UNIVERSITE DE NANTES

FACULTE DE MEDECINE

EXTRACTION DE SIGNATURES COMPLEXES POUR LA DECOUVERTE DE NOUVEAUX MEMBRES DANS DES FAMILLES DE PROTEINES CONNUES

THESE DE DOCTORAT

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

présentée
et soutenue publiquement par

MIKOLAJCZAK Jérôme

Le 04 mai 2005, devant le jury ci-dessous

Président **M. LEGER Jean, Directeur de Recherche, INSERM U533, NANTES**

Rapporteurs **M. GUENOCHE Alain, Chargé de Recherche CNRS, IML, MARSEILLE**
M. PINEAU Charles, Chargé de Recherche, INSERM U625, RENNES

Examinateurs **M. GASCAN Hugues, Directeur de Recherche, INSERM U564, ANGERS**
M. RAMSTEIN Gérard, MC, Ecole Polytechnique Universitaire de NANTES

Directeur de thèse **: M. JACQUES Yannick, Directeur de Recherche, INSERM U601, NANTES**

Résumé

Cette thèse se situe à l'interface de la biologie et de l'informatique et plus précisément dans le domaine de l'extraction des connaissances à partir des données biologiques.

L'objectif principal de ce travail a été d'obtenir des modèles de classification pour notre ensemble de familles de cytokines d'intérêt qui rassemble les familles des interleukines à chaîne courtes, des interleukines à chaînes longues, et des interleukines à six hélices α « IL-10/interférons ».

Notre démarche réside en l'élaboration de stratégies de caractérisation de nos familles d'interleukines au moyen d'un ensemble de signatures caractéristiques. Nous nous sommes inspirés des connaissances acquises en apprentissage artificiel et classification des données dans le but d'obtenir des modèles de classification fiables et performants.

La première partie de ce travail a consisté à établir une approche génétique en trois étapes. La première phase est consacrée à la découverte de motifs hiérarchiques basés sur une classification hiérarchique des acides aminés en fonction de leurs propriétés physico-chimiques. Dans l'étape suivante les signatures sont ensuite définies par les séquences des motifs précédents. Finalement, l'ensemble des signatures est optimisé. Nous montrons que l'ensemble des signatures optimal respecte les contraintes de cardinalité, de support et de spécificité et cible spécifiquement notre ensemble d'intérêt.

La deuxième partie de ce travail à permis d'élaborer une approche discriminante dont l'originalité repose sur l'utilisation d'un algorithme de découverte de motifs suivant le paradigme de la classification hiérarchique. L'ensemble des motifs définit un espace de représentation dans lequel les interleukines sont représentées par un vecteur indiquant la présence ou l'absence de chaque motif dans leurs séquences respectives. Nous utilisons la technique d'apprentissage par vecteur de support (SVM) pour discriminer notre famille d'intérêt par rapport à des séquences non apparentées. Nous montrons que le modèle SVM issu de notre méthode de classification est plus performant sur notre famille de protéines que d'autres méthodes de classification.

Ce travail a donc contribué à la caractérisation de nos familles d'interleukines d'intérêts et ouvre la voie à des travaux d'extraction de séquences candidates à partir des données disponibles dans les bases de données génomiques publiques.

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.

The major purpose of this work was to obtain classification models from our cytokines families set of interest which assembles the short chains interleukins, the long chains interleukins and the six helices “IL-10/interferons” family.

Our reasoning consisted in the conception of characterization strategies of our interleukins families by the way of a set of representatives signatures. We used the scientific knowledge from Machine Learning and Data Classification to obtain robust and powerful models of classification.

The first part of this work was the conception of a genetic approach following a three steps process. Firstly, a top-down strategy is executed to extract hierarchical patterns based on an alphabet including the amino acid set and their own physicochemical properties. Secondly, a discovery algorithm of sequential itemsets searchs for sequence of patterns. Finally the set of signature found in the previous step is reduced. We show that the optimal set of signatures respects the constraints of cardinality, completeness and specificity and specifically targets our biological set of interest.

The second part of our work consisted in a discriminative approach which has the originality to propose an algorithm for discovering motifs based on the ascending hierarchical paradigm. The set of motifs defines a feature space of sequences where each interleukin is transformed into a vector that indicates the possible presence of the motifs on their sequences. We used the Support Vector Machines (SVM) to discriminate our set of interest versus counter-examples from other distinct protein families. We show that the SVM model obtained with this method performs better on our family set than other classification methods used in remote protein classification.

This work contributes to the characterization of our interleukin families of interest and opens the way to future studies toward the extraction of putative sequences of interleukin from the genomic data available in the public databases.

Table des matières

Notations cytokines

β-TG : β-thromboglobuline

Act : Activines

AICD : Activation Induced Cell Death

BCL-2 : B-Cell Lymphoma 2

BCL-XL : Bcl-2 like 1 protein

BDNF : Brain-Derived Neurotrophic Factor

BMP : Bone Morphogenetic Proteins

BSF-3 : B cell Stimulating Factor 3

BTC : betacelluline

CPA : cellule présentatrice de l'antigène

CBM : Cytokine Binding Module

CD27L : Cluster of Differenciation 27L

CHR : Cytokine binding Homology Region

CLC : Cardiotrophin Like Cytokine

CMH : complexe majeur d'histocompatibilité

CNTF : Ciliary Neurotrophic Factor

CNTF-R : Ciliary Neurotrophic Factor Receptor

CRF2 : Cytokine Receptor Class II

CRH : cytokine receptor homolog

CT-1 : cardiotrophine 1

Notations cytokines

CTLA-4 : Cytotoxic T Lymphocyte associated Antigen 4

EGF : Epidermal Growth Factor

EPO : érythropoïétine

F4P : facteur 4 plaquettaire

FADD : Fas-Associated Death Domain

FasL : Fas Ligand

FVIIa : facteur de coagulation VIIa

FGF : Fibroblast Growth Factor

FERM : Four-point-one, Erzin, Radixin, Moesin

FISP : IL-4-Induced Secreted Protein

FKN : Fractalkine

FLIP : FLICE/Caspase-8 Inhibitory Protein

FNIII : Fibronectin III

GAS : Gamma-Activated Sequence

G-CSF : Granulocyte Colony Stimulating Factor

GH : Growth Hormon

GM-CSF : Granulocyte Macrophage Colony Stimulating Factor

GPI : Glycosylphosphatidylinositol

Grb2 : Growth factor receptor-bound protein 2

HB-EGF : Heparin Binding-EGF like Factor

IGF : Insulin Like Growth Factor

IL- : interleukine

IL1-RA : IL-1 Receptor Antagonist

IL-TIF : IL-10 related T-cell-derived inducible factor

IL-22BP : IL-22 Binding Protein

IFN : interféron

Inh : Inhibine

IP10 : gamma interferon inducing protein

IRF-9 : IFN regulatory factor 9

ISGF3 : IFN-stimulated gene factor 3

ISRE : Interferon-Stimulated Response Elements

JAK : Janus Associated Kinase

KGF : Keratinocyte Growth Factor

LAK : Lymphokine-Activated Killer

LCR : Locus Control Region

Lep : leptine

LIF : Leukemia Inhibitory Factor

LPS : lipopolysaccharides

LTA : Lymphotoxine Alpha

LTN : Lymphotactine

MAPK : Mitogen-Activated Protein Kinase

MAPKKK : MAPK Kinase Kinase

Notations cytokines

MCP : Monocytes Chemoattractant Proteins

M-CSF : Macrophage Colony Stimulating Factor

MDA-7 :Melanoma Differenciation associated protein 7

MDC : Macrophage-Derived Chemottractant

MEK : MAPK/ERK Kinase

MIG : Monokine Induced by Gamma Interferon

MIP : Macrophage Inflammatory Protein

MIP1α/β : Macrophage Inflammatory protein 1α et β

MMP : métalloprotéinase

NF-κB : Nuclear Factor Kappa B

NGF : Nerve Growth Factor

NK : Natural Killer

NNT-1 : Novel Neurotrophin 1

NOE RMN : résonance magnétique nucléaire à effet overhauser nucléaire

Ob : Obesity Factor

OSM : Oncostatin

PARC : Pulmonary and Activation-Regulated Chemokine

PBP : Platelet Basic Protein

PDGF : Platelet Derived Growth Factor

PDK1 : 3-Phosphatidyl inositol-Dependent protein Kinase 1

PGN : peptidoglycane

PIAS : Protein Inhibitor of activated STAT

PIP2 : Phosphatidyl inositol 4,5-biphosphate

PIP3 : Phosphatidyl inositol 4,5-triphosphate

PMN : polymorphonucléaire

PRL : prolactine

RANTES : Regulated Upon Activation Normal T Expressed and Secreted

SBE : STATS Binding Element

SCF : Stem Cell Factor

SDF1 : Stromal cell-Derived Factor 1

SH2 : Src-Homology 2

SHc : SH2-containing collagen-related protein

SHP1 : SH2-containing Phosphatase 1

SHP2 : SH2-containing Phosphatase 2

SOCS1 : Supressor Of Cytokine Signaling 1

Sos : Son of sevenless

STAT : Signal Transducers and Activator of Transcription

TARC : Thymus and Activation-Regulated Chemokine

TGF β : Transforming Growth Factor β

TH (lymphocyte) : T Helper

TIMP : Tissu inhibitor of métalloprotéinases

TNF : Tumor Necrosis Factor

Notations cytokines

TPO-R : thrombopoïétine

TPO-R : Thrombopoitin Receptor

TSLP : Thymic Stromal Lymphopoietin protein

VEGF : Vascular Endothelial Growth Factor

WSX-1 : IL-27 receptor

Notations SVM

SVM : Support Vector Machine (= Séparateur à Vaste Marge)

χ : ensemble des exemples

x : exemple

u : classe de l'exemple

N_x : cardinal de l'ensemble des exemples

h : hypothèse de classification

F : ensemble des fonctions de décision

f : fonction de décision

R : risque réel de mauvaise classification

L : fonction de coût

$Remp$: risque empirique de mauvaise classification

G : borne maximale de complexité

dVC : dimension Vapnik-Chervonenkis

ε : écart entre le risque réel et le risque empirique

w : vecteur de poids

b : biais

F : espace de redescription

Φ : fonction de redescription

$f(w)$: fonction objective

$g_i(w)$: contraintes d'inégalité

Notations SVM

$h_j(w)$: contraintes d'égalité

w : variable primale

Ω : ensemble de définition de w

w^* : solution du problème d'optimisation primal

$\Re^n$: ensemble des réels de dimension n

C^1 : classe de fonction

$L(w,\alpha)$: fonction Lagrangienne

α_i : multiplicateur de Lagrange

α^* : solution du problème d'optimisation dual

β_j : multiplicateurs de Lagrange pour les contraintes d'inégalité

K : fonction noyau

$K(x_i, x_j)$: fonction noyau

K_{ij} : matrice des fonctions noyaux

ξ_i : variable ressort

C : poids d'erreur, constante de régularisation,

w_0 : biais de l'hyperplan optimal dans l'expression primale du problème des SVM

w_0^* : biais obtenu pour tout exemple critique (x_c, u_c) dans l'expression duale du problème d'optimisation des SVM

x_c : exemple critique (vecteur support)

u_c : classe de l'exemple critique

Introduction

1.1. L'extraction des connaissances à partir des données biologiques

Depuis l'apparition en biologie des techniques d'expérimentation à haut débit (le système double hybride, le séquençage à haut débit des « Sequence Tagged Sites » en génomique, les « Expressed Sequence Tags », les méthodes de spectrométries de masse en protéomique, les puces à ADN et les données d'expressions) et de la mondialisation des informations grâce à Internet, nous vivons une explosion de la quantité de données issues des expériences de biologie, de médecine et des biotechnologies. Cette évolution a atteint une telle vitesse qu'il n'est plus possible depuis longtemps de se passer des outils informatiques pour le stockage des données, leur traitement et leur analyse.

Le domaine de la bioinformatique participe à ce mouvement de deux manières :

- Le développement des outils informatiques destinés aux spécialistes des domaines scientifiques concernés (biologiste, médecins) pour la manipulation de leurs données,

- le travail de recherche sur les données « in silico » pour extraire de nouvelles connaissances scientifiques.

En informatique, l'apprentissage automatique propose des solutions pour répondre aux problèmes de classification des données, conjointement aux domaines de la fouille de données et de l'extraction des connaissances. Ces dernières années, les avancées scientifiques dans cette discipline ont permis d'aboutir à des méthodes de classifications de plus en plus performantes qui n'ont jamais tardées à trouver des applications en bioinformatique, que ce soit dans la recherche de séquences promotrices dans les génomes, les analyses d'expressions des puces à ADN ou la recherche de séquences homologues lointaines. On distingue les techniques d'apprentissage dites génératives, car basées sur l'apprentissage à partir d'un ensemble monoclasse (par exemple les modèles de Markov cachés « HMM ») et les méthodes discriminantes qui utilisent un jeu d'apprentissage multiclasse (réseaux de neurones, SVM). Souvent, les ensembles de données des méthodes discriminantes sont biclasses : une classe d'exemples et une classe de contre-exemples.

1.2. Les cytokines et les trois sous-familles des interleukines à chaîne longues, à chaînes courtes et IL-10/IFNs

Les membres de la superfamille des cytokines sont des glycoprotéines solubles de faibles poids moléculaires qui interviennent dans la régulation de la réponse immunitaire. Elles ont pour fonction d'assurer la médiation des signaux de prolifération, de différenciation, et d'activation entre les différentes cibles cellulaires. Nous nous intéressons plus particulièrement aux interleukines à hélices courtes (IL-2), hélices longues (IL-6) et aux interleukines à six hélices de la famille des IL-10/IFNs. La fonction pivot des cytokines dans l'activation des voies de l'immunité ainsi que dans l'apoptose leur confère un intérêt de tout premier plan dans la compréhension des mécanismes de la défense du soi. Leur implication dans de nombreuses formes de maladies autoimmunes, d'autoinflammation et de cancers, confirme l'importance des efforts de recherche entrepris dans la découverte et l'amélioration des traitements thérapeutiques.

1.3. Classification et apprentissage automatique

Qu'il s'agisse de données biologiques ou non, la classification s'articule autour de plusieurs étapes importantes :

- La sélection des données. Il s'agit de sélectionner l'ensemble des données de départ les plus intéressantes en fonction des objectifs à atteindre.

- Le prétraitement des données. Les données brutes récupérées ne sont pas toujours directement exploitables. Certaines techniques nécessitent une mise en forme des données en entrée (tableau de données, vecteur). D'autres ne fonctionnent que sur des données catégorielles ou numériques. Il est souvent nécessaire de filtrer les données dupliquées et les données manquantes, et d'optimiser les attributs.

- La fouille de données. Cette étape concerne dans notre cas la construction d'une hypothèse de classification qui servira à prédire la classe des données présentes dans notre échantillon. Elle est difficile à mettre en oeuvre et nécessite de tester la qualité de l'hypothèse de

classification en accord avec les objectifs qui ont été définis. L'évaluation de la qualité de l'hypothèse de classification peut prendre en compte différents critères tels que la rapidité de son obtention et de sa mise en œuvre, sa précision, sa compréhension par l'utilisateur, sa fiabilité, et la stabilité de sa performance dans le temps.

- L'interprétation et l'évaluation des connaissances acquises. Dans cette étape la performance du classifieur et la prédiction sont validées par des méthodes statistiques ou par l'expert. Les résultats permettent d'optimiser l'hypothèse de classification en améliorant les étapes précédentes.

1.4. Motivation de la thèse

La recherche d'homologies éloignées est un axe de recherche en bioinformatique dont le but est d'extraire des homologies de plus en plus faibles entre des membres à partir de l'information contenue dans les séquences. Ce travail nécessite la mise au point de stratégies et d'algorithmes capable de descendre sous le seuil à partir duquel il n'est plus possible de corréler le taux de similarité entre deux gènes ou deux protéines et leur niveau d'homologie. Certaines familles de protéines telles que les cytokines sont caractérisées par des taux de similarités de séquences faibles entre les membres au point de ne pas pouvoir obtenir de résultats avec les algorithmes d'alignement de séquences standard. Parallèlement, on a regroupé sous le terme d'apprentissage automatique des algorithmes de classification de plus en plus performants. Les techniques de classifications supervisées basées sur l'apprentissage à partir d'exemples et de contre-exemples, telles que les SVM, ont prouvé leur robustesse et leur performance dans cette problématique. L'objectif principal de ce travail de thèse est de mettre en œuvre de nouvelles stratégies de classification automatique pour aboutir à des modèles de classification des familles d'interleukines IL-2, IL-6 et IL-10/IFNs qui seront utilisés dans l'extraction de nouveaux candidats à partir des banques génomiques publiques.

1.5. Contribution à la thèse

L'innovation de cette thèse consiste à avoir obtenu des modèles de classification de l'ensemble des sous-familles d'interleukine d'intérêts au moyen de deux stratégies différentes. Nous avons défini les signatures comme étant constituées par des suites de motifs vérifiant les contraintes de cardinalité, de support et de spécificité sur la famille de protéines d'intérêt. Nous avons utilisé une hiérarchie basée sur une classification des acides aminés en fonction de leurs propriétés physico-chimiques pour caractériser l'ensemble des protéines d'intérêt par un ensemble de signatures et par un ensemble de motifs spécifiques. La première méthode se base sur des algorithmes génétiques et a pour objectif d'extraire l'ensemble optimal des signatures à partir d'un premier ensemble de motifs. La deuxième méthode consiste à effectuer une classification hiérarchique ascendante sur les séquences pour obtenir l'ensemble des motifs qui vérifient le mieux les membres de notre famille d'intérêt. Ces motifs permettent la vectorisation d'exemples et de contre-exemples avant l'apprentissage d'un modèle de classification discriminante basée sur les machines à vecteurs de support. Ces deux stratégies ont montré leurs potentiels respectifs à caractériser l'ensemble des trois familles d'interleukines humaines et leur très bonne sélectivité sur un ensemble de contre-exemples par rapport à d'autres méthodes de classifications.

1.6. Plan

La première partie de ce mémoire est dédiée à la superfamille des cytokines, en particulier celles présentes chez l'homme. Le chapitre 2 sert d'introduction à ces protéines solubles impliquées dans la transmission des signaux de la réponse immunitaire. Nous y décrivons leurs fonctions biologiques et leurs voies de signalisation. Les chapitres suivants sont consacrés aux trois familles structurales d'interleukines qui nous intéressent : les interleukines à chaînes longues (IL-6), les interleukines à chaînes courtes (IL-2) et les interleukines du type IL-10/interférons. Le chapitre 3 détaille les classifications des interleukines en fonction de la nature de leurs récepteurs membranaires respectifs et de la structure de leurs gènes dans le génome humain. Les chapitres 4, 5 et 6 décrivent à tour de rôle les structures tridimensionnelles conservées chez les interleukines de nos trois sous-familles d'intérêt.

A la suite de la première partie plutôt axée vers la biologie, nous passons dans le domaine de l'informatique avec une deuxième partie consacrée à l'apprentissage automatique, domaine

dans lequel se situe davantage les idées présentes dans cette thèse. Le chapitre 7 traite de la classification des données. Les thèmes suivants sont décrits : la nature des données, le travail sur les attributs, les méthodes de classification supervisées et non supervisées ainsi que les techniques de tests de performances sur les classifieurs. Le chapitre 8 est consacré aux algorithmes génétiques que nous utilisons dans notre stratégie de caractérisation des signatures. Le chapitre 9 fait une présentation détaillée de la technique de classification discriminante des machines à vecteurs de support SVM que nous utilisons dans notre deuxième méthode pour classifier nos séquences exemples vectorisés.

La troisième et dernière partie de ce mémoire concerne les méthodes et résultats obtenus pendant cette thèse et se découpe en quatre chapitres. Le chapitre 10 est destiné à donner les clefs de ce travail au moyen des définitions des concepts de motifs hiérarchiques et de signatures ainsi que la présentation de la classification des acides aminés que nous avons utilisée pour ce travail. Le chapitre 11 est consacré à l'explication de notre première méthode d'extraction de signatures complexes au moyen d'algorithmes génétiques et aux résultats que nous avons obtenus sur une base de séquences de contre-exemples n'appartenant pas aux trois familles d'interleukines. Le chapitre 12 décrit notre deuxième méthode basée sur la vectorisation des séquences exemples au moyen d'un ensemble de motifs hiérarchiques et servant à l'apprentissage d'un modèle de classification SVM. Nous montrons les résultats obtenus sur la performance de notre méthode par rapport à d'autres méthodes basées sur des SVM et une implémentation des « k plus proches voisins ». Le chapitre 13 présente une comparaison entre les deux algorithmes. Ce travail permet de définir les conditions pour lesquelles chaque méthode est adaptée.
Le chapitre final de cette thèse constitue la conclusion des travaux menés au cours de ces trois ans de thèse et les perspectives amenés par ce travail.

Partie 1. Les cytokines

Chapitre 2. La superfamille des cytokines

Les cytokines sont des glycoprotéines solubles de faible poids moléculaire (15 à 30 kDa) qui interviennent dans la régulation de la réponse immunitaire chez les mammifères. Elles ont pour fonction d'assurer la médiation des signaux de prolifération cellulaire, de différenciation et d'activation entre les différentes cibles cellulaires. On peut estimer à 130 le nombre des protéines classées comme cytokines pour l'espèce *Homo sapiens*. Les cytokines agissent sur leurs cibles cellulaires par l'intermédiaire de récepteurs spécifiques exprimés à la surface de ces cibles. Ces récepteurs ont pour rôle d'activer les voies de transduction du signal à l'intérieur de la cellule.

Les cytokines sont sécrétées principalement par les cellules impliquées dans les réponses immunitaires (Lymphocytes T, Lymphocyte B ou les cellules présentatrices de l'antigène), mais aussi dans d'autres types cellulaires. Chacune peut agir sur différents types cellulaires et selon différents modes d'actions possibles : paracrine (action au voisinage des cellules sécrétrices, autocrines (action sur les cellules sécrétrices), juxtacrine (action impliquant un contact cellulaire entre la cellule émettrice et la cellule réceptrice) ou endocrine (le médiateur utilise la circulation sanguine pour atteindre les cellules cibles).

Ces propriétés les distinguent des hormones qui sont secrétées par un seul type de cellule et agissent à distance (endocrine) sur un type de cible unique.

Le rôle important que les cytokines jouent dans l'activation des voies de la réponse immunitaire ainsi que dans l'apoptose (la mort cellulaire programmée) leurs confère un intérêt de tout premier plan dans la compréhension des mécanismes de défense du soi, pour la découverte et l'amélioration des traitements actuels dans le contexte des maladies inflammatoires, autoimmunes, infectieuses et le cancer.

Les cytokines sont classées par familles en fonction de la nature des gènes qui les expriment, de la classe de leur récepteurs, du type cellulaire qui les sécrètent. Il est aussi possible de les classer en fonction de leur repliement tridimensionnel dans l'espace à partir des observations obtenues par les méthodes de résonance magnétique nucléaire, de cristallographie aux rayons X ou de modélisation structurale par homologie. Plus précisément, les interleukines étudiées dans cette thèse possèdent des structures tertiaires similaires à un repliement consensus alors

qu'elles ne présentent que de faibles homologies de séquences primaires. Les sous-familles d'interleukines IL-6, IL-2, IL10/interférons se caractérisent au niveau structural par un repliement « up-up-down-down » de quatre hélices α .

La classe biologique des interleukines regroupe les cytokines dont la synthèse a été initialement observée chez différents types de leukocytes. La régulation de l'expression des interleukines est très stricte et rarement constitutive. La sécrétion se fait souvent en réponse à l'activation de la cellule par un stimulus qui peut être par exemple une autre interleukine.

Nous traitons dans ce chapitre des connaissances biologiques sur la superfamille des cytokines et des classifications disponibles pour ces nombreux membres. Nous présenterons ensuite les trois sous-familles d'interleukines étudiées qui sont celles des interleukine 6, interleukine 2 et interleukine 10/interférons.

2.1. Définition de la superfamille des cytokines et de la famille des interleukines

L'étymologie du nom cytokine provient du nom grec *kutos*, cellule et du verbe *kineo*, stimuler. La première cytokine découverte est l'interféron α en 1957 (Isaacs and Lindenmann 1957). C'est ensuite l'interféron γ qui a été décrit en 1965 (Wheelock 1965) . Le terme lymphokine a été proposé en 1969 pour caractériser les facteurs immunitaires qui n'étaient ni des anticorps ni des molécules du complément, et exprimés par les cellules lymphocytaires activées par un antigène. Le terme monokine a permis de distinguer les médiateurs exprimés par les monocytes et les macrophages. L'appellation « cytokine » a été proposée à partir de 1979, lorsqu'une activité lymphokine a été observée dans des cultures cellulaires de rein infectées par le virus de la fièvre catharrale maligne du mouton (virus *blue tongue*) . Les cytokines représentent alors une grande classe de médiateurs impliqués dans la régulation de la défense du soi et sécrétés par différents types de cellules.

Les cytokines sont des glycoprotéines de faible poids moléculaires. La grande majorité de ces molécules sont solubles. Cependant, certaines cytokines ainsi que des isoformes possèdent des séquences hydrophobes à leur extrémité N-terminale qui permettent leur ancrage aux membranes des cellules qui les expriment (Gordon 1991; Bosenberg and Massague 1993).

Ces molécules ont une action de messager ou de médiateur des signaux de l'immunité. La transduction du message peut se faire de façon autocrine, paracrine et juxtacrine dans la plupart des cas, moins souvent de manière endocrine. Les cytokines font preuve de pléiotropie, c'est à dire que chacune d'entre elles possèdent plusieurs cibles cellulaires différentes. Elles sont aussi caractérisées par leur redondance fonctionnelle dans les voies de régulation des mécanismes de l'immunité, l'activité inflammatoire ou la croissance des cellules. En effet, plusieurs cytokines peuvent médier le même signal et il existe de nombreux cas pour lesquels l'inhibition ou la neutralisation d'une cytokine n'empêche pas la transduction du signal du fait de cette redondance.

Les résultats expérimentaux qui ont montré que la régulation de l'expression d'une cytokine est souvent modifiée par d'autres cytokines ont permis d'élaborer le concept du « réseau des cytokines » (Balkwill and Burke 1989) et d'expliquer leurs effets additifs, synergiques ou antagonistes dans la médiation du signal. Les cytokines sont exprimées par les cellules de l'immunité : les leukocytes inflammatoires, les globules blancs, ainsi que quelques cellules non leukocytaires telles que des fibroblastes, et des cellules endothéliales. Leur action peut s'effectuer à des concentrations faibles par leur interaction sur des récepteurs extra cellulaires spécifiques. L'interaction de la cytokine sur son récepteur est en général de haute affinité. Cependant les cytokines n'ont pas de fonctions enzymatiques.

Il est possible de classer les cytokines en fonction de leur activité biologique : les interleukines, les facteurs de stimulation des colonies, les interférons, les facteurs de nécrose des tumeurs, les facteurs de croissances et les chimiokines. Il existe aussi une classification qui tient compte de la classe des récepteurs ainsi qu'une classification en fonction du type de repliement dans l'espace.

Le nom « interleukine » désigne une famille de cytokines dont les membres sont exprimés par les cellules leukocytaires, en particulier par les lymphocytes ainsi que par d'autres types cellulaires tels que les phagocytes polymorphonucléaires (PolyMorphoNuclear phagocytes) ou des cellules auxiliaires. La nomenclature des interleukines correspond à l'abréviation « IL- » suivie d'un nombre. Ces protéines possèdent un large spectre d'activité. Elles sont impliquées dans la croissance et la différenciation cellulaire. Les interleukines sont des éléments très importants du réseau des cytokines dans les voies de l'immunité et de l'inflammation. Chaque interleukine agit sur un nombre limité et spécifique de cellules cibles qui expriment les récepteurs spécifiques à leur surface.

2.2. Fonctions biologiques des cytokines

2.2.1. Dans les réponses immunitaires

Les cytokines de la réponse immunitaire comprennent l'ensemble des interleukines ainsi que l'IFN-γ et les deux formes du facteur de nécrose des tumeurs TNF-α et TNF-β.

Les deux sous populations de lymphocytes T auxiliaires T CD4 (helper) jouent un rôle central dans la régulation de la réponse immunitaire . Les TH1 et TH2 se différencient par la nature des cytokines secrétées . La population primitive TH0 exprimerait les IL-2, IFNγ, IL-4 et IL-5. Les lymphocytes TH1 sécrètent IL-2, IFNγ et TNF-β et les lymphocytes TH2 sécrètent IL-4, IL-5, IL-6, IL-10 et IL-13 . L'IL-3, le TNF-α et le GM-CSF sont exprimés par les deux sous-populations . Lorsque le lymphocyte TH2 cesse de produire IL-2, il utilise l'IL-4 comme facteur de croissance.

La sous population des lymphocytes régulateurs TH3 CD4+CD25+ sécrète le TGFβ, l'IL-10 et l'IL-4. Cette population lymphocytaire possède la propriété de diminuer les réactions d'autoimmunité.

Les lymphocytes TH1 vont être spécialisés dans l'immunité cellulaire tandis que la sous population TH2 est impliquée préférentiellement dans la réponse humorale . La différentiation des lymphocytes TH0 primitifs s'effectue dans le thymus. La voie de différenciation pour les TH1 est induite au contact de l'IL-12, relayée par le facteur de signalisation STAT4. L'IL-4 joue le rôle de facteur de différenciation pour les TH2. Le signal de différenciation est dans ce cas relayé par STAT6.

L'IL-2 est le facteur de croissance des lymphocytes T. Elle agit de manière autocrine et active l'expression de l'IFN-γ et de la lymphotoxine. L'IL-2 agit aussi sur la croissance des cellules NK et les lymphocytes B et sur l'activité des cellules NK.

L'IL-9 est aussi un facteur de croissance des lymphocytes. Elle est sécrétée par les lymphocytes T CD4 et les mastocytes mais pas par les lymphocytes T CD8. L'IL-15 est la

cytokine principale dans la différenciation des cellules NK. Elle partage aussi des propriétés redondantes avec l'IL-2.

2.2.2. Dans les maladies autoimmunes

L'autoimmunité correspond au dérèglement du système immunitaire qui entraîne une réponse immunitaire contre le soi. Les maladies autoimmunes sont souvent dues à l'activation anormale des lymphocytes. Les macrophages et les cellules dendritiques peuvent aussi jouer un rôle dans ces pathologies. Les cytokines jouent un rôle dans le développement des cellules immunitaires, l'immunorégulation, et les fonctions des effecteurs de l'immunité. Cependant les mécanismes de l'autoimmunité sont complexes et les cytokines pro-inflammatoires susceptibles d'activer la réponse immunitaire peuvent avoir des propriétés immunosuppressives (Tableau 1). On observe ce partage des propriétés chez l'IL-2, les IFNs et le TNF-α (O'Shea, Ma et al. 2002).

Tableau 1. Exemples d'interleukines impliquées dans des maladies

Groupes pathologiques	type pathologique	Cytokines impliquées	utilisation thérapeutique
Inflammation	Arthrite rhumatoïde	TNF, IL-1, IL-6, IL-8	Antagoniste du TNF
	Maladie de Crohn	TNF, IL-1, IL-8, MCP-1	Antagonistes du TNF
	Sclérose en plaques	TNF, IL-1, IL-6, IL-10, Mig, RANTES, IL-2, IFN-beta	IFN-beta
	Psoriasis	TNF, IL-8	
	Fibrogenic lung disease	TNF, TGF-alpha, TGF-beta, PDGF	
Allergies	Asthme	IL-4, eotaxin, IL-15	
Infections	choc endotoxique	TNF, IL-1, IL-6, HMG-1, IL-8, IL-10	Antagonistes du TNF
	Méningite	TNF, IL-1, IL-8, MCP-1, IP-10	
	Pneumonie bactérienne	IL-8, IP-10	
	Malaria cérébrale	TNF, IL-1, IL-6	
Cancer	Cancer des ovaires	TNF, MCP-1	IFN-alpha
	Myeloma	IL-6, IL-1beta	

L'IL-2 joue le rôle pivot dans l'homéostasie lymphoïde et la tolérance périphérique. Si la déficience d'IL-2 conduit dans la plupart des cas à l'apparition de maladies autoimmunes, il n'a pas été mis en évidence que l'absence d'IL-2 est le principal mécanisme sous-jacent aux maladies autoimmunes communes. En parallèle de ces fonctions proinflammatoires, l'IL-2

promeut l'apoptose des lymphocytes en activant les membres de la famille BCL-2 (Lenardo 1991). Cette interleukine active aussi l'AICD Fas dépendante (Activation Induced Cell Death) (Refaeli, Van Parijs et al. 1999; Bleesing, Straus et al. 2000). IL-2 active la transcription de Fas, FasL et FADD (Refaeli, Van Parijs et al. 1998). Elle effectue un contrôle négatif de l'expression de la protéine FLIP (FLICE/Caspase-8 Inhibitory Protein) qui interfère avec l'apoptose médiée par Fas (Van Parijs, Refaeli et al. 1999).

Il existerait un lien entre l'IL-2 et la sous population de lymphocytes TH3 CD4+CD25+ (Suzuki, Zhou et al. 1999; Sakaguchi 2000; Wolf, Schimpl et al. 2001). Cette sous population lymphocytaire est absente chez les souris KO IL-2 -/-. Les lymphocytes CD4+CD25+ produisent de façon constitutive la protéine CTLA-4 (Cytotoxic T-Lymphocyte-Associated Protein 4) (Suzuki, Zhou et al. 1999). Cette molécule est sur-régulée par l'IL-2. Les souris CTLA-4 déficientes développent une maladie lymphoprolifération fatale (Oosterwegel, Greenwald et al. 1999).

L'IL-2 induit aussi la production d'IFN-γ et de TNF qui ont eux-mêmes des propriétés immunosuppressives en plus de leurs activités proinflammatoires respectives .

L'activité proinflammatoire du TNF a été décrite en particulier pour l'activation des macrophages, des cellules endothéliales, des synoviocytes et d'autres types de cellules. Le TNF induit la production des cytokines proinflammatoires IL-6 et IL-1 ainsi que des chimiokines. L'inhibition du TNF s'est montrée efficace pour le traitement de la maladie de Crohn, pour l'arthrite rhumatoïde (Kollias, Douni et al. 1999; Owens, Wekerle et al. 2001) ou sur le diabète de type 1. La surproduction de TNF a été observée lors de l'arthrite rhumatoïde et la sclérose en plaques. On observe les effets immunosuppresseurs du TNF sur l'inhibition de la signalisation du TCR (Campbell, O'Donnell et al. 2001), la promotion de l'apoptose des cellules T lymphoïdes, l'inhibition des cellules dendritiques. Le TNF- induit la sécrétion d'IL-10, d'IL-6 et et de TGF-β chez les lymphocytes TH2. En induisant IL-6, le TNF active la différenciation des lymphocytes TH2 dépendants de l'IL-4 et atténue la différenciation des lymphocytes TH1 IL-4 indépendants mais SOCS1 (Supressor Of Cytokine Signaling 1) dépendants.

Les IFN-α/β appartiennent à la classe des interférons de type 1. Leur activité immunostimulatrice intervient dans la réponse immunitaire antivirale. Les interférons sont

administrés au cours des traitements anti-infectieux avec des risques de complications vers la déclaration de maladies autoimmunes.

Ces interférons sont impliqués dans l'activation de la présentation à l'antigène et l'activation de l'expression d'IFN-γ ainsi que l'activation de STAT4 et la différenciation des lymphocytes TH1 chez l'homme. Ces interleukines induisent l'IL-15 qui promeut le développement et la différenciation des cellules NK et des lymphocytes T mémoires. Les IFN-α/β activent la maturation des cellules dendritiques. A l'opposé, l'activité immunosuppressive des interférons α/β se caractérise par une activation de Fas et de l'IL-10 et par une inhibition de l'IL-12 STAT1 dépendante. On constate aussi une activité stimulatrice de l'IFN-α avec l'IL-10 sur les cellules T régulatrices CD4+CD25+ (Levings, Sangregorio et al. 2001).

L'interféron γ est l'unique membre de la classe des interférons de type 2. L'IFN-γ possède une activité immunostimulatrice en activant les macrophages, en stimulant la présentation des antigènes et l'expression des molécules du CMH de classe 2 . Il promeut la différentiation des TH1. Son activité immunosuppressive se traduit au niveau de la surrégulation de plusieurs prorteines SOCS « suppressor of cytokine signaling », bien que lui-même soit régulé par SOCS1 (Gadina, Hilton et al. 2001). L' IFN-γ possède des effets anti-prolifératifs sur les cellules myéloïdes et lymphoïdes (Matthys, Vermeire et al. 2001).

2.2.2.1. La sclérose en plaques

La sclérose en plaques est une maladie chronique dont la manifestation caractéristique est la démyélinisation du système nerveux central. Au début, les lésions sont caractérisées par l'accumulation locale de cellules T CD8 et T CD4 activées. Ce phénomène est suivi par une inflammation périvasculaire et par la dégénération de la myéline qui se produit en réponse à la réaction immunitaire active. Le TNF est la cytokine pivot dans cette maladie . Elle accroît l'infiltration des cellules inflammatoires, la gliosie et la démyélinisation. Une concentration élevée de TNF est constatée dans le fluide cérébrospinal et le sérum. Cette élévation semble corrélée avec la progression de la maladie vers sa forme chronique et l'exacerbation des rechutes de sclérose en plaques. Le TNF-α induit la dégradation de la myéline et la nécrose des oligodendrocytes (Zajicek, Wing et al. 1992). Il altère la barrière sang-cerveau en

surrégulant les molécules d'adhésion aux cellules endothéliales cérébrales. On observe une expression élevée de cytokines sécrétées par les lymphocytes TH1 pendant la sclérose en plaques telles que LT, IL-1, IL-2, IFNγ . L'élévation des taux de TNF-α, IL-1, et LT précède chaque crise . Les voies de traitements thérapeutiques consistent à inhiber l'action des cytokines pro-inflammatoires par les cytokines anti-inflammatoires. Des traitements thérapeutiques basés sur l'IFN-β montrent des résultats encourageants.

2.2.2.2. Le Psoriasis

Entre 2 et 3% de la population mondiale est affectée par le psoriasis. Il s'agit d'une affection cutanée chronique. Cette maladie a un caractère héréditaire même si elle ne suit pas les profils d'héritages classiques des maladies mendéliennes autosomales.

Les manifestations cliniques de cette maladie sont principalement une hyper prolifération des kératinocytes de la couche basale de l'épiderme. L'hyperkératose s'accompagne de la disparition de la couche granuleuse. On observe l'infiltration des lymphocytes, de cellules dendritiques, et de macrophages dans le derme. Dans les zones, les vaisseaux sanguins sont dilatés et tortueux. La meilleure hypothèse quant au mécanisme déclenchant le psoriasis concerne le dérèglement du réseau des cytokines (Nickoloff 1991). Un évènement, tel qu'un traumatisme exogène ou endogène, déclenche l'expression en cascade de cytokines entraînant l'ensemble des phénomènes cellulaires. Deux des premières cytokines sont le TNF-α exprimé par les cellules dendritiques et d'autres CPAs et l'IFN-γ produit par les lymphocytes TH1 activés. Lors de la lésion, les lymphocytes T mémoires sont activés par les cellules dendritiques présentant l'antigène. Lorsque les lymphocytes sont activés après la stabilisation de l'interaction CPA:T par la synapse immunitaire, un réseau complexe de médiateurs de l'immunité se crée, constitué par les cytokines, les chimiokines et les facteurs de croissance. Ce réseau des cytokines facilite l'activation cyclique des lymphocytes T et des cellules dendritiques. Les cytokines exprimées pendant les phases chroniques de psoriasis et qui participent à la création et au maintient des plaques cutanées d'inflammation sont les sTNF-α, IFN-γ, IL-1, IL-6, IL-8, IL-15, IL-17, IL-18, IL-20, IL-23. On trouve aussi de nombreuses chimiokines de type CC et CXC, ainsi que les facteurs de croissances TGF-β, IGF-1, KGF,VEGF, NGF et l'amphiréguline (Krueger 2002).

Une première stratégie thérapeutique concerne l'utilisation de la cyclosporine A dans le but de bloquer la synthèse de cytokine par les immunocytes. Les immunosuppresseurs, comme le sirolimus (rapamycin) (Reitamo, Spuls et al. 2001) et le pimecrolimus (ASM981) (Rappersberger, Komar et al. 2002), sont également utilisés. Une autre stratégie consiste à utiliser des anticorps synthétiques anti TNF-α tel que l'infliximab (Antoni and Manger 2002) Dans le même créneau thérapeutique, il s'agit d'utiliser un récepteur soluble de synthèse capable de fixer le TNF-α et de bloquer son action (Mease, Goffe et al. 2000). D'autres recherches concernent le ciblage de l'IFN-γ et de l'unité p40 commune à IL-12 et IL-23. Des tests sont effectués avec des cytokines immunosuppressives synthétiques de type TH2 telles que l'IL-10 (Asadullah, Sterry et al. 1998). l'IL-11 (Trepicchio, Ozawa et al. 1999; Oestreicher, Walters et al. 2001), et l'IL-4 (Ghoreschi, Thomas et al. 2003).

2.2.3. Dans les réponses anti-virales

La réponse immunitaire face à l'infection de l'organisme par un virus fait intervenir les IFN-α/β, l'IFN-γ et l'IL-16 (Baier, Werner et al. 1995). Les IFN-α/β sont capables d'induire un état de résistance à la multiplication virale. Leur synthèse peut aussi être induite par des microorganismes pathogènes tels que les bactéries *Listeria monocytogenes, Haemophilus influenzae, Brucella* … Dans les premières heures après l'infection, les interférons sont sécrétés dans le milieu extra cellulaire par les cellules infectées. Ils vont se fixer aux récepteurs spécifiques des cellules voisines et induisent la synthèse de protéines qui rendent les cellules réfractaires à l'infection.

L'IFN-α est produit par les monocytes/macrophages, les leukocytes et les cellules NK. L'IFN-β est produit exclusivement par les fibroblastes à la suite de l'infection virale. L'IFN-γ est exprimé par les lymphocytes T activés après contact avec l'antigène viral et par les cellules NK.

2.2.4. Dans les processus inflammatoires

Il existe deux classes de cytokines de l'inflammation : les cytokines pro-inflammatoires et les cytokines anti-inflammatoires. On rassemble dans le premier groupe le TNF-α, IL-1, IL-2, IL-6, IL-8, les interférons. Le TNF-α, L'IL-1 et l'IL-6 sont impliqués dans le symptôme de fièvre. Le deuxième groupe rassemble l'IL-4, IL-10, TGFβ, IL1-RA.

La réaction inflammatoire est un ensemble de mécanismes physiologiques de défense mis en place par l'immunité naturelle qui a pour objectif de circonscrire et de réparer les lésions tissulaires. L'inflammation est la conséquence de l'action des cytokines synthétisées par les cellules de la réponse immunitaire naturelle et plus particulièrement par les macrophages activés.

La phase aiguë d'un syndrome inflammatoire comporte quatre phases : l'induction, l'activation-amplification de la réponse inflammatoire, la stabilisation et la résolution.

Pendant la phase aiguë, les monocytes et macrophages activés vont libérer des cytokines : IL-1, IL-2, IL-6, IL-8, TNF … qui vont induire la production hépatique des protéines de l'inflammation. Les causes de l'inflammation sont très variées : les bactéries, les virus, les réactions d'hypersensibilité aux parasites, les agents physiques comme les traumatismes, les irradiations, la chaleur ou le froid, les produits chimiques toxiques ou corrosifs, l'ischémie. La phase aiguë fait intervenir des phénomènes vasculaires : augmentation du flux sanguin vers la région lésée, l'augmentation de la perméabilité membranaire, la formation d'œdème et d'exsudat. On observe les phénomènes cellulaires suivants : adhésion à la paroi vasculaire des neutrophiles , la transmigration de l'endothélium, la migration cellulaires sous l'effet des chimiokines, la phagocytose, la destruction de l'agent responsable de l'infection, la cicatrisation des lésions tissulaire et l'apoptose des neutrophiles.

Les symptômes de l'inflammation sont nombreux. La fièvre est due à l'action des cytokines pro-inflammatoires TNF-α, IL-1 et IL-6. Le trouble métabolique et l'amaigrissement sont le fait du TNF-α et de l'IL-6. La leucocytose est due à l'effet endocrine du GM-CSF ainsi que du G-CSF. Sur le site inflammatoire, le G-CSF, le GM-CSF le TNF-α et l'IL-8, en conjugaison avec d'autres protéines de la phase aiguë (leucotriène, C5a, facteur d'activation de plaquettes, histamine) vont induire l'activation et la dégranulation des polynucléaires neutrophiles.

Les maladies autoimmunes ont souvent comme origine le dérèglement du phénomène de l'inflammation corrélé à un dysfonctionnement du réseau de cytokines.

2.2.4.1. Le choc septique

La lutte de l'organisme contre l'infection microbienne inclut l'expression des cytokines : TNF-α, IL-1, IL-6, IL-8, IL-10, IFNγ, deux types de récepteurs solubles du TNF et IL1-RA (Karima, Matsumoto et al. 1999).

Le syndrome de la sepsis correspond à une inflammation systémique induite par une infection bactérienne, fongique ou parasitaire (Karima, Matsumoto et al. 1999; Van Amersfoort, Van Berkel et al. 2003). Le choc septique est une des complications entraînées par une sepsis.

La réponse immunitaire contre les bactéries gram-négatives implique en particulier la reconnaissance des LPS « lipopolysaccharides » qui sont des constituants membranaires de ce genre bactérien. Cette étape s'effectue au travers de l'interaction entre les molécules de LPS et le récepteur membranaire spécifique CD14. Elle entraîne l'activation des leukocytes et la production de TNF-α, IL-1, et IL-6. La sécrétion d'exotoxines produites par les bactéries gram-positives activent les Lymphocytes T et s'accompagne de l'expression de TNF-α, IL-1, et IL-6 en faibles concentrations et de l'expression de l'IL-8 à des concentrations croissantes (Bone 1994; Opal and Cohen 1999; Sriskandan and Cohen 1999).

Chez la souris, l'activation des macrophages du foie par un LPS provoque la libération de TNF-α (Grewe, Gausling et al. 1994) dont le gène est transcrit de façon constitutive, puis de l'activation de la transcription de l'IL-1 et de l'IL-6. Les concentrations maximales en ARNm pour ces trois interleukines sont atteintes au bout de 40 minutes (Grewe, Gausling et al. 1994; Luster, Germolec et al. 1994).

D'autres médiateurs de l'inflammation sont sécrétés IL-8, IL-12, PAF (Platelet Activating Factor), des chimiokines ainsi que des eicosinoïdes (Qureshi, Honovich et al. 1988; Hack, Aarden et al. 1997; Lukacs, Hogaboam et al. 1999; Katori and Majima 2000). Plusieurs de ces médiateurs et les molécules du complément C3a et C5a vont attirer les PMNs (polymorphonucléaires) de la circulation et les activer (Kawamura, Imanishi et al. 1995; van

Oosten, van de Bilt et al. 1995; Jaeschke and Smith 1997). Les PMNs libèrent des microbicides impliqués dans la lyse des micro-organismes mais qui ont aussi pour effet d'endommager les cellules endothéliales (Chatham and Blackburn 1993). Ces altérations sont observables au cours d'une sepsis (Bone 1991). Les PMNs activés sécrètent CD14 soluble. Les cellules endothéliales, activées par les LPS via le CD14 soluble libèrent IL-1, IL-6, des eicosanoïdes, le « vasoactive agents endothelium derived relaxation factor, l'endothelin-1, des chimiokines et des CSFs (Mahalingam and Karupiah 1999). Les lymphocytes T et B sont attirés par les médiateurs de l'inflammation. Des médiateurs sont sécrétés tardivement : IL-2, IFNγ, et GM-CSF. L'IL-2 et le GM-CSF participe à la prolifération des PMNs et des cellules mononucléaires. Le TNF-γ accroît l'effet des LPS sur les cellules mononucléaires (Yong and Linch 1993).

Les LTA et PGN sont des composants de la paroi des bactéries gram-positives qui causent le relargage de NO, TNF-α et de IFNγ (De Kimpe, Hunter et al. 1995; Kengatharan, De Kimpe et al. 1998; Bowie and O'Neill 2000). Les exotoxines libérées par ces bactéries ont un rôle de super antigène. Elles sont aussi capables d'induire des chocs toxiques (Bannan, Visvanathan et al. 1999).

Le TNF-α et l'IFNγ sont les interleukines clés de la sepsis. Leur sécrétion incontrôlée par les lymphocytes T sont la cause du syndrome de choc toxique (Cameron, Nawijn et al. 2001).

2.2.4.2. L'arthrite rhumatoïde

Le synovium rhumatoïde contient la majorité des cytokines déjà connues. Cette abondance est la conséquence de la présence d'un phénomène d'inflammation sur la zone malade qui contient des lymphocytes activés, des macrophages, des fibroblastes, des cellules endothéliales et des cellules plasmatiques productrices de cytokines. Au niveau de la zone d'inflammation, l'expression des cytokines proinflammatoires est anormalement élevée (Feldmann and Maini 2001). Il a été observé la présence et la production en excès d'IL-1, d'IL-6 et de TNF-α. Cette dernière cytokine est une cible thérapeutique privilégiée compte tenu de son implication dans l'induction et la maintenance de la maladie. En effet, le TNF-α est à l'origine de l'expression des autres cytokines proinflammatrices IL-1, IL-6, GM-CSF et

IL-8. L'IL-15 est aussi présente dans le synovium rhumatoïde (McInnes, al-Mughales et al. 1996). L'administration d'une forme soluble recombinante d'IL-15Rα à des modèles murins a permis de bloquer de développement de l'arthrite rhumatoïde induite par le collagène (Ruchatz, Leung et al. 1998).

En complément aux médiateurs proinflammatoires, on observe une surexpression des cytokines immunosuppressives IL-10, IL-11, IL-1ra ainsi que des récepteurs TNF-s et IL-1 solubles (de Waal Malefyt, Abrams et al. 1991; Deleuran, Chu et al. 1992; Firestein, Berger et al. 1992; Joyce, Gibbons et al. 1994; Katsikis, Chu et al. 1994). L'interleukine anti-inflammatoire IL-4 est très peu présente et L'IL-13 est exprimée de manière très variable. Il existe donc un déséquilibre en faveur des cytokines de l'inflammation.

Les traitements thérapeutiques ciblent principalement l'expression de TNF-α par la synthèse d'anticorps antiTNF-α (infliximab). L'utilisation d'anti-TNF-α conjugué à du MTX (méthrotexate) est une thérapie très efficace (Maini, Breedveld et al. 1998; Maini, St Clair et al. 1999). Des tests cliniques concernent l'administration des cytokines immuno-régulatrices IL-4, IL-10, IL-11 et IFN-β (Moreland, Gugliotti et al. 2001). Une autre voie de traitement thérapeutique concerne l'inhibition du facteur de transcription NF-κB (Foxwell, Browne et al. 1998; Bondeson, Foxwell et al. 1999).

2.2.4.3. La maladie de Crohn

La maladie de Crohn (Andreakos, Foxwell et al. 2002) est l'une des deux formes de la maladie inflammatoire chronique de l'intestin. La deuxième forme pathologique est nommée recto-colite hémorragique. La maladie de Crohn se caractérise par des troubles digestifs, diarrhée, saignement rectal, douleur abdominale, perte de poids, désordres dermiques et oculaires. Elle est aussi la cause de retard de croissance et de maturation sexuelle. Dans les formes de maladie de Crohn modérées, on observe des taux d'IL-15 sécrétés par les PBMC plus faibles que chez les patients atteints de formes aiguës de la maladie ou de recto-colite hémorragique (Kirman and Nielsen 1996). La surexpression de l'IL-15 par les macrophages de la *lamina propria* active les lymphocytes T situés dans le voisinage (Sakai, Kusugami et al. 1998). Ceux-ci induiraient l'expression des TNF-α et d'IL-12 par les monocytes . L'Il-15

indurait l'activation des cellules NK lors de l'inflammation pathologique (Ohta, Hiroi et al. 2002). D'autres cytokines sont impliquées dans la maladie de Crohn : IL-1, IL-8, IL-6, MCP-1. Le TNF-α est présent en excès dans la muqueuse des patients et l'augmentation de l'expression du TNF-α est la cause de la libération des cytokines pro-inflammatoires en particulier IL-1, IL-6 et IL-8. Les traitements se basent sur des anticorps monoclonaux anti-TNF-α ou sur l'ajout des interleukines anti-inflammatoires IL-10 et IL-11.

2.2.5. La cicatrisation

La résolution du phénomène d'inflammation se termine par une phase de réparation qui consiste à éliminer par phagocytose les cellules lésées, à induire un remodelage tissulaire, accompagnée par une étape de néovascularisation et enfin de la cicatrisation. La phase de réparation s'effectue sous le contrôle des cytokines, en particulier le VEGF (Vascular Endothélial Growth Factor), le PDGF (Platelet Derived Growth Factor), le FGF (Fibroblast Growth Factor) et le TGFβ(Transforming Growth Factor β). Le PDGF est responsable du chimiotactisme pour les fibroblastes. Cette cytokine assure les phases de cicatrisation et de réparation tissulaire en activant la production de collagène et de collagénase. Le TGFβ est un accélérateur du processus de cicatrisation. Il est impliqué dans le chimiotactisme des macrophages et des fibroblastes. Il inhibe les MMPs (métalloprotéinases) et active l'expression des inhibiteurs des MMPS appelés TIMPs.

2.2.6. Les cytokines de l'hématopoïèse

L'hématopoïèse constitue le processus complexe de la production de nouvelles cellules sanguines. L'hématopoïèse se déroule dans la moelle osseuse. On retrouve dans la moelle osseuse les cellules souches hématopoïétiques qui sont les précurseurs des cellules hématopoïétiques, lymphocytaires et phagocytaires. En outre, chez l'homme, la moelle osseuse est l'organe lymphoïde primaire pour la lignée B alors que les lymphocytes T achèvent leur maturation dans le thymus. Elle héberge en périphérie une partie des

lymphocytes B activés par l'antigène qui se transforment en plasmocytes sécréteurs d'anticorps.

Les CSFs « Colony Stimulating Factors » contrôlent la différenciation et la maturation des cellules souches hématopoïétiques au sein de la moelle et à la périphérie. On peut inclure dans ce groupe de cytokines l'EPO, l'IL-3, le SCF, le GM-CSF, le M-CSF, le G-CSF, l'IL-7, et l'IL-5.

L'EPO est connue pour induire la production des hématies à partir des cellules souches de la moelle osseuse. L'IL-3 est produit essentiellement par les lymphocytes T CD4 activés. Elle active la prolifération et la différenciation des basophiles et des mastocytes. L'IL-3 agit en synergie avec l'IL-6 sur les cellules souches des lignées hématopoïétiques et les lignés suivantes. Le GM-CSF active les cellules des lignées granulocytaires et monocytaires. Il est exprimé par les lymphocytes T activés, les macrophages activés, les cellules endothéliales, les fibroblastes situé au niveau du stroma et de la moelle. Il n'est pas détecté dans la circulation. Le M-CSF est le facteur de croissances des monocytes macrophages. Le M-CSF n'est pas circulant. Le G-CSF est un facteur de croissance circulant de la lignée granulocytaire. L'IL-7 est un facteur de croissance des progéniteurs B et T, des thymocytes et des lymphocytes CD4 et CD8. L'IL-5 est produit par les lymphocytes T et est impliqué dans l'activation des colonies éosinophiles. Elle active les mécanismes de survie et les éosinophiles matures. C'est un agent chemoattracteur qui provoque la dégranulation des éosinophiles.

2.2.7. Les cytokines, l'apoptose et le cancer

L'apoptose est un mécanisme physiologique par lequel la cellule induit sa mort. L'apoptose est présente pendant l'embryogénèse et au cours des étapes de sélections négatives des lymphocytes T dans le thymus qui permet d'éliminer les lymphocytes T autoréactifs. Elle ne déclenche pas de réponse inflammatoire. Ce mécanisme se traduit par une dégradation de l'ADN chromosomique en fragment de 200 bps, par la condensation du cytoplasme avec rétraction cellulaire et aspect de bullage. La cellule est fragmentée en corps apoptotiques qui sont ensuite phagocytés et éliminés. L'apoptose peut être issue d'un signal membranaire médié par la protéine Fas inductrice d'une cascade de protéase à cystéines appelées caspases pour leur capacité à couper la liaison amine après un résidu aspartate en P1. La cascade de

caspases active des endonucléases qui interviennent lors de la fragmentation de l'ADN, joue un rôle dans la perte de l'assymétrie des phospholipides membranaires avec apparition de phosphatidyl-sérine dans le feuillet externe de la membrane plasmique. Les granzymes sont des cytotoxines préformées par les lymphocytes T cytotoxiques qui transmettent le signal de l'apoptose.

Il existe deux autres voies de signalisation de l'apoptose : la famille Bcl-2 et la famille des récepteurs de TNF.

L'apoptose peut être engagée par différents stimuli tels que les irradiations gamma, les lymphocytes T cytotoxiques, les cytokines cytolytiques telles que le TNF-α. De nombreux facteurs de croissances et des cytokines agissent plutôt comme des facteurs de survie cellulaire dans le but de réguler négativement l'apoptose.

Plusieurs cytokines possède une activité proapoptotique : Le TNF-α, l'IL19, l'IL-24, L'IFN-β2 et l'IFN-γ. D'autres cytokines sont antiapoptotiques : les FGF, IL-3, GM-CSF, IL-15, IL-13, G-CSF, OSM, IL-11, IL23p19, CNTF, IL9. De nombreuse cytokines peuvent jouer un rôle pro ou antiapoptotique : IL-1a/b, IL-2, IL-21, IL-4, IL-5, IL-6, LIF, IL-12, IL-7, IL-10, Ob, et PRL.

Du fait de leurs effets dans l'apoptose, les cytokines sont impliquées dans la croissance et le développement de certains cancers (O'Shea, Ma et al. 2002). Le TNF est très impliqué dans le carcinome ovarien humain (Naylor, Stamp et al. 1993). Il est produit par les îlots tumoraux épithéliaux, et son niveau d'expression croît avec l'aggravation de la maladie. Le rôle des cytokines proinflammatoires dans les mécanismes cancéreux a été mis en évidence à partir de souris KO TNF-α -/- (Moore, Owens et al. 1999). Ces souris sont résistantes au développement de tumeurs épidermiques bénignes et malignes. Dans les stages tumoraux précoces, la croissance des kératinocytes et les effets inflammatoires sont diminués chez les souris KO, alors que le TNF-α semble avoir moin d'effets sur les stages tardifs de la carcinogénèse. Ces données montrent que le TNF-α joue un rôle important dans la carcinogénèse *de novo* et qu'il est important lors des stages précoces de la carcinogénèse. D'autres travaux ont montré que le TNF-α joue un rôle promoteur des tumeurs endogènes. L'IL-6 est une cible thérapeutique dans le cas du myélome malin multiple des cellules B humaines (Klein 1998). L'IL-6 a une fonction paracrine et autocrine sur les cellules myélomateuses. L'IL-6 joue aussi un rôle antiapoptotique sur les cellules myélomateuses et

les protège de nombreuses substances cytotoxiques. Le traitement avec un anticorps anti-IL6 a permis d'améliorer l'état de forme des patients qui présentaient les plus fort taux d'expression d'IL-6. Les recherches thérapeutiques en cours concernent la synthèse l'utilisation d'anticorps anti-IL-6, de récepteur IL-6 soluble ou d'IL6 mutée.

2.2.8. Les chimiokines

Les chimiokines sont des molécules de faible poids moléculaire. Elles sont des médiateurs de l'inflammation et sont impliquées dans le recrutement des cellules immunocompétentes. On distingue la sous famille IL-8 appelée famille des CXC-chimiokines (α-chimiokines). Ces chimiokines sont produites principalement par les monocytes et les macrophages. Les chimiokines de cette famille sont la PBP (Platelet Basic Protein), la β-TG (β-thromboglobuline), le F4P (facteur 4 plaquettaire, l'IP10 (gamma Interferon Inducing Protein), et le SDF1 (Stroma-Derived Factor 1).

La deuxième sous-famille se nomme CC-chimiokines (β-chimiokines). Les membres de cette sous famille sont les MIP1α/β (Macrophage Inflammatory peptides 1α et β), la molécule RANTES (Regulated Upon Activation Normal T Expressed and Secreted) et l'éotaxine. Les membres de cette famille participent au chimiotactisme des cellules de l'immunité telles que les monocytes, les éosinophiles, les basophiles et les lymphocytes mais pas les polynucléaires neutrophiles.

La troisième sous-famille est appelée C-chimiokines ou γ-chimiokines (lymphotactines). Elles sont spécifiques aux lymphocytes T.

La quatrième sous-famille, CX3C ou fractaline est composée de glycoprotéines de membrane. La forme membranaire est exprimée par les cellules endothéliales et joue un rôle dans l'adhésion des monocytes et de lymphocytes T. La forme soluble exerce un rôle chimiotactique sur ces deux types de cellules immunitaires.

2.3. Voies de signalisation des cytokines

La signalisation intra-cytoplasmique fait intervenir des kinases de la famille Jak. Ces éléments peuvent ensuite être associés à trois voies de signalisation liées respectivements aux protéines STAT, aux MAP kinases et à la PI3K.

2.3.1. Le rôle de transduction des protéines JAKs

Les protéines de la famille des JAKs « Janus Associated Kinases » sont au nombre de quatre : JAK1, JAK2, JAK3, et Tyk2. Elles possèdent chacune un domaine de fixation aux régions intracellulaires des chaînes réceptrices des cytokines de type I et II. Ces domaines, situés en N-terminal, possèdent une conformation de type FERM « Four-point-one, Erzin, Radixin, Moesin ». Il est indispensable à la fixation de la kinase JAK au niveau des motifs Box1 et Box2 des chaînes réceptrices. La région Cter des JAKs comporte 3 domaines. Les domaines JH1 et JH2 sont impliqués dans l'activité tyrosine kinase et le domaine JH3 présente une conformation de type SH2 « Src-Homology 2 » capable de lier des tyrosines phosphorylés, mais dont le rôle biologique n'est pas clairement défini.

Les JAKs (Ihle, Witthuhn et al. 1995) sont fixés constitutivement sur les régions intracellulaires des chaînes réceptrices. La formation du complexe cytokine-récepteur entraîne la dimérisation des chaînes réceptrices et le rapprochement des JAKs qui se phosphorylent. sur tyrosines. Les JAKs activées phosphorylent également des tyrosines situées sur les régions intracytoplasmiques des chaînes réceptrices, créant ainsi des sites d'ancrage pour des médiateurs à domaines SH2.

L'activation des JAKs est la première étape de la transduction du signal intracellulaire. La poursute de la signalisation peut être effectuée de trois manières différentes : la voie des STAT « Signal Transducers and Activator of Transcription », la voie des MAPKs « Mitogen-Activated Protein Kinases » et la voie « PI3-K, PDK1, AKT ».

2.3.2. La voie des STATs

La famille STAT (Ihle, Nosaka et al. 1997; Ihle, Thierfelder et al. 1998; Ihle 2001) rassemble sept facteurs de transcription contenant chacun un domaine SH2 de liaison aux phosphotyrosines SH2. Ce domaine permet l'interaction des STATs avec les domaines intracellulaires des régions intracytoplasmiques des récepteurs. Après phosphorylation par les JAKs, les STATS forment des homodimères ou des hétérodimères (STAT1-STAT2 et STAT1-STAT3), étape nécessaire à leur translocation nucléaire. La région N-ter des STATs contient un domaine de type « Leucine Zipper » et un domaine de fixation à l'ADN spécifique des boîtes promotrices de type SBE « STATS Binding Element » telles que les GAS « Gamma-Activated Sequence » et les ISREs « Interferon-Stimulated Response Elements ». Un domaine d'activation transcriptionnelle est situé en C-ter et varie pour chaque STAT. Les différentes combinaisons entre les types des chaînes réceptrices et la nature des JAKs et des STATs influent sur la sous population des gènes cibles qui est activée.

2.3.3. La transduction du signal par les MAP kinases

Le recrutement des sérine/thréonine kinases MAPKs (Kishimoto, Taga et al. 1994) fait intervenir la protéine Grb2 « Growth factor receptor-bound protein 2 » qui se fixe aux tyrosines phosphorylée des régions intracytoplasmiques des chaînes réceptrices. Des protéines adaptatrices, telles que les SHc « SH2-containing collagen-related protein » et SHP2 « SH2-containing Phosphatase 2 », peuvent se fixer sur les sous-unités JAKs phosphorylées. La protéine Grb-2 est capable de s'associer au facteur d'échange Sos « Son of sevenless » qui à son tour va recruter et activer par phosphorylation la sérine/thréonine kinase Ras. La forme phosphorylée de Ras se dimérise avec la kinase Raf qui elle même est phosphorylée à son tour. Raf est une MAPKKK « MAPK Kinase Kinase ». Elle phosphoryle la kinase intracytoplasmique MEK entre autre, qui est une « MAPK Kinase ». MEK a pour rôle de phosphoryler les MAPKs telles que ERK, p38 et junk. Les MAPKs sont capables de migrer dans le noyau et d'activer les facteurs de transcription.

2.3.4. La voie kinase PI3-K, PDK1, AKT

La troisième voie fait intervenir la kinase PI3K. Cette protéine membranaire est un hétérodimère composé d'une petite sous unité p85 possédant deux domaines SH2 et un domaine SH3 ainsi que d'une sous-unité catalytique p110 capable de phosphoryler le PIP2 « Phosphatidyl inositol 4,5-biphosphate » en PIP3 triphosphate (Wymann and Pirola 1998). Le PIP3 interagit au niveau de la membrane plasmique avec la PDK1 « 3-Phosphatidyl inositol-Dependent protein Kinase 1 » et aussi la sérine/thréonine kinase Akt. Le PIP3 induit un changement de la conformation d'Akt qui peut alors être activé par la PDK1. L'Akt activée devient alors cytosolique. Elle est responsable de l'activation du facteur de transcription NFκB (Romashkova and Makarov 1999) et de l'activation de la transcription de ses gènes cibles, de la surexpression de la kinase P70S6K « small subunit ribosomal protein S6 » (Reif, Burgering et al. 1997), qui est impliquée dans les mécanismes de traduction protéique, et de l'inhibition du médiateur anti-apoptotique Bcl-XL (Datta, Dudek et al. 1997).

2.3.5. Les voies inhibitrices de la transduction du signal cytokinique

Il existe trois types d'inhibition de la transduction du signal cytokinique (Wormald and Hilton 2004).

La protéine SHP1 « SH2-connecting Phosphatase 1» est exprimée de manière constitutive par les cellules. Cette phosphatase a la propriété de déphosphoryler les tyrosines situés sur les sous-unités JAKs activées ainsi que sur les chaînes réceptrices.

Les protéines PIAS « Protein Inhibitor of activated STAT » sont aussi exprimées de manière constitutive par les cellules. Certaines PIAS sont capables d'inhiber directement l'assosciation de STAT avec l'ADN. Les membres de cette famille d'inhibiteur possèdent un domaine fonctionnel dont l'activité consiste à entraîner la fixation de la protéine SUMO 1 sur les STATs. Si SUMO 1 est impliqué dans le mécanisme d'ubiquitinilation, le mode d'inhibition de STAT n'est pas encore éclairci.

La troisième classe d'inhibiteur est constituée par les protéines SOCS « Suppressor Of Cytokine Signaling ». Cette famille est composée de sept membres : SOCS 1-7 et la protéine CIS « Cytokine-Inductible-SH2-domain-containing ». Chaque membre possède un domaine « SOCS box » positionné à l'extrémité C-ter et précédé par un domaine SH2. L'extrémité N-ter de ces protéines est variable. Le domaine SH2 permet aux protéines SOCs de se fixer au niveau des phosphotyrosine des sous unités JAKs et des chaînes de récepteurs. Les SOCS inhibent l'activité des JAKs ou entre en compétition avec les protéines STATs. Les SOCS sont aussi capables de recruter les médiateurs de la transduction du signal de dégradation par le protéasome et les médiateurs de l'ubiquitinilation.

Chapitre 3. Les classifications des cytokines

3.1. La classification en fonction du type du récepteur

3.1.1. Récepteurs de cytokines de type I

Les chaînes réceptrices des cytokines hématopoïétiques sont des glycoprotéines de type 1, ce qui indique que l'extrémité NH2 de la chaîne polypeptidique est orientée vers l'extérieur de la membrane (Figure 1). Elles contiennent toutes une seule région transmembranaire, à l'exception du récepteur CNTF-R qui est relié à la membrane par une liaison Glycosyl-phosphatidylinositol (GPI). Les régions extracellulaires des chaînes contiennent toutes au moins un module appelé CRH (cytokine receptor homolog) ou CBM (Cytokine Binding Module). Le CBM est constitué par deux domaines de type hématopoïétique, comprenant chacun sept feuillets β antiparallèles et séparés par une séquence intermédiaire riche en résidu proline formant un coude. Certaines chaînes réceptrices telles que βc et TPO-R sont pourvues de deux CBMs. Le mode de fixation des cytokines sur les récepteurs hématopoïétiques de type 1 a été étudié sur le récepteur à la GH qui est fonctionnel sous sa forme homodimérique (de Vos, Ultsch et al. 1992). Il existe trois sites d'interactions. Le site I du récepteur est de haute affinité et permet la liaison de la cytokine avec le premier monomère. Le site II permet la liaison du deuxième monomère du récepteur à la GH. Il existe un troisième site qui correspond à l'interaction entre les deux chaînes du récepteur. On retrouve un mode de fixation équivalent entre la cytokine et un récepteur hétérodimérique ou trimérique. Dans ce cas, le site I se lie à la chaîne spécifique du récepteur et le site II permet la liaison à la chaîne commune à plusieurs cytokines.

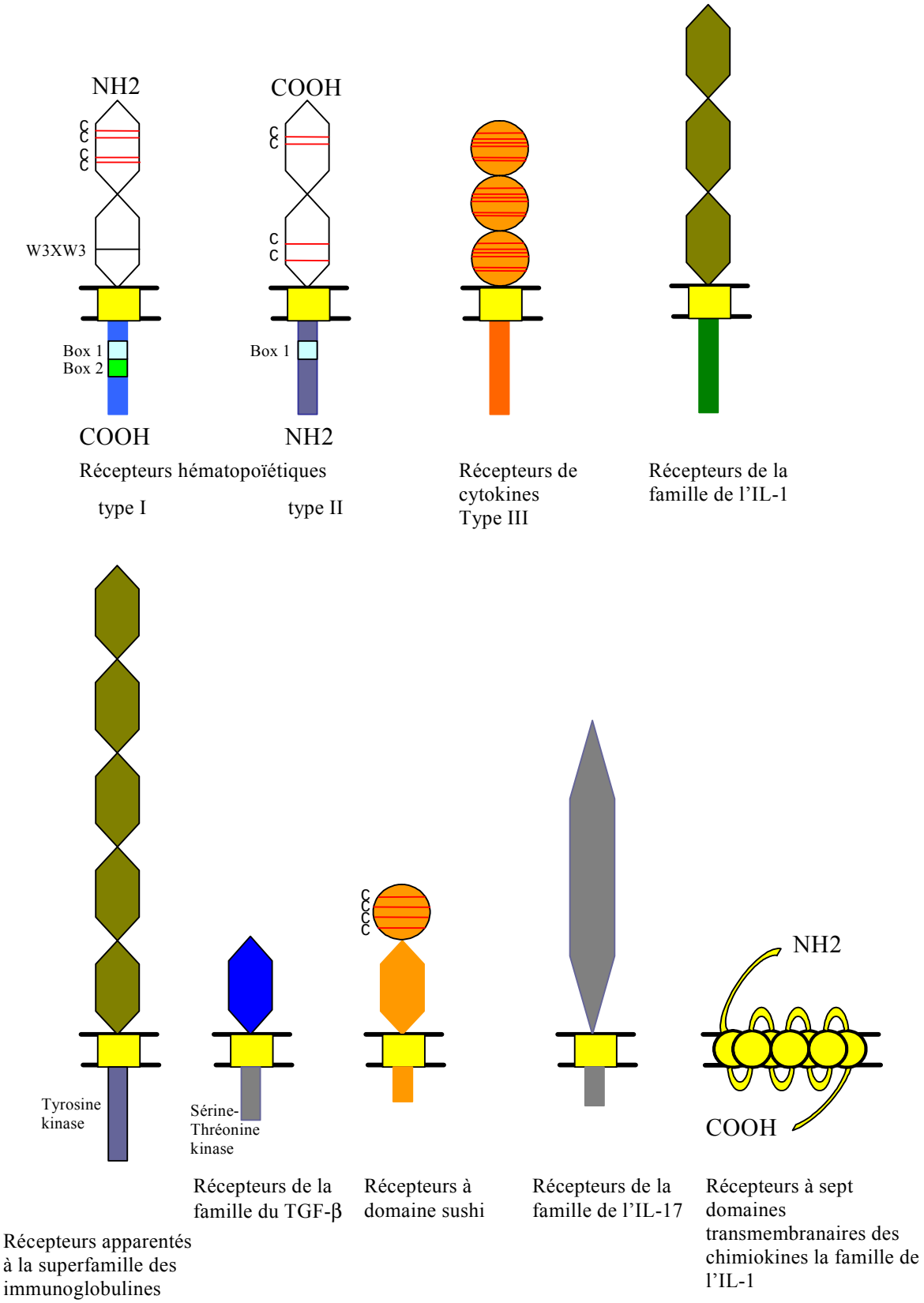


Figure 1. Les familles des chaînes membranaires des cytokines

Le domaine hématopoïétique I est situé à la partie NH2 terminale du CBM. Il est caractérisée par la présence de quatre cystéines qui sont conservées entres les différentes chaînes réceptrices connues. Les cystéines forment deux ponts disulfures intrachaînes successifs qui sont indispensables à la fonction du CRH.

Le domaine hématopoïétique II, en C-terminal du CBM, possède un motif conservé X-S-X-W-S spécifique des chaînes réceptrices hématopoïétiques de type 1. Ce motif est nécessaire au maintien de la fonction du CBM mais n'est pas impliqué de manière directe dans la fixation du ligand.

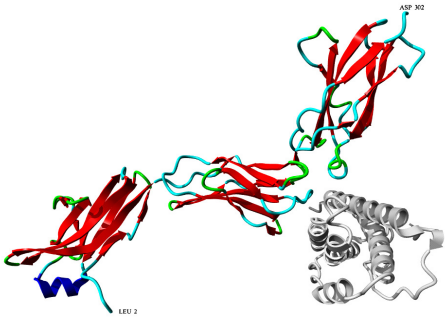
La région extracellulaire des chaînes réceptrices de la famille des cytokines hématopoïétiques de type 1 peut aussi contenir des domaines de type Ig (immunoglobuline) constitués d'une centaine de résidus. Ces domaines peuvent être impliqués dans la stabilisation du complexe cytokine/récepteur. On peut aussi retrouver des domaines FNIII (fibronectine III) longs de 100 résidus en moyenne. Ces domaines sont impliqués dans la stabilisation du complexe entre les chaînes réceptrices.

La région intracellulaire des chaînes réceptrices hématopoïétiques de type 1 a une longueur variable. Les chaînes impliquées dans la transduction du signal possèdent deux motifs box1 et box 2. Le motif Box1 est du type Al-AR-P-X-P-X-P-X-P. Il est nécessaire à la fixation des protéïne tyrosines kinase de la famille des JAKs (Janus-Associated Kinases). Le motif Box2 est moins conservé et est constitué d'une série d'acides aminés hydrophobes suivis d'acides aminés chargés négativement et se termine par deux acides aminés chargés positivement. Le rôle du domaine Box2 serait d'augmenter l'affinité entre la région intracellulaire du récepteur et les protéines JAKs.

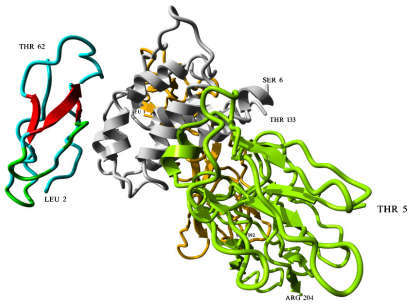
Lorsqu'il s'agit de récepteurs trimériques, les chaînes spécifiquement impliquées dans l'interaction avec la cytokine sont désignées par le caractère « α ». Les chaînes α possèdent en général un domaine intracellulaire réduit et non impliqué dans la fonction du récepteur. La cinétique de formation des complexes cytokines-récepteurs peut se faire de différentes manières. Les chaînes transductrices peuvent être réquisitionnées postérieurement à la fixation de la cytokine sur la chaîne spécifique α. Dans les cas où le complexe chaîne transductrice/protéines Jaks est préformé, il est activé par la liaison de la cytokine au récepteur qui entraîne la modification de la conformation de la chaîne transductrice.

Les interleukines IL-3, IL-5 et GM-CSF partagent la chaîne réceptrice commune βc (Hayashida, Kitamura et al. 1990; Kitamura, Sato et al. 1991; Tavernier, Devos et al. 1991). Le récepteur est constitué d'un hétérodimère constitué de la chaîne βc et de la chaîne α spécifique à l'une de ces cytokines : IL-5Rα, IL-3Rα, GM-CSFRα. Chez la souris, l'IL-3m possède une deuxième chaîne βIL-3 qui est un paralogue de la chaîne βc (Itoh, Yonehara et al. 1990; Miyajima, Mui et al. 1993). La chaîne réceptrice βIL-3 permet une médiation sélective du signal IL-3 en l'absence de βc (Nishinakamura, Miyajima et al. 1996). Contrairement à l'IL-3 et au GM-CSF qui agissent sous la forme monomérique, l'IL-5 agit sous forme homodimérique.

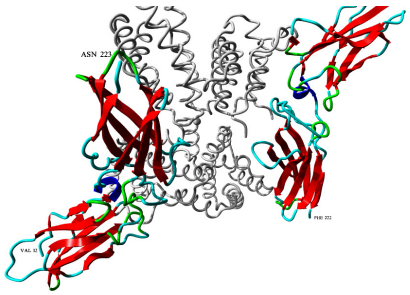
La chaîne gp130 est partagée par sept interleukines distinctes : IL-6, IL-11, LIF, OSM, CNTF, NNT-1, CT-1 (Taga and Kishimoto 1997; Senaldi, Varnum et al. 1999; Bravo and Heath 2000; Miyajima, Kinoshita et al. 2000) (Figure 2). Les complexes des récepteurs sont multimériques. Ces cytokines possèdent des chaînes α spécifiques IL6-Rα, IL-11-Rα, LIF-Rβ, OSM-Rβ, CNTF-Rα. Les interleukines LIF, OSM, CNTF, NNT-1 et CT-1 peuvent utiliser la chaîne LIF-Rβ. Le NNT-1, appelé aussi BSF-3 ou CLC est exprimé en association avec le CLF-1 (cytokine Like Factor 1). Le dimère NNT-1/CLF-1 est un deuxième ligand de la chaîne CNTFR-α (Elson, Lelievre et al. 2000). LIF, OSM, CNTF et NNT-1/CLF-1 partagent également la chaîne gp190.



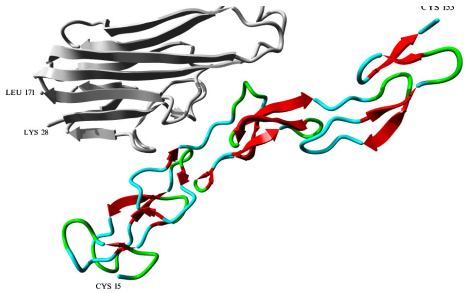
Récepteurs hématopoïétiques de type I :
gp130 [1I1R]



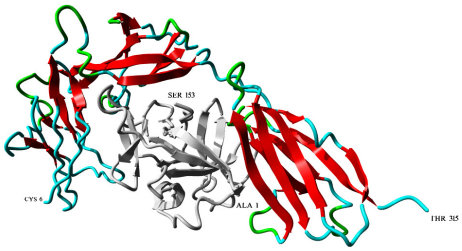
Récepteurs à domaines sushi :
IL-2Rα [1ILN]



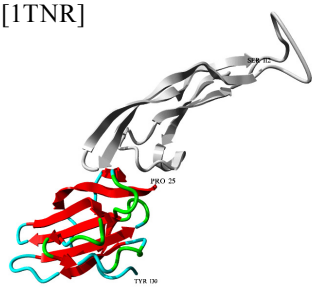
Récepteurs de la famille IL-10/INFs :
INF-γRα [1KTH]



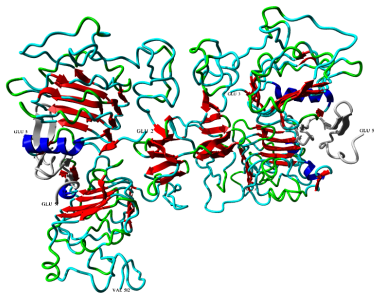
Récepteurs de la famille des TNFs :
TNFβ-R1 [1TNR]



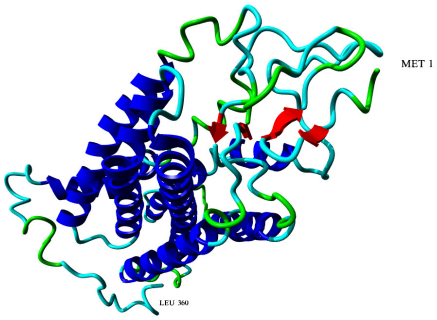
Récepteurs de la famille des IL-1 :
IL-1R [1ITB]



Récepteurs de la famille du TGF-β :
TGF-βR2 [1KTZ]



Récepteurs de la famille des immunoglobulines :
EGF-R [1IVO]



Récepteurs de la famille des
chimiokines : 2CCR-2B [1KP1]

Figure 2. Structures des domaines extracellulaires de différents types de chaîne réceptrices de cytokines

Les chaînes IL-6Rα et IL-11Rα n'ont pas d'activité transductrice mais sont nécessaires à la fonctionnalité du complexe. Pour ces deux interleukines, le complexe cytokine/récepteur est un hexamère (IL/chaîne spécifique/gp130)2. Deux homologues viraux de l'IL-6, KSV-IL-6 (Karposi Sarcoma associated human herpes Virus 8 Interleukin-6) et Rm-IL-6 (Rhesus Macaque rhadinovirus Interleukin-6), ont la propriété de se fixer à la gp130. Le complexe CNTF-R/récepteur serait un hexamère de type (CNTFR2gp130gp190)2. L'OSM utilise la chaîne gp130 en plus de l'OSM-R ou gp190. L'hétérodimère NNT-1/CLF-1 se lie au tétramère (CNTFR α2/gp130/gp190). La chaîne spécifique à CT-1 n'a pas encore été isolée. Celle ci serait liée à la membrane par une liaison GPI de même nature que celle du CNTF-R membranaire. Il existe des formes solubles des chaînes réceptrices spécifiques (sIL-6Rα et sIL-11Rα) capables de se lier, après complexation avec la cytokine, aux chaînes transductrices transmembranaires et de donner des complexes fonctionnels. A l'inverse, des formes solubles des chaînes transductrices sont des antagonistes dont la fonction est la régulation négative des cytokines.

Le récepteur fonctionnel de l'IL-31 (Dillon, Sprecher et al. 2004; Dreuw, Radtke et al. 2004) se compose de la chaîne spécifique IL31-RA et de la chaîne OSM-R. Quatre isoformes d'IL31-RA sont connus, IL31-RAv1 à IL31-RAv4. Les isoformes IL31-RAv3 et IL31-RAv4 sont capables d'activer la voie STAT lorsqu'ils sont couplés à l'OSMR mais pas avec les chaînes gp130, LIF-R, IL-12Rβ1 et IL-12Rβ2.

L'IL-12 et L'IL-23 partagent la même chaîne IL-12Rβ1. Ces deux cytokines et leurs complexes sont très similaires des complexes IL-6/récepteurs solubles/gp130. En effet L'IL12 se présente comme un hétérodimère IL12p35/IL12p40. La sous unité IL-12p35 présente une similarité structurale avec les interleukines de groupe « IL-6 » tandis que la sous unité IL-12p40 se rapproche des chaînes réceptrices solubles. Le récepteur de l'IL-12 est constitué d'un hétérodimère IL-12Rβ1/IL-12Rβ2 (Gately, Renzetti et al. 1998; Trinchieri 1998). De façon semblable à l'IL-12, L'IL-23 est constitué d'une sous unité IL23p19(Oppmann, Lesley et al. 2000) très similaire à l'IL-6 et à la sous unité IL-12p40. Le récepteur de l'IL-23 est le dimère IL-12Rβ1/IL-23R (Parham, Chirica et al. 2002). L'IL-27 est très proche de l'IL-12. Cette interleukine est un hétérodimère issu de l'association de L'IL27p28 (Pflanz, Timans et al. 2002), similaire à l'IL-12p35, et de l'EBI-3 « Epstein-Barr virus-Induced gene » qui

correspond à la chaîne réceptrice soluble. La chaîne spécifique du récepteur transmembranaire correspond à la chaîne orpheline TCCR/WSX-1 (Chen, Ghilardi et al. 2000; Pflanz, Timans et al. 2002). Le récepteur IL-27R correspond à l'association des chaînes WSX-1 et gp130 (Pflanz, Hibbert et al. 2004; Villarino, Huang et al. 2004). La sous-unité IL-12p35 peut se lier à l'EBI-3 mais aucune observation ne permet encore de savoir s'il s'agit d'un hétérodimère fonctionnel.

La chaîne transductrice γc (Takeshita, Asao et al. 1992; Leonard 2001) est commune aux IL-2, IL-4, IL-7, IL-9, IL-15, et IL-21 (Leonard 2001). Cette famille d'interleukines est souvent appelée « famille IL-2 ». Les récepteurs sont hétérodimériques ou hétérotrimériques. Les chaînes réceptrices spécifiques à chacune de ces six interleukines suivent la nomenclature décrite précédemment et sont nommées « α ». Les IL-2 et IL-15 partagent les chaînes IL-2Rβ et γc. Il existe trois types de récepteurs à l'IL-2 : le récepteur de haute affinité IL-2Rα/IL-2Rβ/γc, de moyenne affinité IL-2Rβ/γ et celui de basse affinité IL-2Rα qui n'est pas fonctionnel. L'IL-4 est très proche de l'IL13 même si cette dernière interleukine ne partage pas la chaîne γc. En effet le récepteur de l'IL-4 est l'hétérodimère IL-4Rα/γc (Mosley, Beckmann et al. 1989) mais il existe un autre récepteur fonctionnel IL-4Rα/IL-13Rα1 (Aman, Tayebi et al. 1996; Callard, Matthews et al. 1996; Hilton, Zhang et al. 1996) qui est aussi le récepteur fonctionnel de l'IL-13. L'IL-13Rα1 est très similaire à γc. Une autre chaîne IL-13Rα2 lie spécifiquement l'IL-13 avec une affinité plus élevée mais n'est cependant pas capable de transduire le signal (Caput, Laurent et al. 1996; Donaldson, Whitters et al. 1998). La chaîne IL-7Rα (Pandey, Ozaki et al. 2000; Park, Martin et al. 2000) peut être réquisitionnée par le TSLP et le TSLPR qui possède un taux de similarité très élevé avec la chaîne γc.

3.1.2. Récepteurs de cytokines de type II

Les récepteurs de cytokines de types 2 concernent les membres de la familles IL-10/interférons (Figure 1) (Figure 2) (Figure 3) : IL-10 (Moore, de Waal Malefyt et al. 2001), IL-19, IL-20, IL-22 (IL-TIF), IL-24 (MDA-7), IL-26 (AK155), IFNα, IFNβ, IFNγ, IFNω et IFNτ. Comme pour les récepteurs hématopoïétiques de type 1, les propriétés de redondance et

de pléiotropie sont liées au partage des chaînes réceptrices et transductrices. Toutes les chaînes de récepteurs semblent avoir été découvertes chez l'homme (Kotenko and Langer 2004). Ces récepteurs membranaires sont des homodimères ou des hétérodimères. Les récepteurs de cytokines de classe 2 sont caractérisés par un domaine extracellulaire possédant des patterns très conservés (Bazan 1990; Thoreau, Petridou et al. 1991). En particulier ils sont formés de domaines FNIII (Fibronectine III) en tandem. La plupart des chaînes possèdent deux domaines FNIII en tandem sauf l'IFN-αR1 qui en possède quatre. On dénombre deux chaînes pour les IFNs de type I et deux chaînes pour ceux de type II, deux chaînes pour IL-10 et quatre chaînes transductrices pour les autres interleukines. On ajoute le TF « Tissue Factor » qui interagit avec le FVIIa (facteur de coagulation VIIa) ainsi qu'une chaîne réceptrice orpheline.

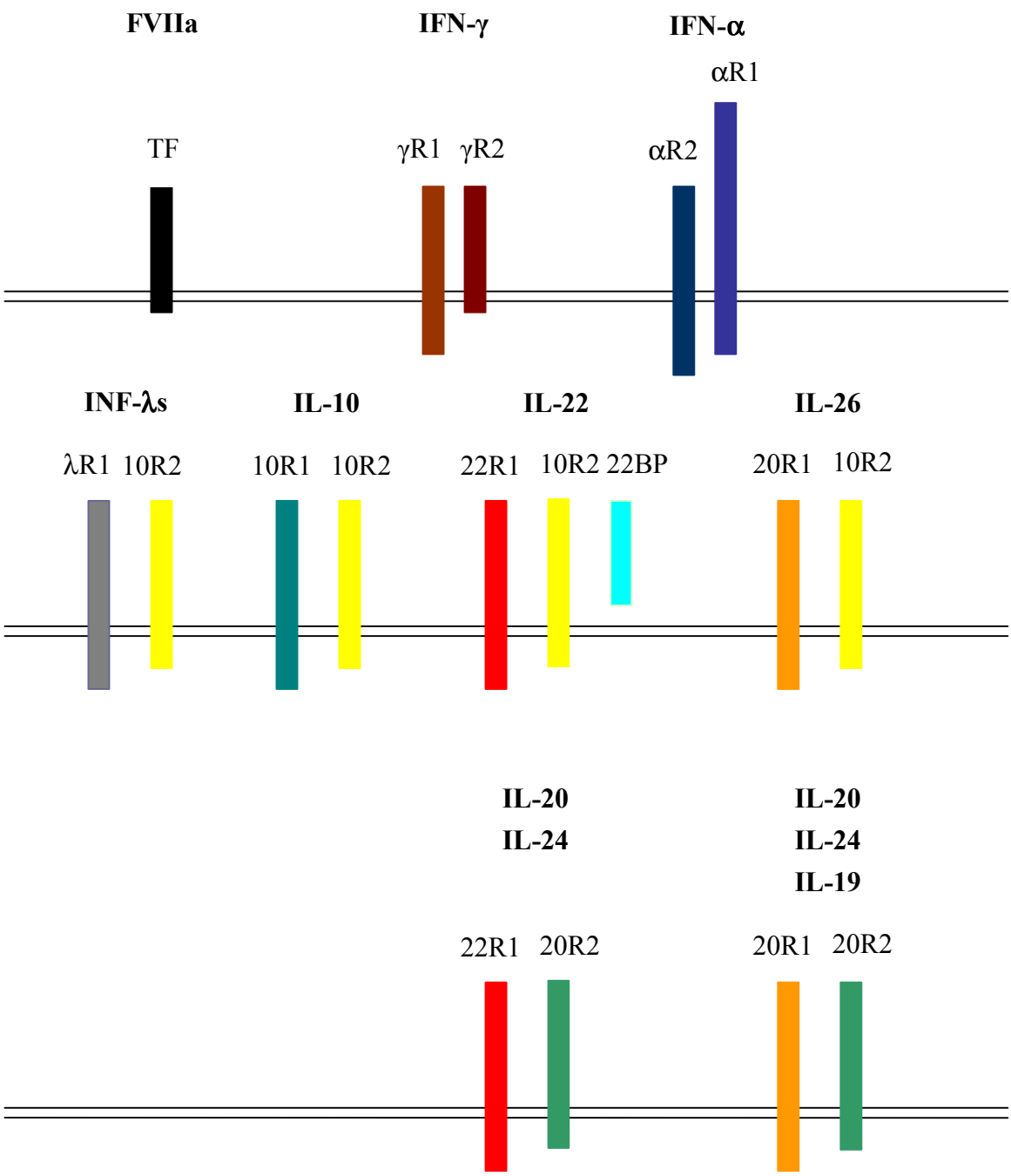


Figure 3. La famille IL-10/IFNs et les récepteurs hématopoïétiques de classe II associés

Chez l'homme, il existe des chaînes réceptrices de types R1 et R2. La nomenclature des chaînes de récepteur correspond au moment de leur découverte : IL-10R2 (CRF2-4), IL-20R1 (CRF2-8), IL-20R2 (CRF2-11), IL-22R1 (CRF2-9), IL-22BP (CRF2-10). Les chaînes de type R1 sont capables de se lier à l'interleukine avec une forte affinité. Elles possèdent des domaines intracellulaires longs qui sont associées avec la tyrosine kinase JAK1. Lors de la formation du complexe cytokine chaîne R1, les résidus Y sont phosphorylés et la chaîne recrute alors des protéines Stats impliquées dans la transduction du signal. Les chaînes R1 sont responsables de la spécificité du signal transduit. Les IL-10 et IFN-γ sont des homodimères (Ealick, Cook et al. 1991; Zdanov, Schalk-Hihi et al. 1995). Ils se lient à leurs chaînes transmembranaires R1 respectives IL-10R1 et IFN-γR1 (Figure 2) (Walter and Nagabhushan 1995; Zdanov, Schalk-Hihi et al. 1995; Josephson, Logsdon et al. 2001). La transduction du signal est activée lorsque les chaînes réceptrices de type R2 (IL-10r2 et IFN-γR2) s'engagent dans le complexe. Les sous-unités R2 ne sont donc pas capables de se lier aux interleukines directement. Elles possèdent un domaine intracellulaire associé à une protéine kinase : Jak2 pour IFN-γR2 (Bach, Tanner et al. 1996) ou Tyk2 pour l'IL-10R2 (Kotenko, Krause et al. 1997). La fonction de cette sous unité serait donc d'apporter une Tyrosine kinase supplémentaire au récepteur et permettre une activation croisée médiée par les kinases Jak (Kotenko, Izotova et al. 1999).

Les complexes des IL-19, IL-20, IL-22 et IL-24 requièrent deux chaînes réceptrices. Le récepteur de l'IL-22 est constitué des chaînes IL-22R1 et IL-10R2 . Cette dernière chaîne est aussi commune à IL-10 (Kotenko, Krause et al. 1997; Xie, Aggarwal et al. 2000; Kotenko, Izotova et al. 2001). IL-20 et IL-24 peuvent transduire leur signal via les récepteurs IL20R1/IL20R2 et IL22R1/IL20R2 (Blumberg, Conklin et al. 2001; Dumoutier, Leemans et al. 2001; Wang, Tan et al. 2002). Le récepteur de l'IL-19 est IL20R1/IL20R2 (Dumoutier, Leemans et al. 2001). L'IL-26 est capable de se lier au complexe IL-20R1/IL-10R2 (Sheikh, Baurin et al. 2004). Si le mode de transduction suit le schéma R1/R2 décrit précédemment pour les récepteurs hématopoïétiques de type 1, les mécanismes de fixation sont souvent différents d'un récepteur à l'autre. Par exemple, IL-22 peut se lier indépendamment à IL-22R1 et IL10R2 (Xie, Aggarwal et al. 2000; Kotenko, Izotova et al. 2001). IL-20 ne se lie qu'à l'hétérodimère IL-20R1/IL-20R2 (Blumberg, Conklin et al. 2001). Un récepteur soluble IL-22BP se lie spécifiquement à IL-22 de façon antagoniste (Kotenko, Izotova et al. 2001).

IL-10, IL-19, IL-20, IL-22 et IL-24 activent la voie Stat3 (Finbloom and Winestock 1995; Weber-Nordt, Riley et al. 1996; Wehinger, Gouilleux et al. 1996; Kotenko, Krause et al.

1997; Dumoutier, Louahed et al. 2000; Dumoutier, Van Roost et al. 2000; Xie, Aggarwal et al. 2000; Aggarwal, Xie et al. 2001; Blumberg, Conklin et al. 2001; Dumoutier, Leemans et al. 2001; Dumoutier, Lejeune et al. 2001; Kotenko, Izotova et al. 2001; Kotenko, Izotova et al. 2001; Wang, Tan et al. 2002) . IL-22 et IL-24 activent la voie Stat1 (Dumoutier, Van Roost et al. 2000; Xie, Aggarwal et al. 2000; Kotenko, Izotova et al. 2001; Kotenko, Izotova et al. 2001; Wang, Tan et al. 2002). IL-22 active la voie Stat5 (Dumoutier, Louahed et al. 2000; Xie, Aggarwal et al. 2000). Le domaine intracellulaire d'IL-22R2 est associé à la tyrosine kinase Tyk2 (Yan, Krishnan et al. 1996). On retrouve les homologues de toutes les chaînes réceptrices chez la souris. Les taux d'homologies sont très élevés sauf pour les domaines extracellulaires des chaînes IL-20R1 et IL-22R1. On trouve des motifs conservés entre la souris et l'homme contenant des résidus tyrosine qui serviraient de sites de liaison pour les protéines contenant des domaines SH2.

3.1.3. Récepteurs de cytokines de type III

Il s'agit des récepteurs TNF-R NGF-R, Fas, CD27, CD 40, et TRAIL-R (Figure 1). Il existe deux types de récepteurs du TNR-R p55 et p75. Les domaines sont riches en cystéines. Il existe un sous-groupe appelé « death-receptors ».

3.1.4. Récepteurs de cytokines de type IV

Cette famille rassemble les deux chaînes réceptrices de l'IL-1, IL-1R1 et IL-1R2 et la chaîne accessoire de l'IL-1R dénommée IL-1Racp (Figure 1) (Figure 2). On y inclut aussi l'IL-18R et la molécule T1/ST2. La région extracellulaire de ces chaînes réceptrices se caractérise par trois domaines Ig-like. La région intracellulaire possède des homologies avec des chaînes réceptrices d'autres organismes, tels que les récepteurs Toll chez la drosophile. Pour cette raison ils sont aussi appelés TLR (Toll-Like Receptors).

3.1.5. Récepteurs des facteurs de croissance apparentés à la superfamillle des immunoglobulines

Les chaînes réceptrices de ce groupe possédent une activité tyrosine kinase intrinsèque dans leur portion intracytoplasmique (Figure 1) (Figure 2). La région extracellulaire et constituée de cinq domaines Ig-like extra-cellulaires. Les membres de ce groupe sont le PDGFR, le M-CSFR, l'EGFR, le SCF (ou ligand de c-kit) et le ligand de flt3. Les récepteurs forment un homodimère après la fixation de leur ligand.

3.1.6. Récepteurs à activité sérine/thréonine kinase des cytokines de la famille du TGFβ

Ce groupe de récepteurs comprend entre autres le TGF-β, Inh-R et Act-R(Figure 1) (Figure 2).

3.1.7. Récepteurs à domaine sushi

Ce groupe est constitué par deux chaînes : l'IL-2Rα et l'IL-15Rα (Figure 1) (Figure 2).

3.1.8. Récepteurs des chimiokines

Les récepteurs de chimiokines sont caractérisés par la présence de sept jonctions transmembranaire dans leurs récepteurs couplés à une protéine G, qui permet la transduction du signal de type β-adrénergique. Les membres de ce groupe sont le XCR1, les CCRs, les CXCRs et le CX3CR-1(Figure 1) (Figure 2).

3.1.9. Récepteurs de la famille des interleukines 17

Ce groupe de récepteurs est caractéristique des IL-17, IL-17Rs (Figure 1).

3.2. Structure des gènes des cytokines chez *Homo sapiens*

Le séquençage des génomes eucaryotes et les expériences de génomique menées pour la localisation des gènes des cytokines permettent de connaître actuellement l'ensemble des locii dans lesquels sont situés les gènes des interleukines de type « IL-6 », « IL-2 » et « IL-10/IFNs ». Cette partie se limite aux gènes au sein du génome humain et ne traite pas de la structure des gènes homologues des interleukines au sein d'autres génomes de mammifères, en particulier celui de *Mus musculus*. Il faut aussi noter que les efforts de recherche sur la structure des promoteurs des gènes d'interleukines et sur les modes de régulation de la transcription sont récents (Holloway, Rao et al. 2002).

3.2.1. Structure des gènes de la famille IL-6

Chez l'homme, le gène *il-6* est situé dans la région chromosomique 7p21, possède une taille de 5 kb et est composé de 5 exons. La région promotrice contient de nombreuse boîtes de régulation de la transcription spécifique des GRE « glucocorticoïd Elements », AP-1 « Activator Protein A », SRE « c-Fos Serum responsive Element », CRE « cyclic AMP responsive Element », et une boîte IL-1 responsive Element sur laquelle peut se fixer le facteur de transcription NF-IL6.

Les gènes *lif* et *osm* se trouvent au niveau de la région chromosomique 22q12.2. Ces deux gènes sont placés en tandem et espacés de 16 kb l'un de l'autre. Ils ont le même sens de transcription.

Le gène *ct-1* est positionné sur le locus 16p11.2-p11.1 et possède une taille de 6 à 7 kb avec 3 exons. Le gène *il-11*se situe en 19q13.3-q13.4 long de 7 kb et consiste en 5 exons et 4 introns.

Le gène *g-csf* est situé dans le locus 17q11.2-q12, et possède 4 introns et 5 exons. L'expression du gène donne deux polypeptides qui diffèrent en longueur de 3 acides aminés mais qui possèdent la même activité.

Le gène *il12a* exprimant l'IL-12p35 est situé sur le locus 3p12-q13.2. Le gène pour la sous unité IL12p40 de l'hétérodimère IL-12 est lui positionné en 5q31-q30.

Le gène *il23p19* est situé dans la région chromosomique 12q13.13

Le gène *il27p28* est situé dans la région chromosomique 16p11.

Le gène *ebi3* a été localisé sur le locus 19p13.3 et est constitué de 4 intron et 5 exons pour une longueur de 8 kb.

Le gène *cntf* se situe dans la région chromosomique 11q12.2 et est constitué d'un seul intron. Il existe une co-transcription du gène *zfp91*,adjacent en 5' et qui code pour une protéine en doigt de zinc, avec le gène *cntf*. Ce transcrit possède la région codante de la protéine en doigts de zinc et une partie incomplète de la région codante du CNTF.

Le gène de la somatotropine *gh1* est situé en position 17q24.2 sur le génome. Un déterminant du LCR (Locus Control Region) est situé 14.5 kb en aval du promoteur de ce gène et joue un rôle critique, spécifique, et non redondant, en facilitant la liaison *trans* des facteurs de transcription sur le promoteur, et dans l'activation de la transcription (Ho, Elefant et al. 2002). Ce déterminant permet l'acétylation d'un domaine de 32 kb qui couvre l'ensemble du LCR et la région promotice du gène *GH1*. Ces résultats supportent un modèle d'activation à distance ciblée du gène *GH1* et médiée par le LCR, accompagnée par la propagation extensive de la méthylation des histones.

Le gène de la prolactine est situé dans la région chromosomique 6p22.2-p21.3. Il possède des zones promotrices alternatives tissus spécifiques positionnées 5,563 kb en aval de la région transcrite. Le promoteur situé en 5' est spécifique de l'expression de la PRL dans les « decidualized human endometrium » et dans les cellules lymphoblastoïdes telles que la lignée cellulaire humaine IM-9-P3. Le promoteur en aval est spécifique de l'expression gènique dans le lactotrope pituitaire est sous le contrôle du facteur PIT1.

Le gène *csh1* « corionic somatomammotropin hormone 1 » code pour l'interleukine PLL. Ce gène est positionné dans la région chromosomique 17q24.2, au sein du locus GH des gènes

codants pour les facteurs de croissance. Il mesure 1,7 kb et compte 5 exons et 4 introns aboutissant à 4 isoformes.

Le gène de la leptine (*ob*) est situé dans le locus 7q31.3. Il a une longueur de 20 kb et est constitué de 3 exons séparés par deux introns. Les 3 exons à eux seuls couvrent 15kb mais la partie codante est partagée entre l'exon 2 et l'exon 3. Le premier exon fait 30 bp et se situe 10,5 kb en amont du codon d'initiation. Les deux introns ont une taille de 2 kb. La transcription de la leptine est activée chez les cellules adipeuses et le promoteur contient les boîtes de transcription Sp1 et C/EBP «CAAT / enhancer binding protein ».

3.2.2. Structure des gènes de la famille IL-2

Plusieurs gènes des membres de la famille des IL-2 appartiennent à un cluster localisé dans la région 5q : centromère-*il-3*, *csf2, il-13, il-4*, *il-5, il-9*-télomère.

Le gène *il-3* est situé dans la région chromosomique 5q31.1. Sa taille est de 2,5 kb et présente 5 exons et 4 introns. Il est situé à 9kb en 5' du gène *csf2* qui code pour le GM-CSF. Le gène de *il-3* est activé par deux enhancers tissus spécifiques (Hawwari, Burrows et al. 2002). Le premier est situé à -4,5 kb et est sensible à la cyclosporine A. Il contient 3 sites NFat, ainsi que des boîtes AML1, AP-1 et Sp1. Le deuxième enhancer est situé à -14 kb du promoteur. Cet enhancer active la transcription du gène *il-3* chez les cellules mastocytaires et un sous ensemble de lymphocytes T. L'Enhancer à -14,5 kb participe, en synergie avec le deuxième enhancer, à l'activation de la transcription de ce gène par les cellules T.

Le gène *csf2* (GM-CSF) est orienté dans le même sens de transcription 5'-3' que le gène *il-3*. Il est composé de 4 exons et 3 introns pour une longueur de 2,3 kb. Le gène *csf2* est un marqueur génétique d'intérêt en raison des délétions dans cette zone qui sont impliquées entre autres dans le syndrome 5q- et la leucémie myélogénique aigue.

Le troisième membre du cluster est le gène *il-4*. Il a une taille de 10 kb et comprend 4 exons. Il est situé 12 kb en 3' du gène *il-13* et dans le même sens de transcription. Ces deux gènes ont une structure similaire et possèdent des boîtes de transcription pour les mêmes facteurs de transcription AP1, AP2, AP3, PEA3, HRE, TCF-1, GATA-3, STAT6 ainsi que des éléments

de type Enhancer (Smirnov, Smirnova et al. 1995; Kelly-Welch, Hanson et al. 2003). Chez l'homme le polymorphisme de la région *il*-4, *il*-13 a été mis en cause dans l'asthme d'origine héréditaire.

Le gène *il-5* est séparé du gène *il-4* par 310kb et possède une structure en 4 exons et 3 introns. Il existe une région de régulation à distance des gènes *il-4*, *il-13* et *il-5* appelée CNS1 « Conserved Non Coding Sequence 1 » (Loots, Locksley et al. 2000). Cette région fait 401 bp et se situe dans la région intergénique entre les gènes *il-13* et *il-4*.

L'*il-9* est l'élément du cluster 5q le plus proche du télomère. Le gène a une longueur de 4 kb et est constitué de 5 exons et de 4 introns.

Le gène *il-15* se situe sur le locus 4q31. Il est long de 34 kb et est constitué de 9 exons et 8 introns. Il existe un exon alternatif situé entre les exons 4 et 5 appelé exon 4a. Cet exon contient 3 codons stop à la suite suivi d'un codon d'initiation. L'isoforme obtenue par épissage alternatif « 1, 2, 3 4, 4a, 5, 6, 7, 8 » aboutit à la forme courte de l'IL15 qui possède un peptide signal raccourci mais fonctionnel. Cet isoforme est appelé SSP-IL15 «Short Signal Peptide IL-15 » en comparaison à l'isoforme long LSP-IL-15 « Long Signal Peptide IL-15 ».

Le gène *il-2* se situe dans la région chromosomique 4q26-q27. Il est situé dans un cluster qui comprend aussi le gène *il-21*. Le gène fait 5,2 kb et compte 4 exon et 3 introns.

Le gène *il-21* est aussi situé dans la région chromosomique 4q26-q27 à une distance de 155 kb et en 5' du gène *il-2*. Le gène a une taille de 8,4 kb et compte 4 exons et 3 introns même si la structure du gène ne correspond pas à celle du gène IL-2.

Le gène *epo* est situé dans la région chromosomique 7q21 il a une taille de 2,9 kb. Il est composé de 5 exons et 4 introns.

Le gène *tpo* est situé dans la région 3q26.3-q27. Il est constitué de 6 exons et 5 introns pour une taille de 6,2kb.

Le gène *il-7* est situé dans la région chromosomique 8q12-q13 et à une taille 72,8 kb. Il est constitué de 5 exons et de 4 introns.

Le gène *tslp* est situé dans la région chromosomique 5q22.1. Il existe deux formes transcrites. La forme longue fait 6,3 kb pour 4 exons et 3 introns. La forme courte contient les exons 3 et 4.

Le gène *il-31* est positionné dans la région chromosomique 12q24.3.

3.2.3. Structure des gènes de la famille des IL-10/IFNs

Les interleukines de la famille IL-10 (Renauld 2003; Kotenko and Langer 2004) sont situés dans des clusters sur le génome humain. Les gènes *il-10*, *il-19*, *il-20* et *il-24* sont situés sur le chromosome 1 dans la région chromosomique 1q32. Le sens de la transcription de l'*il-10* s'effectue vers le centromère à l'inverse des trois autres gènes d'interleukines qui sont à la suite l'un de l'autre et dirigés vers le télomère.

Les gènes *il-22* et *il-26* se situent dans la région chromosomique 12q14 à une distance de 30 kb l'un de l'autre. Ces deux interleukines se situent à 100kb du gène *inf-γ,* situé à la position 12q24.1.

Le gène *inf-γ* a une longueur de 6kb. Des études génétiques ont montrés que le gène *inf-γ* est impliqué dans la forme héréditaire de l'asthme (Cookson 1999). Du fait de leur proximité, les gènes *il-22* et *il-26* seraient également des candidats potentiels dans ce cas de l'asthme.

Il existe 23 gènes d'*inf-α* chez l'homme, de la taille de 1-2kb. Ils sont situés dans un cluster sur le locus 9p22. Il existe plusieurs classes de gènes *inf-α*. La classe 1 exprime des variants de taille comprise entre 156 et 166 résidus tandis que les produits des gènes d'*inf-α* de classe 2 atteignent une taille de 172 résidus. On ne sait pas si ces gènes sont exprimés ensemble par les mêmes voies d'activation, mais les gènes *inf-α-1*, *inf-α-2*, *inf-α-3* sont davantage exprimés que les autres. Ces gènes ne contiennent pas de séquences introniques. Les leucémies lymphoïdes sont souvent déclarées par des cellules qui ont subi des délétions dans la région 9p22.

Le gène de l'*inf-β* humain se trouve aussi dans la région chromosomique 9p22 à proximité du cluster *inf-α*. Le gène mesure 777bp sans introns. On connaît trois paralogues de l'*inf-β* chez le bovin.

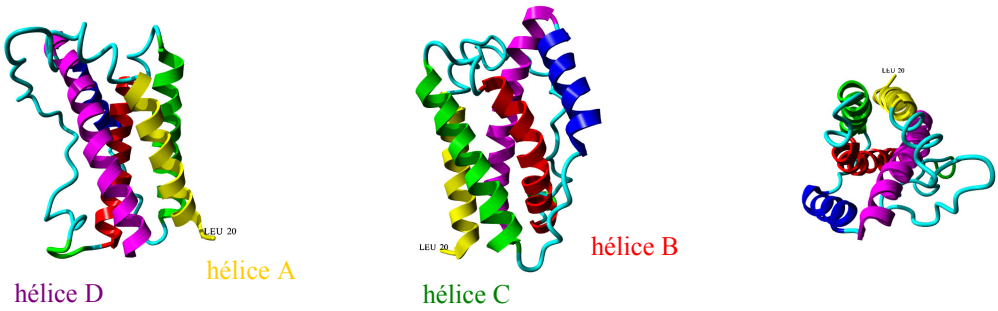
Les trois gènes des *inf-λ* sont situés dans le locus 19q13.

Les gènes de la famille des IL-10 ont des structures identiques composées de 5 exons. La position intron/exon est conservée entre les gènes, seule la taille des introns peut varier. Il existe un exon supplémentaire pour les gènes *il-19*, *il-22* et *il-24* en aval de l'exon 1. Pour *il-19* et *il-24*, il existe deux exons alternatifs. Les exons 1a1 pour *il-19* et 1a2 pour *il-24* possèdent un codon méthionine dans le même cadre de lecture que celui situé dans l'exon 1. Après traduction, les isoformes alternatives de ces deux interleukines possèdent un peptide signal anormal.

3.3. La classification structurale des cytokines

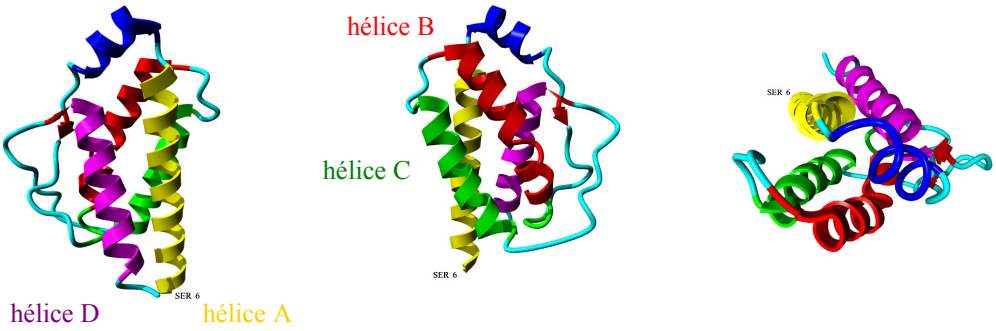
Les cytokines majoritairement en hélice α ont une topologie en faisceau de quatre hélices. Les hélices A et B sont orientées de façon parallèle et séparées par un boucle longue tandis que le boucles C et D sont orientées de façon antiparallèle (Simpson, Hammacher et al. 1997). Si on considère que l'hélice A est dirigée vers le haut (up), les hélices sont orientées en « «up-up-down-down ». De ce fait les hélices A et B sont diagonalement opposées aux hélices C et D. On distingue deux sous-familles structurales : les cytokines à hélices longues (IL-6) (Figure 4), à hélices courtes (IL-2) (Figure 5). Il existe une troisième sous-famille, IL-10/IFNs, dont les membres possèdent deux hélices supplémentaires en N-ter. La première différence entre les deux premières sous-familles se trouve au niveau de la longueur moyenne des séquences qui est inférieure à 150 résidus pour les cytokines à chaîne courte et supérieure à 160 résidus pour les résidus à chaîne longue. Les hélices des cytokines à chaîne longue sont longues de 30 résidus en moyenne tandis que la longueur des hélices des cytokines à hélices courtes mesure entre 10 et 20 résidus. Le paquet d'hélices est plus compact pour la sous familles des cytokines à hélices longues et les croisements des hélices forment des angles inférieurs à 40°. Dans la sous-famille des cytokines à hélices courtes, un de ces angles est supérieur à 40° (GM-CSF) (Harris, Presnell et al. 1994). L'observation des structures tridimensionnelles disponibles montrent que les cytokines à hélices courtes (Figure 5) ont la boucle AB qui passe au dessus de l'hélice D et en dessous de la boucle CD, et contiennent un court segment en feuillet β. Pour les cytokines à hélices longues (Figure 4), la boucle AB

redescend le long de l'hélice A, croise l'hélice D en formant une petite hélice α et passe le long de la boucle CD.



hIL-6 [1IL6]

Figure 4. Structure tridimensionnelle de l'IL-6 humaine [1IL6]

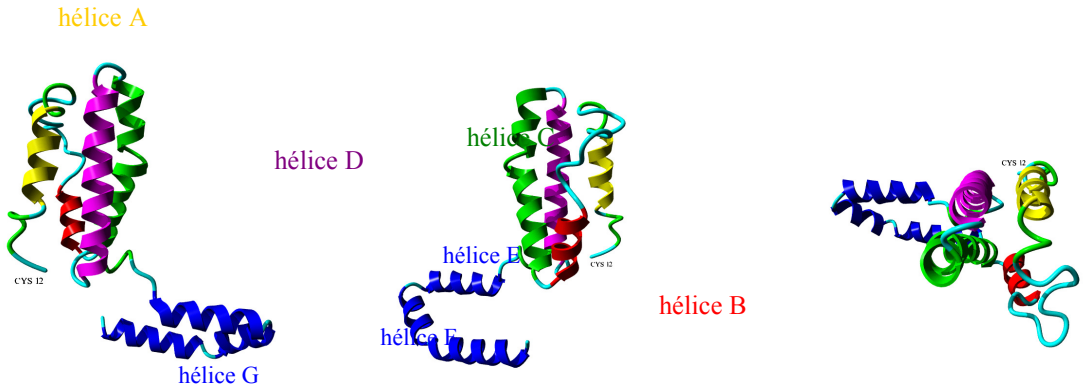


hIL-2 [1M47]

Figure 5. Structure tridimensionnelle de l'IL-2 humaine [1M47]

La troisième sous-famille IL-10/IFNs est caractérisée par des structures en hélices qui ressemblent de préférence à celles de la sous-famille des cytokines à chaînes courtes, tandis qu'elles sont compactées comme celles des cytokines à chaînes longues. L'IFN-β possède une cinquième hélice située dans la boucle CD (Senda, Saitoh et al. 1995) tandis que l'IL-10 (

Figure 6) et l'IFN-γ possédent deux hélices supplémentaires en C-ter qui permettent l'homodimérisation de ces deux cytokines.



ebvIL-10 [1VLK]

Figure 6. Structure de l'homologue de l'IL-10 du virus d'Epstein Barr : *Herpesviridae* (gammaherpes virinae) [1VLK].

On compte huit sous familles de cytokines dont le repliement tridimensionnel est formé par des feuillets β :

1. La famillle des TNFs dans laquelle on compte aussi la lymphotoxine , le CD27L « Cluster of Differenciation », le CD40 et le ligand FasL.
2. La Famille de l'EGF « Epidermal Growth Factor », du TGFβ, de la BTC «betacelluline », et de l'HB-EGF « Heparin Binding-EGF like Factor ».

3. La famille de cytokines tétraédriques à arrêtes en feuillets β de type FGF « fibroblast Growth Factor ».
4. La famille des IL-1/IL-18.
5. La famille IL-17, qui comprend les IL-17A, IL-17B, IL-17C, IL-17D (IL17A), IL-17 (IL-25), et IL-17F (IL-26/ML-1).
6. La famille des cytokines à noyau central compact à ponts disulfures de type NGF « Nerve Growth Factor » et BDNF « Brain-Derived Neurotrophic Factor ».
7. La famille des cytokines à noyau central compact à ponts disulfures de type PDGF « Platelet Derived Growth Factor », VEGF « Vascular Endothelial Growth Factor » et PlGF « Placental Growth Factor »
8. La famille des cytokines à noyau central compact à ponts disulfures de type TGFβ rassemblant les Acts « Activines », les Inhs « Inhibines », les BMPs « Bone Morphogenetic Proteins » et les GDNFs « Glial cell line-Derived Neurotrophic Factor ».

Les structures de type (α/β) posssèdent des domaines structuraux distincts formés exclusivement soit par des feuillets β parallèles (β-α-β) tandis que les structures (α+β) sont composées préférentiellement de feuillets β antiparallèles avec des domaines structuraux en hélices α et feuillets β clairement séparés. S'il n'existe actuellement aucune cytokine ne présentant un repliement de type (α/β). Les chimiokines sont en revanche caractérisées par des repliements de type (α+β) :

1. La LTN « Lymphotactine » est le seul représentant des chimiokines de type C.
2. la famille des chimiokines de type CC qui contient les MCPs « Monocytes Chemoattractant Proteins », les MIPs « Macrophage Inflammatory Proteins », la chimiokine RANTES « Regulated upon Activation Normal T-cell Expressed and presumably Secreted », TARC « Thymus and Activation-Regulated Chemokine », PARC « Pulmonary and Activation-Regulated Chemokine » et MDC « Macrophage-Derived Chemottractant ».
3. La famille des chimiokines de type CXC qui comprend IL-8, SDF « Stromal cell-Derived Factor », MIG « Monokine Induced by Gamma Interferon ».
4. la famille des chimiokines de type CX3C dont le seul membre est la FKN « Fractalkine ».

Chapitre 4. Les interleukines à quatre hélices alpha longues

4.1. L'interleukine 6 (IL-6)

L'IL-6 est une cytokine proinflammatoire (Hirano, Yasukawa et al. 1986; Hibi, Nakajima et al. 1996), (Ishihara and Hirano 2002), (Merberg, Wolf et al. 1992) , (Somers, Stahl et al. 1997). Elle est produite par un grand nombre de cellules de l'immunité telles que les fibroblastes, monocytes, les macrophages, les cellules dendritiques, les lymphocytes TH, les cellules B, les cellules endothéliales et les cellules gliales. L'IL-6 cible les hépatocytes pour induire l'expression des protéines de la phase aiguë de l'inflammation. L'IL-6 promeut aussi la différenciation terminale des lymphocytes B proliférants en cellules plasmatiques capables de libérer les anticorps. L'IL-6 stimule la différenciation des monocytes et la croissance des cellules souches hématopoïétiques. L'IL-6 est capable de promouvoir la différenciation des lymphocytes TH2 et d'inhiber simultanément la polarisation des lymphocytes TH1.

La séquence du précurseur de l'IL-6 humaine a une longueur de 212 résidus dont les 28 premiers constituent le peptide signal. Les séquences homologues on des taux d'identité (et de similarité) de séquences élevés : *Homo sapiens-Mus musculus* 39%(57%), *Homo sapiens- Bos taurus* 52%(71%) (Simpson, Hammacher et al. 1997). La séquence de l'IL-6 humaine contient un seul site de N-glycosylation. Cette interleukine possède deux paires de cystéines (C72-C78) et (C101-C111) très conservées chez les mammifères et les rongeurs et reliées deux à deux par des ponts disulfures intrachaînes. Chez l'homme et la souris les ponts sont situés aux même positions : C44-C50, C73-C83. La connectivité des cystéines est conservée dans les espèces. Ils existe une cinquième cystéine chez les IL-6 du rat et du chien.

La structure NOE RMN de l'IL-6 humaine (Nishimura, Watanabe et al. 1996) a montré que le repliement de la protéine est similaire à celui du G-CSF, pour un taux d'identité de séquence de 16%. Le complexe de basse affinité est un hétérodimère IL-6/IL-6Rα tandis que le complexe de haute affinité IL6, IL-6Rα , gp130 est un hexamère de stochiométrie 2 :2 :2 (Ward, Howlett et al. 1994). Ce dernier se forme en suivant un ordre séquentiel. La formation du complexe implique 4 sites sur l'IL-6. L'hétérodimère IL-6/IL-6Rα se forme en premier et implique le site I de l'IL-6. La chaîne gp130 transmembranaire interagit avec l'hétérodimère

au niveau du site II de l'IL-6 pour former un hétérotrimère. La formation de l'hexamère fait participer les sites III et IV des deux molécules d'IL-6. Le site III d'un des trimère interagit avec la chaîne gp130 de l'autre trimère et les sites IV permettent l'interaction entre les deux sous-unités IL-6 de l'hexamère.

Le site I se situe au niveau des extrémités C-ter de la boucle A-B et de l'hélice D de l'IL-6. Le site II est formé par les hélices A et C. Le site III comprend l'extrémité N-ter de la boucle AB, l'extrémité C-ter de la boucle C-D et l'extrémité N-ter de l'hélice D. Le site IV contient les résidus E106 de la boucle A-B et R113 de l'hélice B (Grotzinger, Kurapkat et al. 1997).

La structure du complexe hexamèrique a été cristallisée (Boulanger, Chow et al. 2003) à partir des régions extracellulaires de l'IL-6Rα et de la gp 130. La région cristallisée de l'IL-6Rα correspond au CHR « Cytokine binding Homology Region » D2-D3. Les auteurs ont utilisé la région de la gp130 contenant la région d'activation du récepteur D1 et le CHR D2-D3. Le site I (IL-6/IL-6Rα) implique les hélices A et D de l'IL-6 et le domaine D » de la chaîne spécifique. Le site II permet la liaison du dimère IL-6/IL-6Rα avec une chaîne gp130 et est divisé en deux domaines IIa et IIb. Le site IIa fait intervenir les hélices A et C de l'IL-6 avec la région située entre les domaines D2 et D3 de la chaîne gp130. Le site IIb permet l'interaction entre les domaines D3 respectifs de l'IL-6Rα et de la gp130. Le site IIb augmente l'affinité du site II. Séparément, l'IL-6 et l'IL-6Rα n'ont pas d'affinité avec la gp130. Le site III est caractéristique des interleukines capables de se lier à la gp130. Il est divisé entre les sites IIIa et IIIb. Le site IIIa est composé de la boucle A-B et de l'extrémité N-ter de l'hélice D de l'IL-6 qui interagissent au niveau du domaine D1 de la gp130 du second complexe ternaire. Le site IIIb se situe entre le domaine D2 de l'IL-6Rα et le domaine D1 de la gp130. L'apposition des interfaces du site IIIb dépend de la modification de la courbure des hélices B et D lors la formation du site I.

4.2. L'interleukine 11 (IL-11)

L'IL-11 (Paul, Bennett et al. 1990) est une protéine de 179 résidus et de poids moléculaire 23 kDa. Elle n'est pas glycosylée et ne contient pas de cystéines. Il existe un taux 88 % d'identité entre les séquences humaines et murines. Il existe 3 sites d'interaction avec les

chaînes du récepteur membranaire (Czupryn, McCoy et al. 1995; Tacken, Dahmen et al. 1999). Le site I interagit avec la chaîne spécifique IL-11Rα et se situe à la fin de la boucle A-B et à l'extrémité C-ter de l'hélice D. Les sites II et III interagissent avec la chaîne gp130. Le site II est localisé au milieu de l'hélice A et à l'extrémité de l'hélice C. Le site III est localisé au début de la boucle A-B et à l'extrémité N-ter de la boucle D.

4.3. Le Granulocyte colony-stimulating Factor (G-CSF)

Le précurseur du G-CSF humain est un polypeptide de 207 résidus de long (Nagata, Tsuchiya et al. 1986). La protéine mature mesure 177 résidus après la coupure du peptide signal situé en position N-terminal. La similarité de séquences entre les protéines homologues humaines et murines s'élève à 72 %. Le G-CSF humain contient 5 cystéines et 2 pont disulfures C36-C42 et C64-C74. La structure tridimensionnelle du G-CSF a été déterminée par cristallographie aux rayons X en 1993 (Hill, Osslund et al. 1993).

4.4. La sous-unité IL-12p35

L'IL-12p75 et un hétérodimère composé des sous-unités IL-12p35 et IL-12p40 (Brombacher, Kastelein et al. 2003). La mesure d'IL-12p40 est très souvent corrélée à la présence d'ILp75. L'IL-12p75 est une cytokine proinflammatoire qui fait le pont entre l'immunité innée et l'immunité adaptative. L'IL-12 est sécrétée principalement par les macrophages et les lymphocytes B et agit en synergie avec l'IL-2 pour cible les cellules T cytotoxiques activées et induire leur différenciation en lymphocytes T cytotoxiques. L'IL-12 cible aussi les cellules NK et les cellules LAKs (Lymphokine-Activated Killer) dans le but de stimuler leur prolifération. Si IL-4 est essentiel au développement de la réponse TH2, IL-12 l'est pour le développement TH1. L'IL-12 joue un rôle important dans la défense de l'hôte contre les infections intracellulaires par les bactéries, les parasites, les mycoses et les virus. L'IL-12 ne serait pas requis lors de l'initiation de la réponse anti-virale de type 1.

La sous-unité IL-12p35 présente de fortes homologies de séquence avec l'IL-6 et le G-CSF (Merberg, Wolf et al. 1992). Cette sous-unité ne semble pas être secrétée sous sa forme libre par les cellules mais en association avec IL-12p40. Le précurseur de l'IL-12p35 humaine mesure 219 résidus auquel il faut retrancher les 22 premiers résidus du peptide signal pour obtenir la forme mature. On dénombre deux sites potentiels de N-glycosylation mais pas de pont disulfure intrachaîne. La cystéine en position 96 permet la formation d'un pont disulfure interchaîne entre IL-12p35 et IL-12p40.

4.5. La sous-unité IL-23p19 (p19)

L'IL-23p19 est capable de se lier à la sous unité IL-12p40 pour former l'hétérodimère fonctionnel IL-23 (Oppmann, Lesley et al. 2000). La séquence de p19 a été découverte par recherche bioinformatique menée à partir du profil des séquences des membres connus de la sous famille IL-6 (Gribskov, McLachlan et al. 1987). Les sous-unités murines et humaines ont 70% d'identités de séquences et contiennent toutes les deux 5 cystéines mais aucun site de N-glycosylation. La séquence de p19 à un taux de similarité élevé avec IL-12p35, IL-6 et au G-CSF.

4.6. L'interleukine IL-27p28 (p28)

La sous unité IL-27p28 a été identifié récemment(Pflanz, Timans et al. 2002; Brombacher, Kastelein et al. 2003; Villarino, Huang et al. 2004) et possède une séquence très similaire à la sous unité IL-12p35. Cette sous unité forme un hétérodimère avec la sous unité EBI-3. Cet hétérodimère a été appelé IL-27, et la sous unité IL-27p28.

Le précurseur humain de l'IL-27p28 a une taille de 243 résidus. La séquence primaire se caractérise par une série de 13 résidus glutamates entre l'hélice C et l'hélice D.

4.7. Le « Leukemia Inhibitory Factor » (LIF)

L'interleukine LIF humaine a été identifiée en 1988 en sondant des banques génomiques humaines avec la séquence d'ADNc du LIF murin (Gough, Gearing et al. 1988; Moreau, Donaldson et al. 1988). Le LIF a par la suite été classé dans la famille structurale des interleukines à 4 hélices α (Bazan 1991; Robinson, Grey et al. 1994). La séquence primaire du précurseur du LIF humain mesure 202 résidus et contient un peptide signal de 22 résidus. La protéine mature mesure 180 résidus. Le LIF humain contient 3 ponts disulfures intrachaînes, C34-C156, C40-C156, C82-C-185, et 6 sites potentiels de N-glycosylations. La séquence du LIF humain possède 79% de similarité de séquence avec son homologue murin. La structure tridimensionnelle du LIF est très similaire à celle de l'OSM. La région structurale composée des résidus N-ter de l'hélice D, l'extrémité C-ter de l'hélice B et la boucle C-D est impliquée dans la fixation à la chaîne LIF-R (Robinson, Grey et al. 1994; Hudson, Vernallis et al. 1996; Hinds, Maurer et al. 1998). Une deuxième région structurale regroupe l'extrémité C-ter de l'hélice D et la boucle A-B et intervient dans la fixation à LIF-R. Une troisième région comprenant des résidus des hélices A et C intervient dans la fixation à la chaîne commune gp130. La rigidité de la boucle C-D du LIF murin (Purvis and Mabbutt 1997), que l'on ne retrouve pas dans les structures des interleukines les plus proches, peut être un facteur de spécificité à l'égard de la chaîne LIF-R.

4.8. L'Oncostatine M (OSM)

L'OSM (Zarling, Shoyab et al. 1986; Malik, Kallestad et al. 1989) est exprimé sous la forme d'un précurseur de 252 résidus. L'OSM mature mesure 196 résidus, est glycosylée et à un poids de 28kDa. Pendant la phase de maturation, le peptide signal N-ter de 18 résidus est clivé ainsi qu'un propeptide hydrophile de 31 résidus en C-ter. L'OSM possède deux sites de N-glycosylation et 5 cystéines dont quatre sont impliquées dans deux ponts disulfures intrachaînes entre les positions C31-C152 et C74-C192. Un modèle de la structure de l'OSM a été obtenu par la technique de modélisation par homologie (Kitchen, Hoffman et al. 1998).

4.9. Le « Ciliary neurotrophic factor » (CNTF)

Le CNTF humain est une interleukine dont la forme mature mesure 182 résidus et qui possède un peptide signal de 18 résidus (Masiakowski, Liu et al. 1991). Il possède 85 % de similarité de séquence primaire avec ses homologues de rat et de lapin (Lin, Mismer et al. 1989; Lam, Fuller et al. 1991; Negro, Tolosano et al. 1991). Il fait partie de la sous famille structurale des interleukines à 4 hélices α(Bazan 1991).

L'IL-6 et Le CNTF ont 6 % de similarité de séquence. Leurs sites respectifs de fixation à leur chaîne réceptrice spécifique se situe au niveau des extrémités C-ter de la boucle A-B et de l'hélice D (He, Chen et al. 1995; McDonald, Panayotatos et al. 1995; Panayotatos, Radziejewska et al. 1995). Le site II de fixation à la gp130 est constitué par les résidus des hélices A et C. Le site II du CNTF est similaire à celui du LIF et est constitué des résidus en C-ter de l'hélice B, des boucles B-C et C-D et de l'extrémité N-ter de l'hélice D. la substitution du site III de l'IL-6, spécifique à la gp130, par le site III du CNTF, capable de lier la chaîne réceptrice LIF-R a permis d'obtenir une molécule chimérique IL-6/CNTF capable de lier l'IL6R-α et de médier le signal via un hétérodimère gp130/LIF-R (Kallen, Grotzinger et al. 1999). Cette expérience a permis d'émettre l'hypothèse que les interleukines sont composées en régions spécifiques pour la fixation des différentes chaînes réceptrices.

4.10. Le « Cardiotrophin Like Cytokine » (CLC)

La CLC est identique à l'interleukine NNT-1 ou BSF-3. Elle a été identifiée à partir de l'exploration d'une base de données d'EST avec un modèle de markov caché basé sur des séquences signal connues (Shi, Wang et al. 1999). Le précurseur de la CLC mesure 225 résidus. La forme mature compte 198 résidus après clivage du peptide signal. La CLC s'associe à la sous-unité CLF-1 pour se fixer au récepteur CNTF-R.

4.11. La Cardiotrophine 1 (CT-1)

La CT-1 humaine mesure 201 résidus et possède un taux d'identité de séquence de 80% avec son homologue murin (Pennica, Swanson et al. 1996). (Hibi, Nakajima et al. 1996). La CT-1 ne possède pas de peptide signal N-terminal. La CT-1 humaine possède deux sites de N-glycosilation potentiels et 2 cystéines qui ne sont pas impliquées dans des ponts disulfures.

4.12. La Leptine (Lep ou ob)

Le précurseur de l'interleukine ob mesure 167 résidus et présente un peptide signal de 21 résidus en N-ter. Les homologues ob présentent des similarité de séquences fortes (Zhang, Proenca et al. 1994). L'ob humaine présente 85% de similarité avec l'ob murine. Il existe 94% d'identité entre l'ob murine et celle de rat. L'ob contient un pont disulfure C117-C167 et pas de site de N-glycosylation. La structure de l'ob humaine a été étudiée par cristallographie aux rayons X (Zhang, Basinski et al. 1997) qui a révélé un repliement en quatre hélices α longues.

4.13. La Prolactine

Le précurseur de la prolactine est un polypeptide de 227 résidus qui comprend un peptide signal long de 28 résidus (Shome and Parlow 1977). La prolactine humaine comporte 3 ponts disulfures C32-C39, C86-C202 et C219-C227 ainsi qu'un site de N-glycosylation potentiel.

4.14. La Somatotropine ou « Growth Hormone » (GH-1)

La GH-1 humaine est une interleukine de 191 résidus obtenus après le clivage du précuseur de 217 résidus et du peptide signal de 26 résidus. Elle contient 2 ponts disulfures C79-C191 et

C208-C215. La GH-1 est très similaire à la prolactine et au « Placental Lactogen ». Les études de cristallisation puis de co-cristallographie (Cohen and Kuntz 1987; de Vos, Ultsch et al. 1992; Sundstrom, Lundqvist et al. 1996) ont permis de mettre en évidence le repliement caractéristique en hélice α et la formation séquentielle du complexe de cette hormone de croissance avec son récepteur.

La GH-1 est capable de se lier aussi avec le récepteur de la prolactine mais pas l'inverse (Somers, Ultsch et al. 1994). La structure de la GH-1 a servi à la modélisation par homologie de l'OSM (Kitchen, Hoffman et al. 1998) et à l'étude de la forme cristallisée du LIF (Robinson, Grey et al. 1994).

Chapitre 5. Les interleukines à quatre hélices alpha courtes

5.1. L'interleukine 2 (IL-2)

L'IL-2 est considérée comme une cytokine centrale dans la régulation de la réponse des lymphocytes T (Bazan 1992), (Villarino, Huang et al. 2004), (Mott, Baines et al. 1995). Elle est sécrétée par les lymphocytes TH1, comme l'IFN-γ et le TNF-β. Ce sous-groupe lymphocytaire est responsable de fonctions classiques médiées par les cellules telles que la DTH « delayed-type hypersensitivity » et l'activation des lymphocytes T cytotoxiques. Les lymphocytes T de type TH2 sécrètent davantage les IL-4, IL-5, IL-6 et IL-10 et agissent de manière plus effective sur les lymphocytes T TH1 comme une aide pour l'activation des cellules B. L'IL2-R est un membre de la sous-famille des récepteurs à l'IL-2, qui inclut aussi les récepteurs aux IL-4, 7, 9 et 15. Il existe 3 formes d'IL-2R qui possèdent des affinités différentes pour l'IL-2. L'IL-2Rα est de faible affinité. Ce récepteur est aussi appelé « TAC antigen » pour son expression exclusive par les lymphocytes T activés. L'IL-2Rβγ est un récepteur dimérique d'affinité moyenne et l'IL-2Rαβγ correspond à un récepteur trimérique de haute affinité. La chaîne γ est commune à tous les récepteurs de la famille des IL2R et est responsable de la transduction du signal. Un défaut de la chaîne γ, due à une mutation du gène, est la cause de la maladie congénitale humaine XSCID « X-linked severe combined immunodeficiency ».

L'activation des lymphocytes T naïfs par le complexe CMH-antigène entraîne la formation des cellules blastocytaires et déclenche un phénomène continu de division cellulaire. L'induction de l'IL-2Rα joue un rôle majeur dans la stimulation initiale de la prolifération cellulaire et dans le support de la croissance cellulaire à long terme. En plus des lymphocytes T, l'IL-2 augmente également l'activité d'autres cellulaires immunitaires telles que les cellules NK, les cellules B, les monocytes/macrophages et les neutrophiles.

L'IL-2 humaine est traduite sous la forme d'un précurseur de 153 résidus. Sa forme mature mesure 133 résidus. Cette interleukine contient un pont disulfure intrachaîne entre la C78 et la C125. La structure de l'IL-2 a été étudiée par la méthode de cristallisation aux rayons X (Brandhuber, Boone et al. 1987). Ce travail a permis de révéler un repliement globulaire à

quatre hélices α. La substitution F42A sur la séquence de l'IL-2 empêche sa fixation à la chaîne IL-2Rα. Cette substitution modifie la conformation du domaine structural de l'IL-2 qui se lie à l'IL-2Rα, en particulier les résidus à l'entourage de la position substituée, mais pas la conformation générale de l'interleukine (Mott, Baines et al. 1995). D'autres expériences de mutagenèse dirigée ont permis d'identifier d'autres résidus de l'IL-2 humaine impliqués dans l'interaction avec la chaîne IL-2Rα : F44, K35, R38 et K43 (Weir, Chaplin et al. 1988; Weigel, Meyer et al. 1989; Sauve, Nachman et al. 1991). Le résidu D20 est important pour la fixation de la chaîne IL-2Rβ (Weigel, Meyer et al. 1989; Moreau, Bossus et al. 1995). Le résidu Q126 est impliqué dans la fixation de la chaîne γc (Buchli and Ciardelli 1993). Il n'existe aucune structure résolue du complexe de l'IL-2 avec son récepteur ou une de ses sous unités. Cependant la structure du complexe ternaire IL-2/IL-2R a été modélisée en prenant comme modèle la structure cristallisée de GH/GHR (Bamborough, Hedgecock et al. 1994). Une étude thermodynamique du complexe IL-2/IL-2R (Rickert, Boulanger et al. 2004) propose différents mécanismes de formation du complexe ternaire.

5.2. L'interleukine 3 (IL-3)

L'IL-3 est produite par une variété de types cellulaires incluant les lymphocytes TH, les cellules NK, les mastocytes. La fonction principale de l'IL-3 est de promouvoir la prolifération et la différenciation des progéniteurs hématopoïétiques. Cette interleukine cible aussi les mastocytes pour stimuler la croissance et la sécrétion d'histamine. L'IL-3 est critique pour le développement, la survie et la fonction des mastocytes et des basophiles sanguins. De ce fait elle peut être impliquée dans les réponses immunitaires dirigées contre les parasites et dans les réactions allergiques. La sécrétion d'IL-3 est influencée par différents facteurs telles que le stress chirurgical, un traumatisme, l'hémorragie, un stress environnemental (choc osmotique, radiation UV, la chaleur) ou des désordres psychiatriques.

L'IL-3 humaine a été identifiée au moyen de la séquence homologue murine (Yang, Ciarletta et al. 1986; Dorssers, Burger et al. 1987; Otsuka, Miyajima et al. 1988). Le précurseur de l'IL-3 mesure 152 résidus dont 133 sont présents dans la forme mature. La séquence contient 2 sites de N-glycosylations ainsi qu'une paire de cystéines C35-C103 formant un pont disulfure intrachaîne. La structure de l'IL-3 a été observée par la technique NOE RMN

(Feng, Klein et al. 1996) et a confirmé l'appartenance de l'IL-3 à la sous-famille des interleukines à hélices courtes. En particulier la comparaison de cette structure avec les structures cristallisées du GM-CSF et de l'IL-5 montre la conservation du type de repliement bien qu'il existe des différences dans la conformation des boucles A-B et C-D. Ce travail a décrit la présence d'une hélice supplémentaire A' dans la boucle A-B et appelée A'.

5.3. L'interleukine 4 (IL-4)

L'IL-4 est sécrétée par les lymphocytes TH2, et dans une moindre mesure par les mastocytes et les lymphocytes NK (Walter, Cook et al. 1992), (Wlodaver, Pavlovsky et al. 1992), (Milburn, Hassell et al. 1993), (Habib, Nelson et al. 2003), (Gallagher, Dickensheets et al. 2000). L'IL-4 aiguille le système immunitaire vers la réaction allergique TH2 dépendante. Cette interleukine est liée aux maladies atopiques telles que les allergies et l'asthme. L'IL-4 peut co-stimuler l'activation des cellules B marquées par l'antigène dans le but d'induire le passage de classe vers les IgE et Ig4 chez l'homme (IgE et IgI chez la souris). Cette interleukine induit aussi l'expression des cytokines TH2 telles que les IL-5, IL-6 et IL-9, certaines chimiokines telles que les éotaxines 1, 2, 3 et des molécules d'adhésion comme VCAM-1. L'IL-4 agit sur les cellules mastocytaires et les basophiles en augmentant l'activation de ces cellules pendant le phénomène allergique.

La séquence du précurseur de l'IL-4 humaine mesure 153 résidus. Elle contient un peptide signal long de 24 résidus. La forme mature mesure 153 aas. L'IL-4 contient 3 pont disulfures intrachaînes C27-C151, C48-C89 et C70-C123 nécessaires à son activité. Il existe deux sites de N-glycosylation. Le taux d'identité de séquences entre les Il-4 humaine et murine s'élève à 41 %.

La structure de l'IL-4 humaine a été résolue par cristallographie aux rayons X (Walter, Cook et al. 1992; Wlodaver, Pavlovsky et al. 1992) et par RMN (Powers, Garrett et al. 1992). Elle adopte un repliement en quatre hélices α similaire à ceux de l'hormone de croissance et du GM-CSF et de l'IFN-β. La structure contient aussi un feuillet β antiparallèle formé par deux brins β dans les boucles A-B et C-D.

5.4. L'interleukine 5 (IL-5)

L'IL-5 est sécrétée par les cellules TH2 et les mastocytes (Milburn, Hassell et al. 1993). Cette interleukine cible les lymphocytes B activés pour stimuler leur prolifération et leur différentiation. L'IL-5 est impliquée dans les réactions allergiques. Au contraire de l'IL-4, l'IL5 induit le passage de classe vers l'IgA. La fonction principale de l'IL-5 consiste en sa capacité à promouvoir la croissance et la différenciation des éosinophiles. Les éosinophiles expriment le récepteur Fcε en abondance. Ce récepteur joue un rôle très important dans la réaction allergique. Les éosinophiles vont fixer les complexes IgE-antigène par l'intermédiaire de la fixation du récepteur Fcε. Ce récepteur fait ainsi le lien entre les actions respectives de l'IL-4 et de l'IL-5 dans la réaction allergique. Des souris IL-5 déficientes ne présentent pas de développement de l'éosinophilie et de d'hyperréactivité aérienne après la mise en présence d'un allergène . L'IL-5 est une cible thérapeutique dans les phénomènes allergiques. Des tests avec un anticorps monoclonal bloquant anti-IL-5 ont montré une diminution du nombre moyen d'éosinophiles dans le sang et le septum et une prévention de l'éosinophilie. Cependant cet anticorps n'a pas montré d'effet significatif sur la réponse asthmatique tardive et l'hyperréactivité aérienne suggérant que les éosinophiles ne participent pas à ces phénomènes.

La séquence du précurseur de l'IL-5 humaine mesure 134 résidus. Le peptide signal a une longueur de 19 résidus et la protéine mature 115 aas. Elle contient deux sites de glycosylation. L'IL-5 humaine a un taux de 70% de similarité de séquence avec l'homologue murin. La structure cristallisée de l'IL-5 humaine (Milburn, Hassell et al. 1993) a montré un repliement voisin de l'IL-4 et en particulier la présence du petit feuillet β antiparallèle au sein des boucles A-B et C-D.

5.5. L'interleukine 7 (IL-7)

L'IL-7 a été identifiée en 1987 (Namen, Schmierer et al. 1988) comme étant un facteur de croissance et de différentiation des cellules B sécrétée par la moelle osseuse. Par la suite il s'est avéré que l'IL-7 agit aussi sur les thymocytes, les monocytes et les macrophages. L'IL-7

est un facteur de croissance des cellules T. L'ADNc de l'IL-7 code pour un polypeptide de 177 résidus et contient deux sites potentiels de glycosylation. L'IL-7 a une masse moléculaire de 25 kDa. Il existe une isoforme non fonctionnelle de l'IL-7. L'IL-7 contient 3 ponts disulfures intrachaînes C27-C166, C59-C154 et C72-C117. Les six cystéines sont conservées chez l'homme et chez la souris, de même que deux sites potentiels de N-glycosilation sur trois. Les IL-7 murines et humaines partagent 60 % d'homologies de séquences. La structure de l'IL-7 a été modélisée (Cosenza, Sweeney et al. 1997; Cosenza, Rosenbach et al. 2000) en prenant la structure de l'IL-4 comme matrice.

5.6. L'interleukine 9 (IL-9)

IL-9 a été découverte en 1988 dans les cellules T murines (Uyttenhove, Simpson et al. 1988). Chez la souris, l'IL-9 est très glycosylée. Sa masse moléculaire est égale à 39 kDa et passe à 14-16 kDa sous sa forme déglycosylée. Les précurseurs de l'IL-9 humaine (Renauld, Goethals et al. 1990) et murine sont formés de 144 résidus dont un peptide signal de 18 résidus et la séquence mature de 126 résidus. Les séquences murines et humaines ont 55 % d'homologie. La séquence murine de l'IL-9 est active sur les cellules humaines mais pas réciproquement.

5.7. L'interleukine 13 (IL-13)

L'IL-13 a été identifiée en 1993. La séquence primaire du précurseur de l'IL-13 murin mesure 151 résidus et son homologue humain 152 résidus. Le peptide signal de l'IL-13 mesure 20 résidus chez les deux espèces. Il y a trois sites de N-glycosylation sur l'IL-13 humaines et quatre chez l'homologue murin. Cependant l'IL-13 humaine est peu glycosylée *in vivo*. Les IL-13 des deux espèces possèdent des activités spécifiques similaires sur les cellules humaines tandis que l'activité de l'IL-13 humaine est dix fois inférieure à l'activité de l'IL-13 murine sur les cellules de souris.

La structure de l'IL-13 en solution a été obtenue par spectroscopie NOE RMN (Eisenmesser, Horita et al. 2001; Moy, Diblasio et al. 2001). Elle est semblable à celle de l'IL-4.

5.8. L'interleukine 15 (IL-15)

L'IL-15 a été identifiée en 1994. Sa forme mature a une masse moléculaire de 14-15 kDa pour une longueur de 114 résidus. L'IL-15 humaine possède 97 % de similarité de séquence avec l'IL-15 de singe et 73 % avec l'IL-15 murine (Grabstein, Eisenman et al. 1994). L'IL-15 possède quatre cystéines qui formeraient deux ponts disulfures intrachaînes C35-C85 et C42-C88. L'IL-15 possède deux sites de N-glycosylations du côté C-ter de la chaîne polypeptidique : N79 et N112. Le résidu D8 de l'IL-15 est indispensable à la fixation de la chaîne IL-2Rβ. Le résidu Q108 est essentiel au site de fixation de la chaîne γc.

L'IL-15 est exprimée sous une forme membranaire par les lignées monocytaires humaines (Musso, Calosso et al. 1999). Le mécanisme d'expression de la forme membranaire n'est pas encore connu mais est indépendant du peptide signal. Une hypothèse serait une co-expression de l'IL-15 et de l'IL-15Rα par les monocytes et la formation du complexe IL15/IL-15Rα au niveau de la membrane plasmique (Dubois, Mariner et al. 2002).

5.9. L'interleukine 21 (IL-21)

L'IL-21 a été isolée à partir d'une banque d'ADNc issue de cellules T CD3+ activées (Parrish-Novak, Dillon et al. 2000; Habib, Nelson et al. 2003). La fréquence de l'ADNc IL-21 dans la banque montre que l'expression de cette interleukine est très fortement régulée. Le polypeptide précurseur de l'IL-21 mesure 162 résidus et la forme mature 131 résidus. La séquence de l'IL-21 est très similaire de celle de l'IL-15. En particulier, ces deux interleukines possèdent quatre cystéines formant deux ponts disulfures intrachaînes. L'un de ces ponts est caractéristique de ces deux protéines dans la sous-famille des interleukines à chaînes courtes.

5.10. Le Granulocyte Macrophage Colony Stimulating Factor (GM-CSF)

Le GM-CSF est un monomère de 12 résidus. Le polypeptide précurseur mesure 144 résidus. La séquence de la forme mature est longue de 127 résidus et contient 2 sites de N-glycosylation. Son poids moléculaire varie entre 14 kDa et 35 kDa en fonction de l'état de N-glycosylation. Le GM-CSF possède deux ponts disulfures (C54/C96) et (C88/C121). Il existe une forme membranaire du GM-CSF impliquée dans la médiation juxtacrine du signal.

5.11. L'Erythropoïétine (EPO)

L'EPO humaine est une interleukine dont le poids moléculaire est égal à 34-37kDa. La protéine contient 3 sites de N-glycosylation N24, N36, N83 et un site d'O-glycosylation sur le résidu S126. Il existe deux ponts disulfures C7/C161 et C29/C33. Chez les rongeurs, le deuxième pont disulfure de l'EPO est absent du fait d'une substitution C33P. Les séquences des EPO présentent de très fortes homologies. La séquence d'EPO humaine à 85% de similarité de séquences avec son homologue murin et 91% avec l'EPO de singe. Il y a 63% des résidus de la séquence EPO qui sont conservés chez les mammifères (Wen, Boissel et al. 1993). La réactivité croisée de l'EPO a été observée entre différents organismes de mammifères. Cependant les EPO d'origine vertébré non-mammifère ne croisent pas chez les mammifères.

5.12. La thrombopoïétine (TPO)

La forme mature de la TPO humaine est longue de 332 résidus après clivage d'un peptide signal de 21 résidus et a un poids de 35 kDa.(Bartley, Bogenberger et al. 1994; de Sauvage, Hass et al. 1994; Foster, Sprecher et al. 1994; Sohma, Akahori et al. 1994). La TPO est constituée de deux régions. La première région, constituée des 154 résidus en N-ter est homologue à l'EPO (taux d'identité de séquence de 23%, conservation des positions des hélices α et du grand pont disulfure N-ter/C-ter). La région C-ter de la TPO contient six sites

de N-glycosylations potentiels qui pourraient jouer un rôle positif dans la stabilité de la protéine *in vivo*. Il existe deux sites protéolytiques potentiels de la forme mature. Le premier est situé entre les positions R153R154 juste après la région EPO-like et le deuxième au niveau de la paire R245R246.

5.13. La « Thymic Stromal Lymphopoïetin » (TSLP)

Le précurseur de la TSLP humaine a une longueur 159 résidus (Quentmeier, Drexler et al. 2001) et sa forme mature mesure 140 résidus. La TSLP humaine présente une similarité de séquence de 43 % avec l'homologue murin (Sims, Williams et al. 2000). Elle contient deux sites potentiels de N-glycosylation qui ne sont pas présents dans la séquence murine. Il existe également 7 cystéines chez l'EPO murine dont 6 sont conservées chez l'humain et sont responsables de la formation de trois ponts disulfures. Les ponts disulfures se font entre les paires de cystéines dans l'ordre (1, 6), (3, 4) et (5, 7).

5.14. L'interleukine 31 (IL-31)

L'IL-31 a été découverte récemment (Dillon, Sprecher et al. 2004). Cette interleukine est secrétée essentiellement par les cellules T activées et davantage par les lymphocytes TH2 que par les lymphocytes TH1. La séquence primaire de l'IL-31 a une longueur de 164 résidus et la forme mature de l'interleukine contient 141 résidus. La prédiction de structure secondaire effectuée sur la forme mature permet de définir quatre hélices α. Les auteurs classent actuellement l'IL-31 dans la sous-famille des interleukines à chaînes courtes. L'IL-31 serait alors la première interleukine de cette sous-famille à utiliser un récepteur dont les chaînes appartiennent à la sous-classe des chaînes réceptrices pour les interleukines à chaînes longues. L'homologue murin de l'IL-31 présente 31% d'identité de séquence avec son homologue humain.

Le signal de l'IL-31 est transduit par le récepteur hétérodimérique IL-31RA/OSM-R. Ce récepteur est exprimé de manière constitutive à la surface des cellules épithéliales et des kératinocytes. Les souris transgéniques qui surexpriment l'IL-31 montre des symptômes de dermatites et de prurit. L'IL-31 serait impliquée dans certaines formes de désordres épithéliaux ou dans des symptômes allergiques tels que l'asthme.

Chapitre 6. Les cytokines de la sous famille structurale de l'interleukine 10 et des interférons (IFNs)

6.1. L'interleukine 10

L'IL-10 est une cytokine anti-inflammatoire (Reineke, Sabat et al. 1998; Harrison and Wedlock 2000; Ball, Vigues et al. 2001; Fickenscher, Hor et al. 2002; Renauld 2003; Donnelly, Sheikh et al. 2004). Elle est sécrétée par les lymphocytes TH2 inhibe la production par les macrophages de cytokines proinflammatoires(IL-12) ainsi que la production par les lymphocytes des cytokines de type TH1 (IFN-γ). La production par les cellules présentatrices de l'antigène de l'IL-10 entraîne également une baisse du taux d'expression des CMH de classe II.

Un modèle du complexe entre l'IL-10 et son récepteur a été obtenu sur la base de la similarité structurale entre l'IL-10 et le IFN-γ (Zdanov, Schalk-Hihi et al. 1996).

6.2. L'interleukine 19

L'IL-19 est une interleukine de 177 résidus après traduction et dont la forme mature mesure 153 résidus. L'IL-19 a un taux d'identité de séquence de 21% avec l'IL-10. Ce taux peut varier entre 20 et 40 % avec les autres interleukines. La séquence de l'IL-19 contient deux sites potentiels de N-glycosylations (Gallagher, Dickensheets et al. 2000). Il existe trois ponts disulfures C10-C103, C57-C109 et C56-C109. La structure de l'IL-19 a été cristallisée (Chang, Magracheva et al. 2003). Elle présente des homologies structurales avec les IL-20, IL-22 et IL-24.

6.3. L'interleukine 20

L'IL20 humaine a été découverte à l'aide d'une méthode bioinformatique basée sur deux profils d'alignements structuraux des interleukines (Blumberg, Conklin et al. 2001). Le premier profil concerne le segment de peptide signal N-ter de précurseur. Le deuxième profil prend en compte une hélice amphiphatique caractéristique des cytokines. Le précurseur de l'IL-20 mesure 176 résidus et sa forme mature 152 résidus. Il existe 3 ponts disulfures C33-C126, C80-C132 et C81-C134 (Dumoutier, Leemans et al. 2001; Fickenscher, Hor et al. 2002; Donnelly, Sheikh et al. 2004).

6.4. L'interleukine 22

L'IL-22 a été découverte chez la souris parmi les gènes induits spécifiquement par l''IL-9 chez les cellules T (Dumoutier, Louahed et al. 2000). Elle a été appelée IL-TIF (IL-10 related T-cell-derived inducible factor). L'IL-22 a été rapidement clonée par la suite chez l'homme du fait de sa forte homologie avec l'orthologue murin (Dumoutier, Van Roost et al. 2000). L'IL-22 humaine est co-exprimée avec l'IFN-γ et l'IL-26 dans les cellules T activées (Wolk, Kunz et al. 2002). L'IL-22 murine possède un taux d'homologie de séquence avec IL-10 de souris de 22%. Le gène de l'IL-22 est situé sur le locus 12q14+3 dans la région des gènes de l'IL-26 et de l'IFN-γ. L'II-22 se lie au complexe IL-22R1/IL-10R2 (Dumoutier, Van Roost et al. 2000; Xie, Aggarwal et al. 2000; Kotenko, Izotova et al. 2001; Donnelly, Sheikh et al. 2004), qui peut activer les voies STAT3 , STAT1 et STAT5 (Kotenko, Izotova et al. 2001; Lejeune, Dumoutier et al. 2002). L'IL-22R est exprimé entre autres dans les reins, le pancréas, le foie (Kotenko, Izotova et al. 2001).

L'IL-22 augmente l'expression de plusieurs agents de la phase aigue de l'inflammation dans les hépatocytes (Dumoutier, Van Roost et al. 2000). L'IL-22 est aussi capable d'augmenter l'expression du gène SOCS3 (Suppressor Of Cytokine signaling-3) par les cellules de la lignée HCL (hepatoma cell line) (Kotenko, Izotova et al. 2001).

Un récepteur soluble appelé IL22-BP (Dumoutier, Lejeune et al. 2001) se lie spécifiquement à l'IL-22 et agit comme antagoniste de la voie STAT3 (Donnelly, Sheikh et al. 2004).

6.5. L'interleukine 24 (IL-24 ou MDA-7)

L'IL-24 humaine (Sauane, Gopalkrishnan et al. 2003) a été identifiée en 1995 sous le nom de MDA-7 comme une protéine dont l'expression augmentait dans les cellules de mélanomes humains en fin de différenciation. Chez le rat, elle a été découverte par deux expériences de « differential display » comme étant un gène surexprimé pendant la phase de cicatrisation ou comme étant un gène induit par l'expression de l'oncogène Ha-ras. Dans cette espèce, elle porte les noms respectifs CAC49A et Mob-5. l'orthologue murin de l'IL-24 est appelé FISP (IL-4-Induced Secreted Protein). L'IL-24 humaine et ses orthologues possèdent des similarités très fortes bien que leur fonction ne soient pas les mêmes au sein de leurs espèces respectives. Les homologues de la souris partagent 83% d'identité de séquence et respectivement 65 % et 70% d'identité avec l'IL-24 humaine. Si ces homologues proviennent d'un même gène ancestral, il semble que leurs fonctions se soient spécialisées au sein de l'espèce.

Le précurseur de l'IL-24 mesure 206 résidus. Sa forme mature est constituée de 155 résidus après le clivage du peptide signal. On dénombre trois sites potentiels de N-glycosylations. La structure de l'IL-24 n'a pas été déterminée.

6.6. L'interleukine 26 (IL-26 ou AKA155)

L'IL-26 (Fickenscher, Hor et al. 2002; Fickenscher and Pirzer 2004; Hor, Pirzer et al. 2004) a tout d'abord été identifiée comme la protéine AK155, lors d'une expérience de clonage. Elle est exprimée par les lymphocytes T humain transformés par le virus HVS « HerpesVirus Saimiri » (Knappe, Hor et al. 2000). La séquence d'ADNc a été identifiée lors d'une expérience de recherche de gène exprimé préférentiellement par le lymphocyte T humain transformé pas le virus simien. De façon surprenante, on n'a observé la surexpression de l'IL-26 que dans les lymphocytes transformés. Dans les lignées de lymphocytes T normaux, la protéine est exprimée très faiblement tout comme dans les cellules sanguines périphériques natives. L'IL-26 n'est pas non plus détectée chez les lymphocytes B. L'IL-26 est un

homodimère tout comme l'IL-10. La protéine AK155 IL-26 présente 24% d'identité et 47% de similarité avec l'IL-10 humaine. La structure prédite par homologie de l'IL-26 montre un repliement de 6 hélices ainsi que 4 résidus cystéines conservés. Le récepteur à l'IL-26 correspond au complexe IL-20R1/IL10R2 (Sheikh, Baurin et al. 2004). Aucun gène orthologue de l'IL-26 n'a été découvert chez la souris (Renauld 2003). Le précurseur de l'IL-26 mesure 171 aas. Une séquence peptide signal a été identifié et correspond aux 21 résidus. L'IL-26 jouerait un rôle de facteur de régulation autocrine pour les cellules T transformées par le HSV. Par extension il semble possible d'envisager l'expression de l'IL-26 lors de l'infection par le γ-herpesvirus sauvage qui implique des interactions très étroites entre les différentes cellules telles que les cellules épithéliales et les lymphocytes.

6.7. Les interleukines IL-28a, IL-28b et IL-29 : la famille des IFN-λ

La famille des IFN-λ regroupe 3 gènes distincts : les IFN-λ1 (IL-29), IFNλ2 (IL-28a) et IFNλ (IL-28b)(Kotenko, Gallagher et al. 2003; Sheppard, Kindsvogel et al. 2003; Donnelly, Sheikh et al. 2004; Langer, Cutrone et al. 2004; Logsdon, Jones et al. 2004). Les séquences IL-28a et IL-28b possèdent 96% d'identité. Les séquences d'IL-28a et d'IL29 humaines ont une identité de séquences de 81% (Sheppard, Kindsvogel et al. 2003) mais ne présentent respectivement que 11 et 13 % d'identité avec IL-10 et 15 et 19 % avec IFN-α et IL-22. Cependant l'alignement multiple de ces différentes séquences met en évidence la conservation des quatre cystéines ainsi qu'une forte compatibilité du profil structural en hélice α amphiphiles.

Par contre ils partagent l'utilisation de la chaîne IL-10R2. Le récepteur est nommé IFN-λR et se compose de la chaîne réceptrice appelée IL-28R qui se lie à l'interleukine et de la chaîne IL-10R2. L'IFN-λR est impliqué dans l'activation de STAT1 et de STAT2. Les protéines STAT1 et STAT2 forment un complexe transcriptionnel avec la protéine IRF-9 (IFN regulatory factor 9, p48) appelé ISGF3 (IFN-stimulated gene factor 3) qui induit la transcription des gènes des interférons de type 1, IFN-α et IFN-β (Donnelly, Sheikh et al. 2004). Les IFN-λ sont co-exprimés avec les IFN-α et IFN-β par les cellules infectées par un virus (Kotenko, Gallagher et al. 2003; Sheppard, Kindsvogel et al. 2003).

6.8. L'interféron α, membre de la classe des interférons de type 1

Outre la nouvelle classe des IFN-λ il existe deux autres classes majeures d'interférons distingués suivant sur leurs sources de sécrétion et leurs fonctions. L'IFN-α et l'IFN-β appartiennent à la classe des interférons de type 1. L'IFN-α est la première interleukine découverte (Isaacs and Lindenmann 1957). Il est principalement sécrété par les leukocytess et inhibe la réplication virale. Les interférons de type 1 ciblent les cellules inflammatoires et engendrent de nombreux effets responsables de l'hypersensibilité retardée « delayed-type hypersensitivity ». Les IFNs sont utilisés dans les traitements des patients souffrant de maladies respiratoires. Les interférons de type 1 se lient au récepteur commun IFN-αR (IFN-αR1/IFN-αR2). Il existe au moins 23 variants communs de l'IFN-α. Les variants ont des homologies de séquence très élevées, certains ne diffèrent que par un ou deux acides aminés. Les précuseurs humains ont une longueur qui peut aller jusqu'à 182 résidus incluant un peptide signal de 23 résidus. Les longueurs des séquences primaires des formes matures sont comprises entre 115 et 166 résidus. Il existe une région conservée située entre les positions 115 et 166. La région N-terminale est plus variable. Il y a deux ponts disulfures intrachaînes entre les positions C1/C98 et C29/C138 des IFN-α matures. Le second pont disulfure est essentiel à la fonction biologique de l'IFN-α. Il existe un site potentiel de glycosylation qui n'est pas fonctionnel dans la majorité des variants. La structure de l'IFN-α 2b humain a été cristallisée et observée à une résolution de 2,9 angstroms (Radhakrishnan, Walter et al. 1996).

6.9. L'interféron β, membre de la classe des interférons de type 1

L'IFN-β est sécrété par les fibroblastes et inhibe aussi la réplication virale. L'IFN-β est une protéine de 166 résidus qui possède 35% d'identité avec la séquence consensus de l'IFN-α et 50% avec la séquence IFN-β orthologue murine. L'IFN-β possède un site fonctionnel de glycosylation. La structure de l'IFN-β a été déterminée par cristallographie à 2,15 angström (Senda, Saitoh et al. 1995) et l'IFN-β2b à 2,2 angströms (Karpusas, Nolte et al. 1997). L'IFN-β possède un pont disulfure entre les résidus C31/C141 qui est nécessaire à son activité biologique.

6.10. L'interféron γ, membre de la classe des interférons de type 2

L'IFN-γ humaine est une protéine soluble de 17kDa composée de 143 résidus. La protéine contient deux sites de glycosylation. Le précuseur IFN-γ est long de 166 résidus, la région supplémentaire correspond au peptide signal. La séquence primaire de l'IFN-γ humain possède 40% d'homologie avec son orthologue murin.

L'IFN-γ fait partie de la classe des interférons de type 2. L'IFN-γ est sécrété par les lymphocytes TH1, les lymphocytes T cytotoxiques, et les cellules NK (Young 1996). La panoplie de fonctions connues de L'IFN-γ est plus étendue que celles des deux autres types d'interférons. L'IFN-γ est impliquée dans l'inhibition de la réplication virale, la surrégulation de l'activité des macrophages, l'augmentation de l'expression des molécules des CMH de classe I et II, l'induction du passage vers l'IgG2a pour les cellules B proliférantes et l'inhibition de la prolifération des cellules TH2 (Benveniste 1998; Gessani and Belardelli 1998; Frucht, Fukao et al. 2001). Le récepteur membranaire de classe 2 IFN-γR est un hétérotétramère constitué de deux sous-unités réceptrices IFN-γR1 associées chacunes à deux sous-unités IFN-γR2. L'IFN-γ se fixe au récepteur sous forme de dimère.

Partie 2. Apprentissage automatique

Chapitre 7. Classification des données

L'extraction des connaissances à partir d'un lot de données peut s'effectuer si l'on dispose d'une classification correcte de ces données. Les algorithmes d'apprentissage par la machine permettent d'opérer cette classification. Si les classes des données sont connues, la classification est dite supervisée. Le classifieur permet alors la reconnaissance d'un exemple dans une classe prédéfinie. Dans ce cas, les algorithmes utilisés ont pour objectif d'apprendre un modèle de la classification observée. Si aucune information préalable sur les classes n'est disponible, la classification est non-supervisée et les algorithmes d'apprentissage cherchent alors à trouver des propriétés partagées pour regrouper les données par similarité.

La première partie de ce chapitre décrit les différents types de données qui peuvent être rencontrés en classification. La section suivante introduit les différentes familles d'algorithmes d'apprentissage. La dernière partie de ce chapitre traite des techniques de mesures des performances des modèles de classification.

7.1. Les données

Une donnée est la représentation codée d'une information dans le but d'être stockée et manipulée. Les données peuvent être de nature qualitatives ou quantitatives.

Un exemple, ou instance, est un couple (x,u) où x est la représentation de l'objet « donnée », $x \in \chi$ et u est la classe de x. La classe $u \in C$, où C est l'ensemble des classes possibles pour un objet.

L'ensemble des instances est décrit au moyen d'un ensemble de caractéristiques ou attributs. Chaque exemple possède ses valeurs d'attributs propres. Les attributs peuvent être indépendants (e.g. couleur des cheveux, pratique d'un sport) ou corrélés entre eux. Les attributs peuvent prendre des valeurs nominales ou numériques. Les attributs nominaux, catégoriels ou discrets prennent leurs valeurs dans des ensembles symboliques prédéfinis (couleur). Les attributs numériques peuvent prendre des valeurs dans un ensemble continu entier ou réel (taille en cm, poids en kg).

Il est possible de décrire un concept de hiérarchie entre les attributs nominaux s'il existe (groupe rhésus). Dans ce cas on parle d'attribut nominal hiérarchique. Un attribut nominal ordonné est utilisé lorsque les valeurs des attributs apportent une information supplémentaire, pour un jeu de cartes par exemple (les cartes de couleurs rouges sont plus fortes que les cartes de couleurs noires). Plusieurs attributs peuvent être liés par une dépendance sémantique (personne mariée, nom de l'époux). Les attributs séquentiels ou chronologiques sont utilisés lorsque l'ordre des attributs à un sens (les noms « poule » et « loupe » n'ont pas le même sens même s'ils sont constitués par les mêmes lettres). Les attributs séquentiels numériques permettent de décrire les évolutions des valeurs d'un exemple (le cours d'une action).

Certains attributs peuvent apporter des informations sur le domaine de connaissances et sont considérés comme des métadonnées. Elles apportent une information sur les données. D'un niveau d'abstraction supérieur, ces attributs font référence à des relations entre les attributs qui caractérisent les données. Les relations peuvent être sémantiques (issues de l'association entre plusieurs données : consommation horaire en électricité d'un ménage), causales (un attribut doit être ajouté à un autre attribut pour renforcer la visualisation de l'information sous-jacente) ou fonctionnelles. Dans ce dernier cas, s'il existe une dépendance fonctionnelle entre un attribut A et un attribut B, alors l'attribut A découle implicitement de l'attribut B.

Le prétraitement des données est une étape à ne pas négliger avant leur utilisation dans un algorithme d'apprentissage de classification. Le premier travail consiste à récupérer toutes les données pertinentes au problème et à les rassembler au sein d'un ensemble d'instances. Ce travail peut conduire à rassembler différentes sources (et à agréger les données).

Le choix des attributs est déterminant pour les performances de la classification. Une image peut ainsi être représentée par son histogramme, son spectre de puissance, utiliser des descripteurs basés sur la texture ou des mesures morphologiques sur des contours. Le type de représentation le plus efficace dépend évidemment de la nature des données et de l'objectif poursuivi (recherche d'images par le contenu, détection de défauts sur des pièces usinées).

Il s'agit aussi d'éliminer les attributs les moins pertinents pour obtenir le sous-ensemble des attributs le plus petit contenant l'essentiel de l'information sur les données (arbres de décision). Il est aussi possible de choisir les attributs par extraction. Il s'agit alors d'appliquer une transformation linéaire (analyse en composantes principales) ou non-linéaire (SVM) pour réduire la dimension de l'espace d'entrée.

La performance de la classification dépend beaucoup de la qualité des données récoltées. En effet, le bruit présent dans les données doit être diminué pour éviter de biaiser les algorithmes de classification. Le premier traitement consiste à repérer et à caractériser les instances qui contiennent des attributs dont les valeurs ne sont pas connues, ou qui n'ont pas pu être obtenues. Il faut aussi faire attention aux valeurs erronées (les erreurs de mesures). Dans ce cas, un expert du domaine est capable de discerner une valeur vraie d'une anomalie. Le tri des données redondantes ne doit pas non plus être négligé.

7.2. Classification

Une classification se base sur des règles. Elles permettent d'affecter une donnée à la classe à laquelle elle appartient. Si le nombre de classes, C, est égal à 1 ou 2, il s'agit de l'apprentissage d'un concept. Dans le premier cas l'apprentissage s'effectue à partir d'exemples uniquement. Si $C = 2$, l'apprentissage se passe sur des exemples et des contre-exemples. Pratiquement, il est souvent plus facile de représenter les règles de classification par des fonctions séparatrices dans un espace d'hypothèses. L'apprentissage revient alors à trouver la meilleure hypothèse dans l'espace des hypothèses.

Les différentes étapes pour établir une classification par apprentissage sont :

- le choix du principe inductif (la minimisation du risque empirique),

- le choix d'une mesure de performance du modèle de classification,

- le choix de l'algorithme d'apprentissage,

- le choix de l'espace des hypothèses (dépendant de l'algorithme d'apprentissage),

- le réglage des paramètres de l'algorithme pour optimiser l'apprentissage.

7.2.1. Classification supervisée

Dans le cas d'une classification supervisée, les exemples de l'ensemble d'apprentissage sont étiquetés. La Figure 7 résume les différentes étapes du processus de classification supervisée. L'expert connaît la classification de l'ensemble d'apprentissage. Le but de l'apprentissage est d'obtenir la meilleure généralisation possible. La démarche consiste à adapter la complexité du système d'apprentissage à la difficulté du problème et à la quantité d'informations dont on dispose pour le résoudre. Dans ce contexte, les étapes de filtrage des données et de sélection des attributs sont des travaux de prétraitements importants pour améliorer la performance des algorithmes de classification. Le sujet de cette sous-partie présente les différentes techniques de classification supervisées.

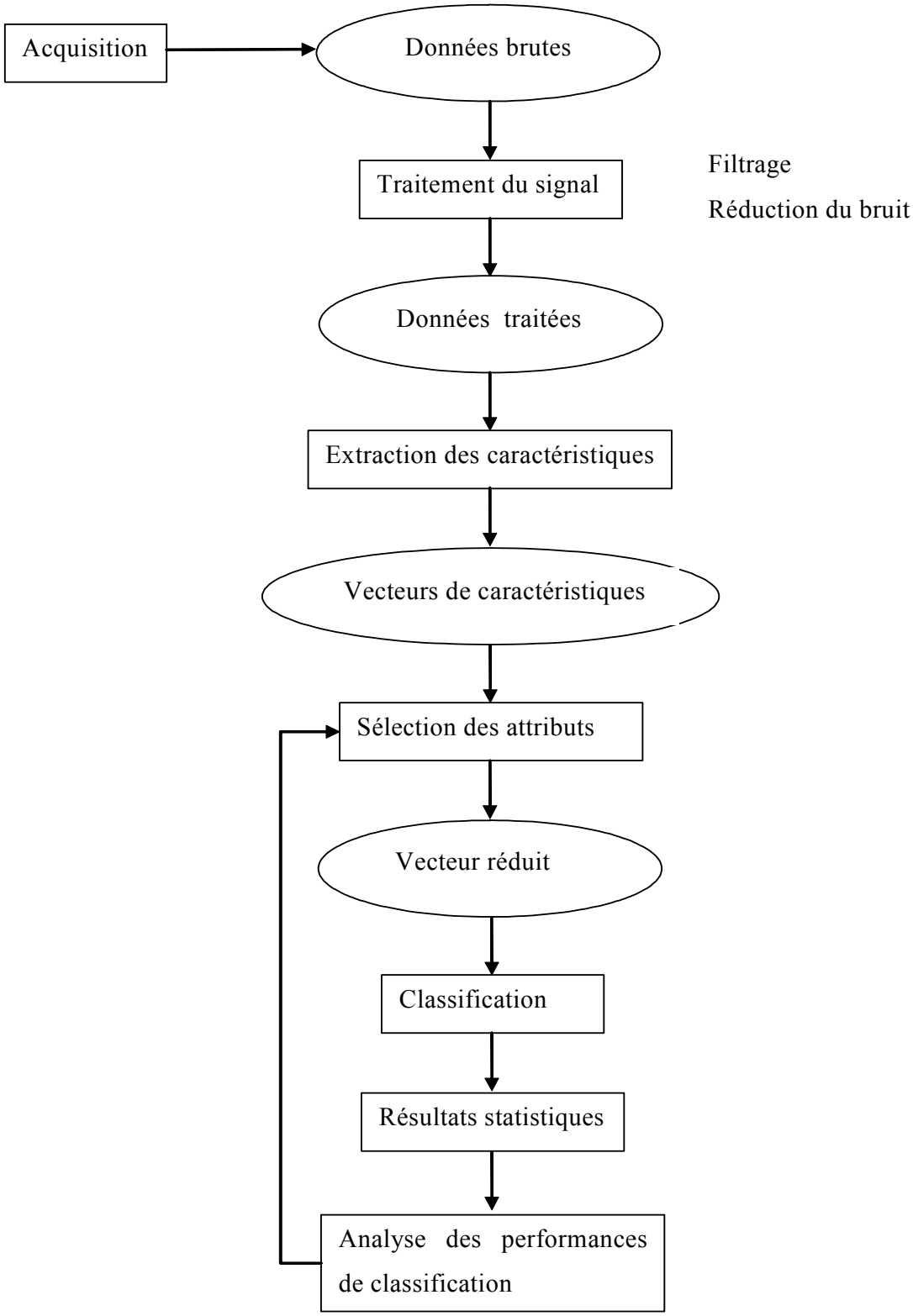


Figure 7. Schéma général d'une classification supervisée

7.2.1.1. Les arbres de décision

Les arbres de décision prennent en entrées les exemples exprimés sous forme de vecteurs d'attributs et donnent la classe prédite en sortie. Cette méthode offre une lisibilité des hypothèses choisies lors de l'apprentissage du modèle, ce qui la rend très attractive.

Un noeud de l'arbre correspond à une hypothèse sur un attribut (ou plusieurs attributs). Une feuille est un nœud qui ne possède pas de fils. Elle correspond à la classe qui est prise par tous les exemples qui arrivent jusqu'à elle. Les branches correspondent aux solutions possibles à chaque nœud (Figure 8). Si un nœud concerne un attribut nominal, les branches peuvent être au nombre des valeurs que peuvent prendre l'attribut. Si l'attribut est numérique, l'hypothèse traitée au niveau d'un noeud peut être la comparaison de sa valeur avec une valeur seuil. Il existe différentes solutions pour prendre en compte les valeurs manquantes, soit ajouter une branche spéciale au niveau du noeud, soit retenir la branche la plus souvent choisie. Un arbre de décision est construit récursivement en choisissant pour racine l'attribut le plus corrélé avec la distribution en classes (qui sépare le mieux les exemples). La construction de l'arbre s'arrête lorsque tous les exemples d'apprentissage sont rattachés à une feuille. La phase d'apprentissage est souvent suivie d'un élagage de l'arbre dans le but le rendre plus simple et limite les erreurs de classification. Les principaux algorithmes de création d'arbres de décision sont IDE3 (Quinlan 1979), C4.5 (Quinlan 1993) et CART (Breiman, Friedman et al. 1984).

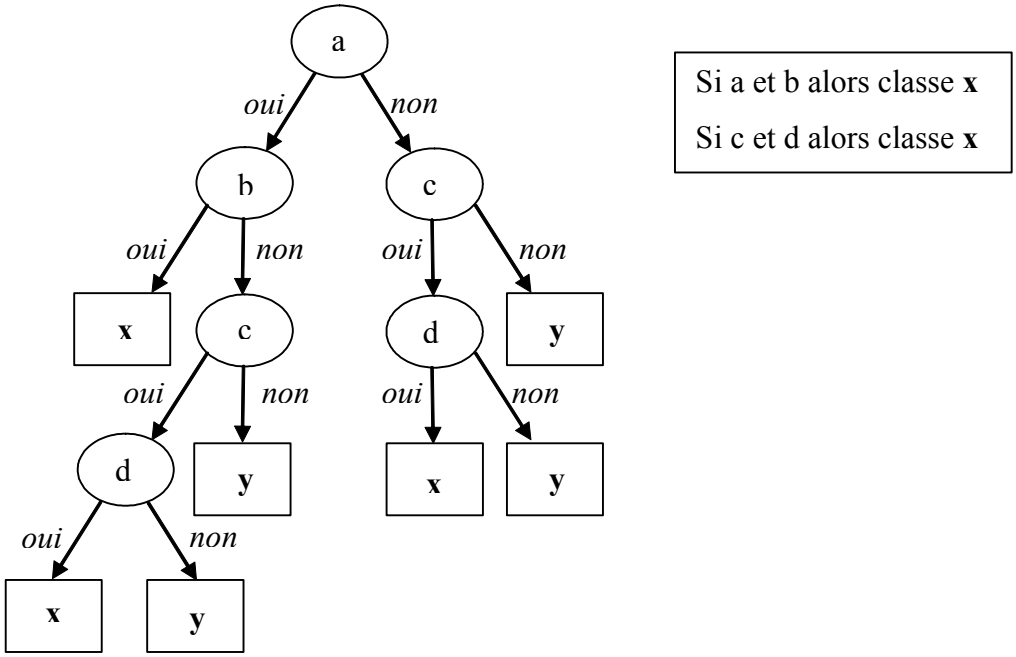


Figure 8. Exemple d'arbre de décision et règles de classification modélisant une disjonction.

On peut remarquer qu'un arbre de désision est un système de règles. Dans ce sens, les règles de classification sont une alternative aux arbres de décision. La précondition de la règle correspond aux hypothèses des nœuds des arbres de décision et la conclusion correspond à la classe donnée par la feuille en bout de chemin. Les règles de classification représentent les disjonctions plus aisément qu'un arbre de décision dans lequel chaque cas va donner lieu à un sous arbre (Figure 8).

Les règles d'association sont proches des règles de classification mais sont utilisées en fouille de données. Elles ne cherchent pas à classer véritablement mais à extraire une connaissance sur les données. Une règle d'association est représentée par la relation du type $A \rightarrow B$ où A est appelé la « prémisse » et B la « conclusion ». Les règles d'association possèdent davantage de liberté que les règles de classification dans le sens où elles sont capables de prédire non seulement la classe d'un exemple mais aussi la valeur d'un attribut associé à un contexte d'attributs donné. Il existe aussi des règles relationnelles capables d'exprimer la relation entre deux attributs (attribut a égal à l'attribut b, plus grand que, plus petit que), ce que ne peut pas non plus décrire un arbre de décision.

7.2.1.2. Les réseaux bayésiens

Dans les réseaux bayésiens, les attributs sont liés par des relations de probabilités conditionnelles (Figure 9). La représentation repose sur un graphe orienté sans circuit dans lequel chaque nœud possède une table de probabilités conditionnelles. Chaque arc représente une dépendance directe entre les variables reliées. Pendant l'apprentissage, le réseau bayésien tient compte des relations d'indépendances conditionnelles pour représenter la distribution de probabilités jointes de l'ensemble des variables de manière compacte. En effet, Il est possible de calculer la probabilité de tout groupe d'attributs à partir d'un autre groupe d'attributs connu, si l'on connaît une propriété locale d'indépendance.

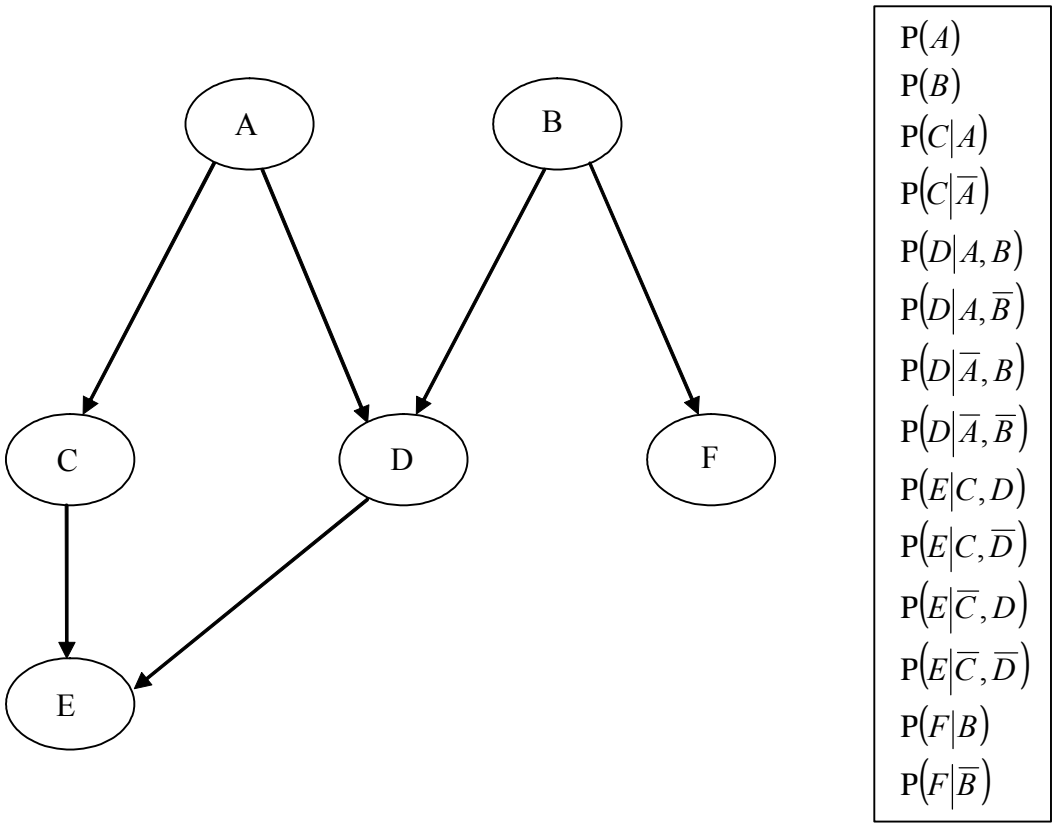


Figure 9. Exemple d'un réseau bayésien et du tableau des propriétés conditionnelles associées aux variables aléatoires

Si la structure du réseau est connue, il est possible d'effectuer un apprentissage à partir d'un ensemble de données incomplet. Les techniques qui permettent d'apprendre simultanément les probabilités conditionnelles et l'architecture du réseau se divisent en deux groupes : celles qui utilisent les indépendances conditionnelles pour construire le réseau et celles qui font appel à une fonction de score pour guider la recherche d'un réseau conforme au données.

7.2.1.3. Les modèles de Markov cachés

Les Modèles de Markov cachés ou HMMs « Hidden Markov model » sont des automates à états qui ont été décrits par Baum (Baum and Petrie 1966). Ils sont notamment utilisés en apprentissage de langue, en reconnaissance de texte ou en biologie pour l'annotation et la classification des séquences génomiques. Les HMMs sont utilisés lorsque des données sont représentées par des séquences d'évènements. A l'issu de l'étape d'apprentissage, le modèle HMM reflète la nature des évènements et la façon de leur enchaînement. On présentera ici l'application des HMMs aux séquences formées de caractères.

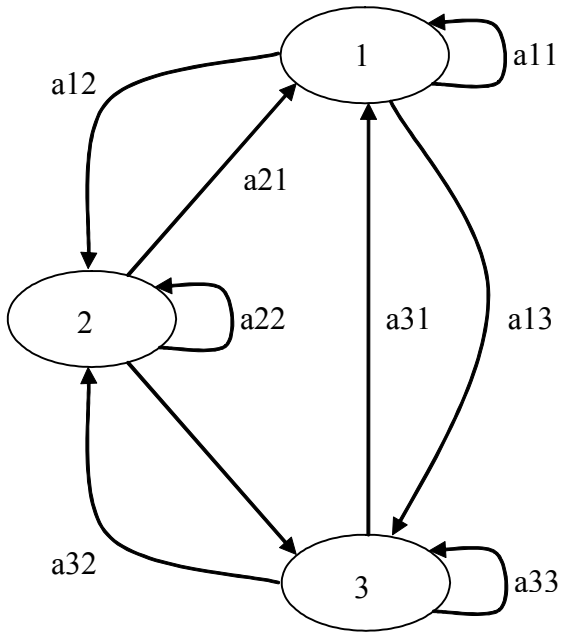
Un modèle de Markov observable est un processus stochastique observable, markovien et stationnaire. Le modèle de Markov caché généralise le modèle de Markov observable. Il est défini par un ensemble d'états et par une matrice de probabilité de transition d'un état vers un autre. Dans un HMM, un état n'est pas lié directement à un évènement qu'il émettrait à coups sûr mais contient une table de probabilité d'émettre chaque évènement (observation) possible.

Pour générer une séquence à partir d'un modèle HMM il faut tout d'abord choisir un état initial. Dans cet état, une observation est émise à partir de la densité de probabilité de l'état initial. Le passage à l'état suivant est choisi en utilisant la table de probabilité des transitions. Le nouvel état courant émet lui aussi une observation au hasard à partir de la densité de probabilité. La transition vers l'état suivant est choisie dans la table de probabilité de transition rattachée à l'état courant. Ces étapes sont répétées jusqu'à atteindre un état final.

Les deux architectures de base des HMM sont les modèles ergodiques et gauche-droite (Figure 10). Les modèles HMM ergodiques ont une structure telle que toutes les transitions

d'un état vers un autre sont possibles. Les HMM gauche-droite possèdent des contraintes nulles sur les transitions de manière à ne pas les autoriser. De ce fait, il n'est pas possible de retourner dans un état inférieur à partir de l'état courant. Les modèles HMM de ces deux catégories ont tous un état final et un état initial.

1.



2.

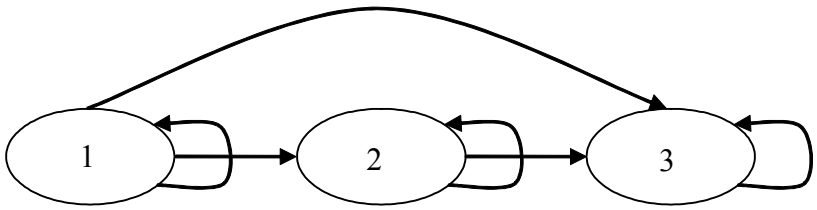


Figure 10. Exemples de structures de modèles d'HMMs. 1. ergodiques à trois états 2. gauche droite à trois états.

Soit x une séquence d'observation de longueur n, $x = x_1 x_2 \ldots x_n$ et soit Q une séquence d'états $Q = (q_1, q_2, \ldots, q_n)$. Les HMMs se basent sur deux hypothèses :

Pour tout modèle HMM λ, les observations sont indépendantes :

$$p(x_i / x_1 x_2 \ldots x_{i-1}, Q, \lambda) = p(x_i / q_i, \lambda)$$

La séquence d'état est décrite par une chaîne de markov d'ordre k :

$$p(q_i / q_1, q_2, \ldots, q_{i-1}) = p(q_i / q_{i-1}, \ldots, q_{i-k})$$

La réalisation des HMMs nécessite la résolution de trois problèmes fondamentaux ;

- La vraisemblance : comment évaluer la probabilité d'observation d'une séquence x étant donné un modèle HMM λ ?

- Le décodage : comment déterminer la meilleure séquence d'états d'un modèle HMM qui maximise la probabilité d'observation de la séquence d'évènements x ?

- L'apprentissage : comment effectuer l'apprentissage des paramètres du modèle HMM à partir d'un ensemble de séquences d'observation ?

Le calcul de la vraisemblance d'une séquence d'observation x connaissant un modèle λ revient à calculer la somme des probabilités jointes $p(x / \lambda)$ pour toutes les séquences d'états Q possibles. Cette sommation peut être calculée au moyen de l'algorithme de programmation dynamique *forward*. Soit la probabilité $\alpha_\tau(i)$ d'observer la séquence d'observation partielle $x_1 x_2 \ldots x_n$ et d'être dans l'état S_i au temps τ :

$$\alpha_\tau(i) = p(x_1 x_2 \ldots x_\tau, q_\tau = S_i / \lambda)$$

Le calcul des probabilités $\alpha_{\tau+1}(i)$ nécessite uniquement de connaître les $\alpha_\tau(i)$ et les probabilités de transition entre les états. Chaque $\alpha_\tau(i)$ peut alors être calculé récursivement.

La recherche de la meilleure séquence d'états équivaut à maximiser la probabilité $p(Q, x / \lambda)$. Elle s'effectue au moyen d'un algorithme similaire à l'algorithme *forward* qui est appelé algorithme de Viterbi (Viterbi 1967). La différence réside dans le fait qu'au lieu de sommer $p(Q, x / \lambda)$ pour toutes les séquences de Q à l'étape de récursion de l'algorithme *forward*, l'algorithme de Viterbi effectue la maximisation de $p(Q, x / \lambda)$.

L'étape d'apprentissage consiste à maximiser $p(x/\lambda)$. L'algorithme d'apprentissage le plus souvent utilisé est celui de Baum-Welch et se base sur les algorithmes *backward* et *forward* . L'algorithme *backward* utilise les valeurs α_τ obtenues par l'algorithme *forward* et les variables β_τ obtenues par l'algorithme *backward.* Cet algorithme permet de calculer la probabilité de la séquence d'observation partielle du temps $\tau+1$ jusqu'à la fin étant donné l'état S_i au temps τ et le modèle λ .

$$\beta_\tau(i) = p(x_{\tau+1}x_{\tau+2}\ldots x_n / q_\tau = S_i, \lambda)$$

L'algorithme de Baum-Welch est une adaptation de l'algorithme EM (Baum, Petrie et al. 1970). Soit λ_i l'ancien modèle HMM, la suite des modèles HMM obtenus par l'algorithme de Baum-Welch vérifie la relation cherchée :

$$p(x/\lambda_{i+1}) \geq p(x/\lambda_i)$$

En bioinformatique, les profils HMM sont utilisés pour la comparaison des séquences biologiques. En particulier ils peuvent modéliser un alignement multiple de séquences d'ADN (Figure 11). Le coeur d'un profil HMM est une collection linéaire d'états match (M), un par colonne consensus de l'alignement multiple.

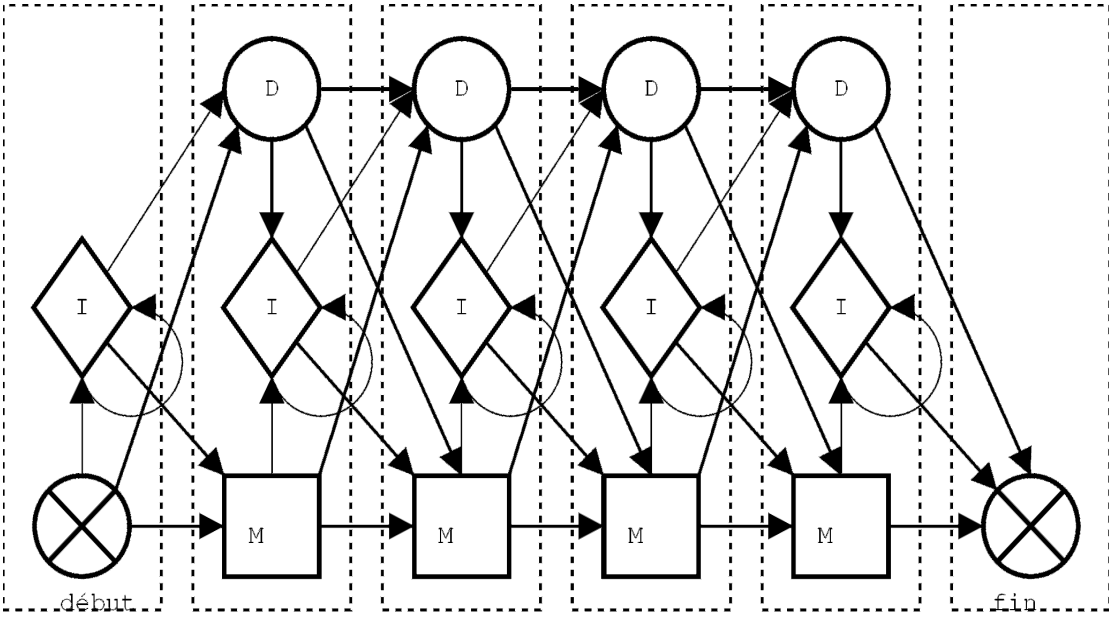


Figure 11. Architecture d'un profil HMM pour un alignement de séquences

Chaque état M émet (s'aligne à) un seul nucléotide avec une probabilité qui est déterminée par la fréquence du nucléotide observée dans la colonne correspondante de l'alignement multiple.

Chaque état M est associé à deux autres états: insertion (I) et délétion (D). Ces deux états supplémentaires permettent de modéliser un domaine conservé complet et non juste un motif sans indels. L'ensemble défini par ces 3 états à la même position de l'alignement multiple est appelé "noeud". Les états sont inter-connectés entre eux par des *probabilités de transition*.

L'architecture des profils incluse dans le logiciel **Hmmer** (Eddy 1995; Eddy, Mitchison et al. 1995) se distingue de celle, standard, brièvement explicitée plus haut par la suppression des probabilités de transition entre les états D $\rightarrow$ I et I $\rightarrow$ D et par la présence d'états supplémentaires qui permettent d'apprendre indifféremment à partir d'alignements multiples locaux ou globaux.

7.2.1.4. Les grammaires

L'inférence grammaticale englobe les techniques qui consistent à extraire une grammaire à partir d'un ensemble de séquences exemples capable d'engendrer un langage. On peut définir une grammaire comme un objet mathématique auquel est associé un processus algorithmique permettant d'engendrer un langage.

On distingue les grammaires régulières « context free » et les grammaires algébriques. Elles modélisent respectivement des langages réguliers et algébriques. Les deux types de grammaires se distinguent par le fait que toutes les grammaires régulières sont représentées par des automates finis. Réciproquement un automate fini représente un langage régulier. Les grammaires algébriques s'appliquent à des espaces de recherches plus complexes et elles ne peuvent pas être représentées par des automates finis.

Une grammaire G est définie par un alphabet de symboles non terminaux N et d'un alphabet terminal $\sum$ disjoint, par un ensemble fini de règles et par un axiome S, $S \in N$.

Soit Σ^* l'ensemble de tous les mots sur Σ, un mot $v \in \sum^*$ est engendré par la grammaire G s'il peut se dériver de l'axiome S.

Le langage $L(G)$ engendré par G est l'ensemble des mots de $\sum^*$ engendré par G.

Un automate fini (Figure 12) est constitué de sous ensembles d'états initiaux et d'états finaux inclus dans l'ensemble des états Q, d'un alphabet fini $\sum$ et d'une fonction de transition qui est une application $Q \times \sum \to 2^Q$.

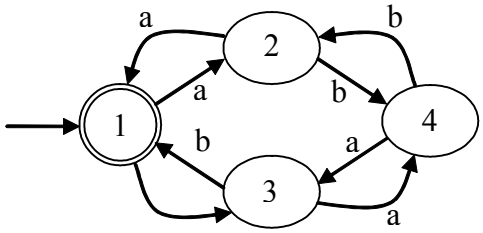


Figure 12. Exemple d'automate à états finis pour un langage régulier. L'automate est capable de représenter les mots composés d'un nombre pair de a et d'un nombre pair de b.

On distingue les automates finis déterministes et les automates finis non déterministes. Un automate fini déterministe contient un seul état initial et chaque fonction de transition d'un couple $(q \in Q, a \in \sum)$ contient au plus un élément.

L'automate déterministe minimal d'un langage désigne l'automate à états finis déterministe qui possède le nombre minimal d'états (Figure 13). Il est démontré que cet automate est unique pour tout langage régulier.

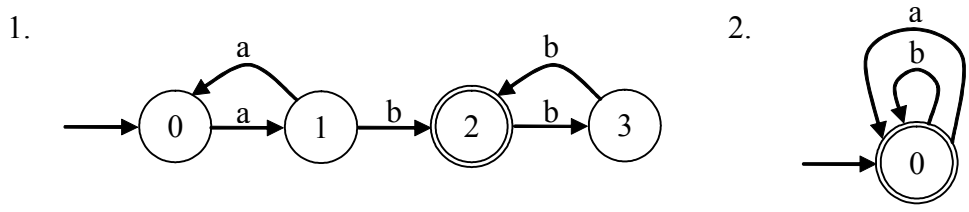


Figure 13. Exemples d'un automate déterministe minimal pour un langage régulier et d'un automate universel. 1. Automate déterministe minimal pour le langage régulier contenant les mots composés d'un nombre impair de caractères a suivi d'un nombre impair de caractères b. 2. Automate minimaliste capables de composer tous les mots de l'alphabet $\{a,b\}$.

7.2.1.5. Les réseaux de neurones artificiels

Le cerveau est un organe constitué par un réseau de neurones interconnectés les uns aux autres au moyen des nerfs eux-mêmes terminés par des synapses. La nature du réseau neuronal dépend du nombre de neurones et de la manière dont ils sont connectés. Chaque connexion entre deux neurones est caractérisée par l'émission d'un potentiel d'action dont l'amplitude va influer ou non sur le neurone récepteur. Par analogie avec la biologie, les réseaux de neurones artificiels sont capables d'effectuer des calculs complexes grâce à la propagation d'informations entre des unités élémentaires de calcul. L'apprentissage d'une classification par les réseaux de neurones équivaut à optimiser les poids sur les interconnexions entre les neurones. De ce fait les réseaux de neurones sont adaptés, en classification, aux données numériques.

Les travaux fondateurs dans cette discipline sont attribués à J. Mc Culloch et W Pitts en 1943 (McCulloch and Pitts 1943). Le réseau le plus simple est issu du percepton (Rosenblatt 1958). Cependant cette unité simple n'est pas capable de représenter la fonction logique XOR. En 1986, les travaux sur les réseaux de neurones sont relancés avec la découverte de l'algorithme de rétropropagation (Rumelhart, Hinton et al. 1986) qui permet l'apprentissage des réseaux de neurones par les exemples.

L'unité de base, le neurone, est un automate à seuil (Figure 14). On distingue 3 types de neurones : les neurones d'entrée, les neurones de sortie et les neurones cachés. Chaque neurone d'entrée prend la valeur d'une des composantes du vecteur d'exemple en entrée tandis que les neurones de sortie vont émettre une décision sur la classification de l'exemple. Les valeurs des neurones de sorties sont rassemblées dans un vecteur de classification qui permet d'obtenir la classe prédite pour l'exemple. Les neurones cachés sont rassemblés dans des couches intermédiaires. Les neurones disposent d'une fonction d'entrée qui donne l'état du neurone σ_i en fonction des poids $w(j,i)$ sur les interconnexions du neurone. Pour les neurones d'entrés, la valeur de σ_i est égale à la valeur de la composante de l'exemple. La nature du signal de sortie est dépendante de l'état du neurone et d'un seuil d'activation. Elle est déterminée au moyen d'une fonction d'activation ϕ. Il existe trois types de fonctions d'activation de base : la fonction seuil, la fonction signe et la fonction sigmoïde (Figure 15).

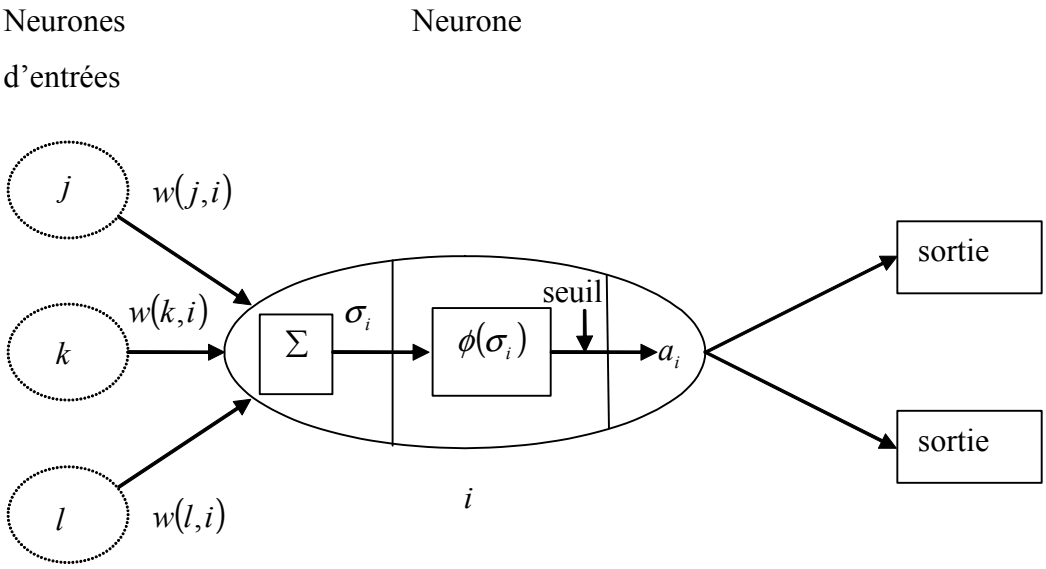


Figure 14. Schémas d'un neurone artificiel

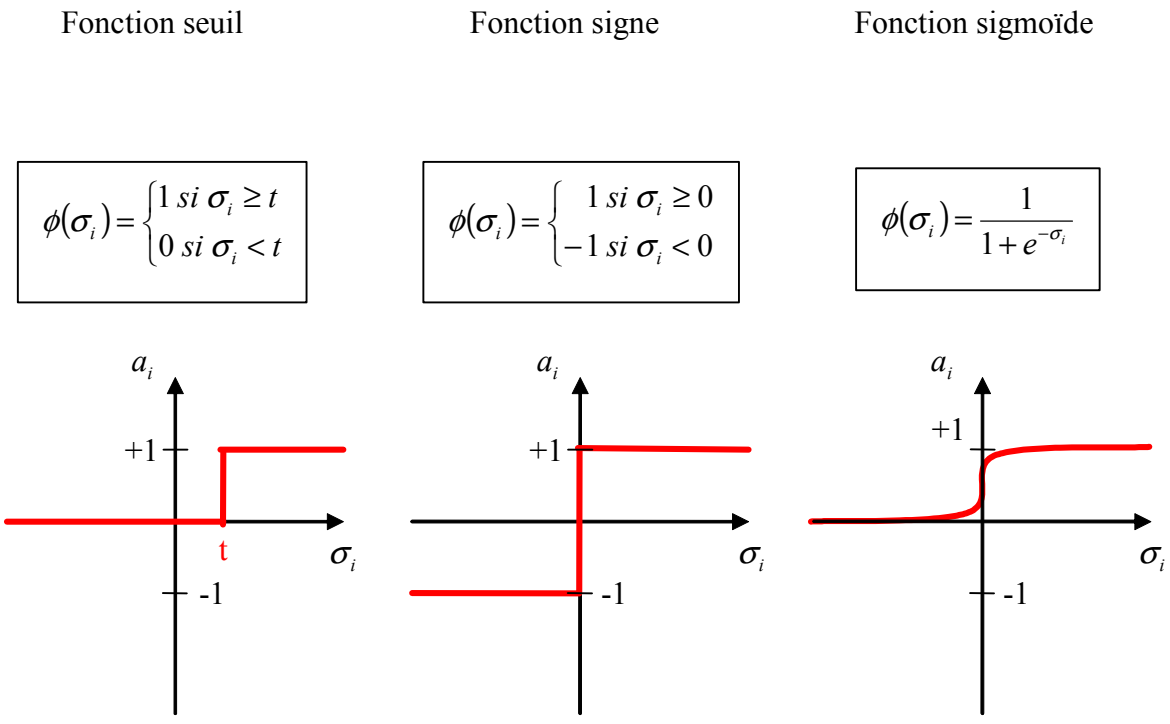


Figure 15. Exemples de fonctions d'activation

Il existe différentes architectures de réseaux de neurones. Le perceptron possède l'organisation la plus simple. Il n'existe pas de couches de neurones cachés entre les neurones d'entrée et les neurones de sortie (Figure 16). Les unités de sortie sont indépendantes les unes des autres et chaque poids est affecté à une seule sortie. Il est possible de limiter l'étude des réseaux de perceptrons (Figure 16) au cas du perceptron seul.

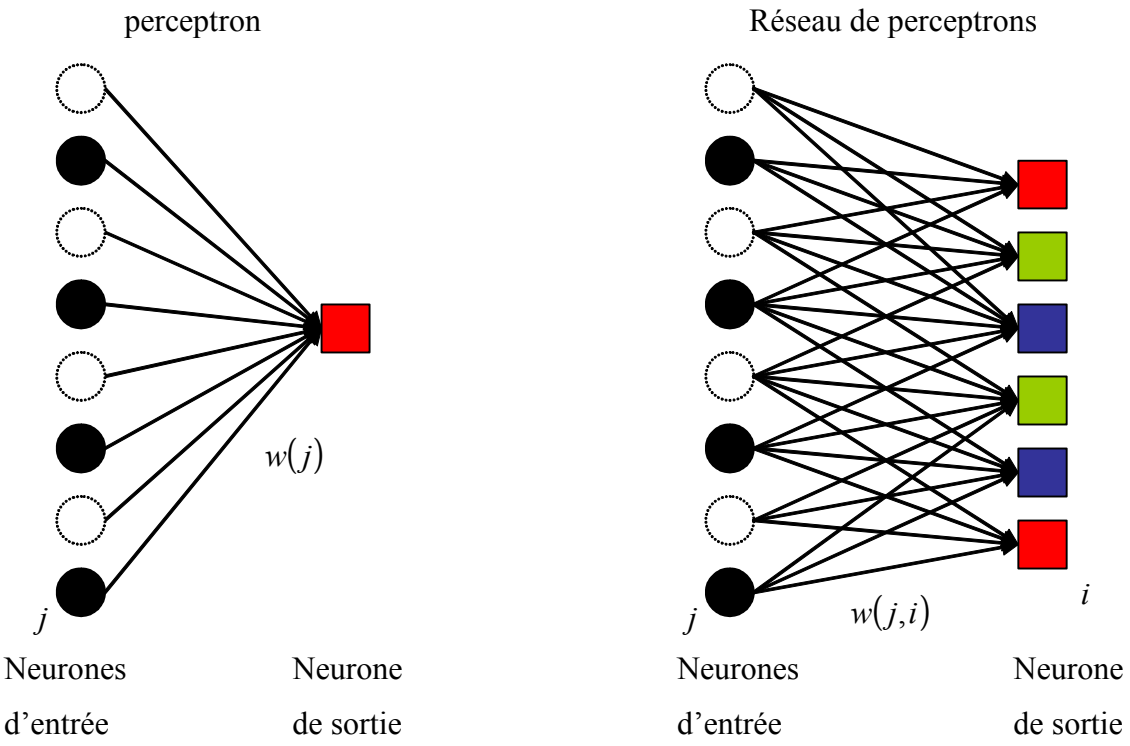


Figure 16. Schémas d'un perceptron et d'un réseau de perceptrons

On distingue deux architectures en réseaux multicouches : les réseaux de perceptrons multicouches et les réseau récurrents. Les réseaux de perceptrons multicouches (Figure 17) sont constitués d'une ou plusieurs couches superposées de neurones cachés. Chaque neurone est lié à des unités des couches voisines et à aucune unité de la même couche. Il n'existe pas non plus de lien vers les couches précédentes ni de connexion qui saute une ou plusieurs couches.

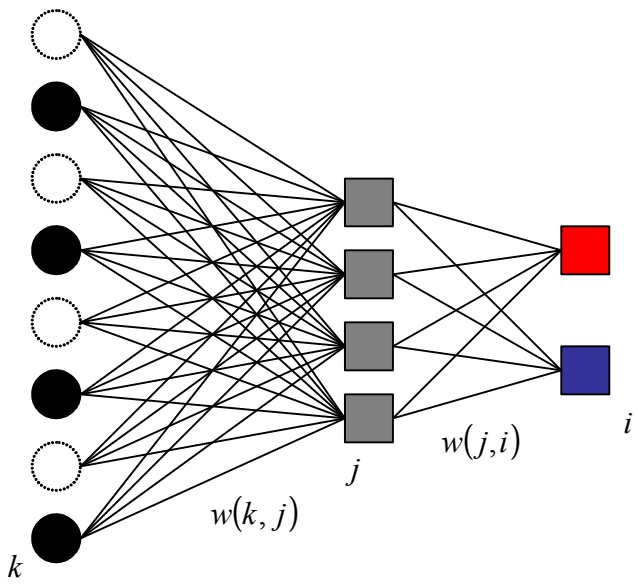


Figure 17. Figure Schéma d'un réseau de perceptrons multicouches.

A l'inverse, les réseaux récurrents sont des réseaux multicouches qui présentent davantage de connectivité. Les réseaux Hopfield (Hopfield 1982; Hopfield and W. 1985) ont une connectivité complète avec des poids des liens entre les neurones symétriques et les machines de Boltzman ont une connectivité complète. Citons aussi les cartes auto organisatrice de Kohonen (Kohonen 1982). Ces cartes sont constituées de deux couches : une couche d'entrée et une couche supérieure dont les neurones sont connectés aux couches d'entrés et aux neurones voisins. Dans cette couche, les poids des connexions latérales sont proportionnels à la distance entre les neurones.

Avec une seule couche de neurones cachés suffisamment étendue, un réseau de perceptrons multicouche est capable de représenter toutes les fonctions continues. Deux couches permettent de représenter aussi les fonctions discontinues.

L'apprentissage des réseaux de perceptrons multicouches consiste à optimiser les poids entre les liens entre les neurones. L'apprentissage des réseaux de neurones est incrémental. Les exemples d'apprentissage sont utilisés de manière séquentielle et plusieurs fois jusqu'à ce que le réseau de neurones généralise le concept de classification. Lors de l'apprentissage, si le

vecteur de classification est conforme à la classification observée, le réseau reste inchangé. Par contre si une erreur est observée, il faut ajuster les poids. L'astuce consiste à minimiser le risque empirique en donnant un blâme pour l'erreur qui sera partagé entre tous le poids qui ont été mis à contribution lors de l'apprentissage. Si cette technique est aisément implémentable pour les réseaux de perceptrons, ce n'est pas le cas pour les réseaux multicouches étant donnée que chaque neurone est connecté en entrée à plusieurs poids et qu'un poids contribue à plusieurs sorties.

C'est en 1986 que le problème d'optimisation des poids des liens entre les neurrones dans les réseaux de perceptrons multicouches a été résolu par la découverte de l'algorithme de rétropropagation. Cet algorithme permet de propager l'erreur obtenue à une unité de sortie à travers le réseau par descente d'un gradient d'erreur dans le sens inverse de la propagation des activations (Rumelhart, Hinton et al. 1986). Seule la fonction d'activation sigmoïde est utilisable par cet algorithme. En effet une des étapes de la rétropropagation nécessite de dériver la fonction d'activation, ce qui ne permet pas d'utiliser les fonctions « seuil » et « signe » qui ne sont pas dérivables. Cet algorithme nécessite également de paramétrer la valeur du taux d'apprentissage qui est comprise entre 0 et 1. L'algorithme de rétropropagation a été adapté pour l'apprentissage de réseaux récurrents.

Comme toute recherche en descente graduelle, l'algorithme de rétropropagation possèdent des problèmes concernant l'efficacité et la convergence. En particulier, il fait converger lentement le risque empirique de classification et est sensible aux minima locaux. Un sous-ensemble de test permet de tester le risque de surapprentissage du réseau de neurones. Plusieurs techniques permettent de déterminer le minimum global : répéter l'apprentissage en initialisant différemment les poids, bruiter les données d'apprentissage ou ajouter une méthode de recuit simulé (*simulated annealing approach*).

Une autre difficulté réside dans le choix de l'architecture du réseau de neurones (nombre de neurones, nombre de couches cachées et nombre de connexions). La méthode la plus simple consiste à tester de façon empirique différentes architectures. Une autre solution consiste à utiliser des algorithmes capables d'étoffer une architecture de réseau minimale ou à l'inverse de démarrer à partir d'un réseau de neurones grand et de diminuer sa complexité. Ce problème peut également être posé comme un problème de recherche dans un espace des possibles. Cette approche a débouché sur de nombreux travaux basés sur des algorithmes génétiques.

Si les réseaux de neurones sont capables de généraliser efficacement les fonctions pour lesquels ils ont été réglés, leur défaut majeur est leur manque de transparence. Ils permettent de modéliser des problèmes de classifications complexes mais la théorie d'apprentissage qui aboutit au modèle de classification est difficilement compréhensible. En effet, contrairement à d'autres algorithmes de classification tels que les arbres de décision, les réseaux de neurones sont comparables à des boîtes noires dont les sorties sont impossibles à expliquer.

7.3. La sélection des attributs

L'algorithme d'apprentissage idéal serait capable d'obtenir une bonne généralisation sans qu'il y ait besoin de sélectionner les attributs. En réalité, l'efficacité des algorithmes d'apprentissage par induction diminue lorsqu'ils sont en présence d'un trop grand nombre d'attributs qui ne sont pas pertinents. La sélection des attributs est donc une étape importante qui peut apporter un gain pour la performance du classifieur. Elle effectue le remaniement des données en entrée du classifieur.

La sélection des attributs est une étape préalable à l'apprentissage. Elle consiste à chercher le sous-ensemble d'attributs qui contient les informations nécessaires à l'algorithme d'apprentissage. De façon générale, le meilleur sous-ensemble d'attributs est celui qui permet d'obtenir au moins la meilleure performance de classification que l'ensemble complet d'attributs. En effet, il est fréquent de se retrouver avec beaucoup trop d'attributs dont la plupart ne sont pas pertinents ou sont redondants. Dans ce dernier cas, un attribut redondant n'apporte aucun gain à la généralisation du modèle de classification s'il est ajouté aux autres attributs. Les attributs les plus pertinents sont les plus discriminants, il est donc utile de définir un degré de pertinence de l'attribut dans le but d'améliorer l'apprentissage des classes. Si les algorithmes de classification sont capables de retrouver les attributs les plus pertinents parmi un ensemble redondant d'attributs, l'étape de sélection des attributs améliore en pratique leur performance. Il faut aussi estimer le coût des attributs non pertinents sur la complexité de l'algorithme.

L'étape de sélection peut être accompagnée de la discrétisation des attributs numériques. Elle intervient principalement lorsque l'algorithme de classification ne peut prendre que des attributs catégoriels en entrée.

L'ordre des attributs peut aussi influencer la performance du classifieur. Si ceux ci sont rangés par ordre décroissant de la pertinence, il est possible d'éliminer les attributs non pertinents en ne retenant que les attributs qui ont une pertinence égale ou supérieure à un seuil de pertinence donné.

Lors de la sélection des attributs, il faut faire la différence entre un sous-ensemble d'attributs qui permet d'obtenir le meilleur modèle et un sous-ensemble qui possède les attributs les mieux corrélés individuellement aux classes.

Il est nécessaire de déterminer la pertinence des attributs par une mesure de qualité. Les attributs qui possèdent peu d'informations discriminantes les uns par rapport aux autres ont alors des valeurs de qualité basses et proches. Il existe différentes manières de calculer la qualité des attributs : soit la mesure de qualité d'un attribut se fait indépendamment des autres soit cette mesure se fait en fonction des autres attributs. Par exemple, la mesure de qualité de l'attribut peut être relative à la contribution de l'attribut à la performance du classifieur. Une solution consiste à comparer la performance du classifieur avec tous les attributs pris en compte par rapport à la performance sans l'attribut d'intérêt. Un attribut idéal est capable de séparer tous les exemples de classes différentes et avoir les mêmes valeurs pour tous les exemples d'une même classe. Il possèderait alors une valeur de qualité élevée.

Il existe de nombreux avantages à la sélection automatique des attributs. En effet il est souvent difficile de prédire manuellement les effets des différents attributs sur la performance de l'apprentissage. De plus, la sélection manuelle des attributs est un travail qui peut s'avérer fastidieux et pas forcément plus efficace pour la performance du classifieur. La méthode automatique laisse toute liberté de choix des meilleurs attributs au classifieur. Il est aussi plus facile d'ajouter un attribut à la volée à un ensemble de données. Si les propriétés de l'ensemble de données se modifient, les attributs peuvent s'adapter. Les attributs peuvent aussi s'adapter à la taille de l'ensemble des exemples. Il faut aussi savoir gérer les attributs les moins pertinents s'il apparaît de nouveaux attributs plus bruyants. Il faut aussi reconnaître les attributs contre-productifs ou « deceptives attributes » que l'algorithme va considérer comme

utiles à la classification mais qui en fait biaisent la généralisation de l'hypothèse de classification au lieu de l'aider.

La sélection des attributs peut se faire par des approches de transformation linéaire des espaces d'entrées. Une méthode couramment utilisée est l'analyse en composantes principales. C'est une méthode non supervisée qui permet d'identifier un sous-ensemble d'attribut décrivant les données en minimisant la perte d'information. La méthode d'analyse en composantes principales communes prend en compte les classes des exemples. Il existe aussi des approches de sélection des attributs non linéaires.

La sélection des attributs peut se faire indépendamment de l'algorithme d'apprentissage ou non. Dans le cas où la sélection est indépendante de l'algorithme d'apprentissage, il faut faire l'hypothèse que la sélection est *a priori* correcte en entrée de l'algorithme d'apprentissage. Les méthodes de sélection des attributs qui utilisent les algorithmes d'apprentissage ont une complexité plus élevée que les méthodes précédentes mais sont aussi plus performantes. Les premières méthodes sont appelées méthodes de filtrage (filter method) et les deuxièmes des méthodes d'enveloppage (wrapper method). Une méthode usuelle de filtrage se base sur l'algorithme RELIEF (Figure 18) (Kira and Rendell 1992). Il existe aussi de nombreuses adaptations de cet algorithme.

R : ensemble des exemples (représentés par un vecteur d'attributs avec son étiquette).
A : attribut.
m : nombre d'instances de R sélectionnées aléatoirement.
Entrée : R , m
Sortie : vecteur W d'estimation de la qualité des attributs
Initialisation de tous les poids W[A] = 0

Pour i=1 à m **faire**
choisir au hasard une instance R ;
H l'instance la plus proche de R et de la même classe;
M l'instance la plus proche de R et de classe différente ;
Pour A =1 à tous les attributs faire
$W[A] = W[A] - \text{dist } A_{RH} / m + \text{dist } A_{RM} / m$
Fin
Fin

Figure 18. Algorithme Relief général.

La méthode d'enveloppage (Figure 19) prend en compte les propriétés de l'ensemble d'apprentissage et suit les biais du système de classification pendant la recherche d'évaluation des poids sur les attributs. Dans cette approche le système d'apprentissage par induction est utilisé par la fonction de mesure de la qualité de l'attribut. La fonction d'évaluation des attributs évalue l'ensemble des attributs en utilisant l'algorithme de classification comme une boîte noire. Elle peut par exemple tenir compte de la performance de l'algorithme par validation croisée. Une mesure de qualité est alors retournée à la fonction de recherche qui décide si le meilleur sous-ensemble d'attributs a été obtenu. A la fin de l'étape de sélection, l'apprentissage est effectué à partir du sous-ensemble d'attributs sélectionnés.

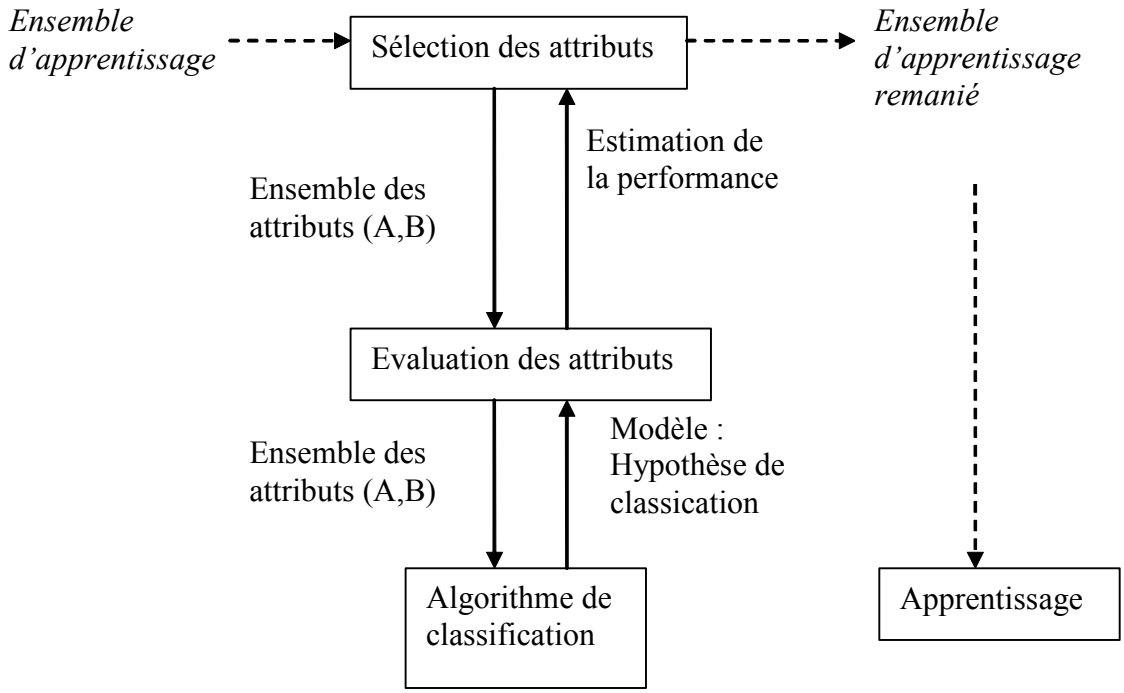


Figure 19. Principe général de la méthode de sélection des attributs par enveloppage

Les travaux en fouille de données axés sur la découverte des relations entre les attributs ont permis de développer des algorithmes qui sont aussi utilisés en sélection des attributs en classification. En effet, ces algorithmes peuvent mettre en évidence qu'un sous-ensemble de données peut être caractérisé par l'association de plusieurs attributs. En particulier, l'algorithme « A priori » (Agrawal, Imielinski et al. 1993; Agrawal and Srikant 1994) (Figure 20) permet de retourner les « itemsets fréquents » présents au sein d'un ensemble de données.

Initialisation : L_1 ensemble des 1-itemsets fréquents dans l'ensemble de données

Tant que le test d'arrêt n'est pas satisfait **faire**

1 Utiliser L_{k-1} pour produire l'ensemble C_k contenant les k-itemsets candidats.

2 Ne conserver que les itemsets de C_k qui sont fréquents dans l'ensemble de données : ils forment l'ensemble L_k.

Fin tant que

Figure 20. Algorithme A priori.

Le terme « item » est donné pour un attribut binaire qui a la valeur 1 pour un exemple donné. Un itemset consiste alors en un ensemble de tous les attributs binaires valant 1 pour un exemple donné. Une association se définit comme une implication disant que deux itemsets sont vrais ensembles pour un nombre suffisant d'exemples. La couverture de l'association est calculée comme le nombre d'itemsets, divisé par le nombre total d'exemples. Quand une couverture est supérieure à un seuil fixé à l'avance par l'utilisateur, on dit que l'itemset constitué par cette intersection est *fréquent*. Le seuil *MinSup* est le support minimal exigé par l'utilisateur. L'algorithme A priori permet de trouver tous les itemsets fréquents qui ont un support plus grand que *MinSup*.

7.4. Tests de performance de l'apprentissage

Le test des méthodes d'apprentissage peut se faire en divisant l'ensemble des données en un ensemble de données d'apprentissage et un ensemble de données de test. Les deux ensembles sont indépendants et l'ensemble des classes C est réparti de manière uniforme entre eux.

Le test de la performance de la classification de l'échantillon de test (ou de l'échantillon d'apprentissage lui-même) est visualisé dans une matrice de confusion $C \times C$ qui permet de visualiser les exemples correctement prédits de ceux qui ont été mal classés. Lorsque les

échantillons d'apprentissage et de test sont grands, il est possible de déterminer l'intervalle de confiance de l'estimation de l'hypothèse à valider qui dépend du nombre d'exemples et de la valeur de risque empirique réel de l'hypothèse. D'autres techniques doivent être utilisées dans les cas où l'ensemble de données est trop petit.

7.4.1. Estimation de la performance par validation croisée

Cette technique consiste à diviser l'ensemble des données en n sous-ensembles indépendants et de même taille. Un sous-ensemble est choisi pour être l'échantillon de test. Les $n-1$ autres échantillons sont rassemblés pour former l'ensemble d'apprentissage. Le risque empirique réel est alors calculé sur l'échantillon de test. Cette technique est répétée n fois et l'erreur moyenne finale est donnée par la moyenne des erreurs mesurées. Si les modèles obtenus donnent des valeurs d'erreurs très variables, il existe un risque que l'hypothèse de classification ne soit pas robuste, à cause de la nature des données, du choix ou de la configuration de l'algorithme de d'apprentissage utilisé.

La méthode du *leave-one-out* correspond à une validation croisée qui prend à chaque fois un seul exemple pour le test du modèle qui est appris à partir des exemples qui restent.

Deux autres techniques dérivées sont adaptés aux petits ensembles d'exemples : la méthode du *bootstrap* (k tirages avec remise) et la méthode du *jackknife*.

Dans le bootstrap (Efron 1979) on retire n fois un exemple qui est placé dans l'ensemble bootstrap. Contrairemnet aux méthodes de validations croisées et de *leave-one-out* le tirage s'effectue avec remise. L'ensemble bootstrap est utilisé pour l'apprentissage. Dans une première phase, les exemples non retenus servent au test. Les erreurs de classification sont stockées dans la variable p_1. Dans un deuxième temps l'ensemble complet des exemples est testé et le nombre d'erreurs de classification est stocké dans la variable p_2. L'ensemble des opérations est répété k fois pour obtenir les moyennes des erreurs : P_1 et P_2. Cette méthode est adaptée aux ensembles qui ne sontiennet qu'un petit nombre d'exemples. Cet inconvénient est compensé par un nombre élevé de répétitions k (supérieur à 100).

Dans le cas général d'un tirage équiprobable avec remise, un exemple a une probabilté d'être choisi à chaque tirage égale à $\frac{1}{n}$ et une probabilité $1-\frac{1}{n}$ de ne pas l'être. La probabiblité qu'un exemple ne soit jamais choisi à l'issue de l'étape de création de l'ensemble d'apprentissage équivaut à : $\left(1-\frac{1}{n}\right)^n \approx e^{-1} = 0.632$; La constante e etant la base du logarithme népérien.

L'estimation du risque réel $R(h)$ se calcule par la relation suivante :

$$R(h) = 0.632 * P_1 + 0.368 * P_2$$

La méthode *jackknife* consiste à énumérer tous les sous-échantillons de l'ensemble des exemples en retirant un seul exemple. Chaque sous-ensemble d'exemples non retirés sert à l'apprentissage et l'exemple retiré sert de test. Le but est de connaître l'influence d'un seul exemple sur l'hypothèse obtenue par classification.

7.4.2. Courbe ROC

Les méthodes décrites précédemment évaluent le risque d'erreurs dans son ensemble par l'estimation du risque réel. D'autres techniques permettent de prendre en compte les taux de faux positifs et de faux négatifs. La courbe ROC « Receiver Operating Characteristic » est utile lorsque le risque d'une mauvaise classification n'est pas symétrique entre les classes. Par exemple, s'il s'agit d'aider à la décision d'un diagnostic médical, le classifieur doit éviter de classer un individu malade dans la classe des individus sains même si en contrepartie il risque de diagnostiquer plus fréquemment la maladie chez un individu sain.

Soit un classifieur et une fonction de décision f tels que la valeur de $f(x)$ détermine la classe prédite pour tous les exemples (x, u_i) d'un ensemble de données. Il est possible d'établir le graphique de probabilité d'appartenance à la classe u_i en fonction de la valeur de $f(x)$. En pratique, la première étape nécessaire à l'établissement de la courbe ROC consiste à

trier les exemples de l'ensemble de test dans l'ordre décroissant de la valeur de $f(x)$. Ce tri permet d'obtenir plus facilement le sous-ensemble des exemples dont les valeurs $f(x)$ sont supérieures ou égales au seuil d'intérêt.

La courbe ROC prend en ordonnée la proportion de classification de vrais positifs et en abscisse la proportion de classification de faux positifs (Figure 21).

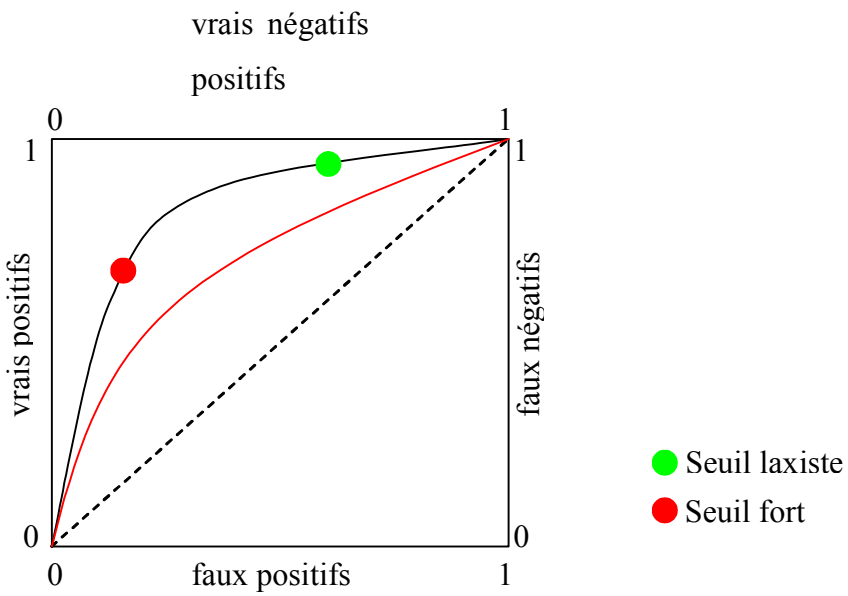


Figure 21. Exemple d'une courbe ROC

Il s'agit de rapporter chaque couple de valeurs pour tous les seuils de séparation $f(x)$. Les valeurs proches de l'origine correspondent aux seuil de critères de décision les plus stringents puisque les valeurs de $f(x)$ sont les plus élevées. Cette région correspond aux modèles qui ne classifient que les exemples de la classe u_i les plus évidents de l'ensemble de données mais aucun faux positif. Plus on s'éloigne de l'origine et plus les seuils permettent d'obtenir des modèles qui prédisent davantage de vrais positifs mais avec une sélectivité décroissante (apparition des faux positifs).

Graphiquement, plus la courbe s'élève, plus le rapport des vrais positifs par rapport aux faux positifs augmente ce qui indique que le classifieur devient plus pertinent. Par contre, si la courbe que l'on obtient est une droite cela indique le taux de classification sera toujours de 50% pour tous les seuils. S'il existe plusieurs modèles basés sur des critères de décision

différents, ils peuvent être comparés par leurs courbes ROC respectives dans les mêmes conditions d'apprentissage.

L'indice de précision (Figure 22) correspond à la proportion d'exemples correctement prédits par l'hypothèse (rapport entre le nombre d'exemples correctement prédit / nombre d'exemples corrects dans l'ensemble de données) (VP/(VP+FN)). Le rappel (couverture) (Figure 22) correspond à la proportion des exemples correctement classés par l'hypothèse (nombre d'exemples prédits correctement / nombre d'exemples prédits) (VP/(VP+FP)). Ces deux indices permettent d'établir la courbe lissée de précision et de rappel. L'utilisateur peut observer la proportion d'exemples correctement prédits pour le taux de rappel qu'il cherche à atteindre. Le score F (Figure 22) correspond à la moyenne pondérée de la précision et de la couverture.

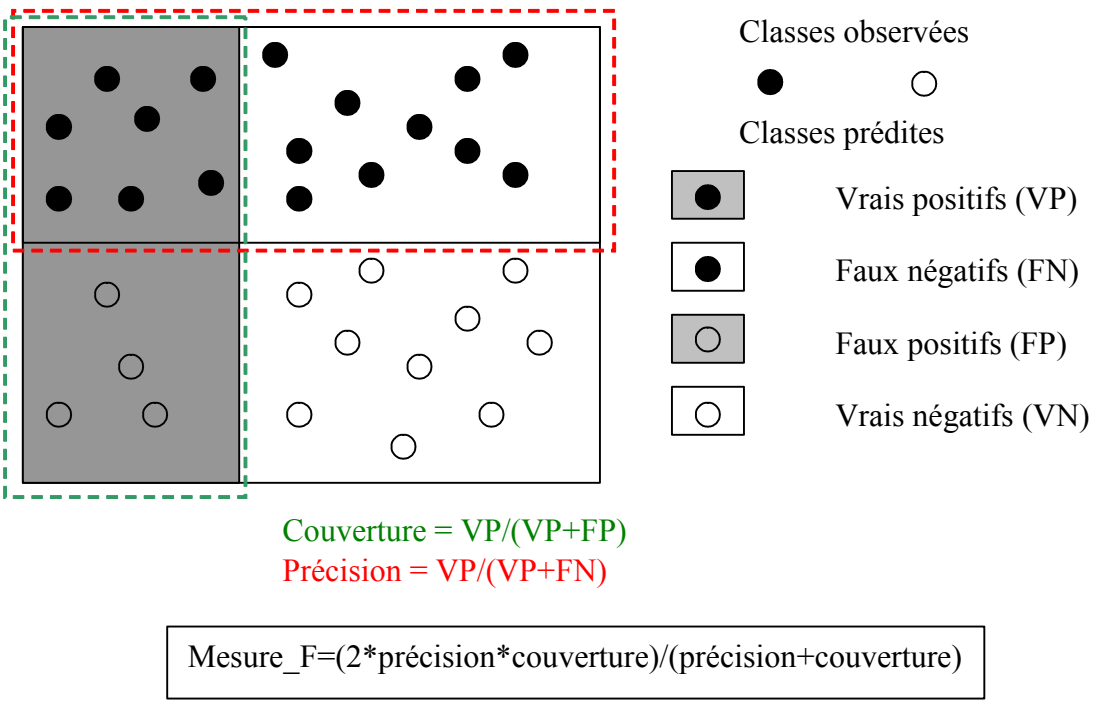


Figure 22. Illustration de la couverture, de la précision et du score F

Chapitre 8. Les algorithmes génétiques

8.1. Généralités sur les algorithmes d'apprentissage par évolution simulée.

Les algorithmes par évolution simulée sont une catégorie d'algorithmes d'apprentissage par exploration. Ils sont spécialisés dans l'exploration des espaces d'hypothèses de grandes dimensions. Si leur but est d'explorer l'espace des hypothèses H à la recherche de la meilleure description de l'espace des exemples χ, leur principe d'exploration se base sur l'analogie avec la théorie de la sélection naturelle. Ils font appel à un espace intermédiaire appelé espace génotypique. L'espace des hypothèses correspond à un espace phénotypique. Chaque génome de l'espace génotypique est couplé à une hypothèse. Une pression de sélection est appliquée sur les membres de l'espace génotypique par l'espace des hypothèses de manière à faire converger les génomes vers le génome optimal. Ce génome représente l'hypothèse qui décrit le mieux l'espace des exemples.

Les algorithmes génétiques ont été développés et popularisés par John Holland (Holland 1975) qui a proposé une analyse théorique appelée « théorie des schémas ». Le principe général de ces algorithmes est à la base de trois autres spécialités de l'apprentissage par évolution simulée : les stratégies d'évolutions (représentation par automate pour la résolution de problèmes d'optimisation de structures mécaniques), la programmation génétique (représentation de programmes sous forme d'arbres) et la coévolution (la fonction de pression de sélection de l'espace des hypothèses est elle-même mise en compétition). Ce chapitre aborde la spécialité des algorithmes génétiques.

8.2. Principe des algorithmes génétiques et définitions.

Les algorithmes génétiques sont conçus sur les principes de l'évolution naturelle. Ils font appel aux notions de sélection et de variabilité. L'espace génotypique est représenté par une

population Pg composée par un ensemble d'individus I . Le nombre d'individus est constant et déterminé de façon empirique à l'initialisation. Un individu possède un génome et chaque génome est caractérisé par un vecteur appelé chromosome. La dimension de ce vecteur est fixe. Chaque dimension de ce vecteur est appelée un gène. Un gène peut prendre des valeurs binaires ou discrètes. Sachant qu'un génome représente une hypothèse dans l'espace H , le choix de la dimension du vecteur influence la taille de l'espace des hypothèses. Par exemple, le choix d'un chromosome composé de 10 gènes binaires revient à explorer un espace H de taille 2^{10} . L'effort effectué sur la représentation de l'espace génotypique tient donc un rôle très important dans l'efficacité des algorithmes génétiques.

L'algorithme génétique général (Figure 23) débute par l'initialisation des individus de la population de départ $Pg(0)$. Les valeurs des gènes sont choisies de manière aléatoire pour obtenir les génomes les plus variés. Si des connaissances à priori sont disponibles, il est possible de biaiser la valeur initiale des gènes. Il faut cependant éviter que ce choix ne soit pas trop contraignant sur la diversité génétique et ne restreigne pas trop l'espace des hypothèses dès le début. En théorie, l'hypothèse optimale est obtenue en temps illimité. Il est donc nécessaire de déterminer le critère d'arrêt qui satisfait l'exploration de l'espace des hypothèses dans un temps limité. Une solution consiste à stopper l'exécution de l'algorithme génétique à partir du moment où la population ne varie plus entre deux générations successives. Une autre solution consiste à choisir un nombre maximal de générations de population en estimant que la convergence vers l'hypothèse optimale est atteinte au bout de ces itérations.

Initialisation de la population $Pg(0)$.

Tant que le critère d'arrêt n'est pas obtenu :

- Calcul de l'adaptation sur chaque individu de la population courante $Pg(t)$ en mesurant chaque hypothèse H associée à chaque génome g .

- Sélection des individus les mieux adaptés de la population courante $Pg(t)$ qui deviendront les parents de la population $Pg(t+1)$.

- Application de la variabilité génétique de la population par des opérations de mutations et de croisements des génomes de la population $Pg(t)$.

- Remplacement des individus de la population $Pg(t)$ par ceux de $Pg(t+1)$.

Fin tant que

Figure 23. Algorithme génétique général.

Les algorithmes génétiques sont constitués de quatre étapes : la détermination de la valeur d'adaptation de chaque individu dans la population, la sélection des descendants, le renouvellement des générations et les opérations de remaniement des génomes. Les trois premières étapes sont nécessaires à la sélection des individus. La dernière étape est celle qui permet de conserver la variabilité des gènes indispensable à l'exploration de tout l'espace des hypothèses (Figure 23).

8.3. La sélection des individus.

Cette partie présente les stratégies qui permettent d'appliquer une sélection naturelle sur les individus.

8.3.1. La pression sélective

L'intérêt d'effectuer une pression sélective sur les individus permet de sélectionner ceux qui expriment les caractères les plus favorables à leur adaptation. Ces caractères sont liés aux gènes ou groupes de gènes présents dans les génomes. Le but des algorithmes génétiques est d'appliquer une pression sélective sur les hypothèses liées aux génomes. Ils mettent en évidence les meilleurs caractères et les conservent au sein de la population pour obtenir les individus possédant l'ensemble des meilleures combinaisons de caractères par rapport à la contrainte de sélection. La pression de sélection est effectuée par l'appel d'une fonction de performance ou d'adaptation « fitness function ». Cette fonction d'adaptation permet d'effectuer le passage de l'espace génotypique vers l'espace phénotypique H pour évaluer la performance de l'hypothèse soutenue par chaque génome. Cette étape d'évaluation nécessite de trouver la fonction d'adaptation la plus discriminante mais la moins coûteuse aussi. Une technique peut consister à utiliser une évaluation approchée et rapide lors des premières générations et plus fine sur les générations suivantes. Le réglage de la force de la pression d'adaptation est important pour éviter la convergence trop rapide due à une pression trop forte ou à l'inverse une convergence trop lente due à une pression trop faible. Cette étape permet d'obtenir une note de qualité pour chaque individu. Il faut ensuite déterminer le sous-ensemble des individus qui seront choisis pour obtenir la génération suivante.

8.3.2. Trois méthodes de sélections

Lorsque l'évaluation de chaque individu est effectuée dans l'espace des hypothèses, il s'agit alors de sélectionner les individus dans l'espace génotypique les plus aptes à la création de la génération suivante. Cette partie présente les principes de trois méthodes couramment utilisées.

Performance (roue de la fortune) : La probabilité de l'individu de participer à l'élaboration de la génération suivante est donnée par sa valeur d'adaptation. Si la dispersion des valeurs de performance est grande alors quelques individus possèdent des performances très élevées. Ce

mode de sélection peut se trouver biaisé et être entraîné vers un maximum local dans l'espace des hypothèses.

Rang (tri des individus) : les individus sont triés en fonction de leur valeur d'adaptation. La probabilité de sélection est alors fonction de leur numéro d'ordre. Ce mode de sélection régule mieux la pression de sélection dans les cas où les variances des performances sont trop élevées ou trop faibles.

Tournoi : cette méthode permet d'accélérer l'étape de sélection. Il s'agit de sélectionner n individus parmi la population totale en effectuant des sélections en parallèle dans n sous-ensembles d'individus disjoints. Les individus retenus sont utilisés pour créer la génération suivante lors de l'étape de remplacement. La méthode des tournois à l'avantage d'être moins sensible aux erreurs de la fonction d'adaptation que les deux autres méthodes.

A l'issue de la sélection, les individus restants sont ceux sur lesquels s'appliquent, dans la plupart des techniques d'algorithmes génétiques, les opérations de remaniement des génomes qui sont la mutation et le croisement.

8.4. Le remplacement des populations

Après avoir obtenu le sous-ensemble des individus sélectionnés, l'étape suivante consiste à renouveler la génération. Il faut remarquer que les étapes de sélection et de remplacement sont indissociables dans la théorie de la sélection naturelle. De ce point de vue, les techniques employées par les algorithmes génétiques sont des techniques de « sélection/remplacement ».

La méthode « d'algorithme génétique générationnel » consiste à effectuer des opérations de remaniement sur les génomes des individus du sous-ensemble de sélection. Les individus obtenus remplacent ensuite leurs propres ascendants dans la génération suivante.

La technique de « génération par remplacement local », sélectionne un sous-ensemble d'individus pour effectuer le remaniement des génomes. A la différence de la première méthode, les individus obtenus vont prendre la place d'un autre sous-ensemble de la population, constitué d'individus peu performants par exemple. Cette technique permet d'implémenter l'algorithme de manière asynchrone.

Les techniques de « stratégies d'évolution » sont des méthodes de remplacements qui ne tiennent pas compte de l'étape de sélection Darwinienne. La stratégie d'évolution (μ,λ) considère que le sous-ensemble de sélection correspond à la population totale. Elle consiste à remplacer totalement l'ensemble de la population à chaque génération. Les descendants sont créés en manipulant tous les génomes de l'ensemble des parents. La nouvelle population est constituée par les meilleurs individus fils. Cependant cette méthode à l'inconvénient majeur de perdre les meilleurs individus des générations précédentes lors de chaque remplacement. Une autre technique $(\mu+\lambda)$ consiste à conserver le sous-ensemble des μ meilleurs parents. Les génomes des ascendants non sélectionnés subissent ensuite des remaniements pour obtenir λ individus qui viennent compléter la population des descendants.

8.5. Les opérateurs sur les individus.

Après sélection sur les individus, l'étape suivante consiste à modifier les génomes des individus pour obtenir des descendants. Ces méthodes sont définies dans l'espace génotypique et sont exécutées de manière stochastique sur les génomes. Si la panoplie des opérateurs de manipulation des génomes peut être infinie, il faut en retenir deux principaux : la mutation et le croisement.

8.5.1. La mutation

L'opérateur de mutation fait référence du principe de la mutation d'un gène biologique. Cet opérateur est unaire et s'applique sur un génome. Il effectue des changements aléatoires sur chacun des gènes en suivant une loi de probabilité de mutation et retourne le génome muté. La fonction de probabilité de mutation est souvent faible pour ne permettre la mutation que de quelques génomes sur une même génération. Elle limite aussi le nombre de mutations possibles simultanément pour un même génome. Cependant la mutation doit être correctement estimée car c'est l'opérateur qui permet le brassage des gènes et donc l'exploration complète de l'ensemble de hypothèses.

8.5.2. Le croisement

L'opérateur de croisement est la transposition d'un phénomène biologique appelé recombinaison génétique. Ce phénomène est fondamental dans la théorie de l'évolution des espèces. Il s'agit de l'échange d'une région sous-génomique commune entre les génomes de deux individus différents.

Cet opérateur est binaire ou n-aire. La fonction de recombinaison est stochastique, ce qui implique que si deux génomes sont sélectionnés par cet opérateur, ils ne seront pas forcément recombinés.

L'opérateur le plus simple sélectionne aléatoirement un point de croisement sur deux génomes pris en entrée et intervertit les régions qui sont en amont du point de croisement pour obtenir deux individus fils.

Une variante plus élaborée, le croisement multipoint, sélectionne au hasard deux points de croisement sur les deux génomes et fait l'échange du segment situé entre les deux cassures.

La technique de croisement uniforme utilise une probabilité souvent proche de 0,5 pour croiser chaque gène des deux génomes.

On peut considérer que deux individus avec une performance remarquable possèdent dans leurs génomes respectifs une fraction de gènes favorables, où primitives, mais qui se situent dans des zones différentes de l'espace des hypothèses, tels deux maximums locaux. L'opérateur de croisement est intéressant puisqu'il donne la possibilité d'obtenir l'ensemble de ces primitives dans le génome d'un des individus fils. Ce dernier est potentiellement mieux adapté que ces parents de telle sorte que l'hypothèse portée par son génome va s'approcher davantage de l'hypothèse optimale.

8.5.3. Paramétrages des effets de mutation et de recombinaison

L'opérateur de croisement permet de tester et de sélectionner des blocs de gènes de plus en plus performants. Ces blocs se perpétuent et incluent d'autres primitives au fur et à mesure des générations. En conséquence, le croisement permet de faire converger les génomes. Mais il accentue la perte de la variabilité génétique de la population et réduit le champ d'exploration de l'espace des hypothèses. En revanche les mutations ont pour effet de favoriser le brassage des gènes et empêche la convergence vers des optimaux locaux. Il est donc nécessaire d'adapter l'importance de ces opérateurs en fonction du problème étudié. Le croisement peut être utilisé lorsqu'il est possible de converger de façon relativement continue vers l'hypothèse optimale. A l'inverse il faut favoriser les mutations s'il est nécessaire d'explorer totalement l'espace des hypothèses, par exemple si l'hypothèse à optimiser est discontinue.

8.6. Accélération de la convergence : méthode du keep best reproduction

La méthode du keep best reproduction (Wiese, Scott et al. 1998; Wiese and Goodwin 1999) permet d'obtenir les meilleurs individus plus rapidement. Le principe de cette méthode repose sur l'observation suivante ; Les individus parents possédant un très bon génotype sont sélectionnés pour l'étape de remaniement. Cependant, dans le cas des algorithmes génétiques standards, les individus de la nouvelle génération peuvent avoir des génotypes remaniés moins adaptés à la pression de sélection que ceux de leurs parents. Le gain apporté par les génotypes parentaux est alors perdu.

La méthode du keep best reproduction évalue d'abord la qualité des génotypes des individus par rapport à ceux de leurs parents. Les plus mauvais individus fils sont alors remplacés par les meilleurs parents. De cette manière le sous-ensemble des individus fils entretient la variabilité génétique tandis que la meilleure information génétique présente chez les parents est conservée. Les individus subissent ensuite des mutations puis sont additionnés à la nouvelle population. Ainsi, cette méthode permet d'accélérer la convergence de l'algorithme génétique tout en conservant sa performance.

Chapitre 9. La classification supervisée, les SVM : Concepts & état de l'art en bioinformatique.

9.1. Introduction

La technique de classification des SVM « Support Vector Machines » ou « Séparateurs à Vaste Marge » fait partie des classifieurs issus du domaine de l'apprentissage statistique. Cette technique de classification est récente. Elle a été décrite en 1995 par Vapnik (Vapnik 1995). Les SVM sont maintenant devenus très populaires et ont montré en pratique leur intérêt et leur performance dans des problèmes de classification très variés. Les concepts impliqués dans la théorie des SVM sont issus de l'apprentissage par optimisation et de l'apprentissage par induction de surfaces séparatrices linéaires, dont est aussi issu la théorie du perceptron.

Dans une première partie nous décrivons les concepts statistiques fondamentaux sur lesquels se base la classification SVM. La suite de ce chapitre est consacrée au détail des trois principes fondamentaux des SVM : la marge, la formulation duale et la fonction noyau. Enfin nous décrivons les résultats déjà obtenus par l'utilisation des SVM en bioinformatique pour la recherche d'homologies de séquences éloignées.

9.2. Les fondements statistiques des SVM : le principe ERM « Empirical Risk Minimization ».

La théorie de classification SVM peut s'expliquer dans le cas de la classification d'un ensemble de données biclasses. Du point de vue statistique la méthode des SVM est dite discriminante dans le sens où il s'agit de trouver une fonction de décision $f : \chi \to \{-1;+1\}$ qui associe à chaque exemple une étiquette positive ou négative. Une part de la théorie des SVM repose sur les concepts théoriques de l'induction en particulier le principe de minimisation du risque empirique.

Soit la fonction de décision f qui vérifie l'hypothèse de classification h. La fonction f appartient à une classe de fonctions F, par exemple la classe des fonctions linéaires.

En théorie de l'induction, la meilleure fonction de décision pour résoudre le problème de classification possède un risque réel R de mauvaise classification minimum. Le risque réel est aussi appelé risque théorique ou risque structural. Il se calcule sur les exemples d'un ensemble de test.

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

La valeur de $f(x)$ correspond à la classe prédite de l'exemple x et u à la classe désirée. Le risque R dépend d'une fonction de coût L et de la distribution de probabilité $\mathrm{P}(x, u)$. La distribution de probabilité est inconnue et ne peut pas être calculée. Il est alors nécessaire de choisir le principe inductif qui permet de décrire la fonction de décision minimisant le risque réel en fonction des observations sur l'ensemble d'apprentissage. Le risque empirique représente le coût d'une mauvaise classification sur les exemples de l'ensemble d'apprentissage.

$$Remp(f) = \frac{1}{N_x} \sum_{i=1}^{N_x} L[f(x_i), u_i]$$

Le principe de minimisation du risque empirique (ERM) est défini de la manière suivante :

La fonction de décision qui s'accorde le mieux aux données d'apprentissage est celle qui permet la meilleure généralisation des concepts du problème de classification. Si la fonction de décision minimise le risque empirique alors elle minimise aussi le risque réel.

Ce principe est à la base de nombreuses techniques d'apprentissage par induction.

Le principe ERM est satisfaisant si le nombre d'exemples est grand. Si le cardinal de l'ensemble d'apprentissage tend vers l'infini, on peut faire converger le risque empirique vers le risque théorique. A l'inverse, si le cardinal de l'ensemble des exemples d'apprentissage est réduit, il existe un risque d'obtenir une fonction de décison qui représente trop exactement les exemples d'aprentissage. Ce phénomène, qui effectue un apprentissage par coeur au détriment de la qualité de la généralisation, est appelé surapprentissage ou « *overfitting* ».

Une solution consiste à réduire la classe de fonctions F pour trouver la fonction de décision f la plus simple. En effet si l'espace des hypothèses est contraint, alors le choix des hypothèses est limité. Dans ce cas, le risque empirique sur la meilleure hypothèse est sûrement proche du risque réel. Si un espace des hypothèses est riche alors il existe de fortes chances que la meilleure hypothèse y soit incluse. Cependant, un espace des hypothèses riche ne garantit pas que le risque empirique soit proche du risque réel.

L'utilisation du risque empirique n'offre pas de garantie contre le sur-apprentissage. Donc le principe ERM ne suffit pas seul à obtenir la meilleure hypothèse avec le risque réel minimum.

La partie suivante est consacrée à un deuxième principe inductif, appelé SRM, qui introduit une borne maximale du risque réel pour une hypothèse donnée en fonction du risque empirique et de la borne supérieure de complexité de la classe des hypothèses.

9.3. Dimension de Vapnik-Chervonenkis et principe SRM (Structural Risk Minimization)

La théorie de Vapnik-Chervonenkis consiste à contrôler la complexité de la classe de fonctions. On note dVC la dimension de Vapnik-Chervonenkis (Vapnik and Chervonenkis 1971). La dimension dVC mesure la richesse de l'espace des hypothèses.

Soit F un ensemble de classes de décisions fini. La dimension dVC exprime la complexité de la fonction de décision f_i par rapport à sa classe de fonctions F_i. La fonction f_i est celle qui minimise le risque empirique dans la classe des fonctions F_i.

Le contrôle de la complexité de la classe de fonction prend en compte la dimension dVC et le risque structurel. Le but de ce principe inductif est de déterminer la fonction de décision la plus simple qui minimise le risque structurel et le risque empirique. Cette stratégie consiste à trouver la classe de fonctions F_i de façon telle que la borne maximale G du risque réel soit minimisée.

$$R(f) \leq Remp(f) + G(N_x, dVC, \varepsilon)$$

$$\text{et : } P\left(\max_{f \in F} \left| R(f) - R_{emp} \right| > \varepsilon \right) < G(N_x, dVC, \varepsilon)$$

La fonction G est la borne maximale de complexité. Elle dépend du nombre d'exemples dans le jeu d'apprentissage composé de N_x données, de la valeur de la dimension dVC et de l'écart ε.

Si on prend le cas d'une classe de fonctions restreinte, dVC étant petit, la borne supérieure complexité prendra des valeurs faibles, mais R_{emp} restera grand. Inversement, si dVC est grand (F de grande complexité), la borne supérieure prend des valeurs très fortes alors que R_{emp} est petit. La borne supérieure de complexité décroît si la fonction de décision permet une bonne séparation des données de l'ensemble d'apprentissage pour une dimension dVC petite.

Le risque empirique R_{emp} et la borne de complexité ont donc des comportements opposés.

La meilleure fonction de décision sera donc celle qui offrira le meilleur compromis entre la complexité corrélée à l'explication des données et le risque empirique. Ce concept correspond au principe SRM « Structural Risk Minimization ».

La partie suivante montre l'implication des principes inductifs ERM et SRM, dans la théorie SVM.

9.4. Hyperplan, marge et espace de redescription

Considérons un ensemble d'apprentissage linéaire, séparable par un hyperplan optimal H. On choisit les fonctions de décisions de la forme $f(x) \leq w.x + w_0$.

Le vecteur w symbolise le vecteur poids normal à l'hyperplan optimal, et w_0 correspond à la valeur de biais.

La marge γ est définie comme la distance minimale entre les exemples du jeu d'apprentissage et l'hyperplan optimal.

On montre que la dimension dVC peut être bornée en fonction de la marge. La marge peut être calculée à partir du vecteur de poids w. Autrement dit, la dimension dVC, la marge γ et le poids w sont liés tels qu'en bornant la marge, on contrôle la dimension dVC.

Soit deux exemples x_1 et x_2 de classes différentes tels que :

$$w.x + w_0 = +1$$

$$\text{et } w.x + w_0 = -1.$$

La marge γ équivaut à la distance entre x_1 et x_2 perpendiculairement à l'hyperplan. Les deux exemples x_1 et x_2 sont appelés vecteurs de support « exemples critiques ». Ils se situent respectivement sur deux hyperplans H_1 et H_2 parallèles à H (Figure 24).

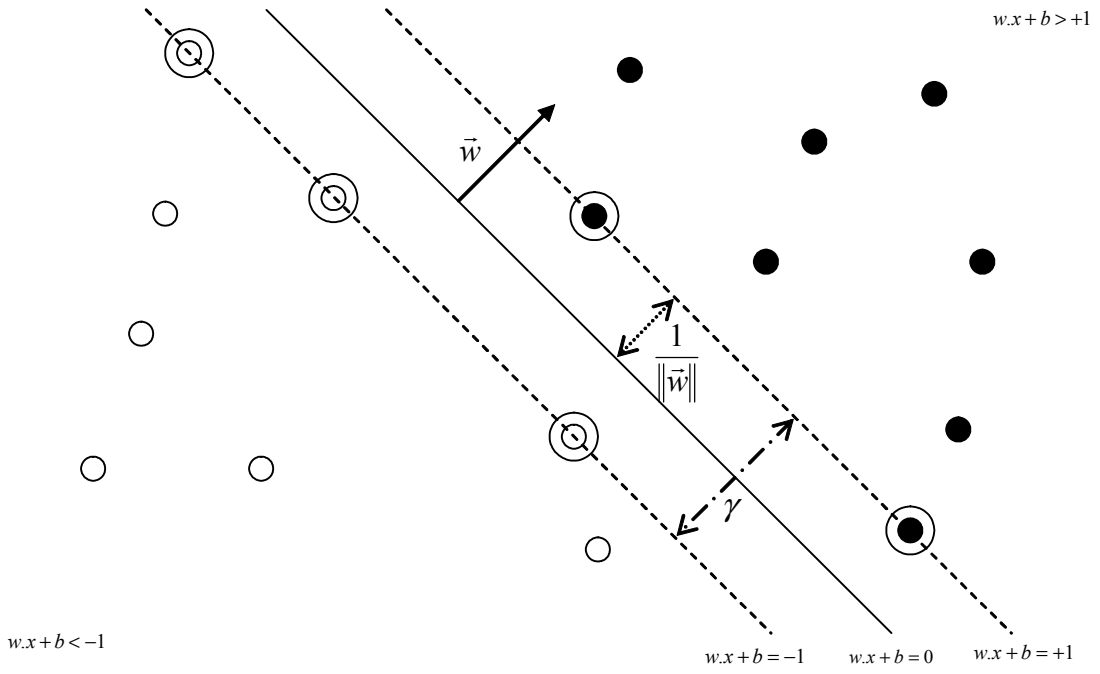


Figure 24. Séparation d'un ensemble linéaire de points par l'hyperplan optimal

L'équation de l'hyperplan optimal est : $w.x + w_0 = 0$.

Les équations des hyperplans H_1 et H_2 sont respectivement $w.x + w_0 = +1$ et $w.x + w_0 = -1$.

Tous les exemples (x_i, u_i) vérifient les deux contraintes suivantes par rapport à l'hyperplan :

$$w.x + w_0 > 0 \quad si \quad u_i = +1$$
$$w.x + w_0 < 0 \quad si \quad u_i = -1$$

Supposons qu'il existe une valeur k qui est la valeur minimale de $|w.x + w_0|$, on obtient alors $w.x + w_0 \geq k$. Il est donc possible de normaliser la valeur de biais w_0 pour avoir les plans support suivant :

$$w.x + w_0 \geq +1 \quad si \quad u_i = +1$$
$$w.x + w_0 \leq -1 \quad si \quad u_i = -1$$

Ces deux contraintes peuvent être traitées sous la forme d'un ensemble d'inéquations :

$$u_i * (w.x_i + w_0) - 1 \geq 0 \quad i = 1, \ldots, N_x$$

Le problème d'optimisation consiste donc à trouver la paire d'hyperplans qui maximise la marge γ. La marge équivaut à la somme des distances entre les hyperplans H_1 et H_2.

$$\gamma = \frac{1}{\|w\|} + \frac{1}{\|w\|} = \frac{2}{\|w\|}$$

Il est possible de reformuler le problème d'optimisation de la marge par un problème d'optimisation quadratique :

minimiser $\frac{1}{2}\|w\|^2$, sous les contraintes énoncées précédemment.

Dans les cas où les classes de l'ensemble d'apprentissage ne sont pas linéairement séparables, on peut effectuer une projection des exemples dans un espace de redescription *F*, éventuellement de plus grande dimension, qui permet de rendre le jeu de données linéairement séparable. On définit la fonction de redescription Φ par :

$$\Phi : \chi \to F$$
$$x \to \Phi(x)$$

L'étape d'apprentissage et de classification se fait ensuite dans l'espace de redescription.

La partie suivante concerne la résolution du problème d'optimisation des SVM et en particulier l'apport de la théorie des Lagrangiens dans cette étape importante de la théorie des SVM.

9.5. Utilisation des expressions primales et duales pour la résolution du problème d'optimisation

Les fonctions de coût L utilisées pour la théorie des vecteurs de support sont des fonctions convexes quadratiques dont les contraintes sont linéaires. Les méthodes d'optimisation sont appelées dans ce cas des programmes quadratiques convexes.

9.6. Formulation du problème général d'optimisation

Soient les fonctions f , g_i , $i = 1,\ldots,k$ et h_j , $j = 1,\ldots,m$, définies sur l'ensemble $\Omega \subseteq \Re^n$, le problème général d'optimisation primal est défini par :

$$\begin{cases} \textit{minimiser} \quad f(w) \quad w \in \Omega \\ \textit{sous les contraintes} \quad g_i(w) \le 0 \quad i = 1,\ldots,k \\ \qquad\qquad\qquad\quad h_j(w) = 0 \quad j = 1,\ldots,m \end{cases}$$

On appelle $f(w)$ la fonction objective, g_i la contrainte d'inégalité et h_i la contrainte d'égalité. La valeur optimale de f est appelée valeur du problème d'optimisation. La région F , telle que f y est définie et que toutes les contraintes sont satisfaites, est appelée région de confiance. w est une variable primale. La solution $w^* \in F$ est appelée minimum global de f si et seulement si il n'existe aucune solution $w \in f$ telle que $f(w) < f(w^*)$. S'il

existe une valeur $\varepsilon > 0$ telle que $f(w) > f(w^*)$ pour tout $w \in \Omega$ avec $\|w - w^*\| < \varepsilon$, alors w^* est un minimum local de f.

Si la contrainte d'inégalité $g_i(w) = 0$ alors g_i est une contrainte active sur w. Les contraintes actives sur w^* sont intéressantes. En effets, si $\{i : g_i(w^*) = 0\}$ est connu alors les contraintes restantes peuvent être ignorées localement.

9.6.1. Définition de la convexité

Un ensemble $\Omega \subseteq \Re^n$ est convexe si $\forall w,u \in \Omega$, et pour chaque point $\theta \in [0,1]$, le point $(\theta w + (1-\theta)u) \in \Omega$. Une fonction continue $f(w)$ est convexe pour $w \in \Re^n$ si $\forall w,u \in \Omega$, et pour chaque point $\theta \in (0,1)$, $f(\theta w + (1-\theta)u) \leq \theta f(w) + (1-\theta) f(u)$.

Si $f(\theta w + (1-\theta)u) < \theta f(w) + (1-\theta) f(u)$ alors $f(w)$ est strictement convexe.

Tous les problèmes de programmation convexe ont la propriété commune que chacune des solutions du problème convexe est une solution globale. De plus, chaque solution est unique si la fonction $f(w)$ est strictement convexe.

La théorie des vecteurs de support se restreint aux cas des contraintes linéaires, des fonctions objectives quadratives convexes dans $\Omega \subset \Re^n$.

9.6.2. Théorie de Lagrange

Le cas de problème d'optimisation le plus simple correspond à l'absence de contraintes d'égalité ou d'inégalité. Dans ce cas la solution est caractérisée par la stabilité de la fonction objective. Ce cas est représenté dans le théorème de Fermat.

La condition nécessaire et suffisante telle que w^* soit un minimum de la fonction convexe $f(w)$, $f \in C^1$ (C^1 : l'ensemble des fonctions convexes) est :

$$\frac{\partial}{\partial w} f(w) = 0.$$

Une fonction Lagrangienne L_P est définie pour prendre en compte les contraintes et la fonction objective. La fonction $L_p(w)$ la plus aisée équivaut à la somme de la fonction objective et de la combinaison linéaire des contraintes d'égalité.

$$L_P(w,\alpha) = f(w) + \sum_{i=1}^{m} \alpha_i h_i(w)$$

Les variables duales α sont appelées multiplicateurs de Lagrange.

Le théorème de Lagrange découle du théorème de Fermat et tient compte des contraintes d'égalité.

La condition nécessaire et suffisante telle que $w^* \in \Re^n$ *soit une solution du problème d'optimisation de la fonction objective convexe* $f(w)$, $f \in C^1$ *est :*

$$\begin{cases} \frac{\partial}{\partial w} f(w^*, \alpha^*) = 0 \\ \frac{\partial}{\partial \alpha} f(w^*, \alpha^*) = 0 \end{cases}$$

Le point optimal correspond à celui pour lequel les contraintes sont nulles : $L_p(w^*, \alpha^*) = f(w^*)$.

La définition de la fonction Lagrangienne généralisée prend en compte les contraintes d'inégalités :

$$L_p(w,\alpha,\beta) = f(w) + \sum_{i=1}^{k} \alpha_i h_i(w) + \sum_{j=1}^{m} \beta_j g_j(w)$$

L'expression duale du problème d'optimisation généralisé équivaut à :

$$\begin{cases} maximiser \quad L_D(\alpha,\beta) = \min_w L_P(w,\alpha,\beta) \\ \quad sous\ la\ contrainte \quad \alpha > 0 \end{cases}$$

Le théorème de Karush-Kuhn-Tucker montre que les formulations primales et duales du problème d'optimisation possède la même solution. Cette solution est appelée point-selle du Lagrangien.

Les conditions nécessaires et suffisantes telle que w^* *soit un optimum du problème d'optimisation primal pour la fonction convexe* $f(w)$, $f \in C^1$, g_i *et* h_i *affines est l'existence des variables duales* α^* *et* β^* *telles que :*

$$\begin{cases} \dfrac{\partial}{\partial w} f(w^*, \alpha^*, \beta^*) = 0 \\ \dfrac{\partial}{\partial \beta} f(w^*, \alpha^*, \beta^*) = 0 \\ \alpha_i^* g_i(w^*) = 0 \quad i = 1, \ldots, k \\ g_i(w^*) \le 0 \quad i = 1, \ldots, k \\ \alpha_i^* \ge 0 \quad i = 1, \ldots, k \end{cases}$$

La relation $\alpha_i^* g_i(w^*) = 0$ est appelée la condition de complémentarité de Karush-Kuhn-Tucker. Elle montre que les contraintes inactives possèdent des multiplicateurs de Lagrange nuls.

La résolution de la forme duale du problème d'optimisation a l'avantage d'être plus simple que la résolution de la formulation primale. Les variables duales sont les inconnues et les variables primales ne sont pas calculées. Les contraintes de maximisation de la forme duale du problème d'optimisation sont aussi plus simples.

La partie suivante traite de l'application de la théorie des Lagrangien au sein de la théorie des SVM. Elle montre qu'un point x_i de l'ensemble d'apprentissage est un vecteur de support (ou exemple critique) si et seulement si le multiplicateur de Lagrange qui lui correspond α_i est strictement positif.

9.7. L'expression primale du problème d'optimisation des SVM

La théorie des Lagrangien permet de résoudre le problème d'optimisation des SVM.

L'hyperplan optimal est celui dont le vecteur poids w vérifie l'équation :

$$\underset{w,w_0}{\operatorname{argmax}} \min\left\{\|x - x_i\| : x \in \Re^d, (w.x + w_0) = 0, i = 1,\ldots,N_x\right\}$$

Sachant que pour cet hyperplan la marge vaut $\frac{2}{\|w\|}$, la recherche de l'hyperplan optimal revient à minimiser $\|w\|^2$.

Le problème d'optimisation s'écrit sous la formulation primale suivante :

$$\begin{cases} \textit{minimiser} \quad \frac{1}{2}\|w\|^2 \\ \textit{sous les contraintes } u_i(w.x_i + w_O) \geq 1, i = 1,\ldots,m \end{cases}$$

Soit :

$$L_P(w, w_0, \alpha) = \frac{1}{2}\|w\|^2 - \sum_{i=1}^{N_x} \alpha_i (u_i \cdot (x_i \cdot w + w_0) - 1)$$

La résolution de cette formulation nécessite le réglage de $d+1$ paramètres, d équivaut à la dimension de l'espace des entrées χ. On peut utiliser des méthodes de programmation quadratique si la dimension d est petite. Mais cette solution n'est pas envisageable pour des valeurs de d supérieures à quelques centaines. Dans ces conditions, il existe la possibilité de transformer le problème d'optimisation sous une formulation duale pour le résoudre.

9.8. L'expression duale du problème d'optimisation des SVM

Si le problème d'optimisation possède une fonction objective et des contraintes convexes, il possède une forme duale. Dans tous les cas, la résolution de la forme duale du problème revient à résoudre le problème original. La formulation duale du problème est la suivante :

$$\begin{cases} \textit{maximiser } L_D(\alpha) = \left\{ \sum_{i=1}^{N_x} \alpha_i - \frac{1}{2} \sum_{i,j=1}^{N_x} \alpha_i \alpha_j u_i u_j (x_i . x_j) \right\} \\ \textit{sous les contraintes} \quad \alpha_i \geq 0, \quad i = 1, \ldots, N_x \\ \qquad\qquad \sum_{i=1}^{N_x} \alpha_i u_i = 0 \end{cases}$$

Les points pour lesquels les multiplicateurs de Lagrange sont **non nuls**, $x_i \cdot w + w_0 = \pm 1$, sont situés sur les hyperplans frontières. Ils correspondent aux vecteurs de support (ou exemples critiques) dans les SVM. L'hyperplan solution correspondant peut alors s'écrire

$$f(x) = w^* \cdot x + w_0^* = \sum_{i=1}^{N_x} \alpha_i^* u_i x . x_i + w_0^*$$

, où les α_i^* sont solutions de la fonction de formulation duale et w_0^* est obtenu pour tout exemple critique (x_c, u_c) dans l'équation :

$$\alpha_i^* \left[u_i \cdot \left(x_i \cdot w^* + w_0^* \right) - 1 \right] = 0$$

On remarque que la résolution de l'hyperplan optimal ne requiert que le calcul du produit scalaire $x . x_i$ entre des vecteurs de l'espace des entrées χ. De plus, la solution dépend du nombre d'exemples critiques, et non plus de la dimension de l'espace des entrées. Ce constat permet d'appliquer en pratique des méthodes de programmation quadratique lorsque le nombre d'exemples critiques est assez petit.

9.9. Classification SVM d'un ensemble des données non linéaire

Appliquons cette formulation au problème des SVM dans le cas d'un espace des entrées χ non linéaire (Figure 25).

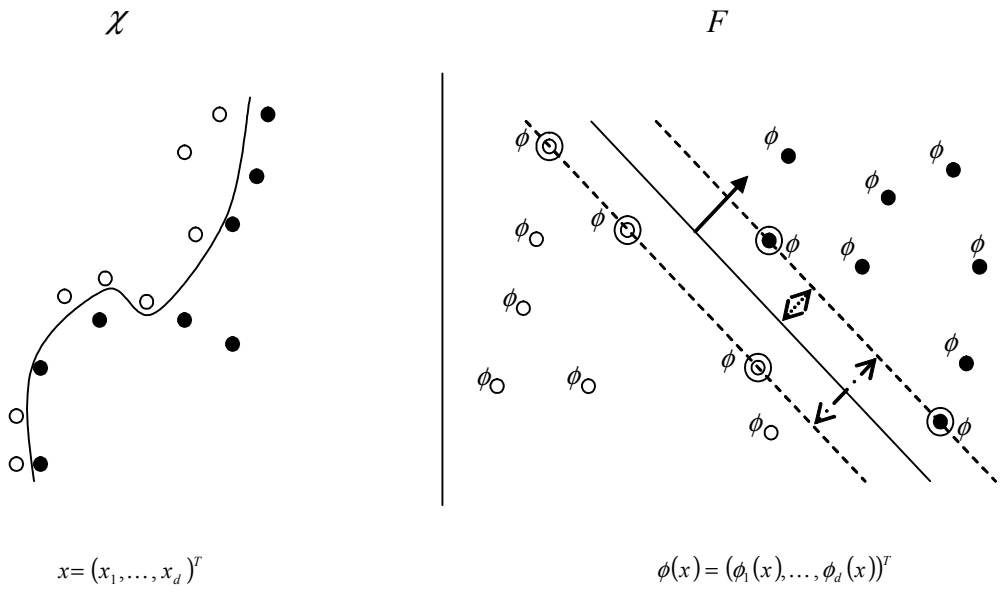


Figure 25. Passage dans un espace de redescription de grande dimension pour un ensemble d'entrées non linéaire.

.

Soit ϕ la fonction de redescription sur les données de l'espace des entrées χ :

$$x=(x_1,\ldots,x_d) \quad \mapsto \quad \phi(x)=(\phi_1(x),\ldots,\phi_d(x))$$

Le problème d'optimisation équivaut à :

$$\begin{cases} \textit{maximiser } L_D(\alpha)=\left\{\sum_{i=1}^{N_x}\alpha_i-\frac{1}{2}\sum_{i,j=1}^{N_x}\alpha_i\alpha_j u_i u_j \phi(x_i).\phi(x_j)\right\} \\ \textit{sous les contraintes} \quad \alpha_i \geq 0, i=1,\ldots,N_x \\ \qquad\qquad \sum_{i=1}^{N_x}\alpha_i u_i = 0 \end{cases}$$

L'hyperplan optimal a alors pour équation :

$$f(x)=w^*\cdot x+w_0^*=\sum_{i=1}^{N_x}\alpha_i^* u_i \phi(x_i).\phi(x_j)+w_0^*$$

Les coefficients α_i^* et w_0^* sont calculés de la même manière que dans le cas général.

Le problème se pose au niveau du calcul du produit scalaire $\phi(x_i).\phi(x_j)$lorsque la dimension de l'espace de redescription F tend vers l'infinie. La solution se trouve dans l'utilisation des fonctions noyaux. Elles possèdent les caractéristiques qui permettent à s'affranchir les traitements qui sont effectués dans l'ensemble de redescription.

9.10. Les fonctions noyaux

Pour résoudre le problème de la recherche de la fonction de redescription et du calcul du produit scalaire $\phi(x_i).\phi(x_j)$, les SVM font appel à des fonctions particulières appelés fonctions noyaux. Elles sont bilinéaires positives symétriques. Ces fonctions ont la propriété de correspondre à un produit scalaire dans un espace de grande dimension. Dans le cas des SVM, ces fonctions permettent de travailler dans l'espace de redescription F sans se préoccuper de la fonction de redescription Φ.

La fonction noyau est notée $K(x_i, x_j)$ telle que $K(x_i, x_j) = (\phi(x_i).\phi(x_j))$.

La résolution du problème d'optimisation correspond alors à :

$$\begin{cases} \textit{maximiser } L_D(\alpha) = \left\{ \sum_{i=1}^{N_x} \alpha_i - \frac{1}{2} \sum_{i,j=1}^{N_x} \alpha_i \alpha_j u_i u_j K(x_i, x_j) \right\} \\ \textit{sous les contraintes} \quad \alpha_i \geq 0, i = 1, \ldots, N_x \\ \qquad\qquad \sum_{i=1}^{N_x} \alpha_i u_i = 0 \end{cases}$$

La solution de l'hyperplan optimal équivaut à :

$$f(x) = w^*.x + w_0^* = \sum_{i=1}^{N_x} \alpha_i^* u_i \cdot K(x_i, x_j) + w_0^*$$

Les conditions pour qu'une fonction noyau K corresponde à un produit scalaire dans un espace sont données dans le théorème de Mercer ;

Soit χ *un sous-ensemble de* $\Re^n$.

Si K *est une fonction noyau continue symétrique tel que pour l'opérateur intégral* $A: L_2(\chi) \longrightarrow L_2(\chi)$ *:*

$$g(y) = Af(y) = \int_a^b K(x,y) f(y) dy + h(x)$$

Vérifiant :

$$\int_{\chi \times \chi} K(x,x') f(x) f(x') dx dx' \geq 0$$

Pour toute fonction $f \in L_2(\chi)$ *(de carré sommable), alors la fonction* K *peut être développée en une série uniformément convergente en fonction des valeurs propres positives* λ_i *et des fonctions propres* ψ_i *:*

$$K(x,x') = \sum_j^N \lambda_j \psi_j(x) \psi_j(x')$$

où N *est le nombre de valeurs propres positives.*

Une condition équivalente aux conditions de Mercer est que la matrice de fonction noyau $K_{ij} := K(x_i, x_j)$ doit correspondre à une matrice de Gram (positives, semi- définies pour tout échantillon d'exemples $\{x_1, \ldots, x_{N_x}\}$).

Il existe des fonctions noyaux simples qui sont couramment utilisées. La fonction noyau linéaire : $K(x_i, x_j) = x_i \cdot x_j + 1$ correspond à la fonction polynomiale de degré 1. Les fonctions polynomiales sont de la forme $K(x_i, x_j) = (x_i \cdot x_j + \tau)^p$, où p est le degré du polynôme.

Le deuxième type de fonction noyau simple s'appelle les fonctions noyaux à base radiale $K(x_i, x_j) = e^{-\frac{\|x_i - x_j\|^2}{2\sigma^2}}$. Le degré p et l'écart type σ sont déterminés de manière empirique.

Une troisième fonction noyau appelée sigmoïde est définie par $K(x_i, x_j) = \tanh(ax_i.x_j - b)$. Cependant les valeurs de a et b doivent prendre des valeurs précises pour vérifier le théorème de Mercer.

9.11. Classification non linéairement séparable dans l'espace de redescription.

Seul le cas des données qui ne sont pas bruitées dans l'ensemble de redescription a été considéré jusqu'à présent. La technique des SVM peut être adaptée lorsque l'erreur empirique R_{emp} est non nulle. La technique des variables ressorts où « slack variables » $\xi_i \geq 0$ permet de diminuer la contrainte sur la marge et donc de trouver un compromis entre la complexité et le risque empirique. Les contraintes sont de la forme :

$$u_i(w \cdot x_i + w_0) \geq 1 - \xi_i, \xi_i \geq 0, \quad i = 1, \ldots, N_x$$

La valeur de ξ_i permet de savoir où se situe l'exemple x_i par rapport à l'hyperplan de séparation optimal. Si la variable $\xi_i > 1$, dans ce cas $u_i(w \cdot x_i + w_0) < 0$. L'exemple x_i est alors un exemple mal classé par rapport à l'hyperplan. Si $0 < \xi_i < 1$ alors x_i est correctement classifié mais se situe à l'intérieure de la marge. Si $\xi_i = 0$ alors x_i est bien classé et se situe à l'extérieur de la marge ou sur le bord de la marge.

La formulation primale du problème revient à :

$$\begin{cases} \textit{minimiser} \quad \frac{1}{2}\|w\|^2 + C\sum_{i=1}^{m} \xi_i \\ \textit{sous les contraintes } u_i(w.x_i + w_O) \geq 1 - \xi_i, \xi_i \geq 0, i = 1, \ldots, N_x \end{cases}$$

, C, appelé le poids sur l'erreur, est une constante strictement positive.

La formulation duale du problème d'optimisation équivaut alors à :

$$\begin{cases} \textit{maximiser} \quad L_D(\alpha) = \left\{ \sum_{i=1}^{N_x} \alpha_i - \frac{1}{2} \sum_{i,j=1}^{N_x} \alpha_i \alpha_j u_i u_j x_i . x_j \right\} \\ \textit{sous les contraintes} \quad 0 \le \alpha_i \le C, i = 1, \ldots, N_x \\ \qquad\qquad\qquad\qquad \sum_{i=1}^{N_x} \alpha_i u_i = 0 \end{cases}$$

On revient au problème d'optimisation dans le cas de l'espace de redescription strictement séparable avec une contrainte sur α_i supplémentaire.

Les conditions d'optimisation de Karush-Kuhn-Tucker (ou conditions de second ordre) établissent des critères nécessaires (suffisants) pour qu'un ensemble de variables soit optimal pour un problème d'optimisation. Pour le problème de formulation duale les conditions KKT sont :

$$\begin{array}{lll} \alpha_i = 0 & \Rightarrow u_i f(x_i) \ge 1 & et \quad \xi_i = 0 \\ 0 < \alpha_i < C & \Rightarrow u_i f(x_i) = 1 & et \quad \xi_i = 0 \\ \alpha_i = C & \Rightarrow u_i f(x_i) \le 1 & et \quad \xi_i \ge 0 \end{array}$$

La valeur de la constante C correspond au poids sur les erreurs. Sa valeur est fixée empiriquement et constitue la valeur de la pénalité accordée aux erreurs. Plus la valeur de C tend vers l'infini, moins on autorise d'erreurs de classification sur les exemples d'apprentissage.

Les exemples correctement classés et situés à l'intérieur de la marge sont associés aux variables ressorts $\xi_i > 0$, tels que $\alpha_i = C$ et $u_i f(x_i) < 1$. Leur distance avec l'hyperplan est inférieure à $\frac{1}{\|w\|}$. Les exemples critiques sont caractérisés par $\xi_i = 0, 0 < \alpha_i < C$ et $u_i f(x_i) = 1$.

9.12. Classification SVM multi classes.

Dans les parties précédentes, la théorie des SVM a été traitée sur des ensembles d'exemples constitués d'au plus deux classes.

Cette partie traite de la classification SVM sur un ensemble d'au moins trois classes. Les deux stratégies exposées dans les deux parties suivantes montrent que l'on peut étendre le problème de la classification biclasse aux problèmes multiclasses.

La première approche consiste en une heuristique qui considère à chaque fois une classe contre toutes les autres classes de l'ensemble d'apprentissage. Soit un ensemble d'apprentissage constitué de n classes. Il s'agit de construire n fonctions de décisions : $f_k(x) = sign((w_k.x) + w_{0,k}), k = 1,\ldots,n$ tel que pour tout exemple x :

$$f_k(x) = \begin{cases} +1 \; si \; x \in k \\ -1 \; sinon \end{cases}$$

Le problème revient à effectuer n classifications SVM biclasses. Cette méthode entraîne le chevauchement des classes. La classe de chaque exemple est déterminée à partir de la méthode du « meilleur gagne ». La distance entre chaque exemple et un des n hyperplans est donnée par $f_k'(x) = (w_k.x) + w_{0,k}$. La classe k d'un exemple est celle pour laquelle ont obtient la valeur maximal de $f_k'(x)$:

$$k^* = \underset{k}{\text{argmax}} \; f_k'(x).$$

Cette heuristique a le mérite d'être facile a implémenter et d'être meilleure que d'autre méthodes de classifications binaires. Cependant, les hyperplans obtenus ne sont pas optimaux. Les marges maximales entre les classes ne sont pas correctes. La seconde approche présentée effectue la recherche de tous les hyperplans optimaux biclasses. Cette heuristique est davantage en adéquation avec les concepts mis en place par les SVM.

Soit la fonction de décision :

$$f_{kl} : \Re \begin{cases} +1 & si \quad x \in l \\ -1 & si \quad x \in k \end{cases}$$ pour la classification entre les classes k et l.

Pour un ensemble de n classes il faut effectuer $\frac{n(n-1)}{2}$ classifications biclasses. La classe de l'exemple x est déterminée par $\underset{k}{\operatorname{argmax}} f_k(x)$ où $f_k(x) = \sum_{l=1}^{n} f_{kl}(x)$.

Cette méthode a pour avantage que les frontières obtenues entre les classes sont des parties des hyperplans optimaux biclasses, contrairement à la première méthode.

Une méthode de classification multi classe directe est possible.

La formulation duale du problème d'optimisation prend la forme suivante :

$$\begin{cases} \textit{maximiser} \\ \quad L_D(\alpha) = \left\{ \sum_{k=1}^{n} \sum_{m \neq k} \left[\sum_{i=1}^{l_k} \alpha_i^{k,m} - \frac{1}{2} \sum_{m^* \neq k} \left(\sum_{i,j=1}^{l_k} \alpha_i^{k,m^*} \alpha_j^{k,m} x_i^k . x_j^k + \sum_{i=1}^{l_m} \sum_{j=1}^{l_{m^*}} \alpha_i^{m,k} \alpha_j^{m^*,k} x_i^m . x_j^{m^*} \right. \right. \right. \\ \qquad \left. \left. \left. - 2 \sum_{i=1}^{l_k} \sum_{j=1}^{l_m} \alpha_i^{k,m^*} \alpha_j^{m,k} x_i^k . x_j^m \right) \right] \right\} \\ \textit{sous les contraintes} \quad 0 \leq \sum_{m \neq k} \alpha_i^{k,m} \leq C, \\ \qquad \sum_{m \neq k} \sum_{i=1}^{l_k} \alpha_i^{k,m} = \sum_{m \neq k} \sum_{j=1}^{l_m} \alpha_j^{m,k} \quad k = 1, \ldots, n \end{cases}$$

La solution de l'hyperplan pour la classe k équivaut à :

$$f_k(x) = \sum_{m \neq k} \sum_{i=1}^{l_k} \alpha_i^{k,m} \left(x_i^k . x \right) + \sum_{m \neq k} \sum_{j=1}^{l_m} \alpha_j^{m,k} x_j^k . x + w_{0,k}$$

Si $n = 2$ on se retrouve la solution au problème de classification biclasse et comme précédemment il suffit de remplacer $x_i^k . x_j^m$ par $K\left(x_i^k, x_j^m\right)$ dans les équations.

9.13. Classification d'ensembles de données déséquilibrées :

Jusqu'à présent, deux classes possédaient un nombre d'exemples égal. En pratique une classe possède moins d'exemples que l'autre. Il faut donc prendre en compte le fait qu'une erreur de classification coûte plus chère à la plus petite des deux classes plutôt qu'à la deuxième.

L'idée de base consiste à donner un poids d'erreurs à chaque classe relatif à leur taille respective. L'effet recherché consiste à éloigner la classe la plus petite de l'hyperplan optimal par rapport à l'autre classe.

Soit $I_+ = \{ i \mid u_i = +1 \}$ et $I_- = \{ i \mid u_i = -1 \}$. On note C^+ et C^- les poids respectifs des deux classes.

La formulation primale s'exprime par :

$$\begin{cases} \textit{minimiser} \quad \frac{1}{2}\|w\|^2 + C^+ \sum_{i \in I_+} \xi_i^k + C^- \sum_{i \in I_-} \xi_i^k \\ \textit{sous les contraintes} \ u_i (w.x_i + w_0) - 1 + \xi_i \geq 0 \quad i = 1, \ldots, l \\ \xi_i \geq 0 \quad i = 1, \ldots, l \end{cases}$$

La formulation duale du problème d'optimisation équivaut à :

$$\begin{cases} \textit{maximiser} \quad L_D(\alpha) = \left\{ \sum_{i=1}^{N_x} \alpha_i - \frac{1}{2} \sum_{i,j=1}^{N_x} \alpha_i \alpha_j u_i u_j x_i . x_j \right\} \\ \textit{sous les contraintes} \quad 0 \leq \alpha_i \leq C^+, u_i = +1 \\ \qquad\qquad\qquad\quad 0 \leq \alpha_i \leq C^-, u_i = -1 \\ \qquad\qquad\qquad\quad \sum_{i,j=1}^{N_x} \alpha_i u_i = 0 \end{cases}$$

Le passage par les fonctions noyaux s'effectue en substituant $x_i . x_j$ par $k(x_i, x_j)$.

Le poids doit être plus grand pour la classe la plus large. Cependant les valeurs de ces paramètres doivent être testés de façon empirique.

9.14. Algorithmes de classification SVM

Les algorithmes de SVM se ramènent à la résolution d'un programme quadratique pour la résolution des formulations duales du problème d'optimisation (Figure 26).

Entrée :

- m exemples d'apprentissages : $\{(x_i, u_i)\}_{i=1,m}$ où $x_i \in \Re^n$ et $u_i \in \{-1,+1\}$
- la fonction noyau K et ses paramètres
- le paramètre C (contrôle de la marge et des erreurs)

Apprentissage

- résoudre le programme quadratique
- obtenir les multiplicateurs de Lagrange α_i
- obtenir les vecteurs support $\alpha_i > 0$
- calculer le biais w_O en se basant sur les vecteurs de support

Classification : (SV ensemble des vecteurs support)

$$f(x) = sign\left(\sum_{i=1}^{SV} u_i \alpha_i K(x_i, x) + w_0\right)$$

Figure 26. Mise en œuvre des SVM

9.15. Application des SVM en bioinformatique

Les SVM sont utilisés dans les techniques de classification dites discriminantes à l'opposé des techniques génératives qui ne se basent que sur les exemples d'une seule classe. Cette partie présente les différents travaux qui concernent l'application des SVM dans la problématique de la détection et la classification de protéine à faible homologie de séquences. La nécessité d'utiliser des stratégies discriminantes est apparue lorsque les méthodes telles que BLAST, PSI_BLAST, les alignements multiples ou les modèles de chaînes de Markov cachés n'étaient pas applicables pour certaines familles de protéines de faibles homologies.

La première application des SVM pour la classification d'homologues éloignées dans les familles protéiques est la méthode des Fisher SVM (Jaakkola, Diekhans et al. 2000).

Cette méthode nécessite tout d'abord l'apprentissage d'un modèle de Markov caché (HMM) sur la famille de protéines d'intérêt (Karplus, Barrett et al. 1998). Ensuite un jeu d'apprentissage constitué d'exemples et de contre-exemples est représenté dans un espace

vectoriel dont chaque dimension est liée à un état ou une transition présent dans le modèle. La valeur de chaque dimension du vecteur est appelée score de Fisher et représente le taux d'utilisation de chaque paramètre du modèle pour modéliser chaque séquence exemple.

Les vecteurs obtenus sont utilisés conjointement avec une fonction noyau particulière, appelée fonction de Fisher, pour l'apprentissage du modèle SVM. Les auteurs ont montré que les classifications par la méthode des Fisher SVM surpassent les méthodes génératives telles que les HMM ou BLAST. Cependant cette méthode trouve ses limites dans la nécessité d'avoir une famille de protéines possédant un nombre important de membres pour pouvoir élaborer le modèle HMM. Elle se caractérise aussi par une complexité temporelle importante autant en phase d'apprentissage que lors de la prédiction de la classe d'une protéine.

La méthode des pairwise SVM (Liao and Noble 2003) utilise l'algorithme d'alignement local de Smith et Waterman. Les exemples sont vectorisés dans un espace dont la dimension correspond au cardinal du jeu d'apprentissage. Chaque dimension du vecteur correspond au score de similarité obtenu par l'alignement local avec une séquence du jeu d'apprentissage.

Les auteurs de cette méthode ont obtenu des performances de classification supérieure à la méthode des Fisher SVMs. Cependant elle ne résout pas le problème de la complexité temporelle et rend la dimension des vecteurs dépendante de la taille du jeu d'apprentissage.

Une alternative au problème de complexité temporelle est le spectre de chaîne (Leslie, Eskin et al. 2002). Dans cette méthode, chaque vecteur représente les fréquences des motifs trouvés sur la séquence de l'exemple. Les vecteurs servent ensuite à l'apprentissage du modèle SVM.

Ce simple procédé s'affranchit des méthodes génératives et permet un gain de temps d'exécution remarquable. Une adaptation de cette méthode prend en compte tous les motifs de taille k fixée et autorise au plus m variations entre deux motifs (Leslie, Eskin et al. 2004).

Cette adaptation prend en compte le concept biologique de mutation des résidus. Les auteurs de cette technique montrent que la performance de leur classification est comparable à la méthode des Fisher SVM tout en optimisant les temps d'exécution. Nos propres expérimentations n'ont pas révélé une amélioration des performances sur notre famille d'intérêt par rapport à la méthode de spectre de chaîne. C'est donc cette dernière méthode qui nous servira de point de comparaison avec notre propre algorithme de classification.

Les méthodes décrites jusqu'à présent ne prennent en compte que l'information située au niveau de la structure primaire des protéines.

La méthode SVM-I-sites (Hou, Hsu et al. 2003) tente d'intégrer une information structurale lors de l'étape de vectorisation des exemples. Une étape préalable consiste à rechercher un profil de séquences à partir de la séquence primaire de l'exemple. Le profil est obtenu au moyen du logiciel PSI-BLAST (Altschul, Madden et al. 1997)et de la base de données Swiss-Prot (Boeckmann, Bairoch et al. 2003).

La création des vecteurs se base ensuite sur la recherche des occurrences de 263 domaines structuraux au sein du profil. Les auteurs indiquent que cette méthode apporte des résultats équivalents à la méthode des pairwise SVM. Elle est aussi plus efficace lors de l'étape de vectorisation.

Les méthodes SVM présentés ci-dessus sont détaillées dans les parties suivantes.

9.15.1. Fisher SVM

La méthode des Fisher SVM est la première approche discriminante des SVM qui cherche à lever le problème des homologies de séquences dans l'objectif de résoudre la classification des protéines dans des familles ou super familles de protéines connues (Jaakkola, Diekhans et al. 1999; Jaakkola, Diekhans et al. 2000). L'idée générale de cette méthode est de partir d'un modèle de Markov caché pour la famille de protéines d'intérêt puis d'effectuer l'apprentissage d'un modèle SVM sur des protéines homologues de la famille et des contre-exemples qui eux ne comportent aucune homologie avec ses membres. La phase préliminaire de la méthode consiste à optimiser les alignements multiples des séquences de la famille de protéines qui doit être modélisée. Les auteurs utilisent le logiciel de modélisation SAMT98 (Karplus, Barrett et al. 1998) sur le profil de la famille de protéines pour concevoir le modèle de Markov caché.

Le score de vraisemblance entre une séquence et le modèle HMM permet d'estimer la distance entre la séquence et le modèle. Cependant, il ne permet pas de discriminer deux séquences différentes qui possèdent le même score de vraisemblance par rapport au modèle

HMM. En d'autres termes cette valeur ne permet pas de savoir quelle est la similarité entre les deux séquences. Pour chaque séquence, les auteurs utilisent l'algorithme appelé *forward-backward* qui permet de récupérer toutes les fréquences (sufficient_statistics) de passage dans un état de transition et de génération d'un acide aminé à partir d'un état donné. La séquence peut alors être représentée sous une forme vectorielle fixe dont chaque dimension correspond à la fréquence de chaque paramètre indépendant du modèle. De ce fait, la comparaison des valeurs de fréquences de chaque dimension permet de déterminer la similarité entre deux séquences différentes.

Ce vecteur de fréquences est une représentation intermédiaire de chaque exemple, il n'est pas calculé directement mais généralisé par le vecteur de score de Fisher :

$$U_x = \nabla_\theta \log P(x|H_1, \theta).$$

Chaque composant du vecteur de score de Fisher U_x correspond à la dérivée du logarithme de la probabilité pour que l'exemple x vérifie l'hypothèse de vraisemblance H_1 (la protéine appartient à la famille d'intérêt) en fonction de chaque paramètre d'état ou de transition, θ, du modèle de Markov caché.

Il est possible de déduire la similarité entre deux exemples x et x' à partir de U_x et $U_{x'}$ qui puisse être mesurée sous la forme d'une distance :

$$D^2(x, x') = \frac{1}{2}(U_x - U_{x'})^T F^{-1}(U_x - U_{x'})$$

, F est la matrice d'information de Fisher. Elle est approchée par $F \approx \sigma^2 I$ telle que I est la matrice d'identité et σ un paramètre d'échelle. La valeur de σ est choisie comme étant la distance médiane entre les scores de Fisher des exemples positifs de l'ensemble d'apprentissage et le plus proche score de Fisher obtenu par un contre-exemple (une séquence non homologue). Cette distance est utilisée dans la fonction noyau gaussienne RBF :

$$K(x, x') = e^{-D^2(x,x')}$$

La fonction de décision correspond alors à :

$$f(x) = \sum_{i \in \chi, H_1} \alpha_i K(x, x_i) - \sum_{i \in \chi, H2} \alpha_i K(x, x_i)$$

La complexité temporelle vaut $O(nmp)$, n est le nombre d'exemples dans l'ensemble d'apprentissage, m est la longueur de la séquence d'apprentissage la plus longue, p est le nombre de paramètres du profil HMM. La phase de classification SVM a une complexité en $O(n^2)$.

9.15.2. Pairwise SVM

La différence principale entre la méthode Fisher SVM et Pairwise SVM (Liao and Noble 2003) réside dans la phase de vectorisation des exemples. Celle-ci se base sur l'algorithme d'alignement de séquences deux à deux de Smith-Waterman (Smith and Waterman 1981).

La technique des SVM requiert la transformation vectorielle des données dans un espace à n dimensions fixes. Ici, cette dimension est donnée par la taille de l'ensemble d'apprentissage N_χ. Chaque séquence x est alignée à l'ensemble des séquences du jeu d'apprentissage au moyen de l'algorithme de Smith-Waterman. La similarité locale entre chacune des deux séquences de l'alignement est représentée par un score $p_{x,i}$ qui équivaut au logarithme de la p-value pour l'alignement de Smith-Waterman de la séquence x avec la séquence d'apprentissage au rang i. La pénalité pour l'ouverture d'un indel vaut 10 et pour l'extension d'un indel 0,05.

Le vecteur obtenu pour la séquence x correspond à :

$$x = (p_{x,1}, p_{x,i}, \ldots, p_{x,N_x})$$

La méthode Pairwise SVM effectue une normalisation telle que l'ensemble des vecteurs ont une norme unitaire dans l'espace de représentation F.

$$K^{norm}(x_i, x_j) = \frac{(x_i . x_j)}{\sqrt{(x_i . x_i)(x_j . x_j)}}$$

La fonction noyau RBF K est appliquée à partir de $K^{norm}(x_i, x_j)$:

$$K(x_i, x_j) = e^{-\frac{K^{norm}(x_i,x_i)-2K^{norm}(x_i,x_j)+K^{norm}(x_j,x_j)}{2\sigma^2}} + 1$$

L'écart type σ est déterminé en prenant la médiane de la distance euclidienne entre tous les $K(x_i, x_j)$ de chaque exemple positif x_i de l'ensemble d'apprentissage au contre exemple x_j le plus proche.

La méthode Smith-Waterman possède une complexité en $O(m^2)$, m est la longueur de la séquence d'apprentissage la plus longue.

La complexité temporelle totale de la méthode Pairwise est en $O(n^2 m^2)$, n est le nombre d'exemples dans l'ensemble d'apprentissage.

9.15.3. Spectrum kernel

La méthode Spectrum Kernel est remarquable par sa simplicité (Leslie, Eskin et al. 2002). Elle se base sur la fréquence des mots de taille k dans les séquences pour représenter chaque séquence sous une forme vectorielle.

Soit A un alphabet de cardinal $|A| = l$, dans le cas des acides aminés l vaut 20, et α un mot de taille $k \geq 1$ constitué de caractères de l'alphabet A. La fonction $\phi_\alpha(x)$ représente le nombre d'occurrences du mot α dans l'exemple x. La fonction de redescription $\phi_k(x)$ indexe tous les mots α, $\alpha \in A^k$, tels que :

$$\begin{cases} \chi \to \Re^{l^k} \\ \phi_k(x) = (\phi_\alpha(x))_{\alpha \in A^k} \end{cases}$$

La fonction noyau correspondante équivaut à :

$$K_k(x_i, x_j) = \phi_k(x_i) . \phi_k(x_j)$$

Le classifieur SVM utilise la fonction noyau normalisée :

$$K_k^{Norm}(x_i, x_j) = \frac{K_k(x_i, x_j)}{\sqrt{K_k(x_i, x_i) K_k(x_j, x_j)}}$$

L'implémentation d'un arbre des suffixes contenant l'ensemble des mots pour les séquences x_i et x_j permet d'effectuer le calcul de $K_k(x_i, x_j)$ en $O(knm)$.

9.15.4. Mismatch kernel

La méthode Mismatch Kernel (Leslie, Eskin et al. 2004) est une extension de la méthode Spectrum Kernel. Soit k la longueur d'une sous séquence (un mot), et A un alphabet de cardinal $|A| = l$. Le symbole α_i représente le caractère de rang i dans A. Un k-mer est une séquence $\alpha = \alpha_1 \alpha_2 \ldots \alpha_k$. Le voisinage de α est noté $N_{(k,m)}(\alpha)$ et représente l'ensemble des séquences β de longueur k formées à partir de l'alphabet A telles qu'elles diffèrent de α d'au plus m caractères. La valeur de m est inférieure ou égale à k. Si m est nul, le problème revient à utiliser la méthode Spectrum Kernel.

Soit la fonction de redescription $\phi_{(k,m)}(\alpha)$:

$$\begin{cases} \phi_{(k,m)}(\alpha) = (\phi_\beta(\alpha))_{\beta \in A^k} \\ \phi_\beta(\alpha) = 1 \quad si \quad \beta \in N_{(k,m)}(\alpha) \\ \phi_\beta(\alpha) = 0 \quad sinon \end{cases}$$

Soit une séquence de protéines x, $\phi_{(k,m)}(x)$ est définie telle que :

$$\phi_{(k,m)}(x) = \sum_{\alpha \in x} \phi_{(k,m)}(\alpha)$$

La fonction noyau correspondante équivaut à :

$$K_{(k,m)}(x_i x_j) = \phi_k(x_i) . \phi_k(x_j)$$

Les auteurs utilisent un arbre des (k,m) mésappariements pour calculer la matrice $K(x_i, x_j)$ $i, j = 1, \ldots, N$ où N est la taille de l'ensemble d'apprentissage. Les k-mers α sont issus de l'exemple x_i et l'ensemble β est donné l'exemple x_j.

La fonction noyau utilisée par le classifieur SVM est normalisée :

$$K^{Norm}_{(k,m)}(x_i, x_j) = \frac{K_{(k,m)}(x_i, x_j)}{\sqrt{K_{(k,m)}(x_i, x_i) K_{(k,m)}(x_j, x_j)}}$$

La complexité temporelle de la méthode de calcul des valeurs de la matrice de fonction noyau vaut $O(k^m l^m N^2)$. Cette méthode est donc plus rapide que Fisher SVM et Pairwise SVM. La complexité spatiale de cette méthode est faible grâce à une implémentation de l'arbre des (k,m) mésappariements qui ne le conserve pas entièrement en mémoire.

9.15.5. I-site kernel

La méthode I-site kernel (Hou, Hsu et al. 2003) inclut pour la première fois des informations de structure locale dans la phase de vectorisation des séquences de protéines alors que les méthodes précédentes ne prennent en compte que l'information contenue dans la structure primaire. Plus précisément cette méthode se base sur la description des protéines dans un espace vectoriel définit par 263 motifs structuraux obtenus par les auteurs.
La phase d'apprentissage d'un classifieur SVM est effectuée sur un ensemble d'exemples et de contre-exemples issus des familles structurales de la base de données SCOP.
La phase de vectorisation d'un exemple consiste à calculer la probabilité de la présence de chaque motif structural au sein de la séquence protéique. Le classifieur SVM utilise la fonction noyau RBF.
La prédiction d'une protéine inconnue consiste à vectoriser sa séquence et de confronter le vecteur aux modèles SVM obtenus lors de l'apprentissage.
La phase de vectorisation de la méthode I-site kernel est quatre fois plus rapide que celle de la méthode pairwise SVM pour une performance de classification équivalente.

Partie 3. Méthodes et Résultats

Chapitre 10. Définitions

Le travail effectué au cours de cette thèse a eu pour objectif d'élaborer des méthodes innovantes en classification des protéines à homologies éloignées, et tout particulièrement celles des sous familles IL-6,IL-2 et IL-10. La première méthode utilise les algorithmes génétiques pour extraire les signatures caractéristiques des interleukines. La seconde extrait un ensemble de motifs spécifiques dans le but d'effectuer l'apprentissage d'un classifieur SVM.

10.1. Définition des alphabets

L'ensemble des acides aminés est défini par :

$$\Omega = \{A, C, D, E, F, G, H, I, K, L, M, N, P, Q, R, S, T, V, W, Y\}.$$

La structure primaire d'une protéine est représentée par la séquence $s = \langle s_1, s_2 \ldots, s_i, \ldots, s_n \rangle$, telle que chaque élément $s_i \in \Omega$.

Soit $P(\Omega)$ l'ensemble des classes d'acides aminés qu'il est possible d'extraire à partir de Ω. Nous retiendrons le sous-ensemble $C(\Omega) \subset P(\Omega)$ de classes possédant des propriétés pertinentes pour notre application :

- l'ensemble ω des singletons $\omega = \{\{A\}, \{C\} \ldots \{Y\}\}$ qui rattache de manière non ambiguë un résidu à une position donnée.

- l'ensemble Ω lui-même, qui représente n'importe quel acide aminé (caractère appelé *joker* ou *wildcard* dans certains algorithmes d'alignement).

- l'ensemble $T(\Omega)$ des classes d'acides aminés partageant certaines propriétés physico-chimiques.

Nous utilisons la classification $T(\Omega)$ établie par Taylor (Taylor 1986) (Figure 27). L'alphabet $C(\Omega)$ correspond à l'union des ensembles Ω, $T(\Omega)$ et des ensembles singletons des acides aminés $\omega = \{\{A\},\{C\}\ldots\{Y\}\}$:

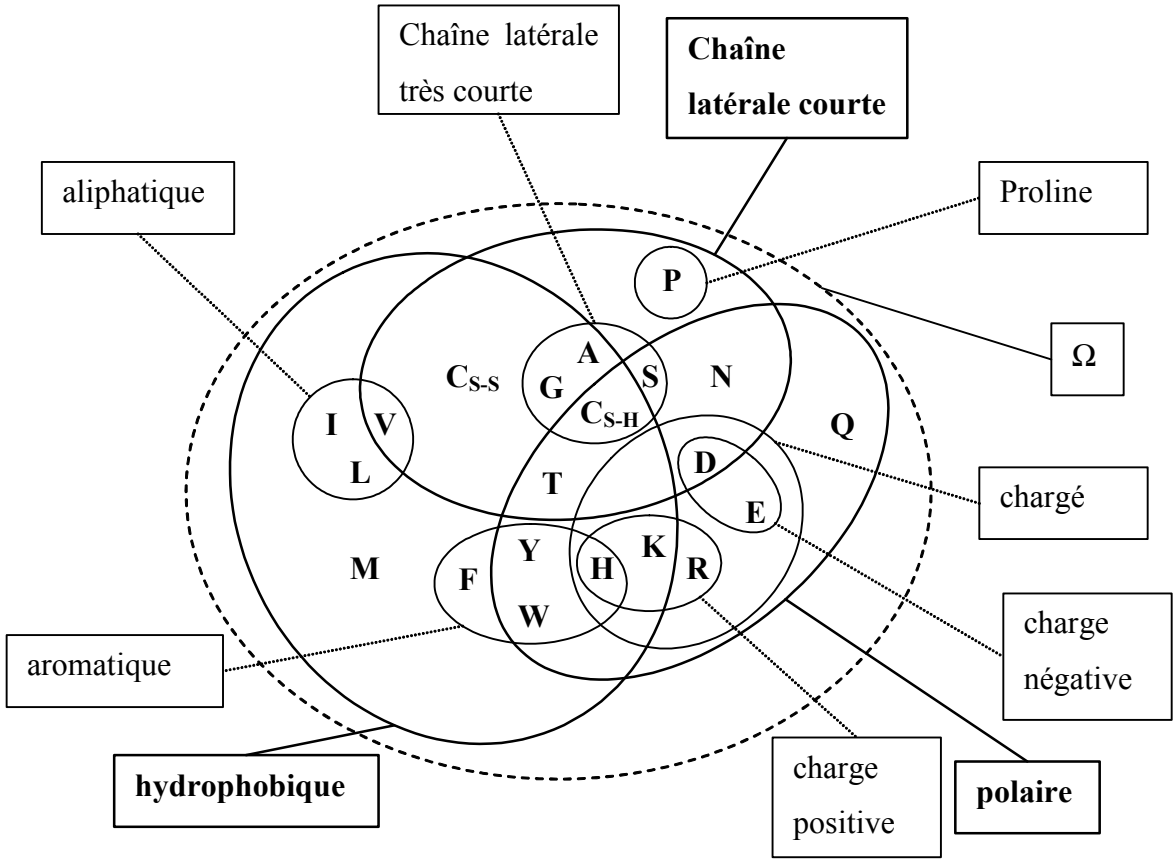


Figure 27. Diagramme de Ven de la classification des acides aminés de Taylor. L'ensemble des acides aminés est représentée par l'ensemble Ω. C_{S-S} : résidu cystéine dont la chaîne latérale est impliquée dans un pont disulfure. C_{S-H} : résidu cystéine avec une chaîne latérale réduite.

$$C(\Omega) = \Omega \cup T(\Omega) \cup \{\{A\},\{C\}\ldots\{Y\}\}.$$

Nous définissons les motifs à partir de cet ensemble de classes prédéfinies. De ce fait, ceci nous permet de réduire considérablement l'espace de recherche. Les motifs qui peuvent être obtenus sont toujours compatibles avec la syntaxe PROSITE (Hulo, Sigrist et al. 2004). Par exemple, le motif $M[aromatique]P$ correspond au motif PROSITE $M[HYFW]P$.

10.2. Classification des acides aminés en fonction de leur propriétés physico-chimiques

Nous avons retenu la classification des acides aminés de Taylor (Figure 27) (Taylor 1986). Les acides aminés y sont classés en fonction de leurs propriétés physicochimiques observées elle tient aussi compte des matrices de substitutions PAM (Dayhoff, Schwartz et al. 1978). Cette classification est simple et robuste. Sa pertinence est observée dans les régions conservées des séquences homologues. Les substitutions sont davantage conservées si elles se font entre des résidus d'une même classe physico-chimique.

L'ensemble des classes physico-chimiques est représenté par $T(\Omega)=\{\alpha,\beta,\gamma,\delta,\varepsilon,\varsigma,\eta,\theta\}$, où chaque classe de Taylor est représentée par un caractère grec (Tableau 2). Les classes physico-chimiques ne sont pas disjointes, par exemple la classe des acides aminés polaires contient la classe des acides aminés chargés. Cette dernière inclut à son tour les classes « charge positive » et « charge négative ». La proline est considérée comme une classe singleton incluse dans les acides aminés à chaînes courtes. Les acides aminés cystéine et thréonine appartiennent aux trois classes principales « chaîne latérale courte », « polaire » et « hydrophobique ». Il est donc possible d'établir une relation de spécialisation généralisation entre les acides aminés et les classes physico-chimiques. Par exemple $I \subseteq \alpha \subseteq \gamma \subseteq \Omega$.

Tableau 2. Tableau de correspondance entre les classes physico-chimiques des acides aminés de Taylor et les symboles caractérisant les propriétés physico-chimiques. Entre parenthèses, les termes anglais désignant les classes physico-chimiques.

Membres	Classe physico-chimique	symbole
I, L, V	Aliphatique (Aliphatic)	α
F, H, W, Y	Aromatique (Aromatic)	β
A, C, F, G, H, I, K, L, M, V, W, Y	hydrophobique (hydrophobic)	γ
D, E, H, K, R	Chargé (charged)	δ
C, D, E, H, K, N, Q, R, S, T, W, Y	Polaire (Polar)	ε
H, K, R	Charge positive (Positive)	ξ
A, C, D, G, N, P, S, T, V	Chaîne latérale courte (small)	η
A, C, G, S, T	Chaîne latérale très courte (tiny)	θ

10.3. Définition d'un motif

Un motif m de taille k est un mot sur l'alphabet $C(\Omega)$ formé de k sous-classes $m_i \in C(\Omega)$: $m = \langle m_1, m_2 \ldots m_k \rangle$. Par exemple le motif de longueur $k = 2$ $L\alpha$ est composé des sous-classes $m_1 = \{L\}$ et $m_2 = \alpha$: la classe de Taylor des résidus aliphatiques (Tableau 2).

Une sous-séquence $\langle s_i, s_{i+1} \ldots s_{i+k-1} \rangle$ dans une séquence s contient un motif m si et seulement si $s_{i+j} \in m_j \forall j, 0 \le j \le k-1$. On dit que la sous-séquence $\langle s_i, s_{i+1} \ldots s_{i+k-1} \rangle$ est l'*occurrence* de rang i du motif m dans la séquence s. Inversement on dit que la séquence s *vérifie* le motif m.

10.4. Définition d'une signature

Une n-signature σ est une série ordonnée de n motifs $\langle m^1, m^2 \ldots m^n \rangle$.

Une séquence s vérifie la signature σ si elle contient n sous-séquences $(s_1, s_2 \ldots, s_n)$ où chaque $s_i = m_i, i \in [1,n]$. Les motifs qui constituent une signature sont ordonnés. Leurs positions respectives, $(p_1, p_2 \ldots p_n)$, dans une séquence s ne se chevauchent pas : $p_{i+1} - p_i \geq |m_i| \; \forall i, 1 \leq i < n$, où $|m_i|$ représente la taille du motif.

10.5. Définition d'une signature complexe

Pour les familles de protéines telles que les interleukines, les algorithmes d'alignements de séquences standard tels que BLAST (Altschul, Gish et al. 1990) ou SSEARCH (Smith and Waterman 1981) ne sont pas suffisamment sensibles pour découvrir des homologies éloignées entre les séquences des membres. Dans d'autre cas, il n'existe pas non plus de motifs biologiques conservés au cours de l'évolution dans les séquences des différents membres.

D'un point de vue biologique, le concept de signature généralise celui de motif biologique et peut être défini comme un ensemble de propriétés biologiques conservées par tous les membres d'une sous famille. On peut, par exemple définir le repliement tridimensionnel conservé en faisceau de 4 hélices α comme la signature structurale des membres des familles IL-6, IL-2 et IL-10.

Dans notre travail, nous considérons qu'une signature complexe est définie sur les séquences primaires par un ensemble de positions dont l'ordre et les propriétés sont conservés entre différents membres d'une même famille.

10.6. Support et spécificité d'un motif, d'une signature, d'un ensemble de motifs et d'un ensemble de signatures

Les critères de support et de spécificité permettent de donner une valeur sur la représentativité d'un motif ou d'une signature pour un ensemble de séquences. Le *support* d'un motif m dans un ensemble de séquences S représente le nombre de séquences dans S qui vérifient m.

Considérons la séquence MH qui vérifie, entre autres, les motifs MH, $M\beta$, $\gamma\lambda$ et $\Omega\Omega$. Le motif MH n'est vérifié que par les séquences contenant une occurrence du motif MH alors que le motif $\Omega\Omega$ est vérifié par toutes séquences de taille supérieiure ou égale à 2. Il importe donc de qualifier la *spécificité* d'un motif en prenant en compte sa probabilité de présence au sein d'une séquence.

Un motif hiérarchique possède aussi les propriétés de support et de spécificité. L'importance de ces facteurs repose dans l'opposition de leur comportement. En effet plus un motif possède un support sur S fort, moins il aura tendance à être spécifique de S. L'exemple le plus extrême de ce cas de figure est constitué par le motif le plus général $m = \Omega$ qui possède une valeur de support maximal et une spécificité nulle. Inversement un k-mot spécifique d'un sous-ensemble de S risque de ne pas représenter toutes les séquences et donc d'avoir un support faible. Par exemple un motif trop spécifique est constitué des singletons qui ne sont présents que dans une seule séquence de S et jamais en dehors. Sa spécificité est forte mais son support est égal à 1.

Un motif est d'autant plus intéressant que son support est maximal dans S et que sa spécificité est minimale en dehors de S.

Une signature caractéristique pour un ensemble de protéines est composée d'une suite ordonnée de motifs qui vérifient certaines contraintes de support et de spécificité. La spécificité d'une signature est définie par la somme des spécificités des motifs qui la composent :

$$\phi(\sigma) = \sum_{i=1}^{k} \phi(m_i)$$

L'estimation de la spécificité d'un motif $\phi(m)$ correspond à la probabilité de le voir apparaître dans une séquence. Comme le calcul exact de cette probabilté est complexe et n'est pas utile pour le classifieur nous avons utilisé la fonction de coût suivante :

$$c(m) = \prod_{j=1}^{k} f(m_j),$$

, où $f(m_j)$ est la fréquence de la classe physicochimique m_j dans l'ensemble des contre exemples du jeu d'apprentissage et constitué des familles de protéines différentes des sous famille d'interleukines présent dans la base SCOP (Murzin, Brenner et al. 1995). On observe une bonne corrélation entre l'estimation $c(m)$ et le support du motif m dans l'ensemble des contre exemples d'apprentissage (Tableau 3).

La spécificité d'un motif m est alors définie par la relation :

$$\phi(m) = -\log(c(m)).$$

Tableau 3. Exemples de motifs avec leur support dans l'ensemble des interleukines, leur spécificité estimée et leur fréquence effective dans la base SCOP

Motif	Support	Spécificité	Fréquence (%)
$\beta\alpha\varepsilon$	1.00	4.4	79.84
$\alpha\delta\varepsilon\alpha$	0.96	5.1	60.95
$\delta\Omega\,\gamma\varepsilon\Omega\,\alpha\xi\varepsilon$	0.65	6.8	24.27
$\Omega\,\alpha\xi\varepsilon\alpha\varepsilon\varepsilon\gamma$	0.54	7.6	11.30
LEE	0.28	7.9	08.93
$\beta\gamma\Omega\,\Omega\,\gamma\xi\varepsilon L$	0.28	8.4	04.33
$\varepsilon\gamma\alpha\xi\delta L\,\Omega\,\varepsilon$	0.37	9.2	03.59
$L\varepsilon\varepsilon\gamma\alpha\varepsilon\delta L$	0.22	10.2	1.07
$\Omega\,\eta L\,\alpha L\,\alpha\Omega\,L$	0.15	11.0	0.19
$F\varepsilon R\,\gamma K\,\varepsilon\Omega\,\gamma$	0.15	11.4	0.18

10.7. Hiérarchisation des motifs

Les motifs peuvent être hiérarchisés selon une relation de généralisation au sein de l'ensemble ordonné $(C(\Omega),\subseteq)$. La relation de généralisation entre deux motifs m^1 et m^2 de longueur k est vérifiée si et seulement si pour une même position dans les deux motifs, la classe de m^1 est identique ou incluse dans la classe de m^2.

La spécificité $\phi(m)$ d'un motif m permet de définir la borne supérieure de deux motifs. Le motif $m^{1,2} = \sup(m^1, m^2)$ est le motif le plus spécifique généralisant m^1 et m^2. Quelque soit m, $m^1 \prec m$ et $m^2 \prec m$, on aura $\phi(m) \leq \phi(m^1, m^2)$. Il est à noter que $m^{1,2}$ est composé de classes vérifiant $m_i^{1,2} = \sup(m_i^1, m_i^2) \forall\, i \in [1,k]$.

L'extraction de motifs hiérarchiques conduit à la découverte de motifs ayant des supports plus grands, au détriment de leur spécificité. Les motifs hiérarchiques m^1 et m^2 possèdent respectivement des supports inférieurs ou égaux à celui de la borne supérieure $m^{1,2}$ et des spécificités supérieures ou égales à $m^{1,2}$.

10.8. Les séquences utilisées

La classification supervisée fait appel à la connaissance à priori des classes des exemples. Le jeu de données est alors constitué par n sous ensembles ou classes. L'ensemble de données se répartit entre l'ensemble des interleukines des familles des IL-6 (Tableau 4), IL-2 (Tableau 5) et IL-10/IFN (Tableau 6), dont les séquences sont rapatriées à partir de la base de données SWISSPROT (Boeckmann, Bairoch et al. 2003) et des contre exemples, les « non interleukines », qui sont présentes dans la base de données de classification structurale SCOP version 1.51 (Murzin, Brenner et al. 1995) ou genbank (Benson, Karsch-Mizrachi et al. 2004). L'ensemble des interleukines humaines regroupe 45 séquences pour 6615 séquences contre exemples.

Tableau 4. Tableau des identifiants des séquences de la famille des interleukines à chaînes longues des IL-6 humaines.

identifiants
sw|P05231|IL6_HUMAN
swall|P09919|CSF3_HUMAN
sw|P15018|LIF_HUMAN
sw|P01241|SOMA_HUMAN
sw|P26441|CNTF_HUMAN
sw|P01243|PLL_HUMAN
sw|P41159|OB_HUMAN
sw|P13725|ONCM_HUMAN
sw|P29459|I12A_HUMAN
swall|Q9H2A5|Q9H2A5 Interleukin 23 p19
sw|P20809|IL11_HUMAN
sw|P01236|PRL_HUMAN
swall|Q16619|CTF1_HUMAN
gi|15394278|emb|CAC60178.1| NNT-1
swall|Q8NEV9|Q8NEV9 IL-27 p28 subunit

Tableau 5. Tableau des identifiants des séquences de la sous famille des interleukines à chaînes courtes (IL-2) humaines.

identifiants
sw|P01588|EPO_HUMAN
sw|P04141|CSF2_HUMAN
sw|P01585|IL2_HUMAN
sw|P08700|IL3_HUMAN
sw|P05112|IL4_HUMAN
sw|P05113|IL5_HUMAN
sw|P35225|IL13_HUMAN
sw|P40933|IL15_HUMAN
sw|P49771|FL3L_HUMAN
sw|P21583|SCF_HUMAN
swall|P13232|IL7_HUMAN
swall|Q969D9|Q969D9 Thymic stromal lymphopoietin
sw|P15248|IL9_HUMAN
swall|P40225|TPO_HUMAN
gi|11093536|gb|AAG29348.1|AF254069_1 interleukin 21

Tableau 6. Tableau des identifiants des séquences de la sous famille des IL-10/IFN humaines à l'exception de l'homologue viral de l'IL-10 BCRF_EBV.

identifiants
sw|P22301|IL10_HUMAN
sw|P01563|INA2_HUMAN
swall|P01574|INB_HUMAN
sw|P01579|ING_HUMAN
sw|Q9UHD0|IL19_HUMAN
sw|Q9NYY1|IL20_HUMAN
sw|Q9GZX6|IL22_HUMAN
sw|Q96PD4|I17F_HUMAN
sw|Q13007|IL24_HUMAN
sw|Q9NPH9|IL26_HUMAN
swall|Q8IZJ0|Q8IZJ0 Interleukin 28A
swall|Q8IU54|Q8IU54 Interleukin 29
sw|P03180|BCRF_EBV
swall|Q8IZI9|Q8IZI9 Interleukin 28B
swall|Q8TAD2|Q8TAD2 Interleukin 27 precursor (IL-17D)

Chapitre 11. Caractérisation des signatures par algorithme génétique

11.1. Algorithme génétique pour l'extraction des motifs

La méthode de caractérisation décrite dans ce chapitre à pour objectif d'extraire un ensemble de signatures caractéristiques des trois sous-familles d'interleukines (Mikolajczak, Ramstein et al. 2004). Notre démarche consiste en une approche génétique pour la découverte des motifs hiérarchiques qui sont présents dans les signatures. L'algorithme suit une démarche descendante en s'appuyant sur les classes physicochimiques des acides aminés. Les signatures sont ensuite extraites au moyen d'un algorithme de découverte des itemsets fréquents dans lequel les motifs jouent le rôle des items. L'ensemble final de signatures est obtenu à l'issue d'une étape d'optimisation.

La recherche des motifs et des signatures révèle un caractère fortement combinatoire qui nous a amené à utiliser les algorithmes génétiques. La description hiérarchique des motifs permet une recherche descendante du motif le plus général vers le plus spécifique. Le motif général $\mu(k)$ est le motif de plus basse spécificité qu'on puisse trouver : $\phi(\mu(k)) = 0$. Il est défini par la série $\langle m_1 m_2 \ldots m_k \rangle$ telle que $m_i = \Omega \; \forall i, 1 < i < k$. Le premier algorithme génétique développé, découverteMotifs (Figure 28), procède en deux étapes.

La première étape vise à découvrir N k-motifs décrits par l'ensemble $\Gamma' = \{\Omega, \alpha, \beta, \gamma, \delta, \varepsilon, \xi, \eta, \theta\}$.

Fonction découverteMotifs (S, N, k) **retourne** Population

entrées

S , l'ensemble de séquences dont on recherche les motifs

N , le nombre de motifs

k , la taille des motifs

$\mu(k)$, le motif général : $\langle \Omega\Omega\cdots\Omega \rangle_k$

$p1, p2, p$: Population ;

#Phase I de la recherche de motifs

$p1 \leftarrow$ AGmotifs $(N, \mu(k), \Gamma', S)$;

Pour tout motif m dans $p1$ faire

#Phase II de la recherche de motifs

$p2 \leftarrow$ AGmotifs (N, m, Γ'', S) ;

#Extraire de $p2$ l'individu le plus apte et l'inclure dans p

$p \leftarrow p \cup \{meilleurIndividu(p2)\}$;

Fait ;

Retourner p ;

Figure 28. Algorithme découverteMotifs.

La valeur du paramètre k , qui correspond la longueur des k-motifs, est choisie au départ par l'utilisateur.

La deuxième étape consiste à reprendre les k-motifs avec l'ensemble des classes $\Gamma'' = \Gamma' \cup \omega$, ω ensemble des singletons, pour rechercher des motifs contenant des acides aminés. La démarche de généralisation-spécialisation des k-motifs permet de réduire l'espace de recherche et évite de voir l'ensemble des k-motifs converger vers des solutions sous-optimales induites par la rareté des motifs incluant les singletons.

Les deux phases de l'algorithme découverteMotifs sont séquentielles et utilisent le même algorithme génétique AGmotifs (Figure 29). AGmotifs présente deux particularités

essentielles. La première est que la population initiale est formée par les individus ou motifs obtenus par copie d'un motif germe. Dans la phase 1, le germe est le motif générique $\mu(k)$ tandis que dans la phase 2, le germe est un des N motifs extraits dans la phase précédente. La deuxième caractéristique est que l'opérateur de mutation est orienté selon le principe de la recherche descendante : une classe physico-chimique c est mutée en un des membres de c en phase 2. Par exemple, la classe physico-chimique α est mutée aléatoirement par la pseudo-classe $\{I\}, \{L\}$ ou $\{V\}$.

Fonction AGmotifs$(N, germe, alphabet, S)$ **retourne** Population

entrées

N, le nombre de motifs

S, l'ensemble de séquences

germe, le motif prototype pour l'initialisation de la première génération

alphabet, la description du motif selon Γ^1 ou Γ^2

parents, *population* : Population ;

#Initialisation de la population

Pour $i = 1$ à N faire $population \leftarrow population \cup \{clone(germe)\}$;

Faire

calculeAdaptation$(population, S)$;

$parents \leftarrow$ selection$(population)$;

$population \leftarrow$ reproduction$(parents, alphabet)$;

Jusqu'à adaptation$(population) \geq seuil$;

Retourner *population* ;

Figure 29. Algorithme AGmotifs

L'opérateur de recombinaison est le deuxième opérateur de reproduction après celui de mutation. Cet opérateur suit le principe de crossing-over. Le passage d'une génération à

l'autre suit la méthode du keep-best-reproduction (Wiese, Scott et al. 1998; Wiese and Goodwin 1999). Cette méthode conserve les meilleurs parents d'une génération à l'autre et améliore la convergence de l'algorithme génétique.

L'étape de sélection de l'algorithme AGmotifs est basée sur la roue de la fortune. L'individu survit d'autant plus sûrement qu'il est bien adapté. La définition de la fonction d'adaptation est un critère déterminant dans tout AG. Dans notre application, un motif est intéressant s'il possède un support suffisant et une spécificité importante. Comme nous l'avons décrit précédemment, ces deux critères ont tendance à avoir des comportements opposés. En effet, un motif avec une spécificité faible aura de forte chance d'avoir un support important. A l'inverse, un motif très spécifique à peu de chance d'être identiquement conservé entre les séquences des membres de la famille d'intérêt (Tableau 3).

La fonction d'adaptation avec laquelle nous avons obtenu les meilleurs résultats sur notre jeu de données est la suivante :

$$a(m) = \begin{cases} 0 & si\ m\ tel\ que\ support(m) < \tau, \\ 1 & si\ m = \mu(k), \\ support(m) * \phi(m)^{\lambda} & sinon \end{cases}$$

La fonction d'adaptation $a(m)$ est proportionnelle au support de m et s'accroît avec la spécificité du motif. Le paramètre λ permet de moduler l'importance de la spécificité par rapport au support. Le paramètre τ correspond au support minimal en deçà duquel le motif est inadapté. Il permet d'éliminer les motifs de fortes spécificités mais avec un support dérisoire. La sélection des individus est effectuée avec une probabilité $a(m)$. Si elle est nulle le motif m disparaît à la génération suivante. Pour éviter qu'il ne disparaisse prématurément dès la première génération, nous avons attribué une adaptation non nulle au motif général $\mu(k)$.

L'algorithme AGmotifs a été adapté de façon à ce que les motifs délivrés par l'AG ne soient par les N meilleurs individus de la dernière génération, mais les N meilleurs individus distincts obtenus pour toutes les générations confondues. Ce principe permet d'éviter les doublons mais n'évite pas le risque de recouvrement des motifs hiérarchiques, par exemple les motifs $\alpha\varepsilon\beta$ et $V\Omega\gamma$. Cependant les cas de redondances de généralisation-spécialisation ne concernent qu'une infime proportion des k-motifs, en particulier ceux de petites tailles. De

plus, ils sont éliminés au cours des étapes suivantes de recherche des signatures, par la contrainte de non recouvrement liée a la définition même des signatures, et au cours de la phase d'optimisation des signatures, par l'élimination des signatures de moindre spécificité.

11.2. Algorithme génétique d'extraction des signatures et raffinement

La découverte d'un ensemble de signatures se divise en deux étapes :

Tout d'abord un ensemble de signatures Σ est extrait à partir de l'ensemble des motifs M obtenu par l'algorithme AGmotifs (Figure 29). Dans un second temps, l'ensemble Σ est réduit pour ne retenir que les signatures les plus représentatives de l'ensemble d'apprentissage en terme de spécificité et de support.

La première étape d'extraction des signatures consiste à adapter l'algorithme Winepi (Manilla and Toivonen 1996) pour déterminer un ensemble de signatures candidates de support minimal τ. Cet algorithme est adapté de l'algorithme Apriori (Agrawal and Srikant 1995). Il explore tous les itemsets fréquents, en partant de ceux de taille $k = 1,2,\ldots$ jusqu'à ce qu'il n'existe plus de signatures de support supérieur ou égal à τ. A chaque étape, nous obtenons ainsi un ensemble Σ^k de k-signatures que nous mémorisons, jusqu'à l'arrêt de l'algorithme, marqué par l'ensemble vide $\Sigma^{k\max}$. Nous appellerons ExtraireSignaturesCandidates(S, M) l'algorithme qui délivre à partir de S de M un ensemble de signatures $\Sigma = \bigcup_{k=1}^{k\max-1} \Sigma^k$.

La deuxième étape consiste à modéliser l'ensemble S par un sous-ensemble Σ' de Σ. Cette phase permet de représenter une famille de séquences connues sous une forme condensée. Ce modèle servira à classer des séquences inconnues dans cette famille ou à les rejeter. L'ensemble Σ' doit satisfaire les contraintes suivantes :

- La contrainte de complétude : chaque séquence $s_i \in S$ vérifie au moins une signature du sous ensemble Σ' de Σ. Le support de Σ' doit être égal au cardinal X de la famille S.

- La contrainte de cardinalité : le cardinal de Σ' doit être la plus petit possible (idéalement à 1).

- La contrainte de spécificité : la spécificité de Σ', définie comme la somme des spécificités de ces éléments, doit être la plus grande possible.

L'algorithme « ExtraireSignaturesCandidates » fournit une solution de grande cardinalité qui ne garantit pas la vérification de la première condition (sauf à fixer le support minimal τ à X, ce qui induirait des signatures de faible spécificité).

Dans la pratique, on supposera que la contrainte de complétude est vérifiée par Σ. si tel n'est pas le cas, il suffit de compléter Σ par la signature universelle $\langle \mu(k) \rangle$. Soit $\Sigma(\varphi)$ l'ensemble des signatures de Σ de spécificité supérieure ou égale à φ. Un premier prétraitement « éliminerRedondances » (Figure 30) consiste à rechercher la spécificité maximale $\varphi_{\max}$ telle que le support de $\Sigma(\varphi_{\max})$ soit égal au cardinal X, puis éliminer de cet ensemble les signatures redondantes et de moindre spécificité.

procédure éliminerRedondances (Σ, φ, S)
entrées
S, ensemble des séquences
Σ, l'ensemble des signatures candidates
φ, le seuil de spécificité maximale autorisé
$\Sigma \leftarrow \Sigma(\varphi)$;

Pour toute signature $\sigma \in \Sigma$ faire
    Soit $E(\sigma) = \{s_i \in S \text{ vérifiant } \sigma\}$;
    Rechercher dans Σ s'il existe une signature $\sigma' \neq \sigma$
    Telle que $E(\sigma) = E(\sigma')$ et $\varphi(\sigma') > \varphi(\sigma)$.
    Si oui, alors $\Sigma \leftarrow \Sigma - \{\sigma\}$;
fait ;

Figure 30. Algorithme éliminerRedondances

Un deuxième traitement permet de retenir les signatures qui couvrent au mieux l'ensemble S. Cette méthode est basée sur une heuristique, appelée rechercheParCardinal, (Figure 31) qui consiste à prendre la signature de support maximal, puis à retenir la signature dont l'union avec la signature précédente ait un support maximal, et ainsi de suite jusqu'à ce que l'ensemble des signatures Σ_{rc}, formé par la méthode rechercheParCardinal, ait un support égal à X. Cette heuristique privilégie la contrainte de cardinalité. Cependant elle ne donne pas le plus petit ensemble de signatures qui vérifie S .

fonction rechercheParCardinal (Σ, S) **retourne** ensembleSignatures

entrées

Σ, l'ensemble des signatures candidates

S, l'ensemble de séquences

$X = |S|$

$\Sigma_{rc} \leftarrow \phi$;

Tant que $support(\Sigma_{rc}) \neq X$ faire

Soit $E = \{\Sigma_{rc} \cup \{\sigma_i\}, \sigma_i \in \Sigma\}$;

Soit $E' = \{e \in E$ tel que le support de e soit maximal dans $S\}$;

Soit $e_{\max} \in E'$ l'ensemble de signatures de spécificité maximale ;

Soit $\sigma_{\max}$ la signature telle que $e_{\max} = \{\Sigma_{rc} \cup \{\sigma_{\max}\}\}$;

$\Sigma_{rc} \leftarrow \Sigma_{rc} \cup \{\sigma_{\max}\}$;

$\Sigma \leftarrow \Sigma - \{\sigma_{\max}\}$;

fait ;

retourner Σ_{rc} ;

Figure 31. Algorithme rechercheParCardinal. Σ_{rc} : ensemble des signatures retournées par la méthode rechercheParCardinal.

Si la valeur du cardinal $c\max = |\Sigma_{rc}|$ est élevée, nous proposons de faire appel à un algorithme génétique comparable à AGmotifs. Cette fois ci, l'algorithme « AGsignature » (Figure 32) recherche des ensembles Σ' de cardinalités k décroissantes, à compter de $k = cmax$. L'algorithme s'arrête lorsqu'on ne trouve plus d'individus dont le support est supérieur ou égal à $\widetilde{X} = X$ pour assurer la contrainte de complétude. A l'initialisation, les individus de la première génération sont des copies du germe Σ_{rc}. Lorsqu'un individu est jugé suffisamment apte, l'algorithme génétique poursuit sa recherche en décrémentant k d'une unité. Cette procédure est réalisée au moyen de la méthode « optimise » qui élimine dans chaque individu de la population une des k signatures qu'il contient. La signature sélectionnée est celle qui assure au nouvel individu ainsi formé un support maximal dans S.

L'opérateur de mutation de l'algorithme AGsignature remplace aléatoirement une des signatures d'un individu par une signature aléatoire de Σ qui n'appartient pas à cet individu.

La recombinaison utilise une technique de *crossing-over*. La fonction d'adaptation que nous avons utilisée est la suivante :

$$a(individu) = \begin{cases} support(individu) & si\ support(individu) < \widetilde{C} \\ \lambda\phi(individu) + \gamma & sinon \end{cases}$$

Cette définition tend à rechercher des individus qui respectent la contrainte de complétude, puis à tendre vers une solution de spécificité maximale pour répondre à la contrainte de spécificité. La contrainte de cardinalité est vérifiée par l'individu de cardinalité minimale dans *résultat*. Les paramètres d'ajustements $\lambda \geq 1$ *et* $\nu \geq \widetilde{X}$ permettent d'accroître l'efficacité de l'algorithme génétique.

fonction AGsignatures $N, \Sigma, \Sigma_{rc}, \tilde{X}$ retourne ensembleSignatures

entrées

N , la taille de la population

Σ , l'ensemble des signatures candidates

Σ_{rc} , l'ensemble des signatures obtenues par la méthode rechercheParCardinal

$\tilde{X}$, le plus petit support recherché

$parents, population, résultat$: Population ;

#initialisation de la population

Pour $i = 1$ à N faire $population \leftarrow population \cup \{clone(\Sigma_{rc})\}$;

$k \leftarrow |\Sigma_{rc}|$; #initialisation à $cmax$

Faire

Faire

calculeAdaptation $(population)$;

$parents \leftarrow$ selection $(population)$;

$population \leftarrow$ reproduction $(parents)$;

Jusqu'à adaptation $(population) \geq seuil$;

Soit ind , l'individu le plus adapté dans $population$

Si le support $(ind) \geq \tilde{X}$ alors $résultat \leftarrow résultat \cup \{ind\}$;

#initialisation de la première génération suivante

$population \leftarrow$ optimise $(population)$;

$k \leftarrow k - 1$;

Tant que support $(ind) \geq \tilde{X}$;

Retourner $résultat$;

Figure 32. Algorithme AGsignature

11.3. Tests de performance

11.3.1. Test de performance avec un ensemble de 45 séquences.

A partir des 45 séquences d'interleukines, l'algorithme AGmotifs a extrait 300 motifs de taille $k = 3$ à 8. Ensuite l'ensemble des signatures candidates avec un seuil de support égal à $\tau = 6$ pour les deux recherches. Le cardinal d'Σ est égal à 40365 avant filtrage et 1654 après appel de la méthode « élimineRedondances ». Après filtrage, l'ensemble de signatures candidates est réduit à 719. L'heuristique « rechercheParCardinal » détermine un ensemble de signatures Σ_{rc} de 12 signatures et de spécificité moyenne 22,55 (Tableau 7). Seulement 1.75% des contre-exemples de la base SCOP vérifient Σ_{rc}. Le meilleur résultat a été obtenu en utilisant une procédure en trois phases, incluant le filtrage, l'heuristique « rechercheParCardinal » et l'algorithme génétique « AGsignature » : support de 45, spécificité de 23.45 et un taux de faux positifs de 1.38%. Le résultat est optimum pour $\varphi_{\max} = 21.09$.

Tableau 7. Tableau des résultats obtenus avec un ensemble d'apprentissage de 45 interleukines : $\varphi_{\max} = 21.09$ et nombre de signatures = 1654 cardinal = 45 sur 40365 signatures Nombre de signatures après filtrage = 719 cardinal = 45. Er : optimisation EliminerRedondances. rc : .optimisation rechercheParCardinal. ag : optimisation AGsignature

Card	Méthode	Support	Spécificité	Fréquence SCOP (%)
12	$\mathrm{Er}(\varphi = \varphi_{\max})$+rc	45	22.55	1.75
12	$\mathrm{Er}(\varphi = \varphi_{\max})$+rc+ag$(\tilde{c} = c)$	45	23.45	1.38
11	$\mathrm{Er}(\varphi = \varphi_{\max})$+rc+ ag$(\tilde{c} = c)$	45	22.99	1.53
10	$\mathrm{Er}(\varphi = \varphi_{\max})$+rc+ag$(\tilde{c} < c)$	44	22.66	1.48
9	$\mathrm{Er}(\varphi = \varphi_{\max})$+rc+ ag$(\tilde{c} < c)$	43	21.62	1.36
6	$\mathrm{Er}(\varphi = 0)$+rc	45	17.16	13.05

Chapitre 12. Méthode déterministe « motifSVM » pour la classification des cytokines par la méthode SVM.

12.1. Algorithme de découverte de motifs

Cette deuxième méthode repose sur une modélisation des familles d'interleukines par un ensemble M de motifs intégrant les propriétés physico-chimiques des résidus. L'algorithme suit le paradigme de la classification hiérarchique ascendante. L'ensemble M définit un espace de représentation des séquences : chaque séquence est transformée en un vecteur indiquant la présence ou l'absence de chaque motif appartenant à M. La vectorisation des séquences est utilisée pour l'apprentissage de la classification des familles d'intérêt par les SVM (Mikolajczak, Ramstein et al. 2004; Mikolajczak, Ramstein et al. 2004; Mikolajczak, Ramstein et al. 2004; Mikolajczak, Ramstein et al. 2004).

La recherche de motifs hiérarchiques procède en deux étapes : l'extraction de motifs germes et la génération des motifs hiérarchiques.

La première étape consiste à extraire des motifs germes à partir de la famille S. Un motif germe est un motif qui ne présente pas de minorants. Ils sont formés uniquement de classes de singletons (résidus). Une façon d'obtenir la liste des motifs germes est de relever l'ensemble des k-sous-séquences potentiellement intéressantes. Cette première étape consiste à croiser chaque k-sous-séquences avec une autre : chaque couple de k-sous-séquences fournit un motif. Si ce dernier possède un support suffisant, les k sous-séquences vérifiant ce motif sont retenues. Avec un support minimal de 3 nous avons pu réduire ainsi le nombre de motifs germes d'un facteur 8.

La seconde étape opère un appariement des motifs pour déterminer leurs bornes supérieures. Cette méthode consiste à généraliser les motifs afin de rechercher des caractéristiques de support suffisamment représentatives. La définition de borne supérieure permet de définir le motif commun le plus spécifique à partir de deux motifs. Par itérations successives, il est possible d'agréger des motifs afin d'obtenir des motifs de support supérieur et de spécificité maximum. L'algorithme « ExtraireMotifs » (Figure 33), s'inspire de la technique de

classification hiérarchique ascendante pour former des clusters de motifs généraux à partir de motifs germes composés uniquement de singletons.

Algorithme ExtraireMotifs

entrées

M , l'ensemble des motifs germes

$supMin$, le seuil de support minimal recherché

$speMin$, le seuil de spécificité minimale recherchée

sortie

E , l'ensemble des motifs hiérarchiques

$E = \{m \in M | support(m) \geq supMin \text{ et } \phi(m) \geq speMin\}$;

Répéter

Soit m^1 et m^2 la paire de motifs de M telle que :

1. $m^{1,2} = \sup(m^1, m^2)$
2. $\phi(m^{1,2}) \geq \phi(m^{i,j})$ pour tout m^i et m^j dans M

$M \leftarrow M - \{m^1, m^2\}$;

$M \leftarrow M \cup \{m^{1,2}\}$;

si $support(m^{1,2}) \geq supMin$ et $\phi(m^{1,2}) \geq speMin$

alors $E \leftarrow E \cup \{m^{1,2}\}$;

jusqu'à $cardinal(M) = 1$ ou $\phi(m^{1,2}) < speMin$;

Figure 33. Algorithme ExtraireMotifs

La table (Tableau 8) représente les motifs hiérarchiques de taille $k = 4$ relatifs aux séquences aux séquences HIWY, HIDY, KLTY, HVSG et DARG. Les motifs m^1 à m^5 sont les motifs germes extraits des sous-séquences de taille 4 dans le jeu de séquences précédent. La figure (Figure 34) montre la construction hiérarchique des motifs par l'algorithme « ExtraireMotifs ». Par exemple si le support minimal fixé est de 3 et une spécificité de 3.0, seul le motif m^7 serait retenu.

Tableau 8. Motifs de taille $k = 4$ extraits par l'algorithme découverteMotifs.

Motif	Chaîne	Support	Spécificité
m^1	HIWY	1	14.25
m^2	HIDY	1	12.83
m^3	KLTY	1	11.44
m^4	HVSG	1	11.79
m^5	DARG	1	10.96
$m^6 = m^{1,2}$	HI ε Y	2	10.64
$m^7 = m^{3,6}$	$\xi\alpha\varepsilon$ Y	3	7.55
$m^8 = m^{4,5}$	$\delta\eta\varepsilon$ Y	2	5.26
$m^9 = m^{7,8}$	$\delta\Omega\varepsilon\gamma$	5	2.56

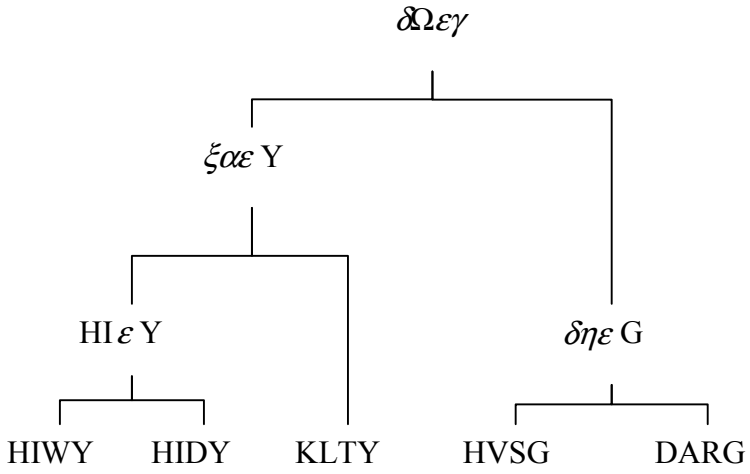


Figure 34. Hiérarchie de motifs à partir de 5 motifs germes

L'algorithme diffère dans sa finalité des techniques usuelles d'extraction de motifs. Alors que ces dernières visent à découvrir un nombre restreint de motifs avec le support le plus large

possible, nous cherchons au contraire un ensemble élevé M de motifs au support très variable. Le caractère hiérarchique de l'algorithme nous permet de trouver un nombre conséquent de motifs tout en s'affranchissant des problèmes de combinatoire.

12.2. Vectorisation

Pour utiliser les machines à vecteurs de support, nous allons transformer une séquence en vecteur booléen de dimension n (n désignant le cardinal de M). L'élément de rang i est vrai si et seulement si le motif de rang i dans M est présent dans la séquence. Il est à noter que cette vectorisation s'effectue en $O(N)$, où N désigne la taille de la séquence. Une autre méthode de vectorisation aurait consisté à dénombrer les occurrences des motifs hiérarchiques. Si l'apprentissage est relativement complexe, la classification est peu coûteuse en temps d'exécution. La table ci-dessous (Tableau 9) montre les valeurs des vecteurs associés aux séquences de l'exemple précédent en considérant l'ensemble des motifs données dans la table MOTIFS (support et spécificité minimaux fixés à 0).

Tableau 9. Représentation vectorielle des séquences à partir des motifs extraits. Le vecteur représente neuf valeurs booléennes (1 pour vrai, 0 pour faux) correspondant à la présence du motif $m^i, i = 1 \grave{a}\ 9$, dans la séquence considérée.

Séquence	Représentation vectorielle
HIWY	100001101
HIDY	010001101
KLTY	001000101
HVSG	000100011
DARG	000010011

12.3. Tests de performance

La classification des protéines passe par une première étape de vectorisation suivie de la prédiction proprement dite par SVM. Dans notre application, la transformation des séquences s'opère par la détection des motifs retenus par l'algorithme « ExtraireMotifs ». Les séquences étant représentées par des vecteurs booléens de taille fixe, il est possible de définir le produit scalaire entre deux séquences s et s' par le nombre de motifs communs dans les vecteurs booléens x et x' associés respectivement à s et s'. Le choix de la fonction noyau a été dicté par certaines propriétés spécifiques à notre espace de redescription. En effet, soit O le vecteur nul, la fonction noyau doit vérifier les propriétés suivantes :

- Si $x \approx O$ et $x' \approx O$ alors $K(x,x')$ doit avoir une valeur proche de la valeur maximale de la fonction noyau. En effet, dans ce cas, les deux séquences appartiennent toutes deux à la classe des contre-exemples.

- Si $x \approx O$ et x' appartient à la classe des exemples, alors la valeur de $K(x,x')$ doit être d'autant plus faible que l'exemple contient un nombre de motifs important.

Nous avons retenus les fonctions à bases radiales qui vérifient bien ces propriétés. Elles reposent sur le calcul de $\|x - x'\|$, contrairement aux fonctions polynomiales basées sur le calcul de $x.x'$. Les fonctions à base radiales donnent généralement de meilleurs résultats que les fonctions polynomiales et sigmoïdes.

Si l'utilisation des SVM ne pose pas de problème d'adaptation, il est cependant important d'observer qu'à l'origine les SVM définissent une frontière optimale entre deux classes d'individus d'intérêt égal. Dans notre type d'application, l'importance égale de ces deux classes soulève une difficulté. S'il est facile de définir la famille d'intérêt, la définition des contre-exemples est moins aisée. Une solution consiste à utiliser des représentants dans chaque superfamille définie dans la base SCOP. Ces tests montrent que les performances de la classification sont peu sensibles au tirage effectué, mais se dégradent lorsque le nombre de contre-exemples s'accroît par rapport au nombre d'exemples positifs. Une autre solution consiste à utiliser les SVM basés sur une classe unique (Scholkopf, Platt et al. 2001). Les résultats obtenus avec cette méthode se sont montrés inférieurs à ceux des SVM bi-classes. Nous interprétons ces résultats par le nombre restreint d'exemples présenté en apprentissage

ainsi qu'à leur grande dispersion (faible homologie intra-classe). Le classifieur SVM a été implémenté avec le logiciel libre libsvm (Chang and Lin 2001).

La comparaison des performances des différentes méthodes de classification est obtenue avec la technique de *leave one out* (Tableau 10). La ligne KNN présente les résultats obtenus avec la méthode des « k plus proches voisins » ($k = 3$). La notion de voisinage se réfère à la proximité entre spectres de chaîne : la mesure de similarité utilisée est le produit cartésien des vecteurs normalisés. Les SVMs à base de spectres de chaînes donnent de meilleurs résultats (SCSVM) que les KNNs, ce qui confirme l'intérêt des SVM. Les résultats sont identiques pour les fonctions linéaires et à bases radiales ; la seule différence consiste en une légère amélioration des performances sur SCOP pour la fonction à bases radiales. Notre technique MotifSVM surpasse largement la technique SCSVM, à condition d'optimiser le type des motifs. Le seuil de spécificité 14 apparaît comme le meilleur compromis : au-delà de cette valeur, on ne découvre pas assez de motifs sur certaines cytokines, en deçà, les motifs ne sont pas assez sélectifs. Un seuil de support minimal de 2 en *leave one out* (3 en apprentissage normal) est nécessaire pour obtenir de bons résultats dans notre jeu particulier d'applications. Il s'avère que les exemples présentés possèdent peu de séquences proches. Nous avons pu vérifier ce fait en calculant les valeurs de fonction noyau à partir de spectres de chaîne : la valeur de fonction noyau tend rapidement vers 0 à partir de la deuxième plus proche voisine.

Sur les 6615 séquences de SCOP (Tableau 11) , MotifSVM en a mal classé 8. Le faible pourcentage de faux positifs autorise l'emploi de motifSVM pour rechercher de nouveaux membres dans une base de données génomiques.

Tableau 10. Résultats de performance de classification en *leave one out*.

Méthode de classification	Vrais positifs (%)	Faux négatifs (%)	Vrais Négatifs (%)	Faux positifs (%)
KNN	88.9	11.1	73.3	26.7
SCSVM linéaire	84.4	15.6	88.9	11.1
SCSVM RBF	84.4	15.6	88.9	11.1
motifSVM 13	95.6	4.4	100	0
motifSVM 14	100	0	100	0
motifSVM 15	88.9	11.1	100	0

Tableau 11. Taux de faux-positifs dans la base SCOP

Classifieur	Taux de faux positifs dans SCOP (%)
SCSVM linéaire	4.32
SCSVM RBF	4.08
motifSVM 14	0.12

Chapitre 13. Comparaison des performances entre les deux méthodes

Deux méthodes ont été proposées. La première se base sur la recherche d'un ensemble de signatures (AGsignature) et la seconde repose sur une extraction des motifs caractéristiques accompagnée d'une classification SVM (MotifSVM).

Les résultats obtenus par nos deux méthodes de classification sur notre ensemble d'apprentissage ont montrés que la méthode motifSVM est plus performante que la méthode AGsignature. En effet, la méthode motifSVM à une sélectivité meilleure (Tableau 12).

Tableau 12. Comparaison des résultats de classification des méthodes AGsignature et motifSVM.

	Agsignature	motifSVM 14
Taux de faux positifs sur la base SCOP (%)	1.38	0.12
Support parmi les cytokines	45	45

La différence la plus notable entre les deux méthodes est que la première est stochastique et l'autre déterministe (Tableau 13). La deuxième méthode, dans sa version actuelle, ne peut s'employer que sur un ensemble d'apprentissage de taille raisonnable (de l'ordre d'une centaine de séquences). La méthode AGsignature est plus coûteuse en temps mais moins sensible à la taille de la famille étudiée.

L'extraction des motifs est commune aux deux techniques mais n'en poursuit pas moins des objectifs distincts. Dans le cas de la méthode MotifSVM, les motifs doivent être très spécifiques (spécificité supérieure à 14 dans nos expériences) pour que les SVMs donnent de bonnes performances. Dans le cas de la seconde méthode, cette contrainte est relâchée, puisque les motifs ne sont qu'une étape intermédiaire. C'est la signature qui importe, et donc une séquence de plusieurs motifs qui, pris individuellement, peuvent avoir une spécificité moindre.

Le paramétrage des deux méthodes joue donc un rôle important dans les résultats obtenus sur les motifs (Tableau 14).

Tableau 13. Tableau de comparaison entre les deux méthodes d'extractions motifSVM et AGsignature.

Critères de comparaison	motifSVM	AGsignature
Type méthode	Déterministe	Métaheuristique
Paradigme	Ascendant	Descendant
Extraction	motifs	signatures
Support des motifs	Gradient : faible à élevé	Moyen
Gradient de spécificité	élevée à faible	faible à moyenne

Tableau 14. Tableau des statistiques des deux jeux de motifs respectifs des méthodes motifSVM et AGsignature.

	Nombre de motifs	Support (min ; moy ; max)	Spécificité (min ; moy ; max)
motifSVM	682	2 ; 2.4 ; 6	14 ; 18.53 ; 26.77
AGsignatures	300	6 ; 19.8 ; 42	5.09 ; 7.46 ; 12.56

Nous avons interverti les deux jeux de motifs obtenus dans les deux méthodes pour d'une part trouver les signatures des motifs provenant de motifSVM, et d'autre part vectoriser les séquences à partir des motifs de « AGmotifs ».

L'ensemble des motifs de motifSVM n'est pas directement adapté à la méthode AGsignature. Nous avons retenu les 297 motifs issus de MotifSVM dont les supports sont supérieurs ou égaux à 3 et de spécificité comparables à celles des motifs de taille 8 extraits par la méthode AGsignature (comprise entre 6.2 et 12 .6). Ce jeu donne 1063 signatures, soit 339 signatures de support Supérieur ou égal à 7 et de spécificité supérieure ou égale à 14.1, soit 302 signatures après filtrage. La procédure « rechercheParCardinal » donne 11 signatures vérifiant les 45 cytokines, avec une spécificité moyenne de 15.66 et un taux de 7.48% de faux positifs dans SCOP. Ce résultat est moins performant que celui obtenu avec AGsignature (1.53%). La meilleure performance de AGmotifs peut provenir du jeu de motifs plus large retenu, incluant des motifs de longueurs 3, 4 et 8. Nous avons ensuite appliqué un algorithme génétique pour optimiser ce résultat. Pour k=11 signatures, nous obtenons un support de 45, une spécificité moyenne de 16.65 et un taux de 6.45% de faux positifs dans SCOP. Il est intéressant de noter que les deux méthodes aboutissent à un même nombre minimal de signatures pour couvrir la famille de cytokines.

La seconde série de test consiste à utiliser les motifs de AGsignature et la classification SVM (Tableau 15). Il apparaît tout de suite que les motifs de faibles spécificités favorisent l'apparition des faux-positifs. Nous avons donc limité la taille des vecteurs en retenant les 95 motifs de spécificités supérieures ou égales à 8 (sur les 300 motifs de départ). Le résultat est moins performant que ceux obtenus avec la méthode MotifSVM. Il serait certainement possible d'améliorer de façon sensible ces résultats en extrayant un nombre plus important de motifs de taille 8 et en abaissant le seuil de support minimal.

Tableau 15. Tableau des résultats lors de l'utilisation des motifs AGsignature et classification SVM. Apprentissage sur l'ensemble d'apprentissage 45 séquences d'interleukines / 45 séquences de contre-exemples.

Origine des motifs	Nombre total motifs	Vrai positifs (%)	Faux négatifs (%)	Vrais négatifs (%)	Faux positifs (%)
AGsignature	95	88.9	11.1	91.1	8.9
motifSVM	682	100	0	100	0

Conclusion et Perspectives

Les formes solubles des interleukines des deux sous-familles structurales IL-2 et IL-6 possèdent un repliement tridimensionnel consensus de quatre hélices α impliquées dans une topologie « up-up-down-down ». Cette conformation est aussi retrouvée pour quatre des six hélices α des membres de la troisième sous-famillle des IL-10/IFNs. Cet état de conservation n'est pas observé au niveau des séquences primaires des interleukines qui présentent des taux de similarité intra et inter classes très faibles. La question qui nous a été donnée de résoudre pendant cette thèse a été de d'obtenir des modèles de classification des trois sous-familles d'interleukines, à partir de leur séquences primaires, capables de discriminer les candidats potentiels pour de nouvelles interleukines au sein des bases de données génomiques publiques. Ce travail a contribué à l'élaboration de deux méthodes de caractérisation des membres des sous-familles des interleukines IL-2, IL-6 et IL-10/IFNs humaines.

Notre première stratégie de classification, AGsignature, est basée sur la caractérisation des signatures au moyen d'algorithmes génétiques. Elle nous a permis de lever le problème des très faibles taux de similarité entre les séquences des interleukines et d'obtenir un ensemble de signatures spécifiques de notre famille. La première étape de cette méthode consiste en la découverte de motifs au moyen de deux niveaux hiérarchiques de classification. Cette étape pourrait être améliorée en autorisant les motifs à croître en taille. La définition de la spécifité est perfectible. A l'heure actuelle elle ne peut être considérée comme une estimation de la probabilité de la présence du motif. La contrainte de complétude peut s'avérer coûteuse en nombre de signatures nécessaires pour caractériser une famille. En effet il est possible d'augmenter la spécificité de l'ensemble des signatures en acceptant un nombre limité de faux négatifs.

Notre deuxième statégie, motifSVM, est construite autour d'une classification hiérarchique ascendante des motifs spécifiques à notre ensemble d'interleukines. Nous avons proposé un paramétrage simple du degré de spécificité et nous avons montré qu'une valeur de spécificité moyennement haute ($\phi(m)=14$) donne les meilleures performances par rappoer à nos données. La capacité des SVM à gérer des espaces de grande dimension permet d'obtenir une sélectivité très forte avec une valeur de 0.12% de faux positifs issus de SCOP. L'algorithme peut être amélioré afin d'optimiser l'ensemble des motifs obtenus. Une proportion non

négligeable d'entre eux peut être certainement éliminée sans nuire à la performance du classifieur. Il est aussi possible de filtrer plus finement les motifs extraits et de procéder ensuite à l'étude de leurs co-occurrences dans les séquences. La stratégie motifSVM possède une meilleure performance de classification que la stratégie AGsignature sur notre ensemble d'interleukines qui possède une spécificité plus faible équivalente à 1.38% de faux positifs dans SCOP. En comparant la méthode motifSVM avec la méthode des spectres de chaînes (Leslie, Eskin et al. 2002), nous avons pu montrer que nous obtenions une meilleure performance de classification sur nos familles d'interleukines.

Au cours de ce travail, nous nous sommes exercés à tester nos méthodes sur un seul banc d'essai constitué par les trois familles d'interleukines qui nous intéressent. Cependant les méthodes AGsignature et motifSVM peuvent s'adapter aisément à la classification d'autres familles de protéines. Si la phase d'apprentissage des signatures est coûteuse en temps pour la première méthode, l'étape de prédiction sur des séquences candidates se fait rapidement en fonction de la détection de la présence de la signature dans la séquence. En revanche, la deuxième méthode a une complexité spatiale plus critique, elle s'applique donc à des familles de séquences de taille raisonnable.

Il est envisageable d'élargir la classification des acides aminés à des superclasses qui couvriront plusieurs classes physicochimiques. L'intérêt de ce travail sera d'étudier si la nouvelle classification améliore la performance des classifieurs. En effet l'ajout de nouveaux niveaux de généralisation des motifs permettra de décrire les positions dans les motifs par des classes différentes de la classe générale Ω. Cependant il existe un risque que les membres de ces nouvelles classes ne partagent plus de propriétés physicochimiques communes.

Une autre perspective consiste à améliorer l'étape de prédiction d'un exemple au moyen d'une mesure de confiance sur la classe prédite. Dans le cadre de la stratégie motifSVM, la mesure de confiance sur la prédiction peut être estimée dans l'espace de redescription des exemples en fonction de la proximité de l'exemple à l'hyperplan optimal. Une seconde voie consiste à obtenir une mesure de confiance finale sur la classe prédite d'un exemple à partir des prédictions effectuées par des modèles obtenus au moyen de méthodes de classification différentes.

Le travail effectué et les modèles obtenus par nos deux méthodes ont donc ouvert la voie à une exploration des séquences dans les bases de données génomiques dans le but de

rechercher de nouvelles interleukines. Dans cet objectif, nous avons fourni un travail important pour le développement d'une plateforme originale d'extraction des séquences. L'application est constituée d'un premier module servant au rappatriemnent, à la gestion et au traitement de séquences génomiques ESTs « Expressed Sequence Tags » présentes dans la base de données Unigene (Schuler 1997). Un deuxième module sert à l'extraction des séquences putatives d'interleukines à partir de nos modèles de classification. La sélectivité élevée de la méthode motifSVM a permis d'effectuer en 55 heures et 35 minutes la classification après traduction des ESTs de 3.809.034 séquences humaines et de retenir 7.771 séquences d'EST, soit 0,16 % de notre base de données, parmi lesquelles nous trouvons des séquences appartenant à des interleukines déjà identifiées. A la suite de ces travaux, les efforts d'analyses doivent se porter sur la dernière fraction de séquences inconnues candidates. La validation de ces séquences fait actuellement l'objet d'une étude menée par un doctorant au sein du même laboratoire.

Liste des figures

Liste des tableaux

Bibliographie

Aggarwal, S., M. H. Xie, et al. (2001). "Acinar cells of the pancreas are a target of interleukin-22." J Interferon Cytokine Res **21**(12): 1047-53.

Agrawal, R., T. Imielinski, et al. (1993). Mining Association Rules between Sets of Items in Large Databases. Proceedings of the 1993 ACM SIGMOD International Conference on Management of Data.

Agrawal, R. and R. Srikant (1994). Fast Algorithms for Mining Association Rules. Proc. 20th Int. Conf. Very Large Data Bases (VLDB), Morgan Kaufman.

Agrawal, R. and R. Srikant (1995). Mining Sequential Patterns. Eleventh International Conference on Data Engineering. Taipei, Taiwan., IEEE Computer Society Press.**:** 3-14.

Altschul, S. F., W. Gish, et al. (1990). "Basic local alignment search tool." J Mol Biol **215**(3): 403-10.

Altschul, S. F., T. L. Madden, et al. (1997). "Gapped BLAST and PSI-BLAST: a new generation of protein database search programs." Nucleic Acids Res **25**(17): 3389-402.

Aman, M. J., N. Tayebi, et al. (1996). "cDNA cloning and characterization of the human interleukin 13 receptor alpha chain." J Biol Chem **271**(46): 29265-70.

Andreakos, E. T., B. M. Foxwell, et al. (2002). "Cytokines and anti-cytokine biologicals in autoimmunity: present and future." Cytokine Growth Factor Rev **13**(4-5): 299-313.

Antoni, C. and B. Manger (2002). "Infliximab for psoriasis and psoriatic arthritis." Clin Exp Rheumatol **20**(6 Suppl 28): S122-5.

Asadullah, K., W. Sterry, et al. (1998). "IL-10 is a key cytokine in psoriasis. Proof of principle by IL-10 therapy: a new therapeutic approach." J Clin Invest **101**(4): 783-94.

Bach, E. A., J. W. Tanner, et al. (1996). "Ligand-induced assembly and activation of the gamma interferon receptor in intact cells." Mol Cell Biol **16**(6): 3214-21.

Baier, M., A. Werner, et al. (1995). "HIV suppression by interleukin-16." Nature **378**(6557): 563.

Balkwill, F. R. and F. Burke (1989). "The cytokine network." Immunol Today **10**(9): 299-304.

Ball, C., S. Vigues, et al. (2001). "Rat interleukin-10: production and characterisation of biologically active protein in a recombinant bacterial expression system." Eur Cytokine Netw **12**(1): 187-93.

Bamborough, P., C. J. Hedgecock, et al. (1994). "The interleukin-2 and interleukin-4 receptors studied by molecular modelling." Structure **2**(9): 839-51.

Bannan, J., K. Visvanathan, et al. (1999). "Structure and function of streptococcal and staphylococcal superantigens in septic shock." Infect Dis Clin North Am **13**(2): 387-96, ix.

Bartley, T. D., J. Bogenberger, et al. (1994). "Identification and cloning of a megakaryocyte growth and development factor that is a ligand for the cytokine receptor Mpl." Cell **77**(7): 1117-24.

Baum, L. and T. Petrie (1966). "Statistical inference for probabilistic functions of finite state Markov chains." Annals of Mathematical Statistics **37**: 1554-1563.

Baum, L., T. Petrie, et al. (1970). "A maximization technique occuring in the statistical analysis of probabilistic functions of Markov chains." Annals of Mathematical Statistics, **41**(1): 164-171.

Bazan, J. F. (1990). "Structural design and molecular evolution of a cytokine receptor superfamily." Proc Natl Acad Sci U S A **87**(18): 6934-8.
Bazan, J. F. (1991). "Neuropoietic cytokines in the hematopoietic fold." Neuron **7**(2): 197-208.
Bazan, J. F. (1992). "Unraveling the structure of IL-2." Science **257**(5068): 410-3.
Benson, D. A., I. Karsch-Mizrachi, et al. (2004). "GenBank: update." Nucleic Acids Res **32 Database issue**: D23-6.
Benveniste, E. N. (1998). "Cytokine actions in the central nervous system." Cytokine Growth Factor Rev **9**(3-4): 259-75.
Bleesing, J. J., S. E. Straus, et al. (2000). "Autoimmune lymphoproliferative syndrome. A human disorder of abnormal lymphocyte survival." Pediatr Clin North Am **47**(6): 1291-310.
Blumberg, H., D. Conklin, et al. (2001). "Interleukin 20: discovery, receptor identification, and role in epidermal function." Cell **104**(1): 9-19.
Boeckmann, B., A. Bairoch, et al. (2003). "The SWISS-PROT protein knowledgebase and its supplement TrEMBL in 2003." Nucleic Acids Res **31**(1): 365-70.
Bondeson, J., B. Foxwell, et al. (1999). "Defining therapeutic targets by using adenovirus: blocking NF-kappaB inhibits both inflammatory and destructive mechanisms in rheumatoid synovium but spares anti-inflammatory mediators." Proc Natl Acad Sci U S A **96**(10): 5668-73.
Bone, R. C. (1991). "The pathogenesis of sepsis." Ann Intern Med **115**(6): 457-69.
Bone, R. C. (1994). "Gram-positive organisms and sepsis." Arch Intern Med **154**(1): 26-34.
Bosenberg, M. W. and J. Massague (1993). "Juxtacrine cell signaling molecules." Curr Opin Cell Biol **5**(5): 832-8.
Boulanger, M. J., D. C. Chow, et al. (2003). "Hexameric structure and assembly of the interleukin-6/IL-6 alpha-receptor/gp130 complex." Science **300**(5628): 2101-4.
Bowie, A. and L. A. O'Neill (2000). "The interleukin-1 receptor/Toll-like receptor superfamily: signal generators for pro-inflammatory interleukins and microbial products." J Leukoc Biol **67**(4): 508-14.
Brandhuber, B. J., T. Boone, et al. (1987). "Three-dimensional structure of interleukin-2." Science **238**(4834): 1707-9.
Bravo, J. and J. K. Heath (2000). "Receptor recognition by gp130 cytokines." Embo J **19**(11): 2399-411.
Breiman, L., J. H. Friedman, et al. (1984). Classification and Regression Trees, Wadsworth International Group: Belmont, California.
Brombacher, F., R. A. Kastelein, et al. (2003). "Novel IL-12 family members shed light on the orchestration of Th1 responses." Trends Immunol **24**(4): 207-12.
Buchli, P. and T. Ciardelli (1993). "Structural and biologic properties of a human aspartic acid-126 interleukin-2 analog." Arch Biochem Biophys **307**(2): 411-5.
Callard, R. E., D. J. Matthews, et al. (1996). "IL-4 and IL-13 receptors: are they one and the same?" Immunol Today **17**(3): 108-10.
Cameron, S. B., M. C. Nawijn, et al. (2001). "Regulation of helper T cell responses to staphylococcal superantigens." Eur Cytokine Netw **12**(2): 210-22.
Campbell, I. K., K. O'Donnell, et al. (2001). "Severe inflammatory arthritis and lymphadenopathy in the absence of TNF." J Clin Invest **107**(12): 1519-27.
Caput, D., P. Laurent, et al. (1996). "Cloning and characterization of a specific interleukin (IL)-13 binding protein structurally related to the IL-5 receptor alpha chain." J Biol Chem **271**(28): 16921-6.
Chang, C., E. Magracheva, et al. (2003). "Crystal structure of interleukin-19 defines a new subfamily of helical cytokines." J Biol Chem **278**(5): 3308-13.

Chang, C. C. and C. J. Lin (2001). LIBSVM: a library for support vector machines.
Chatham, W. W. and W. D. Blackburn, Jr. (1993). "Fixation of C3 to IgG attenuates neutrophil HOCl generation and collagenase activation." J Immunol **151**(2): 949-58.
Chen, Q., N. Ghilardi, et al. (2000). "Development of Th1-type immune responses requires the type I cytokine receptor TCCR." Nature **407**(6806): 916-20.
Cohen, F. E. and I. D. Kuntz (1987). "Prediction of the three-dimensional structure of human growth hormone." Proteins **2**(2): 162-6.
Cookson, W. (1999). "The alliance of genes and environment in asthma and allergy." Nature **402**(6760 Suppl): B5-11.
Cosenza, L., A. Rosenbach, et al. (2000). "Comparative model building of interleukin-7 using interleukin-4 as a template: a structural hypothesis that displays atypical surface chemistry in helix D important for receptor activation." Protein Sci **9**(5): 916-26.
Cosenza, L., E. Sweeney, et al. (1997). "Disulfide bond assignment in human interleukin-7 by matrix-assisted laser desorption/ionization mass spectroscopy and site-directed cysteine to serine mutational analysis." J Biol Chem **272**(52): 32995-3000.
Czupryn, M. J., J. M. McCoy, et al. (1995). "Structure-function relationships in human interleukin-11. Identification of regions involved in activity by chemical modification and site-directed mutagenesis." J Biol Chem **270**(2): 978-85.
Datta, S. R., H. Dudek, et al. (1997). "Akt phosphorylation of BAD couples survival signals to the cell-intrinsic death machinery." Cell **91**(2): 231-41.
Dayhoff, M. O., R. M. Schwartz, et al. (1978). A model of evolutionnary change in proteins, matrixes for detecting distant relationships. Atlas of protein sequence and structure. M. O. Dayhoff. Washington, DC, National Biomedical Research Foundation.
De Kimpe, S. J., M. L. Hunter, et al. (1995). "Delayed circulatory failure due to the induction of nitric oxide synthase by lipoteichoic acid from Staphylococcus aureus in anaesthetized rats." Br J Pharmacol **114**(6): 1317-23.
de Sauvage, F. J., P. E. Hass, et al. (1994). "Stimulation of megakaryocytopoiesis and thrombopoiesis by the c-Mpl ligand." Nature **369**(6481): 533-8.
de Vos, A. M., M. Ultsch, et al. (1992). "Human growth hormone and extracellular domain of its receptor: crystal structure of the complex." Science **255**(5042): 306-12.
de Waal Malefyt, R., J. Abrams, et al. (1991). "Interleukin 10(IL-10) inhibits cytokine synthesis by human monocytes: an autoregulatory role of IL-10 produced by monocytes." J Exp Med **174**(5): 1209-20.
Deleuran, B. W., C. Q. Chu, et al. (1992). "Localization of interleukin-1 alpha, type 1 interleukin-1 receptor and interleukin-1 receptor antagonist in the synovial membrane and cartilage/pannus junction in rheumatoid arthritis." Br J Rheumatol **31**(12): 801-9.
Dillon, S. R., C. Sprecher, et al. (2004). "Interleukin 31, a cytokine produced by activated T cells, induces dermatitis in mice." Nat Immunol **5**(7): 752-60.
Donaldson, D. D., M. J. Whitters, et al. (1998). "The murine IL-13 receptor alpha 2: molecular cloning, characterization, and comparison with murine IL-13 receptor alpha 1." J Immunol **161**(5): 2317-24.
Donnelly, R. P., F. Sheikh, et al. (2004). "The expanded family of class II cytokines that share the IL-10 receptor-2 (IL-10R2) chain." J Leukoc Biol **76**(2): 314-21.
Dorssers, L., H. Burger, et al. (1987). "Characterization of a human multilineage-colony-stimulating factor cDNA clone identified by a conserved noncoding sequence in mouse interleukin-3." Gene **55**(1): 115-24.
Dreuw, A., S. Radtke, et al. (2004). "Characterization of the signaling capacities of the novel gp130-like cytokine receptor." J Biol Chem **279**(34): 36112-20.
Dubois, S., J. Mariner, et al. (2002). "IL-15Ralpha recycles and presents IL-15 In trans to neighboring cells." Immunity **17**(5): 537-47.

Dumoutier, L., C. Leemans, et al. (2001). "Cutting edge: STAT activation by IL-19, IL-20 and mda-7 through IL-20 receptor complexes of two types." J Immunol **167**(7): 3545-9.
Dumoutier, L., D. Lejeune, et al. (2001). "Cloning and characterization of IL-22 binding protein, a natural antagonist of IL-10-related T cell-derived inducible factor/IL-22." J Immunol **166**(12): 7090-5.
Dumoutier, L., J. Louahed, et al. (2000). "Cloning and characterization of IL-10-related T cell-derived inducible factor (IL-TIF), a novel cytokine structurally related to IL-10 and inducible by IL-9." J Immunol **164**(4): 1814-9.
Dumoutier, L., E. Van Roost, et al. (2000). "IL-TIF/IL-22: genomic organization and mapping of the human and mouse genes." Genes Immun **1**(8): 488-94.
Dumoutier, L., E. Van Roost, et al. (2000). "Human interleukin-10-related T cell-derived inducible factor: molecular cloning and functional characterization as an hepatocyte-stimulating factor." Proc Natl Acad Sci U S A **97**(18): 10144-9.
Ealick, S. E., W. J. Cook, et al. (1991). "Three-dimensional structure of recombinant human interferon-gamma." Science **252**(5006): 698-702.
Eddy, S. R. (1995). Multiple Alignment Using Hidden Markov Models. Proc. Third Int. Conf. Intelligent Systems for Molecular Biology.
Eddy, S. R., G. Mitchison, et al. (1995). "Maximum discrimination hidden Markov models of sequence consensus." J Comput Biol **2**(1): 9-23.
Efron, B. (1979). "Bootstrap methods: another look at the jacknife." Annals of Statistics **7**(1): 1-26.
Eisenmesser, E. Z., D. A. Horita, et al. (2001). "Solution structure of interleukin-13 and insights into receptor engagement." J Mol Biol **310**(1): 231-41.
Elson, G. C., E. Lelievre, et al. (2000). "CLF associates with CLC to form a functional heteromeric ligand for the CNTF receptor complex." Nat Neurosci **3**(9): 867-72.
Feldmann, M. and R. N. Maini (2001). "Anti-TNF alpha therapy of rheumatoid arthritis: what have we learned?" Annu Rev Immunol **19**: 163-96.
Feng, Y., B. K. Klein, et al. (1996). "Three-dimensional solution structure and backbone dynamics of a variant of human interleukin-3." J Mol Biol **259**(3): 524-41.
Fickenscher, H., S. Hor, et al. (2002). "The interleukin-10 family of cytokines." Trends Immunol **23**(2): 89-96.
Fickenscher, H. and H. Pirzer (2004). "Interleukin-26." Int Immunopharmacol **4**(5): 609-13.
Finbloom, D. S. and K. D. Winestock (1995). "IL-10 induces the tyrosine phosphorylation of tyk2 and Jak1 and the differential assembly of STAT1 alpha and STAT3 complexes in human T cells and monocytes." J Immunol **155**(3): 1079-90.
Firestein, G. S., A. E. Berger, et al. (1992). "IL-1 receptor antagonist protein production and gene expression in rheumatoid arthritis and osteoarthritis synovium." J Immunol **149**(3): 1054-62.
Foster, D. C., C. A. Sprecher, et al. (1994). "Human thrombopoietin: gene structure, cDNA sequence, expression, and chromosomal localization." Proc Natl Acad Sci U S A **91**(26): 13023-7.
Foxwell, B., K. Browne, et al. (1998). "Efficient adenoviral infection with IkappaB alpha reveals that macrophage tumor necrosis factor alpha production in rheumatoid arthritis is NF-kappaB dependent." Proc Natl Acad Sci U S A **95**(14): 8211-5.
Frucht, D. M., T. Fukao, et al. (2001). "IFN-gamma production by antigen-presenting cells: mechanisms emerge." Trends Immunol **22**(10): 556-60.
Gadina, M., D. Hilton, et al. (2001). "Signaling by type I and II cytokine receptors: ten years after." Curr Opin Immunol **13**(3): 363-73.

Gallagher, G., H. Dickensheets, et al. (2000). "Cloning, expression and initial characterization of interleukin-19 (IL-19), a novel homologue of human interleukin-10 (IL-10)." Genes Immun **1**(7): 442-50.

Gately, M. K., L. M. Renzetti, et al. (1998). "The interleukin-12/interleukin-12-receptor system: role in normal and pathologic immune responses." Annu Rev Immunol **16**: 495-521.

Gessani, S. and F. Belardelli (1998). "IFN-gamma expression in macrophages and its possible biological significance." Cytokine Growth Factor Rev **9**(2): 117-23.

Ghoreschi, K., P. Thomas, et al. (2003). "Interleukin-4 therapy of psoriasis induces Th2 responses and improves human autoimmune disease." Nat Med **9**(1): 40-6.

Gordon, M. Y. (1991). "Hemopoietic growth factors and receptors: bound and free." Cancer Cells **3**(4): 127-33.

Gough, N. M., D. P. Gearing, et al. (1988). "Molecular cloning and expression of the human homologue of the murine gene encoding myeloid leukemia-inhibitory factor." Proc Natl Acad Sci U S A **85**(8): 2623-7.

Grabstein, K. H., J. Eisenman, et al. (1994). "Cloning of a T cell growth factor that interacts with the beta chain of the interleukin-2 receptor." Science **264**(5161): 965-8.

Grewe, M., R. Gausling, et al. (1994). "Regulation of the mRNA expression for tumor necrosis factor-alpha in rat liver macrophages." J Hepatol **20**(6): 811-8.

Gribskov, M., A. D. McLachlan, et al. (1987). "Profile analysis: detection of distantly related proteins." Proc Natl Acad Sci U S A **84**(13): 4355-8.

Grotzinger, J., G. Kurapkat, et al. (1997). "The family of the IL-6-type cytokines: specificity and promiscuity of the receptor complexes." Proteins **27**(1): 96-109.

Habib, T., A. Nelson, et al. (2003). "IL-21: a novel IL-2-family lymphokine that modulates B, T, and natural killer cell responses." J Allergy Clin Immunol **112**(6): 1033-45.

Hack, C. E., L. A. Aarden, et al. (1997). "Role of cytokines in sepsis." Adv Immunol **66**: 101-95.

Harris, N. L., S. R. Presnell, et al. (1994). "Four helix bundle diversity in globular proteins." J Mol Biol **236**(5): 1356-68.

Harrison, G. A. and D. N. Wedlock (2000). "Marsupial cytokines. Structure, function and evolution." Dev Comp Immunol **24**(5): 473-84.

Hawwari, A., J. Burrows, et al. (2002). "The human IL-3 locus is regulated cooperatively by two NFAT-dependent enhancers that have distinct tissue-specific activities." J Immunol **169**(4): 1876-86.

Hayashida, K., T. Kitamura, et al. (1990). "Molecular cloning of a second subunit of the receptor for human granulocyte-macrophage colony-stimulating factor (GM-CSF): reconstitution of a high-affinity GM-CSF receptor." Proc Natl Acad Sci U S A **87**(24): 9655-9.

He, C., J. Chen, et al. (1995). "Preparation and a structure-function analysis of human ciliary neurotrophic factor." Neurosci Res **23**(4): 327-33.

Hibi, M., K. Nakajima, et al. (1996). "IL-6 cytokine family and signal transduction: a model of the cytokine system." J Mol Med **74**(1): 1-12.

Hill, C. P., T. D. Osslund, et al. (1993). "The structure of granulocyte-colony-stimulating factor and its relationship to other growth factors." Proc Natl Acad Sci U S A **90**(11): 5167-71.

Hilton, D. J., J. G. Zhang, et al. (1996). "Cloning and characterization of a binding subunit of the interleukin 13 receptor that is also a component of the interleukin 4 receptor." Proc Natl Acad Sci U S A **93**(1): 497-501.

Hinds, M. G., T. Maurer, et al. (1998). "Solution structure of leukemia inhibitory factor." J Biol Chem **273**(22): 13738-45.

Hirano, T., K. Yasukawa, et al. (1986). "Complementary DNA for a novel human interleukin (BSF-2) that induces B lymphocytes to produce immunoglobulin." Nature **324**(6092): 73-6.
Ho, Y., F. Elefant, et al. (2002). "A defined locus control region determinant links chromatin domain acetylation with long-range gene activation." Mol Cell **9**(2): 291-302.
Holland, J. H. (1975). Adaptation in natural and artificial systems, University of Michigan Press. Reprinted in 1992 by the MIT Press.
Holloway, A. F., S. Rao, et al. (2002). "Regulation of cytokine gene transcription in the immune system." Mol Immunol **38**(8): 567-80.
Hopfield, J. J. (1982). "Neural networks and physical systems with emergent collective computational
abilities." Proc Natl Acad Sci U S A **79**(8): 2554-2558.
Hopfield, J. J. and T. D. W. (1985). "Neural computation of decisions in optimization problems." Biological Cybernetics **52**: 141-152.
Hor, S., H. Pirzer, et al. (2004). "The T-cell lymphokine interleukin-26 targets epithelial cells through the interleukin-20 receptor 1 and interleukin-10 receptor 2 chains." J Biol Chem **279**(32): 33343-51.
Hou, Y., W. Hsu, et al. (2003). "Efficient remote homology detection using local structure." Bioinformatics **19**(17): 2294-301.
Hudson, K. R., A. B. Vernallis, et al. (1996). "Characterization of the receptor binding sites of human leukemia inhibitory factor and creation of antagonists." J Biol Chem **271**(20): 11971-8.
Hulo, N., C. J. Sigrist, et al. (2004). "Recent improvements to the PROSITE database." Nucleic Acids Res **32 Database issue**: D134-7.
Ihle, J. N. (2001). "The Stat family in cytokine signaling." Curr Opin Cell Biol **13**(2): 211-7.
Ihle, J. N., T. Nosaka, et al. (1997). "Jaks and Stats in cytokine signaling." Stem Cells **15 Suppl 1**: 105-11; discussion 112.
Ihle, J. N., W. Thierfelder, et al. (1998). "Signaling by the cytokine receptor superfamily." Ann N Y Acad Sci **865**: 1-9.
Ihle, J. N., B. A. Witthuhn, et al. (1995). "Signaling through the hematopoietic cytokine receptors." Annu Rev Immunol **13**: 369-98.
Isaacs, A. and J. Lindenmann (1957). "Virus interference. I. The interferon." Proc R Soc Lond B Biol Sci **147**(927): 258-67.
Ishihara, K. and T. Hirano (2002). "IL-6 in autoimmune disease and chronic inflammatory proliferative disease." Cytokine Growth Factor Rev **13**(4-5): 357-68.
Itoh, N., S. Yonehara, et al. (1990). "Cloning of an interleukin-3 receptor gene: a member of a distinct receptor gene family." Science **247**(4940): 324-7.
Jaakkola, T., M. Diekhans, et al. (1999). "Using the Fisher kernel method to detect remote protein homologies." Proc Int Conf Intell Syst Mol Biol: 149-58.
Jaakkola, T., M. Diekhans, et al. (2000). "A discriminative framework for detecting remote protein homologies." J Comput Biol **7**(1-2): 95-114.
Jaeschke, H. and C. W. Smith (1997). "Cell adhesion and migration. III. Leukocyte adhesion and transmigration in the liver vasculature." Am J Physiol **273**(6 Pt 1): G1169-73.
Josephson, K., N. J. Logsdon, et al. (2001). "Crystal structure of the IL-10/IL-10R1 complex reveals a shared receptor binding site." Immunity **15**(1): 35-46.
Joyce, D. A., D. P. Gibbons, et al. (1994). "Two inhibitors of pro-inflammatory cytokine release, interleukin-10 and interleukin-4, have contrasting effects on release of soluble p75 tumor necrosis factor receptor by cultured monocytes." Eur J Immunol **24**(11): 2699-705.

Kallen, K. J., J. Grotzinger, et al. (1999). "Receptor recognition sites of cytokines are organized as exchangeable modules. Transfer of the leukemia inhibitory factor receptor-binding site from ciliary neurotrophic factor to interleukin-6." J Biol Chem **274**(17): 11859-67.

Karima, R., S. Matsumoto, et al. (1999). "The molecular pathogenesis of endotoxic shock and organ failure." Mol Med Today **5**(3): 123-32.

Karplus, K., C. Barrett, et al. (1998). "Hidden Markov models for detecting remote protein homologies." Bioinformatics **14**(10): 846-56.

Karpusas, M., M. Nolte, et al. (1997). "The crystal structure of human interferon beta at 2.2-A resolution." Proc Natl Acad Sci U S A **94**(22): 11813-8.

Katori, M. and M. Majima (2000). "Cyclooxygenase-2: its rich diversity of roles and possible application of its selective inhibitors." Inflamm Res **49**(8): 367-92.

Katsikis, P. D., C. Q. Chu, et al. (1994). "Immunoregulatory role of interleukin 10 in rheumatoid arthritis." J Exp Med **179**(5): 1517-27.

Kawamura, N., N. Imanishi, et al. (1995). "Lipoteichoic acid-induced neutrophil adhesion via E-selectin to human umbilical vein endothelial cells (HUVECs)." Biochem Biophys Res Commun **217**(3): 1208-15.

Kelly-Welch, A. E., E. M. Hanson, et al. (2003). "Interleukin-4 and interleukin-13 signaling connections maps." Science **300**(5625): 1527-8.

Kengatharan, K. M., S. De Kimpe, et al. (1998). "Mechanism of gram-positive shock: identification of peptidoglycan and lipoteichoic acid moieties essential in the induction of nitric oxide synthase, shock, and multiple organ failure." J Exp Med **188**(2): 305-15.

Kira, K. and L. Rendell (1992). A Practical Approach to Feature Selection. Procs. of the Ninth International Conference on Machine Learning.

Kirman, I. and O. H. Nielsen (1996). "Increased numbers of interleukin-15-expressing cells in active ulcerative colitis." Am J Gastroenterol **91**(9): 1789-94.

Kishimoto, T., T. Taga, et al. (1994). "Cytokine signal transduction." Cell **76**(2): 253-62.

Kitamura, T., N. Sato, et al. (1991). "Expression cloning of the human IL-3 receptor cDNA reveals a shared beta subunit for the human IL-3 and GM-CSF receptors." Cell **66**(6): 1165-74.

Kitchen, D., R. C. Hoffman, et al. (1998). "Homology model for oncostatin M based on NMR structural data." Biochemistry **37**(30): 10581-8.

Klein, B. (1998). "Update of gp130 cytokines in multiple myeloma." Curr Opin Hematol **5**(3): 186-91.

Knappe, A., S. Hor, et al. (2000). "Induction of a novel cellular homolog of interleukin-10, AK155, by transformation of T lymphocytes with herpesvirus saimiri." J Virol **74**(8): 3881-7.

Kohonen, T. (1982). "Self-organized formation of topologically correct feature maps." Biological Cybernetics **43**: 59-69.

Kollias, G., E. Douni, et al. (1999). "The function of tumour necrosis factor and receptors in models of multi-organ inflammation, rheumatoid arthritis, multiple sclerosis and inflammatory bowel disease." Ann Rheum Dis **58 Suppl 1**: I32-9.

Kotenko, S. V., G. Gallagher, et al. (2003). "IFN-lambdas mediate antiviral protection through a distinct class II cytokine receptor complex." Nat Immunol **4**(1): 69-77.

Kotenko, S. V., L. S. Izotova, et al. (2001). "Identification of the functional interleukin-22 (IL-22) receptor complex: the IL-10R2 chain (IL-10Rbeta) is a common chain of both the IL-10 and IL-22 (IL-10-related T cell-derived inducible factor, IL-TIF) receptor complexes." J Biol Chem **276**(4): 2725-32.

Kotenko, S. V., L. S. Izotova, et al. (2001). "Identification, cloning, and characterization of a novel soluble receptor that binds IL-22 and neutralizes its activity." J Immunol **166**(12): 7096-103.
Kotenko, S. V., L. S. Izotova, et al. (1999). "The intracellular domain of interferon-alpha receptor 2c (IFN-alphaR2c) chain is responsible for Stat activation." Proc Natl Acad Sci U S A **96**(9): 5007-12.
Kotenko, S. V., C. D. Krause, et al. (1997). "Identification and functional characterization of a second chain of the interleukin-10 receptor complex." Embo J **16**(19): 5894-903.
Kotenko, S. V. and J. A. Langer (2004). "Full house: 12 receptors for 27 cytokines." Int Immunopharmacol **4**(5): 593-608.
Krueger, J. G. (2002). "The immunologic basis for the treatment of psoriasis with new biologic agents." J Am Acad Dermatol **46**(1): 1-23; quiz 23-6.
Lam, A., F. Fuller, et al. (1991). "Sequence and structural organization of the human gene encoding ciliary neurotrophic factor." Gene **102**(2): 271-6.
Langer, J. A., E. C. Cutrone, et al. (2004). "The Class II cytokine receptor (CRF2) family: overview and patterns of receptor-ligand interactions." Cytokine Growth Factor Rev **15**(1): 33-48.
Lejeune, D., L. Dumoutier, et al. (2002). "Interleukin-22 (IL-22) activates the JAK/STAT, ERK, JNK, and p38 MAP kinase pathways in a rat hepatoma cell line. Pathways that are shared with and distinct from IL-10." J Biol Chem **277**(37): 33676-82.
Lenardo, M. J. (1991). "Interleukin-2 programs mouse alpha beta T lymphocytes for apoptosis." Nature **353**(6347): 858-61.
Leonard, W. J. (2001). "Cytokines and immunodeficiency diseases." Nat Rev Immunol **1**(3): 200-8.
Leslie, C., E. Eskin, et al. (2002). "The spectrum kernel: a string kernel for SVM protein classification." Pac Symp Biocomput: 564-75.
Leslie, C. S., E. Eskin, et al. (2004). "Mismatch string kernels for discriminative protein classification." Bioinformatics **20**(4): 467-76.
Levings, M. K., R. Sangregorio, et al. (2001). "IFN-alpha and IL-10 induce the differentiation of human type 1 T regulatory cells." J Immunol **166**(9): 5530-9.
Liao, L. and W. S. Noble (2003). "Combining pairwise sequence similarity and support vector machines for detecting remote protein evolutionary and structural relationships." J Comput Biol **10**(6): 857-68.
Lin, L. F., D. Mismer, et al. (1989). "Purification, cloning, and expression of ciliary neurotrophic factor (CNTF)." Science **246**(4933): 1023-5.
Logsdon, N. J., B. C. Jones, et al. (2004). "The IL-10R2 Binding Hot Spot on IL-22 is Located on the N-terminal Helix and is Dependent on N-linked Glycosylation." J Mol Biol **342**(2): 503-14.
Loots, G. G., R. M. Locksley, et al. (2000). "Identification of a coordinate regulator of interleukins 4, 13, and 5 by cross-species sequence comparisons." Science **288**(5463): 136-40.
Lukacs, N. W., C. Hogaboam, et al. (1999). "Chemokines: function, regulation and alteration of inflammatory responses." Chem Immunol **72**: 102-20.
Luster, M. I., D. R. Germolec, et al. (1994). "Endotoxin-induced cytokine gene expression and excretion in the liver." Hepatology **19**(2): 480-8.
Mahalingam, S. and G. Karupiah (1999). "Chemokines and chemokine receptors in infectious diseases." Immunol Cell Biol **77**(6): 469-75.
Maini, R., E. W. St Clair, et al. (1999). "Infliximab (chimeric anti-tumour necrosis factor alpha monoclonal antibody) versus placebo in rheumatoid arthritis patients receiving

concomitant methotrexate: a randomised phase III trial. ATTRACT Study Group." Lancet **354**(9194): 1932-9.
Maini, R. N., F. C. Breedveld, et al. (1998). "Therapeutic efficacy of multiple intravenous infusions of anti-tumor necrosis factor alpha monoclonal antibody combined with low-dose weekly methotrexate in rheumatoid arthritis." Arthritis Rheum **41**(9): 1552-63.
Malik, N., J. C. Kallestad, et al. (1989). "Molecular cloning, sequence analysis, and functional expression of a novel growth regulator, oncostatin M." Mol Cell Biol **9**(7): 2847-53.
Manilla, H. and H. Toivonen (1996). Discovering generalized episodes using minimal occurences. Knowledge Discovery and Data Mining**:** 146-151.
Masiakowski, P., H. X. Liu, et al. (1991). "Recombinant human and rat ciliary neurotrophic factors." J Neurochem **57**(3): 1003-12.
Matthys, P., K. Vermeire, et al. (2001). "Mac-1(+) myelopoiesis induced by CFA: a clue to the paradoxical effects of IFN-gamma in autoimmune disease models." Trends Immunol **22**(7): 367-71.
McCulloch, W. S. and W. Pitts (1943). "A logical calculus of the ideas immanent in nervous activity." Bulletin of Mathematical Biophysics **5**: 115-133.
McDonald, N. Q., N. Panayotatos, et al. (1995). "Crystal structure of dimeric human ciliary neurotrophic factor determined by MAD phasing." Embo J **14**(12): 2689-99.
McInnes, I. B., J. al-Mughales, et al. (1996). "The role of interleukin-15 in T-cell migration and activation in rheumatoid arthritis." Nat Med **2**(2): 175-82.
Mease, P. J., B. S. Goffe, et al. (2000). "Etanercept in the treatment of psoriatic arthritis and psoriasis: a randomised trial." Lancet **356**(9227): 385-90.
Merberg, D. M., S. F. Wolf, et al. (1992). "Sequence similarity between NKSF and the IL-6/G-CSF family." Immunol Today **13**(2): 77-8.
Mikolajczak, J., G. Ramstein, et al. (2004). 5th International Conference of Intelligent Data Engineering and Automated Learning (IDEAL 2004). 5th International Conference of Intelligent Data Engineering and Automated Learning (IDEAL 2004), EXETER (UNITED-KINGDOM), Springer-Verlag GmbH.
Mikolajczak, J., G. Ramstein, et al. (2004). Caractérisation de signatures complexes dans des familles de protéines distantes. Extraction et Gestion des Connaissances (EGC'2004), CLERMONT-FERRAND (FRANCE), CEPADUES-Editions.
Mikolajczak, J., G. Ramstein, et al. (2004). Classification de protéines distantes par motifs hiérarchiques. Journées Ouvertes Biologie Informatique Mathèmatiques (JOBIM 2004), MONTREAL (CANADA).
Mikolajczak, J., G. Ramstein, et al. (2004). Détection de faibles homologies de protéines par machines à vecteurs de support. Revue des Nouvelles Technologies de l'Information - Classification et Fouille de données (RNTI-C1). M. Chavent and M. LANGLAIS, CEPADUES-Edition. **1:** 89-100.
Mikolajczak, J., G. Ramstein, et al. (2004). Détection de faibles homologies de protéines par machines à vecteurs de support. 11èmes Rencontres de la Société Francophone de Classification SFC'04, BORDEAUX (FRANCE), SFC.
Milburn, M. V., A. M. Hassell, et al. (1993). "A novel dimer configuration revealed by the crystal structure at 2.4 A resolution of human interleukin-5." Nature **363**(6425): 172-6.
Miyajima, A., T. Kinoshita, et al. (2000). "Role of Oncostatin M in hematopoiesis and liver development." Cytokine Growth Factor Rev **11**(3): 177-83.
Miyajima, A., A. L. Mui, et al. (1993). "Receptors for granulocyte-macrophage colony-stimulating factor, interleukin-3, and interleukin-5." Blood **82**(7): 1960-74.
Moore, K. W., R. de Waal Malefyt, et al. (2001). "Interleukin-10 and the interleukin-10 receptor." Annu Rev Immunol **19**: 683-765.

Moore, R. J., D. M. Owens, et al. (1999). "Mice deficient in tumor necrosis factor-alpha are resistant to skin carcinogenesis." Nat Med **5**(7): 828-31.
Moreau, J. F., D. D. Donaldson, et al. (1988). "Leukaemia inhibitory factor is identical to the myeloid growth factor human interleukin for DA cells." Nature **336**(6200): 690-2.
Moreau, J. L., M. Bossus, et al. (1995). "Characterization of a monoclonal antibody directed against the NH2 terminal area of interleukin-2 (IL-2) and inhibiting specifically the binding of IL-2 to IL-2 receptor beta chain (IL-2R beta)." Mol Immunol **32**(14-15): 1047-56.
Moreland, L., R. Gugliotti, et al. (2001). "Results of a phase-I/II randomized, masked, placebo-controlled trial of recombinant human interleukin-11 (rhIL-11) in the treatment of subjects with active rheumatoid arthritis." Arthritis Res **3**(4): 247-52.
Mosley, B., M. P. Beckmann, et al. (1989). "The murine interleukin-4 receptor: molecular cloning and characterization of secreted and membrane bound forms." Cell **59**(2): 335-48.
Mott, H. R., B. S. Baines, et al. (1995). "The solution structure of the F42A mutant of human interleukin 2." J Mol Biol **247**(5): 979-94.
Moy, F. J., E. Diblasio, et al. (2001). "Solution structure of human IL-13 and implication for receptor binding." J Mol Biol **310**(1): 219-30.
Murzin, A. G., S. E. Brenner, et al. (1995). "SCOP: a structural classification of proteins database for the investigation of sequences and structures." J Mol Biol **247**(4): 536-40.
Musso, T., L. Calosso, et al. (1999). "Human monocytes constitutively express membrane-bound, biologically active, and interferon-gamma-upregulated interleukin-15." Blood **93**(10): 3531-9.
Nagata, S., M. Tsuchiya, et al. (1986). "Molecular cloning and expression of cDNA for human granulocyte colony-stimulating factor." Nature **319**(6052): 415-8.
Namen, A. E., A. E. Schmierer, et al. (1988). "B cell precursor growth-promoting activity. Purification and characterization of a growth factor active on lymphocyte precursors." J Exp Med **167**(3): 988-1002.
Naylor, M. S., G. W. Stamp, et al. (1993). "Tumor necrosis factor and its receptors in human ovarian cancer. Potential role in disease progression." J Clin Invest **91**(5): 2194-206.
Negro, A., E. Tolosano, et al. (1991). "Cloning and expression of human ciliary neurotrophic factor." Eur J Biochem **201**(1): 289-94.
Nickoloff, B. J. (1991). "The cytokine network in psoriasis." Arch Dermatol **127**(6): 871-84.
Nishimura, C., A. Watanabe, et al. (1996). "Folding topologies of human interleukin-6 and its mutants as studied by NMR spectroscopy." Biochemistry **35**(1): 273-81.
Nishinakamura, R., A. Miyajima, et al. (1996). "Hematopoiesis in mice lacking the entire granulocyte-macrophage colony-stimulating factor/interleukin-3/interleukin-5 functions." Blood **88**(7): 2458-64.
O'Shea, J. J., A. Ma, et al. (2002). "Cytokines and autoimmunity." Nat Rev Immunol **2**(1): 37-45.
Oestreicher, J. L., I. B. Walters, et al. (2001). "Molecular classification of psoriasis disease-associated genes through pharmacogenomic expression profiling." Pharmacogenomics J **1**(4): 272-87.
Ohta, N., T. Hiroi, et al. (2002). "IL-15-dependent activation-induced cell death-resistant Th1 type CD8 alpha beta+NK1.1+ T cells for the development of small intestinal inflammation." J Immunol **169**(1): 460-8.
Oosterwegel, M. A., R. J. Greenwald, et al. (1999). "CTLA-4 and T cell activation." Curr Opin Immunol **11**(3): 294-300.
Opal, S. M. and J. Cohen (1999). "Clinical gram-positive sepsis: does it fundamentally differ from gram-negative bacterial sepsis?" Crit Care Med **27**(8): 1608-16.

Oppmann, B., R. Lesley, et al. (2000). "Novel p19 protein engages IL-12p40 to form a cytokine, IL-23, with biological activities similar as well as distinct from IL-12." Immunity **13**(5): 715-25.

Otsuka, T., A. Miyajima, et al. (1988). "Isolation and characterization of an expressible cDNA encoding human IL-3. Induction of IL-3 mRNA in human T cell clones." J Immunol **140**(7): 2288-95.

Owens, T., H. Wekerle, et al. (2001). "Genetic models for CNS inflammation." Nat Med **7**(2): 161-6.

Panayotatos, N., E. Radziejewska, et al. (1995). "Localization of functional receptor epitopes on the structure of ciliary neurotrophic factor indicates a conserved, function-related epitope topography among helical cytokines." J Biol Chem **270**(23): 14007-14.

Pandey, A., K. Ozaki, et al. (2000). "Cloning of a receptor subunit required for signaling by thymic stromal lymphopoietin." Nat Immunol **1**(1): 59-64.

Parham, C., M. Chirica, et al. (2002). "A receptor for the heterodimeric cytokine IL-23 is composed of IL-12Rbeta1 and a novel cytokine receptor subunit, IL-23R." J Immunol **168**(11): 5699-708.

Park, L. S., U. Martin, et al. (2000). "Cloning of the murine thymic stromal lymphopoietin (TSLP) receptor: Formation of a functional heteromeric complex requires interleukin 7 receptor." J Exp Med **192**(5): 659-70.

Parrish-Novak, J., S. R. Dillon, et al. (2000). "Interleukin 21 and its receptor are involved in NK cell expansion and regulation of lymphocyte function." Nature **408**(6808): 57-63.

Paul, S. R., F. Bennett, et al. (1990). "Molecular cloning of a cDNA encoding interleukin 11, a stromal cell-derived lymphopoietic and hematopoietic cytokine." Proc Natl Acad Sci U S A **87**(19): 7512-6.

Pennica, D., T. A. Swanson, et al. (1996). "Human cardiotrophin-1: protein and gene structure, biological and binding activities, and chromosomal localization." Cytokine **8**(3): 183-9.

Pflanz, S., L. Hibbert, et al. (2004). "WSX-1 and glycoprotein 130 constitute a signal-transducing receptor for IL-27." J Immunol **172**(4): 2225-31.

Pflanz, S., J. C. Timans, et al. (2002). "IL-27, a heterodimeric cytokine composed of EBI3 and p28 protein, induces proliferation of naive CD4(+) T cells." Immunity **16**(6): 779-90.

Powers, R., D. S. Garrett, et al. (1992). "Three-dimensional solution structure of human interleukin-4 by multidimensional heteronuclear magnetic resonance spectroscopy." Science **256**(5064): 1673-7.

Purvis, D. H. and B. C. Mabbutt (1997). "Solution dynamics and secondary structure of murine leukemia inhibitory factor: a four-helix cytokine with a rigid CD loop." Biochemistry **36**(33): 10146-54.

Quentmeier, H., H. G. Drexler, et al. (2001). "Cloning of human thymic stromal lymphopoietin (TSLP) and signaling mechanisms leading to proliferation." Leukemia **15**(8): 1286-92.

Quinlan, J. R. (1979). Discovering Rules by Induction from Large Collections of
Examples. Expert Systems in the Micro-Electronic Age. M. D., Edinburgh University Press**:** 168-201.

Quinlan, J. R. (1993). C4.5: Programs for Machine Learning, Morgan Kaufmann.

Qureshi, N., J. P. Honovich, et al. (1988). "Location of fatty acids in lipid A obtained from lipopolysaccharide of Rhodopseudomonas sphaeroides ATCC 17023." J Biol Chem **263**(12): 5502-4.

Radhakrishnan, R., L. J. Walter, et al. (1996). "Zinc mediated dimer of human interferon-alpha 2b revealed by X-ray crystallography." Structure **4**(12): 1453-63.

Rappersberger, K., M. Komar, et al. (2002). "Pimecrolimus identifies a common genomic anti-inflammatory profile, is clinically highly effective in psoriasis and is well tolerated." J Invest Dermatol **119**(4): 876-87.
Refaeli, Y., L. Van Parijs, et al. (1999). "Genetic models of abnormal apoptosis in lymphocytes." Immunol Rev **169**: 273-82.
Refaeli, Y., L. Van Parijs, et al. (1998). "Biochemical mechanisms of IL-2-regulated Fas-mediated T cell apoptosis." Immunity **8**(5): 615-23.
Reif, K., B. M. Burgering, et al. (1997). "Phosphatidylinositol 3-kinase links the interleukin-2 receptor to protein kinase B and p70 S6 kinase." J Biol Chem **272**(22): 14426-33.
Reineke, U., R. Sabat, et al. (1998). "Mapping of the interleukin-10/interleukin-10 receptor combining site." Protein Sci **7**(4): 951-60.
Reitamo, S., P. Spuls, et al. (2001). "Efficacy of sirolimus (rapamycin) administered concomitantly with a subtherapeutic dose of cyclosporin in the treatment of severe psoriasis: a randomized controlled trial." Br J Dermatol **145**(3): 438-45.
Renauld, J. C. (2003). "Class II cytokine receptors and their ligands: key antiviral and inflammatory modulators." Nat Rev Immunol **3**(8): 667-76.
Renauld, J. C., A. Goethals, et al. (1990). "Cloning and expression of a cDNA for the human homolog of mouse T cell and mast cell growth factor P40." Cytokine **2**(1): 9-12.
Rickert, M., M. J. Boulanger, et al. (2004). "Compensatory energetic mechanisms mediating the assembly of signaling complexes between interleukin-2 and its alpha, beta, and gamma(c) receptors." J Mol Biol **339**(5): 1115-28.
Robinson, R. C., L. M. Grey, et al. (1994). "The crystal structure and biological function of leukemia inhibitory factor: implications for receptor binding." Cell **77**(7): 1101-16.
Romashkova, J. A. and S. S. Makarov (1999). "NF-kappaB is a target of AKT in anti-apoptotic PDGF signalling." Nature **401**(6748): 86-90.
Rosenblatt, F. (1958). "The Perceptron: a probabilistic model for information storage and organization in the brain." Psychological Review **65**: 386-408.
Ruchatz, H., B. P. Leung, et al. (1998). "Soluble IL-15 receptor alpha-chain administration prevents murine collagen-induced arthritis: a role for IL-15 in development of antigen-induced immunopathology." J Immunol **160**(11): 5654-60.
Rumelhart, D. E., G. E. Hinton, et al. (1986). Learning internal representations by error propagation. Parallel Distributed Processing: Explorations in The Microstructure of Cognition, MIT Press.
Sakaguchi, S. (2000). "Regulatory T cells: key controllers of immunologic self-tolerance." Cell **101**(5): 455-8.
Sakai, T., K. Kusugami, et al. (1998). "Interleukin 15 activity in the rectal mucosa of inflammatory bowel disease." Gastroenterology **114**(6): 1237-43.
Sauane, M., R. V. Gopalkrishnan, et al. (2003). "Mda-7/IL-24 induces apoptosis of diverse cancer cell lines through JAK/STAT-independent pathways." J Cell Physiol **196**(2): 334-45.
Sauve, K., M. Nachman, et al. (1991). "Localization in human interleukin 2 of the binding site to the alpha chain (p55) of the interleukin 2 receptor." Proc Natl Acad Sci U S A **88**(11): 4636-40.
Scholkopf, B., J. C. Platt, et al. (2001). "Estimating the support of a high-dimensional distribution." Neural Comput **13**(7): 1443-71.
Schuler, G. D. (1997). "Pieces of the puzzle: expressed sequence tags and the catalog of human genes." J Mol Med **75**(10): 694-8.
Senaldi, G., B. C. Varnum, et al. (1999). "Novel neurotrophin-1/B cell-stimulating factor-3: a cytokine of the IL-6 family." Proc Natl Acad Sci U S A **96**(20): 11458-63.

Senda, T., S. Saitoh, et al. (1995). "Refined crystal structure of recombinant murine interferon-beta at 2.15 A resolution." J Mol Biol **253**(1): 187-207.
Sheikh, F., V. V. Baurin, et al. (2004). "Cutting edge: IL-26 signals through a novel receptor complex composed of IL-20 receptor 1 and IL-10 receptor 2." J Immunol **172**(4): 2006-10.
Sheppard, P., W. Kindsvogel, et al. (2003). "IL-28, IL-29 and their class II cytokine receptor IL-28R." Nat Immunol **4**(1): 63-8.
Shi, Y., W. Wang, et al. (1999). "Computational EST database analysis identifies a novel member of the neuropoietic cytokine family." Biochem Biophys Res Commun **262**(1): 132-8.
Shome, B. and A. F. Parlow (1977). "Human pituitary prolactin (hPRL): the entire linear amino acid sequence." J Clin Endocrinol Metab **45**(5): 1112-5.
Simpson, R. J., A. Hammacher, et al. (1997). "Interleukin-6: structure-function relationships." Protein Sci **6**(5): 929-55.
Sims, J. E., D. E. Williams, et al. (2000). "Molecular cloning and biological characterization of a novel murine lymphoid growth factor." J Exp Med **192**(5): 671-80.
Smirnov, D. V., M. G. Smirnova, et al. (1995). "Tandem arrangement of human genes for interleukin-4 and interleukin-13: resemblance in their organization." Gene **155**(2): 277-81.
Smith, T. F. and M. S. Waterman (1981). "Identification of common molecular subsequences." J Mol Biol **147**(1): 195-7.
Sohma, Y., H. Akahori, et al. (1994). "Molecular cloning and chromosomal localization of the human thrombopoietin gene." FEBS Lett **353**(1): 57-61.
Somers, W., M. Stahl, et al. (1997). "1.9 A crystal structure of interleukin 6: implications for a novel mode of receptor dimerization and signaling." Embo J **16**(5): 989-97.
Somers, W., M. Ultsch, et al. (1994). "The X-ray structure of a growth hormone-prolactin receptor complex." Nature **372**(6505): 478-81.
Sriskandan, S. and J. Cohen (1999). "Gram-positive sepsis. Mechanisms and differences from gram-negative sepsis." Infect Dis Clin North Am **13**(2): 397-412.
Sundstrom, M., T. Lundqvist, et al. (1996). "Crystal structure of an antagonist mutant of human growth hormone, G120R, in complex with its receptor at 2.9 A resolution." J Biol Chem **271**(50): 32197-203.
Suzuki, H., Y. W. Zhou, et al. (1999). "Normal regulatory alpha/beta T cells effectively eliminate abnormally activated T cells lacking the interleukin 2 receptor beta in vivo." J Exp Med **190**(11): 1561-72.
Tacken, I., H. Dahmen, et al. (1999). "Definition of receptor binding sites on human interleukin-11 by molecular modeling-guided mutagenesis." Eur J Biochem **265**(2): 645-55.
Taga, T. and T. Kishimoto (1997). "Gp130 and the interleukin-6 family of cytokines." Annu Rev Immunol **15**: 797-819.
Takeshita, T., H. Asao, et al. (1992). "Cloning of the gamma chain of the human IL-2 receptor." Science **257**(5068): 379-82.
Tavernier, J., R. Devos, et al. (1991). "A human high affinity interleukin-5 receptor (IL5R) is composed of an IL5-specific alpha chain and a beta chain shared with the receptor for GM-CSF." Cell **66**(6): 1175-84.
Taylor, W. R. (1986). "The classification of amino acid conservation." J Theor Biol **119**(2): 205-18.
Thoreau, E., B. Petridou, et al. (1991). "Structural symmetry of the extracellular domain of the cytokine/growth hormone/prolactin receptor family and interferon receptors revealed by hydrophobic cluster analysis." FEBS Lett **282**(1): 26-31.

Trepicchio, W. L., M. Ozawa, et al. (1999). "Interleukin-11 therapy selectively downregulates type I cytokine proinflammatory pathways in psoriasis lesions." J Clin Invest **104**(11): 1527-37.
Trinchieri, G. (1998). "Interleukin-12: a cytokine at the interface of inflammation and immunity." Adv Immunol **70**: 83-243.
Uyttenhove, C., R. J. Simpson, et al. (1988). "Functional and structural characterization of P40, a mouse glycoprotein with T-cell growth factor activity." Proc Natl Acad Sci U S A **85**(18): 6934-8.
Van Amersfoort, E. S., T. J. Van Berkel, et al. (2003). "Receptors, mediators, and mechanisms involved in bacterial sepsis and septic shock." Clin Microbiol Rev **16**(3): 379-414.
van Oosten, M., E. van de Bilt, et al. (1995). "Vascular adhesion molecule-1 and intercellular adhesion molecule-1 expression on rat liver cells after lipopolysaccharide administration in vivo." Hepatology **22**(5): 1538-46.
Van Parijs, L., Y. Refaeli, et al. (1999). "Autoimmunity as a consequence of retrovirus-mediated expression of C-FLIP in lymphocytes." Immunity **11**(6): 763-70.
Vapnik, V. N. (1995). The nature of statistical learning theory, Springer-Verlag.
Vapnik, V. N. and A. Y. Chervonenkis (1971). "On the uniform convergence of relative frequencies of events to their probabilities." Th Prob Appl **16**: 264-280.
Villarino, A. V., E. Huang, et al. (2004). "Understanding the pro- and anti-inflammatory properties of IL-27." J Immunol **173**(2): 715-20.
Viterbi, A. (1967). "Error bounds for convolutional codes and an asymtoptically optimum decoding algorithm." IEEE Transactions on Information Theory **13**: 260-267.
Walter, M. R., W. J. Cook, et al. (1992). "Crystal structure of recombinant human interleukin-4." J Biol Chem **267**(28): 20371-6.
Walter, M. R. and T. L. Nagabhushan (1995). "Crystal structure of interleukin 10 reveals an interferon gamma-like fold." Biochemistry **34**(38): 12118-25.
Wang, M., Z. Tan, et al. (2002). "Interleukin 24 (MDA-7/MOB-5) signals through two heterodimeric receptors, IL-22R1/IL-20R2 and IL-20R1/IL-20R2." J Biol Chem **277**(9): 7341-7.
Ward, L. D., G. J. Howlett, et al. (1994). "High affinity interleukin-6 receptor is a hexameric complex consisting of two molecules each of interleukin-6, interleukin-6 receptor, and gp-130." J Biol Chem **269**(37): 23286-9.
Weber-Nordt, R. M., J. K. Riley, et al. (1996). "Stat3 recruitment by two distinct ligand-induced, tyrosine-phosphorylated docking sites in the interleukin-10 receptor intracellular domain." J Biol Chem **271**(44): 27954-61.
Wehinger, J., F. Gouilleux, et al. (1996). "IL-10 induces DNA binding activity of three STAT proteins (Stat1, Stat3, and Stat5) and their distinct combinatorial assembly in the promoters of selected genes." FEBS Lett **394**(3): 365-70.
Weigel, U., M. Meyer, et al. (1989). "Mutant proteins of human interleukin 2. Renaturation yield, proliferative activity and receptor binding." Eur J Biochem **180**(2): 295-300.
Weir, M. P., M. A. Chaplin, et al. (1988). "Structure-activity relationships of recombinant human interleukin 2." Biochemistry **27**(18): 6883-92.
Wen, D., J. P. Boissel, et al. (1993). "Erythropoietin structure-function relationships: high degree of sequence homology among mammals." Blood **82**(5): 1507-16.
Wheelock, E. F. (1965). "Interferon-like virus-inhibitor induced in human leukocytes by phytohemagglutinin." Science **149**: 310-311.
Wiese, K. and S. D. Goodwin (1999). Convergence characteristics of keep-best reproduction. Proceedings of the 1999 ACM symposium on Applied computing. Sn Antonio, Texas, United States**:** 312-18.

Wiese, K., D. Scott, et al. (1998). Keep-best Reproduction: a selection strategy for genetic algorithms. Proceedings of the 1998 ACM symposium on Applied computing, Atlanta, Georgia, United States.

Wlodaver, A., A. Pavlovsky, et al. (1992). "Crystal structure of human recombinant interleukin-4 at 2.25 A resolution." FEBS Lett **309**(1): 59-64.

Wolf, M., A. Schimpl, et al. (2001). "Control of T cell hyperactivation in IL-2-deficient mice by CD4(+)CD25(-) and CD4(+)CD25(+) T cells: evidence for two distinct regulatory mechanisms." Eur J Immunol **31**(6): 1637-45.

Wolk, K., S. Kunz, et al. (2002). "Cutting edge: immune cells as sources and targets of the IL-10 family members?" J Immunol **168**(11): 5397-402.

Wormald, S. and D. J. Hilton (2004). "Inhibitors of cytokine signal transduction." J Biol Chem **279**(2): 821-4.

Wymann, M. P. and L. Pirola (1998). "Structure and function of phosphoinositide 3-kinases." Biochim Biophys Acta **1436**(1-2): 127-50.

Xie, M. H., S. Aggarwal, et al. (2000). "Interleukin (IL)-22, a novel human cytokine that signals through the interferon receptor-related proteins CRF2-4 and IL-22R." J Biol Chem **275**(40): 31335-9.

Yan, H., K. Krishnan, et al. (1996). "Molecular characterization of an alpha interferon receptor 1 subunit (IFNaR1) domain required for TYK2 binding and signal transduction." Mol Cell Biol **16**(5): 2074-82.

Yang, Y. C., A. B. Ciarletta, et al. (1986). "Human IL-3 (multi-CSF): identification by expression cloning of a novel hematopoietic growth factor related to murine IL-3." Cell **47**(1): 3-10.

Yong, K. L. and D. C. Linch (1993). "Granulocyte-macrophage-colony-stimulating factor differentially regulates neutrophil migration across IL-1-activated and nonactivated human endothelium." J Immunol **150**(6): 2449-56.

Young, H. A. (1996). "Regulation of interferon-gamma gene expression." J Interferon Cytokine Res **16**(8): 563-8.

Zajicek, J. P., M. Wing, et al. (1992). "Interactions between oligodendrocytes and microglia. A major role for complement and tumour necrosis factor in oligodendrocyte adherence and killing." Brain **115 (Pt 6)**: 1611-31.

Zarling, J. M., M. Shoyab, et al. (1986). "Oncostatin M: a growth regulator produced by differentiated histiocytic lymphoma cells." Proc Natl Acad Sci U S A **83**(24): 9739-43.

Zdanov, A., C. Schalk-Hihi, et al. (1995). "Crystal structure of interleukin-10 reveals the functional dimer with an unexpected topological similarity to interferon gamma." Structure **3**(6): 591-601.

Zdanov, A., C. Schalk-Hihi, et al. (1996). "Crystal structure of human interleukin-10 at 1.6 A resolution and a model of a complex with its soluble receptor." Protein Sci **5**(10): 1955-62.

Zhang, F., M. B. Basinski, et al. (1997). "Crystal structure of the obese protein leptin-E100." Nature **387**(6629): 206-9.

Zhang, Y., R. Proenca, et al. (1994). "Positional cloning of the mouse obese gene and its human homologue." Nature **372**(6505): 425-32.

Annexes

Annexe 1.

Jérôme Mikolajczak, Gérard Ramstein & Yannick Jacques,

Caractérisation de signatures complexes dans des familles de protéines distantes,

In *Extraction et gestion des connaissances (EGC'2004), Revue des Nouvelles Technologies de l'Information (RNTI), numéro spécial EGC'2004*, Georges Hébrail, Ludovic Lebart, Jean-Marc Petit, Editors, CEPADUES-Editions, Publisher, CLERMONT-FERRAND (FRANCE), January 2004, Vol. 2 : 317-328, ISBN: 2-85428-633-2.

Caractérisation de signatures complexes dans des familles de protéines distantes

Jérôme Mikolajczak*, Gérard Ramstein**
Yannick Jacques*

* Département de Cancérologie, Institut de Biologie
9 Quai Moncousu, F-44035 Nantes cedex
jerome.mikolajczak@nantes.inserm.fr, yjacques@nantes.inserm.fr
**IRIN, Equipe C.I.D. Ecole polytechnique de l'Université de Nantes
Rue Christian Pauc, BP 50609 44306 Nantes cedex 3
gerard.ramstein@polytech.univ-nantes.fr

Résumé. L'identification de signatures de protéines est un problème majeur pour la découverte de nouveaux membres dans des familles de protéines connues. Le concept de signature qui permet de caractériser ces familles est généralement basé sur la définition de motifs communs. Il s'avère que les familles distantes sont trop hétérogènes pour qu'on puisse identifier des régions conservées à partir des algorithmes classiques de la bioinformatique. Nous proposons une approche génétique pour la découverte de motifs hiérarchiques ; l'algorithme suit une démarche descendante en s'appuyant dans une première phase sur les classes physicochimiques des acides aminés. Les signatures sont ensuite définies par des séquences des motifs ainsi obtenus. Elles sont extraites au moyen d'un algorithme de découverte d'itemsets séquentiels où les motifs jouent le rôle d'items. Une dernière étape consiste à fouiller dans cette base d'itemsets pour n'en retenir qu'un ensemble réduit de signatures. Plusieurs stratégies sont proposées pour déterminer un ensemble optimal de signatures qui respecte des contraintes de complétude, de cardinalité et de spécificité. Nous appliquons notre démarche sur la famille des cytokines. L'analyse de la base de protéines SCOP a montré que le groupe de signatures que nous avons extrait cible spécifiquement cette famille d'intérêt.

1 Introduction

Les protéines qui constituent les briques élémentaires du vivant se regroupent par familles ayant des propriétés ou fonctions similaires. Ces molécules ont évolué dans le temps à partir d'ancêtres communs, ce qui explique la présence de motifs semblables dans les différents membres d'une même famille. Ces motifs sont des régions bien conservées au sein de la structure primaire des séquences biologiques. La structure primaire d'une protéine est représentée par une séquence $s = \langle s_1 s_2 \dots s_n \rangle$ où chaque $s_i \in \Omega$, l'ensemble des acides aminés : $\Omega = \{A,C,D,E,F,G,H,I,K,L,M,N,P,Q,R,S,T,V,W,Y\}$. Plusieurs définitions du motif ont été proposées ; la syntaxe PROSITE est la plus connue [Bucher et Bairoch, 1994]. Le motif $W[ILV]Y$ y désigne une sous-séquence constituée d'un W, suivie immédiatement par un I, L ou un V, et terminée par un Y. Le symbole

$x(i,j)$ est également utilisé pour autoriser un intervalle de taille minimale i et maximale j séparant les acides aminés du motif. Par exemple, le motif $Wx(1,3)Y$ se retrouve aussi bien dans la chaîne WFY que dans la chaîne $WAEY$.
La présence de régions conservées est d'une grande utilité pour la découverte de nouveaux membres d'une famille connue de protéines [Servant *et al.*, 2002]. La recherche de motifs biologiques est un problème majeur en bioinformatique [Dsouza *et al.*, 1997, Ramstein *et al.*, 2000, Sadowski et Parish, 2003]. On utilise le terme de signature pour désigner un motif ou un ensemble de motifs permettant de caractériser un ensemble $\mathcal{S}$ de protéines. Une signature est d'autant plus spécifique qu'elle exclut tout autre protéine. La connaissance d'une signature fournit donc une clé pour l'identification de protéines dont les fonctions sont encore inconnues.

Malheureusement, certaines familles sont si distantes qu'il est impossible d'en extraire une signature en utilisant les logiciels de découverte de motifs biologiques, tels que MEME [Bailey et Elkan, 1994] ou PRATT [Brazma *et al.*, 1996]. Notre processus de découverte procède en trois étapes. Dans la première, nous étendons l'alphabet qui compose le motif par des classes d'acides aminés. Cette description hierarchique permet l'obtention d'un ensemble de motifs. Comme la spécificité de ces derniers est insuffisante pour un criblage génomique, nous recherchons dans une deuxième étape des signatures définies par des séquences de motifs. Nous obtenons ainsi un grand nombre de signatures candidates ciblant chacune un sous-ensemble particulier de la famille d'intérêt $\mathcal{S}$. La troisième étape consiste donc à ne retenir qu'un ensemble réduit de signatures qui recouvre la totalité des membres de $\mathcal{S}$. Après avoir exposé les trois phases de notre méthode, nous terminerons par une application concernant la superfamille des cytokines.

2 Définition des signatures

Nos algorithmes reposent essentiellement sur une hierarchisation des acides aminés. Nous allons dans un premier temps décrire l'alphabet étendu que nous utilisons, avant d'aborder les concepts de motifs et de signatures.

L'ensemble Ω est subdivisé en sous-ensembles non disjoints associés à des propriétés physico-chimiques particulières. Plusieurs variantes de systèmes de classes ont été proposées ; nous avons opté pour celui présenté en table 1 [Taylor, 1986]. La pertinence de cette classification se vérifie par l'étude des régions conservées : on observe que les mutations s'opèrent généralement au sein d'une même classe (par exemple, les acides aminés I, L et V appartenant à la classe aliphatique sont très fréquemment interchangés). Soit Γ^1 le super-ensemble formé par les classes définies dans la table 1 : $\Gamma^1 = \{\Omega,\alpha,\beta,\gamma,\delta,\varepsilon,\zeta,\eta,\theta\}$. Nous appelerons $\Gamma^2 = \Gamma^1 \cup \omega$ le super-ensemble qui contient Γ^1 et le super-ensemble ω des singletons de Ω : $\omega = \{\{A\},\{C\},\ldots,\{Y\}\}$.

Un motif $m = \langle m_1 m_2 \ldots m_k \rangle$ est une k-séquence formée d'ensembles $m_i \in \Gamma^2$. Nous appelerons *occurrence* d'un motif m une sous-séquence $\langle s_{i+1} s_{i+2} \ldots s_{i+k} \rangle$ de s telle que $s_{i+j} \in m_j \ \forall j,\ 1 \leq j \leq k$. On dira que la séquence s *vérifie* le motif m. Le *support* d'un motif m dans $\mathcal{S}$ est le nombre de séquences qui vérifie m. La séquence MH vérifie

Symbole	Classe	Membres
α	aliphatique	ILV
β	aromatique	$FHWY$
γ	non polaire	$ACFGHIKLMVWY$
δ	chargé	$DEHKR$
ε	polaire	$CDEHKNQRSTWY$
ζ	charge positive	HKR
η	chaîne latérale courte	$ACDGNPSTV$
θ	chaîne latérale très courte	$ACGST$

TAB. 1 – *Classes d'acides aminés basées sur des propriétés physico-chimiques*

ainsi 21 motifs de taille 2, dont les motifs MH, $M\beta$, $\gamma\delta$, et $\Omega\Omega$. Le motif MH ne peut être vérifié que par une seule sous-séquence, tandis que le motif $\Omega\Omega$ est vérifié pour n'importe quelle séquence de taille supérieure ou égal à 2. Il importe donc de qualifier la spécificité d'un motif en prenant en compte la probabilité de le voir apparaître dans une séquence. L'estimation de cette probabilité étant coûteuse en temps de calcul, nous avons opté pour la fonction de coût suivante: $c(m) = \prod_{i=1}^{k} f(m_i)$ où $f(m_i)$ est la fréquence de la classe m_i dans une base d'apprentissage comprenant de nombreuses familles de protéines différentes. Dans la pratique, on observe une bonne corrélation entre l'estimation $c(m)$ et le support effectif de m dans la base d'apprentissage. La spécificité d'un motif m sera définie par $\phi(m) = -\log(c(m))$. La table 2 indique quelques exemples de motifs ainsi que leur spécificité et leur support.

motif	support	spécificité	fréquence
$\beta\alpha\varepsilon$	1.00	4.4	0.7984
$\alpha\delta\varepsilon\alpha$	0.96	5.1	0.6095
$\delta\Omega\gamma\varepsilon\Omega\alpha\zeta\varepsilon$	0.65	6.8	0.2427
$\Omega\alpha\zeta\varepsilon\alpha\varepsilon\varepsilon\gamma$	0.54	7.6	0.1130
LEE	0.28	7.9	0.0893
$\beta\gamma\Omega\Omega\gamma\zeta\varepsilon L$	0.28	8.4	0.0433
$\varepsilon\gamma\alpha\zeta\delta L\Omega\varepsilon$	0.37	9.2	0.0359
$L\varepsilon\varepsilon\gamma\alpha\varepsilon\delta L$	0.22	10.2	0.0107
$\Omega\eta L\alpha L\alpha\Omega L$	0.15	11.0	0.0019
$F\varepsilon R\gamma K\varepsilon\Omega\gamma$	0.15	11.4	0.0018

TAB. 2 – *Exemples de motifs avec leur support dans la famille des cytokines, leur spécificité estimée et leur fréquence effective dans la base SCOP*

Notre définition de motifs à partir de classes prédéfinies nous permet de diminuer considérablement l'espace de recherche. Notons qu'on peut toujours transformer un motif hiérarchique sous la forme PROSITE. Du motif $M\beta P$ et des sous-séquences MHP et MYP, on en déduit aisément le motif PROSITE $M[HY]P$.

Une n-signature σ est une séquence de n motifs $\langle m^1 m^2 \dots m^n \rangle$. Une séquence s vérifiant σ contient au moins n sous-séquences $(s_1, s_2 \dots, s_n)$ où chaque s^i vérifie le

motif m^i correspondant. Les positions respectives (p_1,p_2,...,p_n) de ces sous-séquences dans s sont telles que $p_{i+1} - p_i \geq |m_i| \ \forall i, 1 \leq i < n$, où $|m_i|$ désigne la taille du motif. La spécificité d'une signature sera définie par la somme des spécificités de ces motifs :

$$\phi(\sigma) = \sum_{i=1}^{k} \phi(m_i)$$

3 un algorithme génétique de découverte de motifs

Le caractère fortement combinatoire de notre problème nous a amené à utiliser les algorithmes génétiques (AG). De nombreux algorithmes de bioinformatique utilise cette métaheuristique [Sadowski et Parish, 2003]. La description hiérarchique permet une recherche selon une approche descendante qui part du motif le plus général vers le plus spécifique. Dans le cas le plus pessimiste, il existe dans toute famille $\mathcal{S}$ comprenant des k-séquences au moins un k-motif $\mu(k) = \langle m_1 m_2 \dots m_k \rangle$ avec $m_i = \Omega \ \forall i, 1 \leq i \leq k$. Ce motif trivial est le motif de plus basse spécificité qu'on puisse trouver ($\phi(\mu(k)) = 0$). L'algorithme decouverteMotifs que nous avons développé procède en deux phases distinctes. La première vise à découvrir N k-motifs décrits par les classes de Γ^1. Ces motifs généraux sont repris un par un dans une deuxième phase pour rechercher des motifs contenant d'éventuels acides aminés (motifs décrits par Γ^2). Un même AG, dénommé AGmotifs, est utilisé dans les deux phases.

fonction decouverteMotifs($\mathcal{S}$, N, k) **retourne** Population
entrées
$\mathcal{S}$, l'ensemble de séquences dont on recherche les motifs
N, le nombre de motifs
k, la taille des motifs

$p1$, $p2$, p: Population;
/* recherche des motifs de la phase 1 */
$p1 \leftarrow$ AGmotifs(N,$\mu(k)$,Γ^1,$\mathcal{S}$);
Pour tout motif m dans $p1$ faire
    /* recherche des motifs de la phase 2 */
    $p2 \leftarrow$ AGmotifs(N,m,Γ^2,$\mathcal{S}$);
    /* extraire de $p2$ l'individu le plus apte et l'inclure dans p */
    $p \leftarrow p \cup \{$meilleurIndividu($p2$)$\}$;
fait;
Retourner p;

Cette démarche descendante présente l'avantage de restreindre l'espace de recherche et d'éviter de converger vers des solutions sous-optimales induites par la rareté des motifs incluant des singletons.
AGmotifs possède deux particularités essentielles. La première est que la population initiale est formée par des individus ou motifs obtenus par copie d'un motif germe. Dans

la phase 1, le germe est le motif générique $\mu(k)$ tandis que dans la phase 2, le germe est un des N motifs extraits dans la phase précédente. La deuxième caractéristique est que l'opérateur de mutation est orienté, selon le principe de la recherche descendante : une classe Ω sera mutée en une classe physico-chimique dans les phases 1 et 2, une classe physico-chimique c sera mutée en un des membres de c en phase 2 (par exemple, α sera muté aléatoirement par la pseudo-classe $\{I\}$, $\{L\}$ ou $\{V\}$).

fonction AGmotifs(N, *germe*, *alphabet*, $\mathcal{S}$) **retourne** Population
entrées
N, le nombre de motifs recherchés
germe, le motif prototype pour l'initialisation de la première génération
alphabet, la description du motif selon Γ^1 ou Γ^2
$\mathcal{S}$, l'ensemble de séquences

parents, *population* : Population;
/* initialisation de la population */
Pour $i = 1$ à N faire *population* $\leftarrow$ *population* $\cup$ {clone(*germe*)};
Faire
calculeAdaptation(*population*, $\mathcal{S}$);
parents $\leftarrow$ selection(*population*) ;
population $\leftarrow$ reproduction(*parents*,*alphabet*) ;
jusqu'à adaptation(*population*) $\geq$ *seuil* ;
retourner *population*;

Reproduction
Deux opérateurs sont utilisés pour faire évoluer la population ; celui de la mutation que nous avons déjà décrit et celui de la recombinaison qui opère selon la technique dite de *crossing-over*. Le passage d'une génération à l'autre suit la méthode appelée *keep-best reproduction*(KBR). Cette méthode repose sur l'idée que les fils peuvent être moins bien adaptés que leurs parents et qu'il serait désastreux de les remplacer. Une solution intermédiaire consiste à supprimer le fils le moins adapté et de garder le meilleur parent. KBR assure que la nouvelle génération se voit enrichie d'un pool génétique nouveau tout en préservant le matériel génétique performant de l'ancienne génération. Selon leurs auteurs, KBR permet de trouver plus rapidement de meilleurs solutions [Wiese et Goodwin, 1999].

Fonction d'adaptation
La sélection est basée sur la technique de la roue de la fortune, principe selon lequel l'individu survit d'autant plus sûrement qu'il est bien adapté. La définition de la fonction d'adaptation est un critère déterminant dans tout AG. Dans notre application, un motif intéressant est un motif ayant une grande spécificité et un support important. Les deux critères s'opposent : un motif de faible spécificité est largement représenté (le motif $\mu(k)$ a un support égal à $C = |\mathcal{S}|$, le cardinal de la famille considérée) et inversement un motif très spécifique a peu de chances d'être identiquement conservé dans la

famille, comme le montrent les exemples de la table 2. La fonction d'adaptation qui a donné les meilleurs résultats sur notre jeu de données est la suivante :

$$a(m) = \begin{cases} 0 & \text{si } m \text{ tel que } support(m) < \tau, \\ 1 & \text{si } m = \mu(k), \\ support(m) * \phi(m)^{\lambda} & \text{sinon.} \end{cases}$$

Notre fonction d'adaptation $a(m)$ est proportionnelle au support de m et s'accroît avec la spécificité du motif. Le paramètre λ permet de moduler l'importance de la spécificité par rapport au support. Nous avons introduit un support minimal τ en deçà duquel le motif est jugé inadapté. Ce paramètre évite de retrouver des motifs de haute spécificité mais de support dérisoire (le cas limite étant évidemment celui du support 1 où toute k-sous-séquence de $\mathcal{S}$ serait retenue). Notons en effet que si $a(m) = 0$, le motif m disparaîtra à la génération suivante, puisque la sélection des individus est opérée avec une probabilité $a(m)$. C'est la raison pour laquelle nous avons attribué au motif $\mu(k)$ une adaptation non nulle, afin qu'il survive au delà de la première génération.

Redondances des motifs découverts
Nous avons légèrement modifié AGmotifs pour éviter les doublons ; les motifs délivrés par l'AG ne sont pas les individus de la dernière génération, mais les N meilleurs individus distincts obtenus, toutes générations confondues. Ce filtre trivial n'évite pas des redondances de motifs liées au recouvrement de certaines classes de Γ^1. Certains motifs sont en effet en relation de généralisation/spécialisation (comme les motifs $\alpha\Omega\gamma$ et $V\varepsilon\beta$) ou contiennent partiellement des classes qui participent de cette relation (comme les motifs $\alpha\varepsilon\beta$ et $V\Omega\gamma$). Il n'est pas possible de filtrer directement ces redondances potentielles, sans passer par une analyse précise de chacune de leurs occurrences dans les séquences. Cette étape serait fastidieuse et, d'après nos tests, non désirable. En effet, ces redondances ne concernent qu'une part infime des motifs, essentiellement des motifs de petite taille. D'autre part, on peut remarquer que ces motifs seront éliminés dans les deux phases suivantes : dans la phase de recherche des signatures (par la contrainte de non recouvrement liée à la définition même des signatures) et dans la phase de réduction de l'ensemble des signatures (par l'élimination des signatures de moindre spécificité).

4 Extraction d'un ensemble optimal de signatures

La découverte d'un ensemble de signatures ciblant une famille de séquences procède en deux étapes :

1. L'extraction d'un ensemble de signatures Σ à partir de l'ensemble des motifs $\mathcal{M}$ obtenus par l'algorithme décrit précédemment ;
2. La réduction de Σ pour n'en retenir que les membres les plus représentatifs en terme de support et de spécificité.

La première étape d'extraction de signatures s'apparente à la recherche d'itemsets séquentiels fréquents dans des séquences. Dans le champ d'application qui nous

préoccupe, le nombre forcément restreint de motifs rencontrés diminue les risques d'explosion combinatoire que nous avions rencontré dans la recherche de motifs. Nous avons adapté l'algorithme Winepi [Manilla et Toivonen, 1996] pour déterminer un ensemble de signatures candidates de support minimal τ. L'algorithme, inspiré d'AprioriAll [Agrawal et Srikant, 1995], explore tous les itemsets fréquents, en partant de ceux de taille $k = 1$ (les motifs eux-mêmes), puis par jointure, recherche itérativement ceux de taille $k = 2, 3, \ldots$ jusqu'à ce qu'il n'existe plus de signatures de support supérieur ou égal à τ. A chaque étape, nous obtenons ainsi un ensemble Σ^k de k-signatures que nous mémorisons, jusqu'à l'arrêt de l'algorithme, marqué par $\Sigma^{kmax} = \emptyset$. Nous appellerons ExtraireSignaturesCandidates($\mathcal{S}$,$\mathcal{M}$) l'algorithme qui délivre à partir de $\mathcal{S}$ et de $\mathcal{M}$ un ensemble de signatures $\Sigma = \bigcup_{k=1}^{kmax-1} \Sigma^k$.

La deuxième étape consiste à modéliser l'ensemble $\mathcal{S}$ par un sous-ensemble Σ' de Σ. Cette phase permet de représenter une famille de séquences connues sous une forme condensée. Ce modèle servira à classer des séquences inconnues dans cette famille ou à les rejeter. L'ensemble Σ' doit satisfaire les contraintes suivantes :

1. *contrainte de complétude* :
   chaque séquence $s_i \in \mathcal{S}$ vérifie au moins une signature de Σ'. Le support de Σ' doit être égal au cardinal C de la famille $\mathcal{S}$;
2. *contrainte de cardinalité* :
   le cardinal de Σ' doit être le plus petit possible (idéalement de 1) ;
3. *contrainte de spécificité* :
   la spécificité de Σ', définie comme la somme des spécificités de ces éléments, doit être la plus grande possible ;

On constatera aisément que l'algorithme ExtraireSignaturesCandidates fournit une solution de grande cardinalité qui ne garantit pas la vérification de la première condition (sauf à fixer le support minimal τ à C, ce qui induirait des signatures de faible spécificité).

Dans la pratique, on supposera que la contrainte de complétude est vérifiée par Σ (si tel n'est pas le cas, il suffit de compléter Σ par la signature $\langle \mu(k) \rangle$). Soit $\Sigma(\varphi)$ l'ensemble des signatures de Σ de spécificité supérieure ou égale à φ. Un premier prétraitement de Σ consiste à rechercher la spécificité maximale φ_{max} telle que le support de $\Sigma(\varphi_{max})$ soit égal à C, puis à éliminer de cet ensemble les signatures redondantes et de moindre spécificité.

procédure eliminerRedondances(Σ,φ)
entrées
Σ, l'ensemble des signatures candidates
φ, le seuil de spécificité maximale autorisé
$\Sigma \leftarrow \Sigma(\varphi)$;
Pour toute signature $\sigma \in \Sigma$ faire
Soit $E(\sigma) = \{s_i \in \mathcal{S}$ vérifiant $\sigma\}$;
Rechercher dans Σ s'il existe une signature $\sigma' \neq \sigma$
telle que $E(\sigma) = E(\sigma')$ et $\varphi(\sigma') > \varphi(\sigma)$. Si oui, alors $\Sigma \leftarrow \Sigma - \{\sigma\}$;
fait ;

Malgré ce filtrage, l'exploration exhaustive des sous-ensembles de Σ peut demeurer problématique. Nous proposons une stratégie permettant de rechercher rapidement une solution potentiellement satisfaisante. La méthode rechercheParCardinal est une heuristique qui vise à retenir les signatures qui "couvrent" au mieux $\mathcal{S}$. Elle consiste à prendre la signature de support maximal, puis à retenir la signature dont l'union avec la signature précédente ait un support maximal, et ainsi de suite jusqu'à ce que l'ensemble Σ_{rc} ainsi formé ait un support égal à C.

fonction rechercheParCardinal(Σ, $\mathcal{S}$) **retourne** ensembleSignatures
entrées
Σ, l'ensemble des signatures candidates
$\mathcal{S}$, l'ensemble de séquences

$\Sigma_{rc} \leftarrow \emptyset$;
Tant que support(Σ_{rc}) $\neq C$ faire
Soit $E = \{\Sigma_{rc} \cup \{\sigma_i\}, \sigma_i \in \Sigma\}$;
Soit $E' = \{e \in E$ tel que le support de e soit maximal dans $\mathcal{S}\}$;
Soit $e_{max} \in E'$ l'ensemble de signatures de spécificité maximale;
Soit σ_{max} la signature telle que $e_{max} = \{\Sigma_{rc} \cup \{\sigma_{max}\}\}$;
$\Sigma_{rc} \leftarrow \Sigma_{rc} \cup \{\sigma_{max}\}$;
$\Sigma \leftarrow \Sigma - \{\sigma_{max}\}$;
fait ;
retourner Σ_{rc};

Notons que si cette heuristique privilégie la contrainte de cardinalité, elle ne donne pas nécessairement le plus petit ensemble de signatures qui vérifie $\mathcal{S}$. Si la valeur du cardinal $cmax = |\Sigma_{rc}|$ est élevée, nous proposons de faire appel à un AG comparable à celui que nous avons utilisé dans la section 3. Nous allons rechercher des ensembles Σ' de cardinalités k décroissantes, à compter de $k = cmax$. L'algorithme s'arrête lorsqu'on ne trouve plus d'individus dont le support est supérieur ou égal à $\tilde{C}$ ($\tilde{C} = C$ pour assurer la contrainte de complétude). A l'initialisation, les individus de la première génération sont des copies du germe Σ_{rc}. Lorsqu'un individu est jugé suffisamment apte, l'AG poursuit sa recherche en décrémentant k d'une unité. Cette procédure est réalisée au moyen de la méthode *optimise* qui élimine dans chaque individu de la population une des k signatures qu'il contient. La signature sélectionnée est celle qui assure au nouvel individu ainsi formé un support maximal dans $\mathcal{S}$.

fonction AGsignatures(N,Σ,Σ_{rc}, $\tilde{C}$) **retourne** ensembleSignatures
entrées
N, la taille de la population
Σ, l'ensemble des signatures candidates
Σ_{rc}, l'ensemble des signatures obtenues par la méthode rechercheParCardinal
$\tilde{C}$, le plus petit support recherché

parents, *population*, *resultat* : Population;
/* initialisation de la population */

Pour $i = 1$ à N faire $population \leftarrow population \cup \{\text{clone}(\Sigma_{rc})\}$;
$k \leftarrow |\Sigma_{rc}|$; /* initialisation à $cmax$ */
Faire
  Faire
    calculeAdaptation($population$);
    $parents \leftarrow$ selection($population$) ;
    $population \leftarrow$ reproduction($parents$) ;
  jusqu'à adaptation($population$) $\geq seuil$;

  Soit ind, l'individu le plus adapté dans $population$
  si le support(ind) $\geq \tilde{C}$ alors $resultat \leftarrow resultat \cup \{ind\}$;
  /* initialisation de la première génération suivante */
  $population \leftarrow$ optimise($population$);
  $k \leftarrow k - 1$;
tantque support(ind) $\geq \tilde{C}$;
retourner $resultat$;

L'opérateur de mutation remplace aléatoirement une des signatures d'un individu par une signature aléatoire de Σ qui n'appartient pas à cet individu. La recombinaison utilise une technique de *crossing-over*. La fonction d'adaptation que nous avons utilisée est la suivante :

$$a(individu) = \begin{cases} support(individu) & \text{si } support(individu) < \tilde{C}, \\ \lambda\phi(individu) + \nu & \text{sinon.} \end{cases}$$

Cette définition tend à rechercher des individus qui respectent la contrainte de complétude, puis à tendre vers une solution de spécificité maximale (contrainte de spécificité). Notons que la contrainte de cardinalité est vérifiée par l'individu de cardinalité minimale dans *resultat*. Les paramètres d'ajustement $\lambda >= 1$ et $\nu >= \tilde{C}$ permettent d'accroître l'efficacité de l'AG.

5 Application à la superfamille des cytokines

Les membres de la superfamille des cytokines sont des glycoprotéines solubles de faible poids moléculaire qui interviennent dans la régulation de la réponse immunitaire. Elles ont pour fonction d'assurer la médiation des signaux de prolifération, de différenciation, et d'activation entre les différentes cibles cellulaires. La fonction pivot des cytokines dans l'activation des voies de l'immunité ainsi que dans la mort cellulaire programmée (apoptose) confère aux cytokines un intéret de tout premier plan dans la compréhension des mécanismes de la défense du soi, dans la découverte et l'amélioration les traitements anticancéreux actuels. Dans cet article, nous nous intéressons plus particulièrement aux interleukines à hélices courtes (IL-6), hélices longues (IL-2) et aux interleukines de type IL-10 dont les précurseurs possèdent six hélices. Nous avons retenu 46 séquences primaires relatives à la famille des cytokines chez l'homme.

Cardinal	Méthode	Support	Spécificité moyenne	Fréquence SCOP
11	er($\varphi = \varphi_{max}$)+rc	46	21,92	2,54
11	er($\varphi = \varphi_{max}$)+rc+ag($\tilde{c} = c$)	46	22,22	2,42
10	er($\varphi = \varphi_{max}$)+rc+ag($\tilde{c} = c$)	46	21,84	2,52
9	er($\varphi = \varphi_{max}$)+rc+ag($\tilde{c} < c$)	45	22,04	2,36
7	er($\varphi = \varphi_{max}$)+rc+ag($\tilde{c} < c$)	41	22,43	1,09
6	er($\varphi = 0$)+rc	46	16,79	10,02
11	er($\varphi = \varphi_{max} - \epsilon$)+rc	46	21,81	1,77
11	er($\varphi = \varphi_{max} - \epsilon$)+rc+ag($\tilde{c} = c$)	46	22,99	1,53
10	er($\varphi = \varphi_{max} - \epsilon$)+rc+ag($\tilde{c} = c$)	46	21,77	2.31
13	er'(support ≤ 10)+rc	46	20,59	1,21
13	er'(support ≤ 10)+rc+ag($\tilde{c} = c$)	46	22,67	0,77

TAB. 3 – *Résultats obtenus en utilisant différentes stratégies. La colonne Cardinal représente* $|\Sigma'|$*; les méthodes sont: elimineRedondances (er), rechercheParCardinal (rc), AGsignatures (ag), une version er' de er dont le filtrage est basé sur le support des signatures dans SCOP ; la colonne Support indique le nombre de séquences de* $\mathcal{S}$ *vérifiant* Σ' *; Fréquence SCOP est le pourcentage de séquences de SCOP vérifiant* Σ'*. La marge* $\epsilon = 0{,}69$ *est celle qui minimise la fréquence SCOP avec la méthode rc.*

La base d'apprentissage qui nous a servi à l'estimation de la spécificité est issu de la base de données SCOP (Structural Classification Of Proteins, [Murzin *et al.*, 1995]). Celle-ci met à la disposition des chercheurs une classification structurale des meilleures structures de protéines publiées et disponibles dans la PDB (Protein Data Bank). Les séquences de SCOP forment donc un échantillon qui recouvre un large spectre de protéines. Après suppression des interleukines, notre base d'apprentissage comporte 6615 séquences.

Des 46 séquences de cytokines, l'algorithme AGmotifs a extrait 300 motifs de taille $k = 3$ à 8. Nous avons ensuite établi les signatures candidates (le support minimum a été fixé à $\tau = 7$ pour les deux recherches). Le cardinal de Σ est égal à 40 365 avant filtrage et à 12 883 après appel de la méthode elimineRedondances (notée er). L'ensemble $\Sigma(\varphi_{max})$ comporte initialement 1654 signatures ($\varphi_{max} = 21{,}09$). Après filtrage, le jeu de signatures candidates est réduit à 740. L'heuristique rechercheParCardinal (notée rc) détermine un ensemble Σ_{rc} de 11 signatures de spécificité moyenne 21,92. Seulement 2.53% des séquences de la base SCOP vérifient Σ_{rc}. La table 3 montre que le seuil φ_{max} marque une coupure trop stricte et qu'il est préférable de baisser légèrement ce seuil. Inversement, un seuil trop bas dégrade la spécificité de Σ_{rc}, comme le montre le mauvais résultat obtenu avec $\Sigma(\varphi = 0)$. L'AG apporte un résultat pour $k = 11$ à peine meilleur que la méthode rc, mais il réduit à $k = 10$ le cardinal de Σ'.

Le résultat le plus satisfaisant, tant en spécificité (22,99) qu'en pourcentage de faux positifs (1,53%), a été obtenu en utilisant une procédure en trois phases, incluant le filtrage, l'heuristique rechercheParCardinal et l'algorithme génétique. Il est possible d'améliorer encore ces scores en ne retenant de Σ que les signatures très peu vérifiées par la base SCOP. Nous avons modifié la méthode er en remplaçant $\Sigma(\varphi)$ par l'ensemble des signatures de Σ' dont le support dans SCOP est inférieur ou égal à 10 séquences (méthode er'). Cette dernière méthode conduit à de meilleurs résultats (0,77% de faux-positifs), mais induit évidemment un biais : la base de test est utilisée lors de la phase d'apprentissage.

6 Perspectives

Les bons résultats expérimentaux que nous avons obtenu sur la superfamille des cytokines nous conforte dans l'intérêt d'une description hiérarchique des motifs. Plusieurs pistes d'améliorations peuvent cependant être apportées à nos algorithmes. L'AG de recherche de motifs n'est basé que sur deux niveaux hierarchiques ; le nombre de niveaux peut être étendu en intégrant l'imbrication de certaines classes. L'algorithme pourrait par ailleurs être rendu plus performant en autorisant les motifs candidats suffisamment adaptés à croître en taille. D'autre part, notre définition de la spécificité est perfectible ; la version actuelle ne peut être considérée comme une estimation de la probabilité de présence d'une signature. Nous avons également imposé une contrainte de complétude qui peut s'avérer coûteuse en nombre de signatures nécessaires pour caractériser une famille. En autorisant un nombre limité de faux-négatifs, on augmenterait la spécificité tout en diminuant la cardinalité de notre modèle.

Références

[Agrawal et Srikant, 1995] R. Agrawal et R. Srikant. Mining sequential patterns. In IEEE Computer Society Press 1995, editor, *Proceedings of the eleventh Intenational Conference on Data Engineering*, pages 3–14, 1995.

[Bailey et Elkan, 1994] T.L. Bailey et C. Elkan. Fitting mixture model by expectation maximization to discover motifs in biopolymers. In ISMB-94, editor, *Proceedings of the Second International Conference on Intelligent Systems for Molecular Biology*, pages 28–36. AAAI Press, 1994.

[Brazma *et al.*, 1996] A. Brazma, I. Jonassen, E. Ukkonen, et J. Vilo. Discovering patterns and subfamilies in biosequences. In ISMB-96, editor, *Proceedings of the Fourth International Conference on Intelligent Systems for Molecular Biology*, pages 34–43. AAAI Press, 1996.

[Bucher et Bairoch, 1994] P. Bucher et A. Bairoch. A generalized profile syntax for biomolecular sequences motifs and its function in automatic sequence interpretation. In ISMB-94, editor, *Proceedings of the second International Conference on Intelligent Systems for Molecular Biology*, pages 53–61. AAAI Press, 1994.

[Dsouza *et al.*, 1997] M. Dsouza, N. Larsen, et R. Overbeek. Searching for patterns in genomic data. *Trends in Genetics*, 13(12):497–498, 1997.

[Manilla et Toivonen, 1996] H. Manilla et H. Toivonen. Discovering generalized episodes using minimal occurrences. In *Knowledge Discovery and Data Mining*, pages 146–151, 1996.

[Murzin *et al.*, 1995] A.G. Murzin, S.E.Brenner, T. Hubbard, et C. Chothia. Scop: a structural classification of proteins database for the investigation of sequences and structures. *Journal of Molecular Biology*, 247:536–540, 1995.

[Ramstein *et al.*, 2000] G. Ramstein, P. Bunelle, et Y. Jacques. Discovery of ambiguous patterns in sequences: application to bioinformatics. In *Fourth European Conference of Principles of Data Mining and Knowledge Discovery*, pages 581–586, 2000.

[Sadowski et Parish, 2003] M.I. Sadowski et J.H. Parish. Automated generation and refinement of protein signatures : case study with g-protein coupled receptors. *Bioinformatics*, 19(6):727–734, 2003.

[Servant *et al.*, 2002] F Servant, C. Bru, S. Carrère, E. Courcelle, J. Gouzy, D. Peyruc, et D. Kahn. Prodom: Automated clustering of homologous domains. *Briefings in Bioinformatics*, 3(3):246–251, 2002.

[Taylor, 1986] J. Taylor. Classification of amino acid conservation. *Theorical Biology*, 119:205–218, 1986.

[Wiese et Goodwin, 1999] K Wiese et S.D. Goodwin. Convergence characteristics of keep-best reproduction. In *Proceedings of the 1999 ACM symposium on Applied computing*, pages 312–318. San Antonio, Texas, United States., 1999. February 28-March 02.

Summary

Signatures are sequences of patterns that are common to a given set of proteins. By mining genomic databases, signatures may lead to the discovery of new proteins. Unfortunately, bioinformatics algorithms fail to identify conserved regions in the case of remote protein families. This paper presents a genetic approach following a three-step process. Firstly, we extract hierarchical patterns according to a top-down strategy. Patterns are described by a new alphabet that comprise the aminoacid set as well as the set of their physicochemical classes. Secondly, an algorithm of discovery of sequential itemsets searchs for sequences of patterns. Thirdly, the set of signatures obtained in the previous step is reduced. We propose several strategies to determine on optimal set with respect to the constraints of completness, cardinality and specificity. The experimental results on the family of cytokines demonstrate the high specificity of the extracted signatures.

Annexe 2.

Jérôme Mikolajczak, Gérard Ramstein & Yannick Jacques,

Détection de faibles homologies de protéines par machines à vecteurs de support,

Classification et fouille de données (RNTI-C1), Revue des Nouvelles Technologies de l'Information, Marie Chavent, Marc Langlais, editor, CEPADUES-Editions, October 2004, Vol. 1 : 89-100, ISBN: 2-85428-667-7.

Jérôme Mikolajczak, Gérard Ramstein & Yannick Jacques,

Détection de faibles homologies de protéines par machines à vecteurs de support,

In *11èmes Rencontres de la Société Francophone de Classification, SFC'04,* Marie Chavent, Olivier Dordan, Chantal Lacomblez, Marc Langlais, Brigitte Patouille, Editors, SFC, Publisher, BORDEAUX (FRANCE), September 2004, 251-254.

Détection de faibles homologies de protéines par machines à vecteurs de support

Jérôme Mikolajczak*, Gérard Ramstein**
Yannick Jacques*

* Département de Cancérologie, Institut de Biologie
9 Quai Moncousu, F-44035 Nantes cedex
jmikolaj@nantes.inserm.fr, yjacques@nantes.inserm.fr
**LINA, équipe EGC, Ecole polytechnique de l'Université de Nantes
Rue Christian Pauc, BP 50609 44306 Nantes cedex 3
gerard.ramstein@polytech.univ-nantes.fr

Résumé. Cet article décrit une approche discriminative pour la recherche de nouveaux membres dans des familles de protéines à faibles homologies de séquences. L'originalité de la méthode repose sur une modélisation de ces familles par un ensemble M de motifs intégrant les propriétés physicochimiques des résidus. Nous proposons un algorithme de découverte de motifs suivant le paradigme de la classification hiérarchique ascendante. L'ensemble M définit un espace de représentation des séquences : chaque séquence est transformée en un vecteur indiquant la présence ou l'absence de chaque motif appartenant à M. Nous utilisons la technique d'apprentissage par machine à vecteurs de support (SVM) pour discriminer la famille d'intérêt vis à vis des séquences non apparentées. Cette méthode est testée sur la famille biologique des interleukines dont les membres possèdent des homologies de séquences faibles en dépit d'un repliement tridimensionnel en hélices alpha très conservé. Nous montrons que l'ensemble des motifs hiérarchiques modélise spécifiquement les interleukines par rapport aux autres familles structurales de la base de données SCOP (1.51). Notre classifieur est en effet plus performant sur notre famille de protéines que d'autres méthodes de classification dont le SVM basé sur les spectres de chaîne.

1 Introduction

La découverte de nouveaux membres d'une famille de protéines repose sur deux types de techniques. La plus courante est basée sur une mesure d'homologie de la protéine candidate avec un motif spécifique caractéristique de la famille d'intérêt. Cette méthode consiste à fouiller le génome à partir d'outils bioinformatiques tels que BLAST [Altschul *et al.*, 1990]. Certaines familles de protéine sont trop hétérogènes pour qu'on puisse retrouver des régions conservées au niveau de leur structure primaire. Pour lever cette difficulté, une démarche alternative a été suggérée par plusieurs auteurs [Jaakola *et al.*, 2000]. Elle est fondée sur des méthodes d'apprentissage dans lesquelles les séquences de protéines sont étiquetées selon leur appartenance ou non à la famille recherchée. Les exemples positifs (étiquette +1) regroupent les membres connus de la

famille. Les contre-exemples (étiquette -1) peuvent être extraits au sein de familles non apparentées. Une approche particulièrement prometteuse dans le domaine de la classification supervisée repose sur les machines à vecteurs de support [Vapnik, 1995] (ou *support vector machines*, nommées SVMs par la suite). Dans cette technique, le jeu d'apprentissage subit une transformation en un ensemble de vecteurs de taille fixe. Dans notre classe d'application, les séquences primaires des protéines seront donc projetées dans un espace vectoriel. Plusieurs espaces vectoriels ont été proposés avec des performances remarquables. Une méthode particulièrement efficace et rapide utilise des spectres de chaîne [Leslie *et al.*, 2002]. Ce type de représentation est abondamment utilisé en fouille de textes [Jalam et Teytaud, 2001]. Un spectre de chaîne regroupe toutes les combinaisons possibles de séquences de n caractères (ou n-gramme) à partir d'un alphabet Ω. Le spectre de chaîne d'une séquence est donc un vecteur représenté par les occurrences de ses k-sous-séquences. Il est à noter que l'espace de représentation est de haute dimension ($|\Omega|^n$ combinaisons possibles de n-grammes). La technique du spectre de chaîne est très simple à mettre en oeuvre et peu coûteuse en temps d'exécution. Les auteurs montrent que la performance de leur algorithme est comparable avec celle faisant intervenir des méthodes complexes, comme les HMMs [Karplus *et al.*, 1998]. Nos propres expérimentations sur la famille des cytokines démontrent l'efficacité de cette méthode en terme de classification. Il se trouve que notre famille d'intérêt possède des membres très éloignés entre eux en terme d'homologie de séquence. Nous proposons dans cet article d'utiliser un espace de représentation de faible dimension qui cible des propriétés spécifiques de notre famille d'intérêt. Nous allons dans un premier temps décrire le concept de motif hiérarchique, puis nous donnerons un algorithme d'extraction de ces motifs. Nous rappelerons ensuite les principes des SVMs avant de discuter des résultats obtenus sur la famille des cytokines.

2 Motifs hiérarchiques

La structure primaire d'une protéine est représentée par une séquence $s = \langle s_1 s_2 \dots s_n \rangle$ où chaque s_i appartient à Ω, l'ensemble des acides aminés :

$$\Omega = \{A,C,D,E,F,G,H,I,K,L,M,N,P,Q,R,S,T,V,W,Y\}$$

Soit $P(\Omega)$ l'ensemble des parties de Ω. Certaines de ces parties possèdent des résidus partageant des propriétés physico-chimiques particulières. Plusieurs variantes de systèmes de classes ont été proposées ; nous avons opté pour celui de Taylor [Taylor, 1986] présenté en table 1. La pertinence de cette classification se vérifie par l'étude des régions conservées : on observe que les mutations s'opèrent généralement au sein d'une même classe (par exemple, les acides aminés I, L et V appartenant à la classe aliphatique sont très fréquemment interchangés). Les classes physico-chimiques définissent un sous-ensemble de $P(\Omega)$, auquel nous ajoutons l'ensemble des singletons de Ω ainsi que l'ensemble Ω lui-même. Nous noterons $C(\Omega)$ l'alphabet suivant :

$$C(\Omega) = \{\{A\},\{C\},\dots,\{Y\}\} \cup \{\alpha,\beta,\gamma,\delta,\varepsilon,\zeta,\eta,\theta\} \cup \Omega$$

On considérera l'ensemble ordonné $(C(\Omega), \subseteq)$ qui forme un sup-demi-treillis : toute paire (x,y) de $C(\Omega) \times C(\Omega)$ possède une borne supérieure, que l'on notera $sup(x,y)$.

Symbole	Classe	Membres
α	aliphatique	ILV
β	aromatique	$FHWY$
γ	non polaire	$ACFGHIKLMVWY$
δ	chargé	$DEHKR$
ε	polaire	$CDEHKNQRSTWY$
ζ	charge positive	HKR
η	chaîne latérale courte	$ACDGNPSTV$
θ	chaîne latérale très courte	$ACGST$

TAB. 1 – *Classes d'acides aminés basées sur des propriétés physico-chimiques*

Un motif $m = \langle m_1 m_2 \dots m_k \rangle$ est une k-séquence formée d'ensembles $m_i \in C(\Omega)$. Pour la simplicité de la notation, on notera le singleton $\{R\}$ par R directement ; le motif $K\alpha$ désignera ainsi le motif composé de la classe $\{K\}$ suivie de la classe $\{I,L,V\}$.

Bien que rien n'interdise dans notre méthode d'utiliser des motifs de taille k variable, nous supposerons pour la simplicité de l'exposé que k est fixe. Nous appelerons *occurrence* d'un motif m une sous-séquence $\langle s_{i+1} s_{i+2} \dots s_{i+k} \rangle$ de s telle que $s_{i+j} \in m_j \ \forall j,\ 1 \leq j \leq k$. On dira que la séquence s *vérifie* le motif m. Le *support* d'un motif m dans un jeu de séquences $\mathcal{S}$ est le nombre de séquences de $\mathcal{S}$ qui vérifie m. La séquence MH vérifie ainsi 21 motifs de taille 2, dont les motifs MH, $M\beta$, $\gamma\delta$, et $\Omega\Omega$. Le motif MH ne peut être vérifié que par une seule sous-séquence, tandis que le motif $\Omega\Omega$ est vérifié pour n'importe quelle séquence de taille supérieure ou égale à 2. Il importe donc de qualifier la spécificité d'un motif en prenant en compte la probabilité de le voir apparaître dans une séquence. Comme l'estimation précise de cette probabilité est complexe et inutile pour notre classifieur, nous avons opté pour la fonction de coût suivante : $c(m) = \prod_{i=1}^{k} f(m_i)$ où $f(m_i)$ est la fréquence de la classe m_i dans une base d'apprentissage comprenant de nombreuses familles de protéines différentes. La spécificité d'un motif m sera définie par $\phi(m) = -\log(c(m))$.

La table 2 montre qu'on observe une bonne corrélation entre l'estimation $\phi(m)$ et le support effectif de m dans la base de contre-exemples de séquences issues de SCOP [Murzin *et al.*, 1995].

Nous avons défini deux propriétés d'un motif hiérarchique : son support et sa spécificité. Il est important de remarquer que ces deux caractéristiques sont généralement opposées. Plus un motif possède un support élevé, moins il est spécifique, comme le montre l'exemple extrême d'un motif composé uniquement de la classe Ω. A l'inverse, un motif uniquement composé de singletons est fortement spécifique mais aura peu de chances d'être découvert.

Les motifs peuvent être hiérarchisés selon une relation de généralisation. Soit deux k-motifs m^1 et m^2. Nous noterons $\preceq$ la relation d'ordre suivante :

$m^1 \preceq m^2$ ssi pour tout $i \in [1,k]$ on a $m_i^1 \subseteq m_i^2$. L'estimateur $\phi(m)$ est construit de sorte que $\phi(m^1) \geq \phi(m^2)$ pour toute paire de motifs vérifiant $m^1 \preceq m^2$. En spécialisant un motif, on augmente sa spécificité. Nous appelerons borne supérieure des motifs m^1 et m^2 (notée $sup(m^1,m^2)$) le motif $m^{1,2}$ vérifiant :

$m_i^{1,2} = sup(m_i^1, m_i^2)$ pour tout $i \in [1,k]$.

motif	support	spécificité	fréquence
$\beta\alpha\varepsilon$	1.00	4.4	0.7984
$\alpha\delta\varepsilon\alpha$	0.96	5.1	0.6095
$\delta\Omega\gamma\varepsilon\Omega\alpha\zeta\varepsilon$	0.65	6.8	0.2427
$\Omega\alpha\zeta\varepsilon\alpha\varepsilon\varepsilon\gamma$	0.54	7.6	0.1130
LEE	0.28	7.9	0.0893
$\beta\gamma\Omega\Omega\gamma\zeta\varepsilon L$	0.28	8.4	0.0433
$\varepsilon\gamma\alpha\zeta\delta L\Omega\varepsilon$	0.37	9.2	0.0359
$L\varepsilon\varepsilon\gamma\alpha\varepsilon\delta L$	0.22	10.2	0.0107
$\Omega\eta L\alpha L\alpha\Omega L$	0.15	11.0	0.0019
$F\varepsilon R\gamma K\varepsilon\Omega\gamma$	0.15	11.4	0.0018

TAB. 2 – *Exemples de motifs avec leur support dans la famille des cytokines, leur spécificité estimée et leur fréquence effective dans la base SCOP.*

Le motif $m^{1,2}$ représente le motif le plus spécifique qui généralise m^1 et m^2 : toute sous-séquence vérifiant m^1 ou m^2 vérifiera $m^{1,2}$. La table 3 donne des exemples de bornes supérieures pour des motifs de taille 4. La section suivante présente comment extraire les motifs de spécificité minimale.

3 Découverte de motifs hiérarchiques

La recherche de motifs hiérarchiques procède en deux étapes :

1. L'extraction de motifs germes,
2. La génération des motifs hiérarchiques.

La première étape consiste à extraire des motifs germes à partir de la famille $\mathcal{S}$. Un motif germe est un motif qui ne possède pas de minorants. Plus pratiquement, les motifs germes sont les motifs formés uniquement de classes singletons (résidus). Une façon triviale d'obtenir la liste des motifs germes est de relever l'ensemble des k-sous-séquences présentes dans le jeu d'apprentissage. Nous avons procédé à un filtrage en ne considérant que les k-sous-séquences potentiellement intéressantes. Cette première étape consiste à croiser chaque k-sous-séquence avec une autre : chaque couple de k-sous-séquences fournit un motif. Si ce dernier possède un support suffisant, les k-sous-séquences vérifiant ce motif sont retenues. Avec un support minimal de 3, nous avons pu réduire ainsi le nombre de motifs germes d'un facteur 8.

La seconde étape opère un appariement des motifs pour déterminer leurs bornes supérieures. L'algorithme de découverte de motifs consiste donc à généraliser les motifs afin de rechercher des caractéristiques de support suffisamment représentatives. La définition précédente de borne supérieure permet de définir le motif commun le plus spécifique à partir de deux motifs. Il est donc possible par itérations successives d'agréger des motifs afin d'obtenir des motifs de support supérieur. La deuxième étape de notre algorithme s'inspire de la technique de la classification hiérarchique ascendante pour former des clusters de motifs généraux à partir de motifs germes composés

uniquement de singletons. La deuxième étape de l'algorithme de découverte de motifs est présentée ci-dessous :

algorithme découverteMotifs
entrées
M, l'ensemble de motifs germes obtenus lors de l'étape 1
$supMin$, le seuil de support minimal recherché
$speMin$, le seuil de spécificité minimale recherchée
sortie
E, l'ensemble de motifs hiérarchiques

$E = \{m \in M \mid support(m) \geq supMin \text{ et } \phi(m) \geq speMin\}$;
Répéter
Soit m^1 et m^2 la paire de motifs de M telle que :
1. $m^{1,2} = sup(m^1, m^2)$
2. $\phi(m^{1,2}) \geq \phi(m^{i,j})$ pour tout m^i et m^j dans M
$M \leftarrow M - \{m^1, m^2\}$;
$M \leftarrow M \cup \{m^{1,2}\}$;
si $support(m^{1,2}) \geq supMin$ et $\phi(m^{1,2}) \geq speMin$
alors $E \leftarrow E \cup \{m^{1,2}\}$;
jusqu'à $cardinal(M) = 1$ ou $\phi(m^{1,2}) < speMin$;

La table 3 présente les motifs hiérarchiques de taille $k = 4$ relatifs aux séquences HIWY, HIDY, KLTY, HVSG et DARG. Les motifs m^1 à m^5 sont les motifs germes extraits des sous-séquences de taille 4 dans le jeu de séquences précédent. La figure 1 montre la construction hiérarchique des motifs par l'algorithme découverteMotifs. A supposer que l'on se soit fixé un support minimal de 3 et une spécificité de 3.0, seul le motif m^6 serait retenu.

motif	chaîne	support	spécificité
m^1	HIWY	1	14.25
m^2	HIDY	1	12.83
m^3	KLTY	1	11.44
m^4	HVSG	1	11.79
m^5	DARG	1	10.96
$m^6 = m^{1,2}$	KIεY	2	10.64
$m^7 = m^{3,6}$	$\zeta\alpha\varepsilon$Y	3	7.55
$m^8 = m^{4,5}$	$\delta\eta\varepsilon$G	2	5.26
$m^9 = m^{7,8}$	$\delta\Omega\varepsilon\gamma$	5	2.56

TAB. 3 – *Motifs extraits par l'algorithme découverteMotifs.*

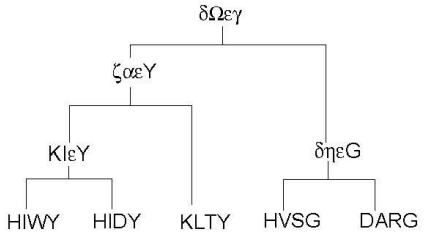


FIG. 1 – *Hiérarchie de motifs à partir de 5 motifs germes*

Notre algorithme diffère dans sa finalité des techniques usuelles d'extraction de motifs. Alors que ces dernières visent à découvrir un nombre restreint de motifs ayant le support le plus large possible (ce qui s'avère quasi impossible pour des familles fortement hétérogènes), nous cherchons au contraire un ensemble élevé M de motifs au support très variable. Le caractère hiérarchique de l'algorithme nous permet de trouver un nombre conséquent de motifs tout en s'affranchissant des problèmes de combinatoire.

Pour utiliser les machines à vecteurs de support, nous allons transformer une séquence quelconque en un vecteur booléen de dimension n (n désignant le cardinal de M). L'élément de rang i est à vrai ssi le motif de rang i dans M est présent dans la séquence (une alternative consiste à compter les occurrences des motifs, mais dans la pratique, il est fort rare de rencontrer deux fois le même motif dans une même séquence pour des valeurs de k importantes). Il est à noter que cette vectorisation s'effectue en $O(N)$, où N désigne la taille de la séquence. Si l'apprentissage est relativement complexe, la classification est peu coûteuse en temps d'exécution. La table 4 montre les valeurs de vecteurs associés aux séquences de l'exemple précédent en considérant l'ensemble des motifs donnés en table 3 (support et spécificité minimaux fixés à 0).

séquence	représentation vectorielle
HIWY	100001101
HIDY	010001101
KLTY	001000101
HVSG	000100011
DARG	000010011

TAB. 4 – *Représentation vectorielle des séquences à partir des motifs extraits. Le vecteur représente neuf valeurs booléennes (1 pour vrai, 0 pour faux) correspondant à la présence du motif m^i, $i = 1$ à 9, dans la séquence considérée.*

4 Machines à vecteurs de support

Cette section introduit la méthode des machines à vecteurs de support [Vapnik, 1995], [Vapnik, 1998]. Cette technique de classification supervisée est basée sur l'apprentissage d'une frontière de décision linéaire qui discrimine au mieux deux classes formées par des exemples et des contre-exemples. Le SVM optimise la frontière de décision par un hyperplan qui maximise la distance entre les points voisins (ou vecteurs de support) durant la phase d'apprentissage. Dans la phase de classification, un point est classé en exemple ou contre-exemple selon sa position par rapport à la frontière de décision. Dans la plupart des problèmes concrets, il y a peu de chances qu'on puisse trouver une séparation linéaire parfaite dans l'espace χ des données. L'efficacité remarquable des SVMs repose sur une transformation non linéaire de l'espace d'entrée χ en un nouvel espace $\Phi(\chi)$ de redescripteurs de plus grande dimension dans lequel les points sont séparables par un hyperplan.

Des travaux théoriques ont permis de montrer que le problème d'optimisation des SVMs revient à résoudre un problème de forme duale qui dépend des produits scalaires entre les vecteurs des exemples $\Phi(x)$ de l'ensemble d'apprentissage dans l'espace de redescription. De ce fait, toute la difficulté repose sur le calcul des produits scalaires $< \Phi(x),\Phi(y) >$ entre chaque couple (x,y) de vecteurs d'exemples dans un espace de redescription $\Phi(\chi)$ de dimension très élevée, voire infinie. Cette difficulté est levée grâce à l'introduction de fonctions bilinéaires symétriques positives $K(x,y)$ appelées *fonctions noyaux*. Les fonctions noyaux permettent d'effectuer tous les calculs nécessaires dans l'espace des données sans jamais devoir passer par l'espace des descripteurs : $K(x,y) =< \Phi(x),\Phi(y) >$. Il existe trois types de fonctions noyaux K simples : les fonctions polynomiales, les fonctions à bases radiales (RBF) et les fonctions sigmoïdes. Les fonctions polynomiales sont définies par $K(x,y) = (x.y + 1)^p$. Le degré du polynome est choisi par l'utilisateur. Les fonctions à bases radiales RBF sont définies par $K(x,y) = e^{\|x-y\|^2/2\sigma^2}$. La valeur de l'écart type σ est estimée empiriquement. Les fonctions sigmoïdes sont de la forme $K(x,y) = tanh(a(x.y - b)$. En pratique, il s'agit de tester les différentes fonctions noyaux pour déterminer celle avec laquelle on obtient l'hyperplan optimal et la marge minimale. Dans le cadre de ce travail, nous avons testé la fonction noyau linéaire (polynome de degré 1) et les fonctions à bases radiales.

La première application des SVMs pour la classification d'homologues éloignées dans les familles protéiques est la méthode des Fisher SVMs [Jaakola *et al.*, 2000]. Cette méthode nécessite tout d'abord l'apprentissage d'un modèle de Markov caché (HMM) sur la famille de protéines d'intéret [Karplus *et al.*, 1998]. Ensuite un jeu d'apprentissage constitué d'exemples et de contre-exemples est représenté dans un espace vectoriel dont chaque dimension est liée à un état ou une transition présent dans le modèle. La valeur de chaque dimension du vecteur est appelée score de Fisher et représente le taux d'utilisation de chaque paramètre du modèle pour modéliser chaque séquence exemple. Les vecteurs obtenus sont utilisés conjointement avec une fonction noyau particulière, appelée fonction de Fisher, pour l'apprentissage du modèle SVM. Les auteurs ont montré que les classifications par la méthode des Fisher SVMs surpassent les méthodes génératives telles que les HMMs ou BLAST. Cependant cette méthode trouve ses limites dans la nécessité d'avoir une famille de protéines possédant

un nombre important de membres pour pouvoir élaborer le modèle HMM. Elle se caractérise aussi par une complexité temporelle importante autant en phase d'apprentissage que lors de la prédiction de la classe d'une protéine. La méthode des pairwise SVMs [Li et Noble, 2003] utilise l'algorithme d'alignement local de Smith et Waterman. Les exemples sont vectorisés dans un espace dont la dimension correspond au cardinal du jeu d'apprentissage. Chaque dimension du vecteur correspond au score de similarité obtenu par l'alignement local avec une séquence du jeu d'apprentissage. Les auteurs de cette méthode ont obtenu des performances de classification supérieure à la méthode des Fisher SVMs. Cependant elle ne résoud pas le problème de la complexité temporelle et rend la dimension des vecteurs dépendante de la taille du jeu d'apprentissage. Une alternative au problème de complexité temporelle est le spectre de chaîne ou *String Spectrum* [Leslie *et al.*, 2002], dont nous avons présenté le principe en introduction. Dans cette méthode, chaque vecteur représente les fréquences des n-grammes trouvés sur la séquence de l'exemple. Les vecteurs servent ensuite à l'apprentissage du modèle SVM. Ce simple procédé s'affranchit des méthodes génératives et permet un gain de temps d'exécution remarquable. Une adaptation de cette méthode prend en compte tous les motifs de taille n fixée et autorise au plus m variations entre deux motifs [Leslie *et al.*, 2004]. Cette adaptation prend en compte le concept biologique de mutation des résidus. Les auteurs de cette technique montrent que la performance de leur classification est comparable à la méthode des Fisher SVMs tout en optimisant les temps d'exécution. Nos propres expérimentations n'ont pas révélé une amélioration des performances sur notre famille d'intérêt par rapport à la méthode de spectre de chaîne. C'est donc cette dernière méthode qui nous servira de point de comparaison avec notre propre algorithme de classification. Les méthodes décrites jusqu'à présent ne prennent en compte que l'information située au niveau de la structure primaire des protéines. La méthode des SVM-I-sites [Hou *et al.*, 2003] tente d'intégrer une information structurale lors de l'étape de vectorisation des exemples. Une étape préalable consiste à rechercher un profil de séquences à partir de la séquence primaire de l'exemple. Le profil est obtenu au moyen du logiciel PSI-BLAST [Altschul *et al.*, 1997] et de la base de données Swiss-Prot [Boeckmann *et al.*, 2003]. La création des vecteurs se base ensuite sur la recherche des occurences de 263 domaines structuraux au sein du profil. Les auteurs indiquent que cette méthode apporte des résultats équivalents à la méthode des pairwise SVMs. Elle est aussi plus efficace lors de l'étape de vectorisation.

Les SVMs ont démontré leur efficacité dans la détection d'homologies éloignées. Les différents types de classification présentés mettent en évidence l'importance de l'étape de vectorisation des exemples dans la performance de ce type de classifieur.

5 Classification des protéines utilisant les SVMs

La classification des protéines passe par une première étape de vectorisation suivie de la prédiction proprement dite par SVM. Dans notre application, la transformation des séquences s'opère par la détection des motifs retenus par l'algorithme découverteMotifs. Les séquences étant représentées par des vecteurs booléens de taille fixe, il est possible de définir le produit scalaire entre deux séquences s et s' par le nombre d'éléments à vrai dans le vecteur $x \wedge x'$ où x et x' désignent les vecteurs

booléens associés respectivement à s et s'. Le choix de la fonction noyau a été dictée par certaines propriétés spécifiques à notre espace de redescription. En effet, soit O le vecteur nul. La fonction noyau doit vérifier les points suivants :

- si $x \simeq O$ et $x' \simeq O$ alors $K(x,x')$ doit avoir une valeur proche de la valeur maximale de la fonction noyau. En effet, dans ce cas, les deux séquences appartiennent toutes deux à la classe des contre-exemples.
- si $x \simeq O$ et x' appartient à la classe des exemples, alors la valeur de $K(x,x')$ doit être d'autant plus faible que l'exemple contient un nombre de motifs important.

Nous avons retenu les fonctions à bases radiales qui vérifient bien ces propriétés (parce qu'elles reposent sur le calcul de $\| x - x' \|$, contrairement aux fonctions polynomiales basées sur le calcul de $x.x'$). Il est à noter que les fonctions à bases radiales donnent généralement de meilleurs résultats que les fonctions polynomiales et sigmoïdes.

Si l'utilisation des SVMs ne pose aucun problème d'adaptation, il est cependant important d'observer qu'à l'origine les SVMs définissent une frontière optimale entre deux classes d'individus d'intérêt égal. Dans notre type d'application, l'importance égale de ces deux classes soulève une difficulté. S'il est en effet facile de définir ce qu'est la famille d'intérêt, la définition des contre-exemples est moins évidente. Une solution consiste à d'utiliser des représentants dans chaque superfamille défini par la base SCOP. Il faut alors veiller à la robustesse de ces choix et à la sensibilité des SVMs face au déséquilibre entre le nombre des exemples et celui des contre-exemples. Nous avons effectué des tests portant sur des échantillons aléatoires de contre-exemples de même taille que le jeu d'exemples positifs et répartis uniformément parmi les classes structurales de la base SCOP. Ces tests montrent que les performances de la classification sont peu sensibles au tirage effectué, mais se dégradent lorsque le nombre de contre-exemples s'accroît par rapport au nombre d'exemples positifs. Une autre solution consiste à utiliser la version de SVMs basée sur une classe unique, proposée par B. Schölkopf [Schölkopf *et al.*, 2001]. Les résultats obtenus avec cette méthode se sont montrés inférieurs à ceux des SVMs à deux classes. Nous interprétons ce phénomène par le nombre restreint d'exemples présenté en apprentissage ainsi qu'à leur grande dispersion (faible homologie intra-classe).

Nous avons utilisé le logiciel open-source libsvm [C.Chang et Lin, 2001] pour implémenter notre classifieur SVM.

6 Application à la superfamille des cytokines

Les cytokines ont pour fonction d'assurer la médiation des signaux de prolifération, de différenciation, et d'activation entre les différentes cibles cellulaires. Dans cet article, nous nous intéressons plus particulièrement aux interleukines à hélices courtes (IL-2), hélices longues (IL-6) et aux interleukines de type IL-10 dont les précurseurs possèdent six hélices. Nous avons retenu 45 séquences primaires relatives à la famille des cytokines chez l'homme.
La base d'apprentissage qui nous a servi à l'estimation de la spécificité est issue de la base de données SCOP (Structural Classification Of Proteins, [Murzin *et al.*, 1995]). Les séquences de SCOP forment donc un échantillon qui recouvre un large spectre

de protéines. Après suppression des interleukines, notre base de test comporte 6615 séquences.

La famille qui nous intéresse est caractérisée par une structure secondaire riche en hélice alpha. Le pas de cycle d'une hélice étant de 3.6, nous avons considéré des motifs de taille 8 (des motifs de taille 4 seraient moins discriminants).

La table 5 présente les résultats obtenus avec différents classifieurs. Les valeurs présentées sont des moyennes de performances obtenues par la technique de *leave-one-out* (pour laquelle une séquence sert de test et les autres pour l'apprentissage). La ligne KNN présente les résultats obtenus avec la méthode des k plus proches voisins ($k = 3$). La notion de voisinage se réfère à la proximité entre spectres de chaîne : la mesure de similarité utilisée est le produit cartésien des vecteurs normalisés. Les SVMs à base de spectre de chaînes donnent de meilleurs résultats (ligne SCSVM) que les KNNs, ce qui confirme l'intérêt des machines à vecteurs de support. Les résultats sont identiques pour les fonctions noyaux linéaires et à bases radiales ; la seule différence consiste en une légère amélioration des performances sur la base SCOP pour la fonction à base radiale (table 6). Notre méthode MotifsSVM surpasse largement la technique à base de spectre de chaîne (100% de bonne classification) à condition de bien optimiser le type des motifs. Un seuil de spécificité de 14 apparaît comme le meilleur compromis ; au delà de cette valeur, on ne découvre pas assez de motifs sur certaines cytokines, en deçà, les motifs ne sont assez sélectifs. Un seuil de support minimal de 2 en *leave-one-out* (de 3 en apprentissage normal) est nécessaire pour obtenir de bons résultats dans notre jeu particulier d'application. Il s'avère que les exemples présentés possèdent peu de séquences proches. Nous avons pu vérifier ce fait en calculant les valeurs de fonction noyau à partir de spectres de chaîne : la valeur de fonction noyau tend rapidement vers 0 à partir de la deuxième plus proche voisine.

classifieur	taux d'erreurs	VP	FN	VN	FP
KNN	18.9	88.9	11.1	73.3	26.7
SCSVM linéaire	13.3	84.4	15.6	88.9	11.1
SCSVM RBF	13.3	84.4	15.6	88.9	11.1
MotifsSVM 13	2.2	95.6	4.4	100	0
MotifsSVM 14	0	100	0	100	0
MotifsSVM 15	5.5	88.9	11.1	100	0

TAB. 5 – *Résultats de classification. VP, FN, VN, FP sont les pourcentages respectivement des vrais positifs, faux négatifs, vrais négatifs, faux positifs.*

La table 6 indique le pouvoir discriminant de notre classifieur sur les séquences négatives extraites de la base SCOP. Sur les 6615 séquences de la base SCOP, le classifieur MotifsSVM a en mal classé 8. Le faible pourcentage de faux-positifs autorise l'emploi de MotifsSVM pour rechercher de nouveaux membres dans le génome.

classifieur	taux de FP dans SCOP
SCSVM linéaire	4.32
SCSVM RBF	4.08
MotifsSVM 14	0.12

TAB. 6 – *Taux de faux-positifs dans la base SCOP.*

7 Conclusion et perspectives

Les excellents résultats obtenus en classification dans la famille des cytokines démontrent la pertinence d'une description hiérarchique des motifs. Ces derniers assurent un rôle de signature, au sens où ils sont spécifiques à la famille étudiée. Nous avons proposé un paramétrage simple du degré de spécificité souhaité et avons observé qu'une valeur moyennement haute donne les meilleurs performances ($\phi(m) \sim 14$). La capacité des SVMs à gérer des espaces de grande dimension nous permet d'obtenir une classification sans erreurs sur les exemples positifs et un très faible taux d'erreurs sur les exemples négatifs issus de SCOP (0.12%). Certaines améliorations de l'algorithme d'extraction restent à mettre en oeuvre, afin d'optimiser le nombre de motifs retenus. Sur les quelques 600 motifs extraits, une proportion non négligeable d'entre eux peuvent certainement être éliminés sans nuire à la performance du classifieur. Nous envisageons dans un premier temps de filtrer plus finement les motifs extraits et de procéder ensuite à une étude de leurs co-occurrences dans les séquences. Cette analyse permettra de déterminer des patrons de motifs propres à une famille de protéines (un patron est défini comme une séquence de motifs situés à des intervalles variables sur une même séquence).

Références

[Altschul *et al.*, 1990] S. F. Altschul, W. Gish, W. Miller, E. W. Myers, et D. J. Lipman. A basic local alignment search tool. *Journal of Molecular Biology*, 215:403–410, 1990.

[Altschul *et al.*, 1997] S. F. Altschul, T. L. Madden, A. A. Schäffer, J. Zhang, Z. Zhang, W. Miller, et D. J. Lipman. Gapped blast and psi-blast : a new generation of protein database search programs. *Nucleic Acid Research*, 25(3389-3402):17, 1997.

[Boeckmann *et al.*, 2003] B. Boeckmann, A. Bairoch, R. Apweiler, M.-C. Blatter, A. Estreicher, E. Gasteiger, M. J. Martin, K. Michoud, C. O'Donovan, I. Phan, S. Pilbout, et M. Schneider. The swiss-prot protein knowledge base and its supplement trembl in 2003. *Nucleic Acid Research*, 31:365–370, 2003.

[C.Chang et Lin, 2001] C. C.Chang et C. J. Lin. *LIBSVM: a library for support vector machines*, 2001. Software available at `http://www.csie.ntu.edu.tw/~cjlin/libsvm`.

[Hou *et al.*, 2003] Y. Hou, W. Hsu, M. L. lee, et C. Bistroff. Efficient remote homology detection using local structure. *Bioinformatics*, 19(17):2294–2301, 2003.

[Jaakola *et al.*, 2000] T. Jaakola, M. Diekhans, et D. Haussler. A discriminative framework for detecting remote protein homologies. *Journal of Computationnal Biology*, 7(1-2):95–114, 2000.

[Jalam et Teytaud, 2001] Radwan Jalam et Olivier Teytaud. Identification de la langue et catégorisation de textes basées sur les n-grammes. *Extraction de Connaissance et Apprentissage*, 1(1-2):227–238, Janvier 2001.

[Karplus *et al.*, 1998] K. Karplus, C. Barret, et R. Hugley. Hidden markov models for detecting remote protein homologies. *Bioinformatics*, 14:846–856, 1998.

[Leslie *et al.*, 2002] C. Leslie, E. Eskin, et W. S. Noble. The spectrum kernel : a string kernel for svm protein classification. In *Proceedings of the Pacific Biocomputing Symposium*, pages 564–575, 2002.

[Leslie *et al.*, 2004] C. Leslie, E. Eskin, D. Zhou, et W. S. Noble. Mismatch string kernel for svm protein classification. *Bioinformatics*, 2004. à paraître.

[Li et Noble, 2003] L. Li et W. S. Noble. Combining pairwise sequence similarity and support vector machines for detecting remote protein evolutionnary and structural relationships. *Journal of Computationnal Biology*, 10(6):857–868, 2003.

[Murzin *et al.*, 1995] A.G. Murzin, S.E.Brenner, T. Hubbard, et C. Chothia. Scop: a structural classification of proteins database for the investigation of sequences and structures. *Journal of Molecular Biology*, 247:536–540, 1995.

[Schölkopf *et al.*, 2001] B. Schölkopf, J. Platt, J. Shawe-Taylor, A. J. Smola, et R. C. Williamson. Estimating the support of a high-dimensional distribution. *Neural Computation*, 13(7), 2001.

[Taylor, 1986] J. Taylor. Classification of amino acid conservation. *Theorical Biology*, 119:205–218, 1986.

[Vapnik, 1995] V. N. Vapnik. *The nature of statistical learning theory*. Springer-Verlag, 1995.

[Vapnik, 1998] V. N. Vapnik. *Statistical Learning Theory*. Springer, 1998.

Summary

This article presents a discriminative approach to the protein classification in the particular case of remote homology. The protein family is modelled by a set M of motifs related to the physicochemical properties of the residues. We propose an algorithm for discovering motifs based on the ascending hierarchical classification paradigm. The set M defines a feature space of the sequences : each sequence is transformed into a vector that indicates the possible presence of the motifs that belongs to M. We then use the SVM learning method to discriminate the target family. Our hierarchical motif set specifically modelises interleukins among all the structural families of the SCOP (1.51) database. Our method yields significantly better remote protein classification compared to spectrum kernel techniques.

Annexe 3.

Jérôme Mikolajczak, Gérard Ramstein & Yannick Jacques,

Classification de protéines distantes par motifs hiérarchiques,

In *5èmes Journées Ouvertes Biologie Informatique et Mathématiques (JOBIM 2004)*,

MONTREAL (CANADA), June 2004, (actes disponibles sur CD-ROM).

Classification de protéines distantes par motifs hiérarchiques

Jérôme Mikolajczak*, Gérard Ramstein**
Yannick Jacques*

* INSERM U601, Département de Cancérologie, Institut de Biologie
9 Quai Moncousu, F-44035 Nantes cedex
jmikolaj@nantes.inserm.fr, yjacques@nantes.inserm.fr
**LINA, équipe EGC, Ecole polytechnique de l'Université de Nantes
Rue Christian Pauc, BP 50609 44306 Nantes cedex 3
gerard.ramstein@polytech.univ-nantes.fr

Résumé. Cet article décrit une approche discriminative pour la recherche de nouveaux membres dans des familles de protéines à faibles homologies de séquences. L'originalité de la méthode repose sur une modélisation de ces familles par un ensemble M de motifs intégrant les propriétés physicochimiques des résidus. Nous proposons un algorithme de découverte de motifs suivant le paradigme de la classification hiérarchique ascendante. L'ensemble M définit un espace de représentation des séquences : chaque séquence est transformée en un vecteur indiquant la présence ou l'absence de chaque motif appartenant à M. Nous utilisons la technique d'apprentissage par machine à vecteurs de support (SVM) pour discriminer la famille d'intérêt vis à vis des séquences non apparentées. Cette méthode est testée sur la famille biologique des interleukines dont les membres possèdent des homologies de séquences faibles en dépit d'un repliement tridimensionnel en hélices alpha très conservé. Nous montrons que l'ensemble des motifs hiérarchiques modélise spécifiquement les interleukines par rapport aux autres familles structurales de la base de données SCOP (1.51)[Murzin *et al.*, 1995]. Notre classifieur est en effet plus performant sur notre famille de protéines que d'autres méthodes de classification dont le SVM basé sur les spectres de chaîne.

1 Introduction

La découverte de nouveaux membres d'une famille de protéines repose sur deux types de techniques. La plus courante est basée sur une mesure d'homologie de la protéine candidate avec un motif spécifique caractéristique de la famille d'intérêt. Cette méthode consiste à fouiller le génome à partir d'outils bioinformatiques tels que BLAST [Altschul *et al.*, 1990]. Certaines familles de protéine sont trop hétérogènes pour qu'on puisse retrouver des régions conservées au niveau de leur structure primaire. Pour lever cette difficulté, une démarche alternative a été sugggérée par plusieurs auteurs [Jaakola *et al.*, 2000]. Elle est fondée sur des méthodes d'apprentissage par l'exemple et le concept de discrimination : les séquences de protéines sont étiquetées selon leur appartenance ou non à la famille recherchée. Les exemples positifs (étiquette +1) regroupent

les membres connus de la famille. Les contre-exemples (étiquette -1) peuvent être extraits au sein de familles non apparentées. Une approche particulièrement prometteuse dans le domaine de la classification supervisée repose sur les machines à vecteurs de support [Vapnik, 1995] (ou *support vector machines*, nommées SVMs par la suite). Dans cette technique, le jeu d'apprentissage subit une transformation en un ensemble de vecteurs de taille fixe. Dans notre classe d'application, les séquences primaires des protéines seront donc projetées dans un espace vectoriel. Plusieurs espaces vectoriels ont été proposés avec des performances remarquables. Une méthode particulièrement efficace et rapide utilise des spectres de chaîne [Leslie *et al.*, 2002]. Ce type de représentation est abondamment utilisé en fouille de textes [Jalam et Teytaud, 2001]. Un spectre de chaîne regroupe toutes les combinaisons possibles de séquences de n caractères (ou n-mot) à partir d'un alphabet Ω. Le spectre de chaîne d'une séquence est donc un vecteur représenté par toutes les occurrences de ses n-mots présents de manière contiguë dans la séquence. Il est à noter que l'espace de représentation est de haute dimension ($|\Omega|^n$ combinaisons possibles de n-mots). La technique du spectre de chaîne est très simple à mettre en oeuvre et peu coûteuse en temps d'exécution. Les auteurs montrent que la performance de leur algorithme est comparable avec celle faisant intervenir des méthodes complexes, comme les HMMs [Karplus *et al.*, 1998]. Nos propres expérimentations sur la famille des cytokines démontrent l'efficacité de cette méthode en terme de classification. Il se trouve que notre famille d'intérêt possède des membres très éloignés entre eux en terme d'homologie de séquence. Nous proposons dans cet article d'utiliser un espace de représentation de faible dimension qui cible des propriétés spécifiques de notre famille d'intérêt. Nous allons dans un premier temps décrire le concept de motif hiérarchique, puis nous donnerons un algorithme d'extraction de ces motifs. Nous rappelerons ensuite les principes des SVMs avant de discuter des résultats obtenus sur la famille des cytokines.

2 Motifs hiérarchiques

La structure primaire d'une protéine est représentée par une séquence $s = \langle s_1 s_2 \dots s_n \rangle$ où chaque s_i appartient à Ω, l'ensemble des acides aminés :

$$\Omega = \{A, C, D, E, F, G, H, I, K, L, M, N, P, Q, R, S, T, V, W, Y\}$$

Soit $P(\Omega)$ l'ensemble des parties de Ω. Certaines de ces parties possèdent des résidus partageant des propriétés physico-chimiques particulières. Plusieurs variantes de systèmes de classes ont été proposées. Nous avons opté pour celui présenté en table 1 [Taylor, 1986]. La pertinence de cette classification se vérifie par l'étude des régions conservées : on observe que les mutations s'opèrent généralement au sein d'une même classe (par exemple, les acides aminés I, L et V appartenant à la classe aliphatique sont très fréquemment interchangés). Les classes physico-chimiques définissent un sous-ensemble de $P(\Omega)$, auquel nous ajoutons l'ensemble des singletons de Ω ainsi que l'ensemble Ω lui-même. Nous noterons $C(\Omega)$ l'alphabet suivant :

$$C(\Omega) = \{\{A\}, \{C\}, \dots, \{Y\}\} \cup \{\alpha, \beta, \gamma, \delta, \varepsilon, \zeta, \eta, \theta\} \cup \Omega$$

On considérera l'ensemble ordonné $(C(\Omega), \subseteq)$ qui forme un sup-demi-treillis : toute paire (x, y) de $C(\Omega) \times C(\Omega)$ possède une borne supérieure, que l'on notera $sup(x, y)$.

Symbole	Classe	Membres
α	aliphatique	ILV
β	aromatique	$FHWY$
γ	non polaire	$ACFGHIKLMVWY$
δ	chargé	$DEHKR$
ε	polaire	$CDEHKNQRSTWY$
ζ	charge positive	HKR
η	chaîne latérale courte	$ACDGNPSTV$
θ	chaîne latérale très courte	$ACGST$

TAB. 1 – Classes d'acides aminés basées sur des propriétés physico-chimiques

Un motif $m = \langle m_1 m_2 \dots m_k \rangle$ est une k-séquence formée d'ensembles $m_i \in C(\Omega)$. Pour la simplicité de la notation, on notera le singleton $\{R\}$ par R directement ; le motif $K\alpha$ désignera ainsi le motif composé de la classe $\{K\}$ suivie de la classe $\{I, L, V\}$.

Bien que rien n'interdise dans notre méthode d'utiliser des motifs de taille k variable, nous supposerons pour la simplicité de l'exposé que k est fixe. Nous appelerons *occurrence* d'un motif m un k-mot $\langle s_{i+1} s_{i+2} \dots s_{i+k} \rangle$ de s tel que $s_{i+j} \in m_j \ \forall j,\ 1 \leq j \leq k$. On dira que la séquence s *vérifie* le motif m. Le *support* d'un motif m dans un jeu de séquences $\mathcal{S}$ est le nombre de séquences de $\mathcal{S}$ qui vérifie m. La séquence MH vérifie ainsi 21 motifs de taille 2, dont les motifs MH, $M\beta$, $\gamma\delta$, et $\Omega\Omega$. Le motif MH ne peut être vérifié que par un seul k-mot, tandis que le motif $\Omega\Omega$ est vérifié pour n'importe quelle séquence de taille supérieure ou égale à 2. Il importe donc de qualifier la spécificité d'un motif en prenant en compte la probabilité de le voir apparaître dans une séquence. Comme l'estimation précise de cette probabilité est complexe et inutile pour notre classifieur, nous avons opté pour la fonction de coût suivante : $c(m) = \prod_{i=1}^{k} f(m_i)$ où $f(m_i)$ est la fréquence de la classe m_i dans une base d'apprentissage comprenant de nombreuses familles de protéines différentes. La spécificité d'un motif m sera définie par $\phi(m) = -\log(c(m))$.

La table 2 montre qu'on observe une bonne corrélation entre l'estimation $\phi(m)$ et le support effectif de m dans la base de contre-exemples de séquences issues de SCOP [Murzin *et al.*, 1995].

Nous avons défini deux propriétés d'un motif hiérarchique : son support et sa spécificité. Il est important de remarquer que ces deux caractéristiques sont généralement opposées. Plus un motif possède un support élevé, moins il est spécifique, comme le montre l'exemple extrême d'un motif composé uniquement de la classe Ω. A l'inverse, un motif uniquement composé de singletons est fortement spécifique mais aura peu de chances d'être découvert.

Les motifs peuvent être hiérarchisés selon une relation de généralisation. Soit deux k-motifs m^1 et m^2. Nous noterons $\preceq$ la relation d'ordre suivante :

$m^1 \preceq m^2$ ssi pour tout $i \in [1, k]$ on a $m_i^1 \subseteq m_i^2$. L'estimateur $\phi(m)$ est construit de sorte que $\phi(m^1) \geq \phi(m^2)$ pour toute paire de motifs vérifiant $m^1 \preceq m^2$. En spécialisant

motif	support	spécificité	fréquence
$\beta\alpha\varepsilon$	1.00	4.4	0.7984
$\alpha\delta\varepsilon\alpha$	0.96	5.1	0.6095
$\delta\Omega\gamma\varepsilon\Omega\alpha\zeta\varepsilon$	0.65	6.8	0.2427
$\Omega\alpha\zeta\varepsilon\alpha\varepsilon\varepsilon\gamma$	0.54	7.6	0.1130
LEE	0.28	7.9	0.0893
$\beta\gamma\Omega\Omega\gamma\zeta\varepsilon L$	0.28	8.4	0.0433
$\varepsilon\gamma\alpha\zeta\delta L\Omega\varepsilon$	0.37	9.2	0.0359
$L\varepsilon\varepsilon\gamma\alpha\varepsilon\delta L$	0.22	10.2	0.0107
$\Omega\eta L\alpha L\alpha\Omega L$	0.15	11.0	0.0019
$F\varepsilon R\gamma K\varepsilon\Omega\gamma$	0.15	11.4	0.0018

TAB. 2 – Exemples de motifs avec leur support dans la famille des cytokines, leur spécificité estimée et leur fréquence effective dans la base SCOP.

un motif, on augmente sa spécificité. Nous appelerons borne supérieure des motifs m^1 et m^2 (notée $sup(m^1, m^2)$) le motif $m^{1,2}$ vérifiant :

$m_i^{1,2} = sup(m_i^1, m_i^2)$ pour tout $i \in [1, k]$.

Le motif $m^{1,2}$ représente le motif le plus spécifique qui généralise m^1 et m^2 : tout k-mot vérifiant m^1 ou m^2 vérifiera $m^{1,2}$. La table 3 donne des exemples de bornes supérieures pour des motifs de taille 4. La section suivante présente comment extraire les motifs de spécificité minimale.

3 Découverte de motifs hiérarchiques

La recherche de motifs hiérarchiques procède en deux étapes :

1. L'extraction de motifs germes,
2. La génération des motifs hiérarchiques.

La première étape consiste à extraire des motifs germes à partir de la famille $\mathcal{S}$. Un motif germe est un motif qui ne possède pas de minorants. Plus pratiquement, les motifs germes sont les motifs formés uniquement de classes singletons (résidus). Une façon triviale d'obtenir la liste des motifs germes est de relever l'ensemble des k-mots présents dans le jeu d'apprentissage. Nous avons procédé à un filtrage en ne considérant que les k-mots potentiellement intéressants. Cette première étape consiste à croiser chaque k-mot avec un autre : chaque couple de k-mot fournit un motif. Si ce dernier possède un support suffisant, les k-mots vérifiant ce motif sont retenus. Avec un support minimal de 3, nous avons pu réduire ainsi le nombre de motifs germes d'un facteur 8.

La seconde étape opère un appariement des motifs pour déterminer leurs bornes supérieures. L'algorithme de découverte de motifs consiste donc à généraliser les motifs afin de rechercher des caractéristiques de support suffisamment représentatif. La définition précédente de borne supérieure permet de définir le motif commun le plus

spécifique à partir de deux motifs. Il est donc possible par itérations successives d'agréger des motifs afin d'obtenir des motifs de support supérieur. La deuxième étape de notre algorithme s'inspire de la technique de la classification hiérarchique ascendante pour former des clusters de motifs généraux à partir de motifs germes composés uniquement de singletons. La deuxième étape de l'algorithme de découverte de motifs est présentée ci-dessous :

algorithme découverteMotifs
entrées
M, l'ensemble de motifs germes obtenus lors de l'étape 1
$supMin$, le seuil de support minimal recherché
$speMin$, le seuil de spécificité minimale recherchée
sortie
E, l'ensemble de motifs hiérarchiques

$E = \{m \in M \mid support(m) \geq supMin \text{ et } \phi(m) \geq speMin\}$;
Répéter
Soit m^1 et m^2 la paire de motifs de M telle que :
1. $m^{1,2} = sup(m^1, m^2)$
2. $\phi(m^{1,2}) \geq \phi(m^{i,j})$ pour tout m^i et m^j dans M
$M \leftarrow M - \{m^1, m^2\}$;
$M \leftarrow M \cup \{m^{1,2}\}$;
si $support(m^{1,2}) \geq supMin$ et $\phi(m^{1,2}) \geq speMin$
alors $E \leftarrow E \cup \{m^{1,2}\}$;
jusqu'à $cardinal(M) = 1$ ou $\phi(m^{1,2}) < speMin$;

La table 3 présente les motifs hiérarchiques de taille $k = 4$ relatifs aux séquences HIWY, HIDY, KLTY, HVSG et DARG. Les motifs m^1 à m^5 sont les motifs germes extraits des k-mots de taille 4 dans le jeu de séquences précédent. La figure 1 montre la construction hiérarchique des motifs par l'algorithme découverteMotifs. A supposer que l'on se soit fixé un support minimal de 3 et une spécificité de 3.0, seul le motif m^6 serait retenu.

L'algorithme précédent permet d'extraire les n motifs les plus intéressants pour caractériser une famille de protéines. Pour utiliser les machines à vecteurs de support, nous allons transformer une séquence quelconque en un vecteur booléen de dimension n. L'élément de rang i est à vrai si et seulement si le motif de rang i est présent dans la séquence (une alternative consiste à compter les occurrences des motifs, mais dans la pratique, il est fort rare de rencontrer deux fois le même motif dans une même séquence pour des valeurs de k importantes). Il est à noter que cette vectorisation s'effectue en $O(N)$, où N désigne la taille de la séquence. Si l'apprentissage est relativement complexe, la classification est peu coûteuse en temps d'exécution. La table 4 montre les valeurs de vecteurs associés aux séquences de l'exemple précédent en considérant l'ensemble des motifs donnés en table 3 (support et spécificité minimaux fixés à 0).

motif	chaîne	support	spécificité
m^1	HIWY	1	14.25
m^2	HIDY	1	12.83
m^3	KLTY	1	11.44
m^4	HVSG	1	11.79
m^5	DARG	1	10.96
$m^6 = m^{1,2}$	HIεY	2	10.64
$m^7 = m^{3,6}$	$\zeta\alpha\varepsilon$Y	3	7.55
$m^8 = m^{4,5}$	$\delta\eta\varepsilon$G	2	5.26
$m^9 = m^{7,8}$	$\delta\Omega\varepsilon\gamma$	5	2.56

TAB. 3 – Motifs extraits par l'algorithme découverteMotifs.

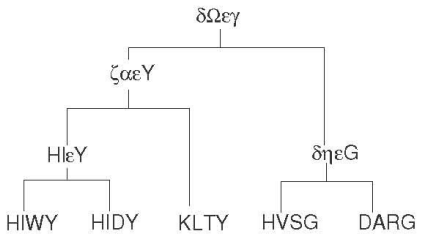


FIG. 1 – Hiérarchie de motifs à partir de 5 motifs germes

4 Machines à vecteurs de support

Cette section introduit la méthode des machines à vecteurs de support [Vapnik, 1995], [Vapnik, 1998]. Cette technique de classification supervisée est basée sur l'apprentissage d'une frontière de décision linéaire qui discrimine au mieux deux classes formées par des exemples et des contre-exemples. Le SVM optimise la frontière de décision par un hyperplan qui maximise la distance entre les points voisins (ou vecteurs de support) durant la phase d'apprentissage. Dans la phase de classification, un point est classé en exemple ou contre-exemple selon sa position par rapport à la frontière de décision. Dans la plupart des problèmes concrets, il y a peu de chances qu'on puisse trouver une séparation linéaire parfaite dans l'espace χ des données. L'efficacité remarquable des SVMs repose sur une transformation non linéaire de l'espace d'entrée χ en un nouvel espace $\Phi(\chi)$ de redescripteurs de plus grande dimension dans lequel les points sont séparables par un hyperplan.

Des travaux théoriques ont permis de montrer que le problème d'optimisation des SVMs revient à résoudre un problème de forme duale qui dépend des produits sca-

séquence	représentation vectorielle
HIWY	100001101
HIDY	010001101
KLTY	001000101
HVSG	000100011
DARG	000010011

TAB. 4 – Représentation vectorielle des séquences à partir des motifs extraits. Le vecteur représente neuf valeurs booléennes (1 pour vrai, 0 pour faux) correspondant à la présence du motif m^i, $i = 1$ à 9, dans la séquence considérée.

laires entre les vecteurs des exemples $\Phi(x)$ de l'ensemble d'apprentissage dans l'espace de redescription. De ce fait, toute la difficulté repose sur le calcul des produits scalaires $< \Phi(x), \Phi(y) >$ entre chaque couple (x, y) de vecteurs d'exemples dans un espace de redescription $\Phi(\chi)$ de dimension très élevée, voire infinie. Cette difficulté est levée grâce à l'introduction de fonctions bilinéaires symétriques positives $K(x, y)$ appelées *fonctions noyaux*. Les fonctions noyaux permettent d'effectuer tous les calculs nécessaires dans l'espace des données sans jamais devoir passer par l'espace des descripteurs : $K(x, y) =< \Phi(x), \Phi(y) >$. Il existe trois types de fonctions noyaux K simples : les fonctions polynomiales, les fonctions à bases radiales (RBF) et les fonctions sigmoïdes. Les fonctions polynomiales sont définies par $K(x, y) = (x.y + 1)^p$. Le degré du polynome est choisi par l'utilisateur. Les fonctions à bases radiales RBF sont définies par $K(x, y) = e^{-\|x-y\|^2/2\sigma^2}$. La valeur de l'écart type σ^2 est estimée empiriquement. Les fonctions sigmoïdes sont de la forme $K(x, y) = tanh(a(x.y - b)$. En pratique, il s'agit de tester les différentes fonctions noyaux pour déterminer celle avec laquelle on obtient l'hyperplan optimal et la marge minimale. Dans le cadre de ce travail, nous avons testé la fonction noyau linéaire (polynome de degré 1) et les fonctions à bases radiales.

La première application des SVMs pour la classification d'homologues éloignées dans les familles protéiques est la méthode des Fisher SVMs [Jaakola *et al.*, 2000]. Cette méthode nécessite tout d'abord l'apprentissage d'un modèle de Markov caché (HMM) sur la famille de protéines d'intéret [Karplus *et al.*, 1998]. Ensuite un jeu d'apprentissage constitué d'exemples et de contre-exemples est représenté dans un espace vectoriel dont chaque dimension est liée à un état ou une transition présent dans le modèle. La valeur de chaque dimension du vecteur est appelée score de Fisher et représente le taux d'utilisation de chaque paramètre du modèle pour modéliser chaque séquence exemple. Les vecteurs obtenus sont utilisés conjointement avec une fonction noyau particulière, appelée fonction de Fisher, pour l'apprentissage du modèle SVM. Les auteurs ont montré que les classifications par la méthode des Fisher SVMs surpassent les méthodes génératives telles que les HMMs ou BLAST. Cependant cette méthode trouve ses limites dans la nécessité d'avoir une famille de protéines possédant un nombre important de membres pour pouvoir élaborer le modèle HMM. Elle se caractérise aussi par une complexité temporelle importante autant en phase d'apprentissage que lors de la prédiction de la classe d'une protéine. La méthode des pairwise

SVMs [Li et Noble, 2003] utilise l'algorithme d'alignement local de Smith et Waterman. Les exemples sont vectorisés dans un espace dont la dimension correspond au cardinal du jeu d'apprentissage. Chaque dimension du vecteur correspond au score de similarité obtenu par l'alignement local avec une séquence du jeu d'apprentissage. Les auteurs de cette méthode ont obtenu des performances de classification supérieure à la méthode des Fisher SVMs. Cependant elle ne résoud pas le problème de la complexité temporelle et rend la dimension des vecteurs dépendante de la taille du jeu d'apprentissage. Une alternative au problème de complexité temporelle est le spectre de chaîne ou *String Spectrum* [Leslie *et al.*, 2002], dont nous avons présenté le principe en introduction. Dans cette méthode, chaque vecteur représente les fréquences des motifs trouvé sur la séquence de l'exemple. Les vecteurs servent ensuite à l'apprentissage du modèle SVM. Ce simple procédé s'affranchit des méthodes génératives et permet un gain de temps d'exécution remarquable. Une adaptation de cette méthode prend en compte tous les motifs de taille k fixée et autorise au plus m variations entre deux motifs [Leslie *et al.*, 2004]. Cette adaptation prend en compte le concept biologique de mutation des résidus. Les auteurs de cette technique montrent que la performance de leur classification est comparable à la méthode des Fisher SVMs tout en optimisant les temps d'exécution. Nos propres expérimentations n'ont pas révélé une amélioration des performances sur notre famille d'intérêt par rapport à la méthode de spectre de chaîne. C'est donc cette dernière méthode qui nous servira de point de comparaison avec notre propre algorithme de classification. Les méthodes décrites jusqu'à présent ne prennent en compte que l'information située au niveau de la structure primaire des protéines. La méthode des SVM-I-sites [Hou *et al.*, 2003] tente d'intégrer une information structurale lors de l'étape de vectorisation des exemples. Une étape préalable consiste à rechercher un profil de séquences à partir de la séquence primaire de l'exemple. Le profil est obtenu au moyen du logiciel PSI-BLAST [Altschul *et al.*, 1997] et de la base de données Swiss-Prot [Boeckmann *et al.*, 2003]. La création des vecteurs se base ensuite sur la recherche des occurences de 263 domaines structuraux au sein du profil. Les auteurs indiquent que cette méthode apporte des résultats équivalents à la méthode des pairwise SVMs. Elle est aussi plus efficace lors de l'étape de vectorisation.

Les SVMs ont démontré leur efficacité dans la détection d'homologies éloignées. Les différents types de classification présentés mettent en évidence l'importance de l'étape de vectorisation des exemples dans la performance de ce type de classifieur.

5 Classification des protéines utilisant les SVMs

Le choix de l'espace de représentation que nous avons adopté nécessite une définition appropriée de la fonction noyau $K(x, x')$. Nous devons en effet tenir compte du nombre plus ou moins grand de motifs découverts dans les séquences. Un produit scalaire $x.x'$ nul signifie qu'aucun motif commun n'existe entre les vecteurs x et x'. Ce cas peut en fait résulter de situations très opposées : si x est un vecteur de support positif avec une norme $\|x\|$ élevée, alors les vecteurs x et x' sont fort dissemblables et une valeur nulle est parfaitement justifiée. Par contre, si x est un vecteur de support négatif, sa norme $\|x\|$ sera faible ou nulle, et tout produit scalaire $x.x'$ possédera une valeur faible ou nulle. Contrairement aux fonctions noyaux RBF, les fonctions noyaux polynomiales ne

répondent pas aux contraintes décrites précédemment. Plusieurs travaux ont révélé que les fonctions noyaux RBF conduisent à de meilleures performances que les les fonctions polynomiales [Leslie *et al.*, 2002]. Elles possèdent des propriétés bien adaptées au cas qui nous intéresse. Soit O le vecteur nul.

1. si $x = O$ et $x' = O$ alors $K(x, x') = 1$. Deux séquences n'ayant toutes les deux aucun motif commun ont une valeur de fonction noyau maximale.

2. si $x = x'$ alors $K(x, x') = 1$. Deux séquences identiques ont une valeur de fonction noyau maximale.

3. si $x' = O$ alors $K(x, x') = e^{-\|x\|^2/2\sigma^2}$. La valeur de fonction noyau est d'autant plus faible que le nombre de motifs dans x est important.

Le paramétrage de la fonction noyau RBF a consisté à appliquer une recherche dichotomique des valeurs optimales pour les paramètres C et $\gamma = \sigma^2$. Finalement, nous avons retenu les valeurs $C = 10$ et $\gamma = 0.5$ pour lesquelles le classifieur obtient des performances comparables à celles obtenues avec une fonction empirique adoptée préalablement. Les fonctions noyaux doivent respecter certaines conditions théoriques (théorème de Mercer) pour pouvoir définir un produit scalaire dans l'espace $\Phi(\chi)$. Ces conditions sont malheureusement difficiles à vérifier dans la pratique et le thèorème de Mercer n'a pas été démontré pour notre fonction empirique. De ce fait nous n'avons conservé que la fonction noyau RBF pour les résultats présentés dans cet article.

La transformation des séquences en vecteurs booléens de taille fixe permet d'utiliser directement les SVMs ; le produit scalaire utilisé dans la fonction noyau RBF entre deux vecteurs booléens x et x' est défini comme le nombre d'éléments à vrai dans le vecteur $x \wedge x'$. Si l'utilisation des SVMs ne pose aucun problème d'adaptation, il est cependant important d'observer qu'à l'origine les SVMs définissent une frontière optimale entre deux classes d'individus d'intérêt égal. Dans notre type d'application, l'importance égale de ces deux classes soulève une difficulté. S'il est en effet facile de définir ce qu'est la famille d'intérêt, la définition des contre-exemples est moins évidente. Une solution consiste à utiliser des représentants dans chaque superfamille définie par la base SCOP. Il faut alors veiller à la robustesse de ces choix et à la sensibilité des SVMs face au déséquilibre entre le nombre des exemples et celui des contre-exemples. Une autre solution consiste à utiliser la version de SVMs basée sur une classe unique proposée par B. Schölkopf [Schölkopf *et al.*, 2001]. Les résultats obtenus avec cette méthode se sont montrés inférieurs à ceux des SVMs à deux classes. Nous interprétons ce phénomène par le nombre restreint d'exemples présenté en apprentissage ainsi qu'à leur grande dispersion (faible homologie intra-classe). Nos tests ont montré qu'en tenant compte d'une juste répartition des contre-exemples dans les familles structurales dans SCOP, on obtient des résultats stables quelque soit le tirage aléatoire effectué. La définition de la fonction noyau RBF assure par ailleurs une grande robustesse face au jeu de contre-exemples.

6 Application à la superfamille des cytokines

Les membres de la superfamille des cytokines sont des glycoprotéines solubles qui interviennent dans la régulation de la réponse immunitaire. Elles ont pour fonction d'assurer la médiation des signaux de prolifération, de différenciation, et d'activation entre les différentes cibles cellulaires. La fonction pivot des cytokines dans l'activation des voies de l'immunité confère aux cytokines un intéret de tout premier plan dans la compréhension des mécanismes de la défense du soi et dans la découverte et l'amélioration les traitements thérapeutiques anticancéreux actuels. Dans cet article, nous nous intéressons plus particulièrement aux interleukines à hélices courtes (IL-2), hélices longues (IL-6) et aux interleukines de type IL-10 dont les précurseurs possèdent six hélices. Nous avons retenu 45 séquences primaires d'interleukines humaines.

La base d'apprentissage qui nous a servi à l'estimation de la spécificité est issue de la base de données SCOP (Structural Classification Of Proteins, [Murzin *et al.*, 1995]). Celle-ci met à la disposition des chercheurs une classification structurale des meilleures structures de protéines publiées et disponibles dans la PDB (Protein Data Bank). Les séquences de SCOP forment donc un échantillon qui recouvre un large spectre de protéines. Après suppression des interleukines, notre base de test comporte 6615 séquences. La famille qui nous intéresse est caractérisée par une structure secondaire riche en hélice alpha. Le pas de cycle d'une hélice étant de 3.6, nous avons considéré des motifs de taille 8 (des motifs de taille 4 seraient moins discriminants).

La table 5 présente les résultats obtenus avec différents classifieurs. Les valeurs présentées sont des moyennes de performances obtenues par la technique de *leave-one-out* (pour laquelle une séquence sert de test et les autres pour l'apprentissage). La ligne KNN présente les résultats obtenus avec la méthode des k plus proches voisins ($k = 3$). La notion de voisinage se réfère à la proximité entre spectres de chaîne : la mesure de similarité utilisée est le produit cartésien des vecteurs normalisés. Les SVMs à base de spectre de chaînes donnent de meilleurs résultats (ligne SCSVM) que les KNNs, ce qui confirme l'intérêt des machines à vecteurs de support. Les résultats sont identiques pour les fonctions noyaux linéaires et à bases radiales ; la seule différence consiste en une légère amélioration des performances sur la base SCOP pour la fonction à base radiale (table 6). Notre méthode MotifsSVM surpasse largement la technique à base de spectre de chaîne (100% de bonne classification) à condition de bien optimiser le type des motifs. Un seuil de spécificité de 14 apparaît comme le meilleur compromis ; au delà de cette valeur, on ne découvre pas assez de motifs sur certaines cytokines, en deçà, les motifs ne sont assez sélectifs. Un seuil de support minimal de 2 en *leave-one-out* (de 3 si on compte l'exemple retiré) est nécessaire pour obtenir de bons résultats dans notre jeu particulier d'application. Il s'avère que les exemples présentés possèdent peu de séquences proches. Nous avons pu vérifier ce fait en calculant les valeurs de fonction noyau à partir de spectres de chaîne : la valeur de fonction noyau tend rapidement vers 0 dès la deuxième plus proche voisine.

La table 6 indique le pouvoir discriminant de notre classifieur sur les séquences négatives extraites de la base SCOP. Sur les 6615 séquences de la base SCOP, le classifieur MotifsSVM a en mal classé 8. Le faible pourcentage de faux-positifs autorise l'emploi de MotifsSVM pour rechercher de nouveaux membres dans le génome.

classifieur	taux d'erreurs	VP	FN	VN	FP
KNN	18.9	88.9	11.1	73.3	26.7
SCSVM linéaire	13.3	84.4	15.6	88.9	11.1
SCSVM RBF	13.3	84.4	15.6	88.9	11.1
MotifsSVM 13	2.2	95.6	4.4	100	0
MotifsSVM 14	0	100	0	100	0
MotifsSVM 15	5.5	88.9	11.1	100	0

TAB. 5 – Résultats de classification. VP, FN, VN, FP sont les pourcentages respectivement des vrais positifs, faux négatifs, vrais négatifs, faux positifs.

classifieur	taux de FP dans SCOP
SCSVM linéaire	4.32
SCSVM RBF	4.08
MotifsSVM 14	0.12

TAB. 6 – Taux de faux-positifs dans la base SCOP.

7 Conclusion et perspectives

Les excellents résultats obtenus en classification dans la famille des cytokines démontrent la pertinence d'une description hiérarchique des motifs. Ces derniers assurent un rôle de signature, au sens où ils sont spécifiques à la famille étudiée. Nous avons proposé un paramétrage simple du degré de spécificité souhaité et avons observé qu'une valeur moyennement haute donne les meilleurs performances ($\phi(m) \sim 14$). La capacité des SVMs à gérer des espaces de grande dimension nous permet d'obtenir une classification sans erreurs sur les exemples positifs et un très faible taux d'erreurs sur les exemples négatifs issus de SCOP (0.12%). Certaines améliorations de l'algorithme d'extraction restent à mettre en oeuvre, afin d'optimiser le nombre de motifs retenus. Sur les quelques 600 motifs extraits, une proportion non négligeable d'entre eux peuvent certainement être éliminés sans nuire à la performance du classifieur. Il est aussi envisageable d'effectuer une pondération des éléments des vecteurs en fonction de la spécificité des motifs. Nous envisageons dans un premier temps de filtrer plus finement les motifs extraits et de procéder ensuite à une étude de leurs co-occurrences dans les séquences. Cette analyse permettra de déterminer des patrons de motifs propres à une famille de protéines (un patron est défini comme une séquence de motifs situés à des intervalles variables sur une même séquence).

Références

[Altschul *et al.*, 1990] S. F. Altschul, W. Gish, W. Miller, E. W. Myers, et D. J. Lipman. A basic local alignment search tool. *Journal of Molecular Biology*, 215 :403–410, 1990.

[Altschul *et al.*, 1997] S. F. Altschul, T. L. Madden, A. A. Schäffer, J. Zhang, Z. Zhang, W. Miller, et D. J. Lipman. Gapped blast and psi-blast : a new generation of protein database search programs. *Nucleic Acid Research*, 25(3389-3402) :17, 1997.

[Boeckmann *et al.*, 2003] B. Boeckmann, A. Bairoch, R. Apweiler, M.-C. Blatter, A. Estreicher, E. Gasteiger, M. J. Martin, K. Michoud, C. O'Donovan, I. Phan, S. Pilbout, et M. Schneider. The swiss-prot protein knowledge base and its supplement trembl in 2003. *Nucleic Acid Research*, 31 :365–370, 2003.

[Hou *et al.*, 2003] Y. Hou, W. Hsu, M. L. lee, et C. Bistroff. Efficient remote homology detection using local structure. *Bioinformatics*, 19(17) :2294–2301, 2003.

[Jaakola *et al.*, 2000] T. Jaakola, M. Diekhans, et D. Haussler. A discriminative framework for detecting remote protein homologies. *Journal of Computationnal Biology*, 7(1-2) :95–114, 2000.

[Jalam et Teytaud, 2001] Radwan Jalam et Olivier Teytaud. Identification de la langue et catégorisation de textes basées sur les n-grammes. *Extraction de Connaissance et Apprentissage*, 1(1-2) :227–238, Janvier 2001.

[Karplus *et al.*, 1998] K. Karplus, C. Barret, et R. Hugley. Hidden markov models for detecting remote protein homologies. *Bioinformatics*, 14 :846–856, 1998.

[Leslie *et al.*, 2002] C. Leslie, E. Eskin, et W. S. Noble. The spectrum kernel : a string kernel for svm protein classification. In *Proceedings of the Pacific Biocomputing Symposium*, pages 564–575, 2002.

[Leslie *et al.*, 2004] C. Leslie, E. Eskin, A. Cohen, J. Weston, et W. S. Noble. Mismatch string kernels for svm protein classification. *Bioinformatics*, 20 :467–476, 2004.

[Li et Noble, 2003] L. Li et W. S. Noble. Combining pairwise sequence similarity and support vector machines for detecting remote protein evolutionnary and structural relationships. *Journal of Computationnal Biology*, 10(6) :857–868, 2003.

[Murzin *et al.*, 1995] A.G. Murzin, S.E.Brenner, T. Hubbard, et C. Chothia. Scop : a structural classification of proteins database for the investigation of sequences and structures. *Journal of Molecular Biology*, 247 :536–540, 1995.

[Schölkopf *et al.*, 2001] B. Schölkopf, J. Platt, J. Shawe-Taylor, A. J. Smola, et R. C. Williamson. Estimating the support of a high-dimensional distribution. *Neural Computation*, 13(7), 2001.

[Taylor, 1986] J. Taylor. Classification of amino acid conservation. *Theorical Biology*, 119 :205–218, 1986.

[Vapnik, 1995] V. N. Vapnik. *The nature of statistical learning theory.* Springer-Verlag, 1995.

[Vapnik, 1998] V. N. Vapnik. *Statistical Learning Theory.* Springer, 1998.

Annexe 4.

Jérôme Mikolajczak, Gérard Ramstein, Yannick Jacques,

SVM-based classification of distant proteins using hierarchical motifs,

In *5th International Conference of Intelligent Data Engineering and Automated Learning (IDEAL 2004)*, Lecture Notes in Computer Science, Zhen Ron Yang, Richard Everson, Hujun Yin, Editors, Springer-Verlag GmbH, Publisher, EXETER (UNITED KINGDOM), October 2004, vol 3177 : 25 – 30, ISBN: 3-540-22881-0.

SVM-based classification of distant proteins using hierarchical motifs

Jérôme Mikolajczak*, Gérard Ramstein**
Yannick Jacques*

* Département de Cancérologie, Institut de Biologie
9 Quai Moncousu, F-44035 Nantes cedex
jmikolaj@nantes.inserm.fr, yjacques@nantes.inserm.fr
**LINA, équipe LEC, Ecole polytechnique de l'Université de Nantes
Rue Christian Pauc, BP 50609 44306 Nantes cedex 3
gerard.ramstein@polytech.univ-nantes.fr

Abstract

This article presents a discriminative approach to the protein classification in the particular case of remote homology. The protein family is modelled by a set M of motifs related to the physicochemical properties of the residues. We propose an algorithm for discovering motifs based on the ascending hierarchical classification paradigm. The set M defines a feature space of the sequences : each sequence is transformed into a vector that indicates the possible presence of the motifs that belongs to M. We then use the SVM learning method to discriminate the target family. Our hierarchical motif set specifically modelises interleukins among all the structural families of the SCOP database. Our method yields a significantly better remote protein classification compared to spectrum kernel techniques.

1 Introduction

Considering that our purpose is to discover new members of a protein family, we need an effi cient and fast classifi er compatible with genome mining. Support vector machines (SVM) are one of the most powerful supervised learning algorithm. An SVM [1], [2] is a binary classifi er that builds a decision boundary by mapping training data from the original input space into a high dimensional feature space. The SVM selects the hyperplan which maximises the separation between two data classes. Due to the high dimensionality of the feature space, the SVM algorithm uses a function called a kernel to compute dot products directly into the input space. SVMs have been successfully applied to different problems in bioinformatics, such as protein fold class prediction and microarray data analysis [3]. SVM-based methods for protein classifi cation transform input protein sequences into fi xed-length feature vectors. This vectorisation step is crucial for the future performances of the classifi er. The SVM-Fisher method [4] uses a hidden Markov model (HMM) trained on a protein family to compute the gradient vector of input sequences. The drawback of this technique is the computation cost due to the scoring of the sequences. The SVM-pairwise method [5] applies a pairwise sequence similarity algorithm using the Smith-Waterman algorithm. The computational expense of this method remains too high for our purpose. A very effi cient algorithm is the spectrum kernel method [6]. The features are the set of all possible susbsequences of

amino acids of fi xed length k. The same author also proposes the mismatch kernel technique [7], an extension of the previous method that authorizes a certain proportion of wildcards into the subsequences of length k. As these two last techniques have a linear time complexity and perform remarkably well, they will serve as reference methods. This paper is organised as follows. Section 2 introduces the concept of hierarchical motif. Section 3 describes the algorithm for discovering motifs. Section 4 presents the SVM classifi er and section 5 presents the results of our experiments on the classifi cation of interleukins.

2 Hierarchical motifs

The primary structure of a protein is represented by a sequence $s = \langle s_1 s_2 \dots s_n \rangle$ where each s_i belongs to Ω, the amino acid set :

$$\Omega = \{A, C, D, E, F, G, H, I, K, L, M, N, P, Q, R, S, T, V, W, Y\}$$

Let $P(\Omega)$ be the set of the subsets of Ω. Some elements of $P(\Omega)$ comprises amino acids sharing particular physicochemical properties of the amino acids. Several classes have been proposed : we have chosen the Taylor code (table 1). The relevance of this classifi cation is verifi ed by the study of conserved regions. For instance, sequence similarities show that mutations often take place inside the same physicochemical class (e.g. the residues I, L, V belonging to the aliphatic class are frequently exchanged). Let $C(\Omega)$ be the union of the physicochemical classes, the singleton set of Ω and Ω :

$$C(\Omega) = \{\{A\}, \{C\}, \dots, \{Y\}\} \cup \{\alpha, \beta, \gamma, \delta, \varepsilon, \zeta, \eta, \theta\} \cup \Omega$$

We will consider the ordered set $(C(\Omega), \subseteq)$ that forms a semi-lattice : every pair (x, y) of $C(\Omega) \times C(\Omega)$ has a least upper bound denoted $sup(x, y)$.

Symbol	Class	Members
α	aliphatic	ILV
β	aromatic	$FHWY$
γ	non-polar	$ACFGHIKLMVWY$
δ	charged	$DEHKR$
ε	polar	$CDEHKNQRSTWY$
ζ	positive	HKR
η	small	$ACDGNPSTV$
θ	tiny	$ACGST$

Table 1: Amino acid classes based on physicochemical properties

A motif $m = \langle m_1 m_2 \dots m_k \rangle$ is a k-sequence of sets $m_i \in C(\Omega)$. For the sake of concision, we denote by R the singleton $\{R\}$ (e.g. the motif $K\alpha$ comprises the class $\{K\}$ followed by the class $\{I, L, V\}$). An *occurrence* of a motif m is a subsequence

$\langle s_{i+1}s_{i+2}\dots s_{i+k}\rangle$ of s such as $s_{i+j} \in m_j\ \forall j,\ 1 \leq j \leq k$. A sequence that contains at least one occurrence of m is said to *verify* m. The *support* of a motif m in a sequence set $\mathcal{S}$ is the number of sequences of $\mathcal{S}$ that verify m. The sequence MH verifi es 21 motifs of size $k = 2$, including MH, $M\beta$, $\gamma\delta$, and $\Omega\Omega$.

Only one subsequence verifi es the motif MH, whereas any sequence of length greater than or equal to 2 verifi es the motif $\Omega\Omega$. We therefore have to consider the specifi city of a motif which is the probability of its occurrence in a sequence. As the estimation of the probability is time consuming, we instead propose the cost function $c(m) = \prod_{i=1}^{k} f(m_i)$ where $f(m_i)$ is the frequency of the m_i class in SCOP database [8]. This database includes a large panel of different protein families. In practice, we observe a good correlation between $c(m)$ and the effective support of m in the training database. The specifi city of a motif m will be defi ned by $\phi(m) = -\log(f(m))$. Our experiments reveal that this specifi ty estimator of a given motif is well correlated with the effective frequence of this motif in a wide panel of proteins belonging to different families.

It is important to note that support and specifi city are opposed concepts : the more a motif presents a high support, the less it is specifi c, and conversely. Motifs can be hierarchised according a relation of generalisation. Let m^1 and m^2 be two k-motifs. We note $\preceq$ the following relation : $m^1 \preceq m^2$ if for all $i \in [1, k]$ $m_i^1 \subseteq m_i^2$.

The defi nition of the estimator $\phi(m)$ implies that $\phi(m^1) \geq \phi(m^2)$ for all pairs of motifs that verify $m^1 \preceq m^2$.

We call least upper bound of motifs m^1 and m^2 (denoted $sup(m^1, m^2)$) the motif $m^{1,2}$ verifying : $m_i^{1,2} = sup(m_i^1, m_i^2)$ for all $i \in [1, k]$. The motif $m^{1,2}$ represents the most specifi c motif that generalises m^1 and m^2 : every subsequence verifying m^1 or m^2 will verify $m^{1,2}$. Table 2 gives some examples of least upper bounds.

3 Hierarchical motif discovery

The discovery algorithm proceeds in two steps :

1. the extraction of seed motifs,
2. the generation of hierarchical motifs.

The fi rst step consists in the extraction of seed motifs from the sequence set $\mathcal{S}$. A seed motif is a motif having no minoring elements. More pratically, a seed motif only contains singleton classes. To defi ne the seed motif set, we use a fi lter that only considers the putative interesting k-subsequences of $\mathcal{S}$. This step consists in the computation of the least upper bound of every pair of k-subsequences. If the least upper bound presents a minimal support, the k-subsequences that verify this motif are kept. With a minimal support of 3, we reduce the size of k-subsequences by a factor of 8.

The second step is inspired by the ascending hierarchical classifi cation technique. A dendrogram of motifs is generated from the seed motifs of the previous step. This

phase proceeds as follows :

algorithm motifDiscovery

inputs

M, the seed motif set obtained at step 1

$supMin$, the minimal support threshold

$speMin$, the minimal specifi city threshold

output

E, the set of hierarchical motifs

$E = \{m \in M \mid support(m) \geq supMin$ et $\phi(m) \geq speMin\}$;

Repeat

Let m^1 and m^2 be the pair of motifs of M such as :

1. $m^{1,2} = sup(m^1, m^2)$
2. $\phi(m^{1,2}) \geq \phi(m^{i,j})$ for all m^i and m^j in M

$M \leftarrow M - \{m^1, m^2\}$;

$M \leftarrow M \cup \{m^{1,2}\}$;

if $support(m^{1,2}) \geq supMin$ and $\phi(m^{1,2}) \geq speMin$

then $E \leftarrow E \cup \{m^{1,2}\}$;

until $cardinal(M) = 1$ or $\phi(m^{1,2}) < speMin$;

Table 2 presents hierarchical motifs of size $k = 4$ obtained from the sequences : HIWY, HIDY, KLTY, HVSG and DARG. Motifs m^1 to m^5 are the seed motifs extracted from the previous sequences. Fig. 1 shows the dendrogram tree built by our algorithm.

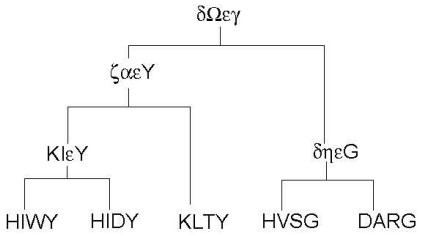


Figure 1: dendrogram of motifs

motif	content	support	specifi city
m^1	HIWY	1	14.25
m^2	HIDY	1	12.83
m^3	KLTY	1	11.44
m^4	HVSG	1	11.79
m^5	DARG	1	10.96
$m^6 = m^{1,2}$	KIεY	2	10.64
$m^7 = m^{3,6}$	$\zeta\alpha\varepsilon$Y	3	7.55
$m^8 = m^{4,5}$	$\delta\eta\varepsilon$G	2	5.26
$m^9 = m^{7,8}$	$\delta\Omega\varepsilon\gamma$	5	2.56

Table 2: Motifs extracted by motifDiscovery algorithm.

4 Application and discussion

The members of the cytokine superfamily are soluble glycoproteins acting in the regulation of immunity responses. The set of positive examples that we use consists in 45 primary sequences of interleukins belonging to the human family of cytokines. We set the size k of our motifs to 8 because of the typical helix secondary structure of this family. The negative examples are extracted from the SCOP database (6615 sequences). As inputs of the SVM are fi xed-length vectors, we associate to a sequence s a boolean vector $v(s)$: the ith element of $v(s)$ is set to true iff the sequence s verifi es the ith element of the set of hierarchical motifs. This vectorisation step has a linear time complexity. Table 3 compares the results obtained with different classifi ers according to the leave-one-out validation technique. The validation set comprises all the 45 positive examples as well as 45 negative examples selected from all the structural superfamilies of SCOP. All SVM classifi ers use the radial basis function as kernel. Line KNN corresponds to K-nearest neighbour method based on the similarity measure corresponding to the dot product of the normalised vectors. The better performance of the spectrum kernel method confi rms the superiority of the SVM algorithm (line SKSVM). We also give the results obtained with the mismatch kernel technique (line MKSVM). Our method (line SVMotifs) surpasses the spectrum and mismatch kernel techniques (100% of good classifi cation), provided that the specifi city threshold is optimised. A value of 14 seems to be the best compromise : beyond this threshold the number of discovered motifs is too low for some interleukins, and below motifs are not specifi c to the interleukin family.

Applied to the SCOP data set, the SKSVM method presents an error rate of 4% while our method only gives 0.12% (8 false positives among 6615 sequences).

The experiments have shown the relevance of hierarchical motifs for the discrimination of protein families. The last ones form a signature in the sense that they are specifi c to the family studied. We have proposed an algorithm that extracts motifs having a desired degree of specifi city. The capacity of SVM to handle high dimensional spaces yields excellent results on the interleukin family : error-free recognition of positives

classifi er	error rate	TP	FN	TN	FP
KNN	18.9	88.9	11.1	73.3	26.7
SKSVM	13.3	84.4	15.6	88.9	11.1
MKSVM	12.0	82.7	17.3	93.3	16.7
SVMotifs 13	2.2	95.6	4.4	100	0
SVMotifs 14	0	100	0	100	0
SVMotifs 15	5.5	88.9	11.1	100	0

Table 3: Classifi cation results. VP, FN, VN, FP are the percentage of true positives, false negatives, true negatives and false positives.

and a very low rate of false negatives on the SCOP database (0.12%). In the future, we intend to work on improving the motif discovery module in order to reduce the cardinal of the motif set.

References

[1] V. N. Vapnik. *The nature of statistical learning theory*. Springer-Verlag, 1995.

[2] V. N. Vapnik. *Statistical Learning Theory*. Springer, 1998.

[3] B. Scholkpf, I. Guyon, and J. Weston. statistical learning and kernel methods in bioinformatics. In P. Frasconi and R. Shamir, editors, *Artificial Intelligence and Heuristic Methods in Bioinformatics*, pages 1–21. IOS Press, 2003.

[4] T. Jaakola, M. Diekhans, and D. Haussler. A discriminative framework for detecting remote protein homologies. *Journal of Computationnal Biology*, 7(1-2):95–114, 2000.

[5] L. Li and W. S. Noble. Combining pairwise sequence similarity and support vector machines for detecting remote protein evolutionnary and structural relationships. *Journal of Computationnal Biology*, 10(6):857–868, 2003.

[6] C. Leslie, E. Eskin, and W. S. Noble. The spectrum kernel : a string kernel for svm protein classifi cation. In *Proceedings of the Pacific Biocomputing Symposium*, pages 564–575, 2002.

[7] C. Leslie, E. Eskin, D. Zhou, and W. S. Noble. Mismatch string kernel for svm protein classifi cation. *Bioinformatics*, 2004. paratre.

[8] A.G. Murzin, S.E.Brenner, T. Hubbard, and C. Chothia. Scop: a structural classifi cation of proteins database for the investigation of sequences and structures. *Journal of Molecular Biology*, 247:536–540, 1995.

EXTRACTION DE SIGNATURES COMPLEXES POUR LA DECOUVERTE DE NOUVEAUX MEMBRES DANS DES FAMILLES DE PROTEINES CONNUES

Résumé

Cette thèse a permis d'obtenir des modèles de classification pour les familles structurales des interleukines à hélices α humaines au moyen d'un ensemble de signatures caractéristiques. Nous avons établi une approche génétique en trois étapes. Les signatures sont définies par des séquences de motifs hiérarchiques préalablement extraits et basés sur une classification hiérarchique des acides aminés en fonction de leurs propriétés physico-chimiques. Après optimisation, l'ensemble optimal des signatures cible spécifiquement notre ensemble d'interleukines. Une seconde approche repose sur l'utilisation originale d'un algorithme de découverte de motifs suivant le paradigme de la classification hiérarchique. L'ensemble des motifs définit un espace de représentation vectoriel basé sur la présence ou l'absence de chaque motif dans les séquences d'interleukines. Nous utilisons la technique des Systèmes à Vastes Marges pour discriminer nos familles. Notre modèle de classification des interleukines est plus performant que d'autres méthodes et ouvre la voie à des travaux d'extraction de nouvelles interleukines dans les bases de données génomiques.

Mots clés

cytokine, interleukine, famille structurale, homologies éloignées, extraction de motifs, apprentissage automatique, classification des données, algorithmes génétiques, systèmes à vastes marges, SVM.

CHARACTERIZATION OF COMPLEX SIGNATURES FOR THE DISCOVERY OF NEW MEMBERS IN WELL-KNOWN PROTEIN FAMILIES

Abstract

This thesis allowed us to obtain classification models from the structural families of all α helices human interleukins by the way of a set of representative signatures. We established a genetic approach following a three steps process. A discovery algorithm of sequential itemsets searchs for sequence of hierarchical patterns previously extracted and based on an alphabet including the amino acid set and their own physicochemical properties. After a reduction step, the optimal set of signatures specifically targets our set of interleukins. The second part of our work consisted in an original discriminative approach which proposes an algorithm for discovering motifs based on the ascending hierarchical paradigm. The set of motifs defines a vectorial feature space that indicates the presence of the motifs in the interleukin sequences. We use the Support Vector Machines to discriminate our set. Our classification model performs better on our interleukins than other remote protein classification methods and opens the way toward the extraction of new interleukins from the genomic public databases.

Key Words

Cytokine, Interleukin, Structural family, remote homologies, Motif Extraction, Machine Learning, Data Classification, Genetic Algorithm, Support Vector Machines, SVM.

Discipline – spécialité

Sciences de la Vie et de la Santé – Bio-informatique

MIKOLAJCZAK Jérôme
"Interactions Récepteurs/Ligands en Immunocancérologie et Immunopathologie "
Département de Cancérologie, INSERM U601, Institut de Biologie, 9 quai Moncousu, 44093 Nantes cedex 1