque candidat, comme par exemple une prédiction de structure tertiaire, seront écartés. Le format de sortie du critère a une importance. Un résultat sera le plus souvent chiffré (un score, une propriété dénombrable comme une température de dénaturation, un poids moléculaire...) pour pouvoir être traité par l'expert automatique. Il faut donc exclure les procédures appelant un contrôle visuel de l'annotateur.

Enfin, il est souhaitable qu'un expert soit interprétable en termes biologiques, *a contrario* des classifieurs qui sont des "boites noires" pour un expert humain. Bien que cette propriété ne soit pas indispensable, elle per-

met d'expliquer certains choix de classement mais aussi de s'affranchir de l'analyse manuelle des critères utilisés par les experts sur les candidats retenus.

Partant de ces propriétés, il est possible de proposer une méthode de choix de critères sur lesquels reposeront les experts automatiques. En premier lieu, il convient de réunir toutes les informations biologiques connues sur la famille d'intérêt. Puis d'exclure :

- les informations déjà mis en oeuvre par les classifieurs,
- les informations ayant une discrimination quasi-nulle,
- les informations dont la signification biologique n'est pas évidente,
- les informations qui ne peuvent être automatisées, qui n'existent pas systématiquement pour tous les candidats ou qui nécessitent des temps de traitement trop important.

Les critères restants pourront être utilisés pour concevoir des experts automatiques.

3.2.2 Exemples de caractéristiques spécifiques aux cytokines à quatre hélices α

Je présente ici huit caractéristiques biologiques des cytokines à quatre hélices α identifiées en 1.6, pouvant être utilisées pour concevoir des experts.

Structure secondaire

Ainsi que je l'ai expliqué en 1.6.1, la structure secondaire des cytokines est très bien conservée dans cette sous-famille, ce qui en fait une excellent critère pour concevoir un expert. Son utilisation présente toutefois deux difficultés.

La première est que les structures secondaires de toutes les protéines humaines ne sont pas résolues. Pour pallier à cela, il est possible de prédire la structure secondaire à partir de la séquence protéique. Des logiciels tels

que PSI-PRED [30] atteignent des taux de bonne prédiction acceptables. La seconde difficulté provient du fait que la structure secondaire est un élément visuel, qui ne se prête pas facilement à une analyse automatique. Pour contourner cette difficulté, la question "le candidat a-t-il une structure secondaire typique des cytokines à quatre hélices α" peut être reformulée en "A quel point la structure secondaire du candidat est-elle proche de celles des cytokines" ce qui revient à chercher un indice de similarité entre structures secondaires. Il existe peu d'indices répondant à cette question dans la littérature, le plus utilisé d'entre eux étant le Segment OVerlapping (SOV, [61]). Il est principalement employé pour mesurer les différences entre une structure prédite et la structure réelle dans le but d'évaluer un outil de prédiction de structure secondaire. Le SOV se calcule de la manière suivante :

$$SOV = \frac{1}{N} \sum_{i \in [H,E,C]} \sum_{Si} (\frac{MinSov(S1;S2) + \delta(S1;S2)}{MaxSov(S1;S2)).Len(S1)}$$

où : $S1$ et $S2$ représentent les segments de structures prédite et observée dont les acides aminés sont sous l'état structurel i choisi parmi hélice (H), feuillet (E) ou coude (C).
$Len(S1)$ désigne le nombre de résidus de $S1$.
$Minsov(S1;S2)$ désigne la taille du segment chevauchant où $S1$ et $S2$ ont des résidus à l'état i.
$Maxsov(S1;S2)$ désigne la taille maximale d'un segment de S1 ou S2 dont les résidus sont tous à l'état i.
Le schéma 3.1 illustre le $MinSov$ et le $MaxSov$. Le paramètre δ est donné par :

$$\delta(S1;S2) = min \begin{cases} Maxsov(S1;S2) - Minsov(S1;S2) \\ Minsov(S1;S2) \\ \frac{Len(S1)}{2} \\ \frac{Len(S2)}{2} \end{cases}$$

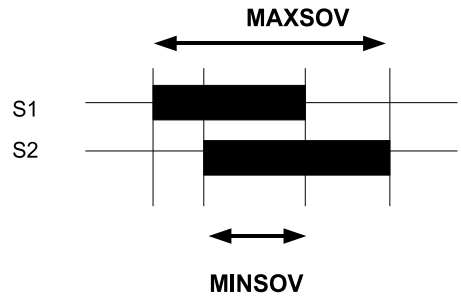

FIG. 3.1 – Représentation du $MinSov$ et du $MaxSov$

N désigne le nombre de résidus dans l'état i dans la structure de plus grande taille.

Le résultat renvoyé est un réel entre 0 et 1 où 0 indique l'absence totale de concordance entre les structures et 1 une concordance parfaite. Cet indice permet de comparer la structure d'un candidat à des structures de cytokines et de déterminer pour chacune leur similarité. Il devient donc possible d'évaluer la proximité de la structure secondaire d'un candidat avec celles des cytokines. Compte tenu de la conservation de la structure secondaire dans cette famille, un candidat ayant un score SOV élevé avec ne serait-ce qu'une cytokine connue aurait de bonnes chances d'appartenir à cette famille. *A contrario* un candidat n'ayant que des scores SOV faibles avec toutes les cytokines aurait peu de chance d'appartenir à cette famille. Cet indice souffre toutefois de deux faiblesses. La première, valable pour tout indice de similarité de structure, est sa dépendance à la qualité de la prédiction. La seconde est due au fait que le SOV se base sur le score d'alignement des deux structures sans directement prendre en compte leurs tailles respectives. Ainsi si une petite protéine ayant une structure en hélice α est comparée à une grande protéine comportant des domaines en hélice α et en feuillet β, le SOV, cherchant à maximiser l'alignement, alignera la petite protéine sur le domaine en hélice de la grande et ces deux structures se verront attribuer un score SOV élevé alors qu'il est facile de les différencier visuellement. Cette technique a toutefois déjà démontré qu'elle pouvait être employée à la recherche d'homologues [17]

Structure génomique

Ainsi que je l'ai mentionné dans la partie 1.6.3, la plupart des gènes de cytokines à 4 hélices α possèdent entre 2 et 6 introns et la structure des gènes est assez bien conservée dans cette famille [4]. Cette caractéristique est d'autant plus intéressante que dans le génome humain un gène comporte en moyenne 8,1 introns ([1]) avec d'importantes variations. Cette spécificité est suffisamment forte pour permettre de définir un expert basé sur le critère "nombre d'introns".
La structure des gènes peut être prédite par des logiciels tels que GENEMARKHMM [37] ou NETGENE2 [7] qui permettent d'automatiser la procédure. Il n'existe pas, comme dans le cas des structures secondaires, d'indice de comparaison de la structure des gènes, toutefois une simple mesure telle que le nombre d'introns dans le gène pourrait servir de critère d'expertise. Des critères plus fin tels que la position sur le gène ou la taille des introns pourraient également être envisagés comme critères.

Taille

Ainsi qu'il a été dit en 1.6.5, les tailles des séquences des cytokines varient entre 132 acides aminés et 352 acides aminés. La taille moyenne se situe à 192,35 acides amminés.
L'utilisation de ce critère peut paraître naïf dans le sens où on ne peut rien affirmer qu'une protéine ayant une taille comparable à celles des cytokines appartienne à cette famille car bon nombre de séquences remplissant cette condition n'en font pas partie. L'utilisation de ce critère ne permet donc pas, dans l'absolu, de réellement discriminer les cytokines mais il offre la possibilité de détecter des faux positifs quand leur taille est en nette discordance avec celles des cytokines. A cela s'ajoute un possible biais des classifieurs vis-à-vis de ce critère. En effet il a été observé que LA kernel et Pairwise ne prenait pas en compte la taille des séquences alignées et que des différences de taille entre ces séquences pouvaient donner lieux à une diminution des performances de ces classifieurs [50]. La prise en considération

de ce critère pourrait donc améliorer les performances globales du système de détection d'homologues.
Cet expert présente également l'avantage d'être facilement automatisable.

Point isoélectrique

La définition donnée dans le chapitre 1.6.5 est que le point isoélectrique est le pH où une protéine est à son état zwitterionique. Cette propriété physico-chimique dépend de la composition en acides aminés dont les chaines latérales sont acides ou basiques, donc de la séquence. On peut raisonnablement penser qu'une famille de protéines ayant toutes le même mode d'action, dans le même milieu, comme les cytokines sera relativement homogène en terme de point isoélectrique. Ce critère offre l'avantage de travailler sur une propriété physico-chimique, en l'occurrence la charge, qui n'est pas prise en compte par les classifieurs.
Ce critère souffre du même défaut que la taille à savoir qu'il ne permet pas réellement d'identifier une cytokine car il est possible qu'une protéine ait un point isoélectrique comparable avec ceux de la famille sans y appartenir. Il est toutefois utilisable pour discriminer négativement des contre-exemples et a également l'avantage d'être facile à automatiser.

Masse moléculaire

Ainsi qu'il l'a été présenté dans le chapitre 1.6.5, la masse moléculaire est définie comme le rapport entre la masse d'une molécule et l'unité de masse des atomes. La masse moléculaire peut être calculée de la façon suivante : $Mm = \sum Ma.In$
où Mm désigne la masse moléculaire, Ma la masse atomique et In l'indice numérique de chaque atome dans la formule brute de la molécule. La masse moléculaire est une propriété physico-chimique des protéines couramment manipulée par un biologiste, elle présente donc une certaine signification. A l'instar du point isoélectrique ou de la taille, la masse moléculaire ne peut discriminer positivement une cytokine, mais peut discriminer négativement

un contre-exemple. C'est également un critère facile à automatiser.

Localisation cellulaire

A l'heure actuelle, toutes les cytokines connues sont des protéines extracellulaires, avec parfois des formes membranaires. Sur le principe, il serait envisageable de diminuer le classement d'un candidat dont la localisation ne serait pas extracellulaire ou, dans une moindre mesure, membranaire. Cette discrimination est possible dans la mesure où il existe des méthodes capables de prédire la localisation d'une séquence avec une bonne précision [13]. La quantification de l'expert serait alors possible par l'utilisation de taux de confiance obtenus par les méthodes de prédiction.

Localisation chromosomique

Ainsi que je l'ai indiqué dans la partie 1.6.4, les gènes de cytokines ont tendance à être organisés en cluster dans le génome humain. Seules deux cytokines sur les 45 connues sont isolées et il existe douze clusters de cytokines dont trois comportent au moins cinq gènes. La localisation chromosomique d'un candidat peut donc être un critère renforçant sa présomption d'être une cytokine.
Ce critère se heurte toutefois à un problème de définition d'un cluster et de quantification de l'expert. Un cluster de gène peut être défini comme un ensemble de gènes situés dans une même région d'un chromosome. On ajoute parfois l'idée que des gènes en cluster partagent certains éléments de régulation transcriptionnelle comme des enhancers. A partir des seules localisations chromosomiques, la question de savoir si deux gènes peuvent être considérés comme faisant partie du même cluster est difficile à trancher car aucun critère de distance n'intervient dans la notion de cluster de gènes. Cette question rend donc délicate l'emploi de ce critère pour concevoir un expert. De plus l'accès à l'information de localisation chromosomique implique la récupération d'informations sur le web, qui ralentissent grandement le traitement, rendant difficile l'emploi de cet expert sur un

grand nombre de séquences.

Présence de cystéines

La présence de cystéines conservées au sein de la séquence des cytokines à quatre hélices α est une caractéristique connue de cette famille. Ces cystéines sont d'autant mieux conservées qu'elles interviennent dans la stabilité de la structure de plusieurs membres de la famille par la formation de deux ponts disulfures. Ces cystéines sont généralement disposées à des positions conservées dans les séquences où elles ont été observées. Cela fait de cette observation un possible critère pour construire un expert. En effet, l'observation de quatre cystéines à des positions similaires à celles des cytokines connues, capables de former deux ponts disulfures chez un candidat renforcerait les présomptions d'appartenance à la famille de ce candidat.
Ce critère est toutefois limité par le manque d'outils capables de prédire avec précision l'existence de ponts disulfures dans une séquence protéique. Quant à la position des cystéines dans la séquence, cette notion pose des difficultés de représentation, similaires à celles évoquées pour la localisation chromosomiques des gènes.

Résumé

Cette partie m'a permis de présenter huit critères biologiques spécifiques aux cytokines ou suceptibles d'éliminer un grands nombre de contre-exemples. Quatre de ces critères (Taille, Structure Secondaire, Point Isoélectrique et Masse Moléculaire) me paraissent particulièrement intéressants et indiqués pour le type de données et d'outils dont je dispose. Je retiens donc ces critères comme suceptibles d'être exploités par des experts.

3.3 Experts automatisés

Après avoir identifié des caractéristiques intéressantes dont l'analyse peut être automatisée, je propose ici une méthode pour construire les ex-

perts basés sur ces dernières. Dans cette partie, je détaillerai les experts issus de certains critères présentés ci-dessus.

3.3.1 Conception générale d'un expert

Les experts pouvant utiliser des informations de natures très différentes, il est intéressant de transformer le résultat de chaque expertise pour le mettre sous une forme commune à tous les experts afin de faciliter leur association avec les classifieurs. A l'image des probabilités générées par ces derniers, on peut définir la probabilité du candidat d'appartenir à la classe d'intérêt selon un critère d'expert.
Disposant des cytokines connues et de SCOP, un jeu de contre-exemples représentatif, il est possible pour n'importe quel critère de calculer sa distribution dans les jeux de cytokines et de contre-exemples. La probabilité pour une séquence inconnue d'appartenir à l'une ou à l'autre de ces distributions peut être obtenue à l'aide de la formule de Bayes :

$$p(f|c) = \frac{p(c|f)p(f)}{p(c)}$$

Où f représente la famille d'intérêt et c le critère utilisé par l'expert (taille, masse moléculaire ...). Ce classifieur bayesien naïf permet de classer un candidat s comme appartenant à f si et seulement si :

$$B(s) = \frac{p(f|c)}{p(\neg f|c)} \geq 1$$

où $p(\neg f|c)$ représente la probabilité que s n'appartienne pas à f. Dans le cas des experts, je n'utilise que le calcul de la probabilité d'appartenance aux cytokines $p(f|c)$, qui représente le score obtenu par cet expert. Ce calcul peut être aisément effectué grâce aux nombreux outils implémentant la notion de classifieur bayesien naïf.

3.3.2 Description des quatre experts spécifiques aux cytokines à quatre hélices α

Expert basé sur la structure secondaire (E-SS)

Pour prédire la structure secondaire, j'ai choisi PSI-PRED car son mode d'utilisation permet une automatisation simple du processus.
J'ai implémenté une version du SOV à partir des sources de l'API QUASAR [5] afin de mesurer l'adéquation de la structure d'une protéine candidate avec celles des cytokines. Pour une classification donnée, les structures des séquences du jeu d'apprentissage ont été prédites par PSIPRED. Les scores SOV de chacune d'entre elles contre les cytokines de ce jeu sont calculés et le score le plus élevé est attribué à cette séquence. Les distributions de scores SOV des cytokines et des contre-exemples sont alors déterminés. De même, lors de la classification du jeu de test, la structure candidate est comparée à l'ensemble des structures de cytokines du jeu d'apprentissage et reçoit comme score, le score SOV le plus élevé obtenu lors de ces comparaisons. Ce score est ensuite utilisé par le classifieur bayésien afin de déterminer une probabilité d'appartenance à la classe des cytokines.

Experts basés sur la taille (E-T), le point isoélectrique (E-PI) et la masse moléculaire (E-MM)

Les valeurs de ces critères sont calculées à partir de la séquence protéique, en utilisant le paquetage proteomics de l'API biojava [1] (pour le point isoélectrique et la masse moléculaire).
Les distributions des valeurs de ces critères pour les cytokines et les contre-exemples sont déterminées et chaque candidat du jeu de test se voit affecter une probabilité d'appartenance à la classe des cytokines comme indiqué plus haut.

[1] http ://biojava.org/wiki/Main_Page

3.4 Evaluation des experts

3.4.1 Données et calculs de probabilités

Les données sont les mêmes que celles utilisées en 2.4.2. Lors de la génération des données, les valeurs des critères utilisés pour chaque expert ont été déterminées pour chaque jeu d'apprentissage et de test. Le calcul des distributions des valeurs de critères pour les cytokines et les contre-exemples et la probabilité d'appartenance à la classe des cytokines a été effectué à l'aide du paquetage bayes, de l'API WEKA [59]. Ces probabilités sont conservées, avec celles des classifieurs pour des traitements ultérieurs.

3.4.2 Méthode d'évaluation

J'ai évalué les performances des experts dans plusieurs situations. Le premier cas concerne simplement l'évaluation des capacités discriminatives de chaque expert considéré isolément, par le calcul du score ROC. J'ai également comparé les taux de Kendall des experts pour mesurer leur corrélation.

Je dispose, comme on l'a vu dans la partie 2.6.1, de 100 couples de jeux d'apprentissage/test. Les jeux d'apprentissage contiennent 15 cytokines et 100 contre-exemples. Les jeux de tests qui leur sont associés contiennent les 30 cytokines restantes et 6393 contre-exemples. Pour évaluer l'efficacité des experts seuls, j'ai procédé comme dans en 2.6.1 à savoir que j'ai calculé les scores ROC et ROC_{200} sur l'ensemble du jeu de test E2.
Afin de faciliter la comparaison des méthodes d'agrégation aux chapitre suivant, j'ai séparé les jeux E2 en deux : un jeu que j'appelle "jeu de paramétrage" (E3) et l'autre que j'appelle "jeu d'évaluation" (E4). Le jeu de paramétrage contient 15 exemples tirés au hasard parmi les 30 du jeu de test et 100 contre-exemples tirés au hasard parmi les 6393 du jeu de test. Le jeu d'évaluation contient le reste, à savoir 15 exemples et 6293 contre-exemples. La figure 3.2 schématise cette répartition des données. Le jeu

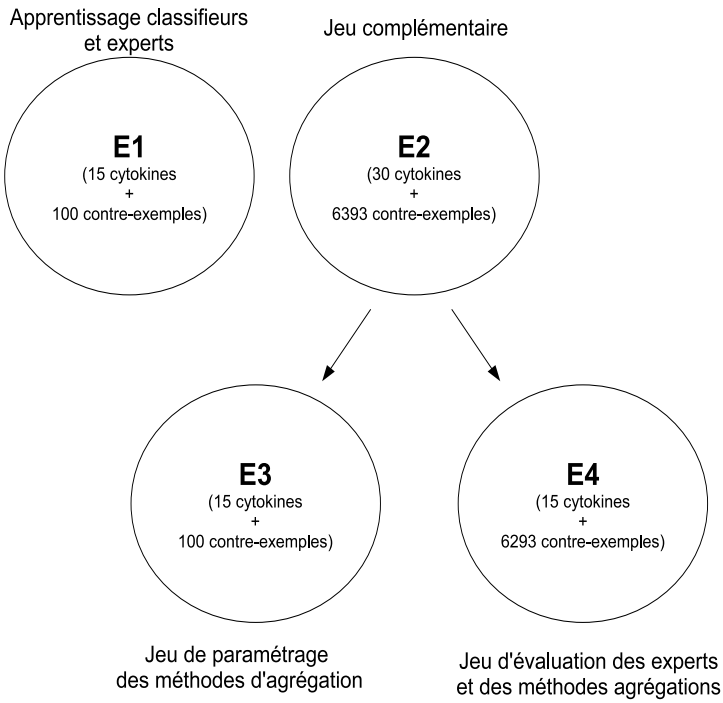

FIG. 3.2 – Répartition des données en jeu de paramétrage (E3) et jeu d'évaluation (E4)

de paramétrage E3 sera utilisé par des méthodes d'agrégation nécessitant un jeu de données connues pour optimiser certains paramètres (*e.g.* les coefficients de pondération d'une moyenne pondérée). Il est ignoré pour les méthodes comme la moyenne non-pondérée qui n'utilise pas de procédure d'optimisation. Ce jeu de données est malgré tout conservé pour ces opérateurs afin de rendre leurs résultats comparables avec ceux des opérateurs pour lesquels E3 est nécessaire.

Pour mesurer l'impact réel des experts sur le classement par les classifieurs, je me suis placé dans le cas où les classifieurs ont des performances de classification faibles. Pour ce faire, j'ai recherché les contre-exemples les mieux classés par les classifieurs (*i.e.* les contre-exemples les plus difficiles pour chaque classifieur).

Ces contre-exemples sont obtenus selon le principe décrit dans l'algorithme suivant :
Soit k le nombre de classifieurs, nb le nombre courant de contre-exemples et c_y classement du classifieur y.
liste := $\oslash$ //liste résultat
rang = $(0, \ldots, 0)$ // vecteur nul de taille k
nb := 0 // nombre de contre-exemples courant

tant que $nb < 200$ faire
    pour $y = 0$ à k faire
        candidat := cherche(c_y, rang[y])
        liste := liste $\{\cup$ candidat.id$\}$
        rang[y] := candidat.rang
        $nb = nb + 1$
    fin pour
fin tant que

La méthode cherche(c_y, r) récupère le premier contre-exemple, absent de la liste finale, du le classement c_y tel que sont rang est supérieur ou égal à r. Elle renvoie son identifiant (id) et son rang (rg) dans le classement c_y
La figure 3.3 illustre la sélection des contre-exemples. Ce processus permet de constituer une liste de contre-exemples ayant des probabilités d'appartenance à la classe des cytokines élevées selon les classifieurs, *i.e.* des négatifs difficiles à discriminer. J'ai fixé à 200 le nombre de contre-exemples ainsi sélectionnés. Ces contre-exemples sont ajoutés aux 15 exemples du jeu d'évaluation, pour former un jeu d'évaluation dit "difficile". C'est sur ce jeu d'évaluation que sont déterminés les scores *ROC* des experts et de l'agrégation avec les classifieurs. Ce processus de création d'un jeu d'évaluation difficile est répété pour chaque jeu d'évaluation, soit 100 fois. Pour agréger les différentes méthodes (classifieurs et experts) J'ai utilisé la moyenne arithmétique. Cette méthode peut paraître simpliste de prime

abord, mais se justifie par la volonté de mesurer l'impact réel des experts sur le classement. La moyenne arithmétique attribue le même coefficient à toutes ses composantes, ce qui permet d'observer réellement l'impact de chaque expert puisque tous ont une contribution égale, entre eux et avec les classifieurs. Compte tenu du fait que l'on travaille sur des probabilités pour les classifieurs et les experts, la stratégie d'agrégation consistant à calculer le produit de ces probabilités peut paraître plus naturel. Il s'avère toutefois que cet opérateur est moins optimal que la moyenne non-pondérée.

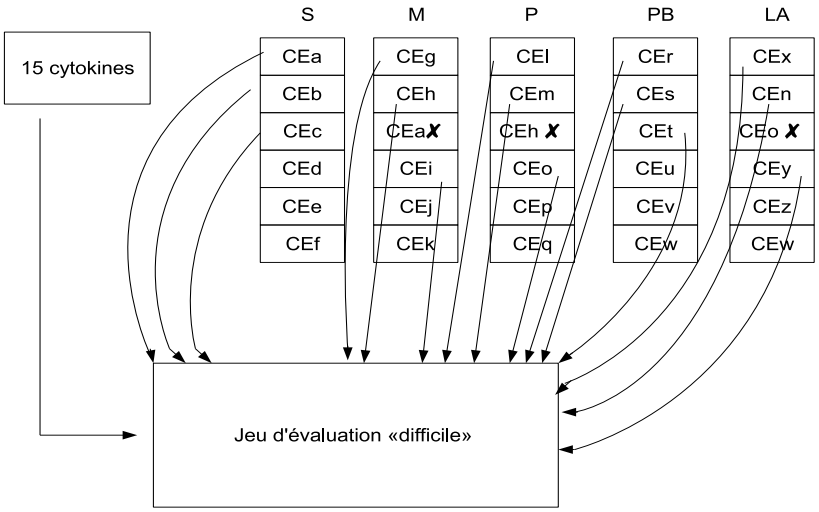

FIG. 3.3 – Exemple de constitution d'un jeu d'évaluation difficile de 15 cytokines et 15 contre-exemples. S = Spectrum Kernel, M = Mismatch Kernel, P = Pairwise, PN = PairwiseBlast, LA = LA kernel. Les contre-exemples, classés par probabilité d'appartenance aux cytokines pour chaque classifieur, sont sélectionné succesivement en partant de la colonne S vers la colonne LA. Dans la colonne M, le candidat CEa n'est pas retenu car il a déjà été sélectionné en première ligne de la colonne S. Le contre-exemples CEi est alors sélectionné Il en va de même pour les contre-exemples CEh, dans la colonne P et CEo dans la colonne LA.

3.4.3 Performances des experts seuls

Le tableau 3.1 présente les scores ROC et ROC_{200} moyens des classements effectués par les experts seuls. On observe que les experts ont des scores ROC et ROC_{200} inférieurs aux classifieurs, confirmant que leur ca-

Experts	E-T	E-SS	E-PI	E-MM
ROC	0.834	0.677	0.895	0.839
ROC_{200}	0.338	0.431	0.858	0.371

TAB. 3.1 – score *ROC* et ROC_{200} moyens des classements obtenu en utilisant les experts comme classifieurs

pacité discriminante est moindre. Ceci est particulièrement frappant pour le ROC_{200} où les experts, à l'exception de l'expert E-PI, obtiennent des scores inférieurs à 0.5.
En terme de classement global, E-PI obtient le meilleur score, les experts E-T et E-MM obtiennent des scores comparables et l'expert E-SS s'avère le moins performant.
En terme de tête de classement, E-PI demeure le plus performant, et présente une faible diminution de score entre le *ROC* et le ROC_{200}, ce qui signifierait que le point isoélectrique est un critère discriminant et stable quand on considère la tête de classement. E-SS obtient le deuxième meilleur score ROC_{200}, E-T et E-MM sont les moins performants, bien que E-MM obtienne un score ROC_{200} légèrement supérieur à celui de E-T. Les résultats très similaires de ces deux experts semblent montrer une certaine corrélation entre eux.

3.4.4 Corrélation entre experts

Les résultats précédents posant la question d'une corrélation entre certains experts, j'ai investigué cette possibilité par l'utilisation du taux de Kendall. Comme décrit dans la partie 2.6.3 pour les classifieurs, j'ai calculé les taux de Kendall du classement des cytokines par experts. Le tableau 3.2 présente ces taux de Kendall pour chaque couple d'experts. On observe que les taux de Kendall de tous les couples, excepté le couple E-T/E-MM, sont proches de 0. Les écart-types importants indiquent que ces taux peuvent varier grandement d'un jeu de données à l'autre sans toutefois réellement s'éloigner de 0. Le taux de Kendall entre E-T et E-MM est en revanche très élevé, indiquant une certaine redondance entre ces deux experts.

	E-T	E-SS	E-PI	E-MM
E-T	/	0.02 (0.082)	-0.052 (0.078)	0.827 (0.048)
E-SS		/	-0.029 (0.074)	0.05 (0.082)
E-PI			/	-0.091 (0.081)

TAB. 3.2 – Taux de Kendall moyen des experts deux à deux. Les écart-types sont indiqus entre parenthèses

3.4.5 Agrégation avec les classifieurs

Pour évaluer l'efficacité de l'agrégation par rapport à l'utilisation des classifieurs seuls, j'ai mesuré la différence, en terme de scores ROC, entre cette dernière et le meilleur des classifieurs.
Soit $S(meill.classif.)$, le score ROC le plus élevé des scores obtenus pour les classifieurs SVM. Le score ROC de l'agrégation des classifieurs avec un ou plusieurs experts est appelé $S(agreg)$. Le gain, G obtenu par l'agrégation avec un ou plusieurs experts est défini comme :
$G = S(agreg) - S(meill.classif)$
Pour chaque jeu d'évaluation difficile, j'ai calculé le gain G. J'ai ensuite déterminé la moyenne des G obtenus sur les 100 jeux d'évaluation. Soit δROC cette moyenne.
Je présente dans le tableau 3.3, le δROC pour les agrégations des cinq classifieurs SVM avec plusieurs combinaisons d'experts sur les jeux d'évaluation difficiles.

Plusieurs observations ressortent de ces résultats. La première est que plusieurs associations entre experts et classifieurs obtiennent des δROC positifs. On constate également que ces δROC sont d'autant plus importants que le nombre d'experts utilisé augmente. Dans le même ordre d'idée, le δROC de l'association de classifieurs seuls est le δROC le plus faible obtenu.
Considérant les effets de l'ajout ou de l'absence d'un expert dans une association, on peut tirer des conclusions concernant l'importance de chacun. L'expert E-SS est le plus efficace puisque chaque association où il est présent obtient des δROC élevés par rapport aux associations contenant

S	M	P	PB	LA	E-T	E-SS	E-PI	E-MM	δROC
⋆	⋆	⋆	⋆	⋆	⋆	⋆	⋆	⋆	0.054
⋆	⋆	⋆	⋆	⋆	⋆	⋆	⋆		0.045
⋆	⋆	⋆	⋆	⋆		⋆	⋆	⋆	0.042
⋆	⋆	⋆	⋆	⋆	⋆	⋆		⋆	0.042
⋆	⋆	⋆	⋆	⋆		⋆		⋆	0.037
⋆	⋆	⋆	⋆	⋆	⋆	⋆			0.035
⋆	⋆	⋆	⋆	⋆		⋆	⋆		0.026
⋆	⋆	⋆	⋆	⋆		⋆			0.022
⋆	⋆	⋆	⋆	⋆	⋆		⋆	⋆	0.009
⋆	⋆	⋆	⋆	⋆	⋆			⋆	-0.002
⋆	⋆	⋆	⋆	⋆	⋆		⋆		-0.006
⋆	⋆	⋆	⋆	⋆				⋆	-0.013
⋆	⋆	⋆	⋆	⋆			⋆	⋆	-0.016
⋆	⋆	⋆	⋆	⋆	⋆				-0.016
⋆	⋆	⋆	⋆	⋆			⋆		-0.038
⋆	⋆	⋆	⋆	⋆					-0.061

TAB. 3.3 – δROC de l'agrégation par la méthode de la moyenne non-pondérée S = Spectrum kernel, M = Mismatch kernel, P = Pairwise, PB = PairwiseBlast, LA = LA kernel

le même nombre d'experts. Plus notable encore, l'association de E-SS avec les classifieurs obtient un δROC 2,4 fois plus important que l'association des trois autres experts avec les classifieurs.
Le deuxième expert le plus intéressant est E-T, ainsi que le montrent plusieurs comparaisons d'association où il est absent avec celles où il est présent. E-MM lui est légèrement inférieur pour plusieurs associations équivalentes, bien que les deux experts semblent avoir un impact comparable. E-PI s'avère clairement le moins efficace des experts en terme de δROC mais apporte tout de même une information positive.

3.4.6 Discussion

Les résultats d'évaluation des experts seuls en ROC et ROC_{200}, de leur taux de Kendall ainsi que la comparaison entre le score ROC du meilleur classifieur et le score ROC des différentes agrégations sur un jeu d'évaluation difficile permettent de tirer plusieurs conclusions.

Performances des experts seuls

La première de ces conclusion est que trois des quatre experts (E-T, E-SS et E-MM) ont une capacité discriminante faible alors que E-PI possède une capacité discriminante importante, y compris pour la tête de classement. Ce résultat est doublement étonnant. D'une part parce que E-PI utilise comme critère le point isoélectrique qui *a priori* devrait permettre d'éliminer des contre-exemples très différents des cytokines mais pas de discriminer positivement les cytokines. On pouvait donc s'attendre à ce qu'un certain nombre de contre-exemples possèdent un point isoélectrique comparable aux cytokines et soit aussi bien classées que ces dernières. Curieusement, les résultats indiquent que le point isoélectrique est un critère assez discriminant, ce qui signifierait que les cytokines forment un groupe assez homogène de ce point de vue.
Au contraire, on aurait pu s'attendre à ce que E-SS soit un critère efficace pour discriminer positivement les cytokines or il s'agit de l'expert le moins discriminant au niveau du classement global et, bien qu'il surclasse E-T et E-MM en terme de classement de tête, son score ROC_{200} reste faible. Une première explication serait que les prédictions de structures secondaires sont entachées d'erreurs, conduisant à un SOV qui ne correspond pas à la réalité. Une seconde, complémentaire de la première, est que le critère de SOV lui même pourrait être biaisé, en attribuant des scores élevés aux structures très similaires aux cytokines mais ne pourrait reconnaître des structures un peu plus éloignées, leur attribuant un score proportionnellement très faible. Les structures très similaires à celles du jeu d'apprentissage seraient donc avantagées mais les cytokines ayant une structure moins similaire seraient rapidement reléguées au niveau des contre-exemples. Cette explication concorde avec l'utilisation classique du SOV à savoir la comparaison entre la structure connue d'une protéine et une prédiction de cette structure, pour évaluer des logiciels de prédiction.
Les performances générales de E-T et E-MM confirment que ces experts apportent une information capable de différencier globalement cytokines et

contre-exemples. L'observation du classement de tête montre toutefois que de nombreux contre-exemples obtiennent des scores comparables aux cytokines, ce qui confirme que ces exeprts ne peuvent discriminer positivement ces dernières.

Performances des experts associés aux classifieurs

La deuxième conclusion importante que l'on peut tirer est que la stratégie d'ajout des experts améliore la capacité de discrimination du système par rapport au meilleur classifieur mais aussi par rapport à l'association des cinq classifieurs uniquement, comme l'indique les résultats de δROC des classifieurs seuls ou associés à des experts. L'apport d'information par les experts est donc, comme prévu, bénéfique pour classer les cytokines.
Il est intéressant de noter que les performances d'agrégation des experts avec les classifieurs présente un profil opposé aux résultats de classifications de ces experts, à savoir que E-SS est de loin l'expert le plus important pour l'agrégation alors que E-PI est le moins intéressant.
Le calcul des taux de Kendall entre les experts ont, quant à eux, soulignés une forte corrélation entre les experts E-T et E-MM. Ces résultats concordent avec le fait que les critères qu'ils utilisent peuvent être corrélés et avec leurs performances, tant comme experts seuls qu'associés aux autres experts et aux classifieurs. Ces derniers résultats suggèrent toutefois que la combinaison de ces deux experts permettrait de maximiser l'agrégation, comme l'atteste le fait que l'association des quatre experts avec les classifieurs est requise pour obtenir la meilleure performance. Il ne parait donc pas judicieux d'écarter *a priori* un de ces deux experts.

3.5 Conclusion

J'ai présenté ici la notion d'expert automatique, destiné à améliorer la classification des séquences biologiques. Cette idée répond à deux nécessités bien distinctes : augmenter les performances des classifieurs par l'ajout d'informations non-exploitées et automatiser le mécanisme d'expertise hu-

maine. Ce concept d'expert automatique permet également d'ajouter une information biologique intelligible.

Dans ce chapitre, j'ai présenté plusieurs exemples de critères, spécifiques aux cytokines à quatre hélices α, pour illustrer cette notion. Parmi ces critères, quatre m'ont paru utilisables pour mettre au point des experts automatiques : la taille de la séquence, la structure secondaire, le point isoélectrique et la masse moléculaire de la protéine.

J'ai proposé une technique basée sur une approche bayesienne pour calculer une probabilité d'appartenance à la famille d'intérêt à partir d'un critère. Cette approche a le mérite de renvoyer une grandeur congruente avec celle obtenue par les classifieurs, assurant une homogénéité du système. J'ai ensuite présenté quatre experts basés sur les critères retenus.

Par les même méthodes que pour les classifieurs, j'ai procédé à l'évaluation des experts proposés. Les évaluations des experts seuls ont montré que ces derniers ne présentaient pas, à l'exception de E-PI, de capacité particulière de discrimination.
Le test de corrélation a montré que E-T et E-MM présentaientt une forte corrélation. Les deux autres experts semblent indépendants entre eux et avec ces deux derniers.
Les tests de combinaison de classifieurs avec les experts ont montré que l'ajout des experts améliorait les capacités de classement du système par rapport au meilleur d'entre les classifieurs ou à une combinaison des classifieurs seuls. Certains experts sont plus importants que d'autres dans cette combinaison. Ainsi, E-SS s'est avérer particulièrement efficace lors de l'agrégation alors que les performances de E-PI sont plus faibles qu'on ne pouvait l'attendre au vu de sa capacité de discrimination. Il semble toutefois qu'aucun expert ne puisse être éliminé sans affecter le résultat de l'agrégation.

La fonction d'agrégation que j'ai utilisée étant par nature très simple, il semble judicieux d'envisager l'utilisation d'autres fonctions d'agrégation plus riches. Cette idée fera l'objet du chapitre suivant.

Chapitre 4

Agrégation de classifieurs

L'évaluation des classifieurs et des experts a montré que, malgré leur efficacité, aucun d'entre eux n'était capable de classer parfaitement les cytokines au milieu d'un vaste jeu de contre-exemples. Combiner leurs informations pourrait donner lieu à une amélioration sensible des performances de classification. Cette stratégie de fusion de méthodes est assez courante en bioinformatique. Dans cette partie, je présenterai tout d'abord les méthodes classiques d'agrégation utilisées en bioinformatique ou dans d'autres disciplines. Je décrirai ensuite différents opérateurs d'agrégation qui seront évalués par les méthodes utilisées dans les chapitres précédents.

4.1 Etat de l'art

Dans cette partie, je présente les méthodes d'agrégation, au sens général, que l'on trouve dans la littérature ainsi que quelques applications connues. Je commencerai par présenter les différents types de méthodes, selon qu'elles sont utilisées sur les classifieurs ou leurs données, avant de discuter de leur applicabilité au problème d'agrégation des classifieurs pour la recherche d'homologues. Dans cette partie et dans la description des méthodes d'agrégations retenues, les experts décrits au chapitre précédent sont assimilés à des classifieurs.

4.1.1 Généralités

J'ai choisi d'exposer d'une manière générale les méthodes d'agrégation avant de détailler les méthodes qui sont applicables aux classifieurs décrits précédemment et utilisées pour résoudre mon problème.
J'utilise le vocable "agrégation de classifieurs" pour désigner des méthodes apportant une réponse unique à un problème de classification en combinant plusieurs classifieurs et/ou les résultats de ces classifieurs. Il existe également des méthodes d'agrégation qui opèrent sur l'architecture des classifieurs ou leur sélection, ces méthodes seront dénommées "méthodes de structuration/sélection des classifieurs" par opposition au vocable précédent, que je conserve pour un usage général.

4.1.2 Agregation de méthodes en bioinformatique

La notion d'agrégation de méthodes en bioinformatique n'est pas une idée neuve. Combiner plusieurs méthodes de détection pour améliorer les performances du système est une pratique courante, proposée en parallèle au développement de nouvelles stratégies. Un exemple récent d'application de ce principe à la recherche d'homologues est le logiciel CHASE [2] qui combine différentes techniques de recherche (HMM, PSIBLAST, MAST ...), chacune renvoyant une e-value des meilleurs candidats. Après normalisation pour les rendre comparable entre elles, ces e-values sont agrégées à l'aide d'une somme pondérée, dont les coefficients sont calculés en fonction des performances de chaque méthode, déterminées par le système PHASE 4 [2].

4.1.3 Décison multicritère

Le champ de la décision multicritère est un champ de l'informatique qui trouve écho dans un grand nombre d'applications réelles, particulièrement dans le domaine du management et de la gestion de ressources où les processus de décision automatiques abondent. Ce domaine, très ancien

(on peut remonter aux travaux de Condorcet, en 1785, sur le système de vote qui portera son nom), s'intéresse à la recherche de solutions pour des problèmes contenant plusieurs variables. En ce sens il rejoint l'agrégation de classifieurs qui peut être vue comme la recherche d'un moyen pour obtenir le ou les objets (assimilables à des solutions) les plus proches de la classe d'intérêt, compte tenu des résultats des classifieurs (assimilables à des critères). Concrètement, on dispose de k critères d'intérêt et x solutions comportant des valeurs pour ces critères, l'objectif étant de trouver la meilleure solution, c'est à dire celle qui répond le mieux aux k critères simultanément. La difficulté réside dans le fait qu'il n'y existe généralement aucune solution qui maximise simultanément les k critères, il faut alors sélectionner une solution sous-optimale, mais supérieure aux autres. Plusieurs stratégies s'offrent alors.

Classement par importance des critères

Une idée intuitive consiste simplement à ordonner les critères par importance puis à classer les solutions selon le nombre et l'importance des critères qu'elles vérifient. Une solution sera d'autant mieux classée qu'elle vérifie un grand nombre de critères importants. Cette approche a pour principale faiblesse la nécessité de classer les critères par ordre d'importance et l'attribution à chacun d'un poids relatif à son importance.

Ensemble de Pareto

Ce concept, développé par Vilfredo Pareto à la fin du XIXème sciècle, tente d'apporter une réponse rationnelle à la question du choix d'une ou plusieurs solutions parmi un ensemble lorsqu'une décision est basée sur de multiples critères. Une présentation de cette notion peut être trouvée entre autre dans [62].
Pour expliciter ce concept, il faut d'abord définir celui de la dominance d'une solution sur une autre. Une solution X est dite dominante sur une solution Y si et seulement si

$\forall i \in (1, \ldots, n), X_i > Y_i$

où X_i et Y_i désignent la valeur du critère i pour les solutions X et Y respectivement et n désigne le nombre de critères considérés. Une fois ce socle posé, il est possible de définir l'ensemble de Pareto comme l'ensemble des solutions non-dominées, c'est à dire qu'elles dominent ou sont égales à toutes les autres solutions pour au moins un critère. L'ensemble de Pareto est rarement composé d'une seule solution, dominant toutes les autres mais dans le cas d'une décision multicritère, cet ensemble a la propriété de contenir la solution optimale en permettant en général une réduction de l'espace de recherche. Il est toutefois nécessaire d'utiliser d'autres méthodes pour identifier cette solution optimale.

Le vote

Une procédure classique en décision, largement utilisée ne serait-ce que dans les systèmes politiques, est le vote. Un vote peut être définit comme l'ordonnancement des solutions selon le nombre de voix accordé par les votants à chaque solution existante. Il existe un grand nombre de procédures de vote, chacune possédant des propriétés propres. Pour illustrer cette multitude, je citerai simplement [14] où les auteurs mentionnent l'existence de 27 systèmes démocratiques dans le monde où 70 procédures différentes de vote ont été appliquées entre 1945 et 1990. Comme exemples de système de vote, on peut citer :

- le vote par majorité : le cas le plus simple de vote où les solutions sont classées selon le nombre de voix qu'elles ont obtenues.
- le vote à majorité absolue : une seule solution est retenue : celle qui reçoit la moitié des voix plus une.
- le vote par simple transférabilité : les solutions obtenant un certain nombre de voix sont retenues. Les voix leur permettant de dépasser ce seuil sont transférées aux autres solutions en fonction d'un vote de "seconde main" des votants. La procédure est répétée jusqu'à ce qu'il n'y ait plus de voix à redistribuer.

– le vote cumulatif : On détermine un nombre de solutions à retenir et chaque votant dispose d'autant de voix. Ils peuvent les distribuer comme bon leur semble entre les solutions, y compris en les cumulant sur certaines d'entre-elles, les solutions retenues étant celles disposant du plus grand nombre de voix.

Ces quelques exemples permettent d'illustrer les différentes configurations de vote existantes, c'est à dire le classement des solutions ou la désignation d'une ou plusieurs d'entre elles. Dans le problème qui nous occupe les méthodes aboutissant à un classement sont celles qui nous intéressent, bien que les méthodes capables de désigner plusieurs candidats soient également applicables si le nombre de candidats à retenir est fixé *a priori* par le biologiste. Il faut toutefois noter que pour pouvoir classer un grand nombre de solutions selon le total des voix recueillies par chacune, il est nécessaire de disposer d'un grand nombre de votants. Si ce nombre est trop faible, la granularité du classement ne sera pas suffisante pour départager les candidats.

La principale limite de ces méthodes de vote est qu'une mauvaise solution peut être favorisée par plusieurs votants commettant le même type d'erreur. *A contrario* un classifieur peut être particulièrement efficace mais isolé par rapport aux autres votants, donc incapable de favoriser suffisamment les bonnes solutions. La sélection et la pondération de chaque votant est donc un problème clé dans les processus de vote.

4.1.4 Agrégation par structuration/sélection de classifieurs

Une stratégie classique en fusion de classifieurs est d'opérer sur l'architecture des classifieurs eux-mêmes afin d'obtenir directement un résultat agrégé. En première intention, une architecture en parallèle est souvent utilisée avant de fusionner les classifieurs ou leurs résultats, mais plusieurs auteurs ont proposé des méthodes tirant partie de constructions plus complexes. L'agrégation peut également prendre la forme d'une sélection du classifieur le plus efficace.

Sélection du meilleur classifieur

La stratégie d'organisation des classifieurs la plus simple, consiste en effet à désigner un unique classifieur dont le résultat tiendra lieu de résultat final. Ceci peut être fait à partir des performances des classifieurs ou à partir de considérations basées sur la connaissance des données, comme par exemple une sélection de type min/max présentées plus loin. Le schéma 4.1 illustre cette notion.

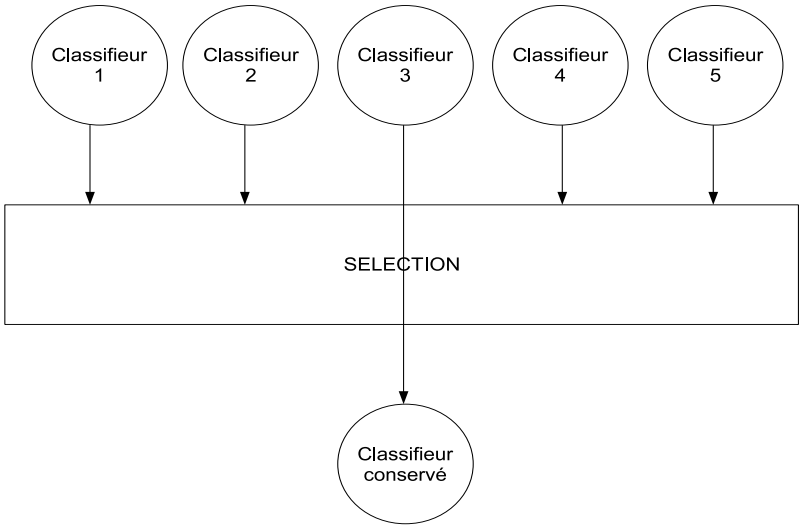

FIG. 4.1 – Sélection de classifieurs

Regroupement de classifieurs

Une autre approche consiste à structurer les classifieurs en groupes et d'appliquer des méthodes d'agrégation différentes pour chaque groupe avant de procéder à l'agrégation finale. Ces méthodes d'agrégation peuvent aussi bien être de la fusion/sélection de classifieurs que de l'agrégation de résultats. Ruta et Gabrys [49] suggèrent que la technique de regroupement de classifieurs sera d'autant plus efficace que les groupes sont constitués de classifieurs utilisant des caractéristiques différentes. L'idée étant de réduire l'erreur de classification globale en augmentant la diversité des erreurs, qui ainsi ne s'additionnent pas lors de l'agrégation finale.

Architecture hiérarchique

L'idée du regroupement de classifieurs peut être poussée encore plus loin par la mise en place d'une architecture hiérarchique de ces classifieurs. Ainsi, les classifieurs (ou les groupes de classifieurs) sont placés en série de façon à ce que les résultats des niveaux précédents soient utilisés par les niveaux suivants. Il est par exemple possible d'imaginer une succession de classifieurs en cascade, voire en structure pyramidale, où le dernier niveau contient un seul classifieur, utilisant les résultats des niveaux précédents comme données à classer. La difficulté de cette technique consiste à optimiser l'ordre des classifieurs pour maximiser les performances de classification. Ceci peut être fait entre autres par des méthodes d'optimisation comme celles décrites plus loin. La figure 4.2 schématise la notion de classifieurs en série.

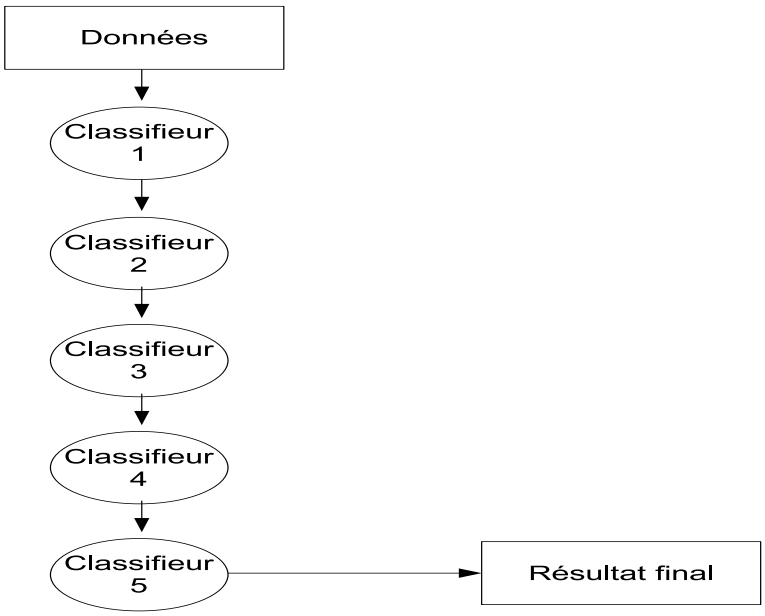

FIG. 4.2 – Cinq classifieurs en série

Arbre de décision

Réemployant le principe de la classification hiérarchique décrit précédemment, les arbres de décision [47] proposent une vision orientée "expertise" de l'agrégation. Dans un arbre de décision, les critères (ici les classifieurs)

sont placés en ordre hiérarchique mais les données en sortie sont traitées en fonctions du résultat obtenu, au lieu d'être envoyées directement au niveau suivant. Ainsi, par exemple, selon la probabilité obtenue par un classifieur bayesien, un candidat pourra être envoyé à un classifieur SVM, à une méthode de traitement des données (réduction des caractéristiques par exemple) ou simplement éliminé. Le cas le plus simple d'arbre de décision est une hiérarchie de classifieurs où chacun élimine les candidats dont le résultat de classification est inférieur (ou supérieur) à une valeur seuil, déterminée pour chaque classifieur. Cette stratégie offre une souplesse de traitement encore plus grande que les hiérarchies décrites précédemment, mais aussi un mécanisme plus complexe d'implémentation et de choix de l'enchaînement optimal de classifieurs.

4.1.5 Agrégation des résultats des classifieurs

Cette stratégie, plus intuitive, consiste à agréger les résultats de classifieurs en parallèle. Plusieurs méthodes sont envisageables, selon le type des données à agréger.

Agrégation de labels

Dans le cas où l'on s'intéresse uniquement aux labels des candidats, il existe plusieurs techniques d'agrégation des résultats des classifieurs.
La principale stratégie est la famille des méthodes de vote, décrites précédemment, qui retient comme label final au candidat le label le plus fréquemment attribuer par les classifieurs. Cette famille de méthodes est une des plus simples à mettre en place, mais pose des difficultés s'il existe un biais entre les votes *e.g.* deux classifieurs basés sur des critères similaires, ce qui a pour conséquence d'augmenter artificiellement le poids du critère.
Une autre méthode appelée "behaviour-knowledge space method" test toutes les combinaison de classifieurs sur différents échantillons du jeu d'apprentissage et conserve ces résultats dans une table. Pour chaque cellule de cette table, le label majoritaire est déterminé. Pour classer un nouvel exemple,

on recherche la cellule qui se rapproche le plus de cet exemple et on lui attribut le label majoritaire. Cette méthode, bien que plus complexe à mettre en place, tire partie de l'apprentissage et n'implique pas l'indépendance des classifieurs, contrairement aux agrégations basées sur un vote.

Agrégation de classement des classes possibles

Ces techniques s'utilisent sur un cas un peu particulier de données, celui de l'existence de plus de deux classes et des classifieurs renvoyant, pour chaque candidat, un classement des labels possibles, du plus probable au moins probable. Ce type de données étant peu utilisé, particulièrement dans le cas de la recherche d'homologues où l'on ne travaille en général que sur deux classes, je n'entrerai pas dans les détails et signalerai simplement l'existence de deux méthodes : la méthode du plus haut rang (qui consiste simplement à appliquer un *min* des rangs des classifieurs sur chaque classe possible pour le candidat et à réordonner ainsi les différentes classes pour choisir la première du nouveau classement) et l'utilisation d'une extension du vote majoritaire appelée "comptage de Borda", du nom du mathématicien français, contemporain de Condorcet, qui mit au point cette méthode.

Agrégation basée sur la mesure d'appartenance à une classe

Par le terme "mesure d'appartenance à une classe", j'entends toute valeur, généralement comprise entre 0 et 1, représentant un score ou une probabilité d'appartenance à une classe. Pour ce genre de données, qui représente le cas le plus général, il est possible de revenir aux techniques décrites précédemment (moyennant une perte d'information) mais également d'en développer d'autres.
On peut dégager un premier groupe de méthodes d'agrégation, que je désignerai par l'expression "méthodes bayesiennes", s'appuyant la combinaison de probabilités conditionnelles provenant des classifieurs.

On trouve également des méthodes évolutives, ayant pour objectif d'optimiser les paramètres d'agrégation. Cette optimisation peut se présenter sous la forme d'une attribution de poids à chaque classifieur dans des méthodes telles que les procédures de vote ou la moyenne pondérée, ou en combinant au mieux des opérateurs simples afin de creer une procédure d'agrégation *ad hoc*.

Il convient également de citer l'usage de "méta-classifieurs" tels que des méta-résaux de neurones ou des métaSVM. Le principe d'un méta-classifieur est d'utiliser les valeurs obtenues des classifieurs comme entrées pour opérer lui-même à une classification. Le résultat peut être donné soit sous la forme d'un label pour le candidat, ou mieux, sous la forme d'un score/probabilité, utilisé pour ordonner les candidats.

Enfin, un autre groupe de méthodes, s'appuyant sur la logique floue (une extension de la logique classique, applicable à des valeurs autres que booléennes), permet d'agréger des classifieurs. Des applications aussi simples qu'une moyenne arithmétique sont en réalité des cas particuliers d'agrégation par logique floue. Il existe certaines méthodes, extrêmement puissantes, comme l'intégrale de Choquet [38] (cf. ci-dessous) qui ajoutant à la moyenne des paramètres supplémentaires, permettent de prendre en compte les interactions entre les classifieurs et les connaissances *a priori* du décideur.

4.1.6 Sélection de méthodes d'agrégation

Dans cet état de l'art, j'ai présenté un ensemble de méthodes d'agrégation, applicables à différents types de problèmes, il convient maintenant de choisir les méthodes les plus adaptées à la fusion de classifieurs dans le cadre de la recherche d'homologues.
En première intention, des méthodes basées sur les labels ou les classements des multiples classes possibles présentent peu d'intérêt car les classifieurs renvoient une probabilité d'appartenance à la classe, donnée ayant un sens

en elle-même qu'il serait dommage de ne pas exploiter, d'autant que la réduction à des labels implique la détermination d'un seuil qui serait un paramètre supplémentaire à intégrer au modèle. Quant à la détermination des meilleurs candidats par la méthode de Pareto, elle ne permet de former qu'un ordre partiel et non un classement complet, ce qui pose un problème de granularité du classement pour la classification de génomes entiers.
A première vue, les méthodes ayant pour objectif de créer une architecture de classifieurs paraissent intéressantes, mais le principe d'architecture s'applique principalement à des classifieurs de natures différentes, afin d'éviter que des erreurs de même type s'additionnent, or les classifieurs ne vérifient pas cette propriété. En effet, le calculs des taux de Kendall (*cf.* 2.6.3 et 3.4.4) montrent que certains classifieurs présentent des corrélations importantes, suggérant qu'ils pourraient avoir des biais similaires. Dans ce cadre, la mise en place de structures hiérarchiques de classifieurs ou d'arbres de décision parait moins intéressante. J'écarte enfin les méthodes de type bayesiennes, déjà employées par les experts. Des méthodes telles que les procédures de vote ou de classement des critères peuvent quant à elles être approchées par l'utilisation de moyennes arithmétiques pondérées

Bien que simplistes, des opérateurs tels que le minimum, le maximum et le produit ont montré de bonnes performances dans certains problèmes d'agrégation. Les méthodes d'agrégation visant à optimiser l'utilisation d'opérateurs, comme la moyenne pondérée, sont également envisageable de par leur spécificité au problème. J'ai également choisi de retenir l'intégrale de Choquet issue de la logique floue, pour sa capacité à gérer les interactions entre classifieurs, tout en proposant également une optimisation du modèle au problème rencontré. Enfin, l'approche "méta-classifieur" me semble être la continuité naturelle de l'utilisation des classifieurs SVM. La validité de l'approche SVM ayant été démontrée dans les chapitres précédents, l'emploi d'un métaSVM comme une solution possible au problème de l'agrégation de classifieurs parait naturelle. Ces méthodes seront évaluées dans la suite

de ce chapitre.

4.2 Méthodes d'agrégation

Dans cette partie, je vais présenter plusieurs types d'opérateurs d'agrégation, dont je comparerai les performances dans la partie suivante. J'ai divisé ces méthodes en deux groupes : les opérateurs classiques qui sont des opérateurs directement applicables sur les résultats des classifieurs et les opérateurs complexes qui impliquent des procédures d'optimisation des paramètres.
Une condition d'utilisation de tous ces opérateurs est la nécessité que les données manipulées *i.e.* les scores retournés par les classifieurs soient commensurables afin de pouvoir être comparé par les opérateurs. Cette propriété est vérifiée par les classifieurs puisque tous renvoient une probabilité d'appartenance à la famille d'intérêt.

4.2.1 Opérateurs classiques

Les opérateurs dits classiques sont des opérateurs simples, qui n'utilisent pas de paramètre. Ces opérateurs sont :

- Le minimum
- le maximum
- le produit

Ils peuvent être utilisés indifféremment sur des scores ou sur les rangs, à l'exception du produit, qui ne s'applique que sur des scores normalisés. Un quatrième opérateur, la moyenne non-pondérée, vient s'ajouter à cette liste. Cet opérateur et ces performances ont été décrits dans le chapitre précédent 3.4.5, page 135.

Le minimum

Le concept du minimum (appelé *min* par la suite) est une philosophie pessimiste qui consiste à considérer que les classifieurs surévaluent les can-

didats et que le classifieur donnant le rang ou le score le plus bas est le plus proche de la réalité. L'idée est donc d'utiliser comme score ou rang global d'un candidat, le score ou le rang le plus faible qu'il ait obtenu parmi l'ensemble des classifieurs. A noter que pour chaque le candidat, un classifieur différent peut être choisi comme *min*.

Le maximum

La philosophie du maximum (désigné *max* par la suite) est exactement l'inverse de celle du *min*. Supposant que les classifieurs sous-estiment les candidats, le classifieur accordant le score/rang le plus élevé aux candidats est le plus proche de la réalité. On désigne donc comme score/rang global d'un candidat le score/rang le plus élevé qu'il a obtenu parmi l'ensemble des classifieurs. Comme pour le *min* un classifieur différent pour chaque candidat peut être désigné comme *max*

Le produit

A l'inverse des opérateurs *min* et *max*, le produit ne sélectionne pas une de ses composantes mais les combinent pour en faire le score final. Cet opérateur d'agrégation diffère également des autres par le fait qu'il ne peut pas travailler sur des rangs, mais permet d'associer des scores pour peu que ces derniers aient été normalisés.
Soient p_i les scores normalisés donnés par chaque classifieur pour un même candidat, le produit P_{agreg} renvoyé par cet opérateur est :

$$P_{agreg} = \prod_{i=1}^{i=n} P_i$$

Malgré sa simplicité, il a été observé que cet opérateur pouvait obtenir de très bons résultats, particulièrement quand les scores manipulés sont des probabilités.

4.2.2 Opérateurs complexes

Ces opérateurs utilisent, comme les opérateurs classiques, les données des classifieurs mais les combinent à l'aide de procédures impliquant le plus souvent une phase d'apprentissage ou d'optimisation. Ces méthodes ne peuvent donc être appliquées qu'à condition de disposer un jeu d'apprentissage, distinct de celui qui a servi à l'apprentissage des classifieurs. Je présente ici les trois méthodes dégagées dans l'état de l'art : La moyenne pondérée, le métaSVM, l'intégrale de Choquet.

La moyenne pondérée

La moyenne pondérée combine les scores ou les rangs de ces derniers afin d'obtenir un score ou rang final. Le calcul de la moyenne s'effectue de la manière suivante :

$$\frac{1}{n}\sum_{i=1}^{i=n}\alpha_i X_i$$

où X_i désigne le score/rang obtenu par le classifieur i et α_i le coefficient de pondération de ce classifieur. La principale question qui se pose pour l'usage de la moyenne pondérée est la valeur des coefficients de pondération de chaque classifieur. La réponse la plus simple est de considérer que chaque classifieur est équivalent aux autres, associant ainsi à chacun une pondération de 1, c'est à dire de calculer la moyenne arithmétique. Cette solution triviale, utilisée dans le chapitre précédent, est un bon choix quand rien n'est connu sur les classifieurs, mais peut être amélioré à l'aide d'informations additionnelles ou par l'usage de méthodes d'optimisation.
Les performances des classifieurs peuvent par exemple être utilisées pour déterminer leurs poids relatifs. Ainsi le classifieur i se verrait attribuer une pondération α_i telle que

$$\alpha_i = \frac{S_i}{\sum_{i=1}^{i=n} S_i}$$

Où S désigne son score *ROC* moyen, comme calculé en 2.4.3. Cette alternative propose une pondération fondée sur des arguments ayant une

signification pour l'expérimentateur mais ne garantit en aucun cas des performances optimales.
De facto, la meilleure solution est de rechercher les coefficients tels qu'ils maximisent les performances de l'agrégation. L'optimisation de plusieurs paramètres à partir d'un critère donné est un problème classique en informatique pour lequel il existe de nombreuses solutions. L'une des plus populaires est l'utilisation d'algorithmes génétiques [32], qui ont des applications variées dans de nombreuses disciplines, y compris la bioinformatique. Le fonctionnement des algorithmes génétiques est présenté ci-après.

4.2.3 Optimisation par méthode évolutionniste

Bien que n'étant pas une méthode d'agrégation, je vais brièvement présenter le fonctionnement général des méthodes évolutionnistes d'optimisations. Dans le cadre de ce travail, ce type de méthode sera employé pour optimiser les coefficients d'une moyenne pondérée.

Cette technique d'optimisation est une métaheuristique ayant pour objectif d'approcher la solution à un problème d'optimisation en un temps raisonnable quand ceci n'est pas possible par des méthodes exactes. Cette stratégie procède par analogie avec la sélection naturelle. Ainsi, les solutions (*e.g.* des jeux de coefficients de moyenne pondérée) sont représentés par des individus possédant des gènes (*e.g.* un coefficient) formant un chromosome (une suite ordonnée de gènes). Ces individus sont amenés à évoluer puis à être sélectionnés en fonction d'une mesure d'adaptation. Les individus qui subissent ce cycle évolution-sélection forment une "génération". On construit ainsi un grand nombre de générations successives qui sont sensées s'approcher de plus en plus de la solution.

Plus précisément, on commence par initialiser au hasard un certain nombre d'individus (*e.g.* tirer au hasard des jeux de coefficients) qui formeront la première génération et sera soumise à un premier cycle évolution-sélection.

Lors des cycles, chaque gène a une certaine probabilité de subir un "événement évolutif". Classiquement les deux événements évolutifs utilisés sont la mutation et la recombinaison. La mutation est une modification aléatoire d'un gène (*e.g.* tirage au hasard d'un nouveau coefficient). La recombinaison est la réassociation des gènes de deux chromosomes d'individus différents. Les chromosomes sont scindés en deux aléatoirement et chaque partie est réassemblée avec une partie du chromosome partenaire.
Une fois cette phase d'évolution effectuée, les individus subissent une phase de sélection. La mesure d'adaptation (*e.g.* le score ROC du classement des candidats après agrégation), est utilisée pour ne conserver qu'un nombre n d'individus destinés à passer à la génération suivante. Cette sélection peut s'opérer de plusieurs façons :

- par rang, *i.e.* les n meilleurs individus sont sélectionnés
- au hasard, *i.e.* n individus sont tiré au hasard pour devenir la génération suivante
- par tournois, *i.e.* k gènes sont tirés aux hasard et les p plus adaptés du point de vue du critère de sélection sont retenus
- par "roulette" *i.e.* on tire au hasard sur une "roue de la fortune" les individus à conserver, sachant que chaque individu est représenté proportionnellement à son adaptation et donc qu'il a d'autant plus de chances d'être sélectionné qu'il est adapté.

Actuellement, les modes de sélection par tournois et par roulette sont les plus prisés. On peut s'étonner que le mode de sélection par rang, où le hasard n'intervient pas, ne soit pas favorisé d'autant qu'il semble le plus efficace pour atteindre un optimum. La raison est que, comme dans la nature, il est plus intéressant de maintenir une certaine diversité, y compris en gardant des individus moins bien adaptés, pour augmenter les possibilités d'évolution de la population. Ceci permet d'atteindre l'optimum en évitant des solutions sous-optimales.
L'algorithme peut s'arrêter sous plusieurs conditions : soit lorsqu'il converge, *i.e.* que l'adaptation moyenne de la population n'évolue plus, soit après un

nombre déterminé d'itérations, soit après qu'un individu ait atteint une valeur prédéterminée (*e.g.* un score *ROC* de 0.95). Le gène le plus adapté, *e.g.* les coefficients obtenant la meilleure agrégation, est alors sélectionné comme solution. La figure 4.3 illustre les différentes étapes d'un algorithme génétique.

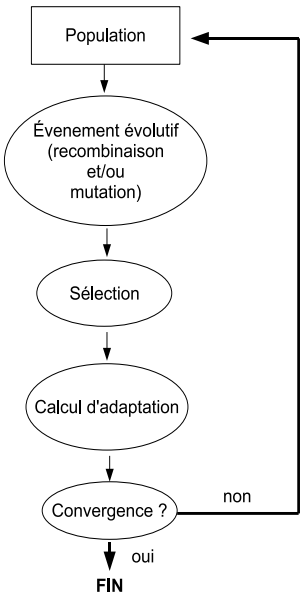

FIG. 4.3 – Schéma d'un algorithme génétique

MétaSVM

L'idée d'un opérateur d'agrégation basé sur un SVM est assez intuitive. En utilisant comme caractéristique les scores obtenus par des classifieurs, on peut construire un classifieur qui discriminerait les négatifs des positifs en traçant une frontière entre les vecteurs de scores des classifieurs individuels. Les SVM ont été plus amplement décrits en 2.2 (p 83). Dans le cas d'un métaSVM, le vecteur est constitué des probabilités d'appartenance à la classe des cytokines du candidat pour chaque classifieur. Il est donc d'une taille égale au nombre de classifieurs. La figure 4.4 schématise le fonctionnement d'un métaSVM. Cette méthode présente tous les avantages des SVM en terme de capacité de sélection des dimensions et des données intéressantes (les vecteur-support), ainsi qu'en terme de généralisation.

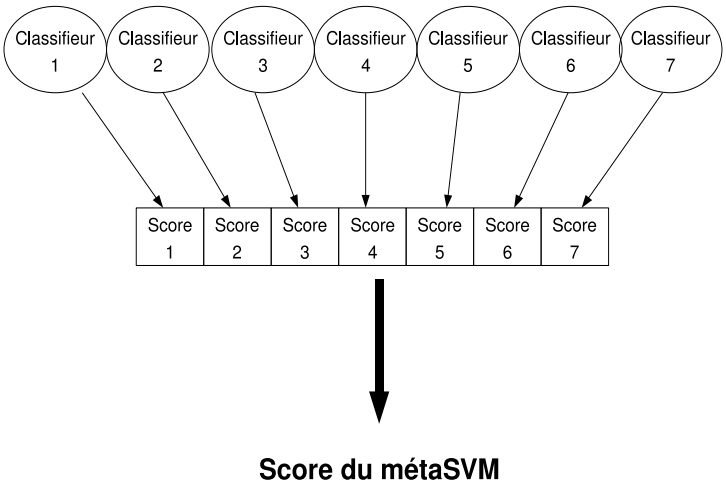

FIG. 4.4 – Fonctionnement général d'un métaSVM : les scores des classifieurs sont utilisés comme composante du vecteur du métaSVM qui renvoie lui-même un score sous forme d'une probabilité d'appartenance à la classe d'intérêt

Comme tout SVM, le métaSVM peut renvoyer une probabilité d'appartenance à la classe, utilisable pour ordonner les candidats.

Intégrale de Choquet

Développée par le mathématicien Gustave Choquet dans les années 50, l'intégrale de Choquet est un opérateur d'agrégation peu connu mais ayant déjà fait ses preuves dans le domaine de la décision multicritère. Une description plus précise de cette méthode est présentée dans les références suivantes : [38] et [19]. L'intégrale de Choquet est basée sur le principe de capacité, qui peut être vu comme une généralisation de la notion de vecteur de poids [38]. La capacité se définie de la façon suivante : soit N un ensemble de critères $N = 1, \ldots, n$ la fonction $\mu : P(N) \rightarrow \mathbb{R}$ est une capacité si

$$\mu(\emptyset) = 0, \mu(N) = 1; \forall S, T \subseteq N, T \subseteq S \Rightarrow \mu(T) \leq \mu(S)$$

$\mu(S)$ et $\mu(T)$ peuvent être considérés comme les poids des sous-ensembles de N, S et T respectivement. En utilisant cette définition de la capacité, on

calcule l'intégrale de Choquet de $X = 1 \ldots, n, \in \mathbb{R}^n$ de la façon suivante :

$$C_\mu(X) = \sum_{i=1}^{n} X_\sigma(i)[\mu(A_\sigma(i)) - \mu(A_\sigma(i+1))]$$

où σ est une permutation sur N telle que $X_\sigma(1) < X_\sigma(n)$ et

$$A_\sigma(i) = \sigma(i), \ldots, \sigma(n), \forall i \in N, A_\sigma(n+1) = \emptyset$$

Il y a plusieurs degrés de complexité à l'intégrale de Choquet. A son niveau le plus simple, l'intégrale de Choquet revient à une moyenne pondérée. A des degrés plus élevés, elle peut prendre en compte certaines propriétés des critères telles que :

- la redondance des critères,
- l'interaction entre critères *i.e.* des critères dont les valeurs croissent et décroissent dans le même sens ou dans des sens opposés,
- les preférences du décideur.

qui sont ignorés par des opérateurs plus classiques tels que la moyenne arithmétique. Concernant les préférences du décideur, il est possible de les modéliser *via* une fonction d'utilité telle que

$$x \prec y \Leftrightarrow U(x) \leq U(y) \forall x, y \in X$$

Où X réprésentent l'ensemble des critères (notes, scores ...) du décideur, x et y deux critères et $U(x)$ et $U(y)$ les fonctions d'utilité des critères x et y respectivement. On dit que si x est préféré à y ($x \prec y$) alors $U(x)$ sera supérieur à $U(y)$.
On détermine alors la capacité par optimisation sous contrainte [51] des fonctions d'utilité. Cette phase utilise un jeu "d'apprentissage" sur lequel l'optimisation est opérée. Le calcul de l'intégrale de Choquet est alors effectué en tenant compte de la capacité et retourne un score qui permet d'ordonner les candidats.

Appliquées au cas qui nous intéresse, aux moins deux propriétés de l'intégrale

de Choquet sont particulièrement intéressantes. La première est la gestion de la redondance entre les critères. Parmi les classifieurs, deux d'entre eux se basent sur la composition en acides aminés de la séquence candidate et les trois autres sur des scores de similarité. On sait, par le calcul des taux de Kendall (2.6.3 et 3.4.4), que les classifieurs présentent une certaine corrélation, la capacité de l'intégrale de Choquet à évaluer et limiter cette redondance est donc ici très intéressante.
La possibilité de tenir compte des interactions entre critères apporte également un plus car cette dernière n'a pas été évaluée chez les classifieurs et les experts or il est possible que de telles relations existent, par exemple entre la similarité de séquences et la similarité de structures.

L'intégrale de Choquet s'avère donc un outil performant, disposant de propriétés intéressantes dans le traitement des critères, particulièrement en ce qui concerne leur possible biais.

4.3 Evaluation des méthodes

Dans cette partie, je présente l'évaluation des différentes méthodes d'agrégation. Le protocole d'évaluation, identique à celui mener pour agréger les classifieurs et les experts par la moyenne pondérée dans le chapitre "Expertises automatiques" est rappelé ci-après.

4.3.1 Matériel et méthodes

Séparation des données

Ainsi qu'on l'a vu dans la partie précédente, les opérateurs complexes d'agrégation nécessitent un jeu d'apprentissage ou d'optimisation qui leur est propre, ce qui implique, pour une évaluation de ces méthodes englobant des classifieurs SVM, de séparer le jeu de données en trois : un jeu d'apprentissage des classifieurs, un jeu d'apprentissage des opérateurs d'agrégation et un jeu de test.

Ainsi que je l'ai expliqué dans le chapitre "Expertise automatiques", Le jeu de test (E2) constitué de 30 cytokines et 6000 contre-exemples peut être partitionné en deux jeux E3 (paramétrage) et E4 (évaluation), constitués de respectivement 15 cytokine et 100 contre-exemples pour le jeu E3 et 15 cytokines et 5900 pour le jeu E4. Le jeu E4 est ensuite restreint aux 15 cytokines et aux 200 contre-exemples les mieux classés par les classifieurs *cf.* 3.4.2.

Optimisation des coefficients de la moyenne

Les coefficients ont été optimisés en implémentant un algorithme génétique. Les paramètres de l'algorithme ont été sélectionnés après plusieurs essais et sont :

- taux de mutation = 0.01
- taux de croisement = 0.8
- condition d'arrêt = 200 itérations

Le mode de sélection choisie est la sélection par tournois.

MétaSVM

Le vecteur du métaSVM comprend neuf composantes. La procédure utilisée pour l'apprentissage du métaSVM est similaire à celle décrite en 2.5 pour les classifieurs SVM .

Intégrale de Choquet

L'intégrale de Choquet a été calculée en utilisant le paquetage kappalab [19] du projet GNU R [27]. L'intégrale de Choquet doit construire une matrice de contraintes or cette matrice ne supporte qu'un nombre restreint de dominances (au sens de Pareto) d'un contre-exemple sur un exemple. Si ces dominances deviennent trop fréquentes, l'optimisation des paramètres devient impossible. C'est le cas pour un grand nombre de jeux de données générés lors de cette étude. L'impossibilité de calculer dans ces cas l'intégrale de Choquet m'a conduit à abandonner cette méthode.

Mesure de performance

Pour chaque méthode, j'ai mesuré les performances de plusieurs combinaisons de classifieurs et d'experts. Pour chacune de ces combinaisons, j'ai généré un jeu d'apprentissage pour les opérateurs d'agrégation (même si la méthode testé, *e.g.* le min, n'en n'avais pas besoin) et un jeu d'évaluation. Après un éventuel paramétrage de la méthode d'agrégation, j'ai mesuré sur le jeu d'évaluation le score ROC de la méthode testée pour l'association donnée ainsi que celui de chaque classifieur. Le meilleur score ROC parmi ceux des classifieurs a été retenu et soustrait au score ROC de l'agrégation pour calculer le gain de score ROC obtenu par cette dernière. Ce processus a été répété sur les 100 jeux de données et la moyenne du gain (δROC) dû à l'agrégation a été calculée. Ce protocole est identique à celui de 3.4.2

4.3.2 Résultats

Je présente dans le tableau 4.1 les δROC de l'association utilisant les cinq classifieurs et les quatre experts pour les méthodes d'agrégation suivantes :

- le min,
- le max,
- le produit,
- la moyenne arithmétique,
- la moyenne pondérée,
- le métaSVM.

Pour l'ensemble de ces résultats, les performances des classifieurs sont du même ordre que celles obtenues en ROC_{200}(2.6.2), c'est à dire entre 0.7 et 0.85. Le classifieur LA kernel est fréquemment le meilleur classifieur, auquel les agrégations sont comparées.

Performances des méthodes d'agrégation

Ayant déterminé au chapitre précédent qu'il n'était pas souhaitable d'écarter des experts, je ne discute pas ici des résultats des différentes

associations d'experts avec les classifieurs. Au demeurant, les δROC de ces associations confirment qu'en général la combinaison des cinq classifieurs et des quatre experts donne les meilleures performances. Le lecteur intéressé par cette discussion est invité à se reporter à l'annexe A, p 183.

On observe que le *min*, la moyenne arithmétique, la moyenne pondérée et

Méthode d'agrégation	δROC
min	0.086
max	-0.08
produit	-0.014
moyenne arithmétique	0.054
moyenne pondérée	0.131
métaSVM	0.122

TAB. 4.1 – δROC obtenu pour chacune des méthodes d'agrégation avec les cinq classifieurs et les quatre experts

le métaSVM obtiennent des δROC positifs alors que le *max* et le produit obtiennent des δROC négatifs.
Par ailleurs les résultats de la moyenne pondérée et du métaSVM surclassent ceux du *min* et de la moyenne arithmétique. Les performances de la moyenne pondérée semblent légèrement supérieures à celles du métaSVM, mais compte-tenu des écart-types, cette différence n'est pas significative.

Coefficients de pondération de la moyenne pondérée

La moyenne pondérée présente l'avantage de rendre compréhensible l'apport de chacune de ses composantes, à travers des coefficients de pondération explicites. Ainsi, j'ai analysé les pondérations de chaque classifieur et expert sur les 100 jeux de données utilisés pour calculer le δROC. La figure 4.5 présente la distribution de ces coefficients.

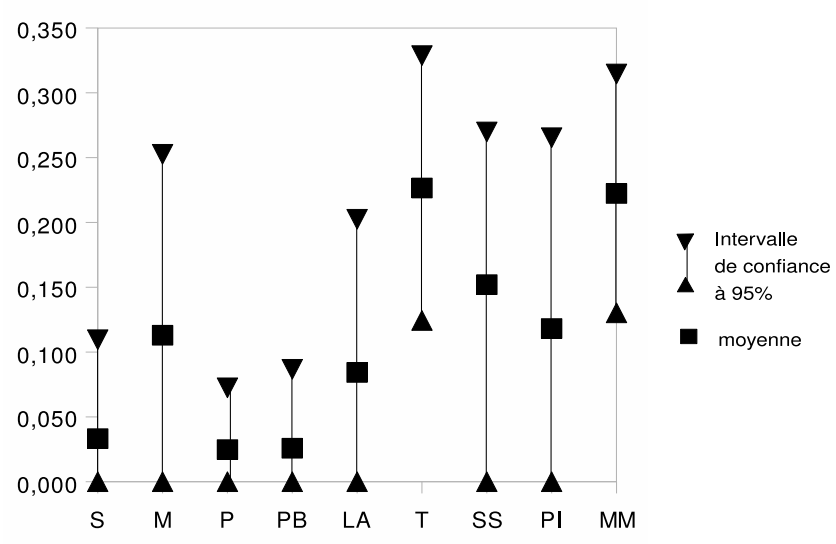

FIG. 4.5 – Distribution des poids des classifieurs et des experts pour 100 jeux de données. S = Spectrum kernel, M = Mismatch kernel, P = Pairwise, PB = PairwiseBlast, LA = LA kernel, T = Expert Taille de séquence, SS = Expert Structure Secondaire, PI = Expert Point Isoélectrique, MM = Expert Masse Moléculaire

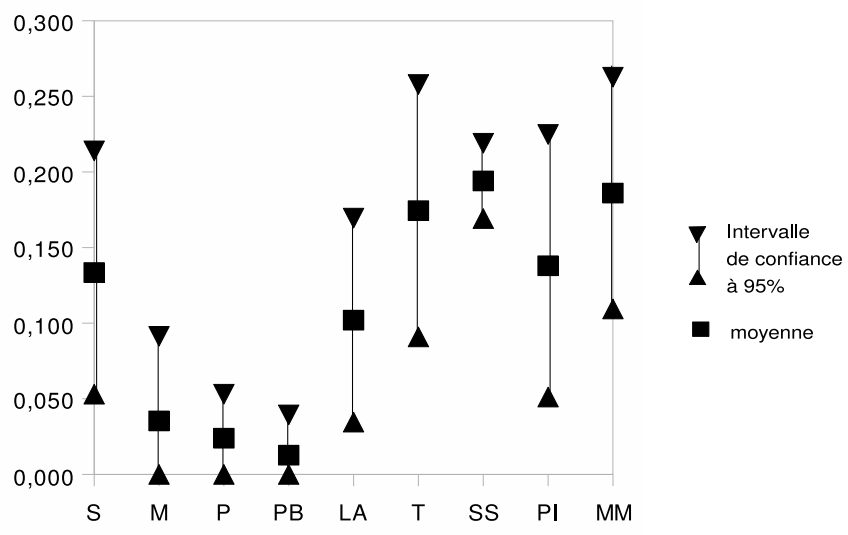

FIG. 4.6 – Distribution des poids des classifieurs et des experts pour 100 optimisations sur un même jeu de données. S = Spectrum kernel, M = Mismatch kernel, P = Pairwise, PB = PairwiseBlast, LA = LA kernel, T = Expert Taille de séquence, SS = Expert Structure Secondaire, PI = Expert Point Isoélectrique, MM = Expert Masse Moléculaire

Sur l'ensemble des 100 jeux de données, on observe que les variations de poids peuvent être très importantes pour un même critère, particulièrement pour Mismatch kernel, LA kernel, E-T, E-SS, E-PI, E-MM.
Ceci pouvant être attribué à la diversité des jeux de paramétrage (E3), présentant des situations très différentes, j'ai également analysé la distribution des coefficients de pondération sur un même jeu de données pour 100 optimisations. Les résultats sont présentés dans la figure 4.6.
On observe que la plupart des classifieurs et experts, à l'exception de Pairwise, PairwiseBlast et E-SS, présentent comme précédemment des variations importantes de pondération.

Ces résultats suggèrent donc que des solutions différentes, représentés par plusieurs maxima dans l'espace de recherche, obtiennent des performances comparables.

4.3.3 Discussion

Les méthodes d'agrégations utilisées ici peuvent être regroupées en deux catégories : les méthodes choisissant une de leurs composantes comme score final (min, max) et les méthodes associant l'ensemble de leurs composantes pour obtenir le score final (produit, moyenne pondérée et métaSVM).
Dans la catégorie des méthodes choisissant une de leurs composantes comme score final, le min présente une amélioration par rapport au meilleur classifieur, ce qui n'est pas le cas du max. Ces deux méthodes sont basées sur le même principe mais ont des philosophies opposées. Le min suppose que les classifieurs et les experts sont optimistes et choisit le plus pessimiste d'entre eux, le max, au contraire, part du principe que les classifieurs et les experts sont pessimistes et choisit le plus optimiste d'entre eux. Les résultats démontrent que la philosophie du min est la plus proche de la réalité.
Parmi les méthodes associant leurs composantes, on observe que le produit obtient des résultats légèrement inférieurs au meilleur classifieur. Les

trois autres méthodes (la moyenne arithmétique, la moyenne pondérée et le métaSVM) obtiennent des performances supérieures au meilleur classifieur. La moyenne arithmétique est toutefois surclassée par la moyenne pondérée et le métaSVM, qui présente des performances comparables. Les δROC obtenus atteignent des valeurs supérieures à 0.1, ce qui représente un gain de performance significatif pour ces méthodes.
On pourra s'étonner qu'une stratégie aussi simple qu'une moyenne pondérée fasse jeu égal avec une autre, beaucoup plus élaborée, comme un métaSVM. L'hypothèse qui pourrait être proposée ici est que les SVM travaillent sur un faible nombre de dimensions, ce qui limite leur capacité discriminante. Cependant, l'analyse de la distribution des coefficients de pondération a montré que la moyenne pondérée présentait plusieurs solutions proches de l'optimum et qu'elle oscillait entre elles, y compris pour un même jeu de paramétrage. Cette instabilité rend son efficacité sur un jeu de données inconnues plus incertain. Je préconise donc plutôt l'emploi d'un métaSVM *a priori* plus stable.

4.4 Conclusion

Ce chapitre clot l'ensemble des étapes nécessaires à la réalisation d'une nouvelle méthode de détection des homologues, en décrivant différentes stratégies d'agrégation des classifieurs et des experts précédemment proposés.
J'ai dans un premier temps présenté un état de l'art des différentes familles de techniques d'agrégation et dégagé quelques méthodes pertinentes pour le problème étudié. Ces méthodes ont été sélectionnées en fonction, entre autre, de leur applicabilité au type de données utilisé ici (des probabilités d'appartenance à la famille d'intérêt).
J'ai ensuite décrit plus précisément ces méthodes en expliquant leur fonctionnement et en dégageant leurs principaux aventages. Les méthodes sélectionnées sont :

– l'intégrale de Choquet.

- le min,
- le max,
- le produit,
- la moyenne arithmétique,
- la moyenne pondérée,
- le métaSVM.

Pour des raisons techniques, l'intégrale de Choquet n'a pu être appliquée a ma problématique, me laissant donc six méthodes d'agrégation.
Dans une troisième partie, j'ai évalué ces méthodes en les comparant aux performances du meilleur classifieur pour chaque jeu de données. Lors de ces évaluations, j'ai pris en compte plusieurs combinaisons d'experts pour trouver l'association optimale de ces derniers avec les classifieurs.
Ces expériences ont montré que la moyenne pondérée, optimisée par un algorithme génétique et le métaSVM sont les meilleures stratégies d'agrégation.
L'ensemble de ces résultats permet donc de mettre au point une stratégie compléte de recherche d'homologues basée sur cinq classifieurs et quatre experts agrégés en utilisant préférentiellement un métaSVM.

La figure 4.7 propose un récapitulatif de l'ensemble de système de classement des candidats pour la recherche d'homologues de cytokines à quatre hélices α.

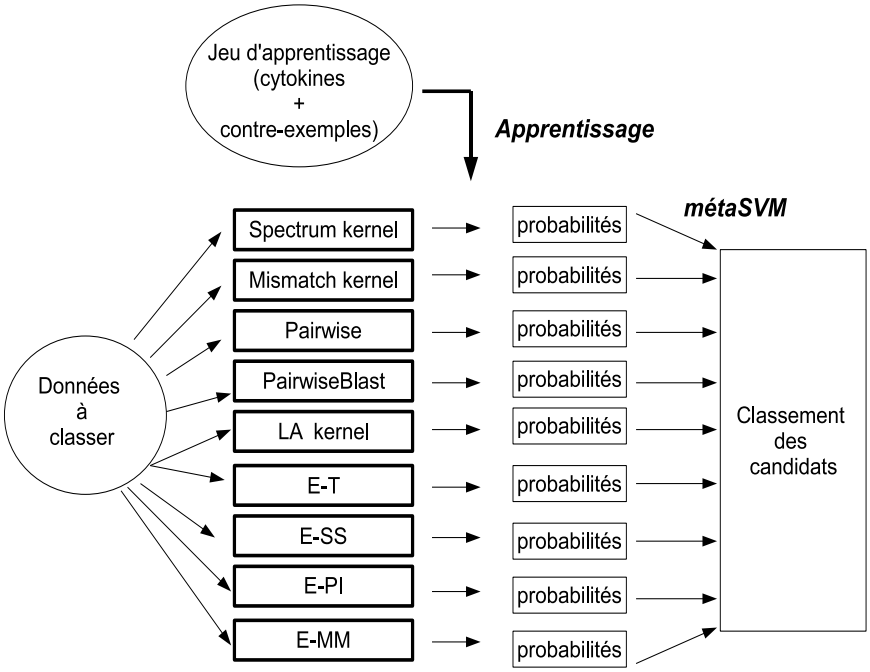

FIG. 4.7 – Schéma général de la classification d'un ensemble de candidats pour la recherche de cytokines à quatre hélices α. moy.pond. = moyenne pondérée.

Conclusion

Cette thèse se positionne comme un travail multidisciplinaire, à l'interface de la biologie et de l'informatique. Son centre d'intérêt principal est les cytokines, une famille de protéines impliquées dans les mécanismes de prolifération, différenciation et mort cellulaire. Cette famille est principalement connue pour son rôle dans le système immunitaire ainsi que son intervention dans un large éventail de pathologies, ce qui en fait une cible de recherche intéressante pour des applications thérapeutiques.
L'objectif de cette thèse est la mise en place d'une méthode de recherche exhaustive des homologues de la famille des cytokines. La recherche d'homologues est une étape cruciale pour l'étude d'une famille de protéines et s'avère également une tâche difficile si les membres de la famille ont de faibles similarités de séquence, comme c'est le cas ici. C'est pour palier à cette difficulté et proposer un système exhaustif et automatique que ce travail de recherche a été entrepris.
Son apport principal est la mise au point d'une stratégie basée sur cinq classifieurs SVM (Spectrum kernel, Mismatch kernel, Pairwise, PairwiseBlast, LA kernel) et quatre experts spécifiques à la famille des cytokines (basées sur les structure secondaire, la taille, la masse moléculaire de la séquence, et le point isoélectrique), pouvant être combinés par un métaSVM ou éventuellement une moyenne pondérée. Le second apport vient du fait que cette méthode peut être généralisée à d'autres problèmes de recherche d'homologues, ainsi que je le montrerai plus bas.

Récapitulatif

Définition de la sous-famille des cytokines à quatre hélices α

La solution proposée pour identifier les homologues de cytokines était de rechercher parmi les séquences du génome humain celles qui s'apparentaient le plus aux membres connus de la famille. J'ai pour cela commencé par étudier la famille des cytokines afin de dégager les caractéristiques de ces protéines ainsi que les difficultés que posent leur classification en famille. En effet, les cytokines semblent former une famille dont les membres n'ont qu'une lointaine parenté, ainsi qu'en attestent par exemple les différents repliments pouvant être adoptés par les membres de la famille. Les cytokines peuvent toutefois être regroupés en sous-familles, comme celle des cytokines à quatre hélices α qui, bien que ne partageant pas une grande similarité de séquence, présentent plusieurs signes d'homologie proche. Dans l'idée d'identifier de nouvelles cytokines, il semble plus intéressant de se pencher sur ces sous-familles que sur la famille en général. Ce travail de recherche s'est donc cantonné à l'étude des cytokines à quatre hélices α.

Recherche d'homologues par classification supervisée

Partant de travaux précédents au laboratoire, je me suis intéressé aux méthodes de classifications supervisées, et plus particulièrement aux SVM, pour rechercher des homologues. Cette stratégie, largement employée dans la littérature, peut donner lieu à plusieurs classifieurs exclusivement dédiés aux séquences protéiques. J'ai évalué cinq classifieurs, issus de la littérature, pour leur capacité à reconnaître les cytokines. Bien que tous soient capables de distinguer les cytokines avec plus ou moins d'efficacité, aucun classifieur n'est capable d'opérer une classification parfaite, particulièrement pour la tête du classement. Pour palier à cette difficulté, j'ai cherché à utiliser des informations supplémentaires, inexploitées par les classifieurs.

Ajouts d'experts biologiques

En plus des classifieurs, j'ai développé la notion d'experts qui désigne des outils incluant des caractéristiques biologiques spécifiques à la famille étudiée. Ces experts n'ont pas des performances de classification optimales mais apportent des informations biologiques supplémentaires, inaccessibles aux classifieurs. J'ai expliqué comment choisir un critère d'expertise et j'ai proposé une définition d'experts, à partir de bayésiens naïfs. J'ai ensuite développé le cas de quatre experts spécifiques aux cytokines à quatre hélices α, basés sur la taille, la structure secondaire, le point isoélectrique et la masse moléculaire des séquences protéiques. J'ai évalué les performances discriminantes de ces experts, qui se sont avérées inférieures à celles des classifieurs, mais pas négligeables. Une agrégation très simple avec les classifieurs a toutefois montré que les experts étaient capables d'améliorer sensiblement les performances de classification du système par rapport au meilleur des classifieurs.

Agrégation

J'ai suggéré dans les parties précédentes l'utilisation des experts en complément des classifieurs et la combinaison de ces derniers pour améliorer encore les performances de classification. Je me suis donc intéressé à une famille de méthodes appelées "méthodes d'agrégation" permettant de combiner les résultats de plusieurs classifieurs et des experts en un seul résultat. Après avoir exploré les différents types de méthodes, j'ai proposé l'évaluation de celles qui semblaient s'adapter le mieux à la question de l'agrégation de scores en un classement final. Certaines méthodes, dont le métaSVM, ont montré une amélioration sensible des performances de classification.

Bilan

L'ensemble de ce travail de thèse a permis de mettre en évidence une nouvelle méthode de recherche d'homologues, basée sur la classification de

séquences candidates par agrégation de classifieurs et d'experts. Cinq classifieurs SVM et quatre experts ont été optimisés pour la recherche d'homologues de cytokines à quatre hélices α et plusieurs méthodes d'agrégations dont la meilleure semble être le métaSVM. Ces méthodes, appliquées aux cytokines à quatre hélices α, peuvent aisément être étendues à d'autres familles de protéines.

Perspectives

Comme tout travail de recherche, ce travail présente plusieurs perspectives, que je présente ci-après.

Implémentation d'une plateforme de recherche d'homologues

La première extension naturelle de ce travail est son application à l'échelle d'un génome. Une plateforme incluant la plupart des outils précédemment décrits a d'ores et déjà été développée. Il s'agirait donc de récupérer l'ensemble des protéines du génome humain et de les tester avec les outils que j'ai décrits. Cette tâche est plus ardue qu'il ne le semble, du fait de la nature des données. En effet, pour avoir accès à toutes les séquences protéiques possibles, y compris des protéines encore non-identifiées, qui ne figureraient pas dans les banques de données, on peut envisager de récupérer l'ensemble des transcrits du génome, présents dans la banque de données Unigene, et les traduire en séquences protéiques. Le problème d'une telle approche est que les EST ou les transcrits incomplets donneraient des séquences protéiques tronquées. L'existence de transcrits possédant encore des traces de vecteurs de clonages donneraient au contraire des séquences avec des résidus n'existant pas de la protéine *in vivo*. Ces biais risquent d'affecter de manière importante les experts, et dans une moindre mesure les classifieurs, qui impliquent de disposer de la séquence protéique exacte. Il conviendra donc, au préalable d'évaluer la qualité des outils sur ces données "biaisées" et/ou de procéder à un nettoyage de ces données.

Cette étape passée, il sera possible d'appliquer les méthodes que j'ai décrites à un génome complet.

Ajout d'experts

J'ai proposé dans les parties 1.6 et 3.2.2, plusieurs critères qui n'ont pas encore été exploités :

- la structure des gènes,
- la localisation chromosomique,
- la localisation cellulaire,
- la présence de ponts disulfures.

Parmi ces critères, certains tels que la structure génomique et la localisation chromosomique paraissent très intéressants pour les cytokines, bien que des contraintes techniques m'aient conduit à les écarter pour l'instant. De même, la localisation cellulaire pourrait permettre d'éliminer un certain nombre de contre-exemples par ailleurs bien classés. Il serait donc important de tester ces critères afin de concevoir de nouveaux experts qui pourraient encore améliorer la qualité du classement des candidats.

Généralisation à d'autres familles de protéines

Une fois les pistes ci-dessus explorées et l'outil appliqué aux cytokines à quatre hélices α, la suite logique de ce travail serait de l'appliquer également à d'autres familles de protéines.

En effet, les différentes méthodes que j'ai exposées dans cette thèse et les développements d'outils basés sur ces méthodes sont transposables à d'autres sujets d'étude que les cytokines à quatre hélices α.

Concernant les classifieurs, il suffit de constituer un nouveau jeu d'apprentissage et de procéder à quelques tests d'optimisation des paramètres pour pouvoir les utiliser avec d'autres familles. Les experts sont quant à eux plus difficiles à transposer tels quels, puisqu'ils ont été proposés pour leur spécificité, mais il est possible de développer de nouveaux experts pour les familles étudiées à partir de connaissances biologiques sur

ces familles. Qui plus est, certains experts, comme la structure secondaire ou la taille et le point isoélectrique, en tant que caractéristiques physico-chimiques, pourraient être applicables à d'autres familles, puisqu'il s'agit de caractéristiques souvent conservées dans une même famille.

Les premières familles de gènes auxquelles les méthodes développées lors de cette thèse pourraient être appliquées sont, fort logiquement, les autres sous-familles de cytokines. Rappelons-le, l'objectif initial était de mettre en évidence toutes les cytokines du génomes or, du fait de l'hétérogénéité de cette "famille" de protéine, il semblait plus judicieux de travailler au niveau de chaque sous-famille. Maintenant qu'une méthode de détection des homologues à été mise au point au niveau d'une de ces sous-familles, il devient possible de les explorer toutes tour-à-tour.

Ce travail de recherche a permis de mettre au point une méthode de détection des homologues basée sur des classifieurs génériques et des experts apportant des informations biologiques spécifiques sur la famille étudiée. Cette stratégie semble donner de bons résultats lorsqu'elle est appliquée aux cytokines à quatre hélices α. Appliquée à l'ensemble du génome humain, elle devrait permettre d'identifier des membres encore inconnus de cette famille et ouvre la voie à d'autres application à d'autres familles de protéines, cytokines ou non.

Table des figures

Liste des tableaux

Annexe A

Comparaisons des associations d'experts

A.1 Résultats d'agrégation par la méthode du *min*

Les résultats des associations des classifieurs et des experts, agrégés par la méthode du *min* sont présentées dans le tableau A.1

On observe en premier lieu que tous les δROC, à l'exception de ceux de l'association des classifieurs entre eux et des classifieurs avec E-SS, sont positifs. L'association des classifieurs entre eux obtient le plus faible δROC. L'association des classifieurs avec E-SS, E-PI et E-MM obtient le meilleur δROC mais la différence avec l'association entre les classifieurs, E-SS et E-MM n'est pas significative.
E-T apparaît souvent dans des associations ayant un δROC élevé mais pas dans les deux meilleurs. *A contrario* E-MM apparaît dans les trois associations ayant les δROC les plus élevés. On notera que la combinaison de cet expert seul avec les classifieurs obtient un δROC plus élevé que plusieurs combinaisons de deux experts et une combinaison de trois experts.
E-SS présente un δROC négatif quand il est associé aux classifieurs seuls, mais il est impliqué dans les cinq combinaisons ayant les δROC les plus élevés.
E-PI seul obtient un δROC quasi nul, mais, associé à d'autres experts, il contribue à l'amélioration du δROC.

S	M	P	PB	LA	E-T	E-SS	E-PI	E-MM	δROC
⋆	⋆	⋆	⋆	⋆		⋆	⋆	⋆	0.099
⋆	⋆	⋆	⋆	⋆		⋆		⋆	0.095
⋆	⋆	⋆	⋆	⋆	⋆	⋆	⋆	⋆	0.086
⋆	⋆	⋆	⋆	⋆	⋆	⋆			0.086
⋆	⋆	⋆	⋆	⋆	⋆	⋆		⋆	0.081
⋆	⋆	⋆	⋆	⋆	⋆		⋆		0.078
⋆	⋆	⋆	⋆	⋆	⋆		⋆	⋆	0.075
⋆	⋆	⋆	⋆	⋆				⋆	0.068
⋆	⋆	⋆	⋆	⋆	⋆	⋆	⋆		0.056
⋆	⋆	⋆	⋆	⋆			⋆	⋆	0.054
⋆	⋆	⋆	⋆	⋆	⋆				0.054
⋆	⋆	⋆	⋆	⋆	⋆			⋆	0.052
⋆	⋆	⋆	⋆	⋆		⋆	⋆		0.031
⋆	⋆	⋆	⋆	⋆			⋆		0.008
⋆	⋆	⋆	⋆	⋆		⋆			-0.034
⋆	⋆	⋆	⋆	⋆					-0.043

TAB. A.1 – δROC de l'agrégation par la méthode du *min*. S = Spectrum kernel, M = Mismatch kernel, P = Pairwise, PB = PairwiseBlast, LA = LA kernel, E-T = Expert Taille, E-SS = Expert Structure Secondaire, E-PI = Expert Point Isoélectrique E-MM = Expert Masse Moléculaire. Les étoiles indiquent les combinaisons de classifieurs et d'experts testés

Plusieurs conclusions peuvent être tirées de ces résultats. En premier lieu, on observe que les classifieurs seuls obtiennent le moins bon δROC, ce qui signifie que les experts jouent un rôle important pour cette méthode d'agrégation.

Parmi les différents experts, E-MM est celui dont la contribution au δROC semble la plus importante. E-SS et, dans une moindre mesure, E-PI contribuent également à obtenir la plus grande efficacité. E-T n'est pas requis pour former la meilleure association de classifieurs et d'experts, malgré de bonnes performances seul ou en association avec d'autres experts. Cela peut être interprété par le fait que, le *min* étant une méthode sélectionnant un classifieur ou un expert comme résultat final, les capacités discriminatrices de deux composantes très corrélées ne sont pas combinées mais misent en compétition et la meilleure des deux donnera une plus grande efficacité à l'agrégation. C'est probablement ce qui se produit ici : E-T et

E-MM présentent une forte corrélation, et il est possible que E-MM soit légèrement plus discriminant que E-T et donc améliore légèrement les performances d'agrégation par rapport à ce dernier. Les résultats obtenus par ces deux experts suggèrent la possibilité qu'ils corrigent un biais de taille de séquences chez les classifieurs.
D'autre part, E-SS semble également requis pour obtenir les performances maximales d'agrégation, ainsi que E-PI dans une moindre mesure.

A.2 Résultats d'agrégation par la méthode du *max*

Les résultats des différentes combinaisons d'experts associés aux classifieurs par la méthode du *max* sont présenté dans le tableau A.2.

Ainsi que je l'ai détaillé dans le chapitre "Agrégation de classifieurs",

S	M	P	PB	LA	E-T	E-SS	E-PI	E-MM	δROC
⋆	⋆	⋆	⋆	⋆	⋆	⋆			-0.075
⋆	⋆	⋆	⋆	⋆	⋆	⋆	⋆		-0.076
⋆	⋆	⋆	⋆	⋆		⋆		⋆	-0.077
⋆	⋆	⋆	⋆	⋆		⋆	⋆	⋆	-0.079
⋆	⋆	⋆	⋆	⋆	⋆	⋆	⋆	⋆	-0.08
⋆	⋆	⋆	⋆	⋆	⋆	⋆		⋆	-0.08
⋆	⋆	⋆	⋆	⋆		⋆	⋆		-0.08
⋆	⋆	⋆	⋆	⋆		⋆			-0.083
⋆	⋆	⋆	⋆	⋆	⋆				-0.162
⋆	⋆	⋆	⋆	⋆				⋆	-0.166
⋆	⋆	⋆	⋆	⋆	⋆		⋆	⋆	-0.168
⋆	⋆	⋆	⋆	⋆			⋆	⋆	-0.168
⋆	⋆	⋆	⋆	⋆	⋆			⋆	-0.17
⋆	⋆	⋆	⋆	⋆			⋆		-0.171
⋆	⋆	⋆	⋆	⋆					-0.172
⋆	⋆	⋆	⋆	⋆	⋆		⋆		-0.174

TAB. A.2 – δROC de l'agrégation par la méthode du *max*. S = Spectrum kernel, M = Mismatch kernel, P = Pairwise, PB = PairwiseBlast, LA = LA kernel, E-T = Expert Taille, E-SS = Expert Structure Secondaire, E-PI = Expert Point Isoélectrique E-MM = Expert Masse Moléculaire. Les étoiles indiquent les combinaisons de classifieurs et d'experts testés

et que le montre les δROC tous négatifs, le *max* n'est pas une méthode

d'agrégation efficace dans le cas présent. Toutefois, on peut en tirer quelques informations sur l'apport de E-SS.
Ainsi, on constate que les meilleures associations comportent toujours E-SS, indiquant que cet expert attribue un score élevé à certaines cytokines. Cela est conforté par le fait que l'absence de E-SS donne un δROC deux fois plus faible qu'en sa présence (-0.083 pour la moins bonne association contenant E-SS contre -0.162 pour la meilleur dont il est absent).
On observe que toutes les associations ne contenant pas l'expert E-SS ont un δROC comparable à celui des cinq classifieurs seuls.

Ces observations montrent une prédominance de l'expert E-SS sur l'agrégation et une influence quasi nulle des autres experts. Cette prédominance confirme le fait que le SOV a tendance à favoriser les structures très proches, donnant à ces dernières un score élevé et donc un bon classement. Cette tendance seule est toutefois insuffisante pour améliorer la classification, probablement parce que cet expert ne favorise que les structures ayant une très forte similarité avec celles des cytokines.

A.3 Résultats d'agrégation par la méthode du produit

Les performances de la méthode du produit pour les différentes associations d'experts avec les classifieurs sont présentés dans le tableau A.3.

On observe qu'aucune association ne parvient à un score ROC au moins égale à celui du meilleur classifieur. Plusieurs associations présentent des δROC presque nuls, dont l'association des classifieurs seuls. D'une manière générale les valeurs de δROC indiquent que le ROC de l'agrégation est proche mais inférieur à celui du meilleur classifieur.

Le fait marquant de ces résultats est que l'ensemble des associations ob-

S	M	P	PB	LA	E-T	E-SS	E-PI	E-MM	δROC
⋆	⋆	⋆	⋆	⋆					-0.007
⋆	⋆	⋆	⋆	⋆	⋆		⋆	⋆	-0.008
⋆	⋆	⋆	⋆	⋆	⋆	⋆			-0.009
⋆	⋆	⋆	⋆	⋆	⋆			⋆	-0.010
⋆	⋆	⋆	⋆	⋆		⋆	⋆		-0.010
⋆	⋆	⋆	⋆	⋆		⋆		⋆	-0.010
⋆	⋆	⋆	⋆	⋆			⋆		-0.010
⋆	⋆	⋆	⋆	⋆		⋆			-0.011
⋆	⋆	⋆	⋆	⋆		⋆	⋆	⋆	-0.013
⋆	⋆	⋆	⋆	⋆	⋆		⋆		-0.013
⋆	⋆	⋆	⋆	⋆	⋆	⋆	⋆	⋆	-0.014
⋆	⋆	⋆	⋆	⋆	⋆	⋆	⋆		-0.014
⋆	⋆	⋆	⋆	⋆			⋆	⋆	-0.015
⋆	⋆	⋆	⋆	⋆	⋆				-0.015
⋆	⋆	⋆	⋆	⋆	⋆	⋆		⋆	-0.018
⋆	⋆	⋆	⋆	⋆				⋆	-0.020

TAB. A.3 – δROC de l'agrégation par la méthode du produit. S = Spectrum kernel, M = Mismatch kernel, P = Pairwise, PB = PairwiseBlast, LA = LA kernel, E-T = Expert Taille, E-SS = Expert Structure Secondaire, E-PI = Expert Point Isoélectrique E-MM = Expert Masse Moléculaire. Les étoiles indiquent les combinaisons de classifieurs et d'experts testés

tiennent des performances similaires ce qui suggère un faible impact des experts, d'autant que les classifieurs seuls obtiennent un δROC comparable à la plupart des autres associations.

A.4 Résultats d'agrégation par la méthode de la moyenne pondérée

Les résultats d'agrégation par la moyenne pondérée sont reportés dans le tableau A.4.

On constate en premier lieu que l'ensemble des δROC sont positifs, y compris l'agrégation entre les classifieurs seuls. On observe clairement la présence de deux types d'associations, celles ayant un δROC entre 0.13 et 0.11 et celles ayant un δROC d'environ 0.06.
Le fait le plus marquant est que les associations ayant un δROC d'environ

S	M	P	PB	LA	E-T	E-SS	E-PI	E-MM	δROC
⋆	⋆	⋆	⋆	⋆	⋆	⋆	⋆	⋆	0.131
⋆	⋆	⋆	⋆	⋆	⋆	⋆			0.129
⋆	⋆	⋆	⋆	⋆	⋆	⋆		⋆	0.129
⋆	⋆	⋆	⋆	⋆	⋆	⋆	⋆		0.123
⋆	⋆	⋆	⋆	⋆	⋆		⋆	⋆	0.115
⋆	⋆	⋆	⋆	⋆		⋆		⋆	0.114
⋆	⋆	⋆	⋆	⋆	⋆			⋆	0.114
⋆	⋆	⋆	⋆	⋆	⋆				0.114
⋆	⋆	⋆	⋆	⋆	⋆		⋆		0.114
⋆	⋆	⋆	⋆	⋆		⋆	⋆	⋆	0.111
⋆	⋆	⋆	⋆	⋆				⋆	0.064
⋆	⋆	⋆	⋆	⋆			⋆	⋆	0.064
⋆	⋆	⋆	⋆	⋆		⋆			0.064
⋆	⋆	⋆	⋆	⋆		⋆	⋆		0.064
⋆	⋆	⋆	⋆	⋆			⋆		0.064
⋆	⋆	⋆	⋆	⋆					0.064

TAB. A.4 – δROC de l'agrégation par la méthode de la moyenne pondérée. S = Spectrum kernel, M = Mismatch kernel, P = Pairwise, PB = PairwiseBlast, LA = LA kernel, E-T = Expert Taille, E-SS = Expert Structure Secondaire, E-PI = Expert Point Isoélectrique E-MM = Expert Masse Moléculaire. Les étoiles indiquent les combinaisons de classifieurs et d'experts testés

0.12 comporte toutes E-T, à l'exception de E-SS+E-MM et E-SS+E-PI+E-MM, et aucune des associations avec un δROC de 0.06 ne comprend cet expert.

Ces résultats indiquent clairement que E-T ou l'association E-SS+E-MM, avec ou sans E-PI, sont nécessaires pour agréger correctement les classifieurs par moyenne pondérée. Ceci est confirmé par la distibution des poids sur 100 jeux de données différents (4.5 p. 164) où E-T et E-MM ont des poids très élevés.

A.5 Résultats d'agrégation par la méthode du métaSVM

Les résultats d'agrégation par le métaSVM sont présenté dans le tableau A.5.

La première observation que l'on peut faire sur ces résultats est que presque

S	M	P	PB	LA	E-T	E-SS	E-PI	E-MM	δROC
⋆	⋆	⋆	⋆	⋆		⋆		⋆	0.123
⋆	⋆	⋆	⋆	⋆	⋆	⋆			0.123
⋆	⋆	⋆	⋆	⋆	⋆	⋆	⋆	⋆	0.122
⋆	⋆	⋆	⋆	⋆	⋆	⋆		⋆	0.116
⋆	⋆	⋆	⋆	⋆		⋆	⋆	⋆	0.112
⋆	⋆	⋆	⋆	⋆	⋆	⋆	⋆		0.111
⋆	⋆	⋆	⋆	⋆	⋆		⋆		0.079
⋆	⋆	⋆	⋆	⋆		⋆	⋆		0.079
⋆	⋆	⋆	⋆	⋆	⋆			⋆	0.078
⋆	⋆	⋆	⋆	⋆	⋆				0.078
⋆	⋆	⋆	⋆	⋆		⋆			0.077
⋆	⋆	⋆	⋆	⋆				⋆	0.076
⋆	⋆	⋆	⋆	⋆			⋆	⋆	0.066
⋆	⋆	⋆	⋆	⋆	⋆		⋆	⋆	0.062
⋆	⋆	⋆	⋆	⋆			⋆		-0.020
⋆	⋆	⋆	⋆	⋆					-0.029

TAB. A.5 – δROC de l'agrégation par la méthode du métaSVM. S = Spectrum kernel, M = Mismatch kernel, P = Pairwise, PB = PairwiseBlast, LA = LA kernel, E-T = Expert Taille, E-SS = Expert Structure Secondaire, E-PI = Expert Point Isoélectrique E-MM = Expert Masse Moléculaire. Les étoiles indiquent les combinaisons de classifieurs et d'experts testés

tous les δROC sont positifs. Ce n'est toutefois pas le cas de l'association des classifieurs seuls et de ces derniers avec E-PI. Les résultats montrent que plusieurs associations obtiennent le δROC maximal.

On peux séparer les associations en trois groupes : les associations dont le δROC est environ de 0.12, les associations dont le δROC est environ de 0.075 et deux association dont le δROC est négatif. L'expert E-SS est présent dans toutes les associations du premier groupe. Seules deux associations comprenant cet expert (E-SS+E-PI et E-SS seul) sont du deuxième groupe.

Les associations du premier groupe contiennent également systématiquement E-T ou E-MM en combinaison avec E-SS. Toutefois les associations comprenant l'un de ces experts où les deux mais pas E-SS sont systématiquement du second groupe.

L'expert E-PI est présent dans des associations du premier, du second et

du troisème groupe. On notera qu'il est le seul expert impliqué dans une association où le δROC est négatif.

Ces résultats mettent en exergue l'importance d'un combinaison E-SS+E-T ou E-SS+E-MM pour obtenir les performances maximales du métaSVM. E-SS seul ainsi que E-T ou E-MM non associés à ce dernier donnent des performances moindres. E-PI semble avoir un impact faible du point de vue de cette méthode d'agrégation.

A.6 Comparaison des apports des différents experts

Les résultats des différentes agrégations montrent que E-SS est presque toujours nécessaire pour garantir de bonnes performances d'agrégation. Cela peut s'expliquer par le fait que cet expert favorise fortement les candidats ayant une structure secondaire très proche de celle de la famille. E-SS peut donc permettre de bien classer d'office les séquences remplissant ce critère, ce qui est cohérant avec la conservation de la structure secondaire dans la famille. *A contrario* E-T et E-SS, ainsi qu'on l'a vu lors de l'agrégation par le *min* semblent capables de pénaliser les candidats incongruents avec les cytokines pour ces critères. Ces deux experts ont également un impact prédominant pour plusieurs méthodes d'agrégation. Les analyses montrent qu'ils sont également parfois interchangeables, comme avec le métaSVM, ce qui confirme les résultats des tests de Kendall (**??** p **??**). E-PI s'avère quant à lui un expert apportant moins d'informations que les autres. Il est toutefois requis pour obtenir les meilleurs performances d'agrégation pour certaines méthodes.

Dans l'ensemble, ces résultats confirment la nécessité de conserver l'ensemble des experts et mettent l'accent sur l'importance de E-T, E-SS et E-MM ainsi que la complémentarité entre E-T ou E-MM et E-SS.

Bibliographie

[1] Finishing the euchromatic sequence of the human genome. *Nature*, 431(7011) :931–945, October 2004.

[2] I. Alam, A. Dress, M. Rehmsmeier, and G. Fuellen. Comparative homology agreement search : an effective combination of homology-search methods. *Proc Natl Acad Sci U S A*, 101(38) :13814–13819, Sep 2004.

[3] S. F. Altschul, T. L. Madden, A. A. Schaffer, J. Zhang, Z. Zhang, W. Miller, and D. J. Lipman. Gapped blast and psi-blast : a new generation of protein database search programs. *Nucleic Acids Res*, 25(17) :3389–3402, Sep 1997.

[4] M. J. Betts, R. Guigo, P. Agarwal, and R. B. Russell. Exon structure conservation despite low sequence similarity : a relic of dramatic events in evolution ? *EMBO J*, 20(19) :5354–5360, Oct 2001.

[5] F. Birzele, J. E. Gewehr, and R. Zimmer. Quasar–scoring and ranking of sequence-structure alignments. *Bioinformatics*, 21(24) :4425–4426, Dec 2005.

[6] J. L. Boulay, J. J. O'Shea, and W. E. Paul. Molecular phylogeny within type i cytokines and their cognate receptors. *Immunity*, 19(2) :159–163, Aug 2003.

[7] S. Brunak, J. Engelbrecht, and S. Knudsen. Prediction of human mrna donor and acceptor sites from the dna sequence. *J Mol Biol*, 220(1) :49–65, Jul 1991.

[8] W. Cai, J. Pei, and N. V. Grishin. Reconstruction of ancestral protein sequences and its applications. *BMC Evol Biol*, 4 :33, Sep 2004.

[9] S. A. Cammer, B. T. Hoffman, J. A. Speir, M. A. Canady, M. R. Nelson, S. Knutson, M. Gallina, S. M. Baxter, and J. S. Fetrow. Structure-based active site profiles for genome analysis and functional family subclassification. *J Mol Biol*, 334(3) :387–401, Nov 2003.

[10] H. Chih-Wei, C. Chih-Chung, and L. Chih-Sen. A practical guide to support vector classification. http ://www.csie.ntu.edu.tw/ cj-lin/libsvm/.

[11] L. Lo Conte, B. Ailey, T. J. Hubbard, S. E. Brenner, A. G. Murzin, and C. Chothia. Scop : a structural classification of proteins database. *Nucleic Acids Res*, 28(1) :257–259, Jan 2000.

[12] T. M. Cover. Geometrical and statistical properties of systems of linear inequalities with applications in pattern recognition. *IEEE Trans. Elect. Comp.*, 14 :326–334, 1965.

[13] P. Donnes and A. Hoglund. Predicting protein subcellular localization : past, present, and future. *Genomics Proteomics Bioinformatics*, 2(4) :209–215, Nov 2004.

[14] Cynthia Dwork, Ravi Kumar, Moni Naor, and D. Sivakumar. Rank aggregation revisited.

[15] A. J. Enright, S. Van Dongen, and C. A. Ouzounis. An efficient algorithm for large-scale detection of protein families. *Nucleic Acids Res*, 30(7) :1575–1584, Apr 2002.

[16] V. Geetha, V. Di Francesco, J. Garnier, and P. J. Munson. Comparing protein sequence-based and predicted secondary structure-based methods for identification of remote homologs. *Protein Eng*, 12(7) :527–534, Jul 1999.

[17] C. Geourjon, C. Combet, C. Blanchet, and G. Deleage. Identification of related proteins with weak sequence identity using secondary structure information. *Protein Sci*, 10(4) :788–797, Apr 2001.

[18] K. Ginalski, J. Pas, L. S. Wyrwicz, M. von Grotthuss, J. M. Bujnicki, and L. Rychlewski. Orfeus : Detection of distant homology using se-

quence profiles and predicted secondary structure. *Nucleic Acids Res*, 31(13) :3804–3807, Jul 2003.

[19] Michel Grabisch, Ivan Kojadinovic, and Patrick Meyer. A review of methods for capacity identification in choquet integral based multi-attribute utility theory : Applications of the kappalab r package. *European Journal of Operational Research*, 127(2) :766–785, April 2008.

[20] C. E. Griffiths and J. J. Voorhees. Psoriasis, t cells and autoimmunity. *J R Soc Med*, 89(6) :315–319, Jun 1996.

[21] J. Grotzinger. Molecular mechanisms of cytokine receptor activation. *Biochim Biophys Acta*, 1592(3) :215–223, Nov 2002.

[22] W. N. Grundy and T. L. Bailey. Family pairwise search with embedded motif models. *Bioinformatics*, 15(6) :463–470, Jun 1999.

[23] J. J. Haddad. Cytokines and related receptor-mediated signaling pathways. *Biochem Biophys Res Commun*, 297(4) :700–713, Oct 2002.

[24] T. Hanada and A. Yoshimura. Regulation of cytokine signaling and inflammation. *Cytokine Growth Factor Rev*, 13(4-5) :413–421, Aug 2002.

[25] E. E. Hill, V. Morea, and C. Chothia. Sequence conservation in families whose members have little or no sequence similarity : the four-helical cytokines and cytochromes. *J Mol Biol*, 322(1) :205–233, Sep 2002.

[26] Y. Hou, W. Hsu, M. L. Lee, and C. Bystroff. Efficient remote homology detection using local structure. *Bioinformatics*, 19(17) :2294–2301, Nov 2003.

[27] Ross Ihaka and Robert Gentleman. R : A language for data analysis and graphics. *Journal of Computational and Graphical Statistics*, 5(3) :299–314, 1996.

[28] T. Jaakkola, M. Diekhans, and D. Haussler. A discriminative framework for detecting remote protein homologies. *J Comput Biol*, 7(1-2) :95–114, Feb 2000.

[29] B. John and A. Sali. Detection of homologous proteins by an intermediate sequence search. *Protein Sci*, 13(1) :54–62, Jan 2004.

[30] D. T. Jones. Protein secondary structure prediction based on position-specific scoring matrices. *J Mol Biol*, 292(2) :195–202, Sep 1999.

[31] K. Karplus, C. Barrett, and R. Hughey. Hidden markov models for detecting remote protein homologies. *Bioinformatics*, 14(10) :846–856, 1998.

[32] John R. Koza. *Genetic programming : on the programming of computers by means of natural selection*. MIT Press, Cambridge, MA, USA, 1992.

[33] C. Leslie, E. Eskin, and W. S. Noble. The spectrum kernel : a string kernel for svm protein classification. *Pac Symp Biocomput*, pages 564–575, 2002.

[34] C. S. Leslie, E. Eskin, A. Cohen, J. Weston, and W. S. Noble. Mismatch string kernels for discriminative protein classification. *Bioinformatics*, 20(4) :467–476, Mar 2004.

[35] L. Liao and W. S. Noble. Combining pairwise sequence similarity and support vector machines for detecting remote protein evolutionary and structural relationships. *J Comput Biol*, 10(6) :857–868, 2003.

[36] W. W. Lin and M. Karin. A cytokine-mediated link between innate immunity, inflammation, and cancer. *J Clin Invest*, 117(5) :1175–1183, May 2007.

[37] A. V. Lukashin and M. Borodovsky. Genemark.hmm : new solutions for gene finding. *Nucleic Acids Res*, 26(4) :1107–1115, Feb 1998.

[38] J.-L. Marichal. An axiomatic approach of the discrete choquet integral as a tool to aggregate interacting criteria. *IEEE Transactions on Fuzzy Systems*, 8(6) :800–807, 2000.

[39] Florian Markowetz. Support vector machines in bioinformatics. Master's thesis, University of Heidelberg, Sept. 2001.

[40] J. Mikolajczak. *Extraction de signatures complexes pour la découverte de nouveaux membres dans des familles de protéines connues.* PhD thesis, école doctorale Chimie-Biologie, 2005.

[41] J. Mikolajczak, G. Ramstein, and Y. Jacques. Caractérisation de signatures complexes dans des familles de protéines distantes. In *proceedings, 4éme Journées Ouvertes de Biologie, Informatique et Mathématiques*, 2003.

[42] S. Mohseni-Zadeh, P. Brezellec, and J. L. Risler. Cluster-c, an algorithm for the large-scale clustering of protein sequences based on the extraction of maximal cliques. *Comput Biol Chem*, 28(3) :211–218, Jul 2004.

[43] W. S. Noble. What is a support vector machine ? *Nat Biotechnol*, 24(12) :1565–1567, Dec 2006.

[44] F. Olosz and T. R. Malek. Structural basis for binding multiple ligands by the common cytokine receptor gamma-chain. *J Biol Chem*, 277(14) :12047–12052, Apr 2002.

[45] K. Ozaki and W. J. Leonard. Cytokine and cytokine receptor pleiotropy and redundancy. *J Biol Chem*, 277(33) :29355–29358, Aug 2002.

[46] J.U. Pontius, L. Wagner, and G.D. Schuler. Unigene : a unified view of the transcriptome. contenu dans le NCBI handbook, août 2003.

[47] J. R. Quinlan. Induction of decision trees. *Mach. Learn.*, 1(1) :81–106.

[48] J. C. Renauld. Class ii cytokine receptors and their ligands : key antiviral and inflammatory modulators. *Nat Rev Immunol*, 3(8) :667–676, Aug 2003.

[49] Dymitr Ruta and Bogdan Gabrys. An overview of classifier fusion methods.

[50] H. Saigo, J. P. Vert, N. Ueda, and T. Akutsu. Protein homology detection using string alignment kernels. *Bioinformatics*, 20(11) :1682–1689, Jul 2004.

[51] Thomas Schiex. Possibilistic constraint satisfaction problems or "how to handle soft constraints ?". In *Proceedings of the Eight International Conference on Uncertainty in Artificial Intelligence*, pages 268–275, Stanford, CA, 1992.

[52] sous la direction de J.M. Cavaillon. *Les cytokines*. Masson, 1997.

[53] A. L. Tarca, V. J. Carey, X. W. Chen, R. Romero, and S. Draghici. Machine learning and its applications to biology. *PLoS Comput Biol*, 3(6) :e116, Jun 2007.

[54] H. van Dam and M. Castellazzi. Distinct roles of jun : Fos and jun : Atf dimers in oncogenesis. *Oncogene*, 20(19) :2453–2464, Apr 2001.

[55] V.N. Vapnik. *The nature of statistical learning theory*. Springer-Verlag, 1998.

[56] E. Vicenzi, P. Biswas, M. Mengozzi, and G. Poli. Role of pro-inflammatory cytokines and beta-chemokines in controlling hiv replication. *J Leukoc Biol*, 62(1) :34–40, Jul 1997.

[57] A. Wallqvist, Y. Fukunishi, L. R. Murphy, A. Fadel, and R. M. Levy. Iterative sequence/secondary structure search for protein homologs : comparison with amino acid sequence alignments and application to fold recognition in genome databases. *Bioinformatics*, 16(11) :988–1002, Nov 2000.

[58] W. WEST, James and Stacey TANNHEIMER. Heterodimeric four helix bundle cytokines, 2007. brevet numero : WO/2007/011670.

[59] Ian H. Witten and Eibe Frank. *Data Mining : Practical Machine Learning Tools and Techniques, Second Edition (Morgan Kaufmann Series in Data Management Systems)*. Morgan Kaufmann, June 2005.

[60] Z. R. Yang. Biological applications of support vector machines. *Brief Bioinform*, 5(4) :328–338, Dec 2004.

[61] A. Zemla, C. Venclovas, K. Fidelis, and B. Rost. A modified definition of sov, a segment-based measure for protein secondary structure prediction assessment. *Proteins*, 34(2) :220–223, Feb 1999.

[62] E. Zitzler, M. Laumanns, and S. Bleuler. A Tutorial on Evolutionary Multiobjective Optimization. In X. Gandibleux et al., editors, *Metaheuristics for Multiobjective Optimisation*, Lecture Notes in Economics and Mathematical Systems. Springer, 2004.

[63] M. H. Zweig and G. Campbell. Receiver-operating characteristic (roc) plots : a fundamental evaluation tool in clinical medicine. *Clin Chem*, 39(4) :561–577, Apr 1993.

Résumé

L'objectif de ce travail est la mise au point d'une méthode de détection d'homologues de cytokines inconnues. J'ai dans un premier temps évalué plusieurs classifieurs SVM. J'ai ensuite proposé d'ajouter, sous la forme d'experts automatiques, des connaissances spécifiques à la famille étudiée. Enfin, afin de maximiser l'efficacité de leur association, j'ai comparé différentes méthodes d'agrégation. Je propose une méthode performante, basée sur la combinaisons de ces classifieurs et de ces experts, généralisable à d'autres familles de protéines.

Identification of four helix cytokine homologs by aggregation of classifiers and automated experts

abstract :

I was working on a particular gene family : the four helix cytokines. The major purpose of this work was to design a new method to detected still unknown members in the human genome. The first part of my work was to compare SVM classifiers, which is known as the best strategy for homologs research, from the literature. During the second part of my work, i designed automatical experts which deals with information like biological features. The last part of my work consisted in evaluating methods to aggregate classifiers and experts. This strategy achieve better results than the best classifer alone and it can easily be adapted to other gene family.

Mots clés : Cytokines à quatr hélices α, recherche d'homologues, classification, SVM, caractéristiques spécifiques, experts automatiques, agrégation

Key words : Four helix cytokines, homologs identification, classification, SVM, specific features, automated experts, aggregation

Nicolas Beaume, équipe "cytokines et récepteurs", INSERM U892 9 quai Moncousu 44093 Nantes cedex 1 / Equipe COD LINA, EPUN, rue Christian Pauc 44000 Nantes .