Recherche de domaines protéiques divergents à l'aide de modèles de Markov cachés : application à Plasmodium falciparum

Nicolas Terrapon

Académie de Montpellier

Université Montpellier II

— Sciences et Technique du Languedoc —

THÈSE

présentée à l'Université des Sciences et Techniques du Languedoc
pour obtenir le diplôme de doctorat

Spécialité	: **INFORMATIQUE**
Formation Doctorale	: **Informatique**
École Doctorale	: **Information, Structure, Système**

Recherche de domaines protéiques divergents à l'aide de modèles de Markov cachés : application à *Plasmodium falciparum*

par

Nicolas TERRAPON

Jury composé de :

M. Daniel Kahn, Directeur de Recherche INRA, LBBE Lyon ..... Rapporteur
M. Jacques Nicolas, Directeur de Recherche INRIA, IRISA Rennes ..... Rapporteur
M. Nicolas Hulo, Chercheur, SIB Genève ..... Examinateur
M. Éric Maréchal, Directeur de Recherche CNRS, LPCV Grenoble ..... Examinateur
M. Olivier Gascuel, Directeur de Recherche CNRS, LIRMM Montpellier ..... Directeur de Thèse
M. Laurent Bréhélin, Chargé de Recherche CNRS, LIRMM Montpellier ..... Co-directeur de Thèse

À Gabriel et sa maman

Table des matières

Première partie

Remerciements

Il faut rendre à César ce qui est à César... et cette thèse n'aurait jamais pu aboutir sans l'aide, le soutien et la bienveillance de nombreuses personnes. Ces personnes agissant dans l'ombre, je tiens en premier lieu à remercier toutes celles que je vais malheureusement oublier de citer dans les lignes qui viennent...

Je tiens tout d'abord à remercier mes directeurs de thèse qui sont les principaux acteurs, metteurs en scène et producteurs de ce travail : Laurent Bréhélin et Olivier Gascuel. Ces dernières années m'ont permis d'apprendre beaucoup en les côtoyant et m'ont fait entrevoir le chemin qu'il me reste à parcourir. Je remercie tout particulièrement Laurent qui m'a formé à la recherche depuis mon stage de DEA, m'a fait voyager en conférence (sérieux des répétitions et décompressions sans complexes) et m'a surtout lu et relu tellement de fois pour contenir mes envolées littéraires.

Mes pensées vont ensuite vers ma famille : ma mère, mes grand-parents, ma compagne, sa famille qui m'a adopté et celle que nous avons construite. Leur soutien sans faille ainsi leur incompréhension totale de ce que j'étais en train de faire m'ont continuellement poussé pour à me surpasser, à clarifier mon propos et à rester humble.

Je remercie ensuite tous les thésards qui ont partagé mon bureau au LIRMM : Denis, Sam, Sylvain, Céline, re-Sam, JB et l'autre Nico. Si certains ont su se faire remarquer par leur discrétion, je dois à Sam d'être aujourd'hui capable de travailler quelque soit la musique ambiante ou l'activité de mes camarades de bureau. Nico et JB lui sont redevable car il est difficile de garder sa concentration face aux propos grivois de "Miss Google traduction". Je n'oublie évidemment pas les autres thésards et les membres de l'équipe MAB qui ont gravité autour de notre bureau (dont j'avoue ne pas être sorti assez souvent) et en particulier à ceux avec qui j'ai le plus discuté, travaillé ou voyagé : Amel, Cécile, Raluca, Éric, Jef, Vincent, et Alban (ou pas), *etc.*

Je remercie très chaleureusement tous les membres du laboratoire iRTSV/LPCV du CEA de Grenoble qui m'ont accueilli pendant ma première année d'ATER. La convivialité des gens des pays froid n'est pas un mythe... mais la recette de la chartreuse restera secrète. Merci donc aux thésards, ingénieurs et chercheurs pour les connaissances qu'ils m'ont apporté en biologie ainsi que pour la valorisation de ma vocation : un bio-informaticien peut aider à faire le lien entre les deux communautés. Merci également pour l'ambiance, les apéros/séminaires et les repas à Emma, Evy, Olivier, Sophie, Djeneb, Maryse et bien sûr Éric. Je salue également l'équipe TIMB du TIMC chez qui j'ai fais un passage éclair durant ma période Grenobloise (Michael, Olivier, Jean-Louis et les autres).

J'ai également eu la chance de m'intégrer dans différentes équipes pédagogiques. Je remercie donc mes collaborateurs d'enseignement à l'IUT de Montpellier (Olivier Cogis, les enseignants dont j'ai assuré les TD ou avec qui je les ai partagé, et tous les personnes présentes aux conseils de classes où j'ai beaucoup appris... et ri), à l'ENSIMAG de Grenoble (Christine, Christophe, Claudia et les thésards Guillaume et Kiev), et à l'Université Montpellier 3 (Alexandre, Corinne, Gwaenaël, Sandra et Sylvain). Ces lieux sont très différents et les publics aussi, mais la joie d'enseigner reste la même.

Si je suis ce que je suis, c'est également grâce à de nombreuses rencontres et amitiés. Certaines datent d'il y a très longtemps. D'autres souffrent des distances. Je remercie Auré,

Gogo, Jéjé, Mama et Didounet pour l'enfance et JC, Ju, Philou et Paul pour la décadence.

Et pour conclure, je tenais à remercier encore une fois Maurice, même s'il a parfois poussé le bouchon un peu loin (comme dans la pub), sa disponibilité et ses efforts émanent de cette thèse.

Deuxième partie

Introduction

Le séquençage d'un génome complet consiste à déterminer l'ordre d'enchaînement des nucléotides constituant l'ADN d'un organisme. Ces dernières années, nous avons assisté à la multiplication des projets de séquençage de génomes complets grâce aux développements technologiques et à la réduction des coûts. Cette dynamique se poursuit et permet un accroissement massif des données disponibles. Cependant, la vitesse d'apparition des données excède de loin les capacités d'analyse de celles-ci, conduisant à un goulot d'étranglement (Benson *et al.*, 2009). L'annotation des données issues de projets de séquençage est l'une des missions prioritaires pour la communauté bioinformatique.

On distingue généralement trois niveaux d'annotations bioinformatiques, permettant respectivement de répondre aux questions : où, quoi, et comment ? (Stein, 2001). Le niveau d'annotation nucléique (où ?) consiste, à partir des séquences d'ADN, à identifier la position des gènes codant. On recherche aussi différents éléments tels que les *single nucleotide polymorphism* (SNP), les ARN non-codants et les régions de régulation par exemple. Le deuxième niveau d'annotation (quoi ?) concerne les protéines, macro-molécules nécessaires à la vie et à la perpétuation des espèces. Synthétisée à partir des gènes, le premier angle d'étude des protéines s'appuie généralement sur la séquence primaire (ordre séquentiel des acides aminés). En comparant les différents domaines du Vivant [1] on a pu observer des caractéristiques communes à ces séquences. On les annote notamment par l'identification de structures secondaires et tertiaires stables. Ces annotations sont fortement liées au dernier niveau d'annotation : l'annotation fonctionnelle (comment ?) qui décrit les processus biologiques dans lesquels sont impliqués les gènes et leurs produits.

L'annotation fonctionnelle est fréquemment synthétisée à l'aide de vocabulaires structurés ou ontologies (par exemple la *Gene Ontology* (Ashburner *et al.*, 2000)). Comprendre et caractériser les différentes fonctions biologiques des protéines est essentiel pour l'étude des processus biologiques, dans une visée fondamentale (avancée des connaissances) ainsi que dans une visée appliquée, qu'elle soit médicale (découverte de médicaments dans le cadre de pathologies humaines) ou économique (développements pour les secteurs agro-alimentaires). Grâce au séquençage systématique de génome, des annotations fonctionnelles ont pu être proposées pour de nombreuses protéines. L'annotation fonctionnelle des protéines par des méthodes bioinformatiques se fait sur le principe du transfert d'annotations, c'est à dire que l'on transfert aux protéines récemment séquencées les connaissances acquises sur des protéines similaires lors de précédentes études. Dans ce cadre, l'une des notions clef pour l'annotation protéique et fonctionnelle est l'identification des domaines protéiques. Les domaines sont les sous-unités structurales des protéines. L'évolution a conduit de nombreux domaines à se combiner pour former des protéines multidomaines ou des complexes protéiques, avec *a priori* un vaste espace de possibilités. Chaque domaine peut être entièrement dupliqué, inversé, transposé, impliqué dans des transferts horizontaux, ou encore faire l'objet de mutations ponctuelles, de délétions et d'insertions. Par conséquent, le domaine protéique est considéré de nos jours comme une unité d'évolution à part entière.

De nombreuses méthodes ont été développées dans le cadre de l'annotation automatique

1. bactéries, archées et eucaryotes

des génomes. Ces méthodes s'appuient principalement sur la *comparaison de séquences* pour identifier les protéines et domaines homologues, c.-à-d. partageant une histoire évolutive commune. La bioinformatique a su apporter des algorithmes rapides et efficaces pour la recherche d'homologie, parmi lesquels on distingue deux catégories d'approches classiques :

- la comparaison de séquences : approches de type BLAST (Altschul *et al.*, 1990), ClustalW (Thompson *et al.*, 1994), *etc.* qui recherchent, dans une base de données de séquences les occurrences d'une séquence requête ;
- la modélisation de familles de protéines : à partir d'un ensemble de séquences homologues regroupées en famille, on utilise une représentation mathématique de l'alignement multiple des séquences (expression régulière, PSSM, HMM, *etc.*) afin de rechercher de nouvelles séquences ressemblant au modèle.

La modélisation est généralement considérée comme une méthode plus performante que la comparaison de séquences deux à deux, pour la détection de séquences homologues (Park *et al.*, 1998). L'avantage de la modélisation réside dans sa capacité à représenter la diversité des séquences et l'information position-spécifique (issue de l'alignement multiple) qui caractérisent les familles d'homologues. Parmi les techniques de modélisation, les modèles de Markov cachés (*Hidden Markov Models* notés HMM) se sont révélés être un outil puissant pour la modélisation et la détection de domaines protéiques et sont utilisées dans de nombreuses bases de données de domaines. La base de données Pfam, par exemple, propose une grande collection de HMM qui couvre environ 74% des protéines répertoriées par SwissProt/TrEMBL. Chaque HMM est un modèle probabiliste représentant un domaine protéique unique. Étant donnée une nouvelle séquence protéique, chaque HMM est utilisé pour calculer un score reflétant la ressemblance de la séquence au modèle. Ce score est alors comparé à un seuil de référence fournit par Pfam au dessus duquel on peut affirmer la présence du domaine dans la protéine. Cependant, cette procédure peut manquer de sensibilité lors de l'étude de protéines fortement *divergentes*.

L'étude des protéines divergentes est l'objet central de cette thèse. La notion de divergence se définit implicitement par rapport à un *standard*, lié ici aux organismes modèles. Les organismes modèles sont les premiers à avoir été séquencés et étudiés pour des raisons historiques, pratiques (facilité de reproduction), économiques ou sociétales. On peut notamment citer les génomes microbiaux d'*Haemophilus influenzae* (grippe) (Fleischmann *et al.*, 1995) et *Mycobacterium tuberculosis* (tuberculose) (Cole *et al.*, 1998), puis chez les eucaryotes *Saccharomyces cerevesiae* (levure) (Goffeau *et al.*, 1996), *Caenorhabditis elegans* (ver) (The *C. elegans* Sequencing Consortium, 1998), *Drosophila melanogaster* (mouche) (Adams *et al.*, 2000) et *Arabidopsis thaliana* (petite plante proche de la moutarde) (The Arabidopsis Genome Initiative, 2000). Les protéines les mieux annotées dans les organismes modèles servent de référence pour l'étude des organismes nouvellement séquencés. Ces protéines définissent les *standards* des bases de données de domaines où elles sont utilisées pour l'apprentissage des modèles de type HMM. C'est pourquoi ces modèles permettent l'identification de domaines dans de nombreuses protéines (proches des standards), mais rencontrent des difficultés face aux protéines divergentes. Une séquence est donc qualifiée de *divergente* si sa séquence d'acides aminés exhibe une évolution spécifique et divergente

par rapport aux séquences standards. Des séquences divergentes sont observées dans la plupart des organismes, y compris les organismes modèles. Néanmoins, la problématique d'annotation se révèle plus contraignante pour l'étude d'organismes composés d'une majorité de séquences divergentes.

C'est notamment le cas de *Plasmodium falciparum*, principal agent de la forme létale du paludisme humain ou *malaria*. Le paludisme est la maladie parasitaire la plus répandue dans le monde. En 2006, le paludisme était endémique dans 109 pays, essentiellement les plus pauvres d'Afrique, d'Asie et d'Amérique latine, et deux milliards d'individus, soit 40% de la population mondiale, étaient exposés. On estime à 500 millions le nombre de personnes atteintes de paludisme et entre 1,5 et 3 millions le nombre de décès causés par la maladie chaque année, principalement des enfants de moins de 5 ans habitant dans les zones d'Afrique sub-saharienne (Rapport de l'Organisation Mondiale de la Santé, 2006). Le séquençage du parasite *Plasmodium falciparum* (Gardner *et al.*, 2002) a révélé un génome atypique, composé à 80% de A et T qui se traduit également par un biais de la composition en acides aminés de ses protéines. Une autre particularité de son génome est la présence de longues insertions de faible complexité (Wootton et Federhen, 1993). Ces atypicités rendent particulièrement difficiles les recherches de séquences homologues. Lors de la publication de son génome (Gardner *et al.*, 2002), plus de 60% des 5484 protéines prédites ne possédaient aucune protéine homologue connue. À titre de comparaison, dans la plupart des organismes séquencés ce nombre s'approche plutôt des 40%. On constate par ailleurs que la recherche de domaines Pfam n'identifie que peu de domaines distincts, et ces domaines concernent à peine 50% de ses protéines. Bien que cette observation puisse s'expliquer par l'existence de gènes spécifiques au genre *Plasmodium*, elle est vraisemblablement exacerbée par l'importante divergence du génome de ce parasite qui rend la détection d'homologie particulièrement difficile avec les outils classiques.

L'objectif de cette thèse est d'apporter des méthodes nouvelles pour affiner la détection de domaines protéiques au sein de protéines divergentes. Deux axes principaux ont été développés. Dans un premier temps, nous avons conçu une approche utilisant les propriétés de *co-occurrence* des domaines protéiques. Différentes études révèlent que la plupart des domaines n'apparaissent qu'avec un nombre limité d'autres domaines "favoris" (Cohen-Gihon *et al.*, 2007). Nous proposons donc d'apprendre une liste de paires de domaines fortement corrélés sur un très grand nombre de protéines par exemple celles de Swiss-Prot/TrEMBL. On s'autorise alors à relâcher les seuils de score requis pour la détection de domaines afin d'obtenir de nombreux nouveaux domaines *potentiels*, parmi lesquels se trouvent un grand nombre de faux positifs. La présence d'un domaine potentiellement présent est *certifiée* s'il existe dans la protéine un domaine dit *validant* tel que la paire (validant, potentiel) appartienne à la liste de paires de domaines fortement corrélés précédemment apprise. Nous utilisons donc les propriétés de co-occurrence des domaines comme un filtre pour retenir les domaines qui ont le plus de chance d'être réellement présents. Cette méthode est accompagnée d'une estimation du taux d'erreur sur l'ensemble des domaines certifiés.

Nous avons ainsi pu mesurer sa capacité à découvrir de nouveaux domaines dans différents organismes divergents. De plus, cette approche semble pouvoir apporter un certain nombre d'information y compris pour l'étude des organismes modèles. Dans un second temps, nous avons proposé un large éventail de méthodes de corrections des HMM pour l'étude d'organismes possédant un fort biais dans la composition moyenne en acides aminés de leurs protéines. Nous avons proposé plusieurs méthodes dont l'objectif est de corriger à l'aide d'approches statistiques et évolutives, les probabilités des modèles pour obtenir des prédictions de domaines plus en adéquation avec les organismes étudiés. Nous montrons que ces techniques offrent un ensemble de résultats complémentaires intéressants et à considérer pour l'étude d'organismes biaisés comme *P. falciparum*.

Dans un premier chapitre, nous présentons les pré-requis biologiques nécessaires pour appréhender les données manipulées au cours de cette thèse. Le domaine protéique, unité d'évolution qui fait sens en terme d'homologie de protéines, en est l'objet central. Ce chapitre est conclu par la présentation des différentes bases de données de familles protéiques. Le second chapitre s'intéresse aux différents modèles mathématiques utilisés pour représenter les familles de séquences. Après une brève présentation des modèles usuels, nous nous concentrons sur les modèles de Markov cachés. Nous détaillons notamment l'approche par HMM profils et le logiciel HMMER qui permet leur manipulation. Le troisième chapitre est consacré au paludisme et au responsable de sa forme la plus létale, *Plasmodium falciparum*. Nous présentons les atypicités du génome de ce parasite ainsi que les approches bioinformatiques qui ont été proposées pour les contourner.

La contribution de cette thèse correspond aux chapitres 4 et 5. Nous y présentons les méthodes que nous avons mises en place pour pallier les limites de la détection de domaines dans les protéines divergentes et en particulier chez *P. falciparum* : la méthode de détection de nouveaux domaines protéiques par co-occurrence (chapitre 4) et différentes méthodes de corrections des modèles probabilistes, conçues afin d'adapter les HMM à l'étude de protéines divergentes (chapitre 5). Nous concluons par une discussion sur l'aboutissement des méthodes et les perspectives envisagées.

Cette thèse s'est inscrite dans le cadre du projet ANR *plasmoexplore* dont l'objectif est l'amélioration de l'annotation de *P. falciparum*. Elle a donné lieu à une publication dans le journal *Bioinformatics* concernant le chapitre 4 (Terrapon *et al.*, 2009) et quatre présentations orales en congrès dont JOBIM 2009 et ISCB Africa ASBCB 2009. Elle a été effectué au Laboratoire d'Informatique, Robotique et Microélectronique de Montpellier (LIRMM), dans l'équipe Méthodes et Algorithmes pour la Bioinformatique (MAB) d'Olivier Gascuel en 2005–2008, et en 2010 au Laboratoire de Physiologie Cellulaire Végétale du CEA de Grenoble, dans l'équipe d'Éric Maréchal, en 2009.

Troisième partie

État de l'art

Chapitre 1

Protéines et domaines protéiques

1.1 Protéines

Les protéines sont des macromolécules essentielles pour la structuration et le fonctionnement des cellules vivantes. Selon leur nature, elles peuvent avoir des fonctions différentes :

- un rôle structurel, comme l'actine ;
- un rôle dans la motilité, comme la myosine ;
- un rôle catalytique (enzyme) ;
- un rôle de régulation de la compaction de l'ADN (histone) ;
- un rôle d'expression des gènes (facteur de transcription) ;
- *etc.*

En fait, l'immense majorité des fonctions cellulaires est assurée par des protéines. De nos jours, la caractérisation de la fonction des protéines est donc une des tâches essentielles en bioinformatique dans une visée fondamentale (avancée des connaissances) ainsi que dans une visée appliquée (identifier par exemple les protéines clefs de certaines pathologies humaines).

1.1.1 Origines

Les protéines furent reconnues comme une classe distincte de molécules biologiques au XVIII$^{\text{ème}}$ siècle. Le terme protéine fut proposé par Gerardus Johannes Mulder (1802-1880) en collaboration avec Jöns Jakob Berzelius (1779-1848). Ce terme est dérivé du grec ancien πρωτειος (*proteios*) qui signifie premier, essentiel. Un extrait d'une lettre de Berzelius à Mulder datée de 1838, reprise par Reynolds et Tanford (2003), donne la précision suivante (en français dans le texte) :

> "Le nom protéine que je vous propose pour l'oxyde organique de la fibrine et de l'albumine, je voulais le dériver de πρωτειος, parce qu'il paraît être la substance primitive ou principale de la nutrition animale".

On remarque également une similitude étonnante du terme "protéine" avec l'adjectif "protéiforme" d'étymologie différente. Ce dernier fait référence au Dieu Grec Protée qui pouvait adopter différentes formes, tout comme les différentes protéines possèdent des formes dis-

tinctes et assurent donc de multiples fonctions, bien que ceci ne fût découvert que bien plus tard, au cours du XXe siècle.

1.1.2 Synthèse

Une protéine se compose d'une ou plusieurs séquences (ou chaînes) d'acides aminés, qui sont assemblées à partir de l'information présente dans les gènes. Les acides aminés sont des molécules organiques possédant un squelette carboné et deux fonctions : une amine (-NH2) et un acide carboxylique (-COOH). Les propriétés physico-chimiques des acides aminés, spécifiques du squelette carboné, sont donc déterminantes pour la structure et la fonction des protéines (Figure 1.2).

La synthèse d'une protéine se fait en plusieurs étapes (*cf.* Figure 1.1) :

- **la transcription** où le gène, présent sur le brin codant de l'ADN [1], est transcrit en ARN messager (ARNm) ;
- **la maturation** où l'ARNm subit un ensemble de modifications post-transcriptionnelles. La plupart des ARNm sont modifiés post-transcriptionnellement, mais la nature des ARN modifiés et des modifications varient entre procaryotes et eucaryotes, et entre le noyau et les organites chez les eucaryotes. Les modifications les plus courantes concernent ses extrémités (ajout d'une coiffe ou polyadénylation par exemple), sa séquence (épissage ou édition) ainsi que la nature chimique de ses atomes (méthylation, pseudouridylation, thiolation, *etc.*). Ces différentes modifications peuvent influer sur différentes caractéristiques de l'ARN, telle que sa stabilité, sa capacité à être traduit ou bien même modifier la séquence à traduire. Par exemple, un épissage alternatif conduit à la création d'ARN matures différents et donc de protéines distinctes à partir d'un même gène. La maturation des ARNm est donc une étape importante du contrôle de l'expression des gènes ;
- **la traduction** où l'ARNm mature est traduit en protéine par le ribosome en fonction du code génétique : à chaque triplet de nucléotides (ou codon) correspond un acide aminé (Crick *et al.*, 1961) (*cf.* Table 1.1). L'assemblage d'une protéine se fait donc acide aminé par acide aminé, de manière séquentielle, par des liaisons peptidiques (covalentes) entre les fonctions carboxylique et amine des acides aminés successifs. Pour symboliser l'ordonnancement de cette séquence d'acides aminés, une protéine se lit de son extrémité N-terminale à son extrémité C-terminale : N faisant référence à l'amine NH2 restée libre du premier acide aminé, et C à l'acide carboxylique COOH non engagé dans une liaison peptidique du dernier acide aminé ;

Une fois synthétisée, la protéine peut également subir des modifications dites post-traductionnelles. Un modification post-traductionnelle est une modification chimique d'une protéine (acétylation, phosphorylation, ubiquitination, *etc.*), réalisée le plus souvent par une

1. L'acide désoxyribo-nucléique (ADN) constitue le génome. Il se compose d'une séquence de nucléotides, c.-à-d. un acide phosphorique lié à un désoxyribose, lui-même lié à une base azotée. Il existe 4 bases azotées : l'adénine (A), la thymine (T), la guanine (G) et la cytosine (C). La structure de l'ADN, deux brins complémentaires composés de paires A-T et G-C s'enroulant en double hélice, a été établi par Watson et Crick (1953).

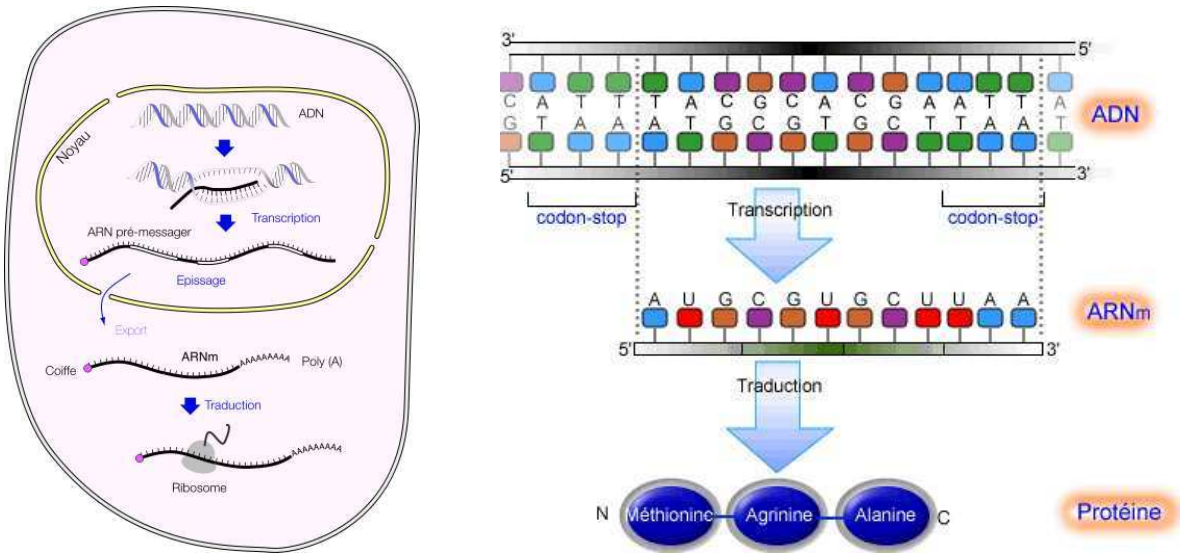


FIGURE 1.1 – **Procédé général de la synthèse d'une protéine.** À gauche, on trouve l'illustration des mécanismes de synthèse à travers les différentes localisations cellulaires dans une cellule eucaryote. À droite, on voit l'évolution de la séquence d'ADN d'un gène à la séquence d'acides aminés d'une protéine, lors de la traduction et de la transcription.

enzyme, après sa synthèse ou au cours de sa vie dans la cellule. Cette modification entraîne généralement un changement de la fonction de la protéine considérée, que ce soit au niveau de son action, de sa demie-vie, ou de sa localisation cellulaire.

1.1.3 Différents niveaux de structure

On distingue communément plusieurs niveaux de structure pour décrire une protéine :

a) Structure primaire : L'ordre dans lequel les acides aminés s'enchaînent constitue la structure primaire de la protéine. On parle alors de séquence d'acides aminés ou séquence protéique. Les acides aminés issus du code génétique sont au nombre de 22 (*cf.* Table 1.1). Cependant, deux de ces acides aminés sont extrêmement rares. La *sélénocystéine* (U) et la *pyrrolysine* (O) sont codés par des codons-stop (codons déclenchant habituellement l'arrêt immédiat de la transcription d'un gène). De plus, la pyrrolysine n'apparaît que chez les archées et les eubactéries.

b) Structure secondaire : Le concept de structure secondaire a été introduit par Linderstrøm-Lang (1952), lors des conférences médicales Lane à Stanford (Schellman et Schellman, 1997). La structure secondaire décrit le repliement de segments courts de la structure primaire. Ce sont des structures locales, stabilisées par des liaisons hydrogènes entre les groupements amide (-NH) et carbonyle (-CO) du squelette peptidique (Pauling *et al.*, 1951). L'existence de structures secondaires vient du fait que les repliements

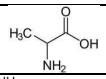
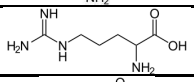
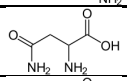
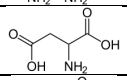
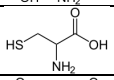
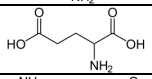
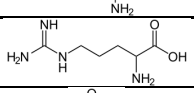
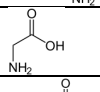
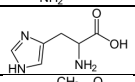
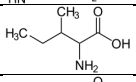
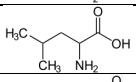
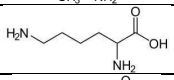
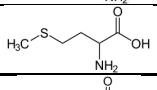
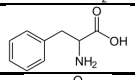
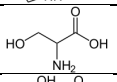
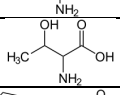
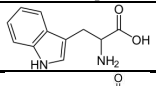
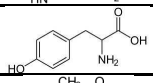
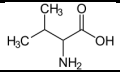
Acide aminé	Code à 3 lettres	Code à 1 lettre	Formule	Traduit par les codons
Alanine	Ala	A		GCU, GCC, GCA, GCG
Arginine	Arg	R		CGU, CGC, CGA, CGG, AGA, AGG
Asparagine	Asn	N		AAU, AAC
Acide aspartique	Asp	D		GAU, GAC
Cystéine	Cys	C		UGU, UGC
Acide Glutamique	Glu	E		GAA, GAG
Glutamine	Gln	Q		CAA, CAG
Glycine	Gly	G		GGU, GGC, GGA, GGG
Histidine	His	H		CAU, CAC
Isoleucine	Ile	I		AUU, AUC, AUA
Leucine	Leu	L		UUA, UUG, CUU, CUC, CUA, CUG
Lysine	Lys	K		AAA, AAG
Méthionine	Met	M		AUG
Phénylalanine	Phe	F		UUU, UUC
Proline	Pro	P		CCU, CCC, CCA, CCG
Sérine	Ser	S		UCU, UCC, UCA, UCG, AGU, AGC
Thréonine	Thr	T		ACU, ACC, ACA, ACG
Tryptophane	Trp	W		UGG
Tyrosine	Tyr	Y		UAU, UAC
Valine	Val	V		GUU, GUC, GUA, GUG

TABLE 1.1 – **Table des 20 acides aminés usuels issus du code génétique.**

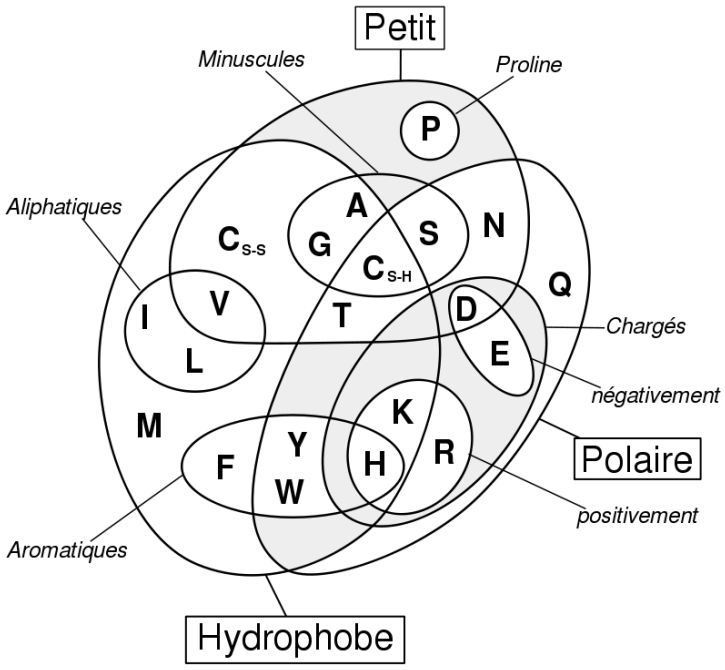


FIGURE 1.2 – **Propriétés physico-chimiques des différents acides aminés** d'après Taylor (1986a).

énergétiquement favorables de la séquence protéique sont limités et que seules certaines conformations sont possibles. Les principaux types de structures secondaires régulières sont les *hélices* α et les *feuillets* β. Ils se trouvent fréquemment dans les protéines mais en proportions et combinaisons variables. On trouve aussi des *coudes* (également nommés *retour en arrière* ou *tournants* — traductions de l'anglais *turns*), ainsi que des *boucles* (*coil*). Ces structures ne sont pas moins ordonnées que les hélices ou les feuillets ; elles sont plus irrégulières et donc plus difficile à décrire. Il ne faut donc pas confondre le terme boucle avec le terme *enroulement au hasard* (*random coil*), qui désigne l'ensemble des conformations non-structurées. Les hélices, feuillets, coudes et boucles comptent pour environ 90% en moyenne dans les protéines standards (dont 31% d'hélices et 28% de feuillets) (Voet et Voet, 2004). Il est connu que certains acides aminés favorisent la formation d'une structure secondaire plutôt qu'une autre (Blout *et al.*, 1960). Par exemple, la proline (P) et la glycine (G) ont une très faible propension à former des hélices α et sont même considérées comme des "briseuses d'hélice" (*helix breaker*), parce qu'elles détruisent la régularité du squelette de l'hélice α (Argos et Palau, 1982). En revanche, elles ont des capacités conformationnelles particulières et se retrouvent fréquemment dans les coudes. Les acides aminés qui favorisent la formation des hélices sont la méthionine, l'alanine, la leucine, le glutamate et la lysine ("MALEK" en code acide aminé à une lettre). À l'inverse, les "gros" acides aminés aromatiques (tryptophane, tyrosine et phénylalanine) et les acides aminés branchés en C^β (isoleucine, valine et threonine) privilégient la conformation en feuillet β. Cependant, ces tendances ne sont pas suffisamment marquées

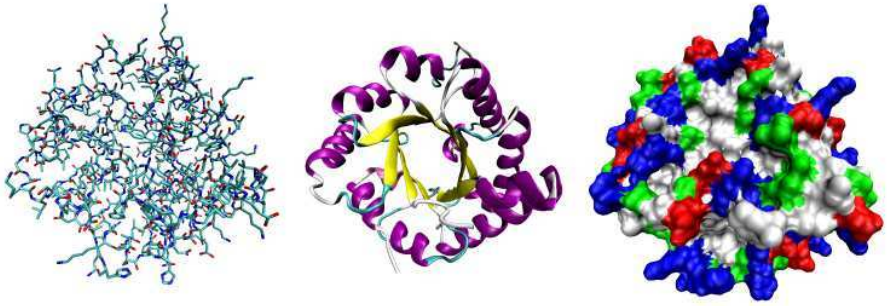


FIGURE 1.3 – Trois représentations possibles de la structure tertiaire de la protéine triose phosphate isomerase. À gauche : représentation de tous les atomes, les différentes couleurs correspondent aux différents types d'atomes. Au milieu : représentation simplifiée de la conformation du squelette de cette protéine et coloration des différentes structures secondaires. Les hélices α sont figurées par des hélices circulaires et les feuillets β par des flèches. À droite : représentation des surfaces accessibles aux solvants, colorées par rapport aux types de résidus (acides en rouge, basiques en bleu, polaires en vert et non polaires en blanc).

pour pouvoir servir de base à la prédiction de structure secondaire, sur la base de la seule séquence en acides aminés.

De nombreuses approches ont été décrites pour prédire la structure secondaire d'une protéine à partir de sa séquence primaire, basées sur les propriétés physico-chimiques, de simples statistiques linéaires, ainsi que de nombreuses méthodes d'apprentissage (réseaux neuronaux, k-plus proches voisins, arbres phylogénétiques, *etc.*) (Cuff et Barton, 2000). Les plus performantes exploitent l'information évolutive disponible grâce aux familles de protéines (Zvelebil *et al.*, 1987; Rost et Sander, 1993; Salamov et Solovyev, 1995; Frishman et Argos, 1996; King et Sternberg, 1996). Il est également possible de calculer la structure secondaire d'une protéine à partir de sa structure tertiaire (voir ci-dessous), par exemple grâce au dictionnaire de structure secondaire de protéines (DSSP) (Kabsch et Sander, 1983), ou grâce à des algorithmes tels que HSSP (*Homology-derived Secondary Structure of Proteins*) (Sander et Schneider, 1991) ou STRIDE (*secondary STRucture IDEntification*) (Frishman et Argos, 1995).

c) Structure tertiaire : La structure tertiaire d'une protéine correspond au repliement de la séquence protéique dans l'espace. On parle plus couramment de structure 3D ou structure tridimensionnelle. La fonction d'une protéine est intimement liée à sa structure 3D. Lorsque cette structure est cassée par l'emploi d'un agent dénaturant ou autre (température, ou pression par exemple), la protéine perd sa fonction (Sela *et al.*, 1957).

La structure tertiaire d'une protéine dépend principalement de sa structure primaire (Anfinsen, 1973). Ainsi, deux protéines homologues [2] ayant une forte similarité de leur sé-

2. L'homologie est une relation rendant compte d'une histoire évolutive commune (par opposition à la

quence primaire auront également des structures très proches et donc des fonctions semblables. Il n'existe pas de règles simples qui permettraient de déduire la structure 3D à partir de la séquence d'acides aminés.

Différentes méthodes expérimentales permettent de déterminer la structure tertiaire des protéines, notamment :

- la cristallographie par rayons X ;
- la spectroscopie par résonance magnétique nucléaire (RMN).

Cependant, ces méthodes sont coûteuses et la détermination de la structure d'une protéine reste un processus lent et complexe. Afin de contourner ce problème, des méthodes automatiques de prédiction de la structure tertiaire des protéines ont été développée et sont aujourd'hui très fiables. Il se dégage deux types de méthodes : les méthodes de modélisation comparative et les méthodes dites *ab initio*, ou *de novo*.

Les méthodes de modélisation comparative ont pour point de départ les structures 3D résolues (*templates*). On distingue deux approches classiques :

- Les méthodes par reconnaissance des repliements, également appelées *Protein Threading*, disposent d'une littérature abondante (Miyazawa et Jernigan, 1985; Sippl, 1990; Bowie *et al.*, 1991; Jones *et al.*, 1992a; Godzik *et al.*, 1992; Maiorov et Crippen, 1992; Bryant et Lawrence, 1993; Johnson *et al.*, 1993; Lathrop, 1994). Il s'agit d'utiliser chaque protéine dont la structure est connue pour reconnaître les séquences susceptibles d'adopter un repliement identique. Ces méthodes adressent le *problème du repliement inverse*, puisqu'elle examinent la compatibilité d'une séquence avec des structures connues, au lieu de prédire la structure à partir de la séquence. L'existence d'un répertoire limité de structures protéiques permet de calculer un potentiel d'énergie (simplifié) du repliement des chaînes d'acides aminés.
- Les méthodes de modélisation probabiliste de séquences homologues dont les structures 3D sont connues, comme le propose les bases de données Gene3D (Wilson *et al.*, 2009) et SUPERFAMILY (Yeats *et al.*, 2008) développées section 1.4.3. L'homologie de séquences protéiques aux modèles de familles permet de prédire leurs structures.

Les méthodes dites *ab initio* ne sont pas limitées aux familles pour lesquelles il existe une structure résolue. Ces méthodes utilisent les principes physico-chimiques afin de prédire la structure 3D d'une protéine, uniquement à partir de la séquence d'acides aminés. Les composantes clefs de ces méthodes consistent à minimiser une fonction d'énergie, soit en simulant le repliement de la protéine (à partir d'une chaîne protéique *à plat*), soit en recherchant dans l'espace des conformations possibles grâce à des méthodes stochastiques (Kolinski *et al.*, 1993). L'inconvénient de ces méthodes est la nécessité d'importantes ressources informatiques. Elles n'ont donc été appliquées avec succès qu'à de petites protéines. Bien que les limitations calculatoire soit fortes, les bénéfices potentiels pour la génomique structurelle font des méthodes *ab initio* de prédiction de structure un champ de recherche actif (Zhang, 2008).

Il existe aussi des protéines dites "non structurées", n'ayant pas de structure 3D particulière sauf lorsqu'elles entrent en interaction avec d'autres facteurs (Dyson et Wright, 2005).

notion d'analogie où la ressemblance provient d'une convergence sans lien évolutif). Des séquences homologues sont issues de l'évolution d'une même séquence ancestrale.

Les protéines intrinsèquement non structurées représenteraient environ 10% des génomes (Tompa, 2002). Plus généralement, environ 40% des protéines eucaryotes posséderaient une région intrinsèquement non structurée (Wright et Dyson, 1999). La base de données Disprot (Sickmeier *et al.*, 2007) fournit des informations supplémentaires sur ces protéines et répertorie toutes les structures intrinsèquement désordonnées connues.

d) Structure quaternaire La structure quaternaire [3] des protéines se réfère à ce que l'on appelle les complexes protéiques. Un complexe protéique regroupe l'association d'au moins deux séquences protéiques — identiques ou différentes — par des liaisons non-covalentes (liaison hydrogène, liaison ionique, interactions hydrophobes), et, plus rarement, des ponts disulfures. Chacune de ces séquences est appelée *monomère* (ou sous-unité) et l'ensemble *oligomère* ou *protéine multimérique*. L'hémoglobine humaine est un exemple de protéine multimérique ; elle est constituée de 4 sous-unités : 2 sous-unités α (de 141 acides aminés) et 2 sous-unités β (de 146 acides aminés). La fonction de la protéine émerge alors de cet agrégat de structures.

1.2 Famille de séquences

L'accroissement exponentiel des données protéiques issues des projets de séquençage de génomes complets aboutit à un goulot d'étranglement (Benson *et al.*, 2009). Identifier la fonction d'autant de protéines par des expérimentations biologiques (étude de leur structure 3D, de leurs interactions, *etc.*) n'est simplement pas réalisable. On ne dispose que de séquences primaires dont il faut tirer le maximum d'information. Le recours actuel pour traiter ces données consiste à utiliser des méthodes bioinformatiques pour prédire une annotation fonctionnelle rapide des protéines récemment séquencées. L'annotation fonctionnelle des protéines passe alors par l'identification de groupes de séquences, plus communément appelés *familles*. Les familles visent à regrouper les protéines homologues, c.-à-d. qui partagent une histoire évolutive commune. Les séquences d'une même famille sont souvent proches et possèdent donc des propriétés semblables. L'intérêt d'une telle classification est de permettre le transfert d'annotations. Lors de l'étude d'une nouvelle protéine (encore non-annotée), on l'assigne à une famille afin de transférer à cette protéine les connaissances acquises sur les autres protéines de la famille (issues de précédentes études). Le regroupement en familles des protéines peut se faire selon la similarité des séquences ou des structures, ou encore la proximité des compositions en domaines protéiques (*cf.* section 1.3). Des méthodes récentes proposent de s'appuyer sur des combinaisons de ces critères avec des données biologiques telles que les voies métaboliques ou les profils d'expression (Hahne *et al.*, 2008; Bréhélin *et al.*, 2010). Cependant, il est difficile d'assigner à une famille les séquences les plus divergentes. Cette thèse s'applique donc à la modélisation et l'identification de familles de domaines protéiques afin d'améliorer l'annotation fonctionnelle des protéines divergentes.

3. terme introduit par (Bernal, 1958), par extension de la terminologie de Linderstrøm-Lang.

1.3 Domaines protéiques

La majorité des protéines, y compris dans les organismes les moins complexes, est composée de plusieurs modules ou *domaines*. Dans un premier temps, nous revenons sur la définition du domaine protéique (section 1.3.1) et la distinction entre domaine et motif protéique (section 1.3.2). Puis nous présentons les principales propriétés révélées par les nombreux travaux sur la combinatoire des domaines, en commençant par l'existence de groupes de domaines conservés au cours de l'évolution, et l'implication de l'apparition de nouvelles combinaisons dans la complexification des organismes (section 1.3.3). Nous verrons qu'il existe un répertoire relativement limité de combinaisons de domaines observables dans la nature (section 1.3.4). Cette propriété sert de point d'ancrage à notre méthode de détection de domaines par co-occurrence (*cf.* Chapitre 4). D'autres travaux se sont intéressés aux mécanismes de création de gènes à travers les phénomènes de recombinaisons de domaines (section 1.3.5). L'évolution a conduit de nombreux domaines à se combiner pour former des protéines multidomaines ou des complexes protéiques, avec *a priori* un vaste espace de possibilités, faisant du domaine une unité d'évolution à part entière (section 1.3.6). On observe une forte corrélation entre la fonction et la composition domaines des protéines (section 1.3.7). Pour conclure sur les domaines protéiques, nous présentons les résultats d'études portant sur les interactions protéine-protéine en terme d'interactions entre domaines (section 1.3.8).

1.3.1 Définition

L'analyse des séquences et des structures de protéines révèle que beaucoup s'organisent en modules structuraux distincts. En effet, les protéines peuvent être vues comme composées d'une ou plusieurs unités fondamentales appelées *domaines protéiques*. Il existe plusieurs définitions du domaine protéique selon l'angle sous lequel on se place. Du point de vue structuraliste, un domaine correspond à une sous-séquence d'acides aminés capable de se replier indépendamment du reste de la protéine. Pour le biochimiste, le domaine est utilisé pour décrire des régions protéiques pour lesquelles une fonction propre a pu être caractérisée : par exemple la fixation d'un ligand, la reconnaissance d'un autre partenaire, l'ancrage membranaire, *etc.* Enfin, pour l'évolutionniste, les domaines sont des sous-séquences homologues, issues d'un domaine ancestral commun et conservés au cours de l'évolution. Bien qu'il n'existe pas de définition unique, ces points de vue ne sont pas incompatibles et les différents spécialistes identifient généralement des zones très proches par leurs techniques respectives (Elofsson et Sonnhammer, 1999; Zhang *et al.*, 2005). En guise de consensus, on peut considérer le domaine protéique comme une unité d'évolution indépendante (*cf.* section 1.3.6).

Un domaine protéique peut constituer à lui seul une protéine. On parle alors de protéine *monodomaine*. Il peut aussi s'associer avec d'autres domaines au sein d'une protéine dite *multidomaine*. Ce domaine conservera sa fonction d'origine ou participera à une fonction différente en collaborant avec les autres domaines. Ainsi, une protéine constituée de plusieurs domaines peut associer plusieurs fonctions distinctes ou en obtenir une nouvelle. L'enjeu des méthodes d'annotation protéique consiste alors à identifier la composition exacte en domaines d'une nouvelle protéine, afin de s'appuyer sur les domaines et les combinaisons de domaines précé-

demment étudiés pour proposer une annotation fonctionnelle de la protéine. L'identification précise des domaines apporte des informations structurelles, fonctionnelles et évolutives (par exemple la localisation cellulaire (Chou et Cai, 2002) ou des annotations dans la *Gene Ontology* (GO — *cf.* paragraphe 1.4.1.c) (Hayete et Bienkowska, 2005; Forslund et Sonnhammer, 2008)).

1.3.2 Domaine et motif protéique

On distingue également au sein des protéines des motifs structuraux ou *motifs protéiques*. Un motif est une séquence courte associée à des interactions bien précises : site actif ou d'ancrage par exemple (Doolittle, 1986). La différence entre domaine et motif est assez mince, et porte souvent à confusion. Elle tient principalement au fait qu'un domaine contenant plusieurs sites d'ancrage est composé de plusieurs motifs. De plus, les motifs n'ont pas forcément de repliement propre. Ils contiennent des résidus essentiels à la fonction et à l'interaction, éventuellement entre-coupés de résidus non-essentiels : il s'agit d'un *pattern*. Cependant, participant à un même site, les acides aminés essentiels peuvent être assez proches dans la structure 3D. Les motifs ont été définis comme des groupes d'acides aminés extrêmement bien conservés entre des séquences globalement différentes (Lesk, 1988). Lesk décrit l'émergence des motifs comme issue des contraintes imposées par l'évolution à certaines portions des séquences pour la conservation de la fonction des sites d'ancrage. Le premier exemple fut le motif appelé "doigt de zinc" (car il fixe l'ion Zn^{2+}) impliqué dans des interactions spécifiques avec l'ADN (Miller *et al.*, 1985). Plus récemment, une publication de Halabi *et al.* (2009) définit le *secteur protéique* comme un groupe d'acides aminés corrélés quasi-indépendants. Trois secteurs sont identifiés dans la famille d'enzyme S1A, dont les acides aminés se trouvent être connectés dans la structure 3D (mais pas dans la séquence primaire). De plus, ces secteurs exhibent des fonctions distinctes et une évolution indépendante au sein de cette famille, ce qui ouvre une réflexion intéressante sur l'émergence de la fonction au sein de certains domaines protéiques.

1.3.3 Conservation de groupes de domaines

Apic *et al.* (2001) publient l'une des premières études portant sur la combinatoire des domaines. Ils montrent qu'un grand nombre des paires de domaines contigus sont sur-représentées à travers les trois domaines du Vivant, et que les recombinaisons de domaines ont été un facteur clef dans la divergence des organismes. L'une des observations majeures concerne la conservation de l'orientation N-C terminale des combinaisons de domaines : 90% des paires de domaines sont toujours dans le même ordre.

Vogel *et al.* (2004) identifient des groupes de domaines conservés (paires et triplets de domaines contigus appelés *supradomaines*) impliqués dans de nombreuses fonctions et dont la moitié sont sur-représentés dans les trois domaines de la vie. Leur étude suivante (Vogel *et al.*, 2005) suggère que ces combinaisons peuvent être considérées comme des unités d'évolution indépendantes, à un niveau supérieur à celui du domaine seul. Ces publications ont permis d'établir le scénario évolutif suivant : la création de nouvelles combinaisons de domaines, associée à l'expansion des familles de domaines et certains mécanismes tels que l'épissage

alternatif chez les eucaryotes, joueraient un rôle prépondérant dans la complexité croissante des organismes (Vogel *et al.*, 2004). La combinaison de domaines serait un processus aléatoire à la suite duquel certaines combinaisons seraient largement dupliquées ou disparaîtraient (Vogel *et al.*, 2005).

Wutchy et Almaas (2005) ont étudié les combinaisons de domaines à l'aide de réseaux de domaines co-occurrents, en se concentrant sur les organismes eucaryotes modèles : *S. cerevisiae*, *C. elegans*, *D. melanogaster*, *M. musculus* et *H. sapiens*. En comparant ces réseaux, ils constatent que les eucaryotes unicellulaires et pluricellulaires possèdent le même nombre de domaines, tandis que la taille des sous-graphes connexes grandit avec l'évolution. Cela suggère que l'augmentation de la complexité des organismes multicellulaires provient de la formation de nouvelles combinaisons de domaines, en particulier chez les métazoaires (Ekman *et al.*, 2007).

Les travaux de Cohen-Gihon *et al.* (2007) portent sur l'étude des protéines de la levure et de ses complexes protéiques. Ils ont proposé une représentation permettant l'identification des combinaisons (de toute taille) de domaines co-occurrents non contigus sur-représentés, appelées *co-occurring domains sets* (CDS). Ils constatent que ces CDS contiennent une fraction significative de domaines anciens, c.-à-d. que l'on retrouve chez les archées et les bactéries. L'observation des domaines communs aux différents CDS ainsi que la comparaison aux protéines monodomaines ont révélé que les protéines hautement modulaires ont tendance à être composées de domaines très abondants et, *a contrario*, les protéines monodomaines à contenir des domaines rares.

1.3.4 Existence d'un répertoire limité de combinaisons

L'observation des combinaisons de domaines montre que peu de types de domaines sont versatiles (Apic *et al.*, 2001; Vogel *et al.*, 2004; Bornberg-Bauer *et al.*, 2005; Weiner *et al.*, 2008). Au sein des protéines, les domaines n'apparaissent qu'avec un nombre très réduit d'autres domaines "favoris", auxquels ils sont liés par une forme de coopération fonctionnelle ou structurelle. Cette propriété est à l'origine d'approches de prédiction d'annotations fonctionnelles de protéines. On peut citer les travaux de :

- Geer *et al.* (2002) dont l'outil CDART permet à l'utilisateur de rechercher les protéines ayant une composition en domaines similaire à une protéine requête ;
- Scott *et al.* (2004) utilisant des réseaux bayésiens de motifs co-occurrents pour prédire la localisation sub-cellulaire de protéines ;
- McLaughlin *et al.* (2007) qui caractérisent des assemblages de domaines (*DOMain ASSembly* — DASSEM), c'est à dire des groupes de domaines qui coopèrent les uns avec les autres pour réaliser une fonction particulière ;
- Forslund et Sonnhammer (2008) qui proposent une approche pour prédire des annotations GO spécifiques pour des groupes de domaines.

Il existe donc un répertoire très limité de combinaisons de domaines dans la nature, qui ne représente qu'une infime partie des combinaisons possibles. Les mécanismes aboutissant à la création de combinaison de domaines sont probablement soumis à une forte pression de sélection (Apic *et al.*, 2003). Apic et ses collaborateurs ont dénombré, parmi 85 génomes com-

plètement séquencées, 796 familles de domaines différentes (base de données SUPERFAMILY version 1.61 — *cf.* section 1.4.3.d), soit environ 634 000 paires possibles en théorie. Or, ils recensent dans ces organismes seulement 2 545 paires différentes, soit 4‰ des combinaisons possibles. Nous avons reproduit cette observation sur les domaines de la base Pfam (*cf.* section 1.4.2.c — version 23.0 datant de Juillet 2008). En considérant la composition en domaines de l'ensemble des protéines d'Uniprot (*cf.* section 1.4.1), nous avons recensé environ 20 000 paires de domaines distinctes, formées sur plus de 5 000 domaines Pfam, soit 1,6‰ des 1,25 millions de paires possibles en théorie. L'approche de certification de domaines par co-occurrence que nous avons conçue exploite cette propriété pour la découverte de nouveaux domaines dans les protéines divergentes (*cf.* Chapitre 4).

1.3.5 Mécanismes évolutifs recombinants

L'existence de groupes de domaines conservés par l'évolution, a conduit à l'étude des mécanismes de recombinaison des domaines protéiques. Björklund *et al.* (2005) proposent une classification des types de réarrangement en différenciant 3 évènements élémentaires :

- la *substitution* de domaine correspond à l'échange d'un domaine par un autre ;
- l'*indel* désigne l'insertion ou la délétion d'un domaine différent des domaines adjacents au point d'insertion dans l'architecture en domaines [4] ;
- la *répétition* représente l'addition d'un domaine identique à l'un des domaines adjacents au point d'addition dans l'architecture en domaines.

Les réarrangement les plus complexes peuvent donc être décrit par une combinaison de ces réarrangements élémentaires. Ils définissent alors une mesure d'évolution définie comme le nombre de domaines sans correspondance dans un alignement de deux architectures en domaines et appelée *distance en domaines*. Leurs résultats ont montré que les indels sont plus fréquents que les répétitions internes, et que les substitutions de domaines sont rares. De plus, les indels et les répétitions sont plus souvent observés aux extrémités N et C-terminales des protéines, tandis qu'ils sont rares entre les domaines. Enfin, selon eux, l'évolution de la majorité des protéines multidomaines pourrait s'expliquer par des insertions de domaines seuls, à l'exception des répétitions de domaines qui réalisent parfois la duplication de plusieurs domaines en tandem. Certaine de ces conclusions sont confirmées par Pasek *et al.* (2006a) dans les architectures multidomaines chez les bactéries. Ces travaux ont mis en évidence qu'un mécanisme majeur à l'origine de la création de nouvelles combinaisons de domaines est fortement lié au jeu des fusions/fissions de gènes. Kummerfeld et Teichmann (2005) ont d'ailleurs estimé qu'il se réalisait 4 fois plus d'évènements de fusion que de fission. Dans une autre publication, Pasek *et al.* (2006b) abordent la redondance des génomes en domaines et pose l'hypothèse d'un lien avec la robustesse des organismes aux mutations silencieuses. La conservation de plusieurs domaines dans les gènes dupliqués partiels ou leur intégration dans une protéine existante permettrait la conservation de la fonction biologique en cas de délétion de la protéine initiale. Ces travaux ont montré que la redondance en domaines est un mécanisme de compensation moins important que la redondance en gènes mais qui n'est pas négligeable. Par ailleurs, Weiner *et al.* (2006) ont choisi de s'intéresser plus spécifiquement aux délétions

4. L'*architecture en domaines* désigne l'ordre séquentiel des domaines composant une protéine.

de domaines. Leur résultats montrent que les pertes de domaines peuvent être expliquée par l'introduction de codons *start/stop* qui rendent la terminaison du domaine non-fonctionnelle et conduisent à sa disparition (n'étant plus soumis à une pression de sélection...). Enfin, ils confirment que la perte et la duplication de domaines ont principalement lieu aux extrémités des protéines, et plus fréquemment à l'extrémité C-terminale quand ces évènements impliquent des protéines monodomaines.

1.3.6 Une unité d'évolution indépendante

Il est maintenant clair que les principaux mécanismes gouvernant l'apparition de nouvelles protéines sont :

- la recombinaison de domaines donnant naissance à une nouvelle architecture en domaines ;
- la duplication de séquences codant un ou plusieurs domaines, suivie de la divergence des séquences (mutations, délétions, insertions d'acides aminés) conduisant à une structure modifiée et sélectionnée par l'évolution.

Chaque domaine peut donc faire l'objet de mutations ponctuelles, être entièrement dupliqué, inversé, transposé, délété, inséré ou impliqué dans des transferts horizontaux. Par conséquent, il convient de considérer le domaine comme une unité d'évolution à part entière. En tant que tel, on peut s'interroger sur la relation d'homologie qui existe entre deux protéines alors que toutes les parties de ces protéines n'ont pas la même histoire (Fitch, 2000). Fitch traite des problèmes liés à la notion d'homologie et qu'il évoque comme "*The recombination problem*". Il arrive à la conclusion que, lorsqu'on veut tenir compte des réarrangements de domaines, le gène n'est pas l'unité adéquate pour parler d'orthologie [5] ou de paralogie [6]. Dans ce cas, le domaine s'avère être une unité plus pertinente. Koonin *et al.* (2000) généralisent cette conclusion en suggérant de revoir l'ensemble des processus évolutifs en terme de domaines plutôt qu'en terme de protéines. Dans sa thèse de doctorat Pasek (2006) transpose des thématiques classiques de la génomique comparative en les appliquant aux domaines plutôt qu'aux gènes. Par exemple, la recherche de zones de synténies en domaines protéiques lui a permis d'identifier des régions correspondantes plus nombreuses et plus larges (Pasek *et al.*, 2005). D'autres thématiques classiques ont également été adaptées au niveau des domaines dans le cadre des réarrangements de domaines, notamment la permutation circulaire (Weiner *et al.*, 2005; Weiner et Bornberg-Bauer, 2006).

5. Orthologie : homologie entre des séquences qui appartiennent à des organismes différents. Issues d'une séquence ancestrale commune, elles ont divergées après un évènement de spéciation (apparition de nouvelles espèces).

6. Paralogie : homologie entre des séquences issues d'une duplication au sein d'un génome. On peut distinguer les *inparalogues* (paralogues au sein d'un même organisme) des *outparalogues* (paralogues entre des espèces distinctes) selon que l'évènement de duplication a lieu respectivement après ou avant la spéciation (Sonnhammer et Koonin, 2002).

1.3.7 À domaines identiques... fonction identique

Il a rapidement été remarqué une similarité de la fonction des protéines composées de groupes de domaines identiques. Ce qui pouvait être une hypothèse assez naturelle, de par la similarité implicite des séquences primaires, confirme l'existence d'une coopération fonctionnelle entre les domaines. Gerstein et Hegyi (2001) ont étudié la similarité fonctionnelle de protéines partageant les mêmes domaines SCOP (*cf.* paragraphe 1.4.3.b) dans différentes espèces eucaryotes. Ils montrent que :

- deux tiers des protéines monodomaines composées du même domaine ont une fonction similaire ;
- 35% des protéines multidomaines possédant un domaine similaire ont des fonctions semblables ;
- ce taux monte à 80% si elles ont deux domaines distincts en commun (sans tenir compte de l'ordre séquentiel de ces domaines) ;
- si elles ont une composition strictement identique (ordre et nombre d'occurrences identiques), 90% des protéines multidomaines ont la même fonction. Les 10% restants peuvent s'expliquer par le fait qu'il existe différentes configurations spatiales pour une même séquence de domaines.

Ces observations sont confirmées par Ye et Godzik (2004) sur les trois domaines du Vivant. À l'aide de réseaux de domaines co-occurrents, ils forment des classes de groupes de domaines récurrents et constatent que les protéines appartenant à une même classe ont tendance à avoir des fonctions similaires. Les travaux de Cohen-Gihon *et al.* (2007) révèlent eux aussi que les protéines contenant des CDS récurrents interagissent fréquemment avec d'autres protéines contenant le même CDS, participent au même processus biologique et sont associées au même complexe protéique.

1.3.8 Interaction Domaine-Domaine

Une thématique qui possèdent déjà une littérature intéressante et riche est l'interaction entre protéines en terme d'interaction entre domaines. Les domaines protéiques constituent une interface de liaison entre les protéines qui interagissent et entre les protéines d'un même complexe (Pawson et Nash, 2003).

Les différents travaux publiés ont en commun la problématique suivante : identifier les paires de domaines susceptibles d'interagir à partir de données d'interactions protéine-protéine à grande échelle (et des compositions en domaines de ces protéines). Cette identification peut servir divers objectifs :

- prédire de nouvelles interactions protéine-protéine, en se basant sur la simple connaissance du contenu en domaines des protéines (Sprinzak et Margalit, 2001; Wojcik et Schächter, 2001; Deng *et al.*, 2002; Kim *et al.*, 2002; Gomez et Rzhetsky, 2002; Ng *et al.*, 2003; Pagel *et al.*, 2004; Chen et Liu, 2005; Jothi *et al.*, 2006) ;
- nettoyer les données d'interactome issues d'expérimentations à grande échelle [7]

7. techniques double-hybride (Fields et Song, 1989) connues pour leur grand nombre de faux positifs et le faible taux de recouvrement entre deux expériences (Uetz *et al.*, 2000; Ito *et al.*, 2001).

(Betel *et al.*, 2004; Riley *et al.*, 2005) ;
– identifier quels sont les domaines qui interagissent dans une interaction protéine-protéine donnée (Nye *et al.*, 2005; Guimarães *et al.*, 2006).

Ces travaux présentent différentes approches pour établir des paires de domaines qui interagissent et en évaluer la consistance. Certaines approches s'appuient sur la signature des séquences (composition en domaines Interpro — *cf.* section 1.4.4) ainsi que sur des scores (Sprinzak et Margalit, 2001; Kim *et al.*, 2002; Ng *et al.*, 2003), des paires de profils des domaines interagissant (IDPP) (Wojcik et Schächter, 2001), l'estimation du maximum de vraisemblance de leur modèle (Deng *et al.*, 2002), des chaînes de Markov Monte Carlo (Gomez et Rzhetsky, 2002), les profils phylogénétiques des domaines conservés (Pagel *et al.*, 2004), des tests statistiques sur les compositions en domaines (Betel *et al.*, 2004; Nye *et al.*, 2005), l'analyse d'exclusion de paires de domaines (DPEA) (Riley *et al.*, 2005), un système de forêts de décisions aléatoires (tenant compte de paires de domaines multiples) (Chen et Liu, 2005), la co-évolution des séquences (Jothi *et al.*, 2006), le principe de parcimonie (Guimarães *et al.*, 2006).

Des bases de données répertorient les interactions connues entre domaines (iPfam (Finn *et al.*, 2005) et 3DID (Stein *et al.*, 2005)) ainsi que les interactions potentielles (DOMINE (Raghavachari *et al.*, 2008)).

La méthode de certification de nouveaux domaines proposé dans cette thèse (chapitre 4) est semblable à certaines de ces approches (Betel *et al.*, 2004; Nye *et al.*, 2005) sur deux niveaux :

– apprendre les paires de domaines corrélés à travers le calcul de P-valeurs ;
– filtrer les vrais positifs dans des données contenant de nombreux faux positifs grâce aux paires précédemment identifiées.

Cependant, notre approche ne s'applique pas aux domaines en interaction entre des protéines distinctes, mais à l'identification de domaines co-occurrents au sein d'une même protéine.

1.4 Les bases de données de familles de protéines

Différentes bases de données proposent une classification des protéines en familles. Celles-ci se basent sur des critères de similarité entre séquences primaires, secondaires ou tertiaires. On peut les opposer par l'expertise automatique ou semi-manuelle choisie pour la création des familles et des alignements, ou encore par les modèles qu'elles utilisent (expressions régulières, profils, HMM — *cf.* Chapitre 2).

Leurs objectifs restent communs : regrouper les connaissances acquises sur les différentes familles protéiques (alignements multiples, structures 3D, annotations fonctionnelles, références bibliographiques, *etc.*), et rendre accessibles leurs modèles pour l'identification et l'annotation de nouvelles séquences. Dans cette section, avant de détailler les différentes bases de données de familles (sections 1.4.2 et 1.4.3) et la *metadatabase* Interpro visant à les unifier (section 1.4.4), nous présentons les données qu'elles utilisent pour construire leurs classifications et décrire fonctionnellement les familles (section 1.4.1).

1.4.1 Données génomiques, structurelles et fonctionnelles

On trouve sur le Web de nombreuses bases de données contenant des informations sur les protéines. Leurs données servent de support à la création de nouvelles méthodes et au maintien d'approches existantes pour la classification et l'annotation fonctionnelle des familles de protéines.

a) Données de séquences primaires, où l'on distingue les bases de données espèce-spécifique et les bases de données universelles.

Les bases de données espèce-spécifique sont dédiées à une espèce ou un taxon particulier. Elles sont développées par des groupes de chercheurs spécialisés dans l'étude de ces organismes. On y trouve généralement le regroupement de toutes les informations essentielles et exhaustives des protéines et des liens vers les autres types de base que nous allons voir par la suite. Pour exemple, la base dédiée aux levures est la *Saccharomyces Genome Database* (SGD [8]) (Cherry *et al.*, 1998), celle pour les espèces plasmodiales est PlasmoDB [9] (Bahl *et al.*, 2003) et celle pour *Arabidopsis thaliana* est TAIR [10] (Swarbreck *et al.*, 2008), *etc.*

Les bases de données universelles : elles visent à regrouper l'ensemble des séquences primaires connues et leurs informations. L'une des plus populaires est maintenue par le consortium UniProt (Apweiler *et al.*, 2004) : collaboration entre l'*European Bioinformatics Institute* (EBI), le *Swiss Institute of Bioinformatics* (SIB) et la *Protein Information Ressource* (PIR). Chaque membre du consortium s'est profondément impliqué dans la maintenance de bases de données (expertise des annotations automatiques, intégration et annotation de nouvelles séquences), le développement et le support de logiciels. Jusqu'en 2002, la PIR produisait la *Protein Sequence Database* (PIR-PSD) tandis que l'EBI et le SIB géraient Swiss-Prot (Bairoch et Apweiler, 1996) et son complément TrEMBL (pour *Translated EMBL Nucleotide Sequence Data Library*) créée à l'origine car les données de séquences étaient générées à un rythme excédant la capacité de Swiss-Prot à annoter correctement ces protéines. L'expertise et les ressources de ces deux sources de données

ont depuis été mises en commun pour former le consortium Uniprot. Cette base donne accès à plus de 9 millions de séquences protéiques (version 15.6 datant de Juillet 2009) dont 95% dans TrEMBL, c.-à-d. ne bénéficiant que d'annotations automatiques (chiffres extraits du site Web d'UniProt [11]). Parmi les autres grandes bases actuelles de séquences protéiques, on trouve notamment GenBank (Benson *et al.*, 2009), ENSEMBL (Hubbard *et al.*, 2009), l'UCSC *Genome Browser* (Kuhn *et al.*, 2009), RefSeq (Pruitt *et al.*, 2007) et Integr8 (Kersey *et al.*, 2005).

b) Données de structures 3D : La *Protein Data Bank* (PDB) (Berman *et al.*, 2000) est la base de données dépositaire de l'information sur la structure 3D de nombreuses molécules biologiques, acides nucléiques ou protéines. On y trouve aujourd'hui,

8. http ://www.yeastgenome.org/
9. http ://plasmodb.org/plasmo/
10. http ://www.arabidopsis.org/
11. http ://www.ebi.uniprot.org/index.shtml

la structure de 54 749 protéines (dont 47 495 déduites par l'utilisation de rayons X et 6 954 par RMN — chiffres extraits du site Web de la PDB [12] le 4 Août 2009). À titre de comparaison, à la fin de ma première année de thèse (Septembre 2006) ces chiffres étaient respectivement de 35 909 protéines, 30 843 déduites par rayons X et 4 897 par RMN. De nombreux efforts sont donc réalisés pour augmenter le nombre de structures connues car elles constituent des informations primordiales pour comprendre la fonction des protéines.

c) Données fonctionnelles : On trouve de nombreuses bases fournissant de précieuses informations sur la fonction des protéines. Par exemple, des bases décrivant des voies métaboliques ou *pathways* (comme KEGG (Kanehisa et Goto, 2000), metaCyc (Caspi *et al.*, 2008)), ou donnant accès à des données d'interactome ou de transcriptome (puces à ADN), *etc.* Un grand nombre de bases propose également une représentation des informations fonctionnelles sous forme d'*ontologie*, c.-à-d. un vocabulaire synthétique et structuré (des fonctions les plus générales aux plus précises ordonnées par des liens de filiation). On peut citer par exemple la *Gene Ontology* (Ashburner *et al.*, 2000) dont l'objectif est de classifier les fonctions moléculaires, les processus biologiques et la localisation cellulaire des produits de gènes. Les termes de la *Gene Ontology* constituent une part importante de l'annotation automatique générée par la découverte de domaines protéiques, et sont donc largement utilisés lors de cette thèse, notamment dans le cadre de notre méthode de détection de nouveaux domaines par co-occurrence (*cf.* Chapitre 4 — sections 4.4.3, 4.5 et 4.6)

1.4.2 Regroupement en familles par séquences primaires

Un certain nombre de bases de données proposent une classification des protéines en familles en s'appuyant sur leur séquence primaire. C'est notamment le cas de COG (Tatusov *et al.*, 1997, 2003), ProtoMap (Yona *et al.*, 2000), Systers (Krause *et al.*, 2000), OrthoMCL (Li *et al.*, 2003), OMA (Schneider *et al.*, 2007), *etc.*, qui regroupent les protéines orthologues grâce à du *clustering* hiérarchique et/ou des approches phylogénétiques. Cependant ces bases ne proposent pas de modélisation de leurs familles, c'est pourquoi elles ne seront pas plus décrites ici. *A contrario*, les bases de données qui accompagne leur classification d'un modèle pour chaque famille protéique sont détaillées dans cette section. Les projets tels que Blocks Henikoff *et al.* (2000) ou Domo Gracy et Argos (1998a,b), ayant respectivement été abandonnés en 2007 et 2006, ne seront pas traités. Toutes ces bases disponibles actuellement font partie de la métabase Interpro (détaillée à la section 1.4.4) et sont présentés dans l'ordre chronologique de leur apparition. Une attention plus particulière sera portée à Pfam (section 1.4.2.c) dont nous affinons les détections de domaines dans nos travaux (Chapitre 4 et 5).

a) Prosite [13], la plus ancienne base de données de famille de modèles, dépend du SIB.

Créée en 1988 par Bairoch (1991), elle contenait alors 58 modèles de familles accompagnés d'une description fonctionnelle. Depuis, Prosite a continué à s'enrichir de modèles et d'annotations (Hulo *et al.*, 2008). À l'origine, cette base construisait des expressions régulières

12. http ://www.rcsb.org/pdb/home/home.do
13. http://www.expasy.org/prosite/

de domaines et de sites fonctionnels (motifs). Mais bien que les expressions régulières soient adaptées aux courtes régions conservées (typiquement site de fixation de groupes prosthétiques, d'ions métalliques, catalyse d'enzyme, *etc.*), le manque de flexibilité de ces modèles constitue une limitation forte pour la détection de domaines/motifs divergents. Prosite a alors développé des profils qui remplacent au fur et à mesure les expressions régulières. De plus, Prosite adjoint à ses modèles un ensemble des règles (ProRule) pour augmenter le pouvoir discriminant de ses modèles (Sigrist *et al.*, 2005). Prosite a été l'une des bases les plus utilisées et des plus référencées ; on peut citer par exemple (Bailey et Elkan, 1995; Jonassen *et al.*, 1995; Grundy *et al.*, 1997; Brazma *et al.*, 1998; Blekas *et al.*, 2005). Elle a notamment servi à la construction des matrices BLOSUM (Henikoff et Henikoff, 1992). En Août 2009, Prosite (version 20.52) dénombrait 1 560 documents d'annotation fonctionnelle, 1 308 expressions régulières, 862 profils et 868 règles ProRule. Chaque modèle est relié à un document d'annotation où l'utilisateur peut trouver différents types d'informations sur la famille : origine du nom, occurrences taxonomiques, architecture, fonction, structure 3D, *etc.* À l'heure actuelle, 53% des protéines de Swiss-Prot ont une référence croisée dans PROSITE (Hulo *et al.*, 2008).

b) Prints[14] (Attwood *et al.*, 1994, 2003) est la seconde plus ancienne base et, comme son nom l'indique, elle recueille des empreintes de protéines. Une empreinte est un groupe de motifs conservés qui est modélisé par une expression régulière. Les modèles sont appris par des parcours itératifs d'Uniprot (par PSI-BLAST (Altschul *et al.*, 1997)) pour caractériser les protéines par les motifs successifs qu'elles contiennent. Les motifs modélisés dans l'expression régulière ne doivent donc pas se chevaucher. Bien que séparés le long de la séquence, les résidus des motifs sont souvent contigus dans l'espace. Prints permet donc de modéliser des sites et des fonctions grâce à une approche plus flexible et complète que des motifs isolés ou indépendants, en exploitant l'information apportée par le contexte des motifs voisins.

c) Pfam[15] (Finn *et al.*, 2010), initialement développée pour l'annotation du génome de *Caernorhabditis elegans* (Sonnhammer *et al.*, 1997), est aujourd'hui une base de données de domaines protéiques incontournable. Elle a été l'objet de publications régulières concernant l'accroissement du nombre de ses modèles (*cf.* Fig. 1.4) et ses améliorations successives (visualisation d'arbres phylogénétiques, parcours taxonomique, architectures en domaines, logo des modèles, intégration des nouvelles informations de structure 3D, données d'interaction, *etc.*) (Sonnhammer *et al.*, 1998; Bateman *et al.*, 1999, 2000, 2002, 2004; Finn *et al.*, 2006, 2008, 2010).

Il existe deux niveaux de qualité dans la base Pfam : Pfam-A et Pfam-B. Pfam-B est un ensemble de familles de domaines inférées automatiquement sans l'intervention d'experts humains. Après avoir longtemps construit les modèles Pfam-B grâce à la classification ProDom (*cf.* paragraphe 1.4.2.d), Pfam utilise désormais la classification automatique ADDA (Heger et Holm, 2003). Les familles de Pfam-A consistent en un alignement expertisé ma-

14. http ://www.bioinf.manchester.ac.uk/dbbrowser/PRINTS/
15. http ://pfam.sanger.ac.uk/

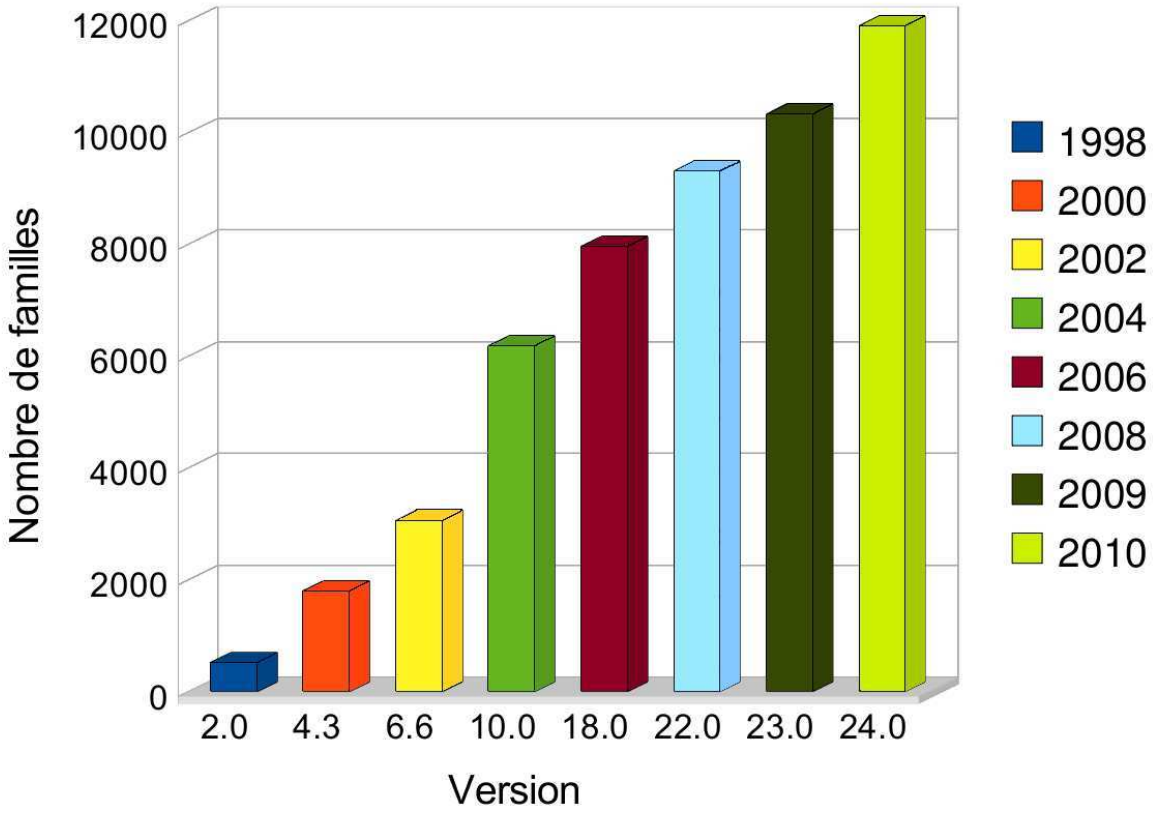


FIGURE 1.4 – **Évolution du nombre de HMM dans Pfam** d'après les publications (Sonnhammer *et al.*, 1998; Bateman *et al.*, 2000, 2002, 2004; Finn *et al.*, 2006, 2008, 2010) (Les HMM de Pfam-B ne sont pas comptabilisés). Les versions correspondantes de la base de données Pfam sont indiquées en abscisses, le nombre de modèles en ordonnées.

nuellement d'un sous-ensemble représentatif de la diversité des séquences du domaine (appelé graine), de deux HMM profils (pour la recherche de similarité globale et locale — voir section 2.4.2) appris sur l'alignement de la graine, et d'un alignement de l'ensemble des protéines d'Uniprot dont la séquence a été reconnue et certifiée comme appartenant à cette famille. Ces familles sont annotées et possèdent de nombreuses informations structurelles, fonctionnelles et évolutives. L'équivalence entre de nombreuses familles Pfam-A et des familles SCOP a été établie à plusieurs reprises (Elofsson et Sonnhammer, 1999; Zhang *et al.*, 2005). Les HMM de Pfam couvrent donc de nombreuses familles de protéines et de domaines : 72% des protéines d'UniProt et 95% des protéines de structure connue sont reconnues par au moins un domaine Pfam. La version 23.0 de Pfam propose dans Pfam-A une collection de 10 340 familles de domaines protéiques. Par défaut, quand on évoque les modèles de Pfam on fait généralement référence à Pfam-A uniquement, les familles Pfam-B étant rarement annotées et de qualité inférieure.

La dernière amélioration de Pfam est la création de *clans* : des ensembles de domaines partageant une origine évolutive commune. Pour le moment, les clan concernent principalement des protéines dont on connaît la structure (66% des protéines concernées). Quand aucune information de structure n'est connue, un clan peut être formé à partir d'indices forts comme des séquences de motifs communs. Pfam propose alors un alignement des séquences du clan, une description, des liens appropriés vers d'autres bases mais pas de modèle du clan.

d) ProDom [16] (Corpet *et al.*, 1998; Servant *et al.*, 2002; Bru *et al.*, 2005), initiée par une collaboration entre le CNRS et l'INRA de Toulouse datant de 1993, est aujourd'hui maintenue par le PRABI (Pôle Rhône-Alpes de BioInformatique). Cette base est une compilation de familles de domaines protéiques, générées de manière automatique par *clustering* de fragments/segments de séquences homologues. La procédure de classification de ProDom (MKDOM2 Gouzy *et al.* (1999)) s'appuie sur des recherches récursives par PSI-BLAST sur les séquences d'Uniprot. ProDom génère un alignement multiple des segments et une séquence consensus pour chaque famille. L'identification de domaines ProDom dans de nouvelles séquences se réalise avec le programme BLAST en comparant cette séquence requête, soit aux consensus des familles (plus rapide), soit à l'ensemble des séquences des familles (plus sensible). Cette base ne propose pas de modèles de familles de domaines mais un *mapping* de ses familles sur les protéines d'Uniprot. De plus, sa classification a longtemps joué un rôle important dans la base Pfam pour créer les modèles de Pfam-B (*cf.* paragraphe 1.4.2.c). C'est pourquoi ProDom possède des liens vers Pfam ainsi que vers Prosite et Interpro (*cf.* section 1.4.4). Enfin, ProDom a développé de nombreux liens entre ses familles et les structures 3D des bases de données PDB et SCOP (Bru *et al.*, 2005) (*cf.* paragraphe 1.4.3.b).

e) SMART [17] (*a Simple Modular Architecture Research Tool*) (Schultz *et al.*, 1998; Letunic *et al.*, 2009) s'intéresse plus particulièrement à l'analyse des architectures en domaines. SMART a développé une interface ergonomique axée sur la recherche de domaines et de combinaisons de domaines spécifiques dans des taxons choisis. Les domaines, issus de protéines de signalisation, de protéines extracellulaires et de protéines associées à la chromatine, sont précisément décrits grâce à leur distribution phylogénétique, leur annotation fonctionnelle, leur structure 3D et les résidus importants pour la fonction du domaine. Pour chacun de ces domaines, des alignements multiples et de nombreuses informations (voies métaboliques, données d'interactions, ontologie, *etc.*) sont disponibles et accompagnés d'une modélisation du domaine par un HMM profil. Chaque domaine, ainsi que les paramètres de recherche et les informations taxonomiques sont stockés dans la base. La dernière version de SMART contient 784 familles de domaines protéiques, expertisées manuellement à partir de 630 génomes complètement séquencés (55 eucaryotes, 46 archées and 529 bactéries).

f) PIRSF [18] (McGarvey *et al.*, 2000; Wu *et al.*, 2004) propose un système de classifica-

16. http ://prodom.prabi.fr/prodom/current/html/home.php
17. http ://SMART.embl-heidelberg.de
18. http ://pir.georgetown.edu/pirsf/

tion complet et non-recouvrant des séquences d'Uniprot sous forme de réseau basée sur les protéines complètes plutôt que sur les domaines qu'elles contiennent. PIRSF prétend ainsi permettre l'annotation des fonctions biochimiques générales et biologiques spécifiques, ainsi que la classification de protéines sans domaines bien définis. La raison de cette politique, plutôt à contre-courant des autres bases, est de permettre de corriger les erreurs d'annotation dues à des similarités locales de domaines en tenant compte de la similarité complète des protéines et de leur histoire évolutive. La classification PIRSF se compose de trois niveaux hiérarchiques dont le niveau central est la *famille homéomorphique*. Les membres d'une famille homéomorphique doivent être à la fois homologues (ayant évolué d'un ancêtre commun) et homéomorphes (partageant leur similarité de séquence sur toute la longueur de la protéine et possédant la même architecture en domaines). Le niveau inférieur est la *sous-famille* qui distingue les spécialisations de la fonction et/ou les variations existantes dans l'architecture en domaines d'une famille homéomorphique. Le niveau supérieur est la *super-famille* qui regroupe des familles homéomorphiques dont la relation évolutive est plus lointaine, ainsi que les protéines orphelines en s'appuyant sur leurs domaines communs. Pour chaque famille homéomorphique, un sous-ensemble représentatif de ses membres (séquences-graines) est choisi afin de générer un alignement multiple utilisé pour la création d'un arbre phylogénétique et d'un HMM profil. L'identification des domaines PIRSF dans les protéines d'Uniprot s'accompagne de nombreuses annotations fonctionnelles et structurelles. Dans une publication récente, Nikolskaya *et al.* (2007) mettent en avant la possibilité d'utiliser les données de PIRSF pour étudier la conservation/spécialisation de fonction dans les protéines multidomaines, ainsi que la convergence de fonction dans des protéines évolutivement non reliées et la divergence de fonction dans des protéines homologues.

g) TIGRFAMs[19] (Haft *et al.*, 2003; Selengut *et al.*, 2007) est développée par *The Institute of Genome Research*[20] (TIGR). Son objectif est de permettre l'annotation de nouvelles séquences par homologie en s'appuyant sur des données expertisées et sélectionnées à la main (lors de la création des familles). Pour cela, TIGRFAMs introduit le concept d'*équivalogue* : des équivalogues sont des protéines homologues dont la fonction a pu être conservée depuis leur ancêtre commun. La librairie compte un nombre réduit de modèles mais dont la fonction moléculaire est établie, ce qui en fait une ressource précise pour l'annotation automatique des protéines issues des nombreux projets de séquençage. TIGRFAMs regroupe donc les protéines équivalogues en familles lorsque c'est possible, complète par des sous- et super-familles le cas échéant, et propose de parcourir sa classification suivant la fonction recherchée. Les familles TIGRFAMs sont caractérisées par des alignements multiples, des modèles de Markov cachés (HMM profils) et des annotations fonctionnelles.

h) PANTHER[21] (*Protein ANalysis THrough Evolutionary Relationships*) (Thomas *et al.*, 2003; Mi *et al.*, 2005, 2007) est une ressource visant à classifier les protéines suivant leur fonction. À cette fin, PANTHER s'appuie sur les résultats d'expérimentations

19. http ://www.jcvi.org/cms/research/projects/tigrfams/
20. fondé en 1992 par J. Craig Venter et devenu une division du *J. Craig Venter Institute* fin 2006
21. http ://www.pantherdb.org/

scientifiques publiés ou sur des relations évolutives en l'absence de preuves expérimentales directes. Les protéines sont classées en famille puis divisées en sous-familles de fonction similaire par expertise manuelle de biologistes. Pour chaque famille de séquences, un arbre phylogénétique est calculé et chacun de ses nœuds (ancêtres) est annoté par une fonction si celle-ci a été conservée par ses descendants. Les sous-familles correspondent aux nœuds annotés et sont caractérisées par leur fonction moléculaire et les processus biologiques auxquels elles participent (*via* la GO). PANTHER génère alors pour chaque sous-famille un alignement multiple de ses séquences et construit un HMM profil. De plus, PANTHER capture les interactions biologiques précises grâce à une ontologie de voies métaboliques, nommée *PANTHER Pathway*, qui peuvent être visualisées de manière interactive. À l'origine axée sur les mammifères (9 espèces), PANTHER a récemment élargi sa sélection de séquences avec une plus grande variété d'espèces (44 espèces de vertébrés, invertébrés, plantes, champignons, protistes et procaryotes) pour lesquelles elle a pu intégrer les annotations fonctionnelles issues des bases de données spécialisées dans chaque organisme. Cette plus grande diversité a notamment permis d'affiner la construction des arbres phylogénétiques et l'identification des sous-familles.

i) HAMAP[22] (Lima *et al.*, 2009) (*High-quality Automated and Manual Annotation of microbial Proteomes*) est une collection de familles de protéines microbiales orthologues créées par expertise manuelle à partir des données de Swiss-Prot. Les développeurs génèrent des profils à partir d'alignements multiples automatiques expertisés manuellement, pour identifier les familles et sous-familles de nouvelles protéines bactériales, archaeales ou codées par des génomes plastidiaux (par exemple des chloroplastes, cyanelles, apicoplastes et plastides non-photosynthétiques).

1.4.3 Regroupement en familles par structure 3D

Certaines bases de données ne s'appuient pas sur la structure primaire pour établir des familles, mais sur la structure 3D des protéines. Cependant, le nombre de séquences pour lesquelles la structure 3D a été résolue ($\sim$ 60 000) est nettement plus faible que celui des séquences primaires connues ($\sim$ 12 000 000). En effet, comme vu précédemment (*cf.* section 1.1.3), déterminer la structure d'une protéine nécessite des expériences biologiques complexes telles que la cristallographie par rayons X ou la spectroscopie par RMN. Il existe plusieurs bases de données qui utilisent les données structurelles de la PDB afin d'établir une classification des protéines en familles. Nous présentons ici les trois plus référencées dans la littérature : FSSP (section 1.4.3.a), SCOP (section 1.4.3.b) et CATH (section 1.4.3.c). Une étude comparative des classifications proposées par ces trois bases recommande de bien comprendre les règles sur lesquelles elles sont bâties pour en faire la meilleure utilisation possible (Hadley et Jones, 1999). S'appuyant sur les classifications de SCOP et de CATH, deux bases de données proposant des modèles mathématiques de familles de protéines basées sur les structures 3D ont vu le jour. Ces bases, nommées SUPERFAMILY et PANTHER, sont détaillées dans les sections 1.4.3.d et 1.4.3.e respectivement.

22. http ://www.expasy.ch/sprot/hamap/

a) FSSP (Holm et Sander, 1994), pour *Fold classification based on Structure-Structure alignment of Proteins*, propose un *clustering* hiérarchique des séquences de la PDB s'appuyant sur la comparaison des ensembles de coordonnées spatiales et le calcul de Z-scores. Un outil supplémentaire adjoint à FSSP, *Dali Domain Dictionary*, permet, étant donné un ensemble de coordonnées donné (requête), de rechercher des structures similaires (parmi les familles FSSP). Aujourd'hui, le serveur Dali [23] est toujours en activité mais la classification FSSP a disparu (aucune publication depuis (Holm et Sander, 1998)). Elle a été remplacée par PDB90, l'ensemble des entrées non-redondantes de la PDB ayant moins de 90% d'identité, utilisé comme référence lors des recherches de similarités (Holm *et al.*, 2008).

b) SCOP [24] (Murzin *et al.*, 1995), pour *Structural Classification of Proteins*, fournit une description détaillée des structures et des relations évolutives des protéines. SCOP regroupe les séquences de structures connues en familles, en se basant sur les liens évolutifs et les principes gouvernant le repliement tridimensionnel des protéines (informations issues de la littérature sur les repliements 3D bien documentés). La méthode pour construire une classification en familles s'appuie essentiellement sur des méthodes de comparaison automatique pour les séquences les plus proches et sur l'inspection manuelle à l'aide d'outils de visualisation pour les séquences les plus distantes. La classification de SCOP distingue trois principaux niveaux hiérarchiques :

- *la famille* regroupe les protéines partageant au moins 30% d'identité de leur séquence primaire, et celles possédant une similarité plus faible mais dont la fonction et la structure sont très proches (par exemple les globines qui ne partagent que 15% d'identité) ;
- *la super-famille* réunit les protéines ayant une plus faible similarité de séquences primaires mais dont la structure et des éléments de fonction suggèrent qu'une origine évolutive commune est probable (par exemple les actines) ;
- *le repliement* ou *fold* associe les protéines possédant approximativement les mêmes structures secondaires majeures, avec des arrangements et connections topologiques semblables.

SCOP comptait à sa création, en 1995, 498 familles, 366 super-familles et 278 types de repliements, pour classer les 3 179 entrées (structures 3D résolues) de la PDB. Actuellement, on dénombre 3 902 familles, 1 962 super-familles et 1 195 types de repliements dans la version 1.75 (Février 2009) pour les 38221 entrées de la PDB (chiffres extraits des statistiques en ligne de SCOP [25]).

c) CATH [26] (Orengo *et al.*, 1997; Cuff *et al.*, 2009), pour *Class, Architecture, Topology and Homology*, filtre les entrées de la PDB pour ne considérer que les structures issues de RMN et les structures de cristallographie ayant une résolution inférieure à 4.0 Ångströms. CATH construit sa classification sur les domaines protéiques individuels, par une combinaison

23. http ://ekhidna.biocenter.helsinki.fi/dali_server
24. http ://scop.mrc-lmb.cam.ac.uk/scop/
25. http ://scop.mrc-lmb.cam.ac.uk/scop/count.html
26. http ://www.cathdb.info/

de procédures automatiques et manuelles. En premier lieu, CATH assigne à toute séquence, les bornes de ses domaines. Pour cela, il procède :

- soit manuellement, en se basant sur les résultats d'une grande variété d'algorithmes de comparaisons (méthodes basées sur la structure 3D : CATHEDRAL (Redfern *et al.*, 2007), SSAP (Orengo et Taylor, 1996), *etc.* ; et les séquences primaires : HMM profils) et la littérature ;
- soit de manière automatique, s'il existe une séquence similaire (plus de 80% d'identité et un score SSAP>=80) dont les domaines ont été identifiés. Les bornes des domaines de la protéine connue sont alors transférés à la nouvelle séquence.

Ensuite, CATH propose une classification à 5 niveaux, des structures 3D connues de protéines :

- *la super-famille* (*homology*) regroupe les structures ayant une forte similarité de séquence primaire et de structure 3D, combinée avec une fonction moléculaire semblable. Ceci autorise à penser que les domaines partagent une histoire évolutive (issus d'un domaine ancestral commun) et doivent donc appartenir à une même super-famille de séquences homologues.
- *la famille* est le niveau introduit pour distinguer les séquences les plus proches et par conséquent les fonctions les plus spécialisées au sein des super-familles. Ce niveau propose donc de diviser chaque super-famille d'homologues en une ou plusieurs familles en fonction d'un pourcentage d'identité des séquences primaires
- *la topologie* (repliement ou *fold*) spécifie la connectivité séquentielle des structures secondaires au cœur du domaine.
- *l'architecture* décrit la forme globale de la structure déterminée par l'orientation spatiale des structures secondaires (par exemple les barils ou les sandwichs) mais en ignorant les connections entre les structures secondaires.
- *la classe* est le niveau le plus bas qui regroupe les domaines possédant la même composition en structures secondaires. Les trois classes principales sont "essentiellement-alpha", "essentiellement-beta" et "alpha-beta". Une quatrième classe comprend les structures ayant peu de structures secondaires.

La procédure pour classifier un nouveau domaine dans les différents niveaux hiérarchiques, consiste à s'appuyer sur l'existence d'un domaine similaire pré-classifié (plus de 35% d'identité et SSAP>=80). S'il en existe un, la classification du nouveau domaine est alors héritée de celle du domaine précédemment étudié. Autrement, le domaine est classifié manuellement en se basant sur l'analyse des résultats d'algorithmes de comparaisons (CATHEDRAL, SSAP, HMM) et la littérature associée à la séquence étudiée. Lors de sa création en 1997, CATH comptait 8078 domaines et distinguait 1 068 familles (S35), 645 super-familles d'homologues, 505 topologies et 31 architectures. La dernière version (version 3.2 datant de Juillet 2008), contient 114 215 domaines repartis en 8 871 familles (S35), 2 178 super-familles d'homologues, 1 110 topologies et 40 architectures.

d) SUPERFAMILY [27] (Gough et Chothia, 2002; Wilson *et al.*, 2009) s'appuie sur la classification SCOP des protéines super-familles. SUPERFAMILY a été utilisé pour identifier

27. http ://supfam.cs.bris.ac.uk/SUPERFAMILY/

les structures 3D de l'ensemble des génomes complètement séquencés et les protéines d'UNIPROT. Cette base propose différents services en ligne s'appuyant sur sa librairie de modèles telle que l'identification des super-familles SCOP dans une séquence requête ; ou l'alignement multiple de séquences sur les modèles (c.-à-d. alignement d'une séquence requête avec un ensemble de séquences de structures 3D connues). SUPERFAMILY possède aussi de nombreuses informations de génomique comparative (sur- et sous-représentation des domaines dans une espèce par rapport à d'autres), d'architectures en domaines (graphes représentant les réseaux de paires de domaines adjacents), *etc.* Le niveau inférieur de la classification SCOP (famille) a également été intégré dans la dernière version, avec la description fonctionnelle des domaines Interpro, les annotations GO, la possibilité de visualiser la distribution taxonomique, *etc.*

e) Gene3D [28] (Buchan *et al.*, 2003; Yeats *et al.*, 2008) est une librairie de HMM profils qui correspond aux super-familles de CATH. L'objectif de cette base est de prédire des annotations fonctionnelles pour l'ensemble de protéines d'UNIPROT, RefSeq et Integr8 en identifiant les structures homologues et en mettant à disposition de nombreuses informations fonctionnelles (annotations GO, voies métaboliques KEGG, liens vers des données d'expressions, classification COG, *etc.*).

1.4.4 La *méta*-base de données InterPro

InterPro [29] (Apweiler *et al.*, 2001; Mulder *et al.*, 2003, 2007; Hunter *et al.*, 2009) est une base de données maintenue par l'EBI (membre du consortium Uniprot dont les données servent de référence à Interpro) visant à unifier les informations issues des différentes bases de données de modèles de familles protéiques.

Chacune des bases participant à InterPro s'appuie sur son propre schéma de l'univers des protéines (bornes des domaines, classification, *etc.*) et possède souvent ses propres outils. InterPro s'applique donc à intégrer l'ensemble des schémas en créant ses propres familles, ou "entrées InterPro" (*Interpro entries*). Les entrées regroupent les domaines équivalents des différentes bases et leurs annotations. Interpro présente les liens évolutifs et fonctionnels entre les différentes entrées par des relations hiérarchiques, et propose des statistiques concernant les recouvrements (en terme de protéines et d'acides aminés) entre ces entrées. Le site Web d'Interpro offrent donc un portail d'accès à toutes ces informations ainsi qu'à des références bibliographiques, des annotations fonctionnelles, la couverture taxonomique et des liens vers les différentes bases du consortium et vers les bases de données d'interaction, d'expression, de structure, *etc.* Une des informations produites par Interpro est l'annotation GO de l'ensemble de ses entrées. Ces annotations sont alors propagées aux familles des différentes bases. La politique d'annotation d'Interpro est qu'un terme GO est attribué à une entrée, si l'ensemble des séquences reconnaissant cette entrée sont annotées par ce terme GO (Mulder *et al.*, 2003). Ainsi, l'identification d'un domaine dans une protéine permet de transférer l'annotation GO du domaine à la protéine.

28. http ://gene3d.biochem.ucl.ac.uk/Gene3D/mainSearch
29. http ://www.ebi.ac.uk/interpro/

Les bases de données appartenant au consortium Interpro ont été décrites en détail précédemment et sont au nombre de onze : Gene3D, SUPERFAMILY, PANTHER, ProDom, TIGRFAMs, PROSITE, Prints, PIRSF, SMART, HAMAP et Pfam. Interpro ne propose pas de modèles pour ses entrées, mais elle s'est dotée d'un programme nommé InterProScan [30] (Zdobnov et Apweiler, 2001), qui combine en une seule ressource les méthodes de reconnaissances de modèles utilisées par les différentes bases du consortium. Cependant, si InterProScan permet une recherche de l'intégralité des domaines des bases participant au consortium Interpro, toutes les familles de ces bases ne sont pas intégrées dans des entrées Interpro. La composition d'Interpro en terme de modèles et le nombre de protéines d'Uniprot reconnaissant au moins l'un de ces modèles, sont donnés respectivement dans les tableaux 1.2 et 1.3 (dont les chiffres sont extraits du site Web d'Interpro [31]).

Cette méta-base de données constitue donc la principale ressource lorsque l'on cherche à identifier les domaines composant une protéine, à accéder à l'ensemble des informations concernant ces domaines, et à annoter de manière automatique la fonction de protéines récemment séquencées. Notre méthode de détection de domaines par co-occurrence (*cf.* Chapitre 4) s'appuie sur l'ensemble des domaines Interpro connus pour découvrir de nouveaux domaines Pfam.

Base de données de modèles	Version	Nombre de modèles	Modèles intégrés à Interpro
HAMAP	28/05/09	1 633	280
PANTHER	6.1	30 127	2 135
Pfam	23.0	10 340	10 336
PIRSF	2.70	3 212	2 691
PRINTS	39.0	1 950	1 928
ProDom	2006.1	1 894	834
PROSITE patterns	20.35	1 316	1 315
PROSITE profiles	20.35	801	779
SMART	5.1	724	720
TIGRFAMs	8.0	3 603	3 581
GENE3D	3.0.0	2 147	1 024
SUPERFAMILY	1.69	1 538	1 090

TABLE 1.2 – **Composition d'Interpro version 19.0 : nombre de modèles intégrés provenant de chaque base**.

30. http ://www.ebi.ac.uk/Tools/InterProScan/
31. http ://www.ebi.ac.uk/interpro/release_notes.html

Base de données de séquences	Version	Nb de séquences	Pourcentage de séquences ayant une correspondance avec :	
			un domaine de l'une des bases d'Interpro	un domaine intégré à une entrée Interpro
UniProtKB	15.6	9 421 896	79.8%	76.3%
UniProt/Swiss-Prot	57.6	495 880	96.5%	94.6%
UniProt/TrEMBL	40.6	8 926 016	78.9%	75.3%

TABLE 1.3 – **Nombre de protéines d'Uniprot ayant un domaine appartenant à l'une des bases d'Interpro, intégré ou non**.

Chapitre 2

Modélisation de familles de protéines

La modélisation permet de dresser un “portrait robot” d'une famille de séquences, en capturant l'information spécifique commune aux protéines de cette famille. Le modèle est alors utilisé pour identifier de nouvelles séquences homologues grâce à leur similarité au “portrait robot”.

Quelles sont les informations spécifiques d'une famille de séquences et comment les identifier ? Ce qui permet de caractériser une famille de séquences, ce sont le degré de conservation et la nature des acides aminés à certaines positions clefs des séquences. Pour identifier ces positions, on s'appuie généralement sur un alignement multiple des séquences de la famille. Les séquences partageant une histoire évolutive commune, l'alignement multiple fait apparaître les positions soumises, ou non, à une pression de sélection et les propriétés physico-chimiques contraintes à ces positions. L'enjeu de la modélisation consiste alors à représenter au mieux, non seulement les positions conservées et les variation acceptables au sein d'une famille, mais également les particularités observables de certaines séquences telles que les insertions ou délétions d'un ou plusieurs acides aminés consécutifs.

Une fois le modèle construit, étant donné une nouvelle séquence, on utilise le modèle pour évaluer la correspondance entre la séquence et la famille. Si la ressemblance est effective, on peut intégrer la séquence à la famille. Il s'agit bien souvent d'un processus itératif : les séquences ajoutées à une famille peuvent permettre la définition d'un meilleur modèle qui servira à son tour à détecter de nouvelles séquences. La performance des approches de modélisation s'est donc renforcée grâce à l'accroissement massif des données protéiques issues des projets de séquençage de génomes complets.

Parmi les représentations mathématiques qui ont été utilisées pour modéliser des familles de protéines on trouve notamment :

- les arbres de suffixes probabilistes,
- les réseaux neuronaux,
- les expressions régulières,
- les profils ou *Position-Specific Scoring Matrices* (PSSM),
- les modèles de Markov cachés (HMM) et les HMM profils.

Les arbres de suffixes utilisés dans (Ron *et al.*, 1996; Sagot et Marsan, 2000), et les réseaux neuronaux (Blekas *et al.*, 2005) n'étant pas à notre connaissance présents dans des bases de données majeures, ils ne seront pas plus détaillés ici. Nous allons présenter, dans un premier temps, les expressions régulières (section 2.1) et les PSSM (section 2.2), qui sont historiquement les premiers modèles à avoir été utilisés pour modéliser les familles de séquences et sont encore présents dans la base de données Prosite. Puis nous portons une attention particulière aux modèles de Markov cachés (section 2.3) et à l'adaptation de ces modèles pour l'étude des séquences biologiques : les HMM profils (section 2.4). Ces derniers, dont l'utilisation est largement répandue dans les bases de données actuelles de famille de séquences protéiques, sont présentés dans le cadre du logiciel HMMER.

2.1 Expressions Régulières

Les expressions régulières ou *patterns* sont des modèles plutôt dédiés aux motifs (*cf.* section 1.3.2) ou à des domaines courts et très conservés. Ces modèles sont issus de la théorie des langages formels (Kleene, 1951). Dans les années 80, l'utilisation d'expressions régulières pour la modélisation de motifs dans des séquences biologiques s'est popularisée (Abarbanel *et al.*, 1984; Bairoch et Claverie, 1988) et a conduit à la création de Prosite Bairoch (1991), la première base de données de modèles regroupant l'ensemble des motifs publiés par la communauté sous forme d'expressions régulières.

La modélisation d'une famille protéique par une expression régulière consiste à représenter le consensus de l'alignement des séquences en respectant une syntaxe prédéfinie. Cette syntaxe utilise les règles des expressions régulières universelles utilisées en informatique (par exemple, la commande `grep` sous Unix) ainsi que des règles plus spécifiques aux séquences biologiques (introduites par Prosite [1]) :

- Les différents acides aminés sont symbolisés par le code standard à une lettre IUPAC (*cf.* Table 1.1) ;
- Le symbole 'x' représente une position où n'importe quel acide aminé est autorisé ;
- Si plusieurs acides aminés sont possibles à une position, ils sont listés entre crochet ;
- Si plusieurs acides aminés ne sont pas envisageables à une position, ils sont listés entre accolades ;
- Chaque position est séparée de la suivante par un tiret '-' ;
- Un acide aminé X répété n fois de manière consécutive s'écrit X(n)
- Si une position est N- ou C-terminale, elle est précédée (resp. suivie) du symbole $<$ (resp. $>$).

Il existe quelques variantes et additifs à ces règles suivant la convention choisie, par exemple on peut représenter une position où l'ensemble des acides aminés hydrophobes serait possible, c.-à-d. [LIVMFYWC], par la lettre Z (*cf.* Table 2.1).

Pour illustrer ces règles syntaxiques, prenons comme exemple l'expression régulière suivante : C-x(2,4)-[DE]-x-{KH}. Ce modèle reconnaît toute séquence possédant un C suivit d'entre 2 et 4 symboles arbitraires puis un D ou un E, un symbole arbitraire et enfin un acide

1. pour plus de détails, voire http ://www.expasy.org/tools/scanprosite/scanprosite-doc.html#patter_syntax

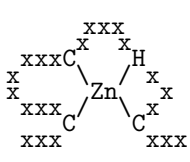
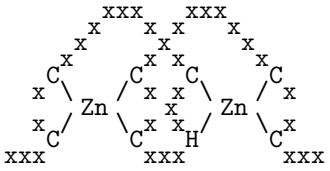
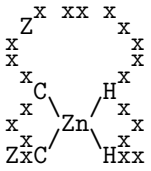
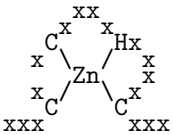
CCCH-type	C3HC4-type	C2H2-type	CCHC-type
C-x(8)-C-x(5)-C-x(3)-H	C-x(2)-C-x(9-39)-C-x(1-3)-H-x(2-3)-C-x(2)-C-x(4-48)-C-x(2)-C	Z-x-C-x(2,4)-C-x(3)-Z-x(8)-H-x(3,5)-H	C-x(2)-C-x(4)-H-x(4)-C
			

TABLE 2.1 – *Plusieurs familles de domaines zinc finger,* correspondant respectivement aux entrées Interpro IPR000571, IPR001841, IPR007087 et IPR001878. La première ligne du tableau indique le type de motifs. La seconde ligne correspond l'expression régulière du motif. La troisième ligne propose une représentation en deux dimensions de la capture d'éléments Zinc par le motif. Dans le motif C2H2, les positions avec un Z (qui représente l'ensemble des acides aminés hydrophobes [LIVMFYWC]) sont importantes pour la stabilité du repliement.

aminé différent de K ou H. Le tableau 2.1 présente quelques exemples d'expressions régulières pour plusieurs familles de motifs *zinc finger*.

Les expressions régulières sont des concepts mathématiques relativement faciles à appréhender. En effet, la lecture d'une expression régulière rend immédiatement compte des positions clés des séquences, et en particulier des propriétés physico-chimiques nécessaires à ces positions ainsi que leur degré de conservation. Elles se distinguent des modèles probabilistes par une réponse déterministe : une expression régulière reconnaît (réponse oui) ou ne reconnaît pas (réponse non) une séquence, il n'y a pas de probabilité associée entre le modèle et la donnée. Il est aussi à noter la grande difficulté pour modéliser d'éventuelles insertions ou délétions.

Le déterminisme de ces modèles rend parfois difficile la construction d'une expression régulière qui ne soit ni trop spécifique, ni trop générale (Hulo *et al.*, 2006). Par exemple, si à une position donnée, l'ensemble des séquences identifiées d'une famille ont soit un I soit un L, une traduction stricte serait [IL]. Toutefois, ce modèle rejette alors toute séquence ayant un V, qui est pourtant une évolution préférentielle de I et L (acides aminés aliphatiques). À l'inverse, une trop forte généralisation des positions où l'on manque d'information, revient à négliger la spécificité des séquences. Dans ces deux cas, on observe que les expressions régulières peuvent conduire à un grand nombre d'erreurs de prédictions (faux négatifs ou faux positifs). Il faut donc chercher un compromis lors de la construction du modèle entre : grande sélectivité/faible sensibilité (grande confiance dans les séquences prédites, mais beaucoup d'homologues réels ne sont pas trouvés) et faible sélectivité/grande sensibilité (où les homologues réels sont noyés parmi de nombreux faux positifs). Cependant, ce compromis n'est pas toujours possible, comme l'ont rapidement remarqué les concepteurs de Prosite pour les familles de globulines, d'immunoglubulines, SH2 et SH3 (Bairoch *et al.*, 1996), à cause de l'extrême divergence des séquences. C'est pourquoi, la base de données Prosite (*cf.* section 1.4.2), utilise non seulement des expressions régulières mais aussi des profils ou PSSM, que nous présentons dans la section

qui suit.

2.2 Profils — PSSM

Les profils ont été introduits par Taylor (1986b) et Gribskov *et al.* (1987), puis généralisés par Bucher et Bairoch (1994). Le terme profil est parfois remplacé par *Position-Specific Scoring Matrices* (PSSM) qui correspond plus précisément à ce que sont ces modèles : des matrices dont les scores sont spécifiques à chacune des positions de l'alignement multiple de la familles et s'accompagnent de pénalités spécifiques pour les insertions et les délétions. Les PSSM permettent donc de capturer les informations spécifiques d'une famille protéique : nature et degré de conservation des positions clefs, ainsi que la longueur et la position des délétions et des insertions conservées dans certaines espèces.

Le formalisme mathématique des profils, défini par Gribskov est le suivant : un profil est une structure composée d'états successifs qui correspondent aux positions les plus conservées de l'alignement multiple de la famille de séquences. On assimile souvent un profil à une matrice $\mathcal{M}$ de taille $(R+1) \times L$, où L est la longueur de l'alignement et R le nombre de résidus possibles — 20 pour les acides aminés — plus un élément dédié aux délétions et insertions. Chaque élément $\mathcal{M}_{ij}$ de cette matrice est un score qui reflète la probabilité d'observer, dans une nouvelle séquence, l'acide aminé i à la $j$ème position du modèle. Pour être plus précis, le score $\mathcal{M}_{ij}$ représente la similarité physico-chimique entre l'acide aminé i de la séquence étudiée et la distribution de probabilités sur les acides aminés observés à la position j dans l'alignement multiple de la famille protéique. Les scores sont donc calculés à l'aide d'une matrice de substitution d'acides aminés et de l'alignement multiple des séquences de la famille protéique. Le score d'un acide aminé i à une position j est défini ainsi :

$$\mathcal{M}_{ij} = \sum_{k=1}^{R-1} W_{kj} \times S_{ki},$$

où S_{ki} est le coût de substitution de l'acide aminé k par i d'après la matrice de substitution (telles que Dayhoff, PAM ou BLOSUM) et W_{kj} un poids pour l'apparition de l'acide aminé k à la position j de l'alignement. Une pondération simple consiste à prendre la distribution moyenne des acides aminés à chaque position de l'alignement : $W_{kj} = n(k,j)/N$ où $n(k,j)$ est le nombre d'apparitions de l'acide aminé k à la position j, et N le nombre d'acides aminés présents à cette position de l'alignement. Les pénalités des délétions/insertions sont elles calculées par rapport au nombre de résidus alignés aux positions concernées et à des coûts d'ouverture et d'extension fixés.

Lors de la comparaison d'une séquence avec un modèle, on calcule un score global en sommant les scores des acides aminés successifs de la séquence aux positions correspondantes du modèle. Ce score global reflète la ressemblance de la séquence au modèle. La comparaison du score à un seuil expertisé permet d'assigner ou non la séquence à la famille modélisée.

Les profils sont utilisés pour modéliser les familles de protéines de la base de données Prosite (*cf.* section 1.4.2), ainsi que pour la détection de motifs dans la méthode MEME (Bailey et Elkan, 1995; Grundy *et al.*, 1997; Bailey *et al.*, 2009).

2.3 Modèles de Markov cachés

Les modèles de Markov cachés (*Hidden Markov Model* ou HMM) sont des modèles probabilistes décrivant un processus probabiliste markovien et caché. Les HMM sont utilisés depuis les années 70 en reconnaissance de la parole et en traitement du langage (Baum *et al.*, 1970; Baker, 1975; Rabiner, 1989), et se sont imposés comme une méthode de modélisation de séquences de référence. Dès la fin des années 80, les HMM ont été utilisés pour l'étude de séquences d'ADN (Churchill, 1989). Nous détaillons dans cette section le formalisme général des HMM (section 2.3.1), ainsi que le cadre probabiliste auquel ils sont attachés (sections 2.3.2 et 2.3.3). Puis nous présentons les applications classiques pour lesquelles les HMM sont utilisés (section 2.3.4) et les algorithmes permettant de les résoudre (section 2.3.5). Enfin, nous abordons les problèmes liés à l'apprentissage de ces modèles (section 2.3.6) et le problème des probabilités nulles (section 2.3.7).

2.3.1 Qu'est-ce qu'un HMM ?

Un HMM se définit comme une structure composée d'un ensemble d'états, de transitions et de distributions de probabilités sur les transitions. De plus, on associe à chaque état générateur une distribution de probabilité sur les symboles d'un alphabet fini (appelées probabilités de génération). Ce type de modèle se différencie d'un automate probabiliste (Cassacuberta, 1990) où les symboles sont générés par les transitions, et où un unique symbole est attaché à chaque transition. Ces modèles sont toutefois fortement apparentés puisqu'on peut simuler tout HMM par un automate probabiliste de même taille (Abe et Warmuth, 1992), la réciproque n'étant pas vraie. Un HMM $\mathcal{H}$ peut être vu comme un quadruplet $(\mathcal{Q}, \mathcal{T}, \Sigma, \mathcal{G})$:

- $\mathcal{Q}$ est un ensemble d'états dont deux sont dits "muets" c'est à dire qu'ils ne génèrent aucun symbole et n'ont donc pas de probabilités de génération associées. Ces deux états sont appelés *Begin* et *End* qui servent respectivement à débuter et conclure une séquence.
- $\mathcal{T} : \mathcal{Q} - \{End\} \times \mathcal{Q} - \{Begin\} \rightarrow [0, 1]$, est l'ensemble des probabilités de transitions entre les états. On note $P(q \rightarrow q')$ la probabilité de transition de l'état q vers l'état q'. Pour chaque état q, on a une distribution de probabilités sur l'ensemble des états : $\sum_{q' \in Q} P(q \rightarrow q') = 1$. Dans la réalité, seules les transitions de probabilité non-nulle sont considérées et forment un *graphe de transitions* pondéré par la probabilité associée à chaque arête.
- Σ est un alphabet fini de symboles (par exemple les 20 acides aminés).
- $\mathcal{G} : \mathcal{Q} \times \Sigma \rightarrow [0, 1]$, est la matrice des probabilités de génération des symboles de Σ par chacun des états. On note $P(s|q)$ la probabilité de générer le symbole s dans l'état q. On a une distribution de probabilités sur les symboles dans chaque état q, c.-à-d. $\sum_{s \in \Sigma} P(s|q) = 1$.

La définition présentée ici n'est pas la définition originale donnée par Baum (Baum *et al.*, 1970), dont elle diffère par l'introduction des états muets *Begin* et *End*. Néanmoins, cette définition est celle habituellement utilisée dans la plupart des applications, dont la modélisation de séquences protéiques. On définit la *structure d'un HMM* comme l'ensemble de ses états,

son graphe de transitions et son alphabet.

2.3.2 Comment un HMM génère-t-il une séquence ?

Le processus de génération d'une séquence de symboles à l'aide d'un HMM consiste à débuter de l'état *Begin*, puis à se déplacer d'états en états en utilisant les probabilités de transition $\mathcal{T}$. Après chaque transition, la distribution de probabilités de génération $\mathcal{G}$ associée à l'état d'arrivée est utilisée pour générer un symbole. Le processus se termine lorsque l'on atteint l'état final *End*. On génère ainsi une séquence de symboles $\mathcal{S} = s_1 \dots s_L$, suivant une séquence d'états, ou *chemin* $\mathcal{C} = q_0 \dots q_{L+1}$ (où q_0 est l'état *Begin* et q_{L+1} l'état *End*). Un HMM définit donc un processus probabiliste non-déterministe, au sens où une même séquence de symboles peut être générée par plusieurs chemins différents. On comprend alors mieux le nom donné à ce modèle. Le processus de génération est un processus :

- ***markovien***, les probabilités de transition et de génération ne dépendent que de l'état actuel et non des états rencontrés précédemment,
- ***caché***, car il est impossible de connaître le processus suivi pour la génération d'une séquence de symboles.

2.3.3 Probabilités de génération d'une séquence $\mathcal{S}$ étudiée par un HMM $\mathcal{H}$ donné

Pour calculer la probabilité de générer la séquences de symboles $\mathcal{S} = s_1 \dots s_L$ à l'aide du HMM $\mathcal{H} = (\mathcal{Q}, \mathcal{T}, \Sigma, \mathcal{G})$, on doit calculer la probabilité de génération de $\mathcal{S}$ pour chaque chemin possible à travers $\mathcal{H}$, et faire la somme de ces probabilités. La probabilité d'avoir généré une séquence $\mathcal{S} = s_1 \dots s_L$ de longueur L par le chemin $\mathcal{C} = q_0 \dots q_{L+1}$ de longueur $L + 2$, où q_0 est l'état *Begin* et q_{L+1} l'état *End*, est définie ainsi :

$$P(\mathcal{S}, \mathcal{C}) = P(q_0 \to q_1) \prod_{l=1}^{L} P(s_i|q_i) P(q_i \to q_{i+1}). \tag{2.1}$$

Soit $\langle \mathcal{C}^{\mathcal{S}} \rangle$ l'ensemble des séquences d'états de $\mathcal{H}$ de longueur $L + 2$ permettant de générer la séquence $\mathcal{S}$. La probabilité de générer la séquence $\mathcal{S}$ avec le HMM $\mathcal{H}$, obtenue en sommant sur l'ensemble des séquences d'états possibles, est donc :

$$P(\mathcal{S}|\mathcal{H}) = \sum_{\mathcal{C} \in \langle \mathcal{C}^{\mathcal{S}} \rangle} P(\mathcal{S}, \mathcal{C}). \tag{2.2}$$

2.3.4 Pour résoudre quels problèmes ?

Les applications classiques des HMM se distinguent en deux catégories qui sont les problèmes de classification et les problèmes de segmentation.

Dans la première catégorie, on trouve, par exemple, les applications de reconnaissance d'un mot parmi un ensemble de mots possibles à partir d'un signal audio (Rabiner, 1989), d'une famille de protéines à partir d'une séquence d'acides aminés (Haussler *et al.*, 1993),

etc. Pour les problèmes de classification, on manipule généralement un ensemble ou *librairie* de HMM, un pour chaque classe à reconnaître, par exemple un HMM par famille de protéines homologues comme le proposent les bases de données de familles décrites précédemment (*cf.* section 1.4). La résolution des problèmes de classification consiste à calculer la probabilité de génération d'une séquence par chacun des HMM de la librairie, et à assigner à cette séquence sa famille la plus probable. Ce type d'application nécessite donc un algorithme efficace pour le calcul de la probabilité de génération d'une séquence, présenté dans la section suivante 2.3.5.

Dans la seconde catégorie, les problèmes de segmentation, on trouve des problèmes tels que le découpage d'un signal musical en notes (Raphael, 1999), la localisation de régions codantes/non-codantes dans une chaîne de nucléotides (Krogh *et al.*, 1994), *etc.* On utilise pour ces applications un HMM accompagné d'un ensemble fini d'étiquettes, chaque état du HMM étant associé à une étiquette. La procédure de segmentation d'un séquence consiste alors à calculer, à l'intérieur du HMM, le chemin qui a la probabilité maximale de générer cette séquence. On associe ensuite à chaque symbole de la séquence l'étiquette de l'état dans lequel il a été généré d'après le chemin de probabilité maximale. Ces applications reposent sur la recherche du chemin optimal parmi l'ensemble des chemins possibles, et nécessitent donc un algorithme efficace pour résoudre ce problème. Cet algorithme est présenté dans la section suivante 2.3.5.

2.3.5 Avec quels algorithmes ?

Comme vu précédemment, la résolution des applications de classification nécessite le calcul de la probabilité de génération d'une séquence pour l'ensemble des HMM d'une librairie. Ce type d'application nécessite donc un algorithme efficace pour le calcul de la probabilité de génération d'une séquence $\mathcal{S} = s_1 \dots s_L$ par un HMM $\mathcal{H} = (\mathcal{Q}, \mathcal{T}, \Sigma, \mathcal{G})$, notée $P(\mathcal{S}|\mathcal{H})$ (*cf.* équation 2.2). En effet, si N est le nombre d'états du HMM, alors le nombre de chemins possibles pour générer une séquence de longueur L est de l'ordre de N^L. Comme pour chaque chemin, le calcul de la formule 2.2 demande de l'ordre de L opérations, le calcul de la probabilité de génération de $\mathcal{S}$ par $\mathcal{H}$ suivant l'équation 2.2 serait donc en $\mathcal{O}(LN^L)$. En prenant des valeurs pour N et L correspondant à la longueur moyenne d'un domaine protéique, soit 200 acides aminés, le calcul de $P(\mathcal{S}|\mathcal{H})$ demanderait approximativement 10^{461} opérations. Heureusement une procédure bien plus efficace existe pour réaliser ce calcul : l'algorithme *forward-backward* (Rabiner, 1989). Le principe de cet algorithme est de considérer une variable *forward*,

$$\alpha_l(q) = P(s_1 \dots s_l, q_l = q|\mathcal{H}),$$

qui exprime la probabilité d'avoir généré la séquence $s_1 \dots s_l$ en partant de l'état *Begin* et d'être arrivé sur l'état q pour générer le $l^{\text{ème}}$ symbole. Cette variable peut être calculée de façon récursive, ce qui permet de proposer une version de l'algorithme de programmation dynamique *forward* (voir algorithme 1).

Algorithm 1: Algorithme *forward*

Données: $\mathcal{S} = s_1 \dots s_L$ une séquence ; $\mathcal{H}$ un HMM
Résultat: Renvoie la probabilité de génération de la séquence $\mathcal{S}$ par le HMM $\mathcal{H}$
pour chaque $q \in \mathcal{Q}$ **faire**
  $\alpha_1(q) = P(Begin \to q)P(s_1|q)$;
pour l *de* 2 *à* L **faire**
  pour chaque $q \in \mathcal{Q}$ **faire**
    $\alpha_l(q) = \left(\sum_{q' \in \mathcal{Q}} \alpha_{l-1}(q')P(q' \to q) \right) P(s_l|q)$;
return $P(\mathcal{S}|\mathcal{H}) = \sum_{q \in \mathcal{Q}} \alpha_L(q)P(q \to End)$;

La complexité de cet algorithme est de l'ordre de $\mathcal{O}(N^2 L)$. En reprenant les valeurs numériques vues précédemment ($L = N = 200$), le calcul de $P(\mathcal{S}|\mathcal{H})$ nécessite alors approximativement $10^{6.9}$ opérations. Cet algorithme est appelé *forward* car la récursion est réalisée en avant : on calcul tout d'abord la probabilité de générer le premier symbole de la séquence, puis à chaque étape de la récursion on rajoute un symbole et on réitère la procédure jusqu'au symbole terminal de la séquence. Un algorithme similaire, l'algorithme *backward*, permet de réaliser ce calcul à l'envers, en utilisant la variable *backward* :

$$\beta_l(q) = P(s_{l+1} \dots s_L, q_l = q|\mathcal{H}),$$

qui exprime la probabilité de générer la séquence $s_{l+1} \dots s_L$ en partant de l'état q et en arrivant sur l'état *End*. La procédure de récursion de l'algorithme *backward* est la suivante :

(1) Initialisation : $\forall q \in \mathcal{Q},\ \beta_L(q) = P(q \to End)$
(2) Récursion : Pour l allant de $L-1$ à 1, $\beta_l(q) = \sum_{q' \in \mathcal{Q}} P(q \to q')P(s_{l+1}|q')\beta_{l+1}(q')$

On obtient ainsi $P(\mathcal{S}|\mathcal{H}) = \sum_{q \in \mathcal{Q}} P(Begin \to q)P(s_1|q)\beta_1(q)$, par un calcul de complexité similaire à l'algorithme *forward* : $\mathcal{O}(N^2 L)$.

Pour les problèmes de segmentation, le problème consiste à trouver, étant donné une séquence de symboles $\mathcal{S} = s_1 \dots s_L$ et un HMM $\mathcal{H} = (\mathcal{Q}, \mathcal{T}, \Sigma, \mathcal{G})$, la séquence d'états du HMM qui a la probabilité maximale de générer $\mathcal{S}$. Ce qui nous préoccupe n'est pas la valeur de la probabilité maximale mais le chemin d'états — appelé *chemin de Viterbi* et noté $\mathcal{C}^*$ — qui permet de générer la séquence $\mathcal{S}$ avec cette probabilité. Ce type de problème possède les mêmes limites combinatoires que les problèmes de classification : nous avons besoin de calculer la probabilité de génération de la séquence suivant tous les chemins possibles avant de choisir celui qui a la probabilité la plus élevée. Ce calcul réalisé directement à partir de la formule 2.2, a une complexité de $\mathcal{O}(LN^L)$ et est donc difficilement applicable. C'est pourquoi on utilise un algorithme de programmation dynamique, appelé algorithme de Viterbi (Rabiner, 1989), très proche de l'algorithme *forward*, pour résoudre ces problèmes. Pour réaliser l'algorithme de Viterbi, on considère tout d'abord la variable $\delta_l(q)$ définie par :

$$\delta_l(q) = \max_{q_0 \cdots q_{l-1}} P(s_1 \dots s_l, q_l = q|\mathcal{H})$$

qui exprime la probabilité de générer la séquence $s_1 \dots s_l$ en suivant un unique chemin de probabilité maximale partant de l'état *Begin* et arrivant sur l'état q pour générer le $l^{\text{ème}}$ symbole. On peut donc calculer les $\delta_l(q)$ de manière récursive, pour obtenir la probabilité maximale de générer $\mathcal{S}$ avec $\mathcal{H}$ par le chemin de Viterbi $\mathcal{C}^*$ (*cf.* algorithme 2). Cependant ce n'est pas la valeur de cette probabilité qui nous intéresse mais la séquence d'états qui permet de générer $\mathcal{S}$ avec cette probabilité. On doit donc, à chaque étape l de la récursion et pour chaque état q, mémoriser l'état q' qui maximise l'arrivée en q, dans une variable $\psi_l(q)$. Une fois les variables $\delta_l(q)$ et $\psi_l(q)$ calculées pour chaque étape de la récursion et pour chaque état, il ne reste plus qu'à lancer une procédure de rétro-propagation pour "dérouler" le chemin de Viterbi — $\mathcal{C}^* = Begin, q_0^*, \dots, q_L^*, End$ —, en partant de l'état *End*. On notera que, mis à part la phase de rétro-propagation, l'algorithme de Viterbi est très similaire à l'algorithme *forward*. La principale différence résulte de la maximisation des probabilités attachées aux états précédents, au lieu du calcul de la somme de ces probabilités dans le cas de l'algorithme *forward*. La phase de rétro-propagation étant en $\mathcal{O}(L)$, la complexité de l'algorithme de Viterbi est donc la même que celle de l'algorithme *forward*, c'est à dire $\mathcal{O}(N^2 L)$.

Algorithm 2: Algorithme de Viterbi

Données: $\mathcal{S} = s_1 \dots s_L$ une séquence ; $\mathcal{H}$ un HMM
Résultat: Renvoie le chemin de Viterbi $\mathcal{C}^*$ de la séquence $\mathcal{S}$ par le HMM $\mathcal{H}$
pour chaque $q \in \mathcal{Q}$ **faire**
  $\delta_1(q) = P(Begin \to q) P(s_1|q)$;
  $\psi_1(q) = Begin$;
pour l *de* 2 *à* L **faire**
  pour chaque $q \in \mathcal{Q}$ **faire**
    $\delta_l(q) = \max_{q' \in \mathcal{Q}} \left(\delta_{l-1}(q') P(q' \to q)\right) P(s_l|q)$;
    $\psi_l(q) = \operatorname{argmax}_{q' \in \mathcal{Q}} \left(\delta_{l-1}(q') P(q' \to q)\right)$;
$s_L^* = \operatorname{argmax}_{q \in \mathcal{Q}} \left(\delta_L(q) P(q \to End)\right)$;
pour l *de* $L-1$ *à* 0 **faire**
  $s_l^* = \psi_l(s_{t+1}^*)$;
return $\mathcal{C}^* = Begin, s_1^*, \dots, s_L^*, End$;

2.3.6 Apprentissage des modèles

Nous venons de voir les algorithmes classiques utilisés pour résoudre des problèmes de classification et de segmentation grâce à des HMM. Ces algorithmes supposent que l'on dispose d'un HMM construit et paramétré de manière à modéliser de façon satisfaisante les séquences que l'on souhaite traiter. La question est donc de savoir comment construire un tel HMM.

Dans le cas le plus favorable, le HMM recherché peut être construit directement à partir des connaissances *a priori* dont on dispose sur les séquences. C'est notamment le cas de

Krogh *et al.* (1994) pour la modélisation de séquences d'ADN d'*Escherichia coli* dans le cadre d'un problème de segmentation. Ils exploitent un certain nombre de connaissances sur les données pour apprendre la structure du HMM et les distributions de probabilités sur les acides nucléiques en fonction de leur rôle/position. Ce genre de modélisation statistique est également utilisé pour la construction de HMM profil à partir d'un alignement multiple, par exemple par la base Pfam. En utilisant uniquement l'alignement des séquences expertisé à la main, la structure du modèle est déterminée et les distributions de probabilités associées aux états sont estimées.

Cependant, dans la plupart des applications, on ne dispose d'aucune connaissance *a priori* sur les données. Le HMM désiré doit être construit à l'aide d'un algorithme d'apprentissage automatique. On peut distinguer, dans le problème de l'apprentissage d'un HMM, deux cas de figure distincts, suivant que la structure est connue ou ne l'est pas.

Lorsque la structure est connue, le problème se réduit à un problème d'*entraînement* consistant à estimer les paramètres numériques — distributions de probabilités de transition et de génération — pour justifier au mieux la génération des séquences d'apprentissage. Ce problème, bien que NP-difficile (Abe et Warmuth, 1992), dispose d'heuristiques classiques telles que l'entraînement de Viterbi et l'entraînement de Baum-Welch, tous deux de complexité $\mathcal{O}(KN^2T)$ où N est le nombre d'états du HMM, T la taille totale des séquences d'apprentissage et K une constante bornant le nombre d'itérations de l'algorithme. L'entraînement de Baum-Welch est issu de la méthode générale d'*Expectation-Maximization* (EM) (Dempster *et al.*, 1977), servant à estimer les paramètres de nombreux modèles probabilistes, et vise à maximiser la vraisemblance des séquences d'apprentissage. L'entraînement de Viterbi est lui une adaptation de l'algorithme EM où la phase d'*Expectation* est remplacée par la recherche des chemins de Viterbi. Pour connaître les détails de ces algorithmes, on pourra consulter (Bréhélin, 2001).

Lorsque la structure est inconnue, le problème de l'apprentissage devient alors encore plus difficile. Il ne suffit plus de paramétrer une structure mais il faut également déduire cette structure des séquences d'apprentissage. Parmi les approches proposées, on peut notamment citer les approches :

- par généralisation ou fusion d'états : dont le principe est de construire un HMM équivalent à la disjonction des séquences (le plus spécifique) puis de le généraliser par des étapes successives de fusion d'états/de transitions (Stolcke et Omohundro, 1994) ;
- par spécialisation ou fission d'états : dont le but est de spécialiser un HMM très général en créant/scindant successivement des états/transitions (Takami et Sagayama, 1992).

Ces deux approches maximisent un critère qui est respectivement la probabilité *a posteriori* et le maximum de vraisemblance.

2.3.7 Le problème des probabilités de génération nulles

Ce problème survient lors de l'apprentissage des paramètres d'un HMM $\mathcal{H}$, lorsque la probabilité de génération d'un symbole $s \in \Sigma$ dans un état q est nulle, c.-à-d. $P(s|q) = 0$. Ce cas pose un problème crucial lors de l'utilisation du HMM sur de nouvelles séquences.

Tout chemin $\mathcal{C}$ passant par q pour générer s, entraîne une probabilité nulle de la séquence $\mathcal{S}$ entière par ce chemin : $P(\mathcal{S}, \mathcal{C}|\mathcal{H}) = 0$. Une solution pour résoudre ce problème est de *lisser* les distributions de probabilités de manière à éviter les probabilités nulles. La technique classiques, appelée *Laplace Smoothing*, consiste, pour chaque symbole, à ajouter une constante à son estimateur [2] une constante l :

$$\widehat{P}(s|q) = \frac{n_{s,q} + l}{|\Sigma| \cdot l + \sum_{s' \in \Sigma} n_{s',q}},$$

où $n_{s,q}$ est l'estimateur du symbole s dans l'état q. Il existe des techniques plus fines, en particulier pour le traitement de séquences biologiques où l'on dispose parfois de connaissance *a priori* sur les symboles de l'alphabet. Nous y revenons dans la section 2.4.2.

2.4 Un HMM dédié aux séquences biologiques : le HMM profil

Les publications d'Haussler *et al.* (1993) et Krogh *et al.* (1994) introduisent l'utilisation des "HMM profils", une spécialisation des HMM dédiée à l'étude des séquences biologiques. Leur utilisation s'inscrit dans le cadre de la modélisation de famille de séquences et la recherche d'homologues. Les HMM profils sont aujourd'hui les modèles de prédilection dans les bases de données de famille de séquences et ils sont devenus des outils standards en bioinformatique (Eddy, 1995; Durbin *et al.*, 1998).

Les deux principaux programmes permettant la manipulation de HMM profils pour l'analyse de séquences biologiques sont HMMER (Eddy, 1995, 1998) et SAM (*Sequence Alignment and Modeling system*) (Hughey et Krogh, 1996; Karplus *et al.*, 1998). L'utilisation de ces programmes s'est largement répandue au sein de la communauté. SAM a été employé pour la construction de la base SUPERFAMILY et HMMER pour celle de Pfam. Au cours de cette thèse nous avons principalement utilisé les modèles de domaines de la base Pfam. C'est pourquoi, après avoir présenté les particularités des HMM profils par rapport aux HMM généraux (section 2.4.1), nous développons dans cette section la manipulation de HMM profil à travers celle du logiciel HMMER (section 2.4.2). Enfin, nous concluons par une comparaison entre les deux programmes concurrents HMMER et SAM (section 2.4.3).

2.4.1 Structure d'un HMM profil

Les HMM profils ont pour objectif de modéliser une famille protéique, c.-à-d. un ensemble de séquences homologues. Nous avons vu en introduction de ce chapitre et dans les précédents modèles qu'un alignement de séquences est la donnée clef dans ce genre de modélisation. La puissance des HMM profils réside donc dans la représentation probabiliste des propriétés induites par l'alignement multiple : à quel point est conservée une colonne de l'alignement, quels résidus y sont préférés, et s'il y a une forte probabilité d'insertions ou de délétions.

2. les estimateurs sont les comptes du nombre observé de symboles à une position de l'alignement multiple, à partir desquels on déduit la distribution de probabilité : $P(s|q) = \frac{n_{s,q}}{\sum_{s' \in \Sigma} n_{s',q}}$.

La première propriété des alignements de séquences est la présence de blocs dits conservés, c.-à-d. un ensemble de positions consécutives où chaque position présente une forte conservation des résidus observés entre les différentes séquences. On s'intéresse donc dans un premier temps à la modélisation de la concaténation des blocs conservés de l'alignement. Cette concaténation de positions conservées est modélisé dans un HMM profil par une succession linéaire d'états appelés *Match*. Chaque état *Match*, M_p doit contenir la distribution de probabilité des acides aminés attendus à la $p^{\text{ème}}$ position conservée de l'alignement. On constate ici la proximité entre les PSSM et les HMM profils : une PSSM peut être représentée par une succession linéaire d'états *Match* séparés par des transitions de probabilités 1. Les PSSM dont l'utilisation a le même objectif que les HMM profils ne disposent cependant pas d'une structure entièrement probabiliste comme celle des HMM profils (héritée des HMM généraux), en particulier pour la représentation des insertions et des délétions.

En effet, les HMM profils offrent un cadre probabiliste permettant d'intégrer dans un modèle unique, la totalité des positions d'un alignement, c'est à dire non seulement les positions conservées mais également les positions d'insertions ou de délétions d'un ou plusieurs acides aminés. Les HMM profils permettent non seulement de représenter l'apparition de tels évènements avec des coûts dépendants de leur position dans la séquence, mais aussi de proposer des coûts d'extensions positions-spécifiques. Pour cela, pour chaque position p d'un HMM profil, l'état *Match* M_p est accompagné :

- d'un état dit *Insert* I_p, pour modéliser l'insertion éventuelle d'un ou plusieurs acides aminés, consécutivement à la $p^{\text{ème}}$ position conservée. Chaque état I_p dispose d'une distribution de probabilité pour générer les acides aminés que l'on s'attend à observer à cette position Cette distribution peut être déduite de l'alignement multiple ou des connaissances *a priori* que l'on a des insertions (voir section suivante). Le coût d'ouverture d'une insertion se traduit par une plus faible probabilité de transition de l'état *Match* M_p vers l'état *Insert* I_p que celle de M_p vers l'état *Match* suivant M_{p+1}. L'extension de cette insertion est permise par la boucle de l'état sur lui-même qui représente le coût d'extension. Enfin, une fois l'insertion terminée on oblige la modélisation à reprendre dans l'état *Match* suivant par une transition de I_p vers M_{p+1}.
- d'un état dit *Délétion* D_p, qui représente une possible délétion à une position conservée dans l'alignement. Les états délétions sont des états muets/silencieux, comme les états *Begin* et *End*. Ils ne possèdent pas de distribution de probabilités de génération sur les acides aminés et ne génèrent donc pas de symboles. L'état *Délétion* D_p autorise à contourner l'état *Match* M_p pour modéliser les séquences ayant subi une délétion d'un acide aminé à cette position. Le coût d'ouverture de la zone de délétion est traduit comme pour les états *Inserts via* la probabilité de transition de l'état *Match* M_p vers l'état *Délétion* D_p (plus faible que celle de M_p vers M_{p+1}). L'extension de la zone de délétion est permise par une transition de l'état *Délétion* D_p vers l'état *Délétion* suivant D_{p+1}. Le coût de cette extension s'exprime par une plus forte probabilité de rejoindre l'état *Match* suivant M_{p+1}, que d'aller dans l'état *Délétion* D_{p+1}.

Cette définition n'est pas exactement celle proposée par Haussler *et al.* (1993) et Krogh *et al.* (1994), dont elle diffère par l'absence de transitions entre les états *Délétions* et les états

Inserts. Elle correspond toutefois au cœur de la structure des HMM profils manipulés par le logiciel HMMER et présents dans la plupart des bases de données de domaines protéiques telles que Pfam.

La structure générale d'un HMM profil est représentée par la figure 2.1. Les HMM profils capturent donc l'information spécifique à chaque position d'un alignement multiple de séquences grâce à leur aspect séquentiel, à la présence de trois types d'états adaptés à la problématique de modélisation de séquences biologiques et à un cadre entièrement probabiliste. Tout cela a contribué à faire du HMM profil le modèle de prédilection lorsqu'il s'agit de représenter des familles de protéines afin d'identifier de nouvelles séquences homologues.

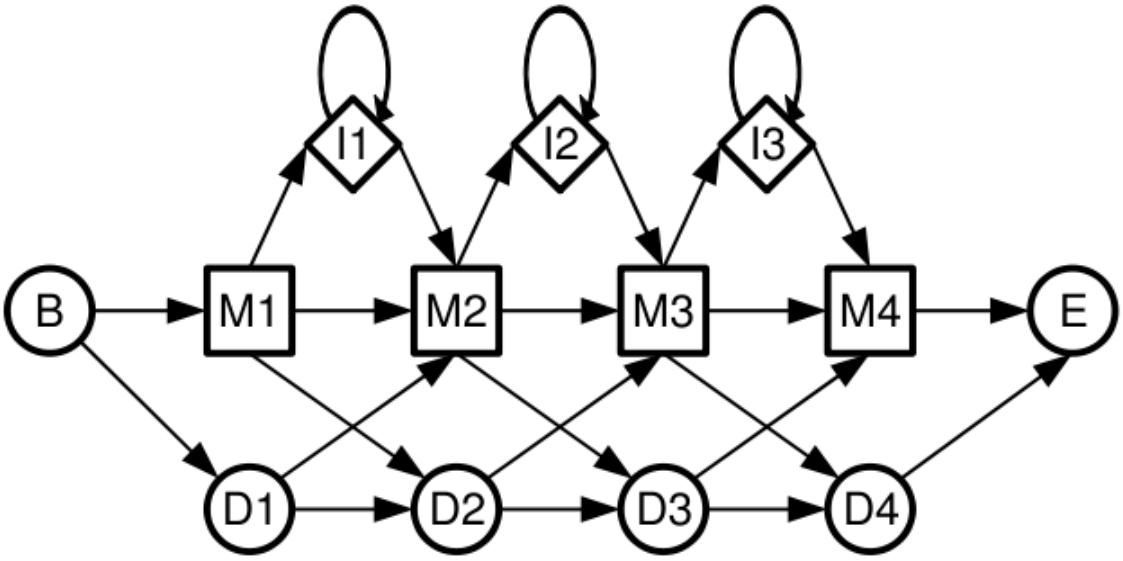


FIGURE 2.1 – **Structure des HMM profils :** l'état *Begin* (B), d'où commence tout(e) chemin (séquence). Puis on suit la succession linéaire des états *Matchs* (M). Chaque état *Match* est accompagné des états *Insert* (I) et *Délétion* (D) correspondants à l'information spécifique pour chaque position de l'alignement multiple des séquences de la famille. Enfin on finit toujours sur l'état *End* (E).

2.4.2 Le logiciel HMMER

Les modèles que nous avons choisis comme référence dans cette thèse sont les HMM profils de la base de données Pfam. Ces modèles sont accessibles librement en ligne, mais dans un format particulier, conçu pour être traité par le logiciel HMMER. Plusieurs bases de données d'Interpro utilisent le logiciel HMMER et proposent donc des HMM profils respectant ce format. Le logiciel HMMER permet de réaliser de nombreuses fonctionnalités standards liées à la manipulation de HMM profils. On peut notamment construire un HMM profil à partir d'un alignement multiple, aligner des séquences sur un HMM (problème de segmentation), générer des séquences correspondant aux probabilités du modèle, et surtout effectuer des recherches de séquences homologues en confrontant un ou plusieurs HMM à un

ensemble de séquences (problème de classification). Cette section n'a pas pour vocation d'être un guide pour l'utilisateur de ce logiciel (dont il existe une version exhaustive (Eddy, 2003)), mais de présenter brièvement le format informatique et la structure détaillée des modèles ainsi que les détails techniques importants pour appréhender la recherche de domaines avec le logiciel HMMER dans sa version 2.3.2.

a) Le format .hmm : Les fichiers des HMM profils manipulés par le logiciel HMMER peuvent être vus comme composés de deux parties : une entête comportant un ensemble d'informations sur le modèle et la famille de protéines qu'il représente ; et le corps du modèle c.-à-d. ses probabilités de génération et de transition. Un exemple de fichier .hmm est représenté dans la figure 2.2.

Dans l'entête, on trouve notamment la version de HMMER, un descriptif de la famille protéique, le nom et l'identification numérique du modèle. Sont ensuite indiqués le nombre de séquences dans l'alignement d'apprentissage, la longueur du modèle (nombre d'états *Matchs*), différents seuils de score pour la recherche de séquences homologues, le modèle *nul* (encodés par une formule avec un logarithme pour éviter les problèmes d'arrondis) sur lequel nous revenons ci-après, *etc.*

Dans le corps du modèle, à partir de la ligne débutant par "HMM" et listant les 20 acides aminés, on trouve les distributions de probabilités de génération des états *Matchs* et *Inserts*, ainsi que l'ensemble des transitions entre les états encodées par rapport au modèle nul pour éviter les problèmes d'arrondis. Cette partie du fichier est à lire par triplet de lignes pour chaque position p de l'alignement multiple d'apprentissage. Les trois lignes correspondent respectivement aux éléments suivant :

- distribution de probabilités de l'état *Match* M_p,
- distribution de probabilités de l'état *Insert* I_p,
- transitions entre les états (M_p, I_p, D_p) de la position p et les états (M_{p+1}, I_{p+1}, D_{p+1}) de la position suivante.

Le corps du modèle est donc composé de $3N + 1$ lignes : une ligne pour les transitions de départs (sortie de l'état *Begin*), suivie de N triplets de lignes où N est la longueur du modèle.

Enfin, la fin du fichier est marquée par la ligne "//". Ainsi, il est possible de concaténer dans un même fichier plusieurs HMM profils successivement, en respectant la description susmentionnée. C'est ainsi que Pfam, propose en libre accès un fichier contenant l'ensemble de ses modèles.

b) La structure Plan7 : La structure des modèles d'HMMER, plus connue sous l'appellation *structure Plan7* est représentée dans la figure 2.3. Cette structure est une adaptation de la structure classique des HMM profils (présentés section 2.3.1) qui étend le pouvoir de modélisation des modèles d'HMMER.

Cette structure s'appuie sur un noyau assimilable à un HMM profil classique qui est recouvert par une armature permettant de modéliser l'intégralité de la séquence d'acides aminés d'une protéine. Cette armature permet de capter les positions en amont et en aval du domaine modélisé, ainsi que les positions qui sépare les occurrences successives d'un type de domaine en cas de domaine répété. Les états **N** et **C** permettent respectivement de modéliser

```
HMMER2.0  [2.3]
NAME  rrm
ACC   PF00076
DESC  RNA recognition motif. (a.k.a. RRM, RBD, or RNP domain)
LENG  77
ALPH  Amino
RF    no
CS    yes
MAP   yes
COM   ../src/hmmbuild -F rrm.hmm rrm.sto
NSEQ  90
DATE  Tue Apr 29 11:01:43 2003
CKSUM 8325
GA    15.2 0.0
TC    15.2 0.3
XT      -8455     -4  -1000  -1000  -8455     -4  -8455     -4
NULT       -4  -8455
NULE      595  -1558     85    338   -294    453  -1158    (...)
   -21   -313     45    531    201    384  -1998   -644
HMM         A      C      D      E      F      G      H    (...)
          m->m   m->i   m->d   i->m   i->i   d->m   d->d   b->m   m->e
           -16      *  -6492
     1  -1084    390  -8597  -8255  -5793  -8424  -8268    (...)  1
     -   -149   -500    233     43   -381    399    106    (...)
     C     -1 -11642 -12684   -894  -1115   -701  -1378     -16      *
     2  -2140  -3785  -6293  -2251   3226  -2495   -727    (...)  2
     -   -149   -500    233     43   -381    399    106    (...)
     C     -1 -11642 -12684   -894  -1115   -701  -1378       *      *
(...)
    76  -2255  -5128   -302    363   -784  -2353   1398    (...)  103
     -   -149   -500    233     43   -381    399    106    (...)
     E     -1 -11642 -12684   -894  -1115   -701  -1378       *      *
    77   -633    879  -2198  -5620  -1457  -5498  -4367    (...)  104
     -      *      *      *      *      *      *      *    (...)
     C      *      *      *      *      *      *      *       *      0
//
```

FIGURE 2.2 – **Extrait d'un fichier au format HMMER**, contenant le modèle du domaine Pfam rrm (PF00076).

les acides aminés précédant le domaine (en N-ter) et succédant au domaine (en C-ter). L'état **J** permet lui de rechercher dans une séquence plusieurs occurrences du domaine modélisé, contiguës ou non, en formant une boucle autour du HMM profil noyau modélisant le domaine.

Cette structure permet aussi de proposer, pour chaque modèle, une deuxième version du HMM pour la recherche de fragments de domaines (au lieu du domaine dans sa totalité). Cette recherche de fragments est possible grâce à la présence dans le squelette de la structure Plan7 de transitions directes (en pointillés sur la figure) :

- soit de l'état **B** (*Begin*) vers les différents états *Matchs* internes,
- soit depuis les états *Matchs* internes vers l'état **E** (*End*).

Ces transitions créent ainsi des courts-circuits, habituellement non-autorisés dans les HMM profils afin de garantir que la perte d'une partie du domaine soit pénalisée *via* les états *Délétions*. On nomme HMM-*fs* (pour *fragment search*) ces versions alternatives des modèles par opposition à la version originale du HMM nommée HMM-*ls* (pour *local search*). La base

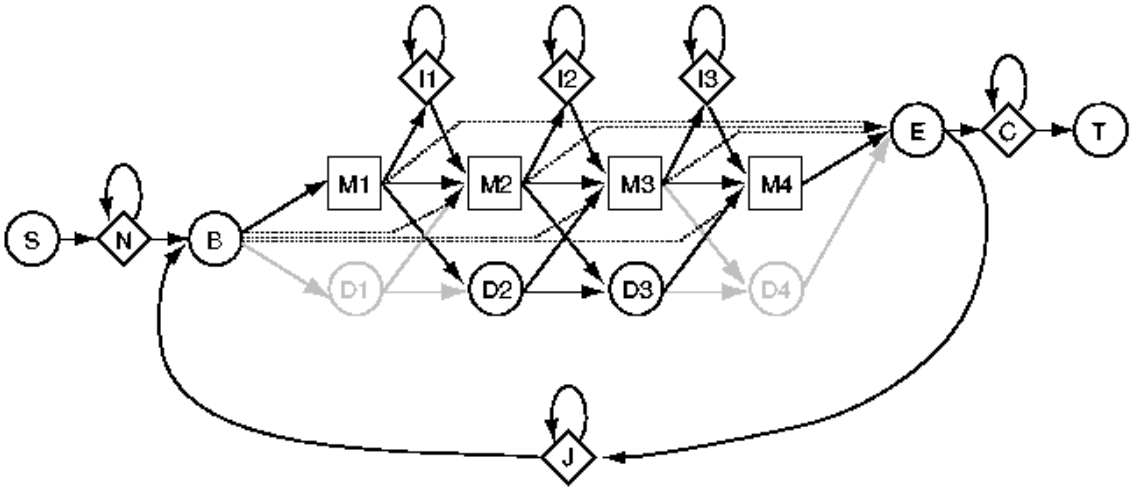


FIGURE 2.3 – **Architecture Plan7 des HMM profils d'HMMER**

de données Pfam (version 23.0) propose en téléchargement ces deux versions des HMM profils pour chacune de ses familles. Cependant, les recherches de domaines réalisées au cours de cette thèse sont exclusivement des recherches de domaines complets (HMM-*ls*) afin d'éviter les domaines tronqués dont on peut douter de la conservation de leur fonctionnalité. Dans une publication récente, Wong *et al.* (2010) décrivent de nombreuses erreurs d'annotation dues à l'identification de fragments de domaines. Selon eux, certaines sous-structures non-globulaires des domaines — telles que les régions transmembranaires ou les peptides signaux — seraient impliquées dans des détections de domaines erronées où seules ces structures sont reconnues. Or la similarité de ces séquences dénote des contraintes physiques et/ou un biais en acides aminés, mais pas nécessairement un lien d'homologie.

c) Construction d'un HMM profil avec HMMER : La première étape pour construire un "bon" modèle consiste à sélectionner l'ensemble de séquences appelées graines (en anglais *seed*) : les séquences les plus représentatives de la diversité évolutive d'une famille de séquences homologues. La sélection des séquences graines permet de créer des modèles à la fois sensibles et sélectifs, c'est à dire qui puissent identifier le maximum d'instances du domaine sans faux-positifs. Pour des raisons historiques, on observe dans les alignements-graines de nombreuses séquences issues d'organismes modèles classiques, qui exhibent une très forte similarité. Les séquences divergentes sont donc sous-représentées dans les alignements-graines des librairies classiques. Pour corriger cet effet, des algorithmes sont utilisés afin d'établir une pondération des séquences ainsi que pour lisser les probabilités de génération en fonction des substitutions en acides aminés préférentielles (*cf.* paragraphe suivant). Cependant, ces algorithmes sont parfois insuffisants pour l'étude de protéines très divergentes. C'est pourquoi nous proposons au chapitre 5 différentes méthodes de correction des modèles, par exemple *via* l'intégration majoritaire de séquences divergentes au sein de l'alignement graine. L'utilisation de séquences homologues pour affiner la construction de modèles est une approche classique

que l'on retrouve notamment dans la base *Fungi*-spécifique FPfam Alam *et al.* (2007) et dans la méthode PSI-BLAST (Altschul *et al.*, 1997).

La deuxième étape est l'apprentissage des paramètres du HMM profil à partir de l'alignement multiple des séquences graines. Pour cela, HMMER identifie tout d'abord les positions les plus conservées (c.-à-d. avec une majorité de séquences sans gap) qui deviendront les états *Matchs*. Ensuite, on estime les probabilités associées aux états *Matchs* en faisant intervenir deux algorithmes :

- *l'algorithme de pondération des séquences* : Les HMM profils ne disposent d'aucune notion de phylogénie. Or, supposer que les séquences observées sont des exemples indépendants et non-corrélés d'un même modèle est clairement faux. C'est pourquoi HMMER pondère les séquences de l'alignement lors de la construction du modèle. L'objectif de ces algorithmes est de relativiser l'impact des séquences les plus semblables et d'accroître le poids des séquences les plus divergentes dont on ne dispose généralement que de peu d'exemples. Intuitivement on comprend que l'information provenant des séquences les plus proches est partagée par celles-ci. On ne doit donc pas leur donner la même influence dans le processus d'estimation des probabilités de génération, qu'à une séquence seule et plus divergente par rapport aux autres séquences de l'alignement. Il existe différents algorithmes pour calculer le poids à attribuer à chaque séquence. On peut notamment citer la pondération de Voronoï (Sibbald et Argos, 1990), par maximum d'entropie (Krogh et Mitchison, 1995) ou encore celle utilisée par défaut dans HMMER, nommée GSC par référence à ses trois auteurs Gerstein/Sonnhammer/Chotia (Gerstein *et al.*, 1994).
- *l'algorithme de lissage des probabilités* : Comme évoqué précédemment pour les HMM généraux, le lissage des distributions de probabilités de génération permet d'attribuer une probabilité faible mais non nulle aux acides aminés qui n'ont jamais été observés à une position donnée de l'alignement. Sans cette étape, on s'expose à des erreurs de prédiction, dues au manque de données lors de l'apprentissage du modèle. En effet, si une séquence possède un acide aminé totalement inédit à une position du modèle, alors la probabilité de tout chemin générant cet acide aminé à cette position est nulle (*cf.* section 2.3.7). Le lissage est traditionnellement réalisé à l'aide de pseudo-comptes. Il existe cependant des approches plus performantes dans le cadre de l'étude des séquences biologiques où l'on dispose de connaissances *a priori* sur les symboles de l'alphabet. Traitant des données protéiques, on peut notamment s'appuyer sur les substitutions préférentielles entre les acides aminés de propriétés physico-chimiques semblables. On trouve donc parmi les méthodes de lissage des distributions, l'utilisation de matrices de substitution entre acides aminés ou des méthodes plus complètes comme l'utilisation de mixtures de Dirichlet (Brown *et al.*, 1993; Sjölander *et al.*, 1996). Les mixtures de Dirichlet combinent la prise en compte des contraintes physico-chimiques (propriété des matrices de substitution), ainsi que de la quantité d'information de départ (propriété des pseudo-comptes).

HMMER détermine également les probabilités de génération associées aux états *Inserts*. Cependant, ces probabilités ne sont pas apprises uniquement à partir de l'alignement des

séquences. En réalité, les positions d'insertions de l'alignement ne servent qu'à pondérer sensiblement une distribution empririque fixée. Tous les états *Inserts* ont donc globalement une distribution de probabilités identique. Cette distribution empirique, représentée figure 2.4, a été choisie par les concepteurs d'HMMER afin de modéliser une caractéristique des insertions dans de nombreuses espèces : elles participent généralement à des "boucles" à la surface des protéines. Il en résulte qu'on observe un biais significatif vers des acides aminés hydrophiles. Les états *Inserts* d'HMMER favorisent donc ce genre de résidus (probabilités de P et S supérieures à la composition moyenne — *cf.* figure 2.5) et pénalisent sensiblement les insertions composés de certains acides aminés hydrophobes (en particulier les aliphatiques et les aromatiques, respectivement représentés en bleu et en turquoise sur les logos).

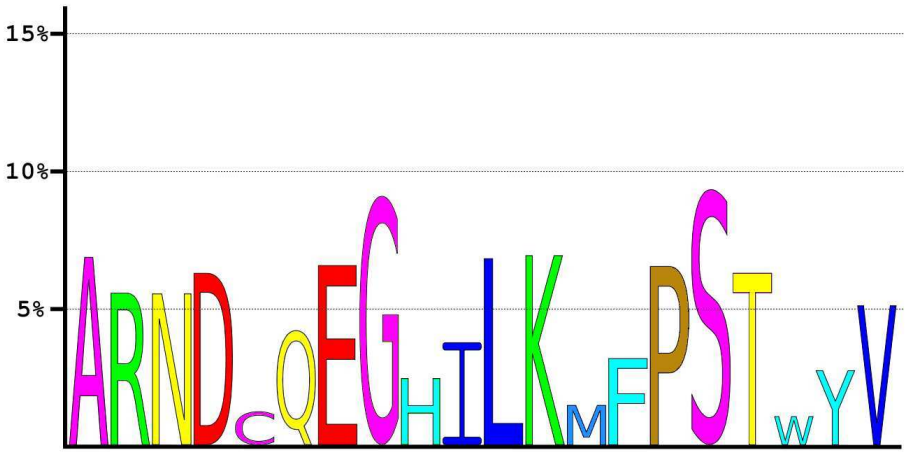


FIGURE 2.4 – **Logo de la composition en acides aminés des états *Inserts* des modèles HMMER**. Le code couleur représente des groupes d'acides aminés aux propriétés identiques : en bleu les aliphatiques (ILV), en turquoise les aromatiques (HFWY), en vert les chargés positifs non-aromatiques (KR), en rouge les chargés positifs (DE), en rose les très petits (ACGS), en jaune les acides aminés polaires n'appartenant pas aux groupes précédents (NTQ) et les deux derniers acides aminés restant : M l'hydrophobe en bleu clair et P l'hydrophile en marron.

d) Calcul d'un score : Une fois le modèle construit, la question est de savoir, étant donné une protéine étudiée, si le HMM profil *reconnaît* cette séquence. Cette reconnaissance peut être évaluée, dans un premier temps, par le calcul d'un score, aussi appelé *log-odds ratio*.

On a vu dans la section 2.3.3, la probabilité de génération d'une séquence $\mathcal{S}$ par un HMM $\mathcal{H}$, notée $P(\mathcal{S}|\mathcal{H})$. Cette probabilité est comparée à la probabilité de générer cette même séquence $\mathcal{S}$ par un modèle dit *nul*, notée $P(\mathcal{S}|nul)$. Le score d'une séquence $\mathcal{S}$ étant donné un HMM profil $\mathcal{H}$ s'apparente au test du rapport vraisemblance et s'obtient par la formule :

$$score(\mathcal{S}|\mathcal{H}) = \log \frac{P(\mathcal{S}|\mathcal{H})}{P(\mathcal{S}|nul)}, \tag{2.3}$$

Dans le logiciel HMMER, le modèle nul est un HMM composé d'un seul état qui boucle sur lui même (*cf.* fig 2.6). Les probabilités de génération de cet état correspondent à la composition moyenne en acides aminés des protéines de Swiss-Prot (*cf.* figure 2.5). Le nombre de fois où l'on reste dans cet état suit une loi géométrique de paramètre 1-p, où p est la probabilité de boucler sur cet l'état (c.-à-d. de ne pas sortir du modèle). L'espérance d'une telle variable est égale à $\frac{1}{1-p}$. Les concepteurs d'HMMER ont donc choisi $p = \frac{350}{351}$ afin que l'espérance du nombre de boucles — et donc de générations — dans cet état soit égal à la longueur moyenne des protéines de Swiss-Prot soit 350 acides aminés (Eddy, 2003).

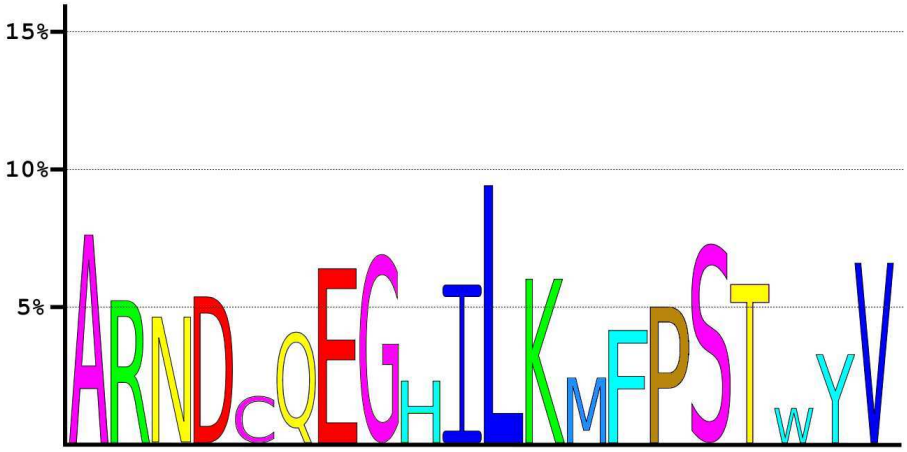


FIGURE 2.5 – **Logo de la composition en acides aminés de l'état du modèle *nul* de HMMER** La description du code couleur est donnée dans la légende de la figure 2.4.

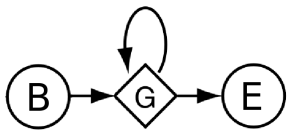


FIGURE 2.6 – **Structure du modèle *nul* de HMMER**

La structure de ce modèle nul est spécifique à ce logiciel. Des études de l'équipe de Karplus (Karplus *et al.*, 2005) suggèrent que le modèle nul le mieux adapté pour une séquence serait de renverser le HMM, ce qui revient à calculer le score de la séquence inversée par le même modèle. Cette approche permet de conserver l'aspect compositionnel (en acides aminés), la longueur de la séquence comparée et celle du modèle. Ce qui, en théorie, permettrait un calibrage plus adapté des E-valeurs (*cf.* section suivante 2.4.2.e "calcul d'une E-valeur"). Cette alternative est implémentée dans le logiciel SAM (concurrent d'HMMER — *cf.* section 2.4.3).

HMMER fait aussi intervenir dans le calcul du score un second modèle nul, appelé *nul2*. La

nécessité d'un second modèle nul s'explique par le problème de l'uniformité des états *Inserts* des HMM profils construits par HMMER. En effet, comme vu dans la section précédente 2.4.2.c, lors de la construction d'un HMM et dans tous les HMM de Pfam, l'ensemble des états *Inserts* ont une distribution de probabilités identique de génération des acides aminés. Le modèle *nul2* est alors utilisé pour éviter qu'une séquence emprunte préférentiellement les états *Inserts* au lieu des états *Matchs*. Comme le premier modèle nul, ce modèle est un HMM composé d'un seul état (*cf.* figure 2.6). Les probabilités de génération associées à cet état sont calculées à la volée, lors de la traversée du HMM par la protéine suivant le chemin de probabilité maximale, en moyennant les distributions associées aux états de ce chemin. Ainsi si une séquence parcours principalement des états *Inserts*, son score sera fortement pénalisé par le modèle *nul2* (car la probabilité de générer la séquence par le HMM et par le modèle *nul2* seront équivalente). L'équation complète du calcul d'un score dans HMMER est donc donnée par la formule :

$$score(\mathcal{S}|\mathcal{H}) = \log \frac{P(\mathcal{S}|\mathcal{H})}{\frac{1}{2}P(\mathcal{S}|nul) + \frac{1}{2}P(\mathcal{S}|nul2)}.$$

Enfin, il est à noter que, par défaut, HMMER dans sa version 2.3.2 ne calcule pas $P(\mathcal{S}|\mathcal{H})$ (la probabilité de générer la séquence $\mathcal{S}$ étant donné le HMM $\mathcal{H}$ obtenue par l'algorithme *forward*), mais $P(\mathcal{S}^*|\mathcal{H})$ la probabilité du chemin de Viterbi dans le HMM $\mathcal{H}$. Cette expression correspond au chemin de probabilité maximale et omet les probabilités de tous les autres chemins permettant de générer $\mathcal{S}$. Bien que la probabilité de ces chemins est fréquemment négligeable par rapport à celle du chemin de Viterbi, le "score de Viterbi" est généralement moins précis que le "score *forward*". Le choix d'HMMER s'explique par le fait que l'on connaît la forme de la distribution des scores de Viterbi mais pas de celle des scores *forward*. Cette distribution des scores étant nécessaire au calcul des E-valeurs (*cf.* section suivante 2.4.2.e), HMMER compense sa perte de précision au niveau des scores par une meilleure estimation des E-valeurs. Nous verrons en conclusion de ce chapitre que cette politique d'HMMER a depuis été modifiée dans la version 3.0 du programme (Eddy, 2010).

e) Calcul d'une E-valeur : Plaçons nous dans dans le cadre de l'étude d'un ensemble de séquences requêtes (typiquement l'ensemble des protéines d'un organisme cible) à l'aide d'une librairie de HMM profils. À partir du score d'une séquence pour un HMM donné, on peut calculer une statistique appelée E-valeur (Eddy, 2003). Cette mesure permet d'évaluer la significativité du score et donc de décider si la séquence a effectivement été reconnue par le modèle. L'E-valeur représente l'espérance du nombre de séquences ayant un aussi bon score que la séquence requête, dans une base de données de séquences aléatoires de taille $\mathcal{M}$, où $\mathcal{M}$ est le nombre de séquences requêtes. Cela entraîne une dépendance directe (proportionnelle) de l'E-valeur d'une séquence à la taille de l'ensemble de séquences dans laquelle on effectue la recherche, d'où la présence du facteur $\mathcal{M}$ dans les équations qui suivent. L'E-valeur peut être calculée de deux façons différentes :

– l'E-valeur brute est calculée de manière analytique et rapide. Elle produit cependant une borne supérieure de l'E-valeur dont l'estimation est trop conservatrice. La formule de cette borne supérieure est décrite par (Barret *et al.*, 1997) et issue des travaux de

Milosavljevic et Jurka (1993) :

$$\text{E-valeur}(\mathcal{S}|H) \leq \mathcal{M}z^{-score(\mathcal{S}|H)},$$

où z est la base du logarithme utilisé dans le calcul du score.

– l'E-valeur empirique (Eddy, 1997) est plus précise mais nécessite un temps de calcul supérieur. Cette E-valeur vient de l'observation de la distribution des scores de Viterbi qui suit une loi de Gumbel, cas particulier (type I) des distributions de valeurs extrêmes — EVD pour *Extreme Value Distribution* — (Gumbel, 1958), de paramètres μ et λ. Cette méthode nécessite donc de calibrer au préalable les paramètres μ et λ de l'EVD pour chaque HMM. HMMER utilise pour cela un histogramme des scores obtenus par des séquences générées artificiellement par son modèle nul, sur lequel il ajuste (*fit*) les paramètres μ et λ. Plus le nombre séquences aléatoires générées pour construire l'histogramme est grand, plus le calibrage est précis, ce qui rend cette méthode plus coûteuse en temps que la précédente. De plus, toutes les familles/HMM ne produisent pas des histogrammes de scores qui suivent parfaitement cette distribution. Une fois les paramètres appris, on obtient l'E-valeur d'une séquence requête par la formule suivante :

$$\text{E-valeur}(\mathcal{S}|H) = \mathcal{M}(1 - z^{-z^{-\lambda(score(\mathcal{S}|H)-\mu)}}).$$

f) Seuils de détection : Il existe deux manières d'affirmer si le modèle a bien reconnu une séquence :

– soit grâce au score obtenu par la séquence contre le HMM : s'il est supérieur à un seuil donné. Par exemple, la base Pfam accompagne chacun de ses modèles de seuils de score inclus dans le fichier du HMM et calibrés afin de minimiser le nombre de faux-positifs. Cependant ces seuils entraînent un manque de sensibilité des modèles pour l'étude de protéines divergentes (faux négatifs — *cf.* Chapitre 4).

– soit par l'E-valeur, si elle est inférieure à un seuil souhaité. Le seuil d'E-valeur est un paramètre dans la recherche de séquence homologue d'HMMER. Par défaut, il vaut 10, mais les résultats contiennent alors souvent des faux positifs, on peut alors envisager 0,1 pour des prédictions relativement sûres mais qui ne dispensent pas d'une étude manuelle des résultats au-delà (et en-deçà) du seuil.

2.4.3 Comparaison SAM/HMMER

Deux suites logicielles proposent les applications nécessaires à la manipulation de HMM profils pour l'étude de séquences biologiques et sont donc en concurrence : HMMER développé par Eddy (1995) à Chevy Chase (Maryland, USA) pour la base de données Pfam, et SAM initié par Hughey et Krogh (1996) à Santa Cruz (Californie, USA) pour les expériences de prédiction de structure CASP et la base SUPERFAMILY. Il existe évidemment de nombreuses différences entre les implémentations de ces deux programmes. Une distinction notable est que SAM

permet de convertir ses modèles au format d'HMMER. La réciproque n'étant pas possible les bases utilisant SAM peuvent proposer deux librairies de HMM pour chacun des programmes. On peut également citer la présence de transitions entre les états *Inserts* et *Délétions* (D→I et I→D), comme implémentation spécifique au programme SAM. Les questions importantes à se poser sont :

- Quel est l'impact de ces différences sur la qualité de ces programmes ?
- Quels sont les points forts/faibles de chacun de ces programmes ?

Plusieurs études ont comparé ces deux programmes concurrents pour tenter de répondre à ces questions. À notre connaissance, la première fut celle de McClure *et al.* (1996) qui cherche à évaluer l'impact des paramètres par défaut et optionnels — l'initialisation des probabilités, la longueur des modèles et la taille des ensembles d'apprentissage — indépendamment pour chaque programme (à l'époque dans leurs premières versions) à travers l'étude de quatre familles protéiques : globines, kinases, protéase d'acide aspartique et ribonuclease H. Ensuite, Lindahl et Elofsson (2000) ont comparé la sensibilité et la spécificité de SAM (version T98) et d'HMMER (version 2.1) ainsi que d'autres algorithmes pour la détection d'homologues distants à partir des familles et des super-familles de la base SCOP. Ils concluent sur de meilleures performances de SAM pour l'étude de super-familles et d'HMMER lorsqu'il s'agit de l'étude au niveau des familles SCOP. Ce résultats est confirmé l'année suivante par Rehmsmeier et Vingron (2001) lors de l'étude de 43 familles SCOP par SAM, HMMER ainsi que leur propre méthode de recherche basée sur des arbres phylogénétiques. Dans cette étude, les auteurs propose une approche originale où la recherche d'homologie se fait grâce à un arbre phylogénétique. À partir d'un arbre appris au préalable sur un alignement d'une famille, chaque séquence d'une banque est intégrée à l'alignement pour construire un nouvel un arbre. Ce dernier est alors confonté à l'arbre de référence pour établir, en fonction de la longueur de branche créée, si la séquence appartient ou non à la famille. Leurs conclusions révèlent que les résultats obtenus par leur méthode seraient supérieurs en terme de minimum de faux positifs à ceux d'HMMER, eux-mêmes supérieurs à ceux obtenus par SAM.

Des études plus récentes ont permis une approche plus systématique des avantages/inconvénients des versions plus récentes de ces programmes. La première, réalisée par Madera et Gough (2002), porte sur deux familles de protéines (globines et cuprédoxines). Elle a permis d'identifier une sensibilité supérieure des modèles construits par SAM, et particulièrement l'importance des alignements-graines (en montrant la qualité des alignements générés par le script T99 de SAM). La supériorité de l'estimation des modèles par SAM a été confirmée par Wistrand et Sonnhammer (2005). Cette étude détaille l'impact des schémas de pondération des séquences de l'alignement graine, ainsi que des mixtures de Dirichlet utilisées pour lisser les distributions de probabilités de génération. En effet, bien que ces deux programmes utilisent lors de la construction de modèles des mixtures de Dirichlet, la mixture *recode3.20comp* de SAM s'appuie sur 20 composantes contre 9 pour celle d'HMMER (Sjölander *et al.*, 1996). La comparaison des algorithmes de pondération par défaut des programmes semble indiquer un léger avantage pour SAM. Cette étude évite la question des alignements-graines en utilisant ceux de Pfam et tranche sur une question non résolue précédemment concernant le programme dont les mesures de scores (et d'E-valeur) semblent les

plus précises : HMMER.

2.4.4 HMMER version 3.0

Un nouvelle version du programme HMMER est disponible depuis quelques mois comportant de nombreuses modifications du programme (Eddy, 2010). Par conséquent nous n'avons pas encore pu mesurer précisément l'impact de ces changements sur les méthodes développées dans cette thèse. Parmi les modifications mineures, on peut citer le format de sortie des recherches de domaines (modifié pour inclure de nouvelles informations) et le format des fichiers .hmm (disparition du modèle nul). Cette dernière modification entraîne notamment l'impossibilité de modifier le modèle nul sans entrer dans le code source d'HMMER. L'un des atouts majeurs de ce nouveau programme est une nouvelle méthode de recherche qui permet de diviser jusqu'à 100 fois le temps de calcul. Pour cela, HMMER3 exploite des calculs intermédiaires de la probabilité de génération de la séquence par le HMM ainsi que des mesures de divergence des séquences pour éliminer de l'ensemble de recherche les protéines n'atteignant pas des seuils fixés (paramétrables). L'autre atout majeur est le passage à des scores *forward*. Grâce à la publication d'Eddy (2008), une distribution des scores *forward* a été proposée. Ils suivraient une loi exponentielle de paramètre λ dans la partie finale des histogrammes. De plus, cette publication approxime une valeur fixe du paramètre $\lambda = log_z(2)$ (z base du logarithme utilisé pour le calcul du score) pour les distributions des scores *forward* et de Viterbi, ce qui permet un calibrage des modèles plus rapide. Ces découvertes nécessitent cependant des modifications de l'architecture de la structure des modèles qui constituent le principal inconvénient de cette nouvelle version. En effet, on ne peut plus contraindre le programme à rechercher des domaines complets : tous les domaines découverts sont potentiellement uniquement des fragments dont la conservation fonctionnelle peut être mise en doute. De plus, on risque de confondre groupe de domaines répétés avec domaine fragmenté.

Chapitre 3

Plasmodium falciparum

3.1 Le paludisme

3.1.1 Une histoire ancienne

Le paludisme, ou *malaria*, est une maladie aux complications parfois mortelles affectant les êtres humains depuis plus de 100 000 ans (Mu *et al.*, 2002; Hay *et al.*, 2004). Des fièvres mortelles, dont probablement certaines d'origine paludique, ont été rapportées dans des traités égyptiens tels que le papyrus d'Ebers (Ebers et Stern, 1875), daté d'environ 1 500 ans avant JC, ainsi qu'en Inde et en Chine plus de 2 000 ans avant JC (Cox, 2002). Ses signes cliniques sont décrits dès l'antiquité en Grèce par Hippocrate (460-377 av. JC) dans "Le Livre des Épidémies". On attribuait alors le paludisme aux miasmes émanant des zones marécageuses comme l'indique son étymologie (*palus* issu du latin "marais"). Giovanni Lancisi, médecin du pape Clément XI, publie en 1717 une étude présentant la preuve que la maladie est transmise par des mouches et introduit le terme malaria, de *mala aria* : mauvais air en italien. Son origine parasitaire n'est cependant découverte qu'à la fin du XIX$^{\text{ème}}$ siècle, par Alphonse Laveran (Laveran, 1880), médecin de l'armée française, qui reçut le prix Nobel de médecine et de physiologie en 1907. L'hypothèse d'un moustique absorbant le parasite puis pondant dans l'eau (que l'homme ingurgite), est émise en 1884 par le Dr Patrick Manson. En 1897, la preuve de la transmission du parasite par la *piqûre* d'une espèce spécifique de moustiques du genre *Anopheles*, est apportée par le médecin britannique Ronald Ross (prix Nobel 1902). Pour un historique plus complet on pourra consulter (Garnham, 1966; Harrison, 1978; Desowitz, 1991).

3.1.2 Une pandémie mondiale

Le paludisme est la maladie parasitaire la plus répandue dans le monde. En 2006, le paludisme était endémique dans 109 pays (essentiellement les plus pauvres d'Afrique, d'Asie et d'Amérique latine) et deux milliards d'individus étaient exposés (soit 40% de la population mondiale). On estime à 500 millions le nombre de personnes atteintes de paludisme parmi lesquelles entre 1,5 et 3 millions décèdent chaque année, principalement des enfants de moins de 5 ans habitant dans les zones d'Afrique sub-saharienne (Rapport de l'Organisation

Mondiale de la Santé, 2006 — lien). En effet, c'est en Afrique subsaharienne, déjà fortement touchée par le VIH, que l'on trouve 85 à 90% des morts du paludisme. De plus, ces deux maladies contribuent à leur propagation mutuelle : le paludisme accroît la charge virale et l'infection par le VIH augmente la probabilité d'une infection paludique (Abu-Raddad *et al.*, 2006). Le paludisme est donc au premier rang des priorités de l'OMS tant par ses ravages directs que par ses conséquences socio-économiques : l'improductivité aboutissant à la sous-alimentation et au sous-développement (Sachs et Malaney, 2002). Au milieu du XX$^{\text{ème}}$ siècle, le paludisme a pu être éradiqué d'une grande partie de l'Europe, de l'Amérique centrale et de l'Amérique du Sud. Cependant, malgré les efforts entrepris pour réduire la transmission de la maladie et améliorer son traitement, il y a eu peu d'évolution depuis le début des années 1990 (Hay *et al.*, 2004). Les zones touchées (dites impaludées) sont essentiellement intertropicales, dans des niches écologiques propices à la reproduction des moustiques (*cf.* Figure 3.3). Plusieurs milliers de cas de paludisme dits d'importation sont aussi recensés chaque année en France métropolitaine. Ils sont principalement consécutifs à des voyages en Afrique subsaharienne (pour 95% des cas). On dénombre également de rares cas de paludisme dit d'aéroport, suite au transport accidentel d'*Anopheles* infectées dans les soutes d'avions en provenance de zones impaludées.

3.1.3 Responsable : le parasite *Plasmodium falciparum*

Les parasites responsables du paludisme sont transmis d'une personne à l'autre par les piqûres de moustiques infectés [1]. Ces parasites sont des eucaryotes unicellulaires appartenant au genre *Plasmodium* du phylum des *Apicomplexa* (dont une cellule type est présentée Figure 3.4). Les apicomplexes sont des parasites intracellulaires dont la majorité sont des agents pathogènes d'espèces métazoaires. Du point de vue phylogénétique, les apicomplexes font partie du règne *Chromalveolata* et plus précisément de la division *Alveolata* (*cf.* Figure 3.1).

Au sein des apicomplexes, la lignée des *Haemasporidia* regroupe tous les parasites malariaux. Ces espèces se caractérisent par l'infection d'un hôte vertébré, la digestion de l'hémoglobine et un cycle de vie complexe comprenant l'ingestion du sang de l'hôte par un moustique vecteur (Valkiūnas, 2004). Les *Haemasporidia* comprennent 4 genres : *Leucocytozoon* infectant une grande variété d'oiseaux, *Haemoproteus* infectant des sauriens (oiseaux et reptiles), *Plasmodium* infectant sauriens et mammifères, et *Hepatocystis* qui infectent des mammifères (Perkins et Schall, 2002; Valkiūnas, 2004). À ce jour, il existe cinq espèces *plasmodiales* recensées infectant l'humain : *Plasmodium falciparum*, *Plasmodium vivax*, *Plasmodium malariae*, *Plasmodium ovale* et *Plasmodium knowlesi*. Les plus répandus sont *P. vivax* et *P. falciparum*. Ce dernier est responsable d'environ 60% à 75% de tous les cas de paludisme, ainsi que de 90% des décès (Hay *et al.*, 2004).

1. Dans le genre *Anopheles*, seules les femelles sont hématophages (par nécessité pour la ponte) et donc vecteurs du paludisme. Les mâles ne piquent pas. Il est à noter que contrairement à l'*Aedes*, moustique vecteur du Chikungunya, de la Dengue et de la Fièvre Jaune, l'*Anopheles* se nourrit de préférence la nuit.

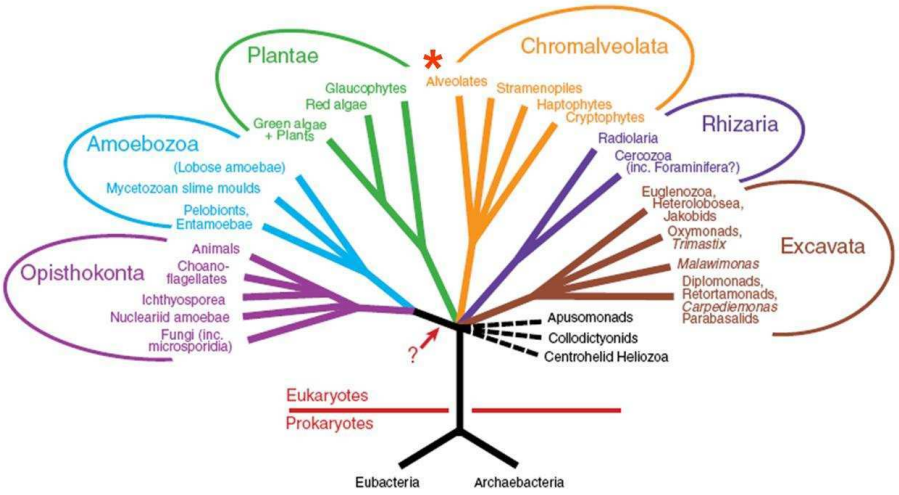


FIGURE 3.1 – **Les règnes des Eucaryotes selon Simpson et Roger (2004).** Ce schéma privilégie les informations moléculaires obtenues à partir d'un ensemble de séquences considérées comme des marqueurs phylogénétiques. La flèche indique la position possible de la racine, fondée sur l'analyse de fusion de gènes marqueurs. Le signe (*) indique le groupe des Alvéolés dans lequel sont classés les apicomplexes.

3.1.4 Cycle parasitaire et effets de l'infection

Le cycle parasitaire des espèces plasmodiales est très complexe (*cf.* Figure 3.2). Il se compose de deux étapes essentielles : une phase de multiplication asexuée (ou schizogonique) chez l'humain et une phase de multiplication sexuée (ou sporogonique) chez le moustique. Au cours de ces phases, le parasite passe par de nombreuses formes distinctes pour lesquelles certaines régulations protéomiques spécifiques ont été mises en évidence (*cf.* section 3.5.1). Lors d'une piqûre, le moustique se nourrit de sang et ingère des parasites sous forme de *gamétocytes* qui se différencient en gamètes dans son estomac. La fécondation des gamètes engendrent la formation d'œufs ou *zygotes*, puis d'*oocystes* qui se placent dans l'interstice des cellules de la paroi stomacale du moustique et libèrent des parasites différenciés en *sporozoïtes*. La durée de la maturation du parasite est étroitement dépendante de la température extérieure : pour *P. falciparum* pas de maturation en dessous de 18°C ou au dessus de 35°C, elle est optimale vers 24°C. Les sporozoïtes remontent alors vers les glandes salivaires du moustique conduisant à l'infection d'un nouvel individu lors de la prochaine piqûre. C'est donc sous la forme de sporozoïtes que le parasite infecte le système sanguin de son hôte humain. Les sporozoïtes migrent très rapidement vers le foie *via* la circulation sanguine pour envahir les cellules hépatiques, où ils se cachent (forme *cryptozoïte*) et se multiplient. Cela donne lieu a une phase exo-érythrocytaire ou hépatique qui va durer environ 6 jours chez *P. falciparum*, 8 jours pour *P. vivax*, 9 jours pour *P. ovale* et 12 jours pour *P. malariae*. Lorsque l'hépatocyte rompt, il donne naissance à plusieurs milliers de *mérozoïtes*. Ces derniers sont relâchés dans le système sanguin et pénètrent dans les érythrocytes (globules rouges). Commence alors le

cycle érythrocytaire de la vie du parasite qui passe par trois stades : le premier stade dit *anneau* (en référence à la forme prise par le parasite), puis le stade *trophozoïte* (où il se nourrit de l'hémoglobine), et enfin le *schizonte* (stade plurinucléé au sein duquel se différencie une nouvelle génération de mérozoïtes). Le cycle érythrocytaire s'achève par l'éclatement du globule rouge et la libération de nouveaux mérozoïtes dans le système sanguin prêts à infecter d'autres érythrocytes. Les éclatements brutaux et synchrones des érythrocytes sont à l'origine des accès de fièvre et s'accompagnent de la libération d'hémozoïne et de différentes endotoxines, qui vont perturber le fonctionnement de l'organisme hôte. Le temps qui s'écoule entre la pénétration d'un parasite dans un globule rouge et l'éclatement de celui-ci est assez constant et atteint chez l'être humain 48 heures pour *P. vivax*, *P. ovale* et *P. falciparum* et 72 heures pour *P. malariae*. Certains mérozoïtes de *P. ovale* ou *P. vivax* peuvent rester cachés dans le foie plusieurs années, voire la vie entière pour *P. malariae*, avant de se réactiver en vagues successives. Durant ces phases dites "dormantes", le parasite ne se réplique pas mais semble en sommeil (forme *hypnozoïte* de Hypnos dieu grec du sommeil). Lors du cycle érytrocytaire certains parasites se différencient en gamétocytes (cellules sexuées) qui ne pourront poursuivre leur développement que s'ils sont ingérés par le moustique. On note que pour le moustique, le cycle de vie du parasite est extracellulaire tandis que chez l'Homme il est intracellulaire.

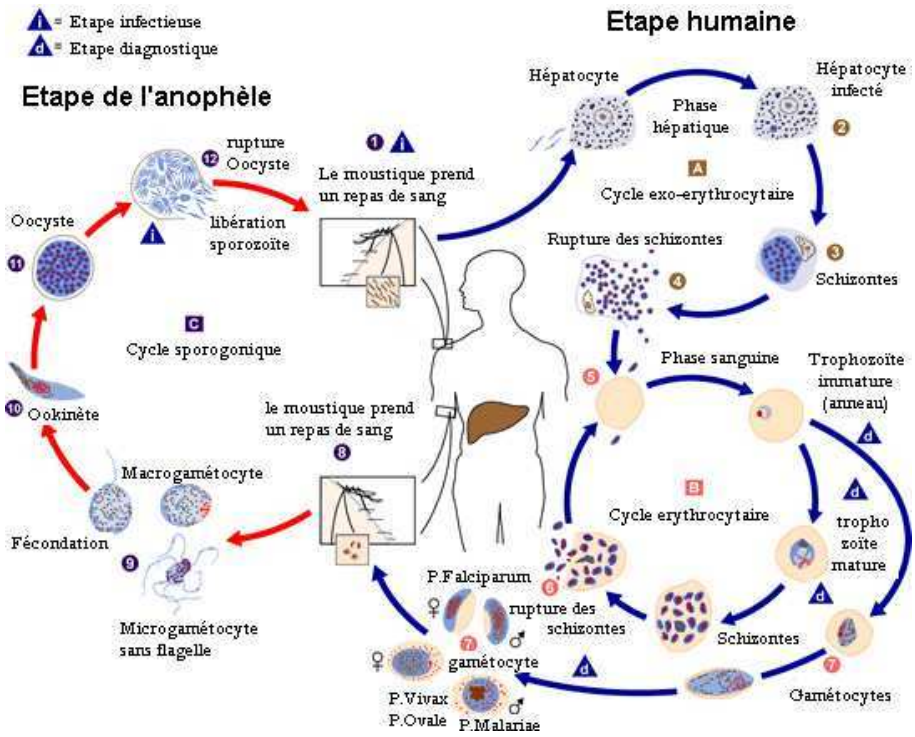


FIGURE 3.2 – **Cycle parasitaire de Plasmodium falciparum.**

Chez l'Homme, le paludisme se manifeste par des accès de fièvre qui peuvent s'accompagner — ou non — de céphalées (maux de tête), de myalgies (douleurs musculaires), d'une

asthénie (fatigue), de vomissements, de diarrhées et de toux. Ces symptômes apparaissent généralement dix à trente jours après la piqûre du moustique. Des cycles typiques alternant fièvre et tremblements avec sueurs froides et transpiration intense, peuvent alors survenir : c'est "l'accès palustre". La périodicité des cycles dépend de l'espèce du parasite en cause, et coïncide avec l'éclatement des globules rouges qui conduit également à l'anémie. En l'absence de traitement, le paludisme à *P. falciparum* peut entraîner rapidement le décès par les troubles circulatoires qu'il provoque.

3.1.5 Cibles thérapeutiques et résistances

Aucun vaccin n'est aujourd'hui disponible mais il existe plusieurs molécules antipaludiques qui peuvent être utilisées en prophylaxie (prévention lors d'un voyage en zone endémique) ou en thérapie.

Au début du XVII$^{\text{ème}}$ siècle, des missionnaires jésuites observent l'utilisation par des populations d'Amérique du Sud (riveraines du lac de Loxa au Pérou) de l'écorce de quinquina (arbuste du genre *Cinchona*) pour soigner les fièvres (Kaufman et Ruveda, 2005). La quinine (alcaloïde végétal qui en est extrait) restera le seul traitement antipaludique connu en Europe et en Amérique jusqu'au XX$^{\text{ème}}$ siècle. Afin de pallier aux difficultés pour se procurer de la quinine, des recherches sont conduites et aboutissent au milieu du XX$^{\text{ème}}$ siècle à la synthèse de ce qui deviendra la chloroquine, utilisée dans les premières campagnes antipaludiques massives de l'OMS. Cependant, dès 1961, des souches de *P. falciparum* résistantes à la chloroquine apparaissent (*cf.* Figure 3.3). Dans les années 70, la combinaison de sulfadoxine et pyriméthamine (antifolates) se substitue à la chloroquine, mais cinq ans seulement ont suffit pour que des foyers de résistances se développent (*cf.* Figure 3.3). Dans ces zones de résistance, seule la quinine reste un antipaludique efficace, mais elle demeure une solution de dernier recours, pour les cas de paludisme sévère, à cause de sa toxicité pour le système nerveux. De plus, on a récemment pu s'apercevoir que ce traitement est lui aussi confronté à de nouvelles résistances du parasite. À l'heure actuelle, il n'existe qu'un seul véritable traitement face au paludisme, à base d'artémisine. Au cours des 20 dernières années, la sécurité et l'efficacité de l'artémisinine a pu être établie. Découvert durant la guerre du Viêt Nam, ce dérivé d'une plante (l'armoise annuelle ou *Artemisia annua*) était employé depuis plus de deux mille ans en Chine sous le nom de *qing hao su*. L'utilisation d'artémisinine est cependant déconseillée en monothérapie pour éviter le développement de résistances par le parasite. Pendant ce temps, un certain nombre de pistes sont explorées pour concevoir de nouvelles thérapies et peut être créer un vaccin. En France, une publication de l'Institut Pasteur vient de confirmer l'intérêt d'un candidat-vaccin contre le paludisme, nommé MSP3 (Roussilhon *et al.*, 2007) et une collaboration entre l'IRD de l'université Paul-Sabatier de Toulouse, le CNRS de Kourou en Guyane, et le Muséum national d'Histoire naturelle, a étudié l'utilisation et les effets de la feuille de thé de *Quassia amaria*, connue en Guyane Française pour ses propriétés antipaludiques (Bertani *et al.*, 2007). On trouvera une compilation des extraits naturels aux propriétés antimalariales (Kaur *et al.*, 2009) ainsi qu'un état des lieux précis des recherches pharmacologiques et vaccinales actuelles dans différentes revues (Good, 2009; Pierce et Miller, 2009).

À l'heure actuelle, les mesures recommandées par l'OMS en matière de lutte contre le

paludisme prévoient :

- un traitement rapide et efficace par des associations médicamenteuses comportant de l'artémisinine (*ACT* pour *Artemisinin-based Combination Therapy*) ;
- l'utilisation de moustiquaires imprégnées d'insecticide et la pulvérisation d'insecticide à effet rémanent à l'intérieur des habitations pour lutter contre les moustiques vecteurs ;
- l'installation d'air conditionné dans les habitations pour faire baisser la température et brasser l'air (afin de perturber le moustique dans ses déplacements et dans sa faculté sensorielle à trouver sa cible) ;
- après le coucher du soleil : application de crème répulsive sur la peau ou les vêtements, port de vêtements amples, longs et de couleur claire et abstinence d'alcool (les anophèles sont aussi bien attirées par les couleurs foncées, plus spécialement le noir, que par les vapeurs d'alcool).

Cependant la situation reste préoccupante car depuis plusieurs années dans de nombreuses régions du monde, les moustiques développent des résistances aux insecticides, notamment le DDT (Dichloro-Diphényl-Trichloréthane) utilisée depuis les années 1960, avec excès dans certaines régions du globe. Des moyens alternatifs pour combattre le vecteur du paludisme ont été mis en place :

- assèchement des marais (sans bouleverser le système écologique), drainage des eaux stagnantes où se développent les larves des anophèles ;
- ensemencement des eaux avec des prédateurs des anophèles ou de leurs larves comme certains mollusques ou poissons (tilapias, guppys, gambusies, aphanius) ;
- réintroduction et protection des variétés de chiroptères insectivores là où elles ont disparu ;
- identification de cibles à partir du séquençage du génome du moustique (Holt *et al.*, 2002), par exemple l'emploi d'insectifuges et d'insecticides ciblés uniquement contre l'anophèle ou encore la dispersion de mâles anophèles stériles dans la nature.

Bien qu'efficace sur un territoire limité, ces mesures sont très difficiles à appliquer à l'échelle d'un continent tel que l'Afrique.

3.2 Publication du génome de *P. falciparum*

Si, au départ, les biologistes travaillaient sur des gènes isolés pour leurs intérêts thérapeutiques, la publication en 2002 du génome complet de *P. falciparum* (Gardner *et al.*, 2002) a profondément dynamisé la recherche sur le paludisme. Le génome de *P. falciparum* contient 14 chromosomes dont la taille varie entre 0,64 et 3,3 Mb (Mégabase = 1 million de paires de bases). De plus, il existe deux génomes dits extranucléaires : un génome mitochondrial compact de 6 kb et un génome plastidial circulaire de 35 kb (kilobase) découvert récemment au sein d'un organite connu sous le nom d'apicoplaste que le parasite a acquit par endosymbiose secondaire d'une algue (Kohler *et al.*, 1997; Maréchal et Cesbron-Delauw, 2001) (*cf.* Figure 3.4). Sept années de travaux ont été nécessaires pour obtenir les 22,8 Mb de la souche *3D7* de *P. falciparum*, et révéler 5 268 gènes codants prédits. À titre de comparaison, le séquençage de *Drosophila melanogaster* (Adams *et al.*, 2000) et de ses 180 Mb réparties sur 4 chromo-

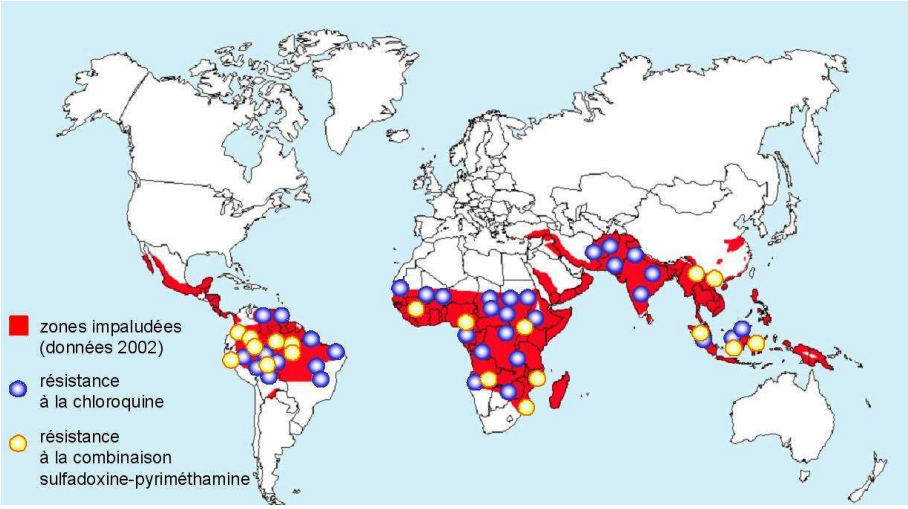


FIGURE 3.3 – **Zones impaludées dans le monde** (zones rouges). Les foyers de résistances aux traitements par la chloroquine (cercles bleus) et aux associations sulfadoxine-pyriméthamine (cercles jaunes) sont répartis sur l'ensemble des zones impaludées.

somes a été réalisée en une seule année. La difficulté d'assemblage et d'analyse du génome de *P. falciparum* est essentiellement due à une particularité extrême, à savoir un taux en A+T, adénine et thymine (deux des quatre nucléotides de l'ADN), supérieur à 80% (Musto *et al.*, 1995). Toutefois, les nouvelles stratégies de séquençage développées pour surmonter ces difficultés ont ouvert la voie pour le séquençage complet de nombreuses espèces de parasites (Lau, 2009). L'étude du génôme de *P. falciparum* a révélé un certain nombre d'atypicités sur lesquelles nous revenons dans la section 3.3 qui suit.

La connaissance du génome a aussi ouvert les perspectives de nouvelles voies de recherche que l'on qualifie de post-génomiques. Elles visent notamment à annoter fonctionnellement les gènes séquencés et à caractériser les produits des gènes par des approches globales. De nombreuses études basées sur les puces à ADN, la spectrométrie de masse et d'autres techniques ont été publiées, permettant l'étude du transcriptome, du protéome, du secrétome et de l'interactome. Pour une revue de ces travaux, ainsi que des comparaisons inter-espèces et un état des lieux des pistes à explorer pour la recherche de nouveaux vaccins, on pourra consulter (Kooij *et al.*, 2006; Winzeler, 2008). Enfin, une base de données dédiée à l'annotation des génomes plasmodiaux a été créée et est régulièrement mise à jour sur le site PlasmoDB [2] (Bahl *et al.*, 2003). Cette base est liée aux sites ApiDB (Aurrecoechea *et al.*, 2007) (regroupant les autres apicomplexes dont *Toxoplasma gondii* (qui dispose également de sa propre base : ToxoDB (Gajria *et al.*, 2007)) et EuPathDB (Aurrecoechea *et al.*, 2009b) pour l'ensemble des pathogènes eucaryotes. PlasmoDB permet d'accéder à l'ensemble des séquences génomiques et protéiques des espèces plasmodiales entièrement séquencées, ainsi qu'à de nombreuses données transcriptionnelles, protéomiques et des informations relatives à la fonction

2. http ://www.plasmodb.org/plasmo/

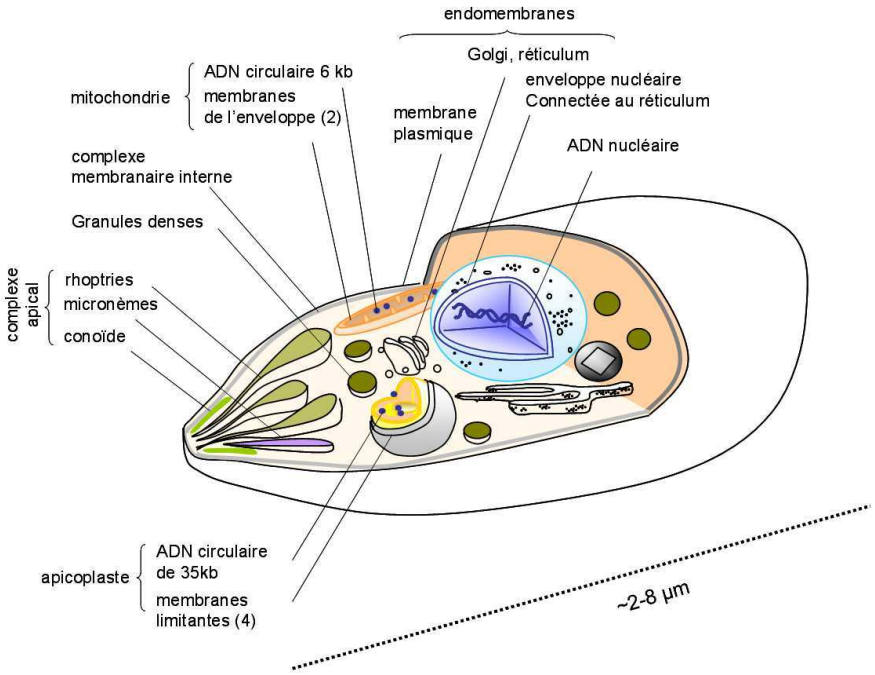


FIGURE 3.4 – **Caractéristiques générales d'une cellule de parasite apicomplexe.** Les zoïtes (formes cellulaires invasives) sont des cellules fortement polarisées. Trois organites sécrétoires vésiculaires, les rhoptries, micronèmes et granules denses participent à l'invasion et à l'élaboration de la vacuole parasitophore. Comme dans les cellules végétales, en plus du noyau et de la mitochondrie, un plaste vestigial ou apicoplaste contient de l'ADN. À la différence des chloroplastes simples des cellules végétales, l'apicoplaste est entouré d'un système membranaire additionnel de nature endosomale. La membrane plasmique est doublée d'un complexe membranaire interne, structure spécialisée dérivée des alvéoles qui caractérisent le groupe des Alvéolés.

putative ou établie des gènes, à des familles d'orthologues, aux domaines protéiques, aux annotations GO, *etc.*

3.3 Atypicités

3.3.1 Biais dans la composition en acides aminés

La principale particularité du génome de *P. falciparum* est donc son taux très élevé en A+T atteignant 90% dans certaines régions (Gardner *et al.*, 2002). Ce taux est remarquable car il est nettement supérieur à celui observé dans la plupart des organismes. Lors de la traduction, ce déséquilibre dans la composition en nucléotides des gènes se traduit par un biais de la composition en acides aminés des protéines. Ce biais constitue un obstacle important à la recherche de séquences homologues et à l'annotation fonctionnelle de *P. falciparum*.

Lorsque l'on parle de biais compositionnel chez *P. falciparum*, on se compare implicitement à un ensemble d'organismes de référence appelés organismes modèles. Les organismes modèles sont les organismes qui ont fait l'objet des études à la paillasse les plus massives et bien souvent les plus anciennes. Par conséquent, ce sont les premiers à avoir été complètement séquencés. Chez la plupart des organismes modèles, comme la levure par exemple, le taux de A+T est généralement d'environ 60%. Le biais compositionnel d'un organisme (AT-riche ou GC-riche) s'observe donc par un déséquilibre dans la fréquence d'utilisation des acides aminés codés par des codons AT-riches (FYMNIK) et GC-riches (GARP) dans l'ensemble de ses protéines. Nous avons comparé la distribution en acides aminés chez *P. falciparum* (d'après les séquences protéiques extraites du site Web de PlasmoDB version 5.5 datant du début 2009), avec celle des organismes actuellement séquencés et répertoriés dans Swiss-Prot (bénéficiant d'une grande qualité d'annotation). La figure 3.5 représente, pour ces deux ensembles de protéines, la distribution des 20 acides aminés, ordonnés par le pourcentage moyen d'A+T permettant leur synthèse (calcul uniquement à partir du code génétique sans tenir compte d'une table d'usage des codons). Cette figure montre clairement la baisse de fréquence des 4 acides aminés GC-riches (GARP) chez *P. falciparum*. On remarque que cette diminution de l'arginine (R), codée par 6 codons différents, est compensé par la lysine (K) bien que codée uniquement par deux codons distincts. Ces acides aminés sont tous deux chargés positivement et facilement interchangeables. On observe également que la leucine (L) codée par 6 codons semble être fréquemment remplacée par l'isoleucine, codée par 3 codons mais AT-riche, qui est également un autre acide aminé aliphatique. Un chiffre à retenir pour illustrer ce déséquilibre est celui de la prédominance des acides aminés (I, K, N). Avec ces trois seuls acides aminés, *P. falciparum* code plus de 35% de ses séquences protéiques.

De plus, il a été suggéré que le biais provient d'une pression d'origine nucléique (Singer et Hickey, 2000; Bastien *et al.*, 2004). On observe en effet que la distribution des nucléotides diffère selon la position dans les codons. Les chiffres extraits de la table d'usage des codons de *P. falciparum* (PlasmoDB 6.5) montre un pourcentage en A+T supérieur en deuxième position d'un codon par rapport à la première position (environ 78% contre 68% respectivement). Le plus fort taux de A+T (>82%) est observé en troisième position des codons (Musto *et al.*, 1995). De plus, on constate une autre particularité en troisième position : l'inversion des tendances entre l'adénine et la thymine. Sur les deux premières positions des codons le taux d'adénine est supérieur de plus de 20% à celui de thymine. En troisième position la fréquence d'une thymine devient plus élevée (de 5%) que celle de l'adénine, ce qui se traduit notamment par une utilisation plus fréquente de l'asparagine N (codée à 86% par le codon AAT et à 14% par AAC) que de la lysine K (codée à 82% par AAA pour 18% de AAG).

3.3.2 Insertions de faible complexité

La seconde particularité que l'on observe chez *P. falciparum* est la longueur de ses protéines, environ 20% plus longues que les protéines homologues d'autres organismes (Pizzi et Frontali, 2001). Quand un alignement multiple est possible, cette différence de taille semble provenir de la présence dans les protéines de *P. falciparum* de longues insertions, allant

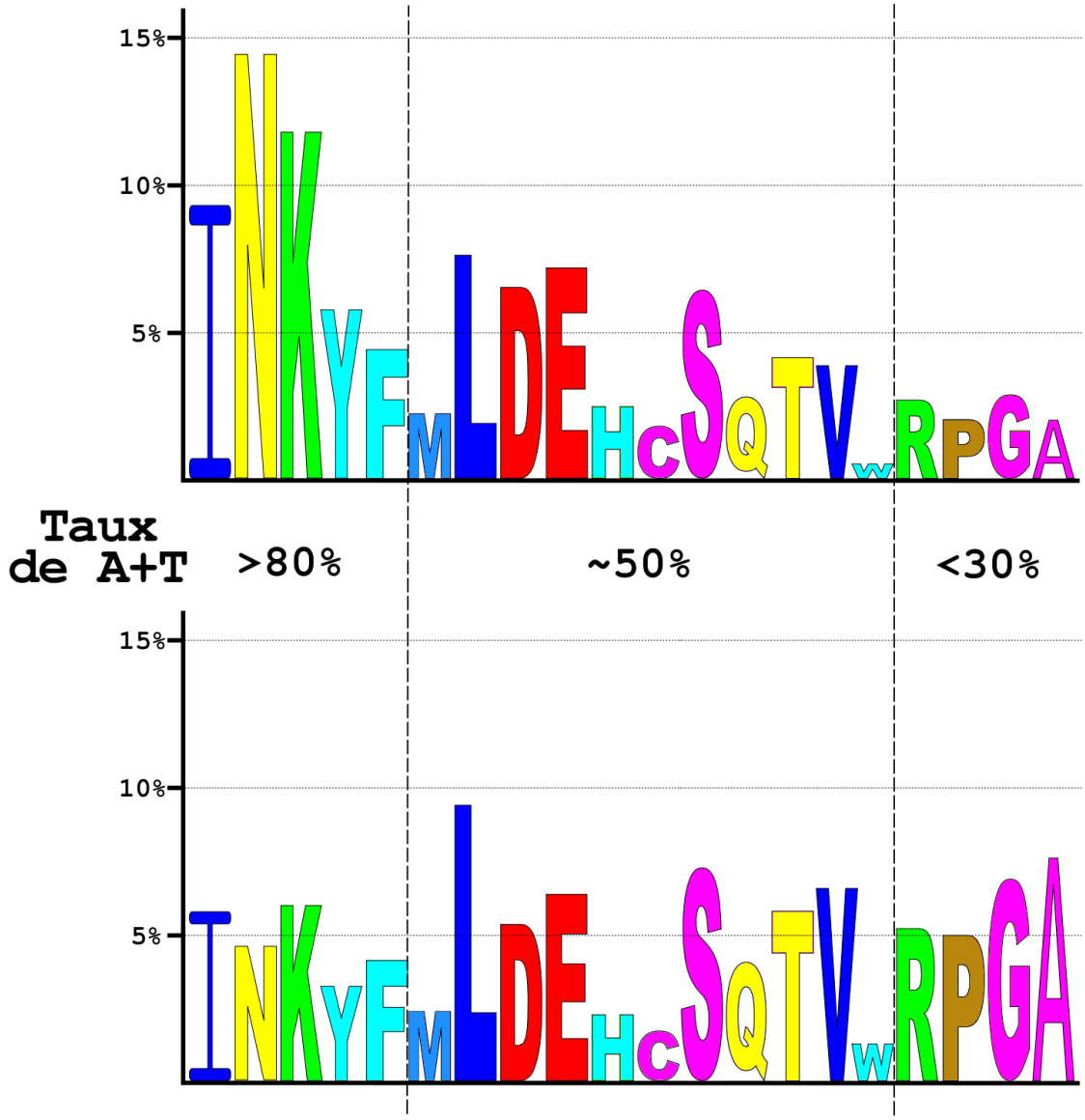


FIGURE 3.5 – **Logos de la composition moyenne en acides aminés des protéines de *P. falciparum* (en haut) et de Swiss-Prot (en bas)**. Les fréquences des acides aminés sont reportées sur l'axe de droite. Les acides aminés sont placés dans l'ordre décroissant (de gauche à droite) en fonction de leur richesse en AT (*cf.* figure 2.4 66 pour le code couleur). L'ordonnancement est calculé selon la proportion en A+T moyenne des codons (indiquée au centre). Seul le code génétique universel est considéré, c.-à-d. que l'usage des codons de *P. falciparum* n'intervient pas.

parfois jusqu'à plusieurs centaines d'acides aminés. Ces insertions séparent parfois des blocs bien conservés qui sont adjacents dans les protéines homologues des autres espèces (*cf.* Figure 3.6) et gênent par la même occasion les recherches de similarité notamment lors du calcul de score BLAST, à cause du biais compositionnel (Bastien *et al.*, 2005).

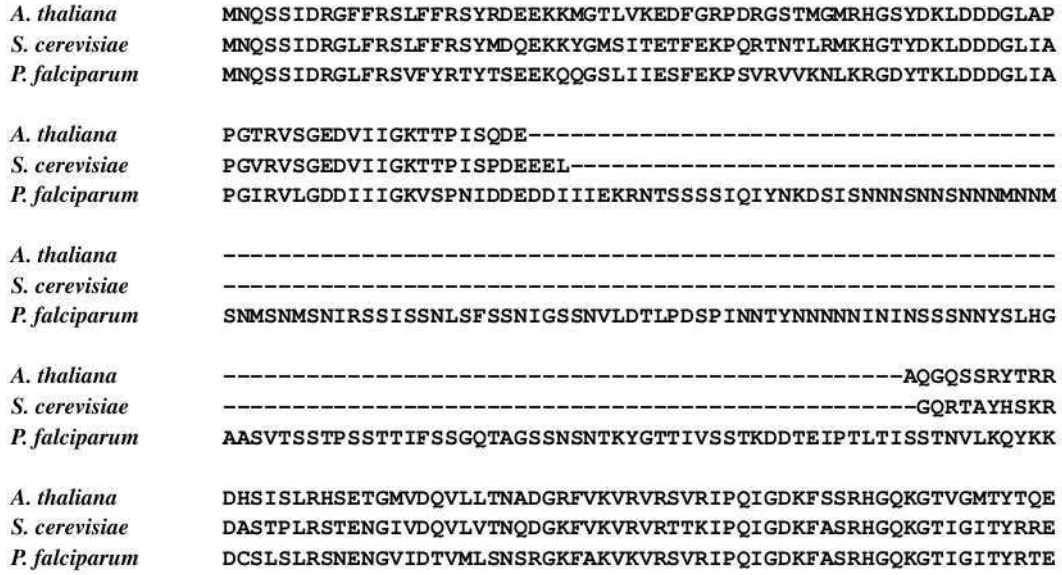


FIGURE 3.6 – **Exemple d'une insertion de faible complexité caractérique des protéines de *P. falciparum*.** Les protéines alignées sont des deuxièmes sous-unités d'ARN polymérases de type II appartenant à *A. thaliana* (AT5G45140 — positions 804 à 958), *S. cerevisiae* (AAA68096 — positions 841 à 997) et *P. falciparum* (PFB0715w — positions 816 à 1174). La zone de faible comlexité se trouve au centre du domaine Pfam RNA_pol_Rpb2_6 (PF00562).

La caractéristique notable de ces insertions est la présence de zones de *faible complexité*, terme issu de la théorie de la complexité (Kolmogorov, 1968; Lempel et Ziv, 1976), appliquée pour la première fois aux séquences protéiques par (Wootton et Federhen, 1993). On observe dans les insertions de longs segments composés d'une répétition intensive d'un seul acide aminé ou de très courts motifs. De plus, il a été observé une divergence rapide des résidus au centre de ces insertions avec une préférence pour des acides aminés hydrophiles. Cependant, les insertions semblent conservées sur leurs bords pour des raisons de contraintes phénotypiques (Pizzi et Frontali, 2001). La sous-représentation d'acides aminés hydrophobes, semble indiquer que ces zones codent des domaines non-globulaires à la surface des protéines plasmodiales. La fonction de ces insertions reste à l'heure actuelle inconnue. Cependant, elles ne semblent pas, *a priori*, altérer le repliement fonctionnel de la protéine. Les régions nucléiques de faible complexité étant autant présentes dans les introns que dans les exons, l'hypothèse d'une pression de sélection au niveau nucléique a été émise. On observe dans ces régions de faible complexité une utilisation de l'asparagine N et de la lysine K qui atteignent respectivement des fréquences de 16,4% et 13,3%, révélant, malgré une préférence pour les codons

A-riches sur les deux premières positions, qu'il existe une pression de sélection chez *P. falciparum* en faveur de l'asparagine N (codée par AAT et AAC) sur la lysine K (codée par AAA et AAG). La caractérisation fonctionnelle et structurale des insertions chez *P. falciparum* reste un champ de recherche ouvert.

3.4 La question du positionnement phylogénétique

La phylogénie de *P. falciparum* est un sujet non-résolu par la communauté scientifique car son origine et sa position au sein des *Haemosporidia* restent obscures. Depuis les années 90, de nombreuses phylogénies moléculaires ont placé *P. falciparum* plus proche des parasites aviaires que des parasites mammifères. L'hypothèse d'un passage récent de l'hôte aviaire à humain était donc préférée à celle d'une histoire ancestrale commune aux parasites de mammifères (Waters *et al.*, 1991; Kissinger *et al.*, 2002; Escalante et Ayala, 1994; McCutchan *et al.*, 1996). Cependant, il a récemment été reconnu que les résultats de ces travaux proviennent d'un manque de séquences analysées. L'étude de gènes uniques ainsi qu'un échantillon de taxons insuffisants ont probablement conduit à un biais affectant ces reconstructions phylogénétiques (Hagner *et al.*, 2007). Depuis environ 5 ans, des études utilisant une quantité de données plus importante ont montré une monophylie de l'ensemble des parasites malariaux des mammifères, y compris *P. falciparum*, contredisant l'hypothèse aviaire. Cette monomphylie mamifère se compose de trois lignées distinctes :

- une lignée composée de *P. falciparum* et de ses plus proches relatifs chez les grands singes : *P. reichenowi* et *P. gaboni* ;
- la lignée des parasites infectant les primates dans laquelle on retrouve les autres espèces plasmodiales infectant l'homme (*malariae, ovale, vivax* et *yoelii*) ;
- la lignée des parasites infectant les rongeurs (*Rodent Malaria Parasites* ou RMPs).

Selon ces études, *P. falciparum* et ses plus proches relatifs chez les grands signes seraient issus de la plus ancienne divergence au sein de la monophylie mammifère (Perkins et Schall, 2002; Martinsen *et al.*, 2007; Roy et Irimia, 2008; Hayakawa *et al.*, 2008; Carlton *et al.*, 2008b). Cependant, une controverse récente voudrait que cette lignée soit plus proche des RMPs que des parasites de primates, suggérant ainsi une émergence plus récente (Perkins, 2008; Blanquart et Gascuel, 2010).

La difficulté des reconstructions phylogénétiques, due au biais compositionnel, est accentuée par un biais dans l'usage des codons (en particulier en troisième position des codons), ainsi que par la divergence extrêmement rapide des gènes lignée- et espèce-spécifiques (Kuo et Kissinger, 2008). Les études phylogénétiques des *Haemasporida* s'appuient donc sur des méthodes de reconstruction phylogénétique adaptées pour permettre la représentation de modèles d'évolution non-homogènes et non-stationnaires (Yang et Roberts, 1995; Galtier et Gouy, 1998; Foster, 2004; Blanquart et Lartillot, 2006) :

- Les modèles non-homogènes visent à éviter les artefacts dits d'*attraction des longues branches*. Lorsque des séquences ont connu une évolution rapide, on observe que les modèles d'évolution homogènes sont parfois enclin à les regrouper sans qu'elles aient pourtant de liens phylogénétiques directs, car le signal phylogénétique est masqué par

les multiples substitutions (Felsenstein, 1978).
- Les modèles non-stationnaires permettent d'éviter des artefacts dûs à l'*attraction des séquences de composition similaire*. Ces séquences ont tendance à être regroupées par les modèles stationnaires car elles partagent des convergences interprétées à tort en tant que signaux phylogénétiques (Lockhart *et al.*, 1992).

D'autres publications proposent des méthodes prenant en compte les biais dans l'usage de codons (Dávalos et Perkins, 2008) et des biais dans la composition en acides aminés des protéines variable le long de la phylogénie (hétérogénéité au cours du temps et au long des sites (Blanquart et Lartillot, 2008)).

3.5 Difficultés d'annotation

La principale conséquence des atypicités des séquences plasmodiales est qu'elles rendent difficile, voire parfois impossible, l'identification — par similarité de séquences — de protéines homologues caractérisées dans d'autres organismes. Cela constitue un frein majeur au transfert d'annotations fonctionnelles aux protéines du parasite et limite donc la désignation de cibles potentielles pour la découverte de vaccins et de médicaments. Des 5 268 protéines initialement prédites, il n'est pas possible d'attribuer la moindre protéine homologue dans des organismes modèles pour plus de 60% d'entre elles. Bien que certaines de ces protéines ont été identifiées comme spécifiques au genre *Plasmodium* ou à *P. falciparum* uniquement, ce taux reste bien plus élevé que dans la majorité des organismes séquencés où moins de 40% de protéines n'ont pas d'homologues identifiés. De plus, de nombreuses familles protéiques s'avèrent être sous-représentées telles que les protéines impliquées dans le métabolisme ou les facteurs de transcription (Coulson *et al.*, 2004; Callebaut *et al.*, 2005). Deux hypothèses sont avancées. La première est qu'une partie de ces protéines sont réellement absentes du génome de *P. falciparum*, conséquence du mode de vie parasitique. Leur absence serait cependant compensée par d'autres protéines, d'autres mécanismes épigénétiques (*RNA decay*) ou des ARN non-codants (*non-coding RNA* ou ncRNA) qui participeraient à des voies métaboliques uniques au genre *Plasmodium* (Li *et al.*, 2007; Mourier *et al.*, 2007). La deuxième hypothèse est que la divergence du génome de ce parasite empêchent la détection de ces protéines par similarité à des protéines homologues issues d'organismes modèles. La question est donc de savoir s'il est possible d'adapter les outils d'annotations classiques à la divergence et au biais compositionnel de *P. falciparum* afin de surmonter les limitations actuelles de ces méthodes.

3.5.1 Gènes spécifiques

Il est reconnu qu'un certain nombre de gènes de *P. falciparum* ne possèdent aucun homologue dans les organismes modèles car ils sont spécifiques au genre ou à l'espèce ; on parle de *gènes orphelins*. Plusieurs études des espèces plasmodiales ont révélé la présence de nombreux gènes dits stade-spécifiques correspondant aux 4 stades sanguins de la vie du parasite (sporozoïte, mérozoïte, trophozoïte et gamétocyte — *cf.* Section 3.1.4). Les deux études de référence par puces à ADN (Le Roch *et al.*, 2003; Bozdech *et al.*, 2003) ont montré des patterns de transcriptions spécifiques à chaque stade, ainsi que la production de gènes spécifiques

à certains stades. On retrouve la même cascade d'expression de gènes que chez Bozdech *et al.* (2003) et dans d'autres souches de *P. falciparum* (Llinás *et al.*, 2006). Dans son étude du protéome, Florens *et al.* (2002) ont découvert que seulement 6% des 2425 protéines identifiées étaient exprimées durant les quatre stades, et que plus de la moitié des gènes exprimés à un stade étaient spécifiques à un stade unique. On recherche donc à accroître le nombre de gènes dont on connaît la fonction parmi ces gènes spécifiques à certains stades en annotant les gènes "putatifs" ou "hypothétiques" notamment grâce aux annotations des gènes exprimés simultanément (profils trascriptomiques similaires), comme réalisé pour les mérozoïtes par (Florent *et al.*, 2004, 2009). Cette approche permettrait de mieux cibler les gènes clés pour un vaccin/médicament en isolant ceux qui interviennent dans certaines phases sensibles du cycle parasitaire, en bloquant par exemple la cascade de transcription. Une découverte récente de ce type a été faite chez *Toxoplasma gondii* (Bougdour *et al.*, 2009) avec l'identification de la protéine FR235222 indispensable à la différenciation des stades *tachyzoïte* à *bradyzoïte* et donc au développement du parasite.

De nombreuses études du protéome et du transcriptome ont été menées. Grâce à la disponibilité d'un certain nombre de séquences provenant de plusieurs espèces plasmodiales proches, notamment les parasites de rongeurs (*rodent malaria parasites* ou RMP), certaines de ces études ont apporté de meilleures connaissances pour l'étude des différents *Plasmodium*. On peut citer par exemple la découverte de 3500 à 3800 orthologues entre *P. yoelii* et *P. falciparum* soit 60% de gènes en commun (Carlton *et al.*, 2002), ou encore l'isolation de familles de gènes spécifiques au genre *Plasmodium* grâces aux zones de synténie entre ces espèces (Kooij *et al.*, 2005) et aux phases communes aux cycles parasitaires (Hall *et al.*, 2005). Récemment, un important travail de comparaison a également été réalisé par Carlton *et al.* (2008a) visant à accroître l'intérêt pour *P. vivax*, seconde espèce plasmodiale la plus répandue et virulente (parfois fatale) et qui a surtout la particularité d'être la moins biaisée en A+T (avec un taux inférieur à 60%).

De plus, il est communément admis que les différentes espèces plasmodiales possèdent des gènes propres à l'infection des espèces qu'elles parasitent (invasion de l'hôte et du vecteur). Chaque espèce plasmodiale, et *P. falciparum* en particulier, se distingue donc par des gènes uniques. On trouve notamment dans les zones subtélomériques des chromosomes, de nombreux gènes orphelins, dont la fonction est liée à l'invasion immunitaire, à la séquestration des cellules hôtes et des variants antigéniques (Janssen *et al.*, 2004). Bien qu'existant également au centre des chromosomes, ces gènes sont majoritairement présents dans les répétitions subtélomériques et forment les plus grandes familles multigéniques des parasites plasmodiaux. Chez *P. falciparum*, on dénombre douze de ces familles (représentant 5 à 10% des gènes) dont seulement cinq se retrouvent chez les RMP (Kooij *et al.*, 2006).

3.5.2 Gènes cachés ou perdus

Parmi les mécanismes biologiques que l'on pense exister chez *P. falciparum*, certains semblent absents alors même qu'on observe parfois leur produit. Il arrive aussi qu'une partie des gènes associés à un complexe soit présents mais que certains éléments indispensables à la fonction biologique restent introuvables. On peut citer par exemple :

- les molécules de transport qui semblent réduites en comparaison des autres espèces vivantes non-parasitaires (Gardner *et al.*, 2002) ;
- les facteurs de transcriptions, où le complexe protéique TFIID pourtant extrêmement bien conservé au sein des eucaryotes, ne trouve quasiment aucune correspondance chez *P. falciparum* (Coulson *et al.*, 2004). L'utilisation d'une méthode de comparaison au niveau de la structure secondaire (HCA) (Callebaut *et al.*, 2005) prédit cependant plus de facteurs de transcription généraux qu'on ne pensait initialement. De plus, ces séquences sembleraient être présentes chez les autres espèces plasmodiales. Mais cette étude souligne encore l'absence d'éléments primordiaux et confirme l'hypothèse de Coulson *et al.* (2004) concernant le rôle majeur des processus de post-transcription dans la régulation des niveaux de protéines chez *P. falciparum*. Il existe à l'heure actuelle un important débat sur les mécanismes de régulation chez *P. falciparum* et notamment sur le rôle de l'épigénétique dans ces mécanismes étant donné le faible nombre de facteurs de transcription identifiés. Nous verrons dans la partie 4.5 (page 108), que les approches proposées dans cette thèse ont permis d'identifier un certain nombre de cibles qui pourraient être impliquées dans le contrôle de l'expression génétique.

3.5.3 Modification des outils d'alignement de séquences

De nombreux efforts ont été réalisés pour identifier ces gènes cachés par la divergence des séquences génomiques, notamment en adaptant les outils d'alignement de séquences. Dans ce type d'approche on recherche à identifier chez *P. falciparum* une protéine homologue à une protéine annotée dans un organisme modèle afin de pouvoir transférer l'annotation fonctionnelle. La similarité des séquences est déterminée par l'alignement des séquences d'acides aminés (requête et cible) à l'aide d'une matrice de score qui représente les similarités physico-chimiques entre les différents acides aminés. La difficulté rencontrée par cette approche lors de l'étude de *P. falciparum* est que les séquences requêtes sont généralement des séquences issues des organismes modèles qui ne possèdent ni le biais compositionnel ni la divergence des protéines plasmodiales. La modification des outils d'alignement de séquence passe donc par la correction des matrices de substitutions d'acides aminés (matrice de score) pour permettre la comparaison d'espèces de compositions différentes. C'est ce que proposent les travaux de Yu *et al.* (2003), Yu et Altschul (2004) et Bastien *et al.* (2005). Ces auteurs ont créé des adaptations de la matrice de substitution BLOSUM62 tenant compte de la différence de composition entre séquences requête et cible. Ces nouvelles matrices ont été appliquées à la comparaison :

- de *fructose-bisphosphate aldolases* chez *P. falciparum* et *Fusobacterium nucleatum* (Yu *et al.*, 2003) ;
- de l'*asparagine synthase* de *P. falciparum* avec la protéine *PurF* de *Mycobacterium tubercolosis* (Yu et Altschul, 2004) ;
- d'orthologues groupés par paires, chez plusieurs espèces (Yu et Altschul, 2004). Les paires d'orthologues extraites de COG forment trois jeux de données. Le premier associe

les séquences de *Clostridium tetani* (AT-riche) avec celles de *M. tuberculosis* (GC-rich), le second *Bacillus Subtilis* et *Lactococcus lactis* (génomes non-biaisés) et le dernier *M. tuberculosis* et *Streptomyces cœlicolor* (tous deux fortement biaisées vers GC) ;

– d'orthologues entre *P. falciparum* et *Arabidopsis thaliana* par (Bastien *et al.*, 2005).

Dans tous ces cas, il existe une amélioration du score de l'alignement par rapport aux alignements standards qui n'est pas dû à un artefact comme le prouvent les résultats de rétrocontrôle de Yu et Altschul (2004). De plus, dans le cadre de la théorie de l'information, Bastien *et al.* (2005) ont montré un gain de l'entropie relative de leurs matrices suggérant un gain théorique en sensibilité. On doit cependant noter qu'aucune de ces matrices n'a été évaluées à plus grande échelle.

3.6 Détection des domaines protéiques

Dans cette thèse, nous nous intéressons à l'utilisation de HMM profils pour la détection des domaines protéiques chez *P. falciparum*. Comme décrit dans le chapitre 1, l'utilisation de librairies de modèles (proposées par des bases de données en libre accès sur le Web) permet d'identifier les domaines qui composent les protéines d'un organisme cible. Cependant, comme pour l'alignement de séquences (Section 3.5.3), la modélisation s'appuie elle aussi sur les séquences issues des organismes modèles. Les HMM profils proposés ne sont donc pas adaptés à la détection de domaines dans des protéines divergentes. La recherche de domaines Pfam révèle seulement 1 421 familles de domaines distinctes dans les protéines de *P. falciparum* sur les 10 340 familles existantes. À titre de comparaison, 2 369 familles de domaines sont répertoriés chez la levure (*cf.* Table 3.1). De plus, la procédure standard de détection de domaines Pfam est incapable d'identifier le moindre domaine dans 47% des protéines de *P. falciparum* (contre 24% chez la levure). Cette tendance est également observée pour d'autres organismes apicomplexes (*cf.* Table 3.1). De plus, il semble que de nombreux domaines classiques des eucaryotes soient absents du répertoire de *P. falciparum*. On peut notamment citer les résultats des travaux de Coulson *et al.* (2004), qui ont montré la limite de l'utilisation des HMM profils de Pfam pour la recherche de facteurs de transcription. Sur les 51 HMM associés à des facteurs de transcription, seuls 17 trouvent une correspondance chez *P. falciparum* dans 69 protéines. Cela représente seulement 1.3% du génome de *P. falciparum* contre 5.7% de moyenne pour les sept autres espèces de cette étude (*S. cerevisiae, S. Pombe, C. elegans, D. melanogaster, A. thaliana, H. Sapiens, A. gambiae*). Parmi les quatre domaines ayant les meilleurs scores pour *P. falciparum*, trois d'entre eux ont la plus faible abondance relative chez les autres espèces, tandis que le quatrième n'apparaît que 11 fois tous les 10 000 gènes chez *P. falciparum*, contre 130 fois tous les 10 000 gènes en moyenne chez les autres eucaryotes.

Cette absence de domaines chez *P. falciparum* par rapport aux organismes eucaryotes modèles peut s'expliquer en partie par son adaptation à un mode de vie parasitique. Cependant, ce phénomène est vraisemblablement sur-évalué à cause d'une modélisation et d'une détection des domaines inadaptées. L'objectif de cette thèse est de proposer des méthodes permettant d'affiner la modélisation et la détection de domaines dans les protéines de *P. falciparum*. Dans la suite, nous présentons dans le chapitre 4 une approche permettant

la détection des domaines divergents grâce à l'exploitation des propriétés de co-occurrence des domaines protéiques. Son application a permis d'identifier de nombreux nouveaux domaines chez *P. falciparum* et ses espèces proches. Le chapitre 5 de cette thèse est consacré à l'intégration du biais compositionnel dans de la modélisation de domaines par des HMM profils. Les résultats obtenus, en adaptant par différentes approches les paramètres des HMM profils au biais compositionnel de *P. falciparum*, ont également permis la découverte de nouveaux domaines chez *P. falciparum*, dont certains uniques à ces librairies *plasmodifiées*.

	Espèce	Nombre de protéines	Familles de domaines	# total de domaines	Couverture	
					Séquences	Résidus
Organismes modèles	*Anopheles Gambiae*	12 347	2 991	18 472	74%	38%
	Arabidopsis thaliana	34 517	3 125	46 796	74%	40%
	Caenorhabditis elegans	22 637	2 953	26 424	65%	37%
	Homo sapiens	40 252	3 914	64 282	68%	39%
	Saccharomyces cerevisae	5 809	2 373	6 939	76%	40%
Apicomplexes	*Cryptosporidium parvum*	3 805	1 214	3 337	54%	20%
	Plasmodium falciparum	5 249	1 421	5 782	53%	18%
	Plasmodium vivax	5 432	1 415	3 470	50%	17%
	Plasmodium yoelii	7 724	1 313	3 925	42%	23%
	Theileria annulata	3 790	1 169	4 061	55%	21%

TABLE 3.1 – **Comparaison inter-espèces du nombre de domaines Pfam et de la couverture du génome**. "Couverture : Séquences" représente le pourcentage de protéines ayant au moins un domaine Pfam identifié. "Couverture : Résidus" représente le nombre d'acides aminés participant à ces domaines (chiffres extraits du site Web de la base Pfam version 23.0). On constate un nombre réduit de domaines distincts et une faible couverture du génome dans le genre *Plasmodium* ainsi que chez les autres apicomplexes (*C. parvum* et *T. annulata*). Cette tendance est également observée pour de nombreuses espèces de pathogènes eucaryotes (Ghouila *et al.*, 2010).

Quatrième partie

Travaux

Chapitre 4

Certification de domaines par co-occurrence

Comme vu lors du premier chapitre, déterminer la composition exacte en domaines d'une protéine est une étape clef pour prédire sa fonction. Nous avons présenté lors du second chapitre les HMM profils, qui se sont révélés être un outil puissant de modélisation des domaines protéiques. Parmi les bases de données proposant ce genre de modèles, l'une des plus populaires est Pfam, qui fournit une importante librairie de HMM profils accompagnés de seuils de score permettant de limiter le nombre de faux positifs lors de la recherche des domaines d'une protéine. Cependant, cette approche peut manquer de sensibilité dans le cas de protéines fortement divergentes. Appliquée à *P. falciparum*, cette stratégie se révèle incapable de détecter le moindre domaine dans 47% de ses protéines, tandis que de nombreux domaines semblent absents du répertoire de *P. falciparum*. De plus, seulement 1 421 types de domaines distincts ont pu être identifiés. À titre de comparaison 2 369 types de domaines sont répertoriés chez la levure, et concernent plus de 76% des protéines (*cf.* Table 3.1 page 89).

Abaisser les seuils requis pour la détection des domaines permettrait de plus nombreuses identifications, mais au prix d'un nombre d'erreurs important. Une solution consiste à utiliser une source d'information supplémentaire afin de filtrer parmi ces nouveaux domaines potentiels ceux dont la présence est la plus vraisemblable. Pour cela, nous proposons dans ce chapitre une méthode exploitant la tendance des domaines protéiques à apparaître préférentiellement avec un nombre réduits d'autres domaines favoris dans une protéine (*cf.* section 1.3.4 page 31). Dans une protéine, quand le score d'un modèle Pfam n'est pas suffisant pour faire accepter le domaine par les seuils recommandés, notre méthode *certifie* la présence du domaine en se basant sur la présence d'autres domaines favoris dans la protéine.

Ce chapitre débute par une présentation détaillée du principe de notre méthode de certification par co-occurrence et de ses paramètres (section 4.1). Cette méthode s'accompagne d'une procédure d'estimation du taux d'erreur qui est décrite en section 4.2. Notre approche est validée grâce à une expérience sur la levure où l'évolution (la divergence) de ses protéines est simulée, et l'impact des différents paramètres est évalué (section 4.3). Puis la méthode est appliquée à l'annotation fonctionnelle des protéines de *P. falciparum* (section 4.4) sur lequel une analyse biologique approfondie des résultats est menée (section 4.5). Nous étendons

alors son application aux espèces proches *P. vivax* et *P. yoelii* où des résultats analogues et congruents sont observés (section 4.6). Dans la section 4.7, nous comparons notre approche aux méthodes mettant à profit le contexte des domaines pour améliorer l'annotation des protéines. Les développements en cours et autres perspectives concluent ce chapitre (section 4.8), avec notamment la présentation de l'interface Web mise à disposition de la communauté pour consulter l'intégralité des résultats de notre approche.

4.1 Présentation de la méthode

L'objectif de cette approche est d'enrichir l'annotation fonctionnelle des protéines d'un organisme par une connaissance plus précise de leur *composition en domaines*. Par composition en domaines, on entend l'ensemble de domaines d'une protéine sans tenir compte ni de l'ordre séquentiel ni du nombre d'occurrences de chacun des domaines. Ce choix s'appuie sur l'hypothèse que la présence d'une famille de domaines est l'information primordiale pour annoter fonctionnellement une protéine (Cohen-Gihon *et al.*, 2007). Le principe de notre approche consiste à utiliser les propriétés de co-occurrence des domaines protéiques pour certifier la présence de nouveaux domaines *potentiels* dans une protéine à partir de la présence d'un autre domaine dit *validant*.

Ce travail repose principalement sur les propriétés de co-occurrence des domaines protéiques, présentées dans la section 1.3.4 (page 31). Notre approche requiert dans un premier temps l'identification de paires de domaines montrant une co-occurrence forte dans de nombreuses protéines, c'est à dire que la présence de l'un des domaines de la paire doit être un indice fort pour la présence de l'autre domaine. Ces paires de domaines conditionnellement dépendants (CDP pour *Conditionaly Dependent Pairs*) sont apprises sur un grand nombre de séquences à l'aide d'un test statistique (test exact de Fisher). Les CDP forment alors une liste de référence qui est utilisée de la manière suivante. Considérons une protéine d'un organisme cible pour laquelle un ou plusieurs domaines *potentiels* sont détectés en relâchant les seuils de score des HMM de Pfam. Si un de ces domaines potentiels forme, avec un autre domaine non-recouvrant de la protéine (dit *validant*), une paire appartenant à la liste des CDP, alors la présence de ce domaine potentiel est considéré comme *certifiée*.

Pour appliquer cette approche de certification par co-occurrence, on a donc besoin d'établir au préalable la liste des CDP (*cf.* section 4.1.1). Il faut ensuite déterminer les ensembles de domaines *validants* V_i et *potentiels* P_i de chaque protéine i de l'organisme étudié (section 4.1.2). Le mécanisme de certification peut alors se formaliser ainsi : on certifie un domaine potentiel $x \in P_i$, grâce à un domaine validant $y \in V_i$, si la paire $(x, y) \in$ CDP.

4.1.1 Sélection des CDP

La liste des paires de domaines conditionnellement dépendantes est calculée à partir de l'ensemble des paires de domaines co-occurents, observées dans un grand nombre de protéines (à l'exception de l'organisme étudié). Ici nous avons choisi de nous appuyer sur la composition en domaines de l'ensemble des protéines d'Uniprot. Les CDP sont utilisées pour certifier dans une protéine cible la présence potentielle d'un domaine grâce à la présence d'un domaine favori

		Domaine A		Totaux
		présent	absent	
Domaine B	présent	x	y	$b = x + y$
	absent	w	z	$d = w + z$
Totaux		$a = x + w$	$c = y + z$	N

TABLE 4.1 – **Table de contingence 2 x 2**

de celui-ci. Elles doivent donc révéler une dépendance conditionnelle entre ces domaines, c.-à-d. que la présence de l'un des domaines doit être un indice fort de la présence de l'autre domaine. Toutes les paires observées dans Uniprot ne satisfont pas ce critère. Par exemple, si deux domaines apparaissent ensembles mais également avec de nombreux autres domaines différents, il est évident que ces deux domaines — dits versatiles — ne forment pas une paire conditionnellement dépendante. Nous devons donc filtrer les paires observées à l'aide d'un test statistique pour obtenir la liste des CDP nécessaire à la procédure de certification.

Tester la dépendance conditionnelle d'une paire de domaines revient à mesurer l'association de deux variables. Une solution à ce problème peut être apportée par un test de corrélation (par exemple un test du χ^2). Nous avons choisi d'appliquer un test exact de Fisher (*one-tailed*), plus précis pour de petits échantillons comme c'est parfois le cas ici. Afin de réaliser ce test, on établit une table de contingence pour chaque paire observée de domaines (A, B), $A \neq B$ (voir table 4.1) qui reporte le nombre de protéines où A et B ont été observés ensemble (noté x), le nombre de protéines possédant le domaine A mais où le domaine B est absent (noté w), le nombre de protéines possédant le domaine B mais où le domaine A est absent (noté y) et le nombre de protéines où A et B sont absents (noté z).

Sous l'hypothèse nulle d'indépendance des domaines A et B, la probabilité d'une telle table correspond à la probabilité d'observer **exactement** x protéines avec les domaines A et B parmi b protéines ayant B, sachant qu'il existe a protéines possédant le domaine A dans l'ensemble total des N protéines observées. Cette probabilité suit une loi hypergéométrique :

$$P(x|a,b,N) = \frac{C_a^x \times C_{N-a}^{b-x}}{C_N^b} = \frac{\frac{a!}{x!(a-x)!} \times \frac{(N-a)!}{(b-x)!(N-a-b+x)!}}{\frac{N!}{b!(N-b)!}} = \frac{a!(N-a)!b!(N-b)!}{x!(a-x)!(b-x)!(N-a-b+x)!N!}.$$

Le raisonnement peut se faire dans l'autre sens, c.-à-d. calculer la probabilité d'observer x protéines avec les domaines B et A parmi les a protéines ayant le domaine A, sachant qu'il existe b protéines ayant le domaines B dans l'ensemble des N protéines observées. Ce calcul conduit aux équations suivantes :

$$P(x|b,a,N) = \frac{C_b^x \times C_{N-b}^{a-x}}{C_N^a} = \frac{\frac{b!}{x!(b-x)!} \times \frac{(N-b)!}{(a-x)!(N-b-a+x)!}}{\frac{N!}{a!(N-a)!}} = \frac{a!(N-a)!b!(N-b)!}{x!(a-x)!(b-x)!(N-a-b+x)!N!},$$

qui se révèlent équivalentes aux précédentes : le test est donc symétrique.

Enfin, la probabilité d'observer **au moins** x protéines avec les domaines A et B parmi $x + y$ protéines ayant B, sous l'hypothèse nulle d'indépendance des domaines, est donnée par

la somme des probabilités de la table observée et des tables plus extrêmes au sens de l'écart à l'indépendance. On obtient ainsi la P-valeur du test exact de Fisher pour une paire (A, B) de domaines, par la somme de lois hypergéométriques suivante :

$$P\text{-}valeur(A,B) = \sum_{n=x}^{min(a,b)} P(n|a,b,N).$$

Une P-valeur est donc calculée pour chaque paire de domaines. Si cette P-valeur est inférieure à un seuil fixé, l'hypothèse nulle est rejetée, les domaines sont considérés comme conditionnellement dépendants, et la paire est ajoutée à la liste des CDP.

Notons pour finir, qu'en accord avec notre objectif, on établit une liste de CDP dont chaque paire est composée de familles de domaines *distinctes*. En effet l'information que nous souhaitons utiliser est la coopération entre domaines et non pas la répétition de domaines.

4.1.2 Domaines potentiels et validants

Une fois apprise la liste des CDP sur un grand nombre de protéines, on se replace au niveau de l'organisme cible. Il nous faut alors déterminer parmi ses protéines où se trouvent les domaines potentiels et validants utilisés dans le processus de certification.

a) Inférence des domaines potentiels : L'ensemble des domaines potentiels (P_i) est inféré à partir des résultats de la recherche de domaines protéiques en utilisant le logiciel `HMMER` et la librairie complète de HMM de Pfam. Nous avons vu précédemment (*cf.* section 2.4.2 pages 66 à 68) qu'étant donné un ensemble de protéines et un HMM, `HMMER` permettait le calcul d'un score reflétant la similarité de chaque séquence au domaine protéique modélisé par le HMM. Ce score est généralement comparé à un seuil calibré manuellement pour authentifier la présence du domaine et garantir l'absence de faux-positifs. De plus, ce score est utilisé pour estimer une E-valeur qui représente l'espérance du nombre de séquences qui obtiendraient un aussi bon score.

La construction de l'ensemble des domaines potentiels nécessite plusieurs étapes dont la première consiste à considérer toutes les occurrences de domaines renvoyées par le programme `HMMER` qui diffèrent des domaines déjà connus et dont l'E-valeur est inférieur à une valeur seuil permissive. Cette valeur est choisie pour être beaucoup moins conservatrice que les seuils de score recommandés par Pfam pour chaque HMM. Une fois fixé le seuil d'E-valeur, on dispose, pour chaque protéine, d'une collection de domaines (avec leur position sur la séquence). Cette collection de domaines n'est pas exempte de chevauchement, c.-à-d. de domaines détectés sur les mêmes positions/acides aminés de la séquence. L'étape suivante consiste, dans un premier temps, à éliminer tous les domaines potentiels qui chevauchent un domaine connu de la protéine, puis à construire une liste de domaines potentiels non-chevauchants. Pour cela, nous avons donc mis en place est une heuristique qui conserve en priorité les domaines potentiels de meilleure E-valeur. Pour chaque protéine, l'heuristique va mémoriser successivement le domaine potentiel ayant la meilleure E-valeur et, s'il existe d'autres domaines potentiels qui le chevauchent, alors on élimine ces domaines de la collection. D'autres critères pourraient être envisagés et font actuellement l'objet d'expérimentations

dans le cadre de la thèse d'Amel Ghouila. Notons que, lors de cette étape, il faut considérer les positions alignées sur des états *Inserts* comme des positions non-occupées par le domaine afin de prendre en compte les phénomènes de domaines *encastrés*. Ce phénomène peut être observé par exemple chez *P. falciparum* dans la protéine PFB0715w, où le domaine Pfam RNA_pol_Rpb2_2 (PF04561) est encastré dans le domaine RNA_pol_Rpb2_1 (PF04563). Ce genre de conformation n'est pas unique à *P. falciparum* : on l'observe pour des protéines orthologues de plusieurs espèces[1] y compris des organismes modèles (levure, drosophile, *etc.*). Il est marginal, sans être unique puisqu'on l'observe fréquemment pour certaines familles de domaines Pfam — par exemple le domaine HHH (PF00633) encastré dans le domaine HhH-GPD (PF00730). Cela peut s'expliquer par un mécanisme d'insertion d'un domaine fonctionnel complet au sein d'un autre domaine, suite par exemple à un évènement d'*exon shuffling* (Gilbert, 1978). À l'issue de la sélection des domaines non-chevauchants, un même domaine peut encore apparaître plusieurs fois dans la protéine. La dernière étape pour l'obtention des ensembles de domaines potentiels (P_i) consiste alors à ne retenir que le nom/identifiant de chaque domaine (sans considérer les positions/occurrences) afin d'éliminer toute redondance.

b) Choix des domaines validants : Le choix de l'ensemble des domaines validants (V_i) est un paramètre très important. En effet, c'est en se basant sur ces domaines que l'on certifie la présence de nouveaux domaines. Trois types de domaines validants ont été considérés dans ces travaux :

– **Les domaines Pfam connus :** La première solution est d'utiliser les domaines Pfam connus dans la protéine (c.-à-d. les domaines détectés par les seuils de score recommandés par Pfam). Cet ensemble peut être obtenu à l'aide du logiciel HMMER ou téléchargé directement depuis la base de données dédiée à l'organisme cible (par exemple la base PlasmoDB pour *P. falciparum*). Cette solution est la plus naturelle et la plus sûre.

– **Les domaines Interpro (non-Pfam) connus :** Une solution complémentaire consiste à considérer l'ensemble des domaines d'InterPro connus dans la protéine, à l'exclusion des domaines issus de Pfam. Cette liste de domaines peut être obtenue à l'aide du programme `InterProScan` ou téléchargée depuis une base de données en ligne. L'utilisation de l'intégralité des bases de données InterPro permet d'accroître considérablement le nombre de domaines validants de chaque protéine. Par conséquent, on s'attend à obtenir un plus grand nombre de domaines certifiés. Cependant, l'hétérogénéité des schémas de domaines des bases d'Interpro risque de conduire à des certifications de moindre qualité par rapport à celles réalisées grâce aux domaines Pfam connus.

– **Les domaines Pfam potentiels :** Les deux précédents ensembles de domaines validants fournissent une base solide pour la certification de domaines potentiels, car la présence de ces domaines est indiscutable. Ils induisent néanmoins une limitation importante : on ne peut certifier un domaine que dans des protéines où la présence d'au moins un autre domaine est déjà connue. Or, les annotations les plus intéressantes sont justement attendues dans des protéines où, jusqu'ici, aucun domaine n'a pu être identifié. Pour surmonter cette

1. *cf.* http ://pfam.sanger.ac.uk/family ?acc=PF04563#tabview=tab1

limitation, une solution est de considérer un troisième ensemble de domaines validants : les domaines potentiels eux-mêmes. Dans cette solution, toutes les paires de domaines potentiels sont énumérées et si une paire appartient à la liste des CDP, les deux domaines sont certifiés. Bien sûr cette procédure est beaucoup plus sujette à certifier de faux positifs que les deux précédentes mais nous allons voir dans la section 4.2 comment cela peut être contrôlé.

Nous venons de définir trois ensembles de domaines validants disjoints et de qualité *a priori* décroissante. Notons pour finir que, pour certifier un domaine potentiel, seuls les domaines validants qui ne recouvrent pas ce domaine seront considérés. Cela permet notamment d'empêcher la certification d'un domaine Pfam par un domaine Interpro équivalent connu à cette position.

4.2 Estimation du nombre d'erreurs

À partir de la liste des CDP apprise sur les protéines d'Uniprot, et des domaines potentiels et validants d'un organisme cible, on est capable de certifier un certain nombre de nouveaux domaines dans cet organisme. Une question est alors d'évaluer le nombre de faux positifs parmi ces nouveaux domaines. Pour cela, nous estimons la probabilité de certifier un domaine potentiel sous l'hypothèse H_0 où tous les domaines potentiels sont prédits de manière aléatoire. Ceci est réalisé par des simulations, à l'aide d'une procédure de ré-échantillonnage des différents domaines potentiels des protéines. Permuter aléatoirement les différents domaines potentiels crée une situation dans laquelle les domaines potentiels sont indépendants des domaines validants, tout en préservant la distribution des types de domaines, ainsi que le nombre de domaines potentiels et validants de chaque protéine.

La procédure de ré-échantillonnage, détaillée dans l'algorithme 3 et dans la figure 4.1, est la suivante. Dans un premier temps, les domaines validants associés aux protéines sont mémorisés et tous les domaines potentiels sont collectés. Puis on redistribue aléatoirement les domaines potentiels à travers les différentes protéines, en respectant le nombre original de domaines potentiels de chaque protéine. On crée ainsi de nouveaux ensembles de domaines potentiels P_i^* de même taille que les ensembles P_i originaux. On applique ensuite notre méthode de certification aux ensembles P_i^*, grâce à la liste des CDP et aux ensembles V_i originaux, pour comptabiliser le nombre de domaines aléatoires certifiés. Cette procédure est répétée un grand nombre de fois (typiquement 1 000) pour obtenir une bonne estimation de l'espérance du nombre de domaines certifiés sous H_0. Ce nombre est alors utilisé pour calculer une estimation du taux de faux positifs (*False Discovery Rate*, ou *FDR*) associé à l'ensemble original de domaines certifiés, par la formule :

$$FDR = \frac{\text{estimation du nombre de certification sous } H_0}{\text{nombre de domaines certifiés sur les données originales}}$$

Cette approche est similaire à celles proposées dans Soriç (1989) et Benjamini et Hochberg (1995) pour contrôler le FDR associé à des tests multiples. Nous verrons qu'en jouant sur le seuil d'E-valeur utilisé pour définir les domaines potentiels, on peut contrôler le *FDR* associé à nos certifications grâce à cette procédure de ré-échantillonnage (*cf.* section 4.3.2).

Algorithm 3: Algorithme de ré-échantillonnage. Cet algorithme prend en entrée les ensembles de domaines validants $\{V_1, \ldots, V_N\}$ et potentiels $\{P_1, \ldots, P_N\}$, la liste des CDP $liste\_CDP$ et un nombre de répétitions $nbBoucle$. À chaque répétition (ligne 3), on rassemble l'ensemble des domaines potentiels dans une collection $\mathcal{P}$ (ligne 4). Ensuite, pour les N protéines de l'organisme cible (ligne 6), on construit un nouvel ensemble de domaines potentiels P_i^* (de même taille que P_i) par des tirages aléatoires sans remise des domaines de $\mathcal{P}$ en s'assurant de ne pas insérer plusieurs fois le même domaine dans P_i^* (lignes 8 à 12). Si le hasard des tirages mène à une solution où il est impossible de redistribuer intégralement les domaines potentiels en respectant cette contrainte, la redistribution est interrompue et une nouvelle redistribution est réalisée (lignes 13-14). Dans le cas contraire, la procédure de certification est lancée sur ces données randomisées et le nombre de domaines certifiés est comptabilisé (ligne 16). L'algorithme se termine et renvoie le nombre moyen de certifications sous H_0.

Données: $\{V_1, \ldots, V_N\}$, $\{P_1, \ldots, P_N\}$, $liste\_CDP$, $nbBoucle$
Résultat: Renvoie l'espérance du nombre de certification sous H_0

1 $NbCertification \leftarrow 0$;
2 $B \leftarrow 0$;
3 **tant que** $B \neq NbBoucle$ **faire**
4 $\quad \mathcal{P} \leftarrow \{P_1,\ P_2\ \ldots\ P_N\}$;
5 $\quad nbTestId \leftarrow true$;
6 $\quad$ **pour** i *de* 1 *à* N **faire**
7 $\quad\quad P_i^* \leftarrow \emptyset$;
8 $\quad\quad$ **tant que** $(taille(P_i^*) \neq taille(P_i)) \wedge nbTestId$ **faire**
9 $\quad\quad\quad$ Tirer au hasard $dom \in \mathcal{P}$;
10 $\quad\quad\quad$ **si** $dom \notin P_i^*$ **alors**
11 $\quad\quad\quad\quad P_i^* \leftarrow P_i^* \cup \{dom\}$;
12 $\quad\quad\quad\quad \mathcal{P} \leftarrow \mathcal{P} - \{dom\}$;
13 $\quad\quad\quad\quad$ **si** $\forall\ d \in \mathcal{P},\ d \in P_i^*$ **alors**
14 $\quad\quad\quad\quad\quad nbTestId \leftarrow false$;
15 $\quad$ **si** $nbTestId$ **alors**
16 $\quad\quad NbCertification \leftarrow NbCertification + certifie(\{V_1, \ldots, V_N\},\ \{P_1^*, \ldots, P_N^*\},\ liste\_CDP)$;
17 $\quad\quad B \leftarrow B + 1$;

return $\dfrac{NbCertification}{nbBoucle}$;

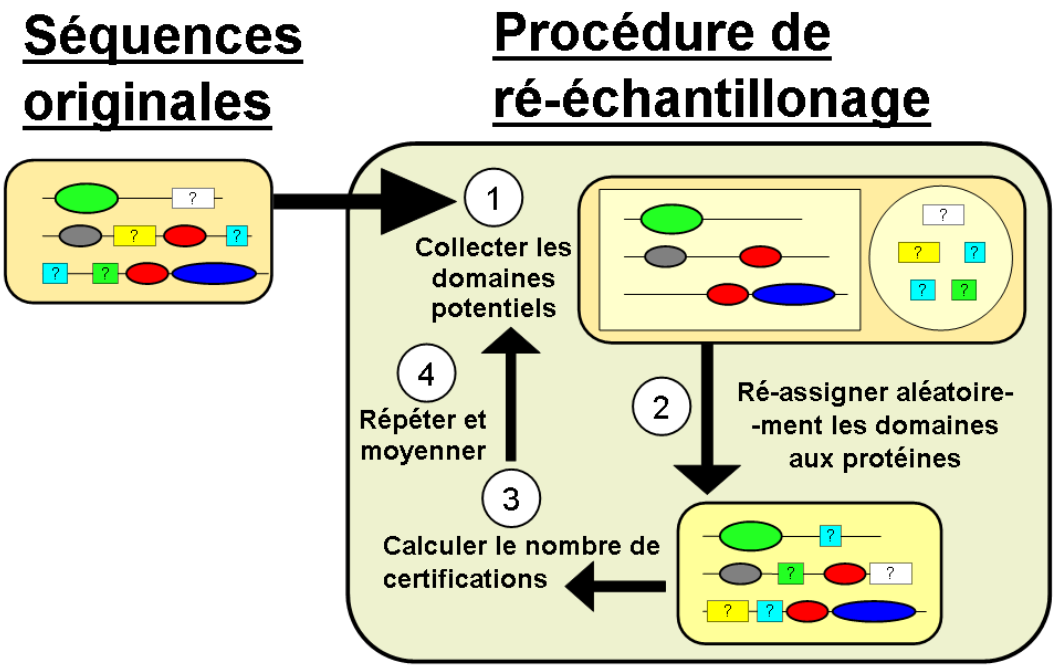


FIGURE 4.1 – **Principe de la procédure de ré-échantillonnage.**

4.3 Expérimentations

4.3.1 Simulations sur la levure

L'objectif de cette première expérience était de s'assurer de la capacité de la méthode à améliorer la sensibilité de la détection de domaines Pfam dans les protéines divergentes. Le protocole de cette expérience se divise en trois étapes :

a) Déterminer l'ensemble des domaines de référence : Dans un premier temps, les HMM de Pfam sont utilisés avec leurs seuils de score pour déterminer l'ensemble des domaines de référence chez la levure *S. cerevisiae* (protéines extraites de la *Saccharomyces Genome Database* ou SGD (Cherry *et al.*, 1998)). Seules les protéines pour lesquelles au moins deux domaines Pfam distincts ont été identifiés sont considérées dans les étapes suivantes.

b) Simuler l'évolution des séquences :

L'étape suivante consiste à simuler l'évolution des protéines afin de modifier leur composition globale en acides aminés pour la rapprocher de celle de *P. falciparum*. Le programme `seqgen` (Rambaut et Grassly, 1997) a été utilisé avec la matrice de taux d'échanges instantanés *WAG* (Whelan et Goldman, 2001), mais en remplaçant la composition en acides aminés standard par celle mesurée chez *P. falciparum* (PlasmoDB 5.5). Par conséquent, à partir de n'importe quelle séquence, en appliquant les substitutions selon la matrice modifiée, on obtient une protéine artificielle dont la composition en acides aminés converge vers celle de *P. falciparum*. En appliquant différents taux de substitution par site — 0.1, 0.25, 0.5 et 0.75 — nous avons créé, à partir des séquences protéiques de la levure, quatre jeux de protéines

artificielles de divergence croissante.

c) Retrouver les domaines divergents par co-occurrence : Enfin, dans la dernière étape de cette expérience, on applique aux quatre ensembles de protéines divergentes la procédure suivante. Chaque HMM est utilisé avec son seuil de score Pfam pour déterminer les ensembles de domaines validants. On s'attend à ce qu'un certain nombre de domaines de référence ne soient plus détectés à cause de la divergence des séquences. Les seuils de Pfam sont alors relâchés à une E-valeur de 10 pour déterminer les ensembles de domaines potentiels et la méthode de certification par co-occurrence est appliquée. On espère ainsi retrouver une partie des domaines précédemment perdus.

Taux subst.	Dom. de référence	Dom. perdus	Potentiellement retrouvables	Domaines retrouvés	*FDR* Estimé	Domaines inédits	Proportion GO connu
0.1	2 407	149	145	134	11.5%	274	97/130
0.25	2 407	346	301	265	9.2%	171	72/93
0.5	2 407	907	645	491	5.4%	60	20/31
0.75	2 407	1 436	747	501	4%	12	7/12

TABLE 4.2 – **Résultats sur la levure après dérive des séquences.** "Taux subst." indique le taux de divergence des séquences, "Dom. de référence" les domaines des protéines multidomaines de la levure originale, "Dom. perdus" correspond aux domaines non retrouvés par les seuils de Pfam sur les séquences divergentes, "Potentiellement retrouvables" indique le nombre de domaines que l'on peut espérer retrouver (c.-à-d. des domaines perdus dans une protéines où au moins un autre domaine est retrouvé par les seuils de Pfam), "Domaines retrouvés" indique les domaines perdus que l'on retrouve par notre méthode de certification, "Domaines inédits" est le nombre de domaines inédits à l'ensemble de référence trouvé en plus par notre méthode, et "Proportion GO connu" indique la proportion de domaines inédits annotés par des annotations déjà connues dans les protéines correspondantes.

Le tableau 4.2 récapitule les résultats de cette expérience. Comme attendu, plus la divergence des séquences est importante, plus les seuils de Pfam se révèlent dans l'incapacité de retrouver certains domaines de référence. Par exemple, pour un taux de substitution de 0.5, 907 domaines sont perdus, soit environ un tiers des domaines de référence. On note que parmi ces 907 domaines, 645 sont potentiellement retrouvables (c.-à-d. sont présents dans une protéine où au moins un autre domaine est encore détecté par les seuils de Pfam), et 491 sont retrouvés par notre méthode. Ainsi, pour un taux de substitution de 0.5, ∼76% des domaines que l'on peut espérer retrouver sont effectivement certifiés, c.-à-d. ∼54% du nombre total de domaines perdus. De plus, 60 domaines inédits (absents des domaines de référence) sont également détectés malgré un faible *FDR* de 5.4%. Ce nombre de nouveaux domaines est encore plus important pour des taux de substitution moins élevé, et peut paraître étonnamment haut pour un organisme aussi bien annoté que la levure. Cela pose la question de la validité de ces nouveaux domaines. Répondre à cette question n'est pas une tâche aisée. Une solution est de se référer aux annotations GO associées aux domaines. En effet, il semble raisonnable de supposer que si les annotations associées aux nouveaux domaines découverts concordent

avec l'annotation de la protéine alors la présence de ces domaines est vraisemblable. Dans la dernière colonne du tableau 4.2, est reportée la proportion de domaines possédant une annotation concordante avec la protéine, parmi les domaines inédits annotés dans la GO. Par exemple pour un taux de substitution de 0.1, des 274 domaines inédits, 130 possèdent une annotation GO parmi lesquels 97 (soit 75%) possèdent une annotation déjà connue dans la protéine. Cette forte proportion suggère qu'une grande partie de ces nouveaux domaines ne seraient pas des faux positifs, mais des domaines réellement présents chez la levure découverts grâce à notre approche.

4.3.2 Impact des paramètres utilisés pour la certification

La deuxième série d'expériences a été appliquée à *P. falciparum*. Elle avait pour but d'évaluer l'impact des paramètres (E-valeur seuil et P-valeur) sur le nombre de nouveaux domaines certifiés par la méthode et sur le *FDR*. Dans ces expériences, les domaines validants sont les domaines Pfam connus issus de la base de données PlasmoDB (version 5.5).

Les résultats présentés à la figure 4.2 montrent l'évolution du *FDR* en fonction de la P-valeur utilisée lors de la construction des CDP. Les courbes ont été réalisées pour différents ensembles de domaines potentiels correspondant à des seuils d'E-valeur fixés à 50, 10, 1 et 0.01. Comme attendu, plus la P-valeur est conservative, plus le FDR associé aux prédictions est faible. Même pour les E-valeurs les plus hautes, une P-valeur de 10^{-3} permet une certification avec un FDR performant. Notons qu'une P-valeur moins conservatrice de 10^{-1} peut aussi être envisagée, puisque le gain en précision a principalement lieu entre 1 et 10^{-1}.

On s'intéresse ensuite à l'impact de l'E-valeur sur la certification. La figure 4.3 représente d'un côté l'évolution du nombre estimé de domaines certifiés sur les données originales et sous H_0 (à gauche), et de l'autre côté celle du *FDR* (à droite), lorsque l'on fait varier l'E-valeur déterminant les domaines potentiels. Ces courbes ont été réalisées pour deux seuils différents de P-valeur (10^{-1} en haut et 10^{-3} en bas). On constate que plus on élève le seuil d'E-valeur — c.-à-d. plus on augmente la taille de l'ensemble des domaines potentiels —, plus le nombre de certifications sur les données réelles et sous H_0 augmente. Cependant, le FDR est lui aussi plus élevé pour les plus grandes E-valeurs. On peut donc contrôler le FDR en calibrant le seuil d'E-valeur en fonction de ce que l'on souhaite privilégier (*FDR* faible ou plus grand nombre de nouveaux domaines).

Dans les sections suivantes, les résultats présentés chez *P. falciparum* et ses orthologues ont été obtenus pour une P-valeur de 10^{-2} et des E-valeurs correspondant à des FDR de 10% et 20%.

4.4 Annotations des protéines de *P. falciparum*

Nous avons ensuite appliqué la méthode de certification par co-occurrence à *P. falciparum* en utilisant les trois types de domaines validants présentés dans la section 4.1.2 : les

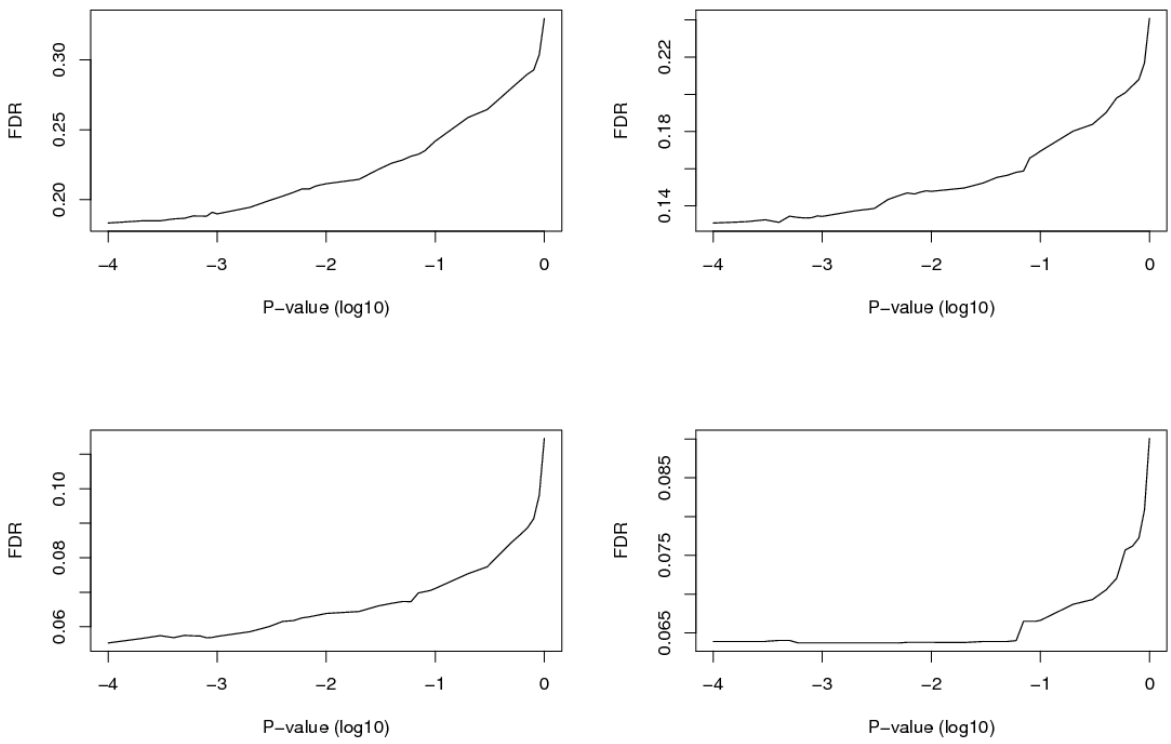


FIGURE 4.2 – **Évolution du *FDR* en fonction de la P-valeur**. Les différentes figures ont été réalisées pour des ensembles de domaines potentiels dont l'E-valeur est inférieure à 50 (en haut à gauche), à 10 (en haut à droite), à 1 (en bas à gauche) et à 0.01 (en bas à droite).

domaines Pfam connus, les domaines Interpro connus (à l'exclusion des domaines Pfam) et les domaines potentiels eux-mêmes. Les domaines Pfam et Interpro connus ont été obtenus à l'aide du logiciel `InterproScan` (Interpro version 19.0) appliqué aux protéines de *P. falciparum* (PlasmoDB version 5.5). Pour les ensembles de domaines potentiels, plusieurs seuils d'E-valeur ont été utilisés afin d'obtenir des prédictions de FDR différents.

4.4.1 Nouveaux domaines certifiés

Le tableau 4.3 récapitule l'ensemble des résultats obtenus pour des FDR inférieurs à 10% et 20% pour les trois différents types de domaines validants. Par exemple, pour un *FDR* inférieur à 20%, 585 nouveaux domaines sont certifiés par notre approche. Cela représente ∼16% des 3 683 domaines déjà connus dans les protéines de *P. falciparum* (une seule occurrence des domaines — nouveaux ou connus — par protéine est comptabilisée ici). Parmi ces domaines, 479 correspondent à une famille de domaines InterPro inédite dans la protéine. Les domaines Pfam connus ont permis de certifier 363 des 585 nouveaux domaines, les domaines

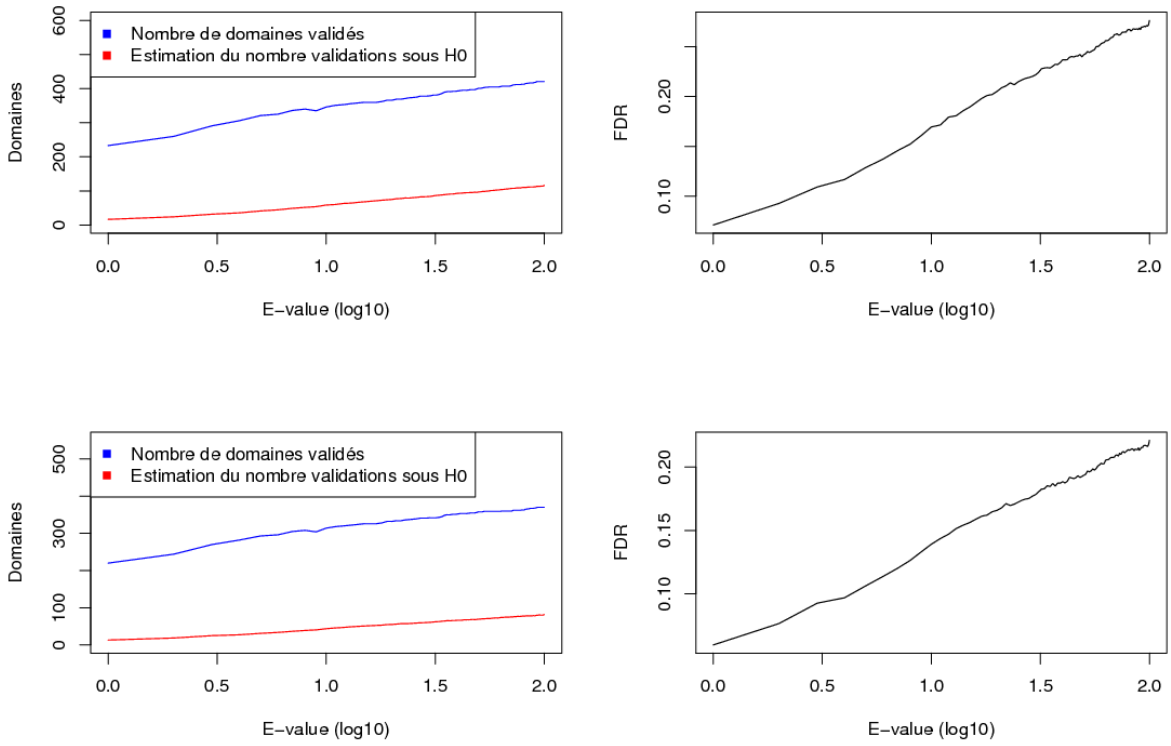


FIGURE 4.3 – **Évolution du nombre de certification et de l'estimation du nombre d'erreurs en fonction de la E-valeur.** Les courbes ont été réalisées pour des P-valeurs fixées à 10^{-1} (en haut à gauche) et 10^{-3} (en bas à gauche). On leur fait correspondre leurs *FDR* respectifs (figures de droite), obtenu par le ratio du nombre estimé d'erreurs sur le nombre de certification (soit courbes rouge sur bleue).

InterPro connus 395, et les domaines potentiels eux-mêmes 130 (avec du recouvrement, certains domaines étant certifiés par deux ou trois types de domaines validants). De plus, 159 nouveaux types de domaines ont été découverts — c.-à-d. qui n'avaient jamais été détectés dans une protéine de *P. falciparum* auparavant —, soit 11% du nombre total de types de domaines connus chez *P. falciparum* (*cf.* Table 3.1). La figure 4.4 présente le nombre de certifications réalisées par chaque ensemble de domaines validants en fonction du FDR. On peut voir que, pour un *FDR* donné, les domaines potentiels permettent de certifier moins de domaines que les deux autres types. Ce n'est pas une surprise étant donné que dans cet ensemble de domaines validants, un grand nombre de domaines sont vraisemblablement faux. Seules les plus faibles E-valeurs permettent d'obtenir de faibles *FDR*, ce qui implique des domaines potentiels-validants moins nombreux. L'utilisation des domaines Pfam connus comme domaines validants permet de certifier un plus grand nombre de domaines que les domaines

FDR	≤10%				≤20%			
Domaines validants	Pfam	Interp.	Pot.	Tous	Pfam	Interp.	Pot.	Tous
Domaines certifiés	259	185	47	340	363	395	130	585
Nvlles entrées InterPro	200	138	39	259	298	318	114	479
Familles inédites chez *Pf*	70	51	13	90	106	105	36	159

TABLE 4.3 – **Résultats de certification par co-occurrence sur *P. falciparum*.** "Domaines validants" indique l'ensemble de domaines validants utilisé pour la certification : "Pfam" les domaines Pfam connus d'après InterProScan ; "Interp." les domaines InterPro (non-Pfam) connus d'après InterProScan ; "Pot." les domaines potentiels eux-mêmes ; "Tous" pour les résultats combinés des trois types. "Domaines certifiés" dénote le nombre de nouveaux domaines certifiés, "Nvlles entrées InterPro" indique le nombre de domaines certifiés appartenant à une entrée InterPro inédite pour la protéine, et "Familles inédites chez *Pf*" le nombre de familles de domaines qui n'avaient jamais été observées auparavant dans une protéine de *P. falciparum*.

Interpro non-Pfam pour des FDR faibles, mais on observe que la tendance s'inverse lorsque le FDR augmente. Cet effet est vraisemblablement lié à l'hétérogénéité des schémas de domaines dans les différentes bases qui est compensé par le potentiel de certification supérieur des domaines Interpro lorsque le nombre de domaines potentiels augmente.

4.4.2 Conservation de la fonctionnalité des nouveaux domaines

Nous avons ensuite soulevé la difficile question de la conservation de la fonctionnalité dans les domaines divergents détectés par notre approche. Si la divergence est importante, quelles preuves avons-nous qu'ils accomplissent toujours la même fonction ? Pour essayer de répondre à cette question, nous avons tout d'abord fait le choix de rechercher uniquement des domaines Pfam complets, et non pas des fragments de domaines (HMM-*ls* et non pas HMM-*fs*, *cf.* section 2.4.2.b page 62). Ce premier point permet d'inférer une annotation fonctionnelle plus robuste. Nous avons ensuite regardé deux indicateurs :

a) Les régions de faible complexité : Comme discuté dans la section 3.3.2, les protéines de *P. falciparum* exhibent de nombreuses des régions de faible complexité. Ces régions peuvent être suspectées de modifier en priorité les parties non-fonctionnelles des protéines. Cependant, une comparaison de la proportion de régions de faible complexité dans les domaines certifiés et dans les domaines connus ne révèle pas de différence significative (*cf.* tableau 4.4).

b) La position des domaines : La position des domaines est également un indicateur d'évolution. Weiner *et al.* (2006) ont montré que les évènements de divergence de domaines (spécialement la perte de domaines due à la perte de fonctionnalité) ont principalement lieu aux extrémités des protéines, et en particulier en position C-terminale (à la fin de la séquence). Nous avons donc mesuré les distances séparant les extrémités des protéines des nouveaux domaines certifiés et des domaines connus (*cf.* figure 4.5). Encore une fois, aucun

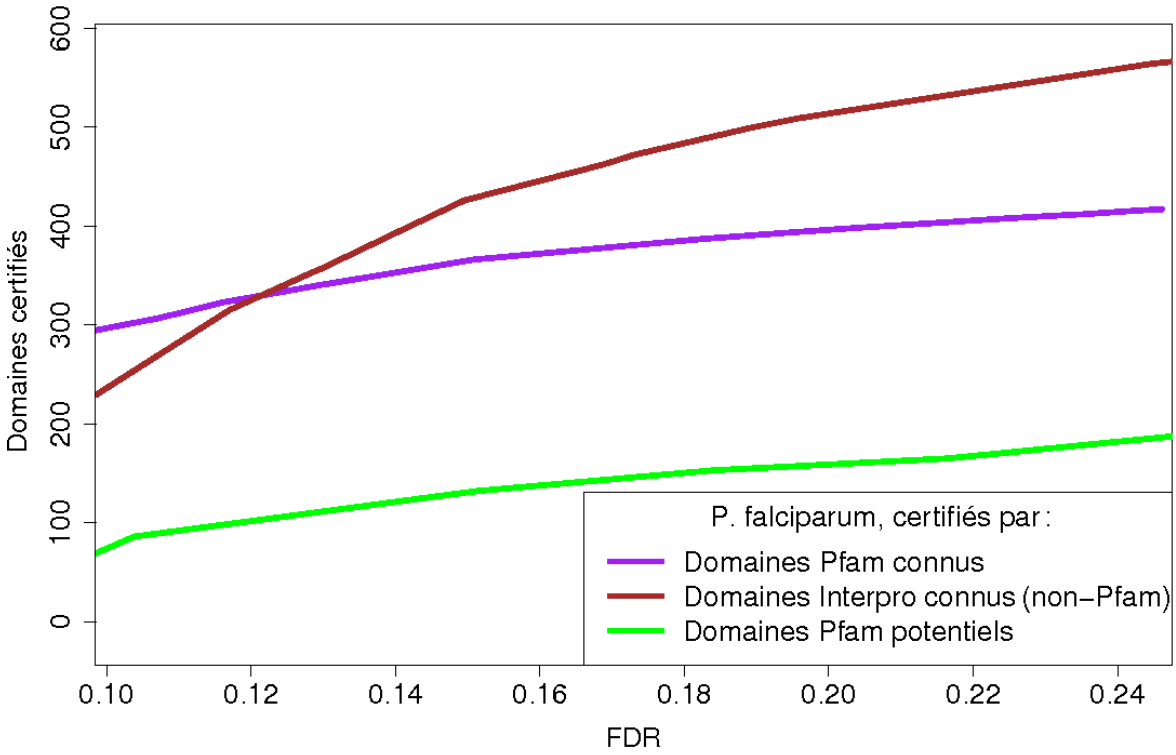


FIGURE 4.4 – **Nombre de certifications réalisées par les trois différents ensembles de domaines validants en fonction du FDR.** Nombre de certifications réalisées grâce aux domaines Pfam connus (courbe violette), aux domaines InterPro non-Pfam connus (courbe marron) et aux domaines Pfam potentiels (courbe verte).

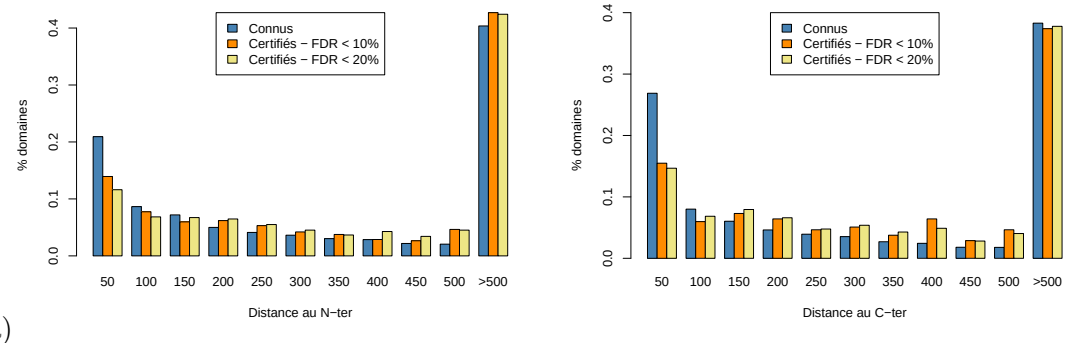
biais vers les extrémités de la protéine n'a été observé pour les nouveaux domaines certifiés par notre méthode. Ces domaines ne sont donc vraisemblablement pas plus consécutifs à des évènements de divergence que les domaines connus ne le sont.

4.4.3 Nouvelles annotations GO

Nous nous sommes ensuite intéressés aux annotations/termes GO qui peuvent être transférés aux protéines de *P. falciparum* grâce aux nouveaux domaines identifiés. Comme décrit précédemment (*cf.* section 1.4.4 page 45), Interpro et les différentes bases de domaines annotent les familles de domaines par des termes GO. La politique d'annotation consiste à associer à un domaine donné, les annotations partagées par l'ensemble des protéines annotées possédant ce domaine. Cette approche permet de transférer l'annotation du domaine à toute protéine où il est nouvellement identifié avec un risque d'erreur minime. En étendant cette politique d'an-

	Domaines dans une région de faible complexité	# moyen de résidus dans une région de faible complexité
Domaines connus	3 811/5 782 (66%)	30
Domaines certifiés avec $FDR \leq 10\%$	292/452 (64%)	30
Domaines certifiés avec $FDR \leq 20\%$	454/818 (56%)	26

TABLE 4.4 – **Proportion de domaines dans des zones de faible complexité.** Ce tableau rapporte la proportion de domaines Pfam connus et certifiés dans les protéines de *P. falciparum*, qui recouvrent une zone de faible complexité (de longueur > 10 résidus) (PlasmoDB 5.5), et le nombre moyen de positions des domaines qui se superposent à une région de faible complexité. Par exemple, 3 811 des 5 782 domaines Pfam déjà connus chevauchent une zone de faible complexité d'au moins 10 acides aminés. De plus, en moyenne, ces régions de faible complexité concernent 30 acides aminés de ces domaines.



(a) (b)

FIGURE 4.5 – **Histogramme des distances entre les domaines et les extrémités des protéines.** Les distances, en abscisse, sont exprimées en nombre de résidus. En ordonnée, on trouve la densité de domaines pour chaque tranche de distance aux extrémités N- et C-terminales (figure (a) et (b)) de la protéine. En bleu : les domaines Pfam connus ; en orange : les nouveaux domaines certifiés avec un $FDR \leq 10\%$; en jaune : les nouveaux domaines certifiés avec un $FDR \leq 20\%$.

notation aux combinaisons de domaines, comme décrit par Forslund et Sonnhammer (2008), de nombreux termes additionnels peuvent être déduits et tranférés aux protéines à partir de combinaisons de deux domaines et plus. À cette fin, nous avons énuméré toutes les combinaisons de domaines Pfam dans les protéines de la base de données Swiss-Prot, et identifié, pour chaque combinaison, les termes GO partagés par toutes les protéines annotées exhibant cette combinaison (seules les combinaisons observées dans au moins 10 protéines sont considérées ici). Au total, 2 235 combinaisons de domaines Pfam ont pu être associées avec un ou plusieurs termes GO n'appartenant pas à l'annotation individuelle des domaines : 2 115 paires de domaines, 119 triplets de domaines et 1 quadruplet. Les associations entre ces combinaisons

FDR	Combin. dom. connus	Dom. certif. seuls	Combin. avec dom. certifiés	Total prot.	Prot. non-annotées
≤10%	122	128	74	194	20
≤20%	122	273	114	267	39

TABLE 4.5 – **Nouvelles annotations GO des protéines de *P. falciparum*.** "Combin. dom. connus" est le nombre d'annotations GO qui peuvent être déduites des combinaisons de domaines Pfam avérés ; "Dom. certif. seuls" est le nombre d'annotations GO inédites distinctes que l'on peut transférer à la protéine d'après les termes GO associés par Interpro (*cf.* section 1.4.4 page 45) aux familles de domaines certifiés ; "Combin. avec dom. certifiés" est le nombre d'annotations GO supplémentaires (différentes de celles des deux précédentes colonnes) qui peuvent être déduites des combinaisons impliquant un nouveau domaine certifié. "Total prot." est le nombre total de protéines concernées par les trois précédentes sources d'annotations, et "Prot. non-annotées" est le nombre de protéines sans aucune annotation GO pour lesquelles une annotation par au moins un terme GO a été proposée.

et les termes GO sont disponibles en ligne [2]. Au final, les domaines seuls et les combinaisons de domaines ont permis d'améliorer l'annotation de nombreuses protéines de *P. falciparum*.

Le tableau 4.5 récapitule les résultats. Par exemple, pour un $FDR \leq 20\%$, les nouveaux domaines certifiés permettent d'ajouter de $273 + 114 = 387$ nouvelles annotations GO, ce qui représente 6% des 5 791 annotations GO connues dans cet organisme (dans les trois ontologies confondues). De plus, 39 de ces protéines ne possédaient aucune annotation GO auparavant. Enfin, à côté de ces résultats, on notera que 122 nouvelles annotations GO supplémentaires ont pu être déduites en combinant simplement les domaines Pfam déjà connus chez *P. falciparum*.

4.5 Caractérisation des résultats obtenus sur *P. falciparum*

Dans un premier temps, nous avons pu observer que les familles de domaines que l'on certifie sont majoritairement des domaines courts. La figure 4.6 illustre la longueur des domaines connus chez *P. falciparum* et des domaines certifiés par notre approche. Cette tendance s'explique par le fait que la divergence des séquences masque plus fortement l'homologie dans des séquences courtes que dans les plus longues.

Nous avons ensuite mené, avec l'aide d'Éric Maréchal, du Laboratoire PCV (CEA de Grenoble), une analyse détaillée des résultats obtenus par notre méthode sur les protéines de *P. falciparum*. L'information fonctionnelle induite par la certification de nouveaux domaines Pfam, parfois confirmée par des annotations GO, montre un important niveau de consistance biologique.

2. http ://www.atgc-montpellier.fr/EuPathDomains/supp.php

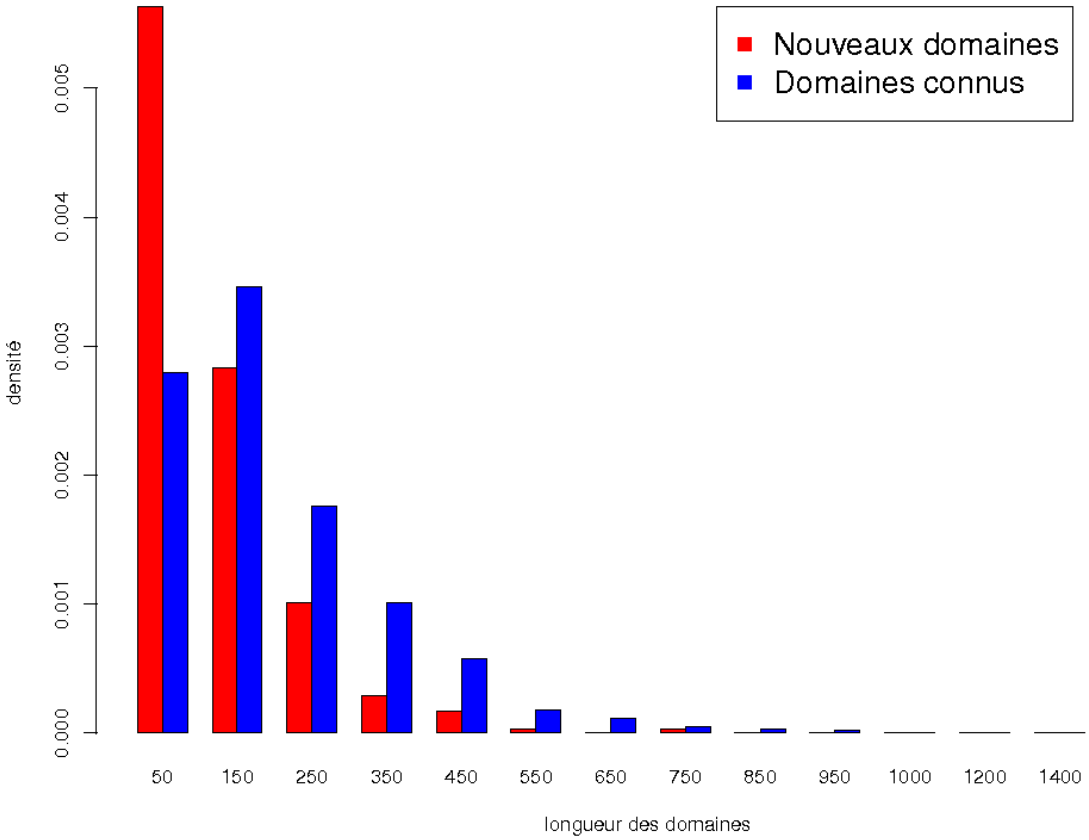


FIGURE 4.6 – **Histogramme de la longueur des domaines connus et certifiés chez *P. falciparum*.** Par longueur des domaines, on entend le nombre d'états *Match* du HMM modélisant la famille de domaine. Les occurrences de domaines Pfam connues sont représentées en bleu et les nouveaux domaines certifiés avec un $FDR \leq 10\%$ en rouge.

4.5.1 Protéines précédemment annotées

a) Confirmation de l'annotation : Dans les séquences de *P. falciparum* déjà annotées, les nouveaux domaines découverts, et les annotations GO qu'ils apportent, confirment toujours les fonctions initialement inférées. Par exemple, l'annotation de la protéine PF10_0122 comme phosphoglucomutase putative, initialement due à la présence des domaines InterPro SSF53738 et G3DSA:3.40.120.10 (nommés "A-D-PHexomutase_a/b/a-I/II/III"), et du domaine PF02878 ("PGM_PMM_I"), est maintenant abondamment soutenue par la certification des domaines Pfam PF02879 ("PGM_PMM_II"), PF02880 ("PGM_PMM_III") et PF00408 ("PGM_PMM_IV"). Globalement, aucune contradiction avec une précédente annotation n'a été mise en évidence dans l'ensemble de nos prédictions.

b) Raffinement de l'annotation : De plus, quelques annotations peuvent être substantiellement enrichies/raffinées. La protéine MAL13P1.83 a été décrite comme "karyophérine"

dans PlasmoDB (version5.5) d'après des alignements de séquences, et contient un seul domaine ayant une annotation GO — le domaine InterPro SSF48371 ("ARM-type_fold"), associé au terme GO:0005488, correspondant à "*binding*" (ontologie *Molecular function*). Ce terme GO est peu informatif car il définit une "interaction avec un ou plusieurs sites spécifiques d'une autre molécule". Cette protéine contient aussi le domaine Pfam PF08389 ("Xpo1") qui est lui aussi peu informatif et ne possède pas d'annotation GO. Cependant, en s'appuyant sur celui-ci, nous avons pu certifié la présence du domaine co-occurrent PF03810 ("IBN_N") dont l'annotation dans la *Gene Ontology* est riche, c.-à-d. GO:0008565, "*protein transporter activity*" ; GO:0000059, "*protein import into nucleus, docking*" ; GO:0006886 ; "*intracellular protein transport*" ; GO:0005634, "*nucleus*" ; GO:0005643, "*nuclear pore*" ; et GO:0005737, "*cytoplasm*". Donc la karyophérine de *P. falciparum*, un transporteur de l'enveloppe nucléaire, est maintenant riche de détails au niveau de l'annotation des domaines et de la GO, ce qui permet de retrouver cette séquence plus facilement lors d'une recherche basée sur les termes GO et la description des domaines. De même, MAL7P1.91 est annotée comme une "serine/thréonine kinase exportée" en se basant sur l'occurrence des domaines PS50011, SSF56112 et PF00069 qui soutiennent une activité de protéine kinase. Une fonction moléculaire plus précise peut lui être assignée : une implication dans l'ubiquitination de protéines. Cette prédiction se base sur la certification du domaine PF04564 ("U-box") — co-occurrent avec les domaines kynases PS5001 et PF00069 — qui apporte les nouvelles annotations : GO:0000151, "*ubiquitin ligase complex*" ; GO:0004842, "*ubiquitin-protein ligase activity*" ; et GO:0016567, "*protein ubiquitination*".

c) Inférence de nouvelles annotations : Dans certains cas, de nouvelles fonctions insoupçonnées peuvent aussi être avancées pour certaines protéines précédemment annotées. Prenons par exemple MAL7P1.12, connue pour être un antigène associé à la membrane de l'érythrocyte (Kun *et al.*, 1991). Elle était initialement composée de trois domaines InterPro sans annotation GO (SSFS2540, SM00487 et G3DSA:3.40.50.300), et aucun domaine Pfam n'avait été détecté. En s'appuyant sur la co-occurrence avec SM00487, nous certifions le domaine Pfam PF00035 ("dsrm") qui est notamment annoté pas le terme GO "*binding to double stranded RNA*" (GO:0003725). Puis, par co-occurrence avec le domaine "dsrm", nous pouvons certifier la présence du domaine Pfam PF04851 ("ResIII") annoté par les termes GO suivants : GO:0003677, "*DNA binding*" ; GO:0005524, "*ATP binding*" ; et GO:0016787, "*hydrolase activity*". La certification de ces nouveaux domaines suggère que MAL7P1.12 pourrait être impliquée dans quelques divers mais essentiels processus cellulaires régulés par de l'ARN double-brin (*dsRNA*) et/ou un mécanisme de défense contre des agents pathogènes impliquant une dégradation des ARN, ou encore un contrôle des niveaux d'ARN de la cellule du parasite au cours de son cycle de vie (Saunders et Barber, 2003). La certification conjointe d'un domaine fixant l'ARN double brin ("dsrm") et d'un domaine impliqué dans des activités d'hydrolase ("ResIII") suggère que MAL7P1.12 pourrait découper les *dsRNA* en fragments plus petits, générant peut être des ARN courts interférants et/ou des micro ARN (*miRNA*). Cela suggère que MAL7P1.12 pourrait accomplir une part de fonction moléculaire correspondant aux protéines eucaryotes *Dicer* (Jaskiewicz et Filipowicz, 2008), qui génèrent des *siRNA* et

des *miRNA* et aident à charger ces fragments dans le *RNA-induced silencing complex* (*RISC*). Aucune protéine *Dicer* ni *RISC* n'a pourtant été détectée chez *P. falciparum* jusque là, et une récente étude de Xue *et al.* (2008) soutient qu'aucun *miRNA* n'est produit dans ce parasite. MAL7P1.12 pourrait alors avoir un rôle proche mais distinct, découpant peut être certains *dsRNA* spécifiques. Étant donné les opinions actuelles concernant la régulation de l'expression du génome de *P. falciparum* au niveau des ARN (phénomènes d'épigénétique incluant un potentiel *RNA decay*), plutôt qu'au niveau transcriptionnel, la certification des domaines Pfam PF00035 et PF04851 dans un antigène associé à la membrane de l'érythrocyte soulève donc des questions concernant la fonction de cette protéine qui devrait maintenant être examinée précautionneusement dans le contexte du contrôle de l'ARN.

4.5.2 Protéines précédemment non-annotées (*unknown function*)

Pour les séquences de *Plasmodium falciparum* listées comme "hypothétique" dans PlasmoDB, ou pauvrement annotées, deux cas principaux sont observés en fonction da la qualité de la description et de l'annotation des nouvelles familles de domaines certifiés.

a) Nouveaux domaines ne possédant pas ou peu d'annotations GO informatives ni d'informations précises dans leur description : Dans ce cas, aucune annotation fonctionnelle précise ne peut être déduite du présent travail, mais la catégorisation structurelle de ces protéines est affinée grâce à l'identification des domaines et pourra servir de preuve lors d'inférences fonctionnelles futures. Par exemple le domaine Pfam WD40 (PF00400) apparaît chez tous les eucaryotes dans des protéines impliquées dans une grande variété de fonctions (du signal de transduction et de la régulation transcriptionnelle, au contrôle du cycle cellulaire et à l'apoptose). Initialement reporté dans 63 protéines de *P. falciparum*, le domaine WD40 a été certifié dans 12 protéines supplémentaires avec un FDR inférieur à 20%. La famille des protéines contenant un domaine WD40 serait donc la troisième plus grande famille du génome de *P. falciparum*. De plus, de nouvelles occurrences de ce domaines ont été certifiées dans 41 des 63 protéines protéines où il était connu. En s'appuyant sur la description de Pfam, les répétitions du domaines WD40 jouent le rôle de site pour l'interaction protéine-protéine et peuvent servir de plateforme pour l'assemblage de complexes protéiques ou de médiateurs d'interaction transitoire. La spécificité fonctionnelles de ces protéines se détermine par la séquence et les domaines se trouvant autour de ces répétitions. La combinaison de WD40 avec les autres domaines permet de classifier les protéines correspondantes, et pourrait faciliter l'annotation de cette famille pour de futurs travaux. Par exemple, il est possible de définir les combinaisons de domaines spécifiques dans les protéines "hypothétiques", comme la combinaison LisH/WD40 observée dans les protéines MAL13P1.54, PFE0540w, PFE0930w et PFE0930w, qui diffère de la combinaison LisH/RanBPM des protéines MAL13P1.182 et MAL13P1.308. De même, le domaine DEAD (PF00270), initialement reporté dans 42 protéines, est certifié dans 11 nouvelles protéines. L'occurrence d'un domaine DEAD indique que la protéine est vraisemblablement impliquée dans divers aspects du métabolisme de l'ARN, incluant la transcription nucléaire, l'épissage de pré-mRNA, la biogenèse du ribosome, le transport nuléoplasmique, la translation, le

RNA decay et l'expression des gènes de l'organelle. La fonction précise des protéines ne peut donc être déduite de la détection seule du domaine DEAD, cependant sa combinaison avec d'autres domaines sera une information essentielle pour les analyses futures de ces protéines.

b) Nouveaux domaines informatifs : De nombreux nouveaux domaines certifiés sont assez informatifs pour permettre de proposer une annotation fonctionnelle à des protéines "hypothétiques" ou faiblement annotées. On peut distinguer trois cas, en fonction de la connaissance antérieure de domaines dans la protéine ainsi qu'en fonction de l'annotation de la protéine et de ces domaines connus :

- **Protéines n'ayant aucun domaine InterPro ou Pfam connu :** Lorsque la fonction d'une protéine n'a pas encore été caractérisée par des expérimentations biologiques, l'unique source d'annotation dont on dispose relève de l'annotation informatique. Dans ce cadre, la plupart des protéines sont annotées uniquement à partir des domaines qu'elles contiennent. Par conséquent, l'absence de domaines connus traduit souvent une ignorance profonde de la fonction de la protéine, d'où l'importance de nos résultats dans ce genre de protéines. Par exemple, PF14_0380, qui est enregistré comme une protéine "hypothétique" dans PlasmoDB, est vraisemblablement une sous-unité du complexe de cohésion impliqué dans la ségrégation des chromatides. Cette prédiction se base sur la découverte des domaines Pfam PF04824 ("Rad21_Rec8") et PF04825 ("Rad21_Rec8_N"). Un autre exemple, les protéines PFF0910c, MAL13P1.78 et PFF1045w seraient des protéines kinases : certification du domaine PF08373 et du domaine PF06743, porteur notamment de l'annotation GO:0004672, "*protein kinase activity*".
- **Protéines annotées "*unknown function*" malgré la connaissance d'un domaine Pfam informatif :** Plusieurs protéines possèdent un domaine Pfam avéré qui n'a pas été pris en compte dans leur annotation fonctionnelle. Dans ces cas, les domaines que nous découvrons confortent le domaine précédemment identifié, et permettent d'inférer avec un certaine confiance une annotation fonctionnelle aux protéines concernées. C'est par exemple le cas de PFF1490w où le domaine PF02882 ("THF_DHG_CYG_C") est détecté par les seuils de Pfam et permet la certification du domaine PF00763 ("THF_DHG_CYG") suggérant que PFF1490w est une tetrahydrofolate dehydrogenase/cyclohydrolase, une enzyme du métabolisme du folate. Ou encore la protéine PF14_0052 où la certification du domaine PF07683 ("CobW_C") confirme la présence du domaine PF02492 ("CobW") et indique que cette protéine serait impliquée dans la synthèse de cobalamine (vitamine B12), une molécule nécessaire au développement du parasite en faible quantité mais exhibant des propriétés antipaludiques en quantité plus importante (Chemaly *et al.*, 2007).
- **Protéines où les domaines connus ne sont pas informatifs :** Dans ce cas, notre approche permet souvent la certification de domaines informatifs en s'appuyant sur la présence de ces domaines connus non-informatifs. On peut citer quelques exemples : PF10_0040 est vraisemblablement une nucléase impliquée dans la réparation de l'ADN (certification des domaines PF00867 et PF00752) ; PF10_0152 une nucléotidyltrans-

férase (domaines PF01909 et PF03828) ; PF11_0244 une protéase ATP-dépendante (domaine PF02190) ; PF11_0276 une lipase (domaine PF00561) ; PF11_0368 un transporteur de l'enveloppe nucléaire (domaine PF03810) ; PF11_0375 une protéine liée aux particules de reconnaissance de signaux (domaine PF08492) ; PF11_0469 une ARN polymérase ADN-dépendante (domaines PF08221 et PF05645) ; PF14_0031 une protéine fixant le Ca^{2+} impliquée dans le trafic vésiculaire (domaines PF00637, PF0008 et PF07645) ; et pour finir un domaine qui semble très intéressant certifié dans la protéine PF14_0479 est le domaine PF05605 ("Di19"). Ce domaine, spécifique au règne végétal, est associé à la réponse des plantes à la sécheresse. Il a vraisemblablement été acquis par *P. falciparum* consécutivement à son endosymbiose secondaire d'un algue (à l'origine de la caractéristique végétale des apicomplexes — *cf.* section 3.2 page 78).

Toutes les annotations fonctionnelles qui peuvent être proposées n'ont pas été listées ici. Toutefois, nous avons pu constater que beaucoup d'entre elles ont révélé des protéines impliquées dans des interactions avec la chromatine (telles que PFF1385w, PFL0975w ou PF07_0106) et de nombreuses protéines associées à des facteurs de transcription. Ce dernier résultat est particulièrement intéressant, compte tenu de la faible quantité avérée de ce genre de protéines chez *P. falciparum* d'après les précédentes études (Coulson *et al.*, 2004; Callebaut *et al.*, 2005). Enfin, nous n'avons pas listé ici l'ensemble des protéines hypothétiques pour lesquelles une fonction a été inférée après le *workshop* de ré-annotation de *P. falciparum* de 2008 (co-organisé par le *Wellcome Trust Sanger Institute* (WTSI) et le consortium EuPathDB), car les domaines certifiés sont consistants avec ces annotations révisées.

4.5.3 Domaines connus et certifiés les plus abondants

Un point intéressant est que la méthode n'ajoute pas une seule séquence à la plus grande familles de protéines plasmodiales, c.-à-d. les protéines contenant des domaines Rifin_STEVOR, et très peu de changement pour à la seconde : la famille des protéines contenant des domaines Duffy-Binding et PFEMP. Ces domaines sont construits par Pfam, uniquement sur des antigènes de surface plasmodiaux dont les séquences sont donc proches de celles de *P. falciparum*. Il est cohérent que notre méthode ne permette pas d'en détecter plus que les seuils recommandés par Pfam car ces domaines ne sont pas divergents par rapport aux modèles et le seuil recommandé est donc adapté.

Les autres familles de protéines les plus fréquentes chez *P. falciparum* contiennent des domaines Pkinase ou des domaines liés à la fixation, la modification et/ou au traitement de l'ARN (par exemple Helicase_C, RRM, DEAD). Pour ces familles, plusieurs nouvelles protéines sont identifiées grâce à notre approche.

Toutefois, les domaines dont on certifie le plus de nouvelles occurrences semblent impliqués dans les interactions protéine-protéine. On note en particulier les familles de protéines contenant le domaine WD40 ainsi que celle possédant des domaines TPR_1 et TPR_2 (identifié respectivement dans 15 et 8 séquences initialement, et certifiées respectivement dans 13 et 16 nouvelles protéines). Notre travail permet donc une analyse en profondeur de ces familles de domaines, ainsi qu'un accroissement de la couverture et de la connaissance des compositions exactes en domaines Pfam des protéines de *P. falciparum*.

4.6 Consistance avec les orthologues de *P. vivax* et *P. yoelii*

Nous avons ensuite appliqué la procédure de certification de domaines aux protéines de *P. vivax* et *P. yoelii*. Ces deux espèces plasmodiales sont entièrement séquencées et disposent d'un assemblage de bonne qualité. *P. vivax* est le deuxième plus virulent des *Plasmodium* infectant l'Homme (Carlton *et al.*, 2008a). Quand à *P. yoelii*, ce parasite de rongeur (RMP) sert notamment de modèle lors d'expérimentations préliminaires de traitement potentiels (Carlton *et al.*, 2005). Ces deux espèces exhibent des séquences très divergentes et possèdent un nombre de domaines connus plus faible que celui de *P. falciparum* (bien que chez *P. vivax* il n'y ait pas de biais en A+T). Elles nous offrent donc un point de comparaison important pour les résultats obtenus chez *P. falciparum*.

Les tableaux 4.6 et 4.7 présentent les statistiques concernant, respectivement, le nombre de domaines certifiés et les annotations GO déduites chez *P. vivax* et *P. yoelii* (correspondant aux tableaux 4.3 et 4.5 chez *P. falciparum*). De plus, pour chaque ensemble de domaines validants, la figure 4.7 permet de comparer le nombre de domaines certifiés dans ces trois espèces en fonction du FDR associé aux certifications. Le nombre de domaines certifiés semble être légèrement plus élevé chez *P. falciparum*, particulièrement par rapport à *P. yoelii*. Cela peut s'expliquer par le fait que l'on dispose d'un plus grand nombre de domaines connus et potentiels chez *P. falciparum*. Cette limitation provient vraisemblablement pour *P. yoelii* d'erreurs d'assemblage (7 724 protéines recensées, contre à peine plus de 5 000 chez *P. falciparum* et *P. vivax*).

	FDR	≤ 10%				≤ 20%			
	Dom. validants	Pfam	Interp.	Pot.	All	Pfam	Interp.	Pot.	All
P. vivax	Dom. certifiés	279	76	65	348	343	253	101	517
	Nvlles entrées	227	46	56	274	290	274	89	417
	Familles inédites	77	26	22	94	106	101	32	150
P. yoelii	Dom. certifiés	233	140	42	298	289	329	123	485
	Nvlles entrées	195	98	32	236	249	267	106	406
	Familles inédites	66	35	10	77	87	103	35	144

TABLE 4.6 – **Nouveaux domaines certifiés dans les protéines de *P. vivax* et *P. yoelii*.** "Dom. validants" indique l'ensemble de domaines validants utilisé pour les certifications : "Pfam", les domaines Pfam connus d'après InterProScan ; "Interp.", les domaines InterPro (non-Pfam) connus d'après InterProScan ; "Pot.", les domaines potentiels eux-mêmes ; "All" indique les résultats obtenus en cumulant les trois types de domaines validants. "Dom. certifiés" dénote le nombre de nouveaux domaines certifiés, "Nvlles entrées" le nombre de domaines certifiés appartenant à une entrée InterPro inédite dans la protéine, et "Familles inédites" indique le nombre de familles de domaines qui n'avaient jamais été observées auparavant dans aucune protéine de l'organisme concerné.

	FDR	Domaines seuls	Combin. avec dom. certifié	Protéines non-annotées
P. vivax	$\leq 10\%$	144	119	37
	$\leq 20\%$	230	142	55
P. yoelii	$\leq 10\%$	122	99	28
	$\leq 20\%$	248	111	44

TABLE 4.7 – **Nouvelles annotations GO des protéines chez *P. vivax* et *P. yoelii*.** "Domaines seuls" est le nombre d'annotations GO inédites distinctes que l'on peut transférer à la protéine d'après les termes GO associés par Interpro (*cf.* section 1.4.4 page 45) aux familles de domaines certifié ; "Combin. avec dom. certifié" est le nombre d'annotations GO supplémentaires (différentes de la précédente colonne) qui peuvent être déduites des combinaisons impliquant un nouveau domaine certifié. "Protéines non-annotées" est le nombre de protéines sans aucune annotation GO pour lesquelles nous proposons une annotation par au moins un terme GO.

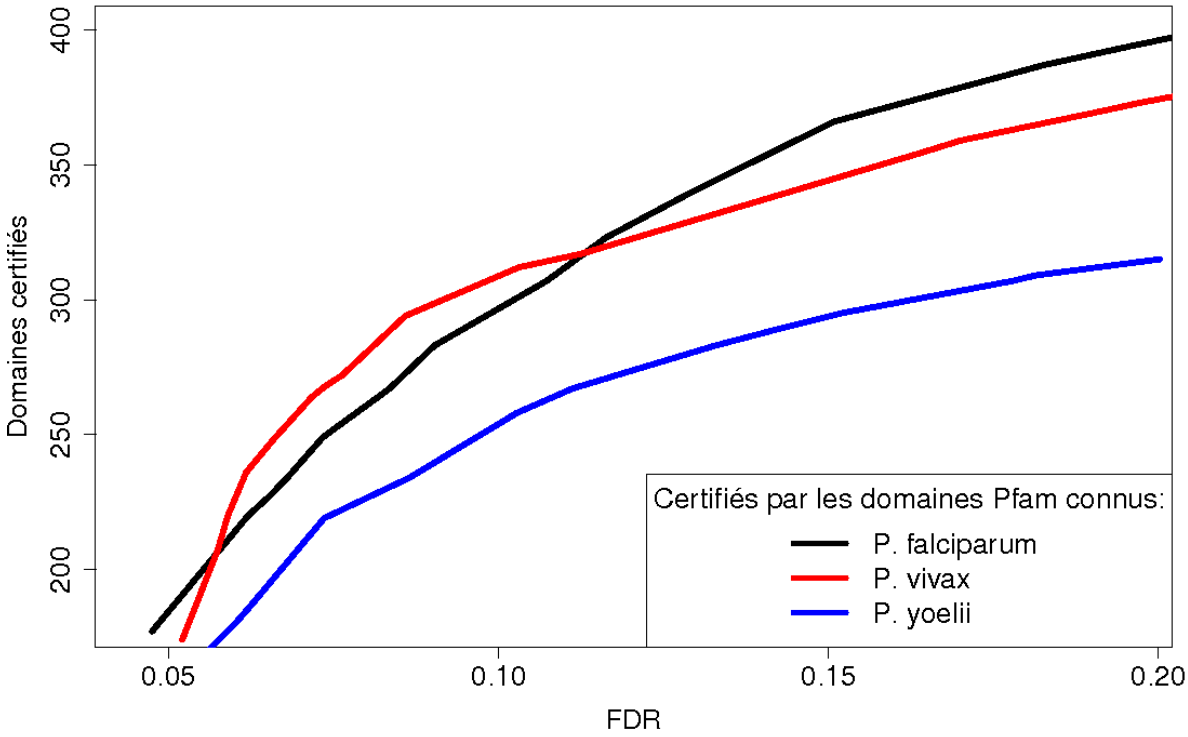
Un point intéressant vient de la comparaison des domaines certifiés dans les protéines orthologues aux protéines de *P. falciparum* chez *P. vivax* et *P. yoelii* (orthologies extraites de PlasmoDB 5.5). Les tableaux 4.8 et 4.9 révèlent qu'une grande partie des domaines certifiés dans une protéine de *P. falciparum* sont aussi certifiées dans une protéine orthologue dans ces espèces. De plus, une partie des domaines certifiés chez *P. falciparum* correspondent à des domaines déjà connus dans leurs protéines orthologues. Par exemple, parmi les nouveaux domaines certifiés avec un $FDR \leq 10\%$ dans les protéines de *P. falciparum* ayant une orthologue chez *P. vivax*, 14% sont des domaines déjà connus dans une protéine orthologue de *P. vivax*, et 63% ont été certifiés avec un FDR équivalent dans les protéines orthologues. Ce qui signifie que 77% des nouveaux domaines certifiés chez *P. falciparum*, sont aussi présents dans des protéines orthologues. De plus, l'observation d'autres paramètres tels que l'éloignement aux extrémités N- et C- terminales ainsi que la taille des domaines certifiés se révèle également congruente (données non-montrées). Tous ces résultats supportent fortement notre méthode et les résultats obtenus sur les protéines des espèces plasmodiales. Ils peuvent également être vus comme un troisième indicateur concernant la conservation de la fonctionnalité des domaines divergents.

		P. vivax		
		Dom. connus	Dom. certif.≤10%	Dom. certif.≤20%
P. falciparum	Dom. connus	2 985/3 143 (95%)	34/3 143 (1%)	55/3 143 (2%)
	Dom. certif.≤10%	46/323 (14%)	205/323 (63%)	222/323 (69%)
	Dom. certif.≤20%	54/548 (10%)	233/548 (43%)	277/548 (51%)

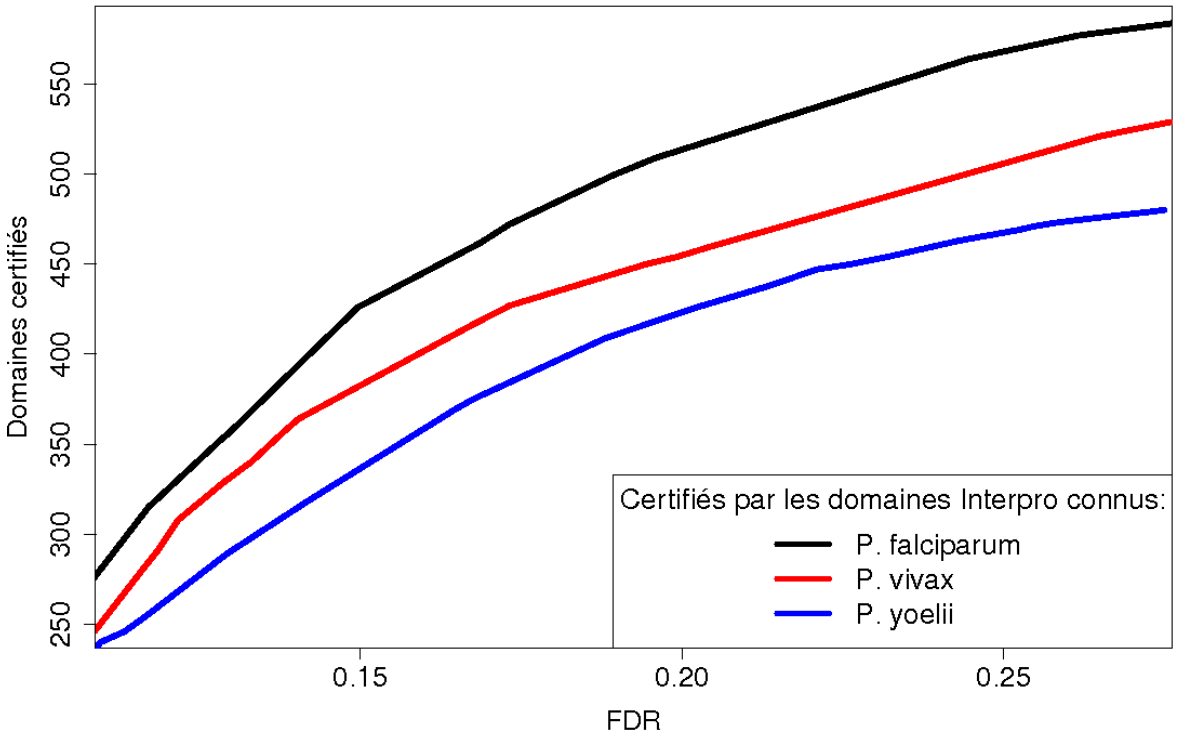
TABLE 4.8 – **Proportion de domaines connus et certifiés chez *P. falciparum* que l'on retrouve dans des protéines orthologues de *P. vivax*.** Ce tableau rapporte le nombre de domaines Pfam — connus et certifiés dans une protéine de *P. falciparum* — pour lesquels le même domaine Pfam est connu/certifié dans une protéine orthologue chez *P. vivax*. Ce nombre est comparé au nombre total de domaines dans les protéines *P. falciparum* qui possède une protéine orthologue chez *P. vivax*. Par exemple, 3 143 domaines Pfam sont connus dans les protéines de *P. falciparum* ayant un orthologue chez *P. vivax*. Parmi ces domaines, 2 985 sont connus et 34 sont certifiés avec un FDR inférieur à 10% dans la protéine orthologue de *P. vivax*. De plus, parmi les 323 nouveaux domaines certifiés chez *P. falciparum* avec un $FDR \leq 10\%$ (dans une protéine ayant un orthologue chez *P. vivax*), 46 sont connus et 205 sont certifiés avec un $FDR \leq 10\%$ dans l'orthologue de *P. vivax*.

		P. yoelii		
		Dom. connus	Dom. certif.≤10%	Dom. certif.≤20%
P. falciparum	Dom. connus	2 700/2 998 (90%)	42/2 998 (1%)	50/2 998 (2%)
	Dom. certif.≤10%	41/314 (13%)	158/314 (50%)	169/314 (54%)
	Dom. certif.≤20%	47/538 (9%)	185/538 (34%)	228/538 (42%)

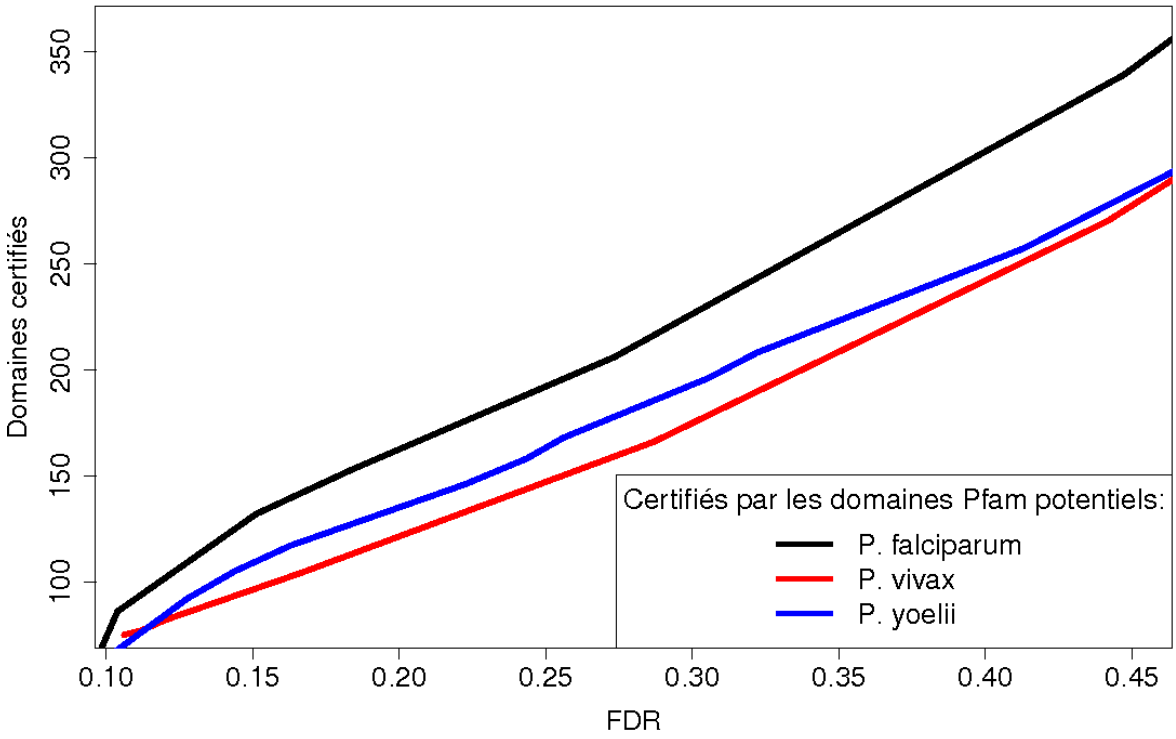
TABLE 4.9 – **Proportion de domaines connus et certifiés chez *P. falciparum* que l'on retrouve dans des protéines orthologues de *P. yoelii*.** Voir tableau 4.8 pour légende.


Domaines certifiés
200
250
300
350
400
0.05
0.10
0.15
0.20
FDR
Certifiés par les domaines Pfam connus:
P. falciparum
P. vivax
P. yoelii

(a)


Domaines certifiés
250
300
350
400
450
500
550
0.15
0.20
0.25
FDR
Certifiés par les domaines Interpro connus:
P. falciparum
P. vivax
P. yoelii

(b)



(c)

FIGURE 4.7 – **Nombre de certifications réalisées par les trois types de domaines validants pour *P. falciparum*, *P. vivax* et *P. yoelii*.** Nombre de domaines certifiés (en ordonnée) en fonction du FDR associé à ces certifications (en abscisse) La méthode a été appliquée en utilisant les domaines Pfam connus (a), les domaines InterPro connus (b), et les domaines potentiels Pfam eux-mêmes (c).

4.7 Comparaison aux travaux antérieurs

Affiner la détection de domaines est une tâche difficile. En pratique, les modèles de domaines sont construits pour assurer la présence d'un domaine grâce à un seuil de score expertisé à la main. Au-delà de ce seuil, l'absence de faux positifs n'est plus garantie. Afin d'améliorer la détection de domaines Pfam, nous avons proposé une méthode pour filtrer les faux positifs parmi les domaines détectés au-delà des seuils en exploitant l'information apportée par le contexte en domaines. À notre connaissance, seuls deux travaux ont été publiés sur cette problématique. Beaussart *et al.* (2007) ont développé un outil pour identifier de possibles artefacts d'annotations, notamment les domaines manquants ou une erreur sur la nature (famille) du domaine. Pour cela, ils créent des classes de protéines dont la composition en domaines est proche, et alignent les protéines de chaque classe en se basant sur l'ordre séquentiel des domaines. Ils peuvent alors détecter un domaine manquant ou discordant dans une protéine, par rapport à la composition en domaines des autres protéines de la classe. Cette stratégie peut se révéler efficace à condition que des protéines homologues soient déjà connues et correctement annotées. Coin *et al.* (2003) ont proposé une approche élégante pour augmenter la sensibilité de la détection de HMM de domaines en incorporant l'information du contexte en domaine des protéines dans le calcul du score des domaines. Plutôt que de détecter indépendamment chaque domaine d'une séquence protéique, les auteurs proposent une modélisation Markovienne qui permet la détection de la composition en domaines de la protéine. Avec cette modélisation, le score obtenu par un domaine à une position donnée dépend non seulement de la séquence d'acides aminés du domaine mais également du contexte, riche de la connaissance des autres domaines potentiels dans la protéine. Une comparaison précise des résultats obtenus par cette approche est difficile car les bases de données qu'ils utilisent – Pfam version 7.7 et Swiss-Prot40 – ont été amplement enrichies depuis sa publication. Pour cela, nous avons travaillé sur les mêmes versions des bases que les auteurs ont utilisées afin de proposer une comparaison objective. Nous avons réalisé l'apprentissage des CDP sur les données de Swiss-Prot40 et l'identification des domaines Pfam potentiels et connus *via* la version 7.7 de Pfam. On dénombre 491 protéines de *P. falciparum* dans Swiss-Prot40. Dans ces protéines, Coin *et al.* (2003) proposent 7 nouveaux domaines (dans 4 protéines) tandis que la méthode de certification identifie 14 nouveaux domaines avec un FDR inférieur à 10% et 12 domaines supplémentaires pour un FDR inférieur à 20% (concernant respectivement 12 et 21 protéines). Parmi les 7 domaines obtenus par Coin *et al.* (2003), deux sont également découvert par notre approche. En ce qui concerne les 5 autres domaines, une inspection approfondie des résultats révèle qu'il s'agit de répétitions de domaines dans la protéine et que, dans trois cas, seul cette famille de domaine est détectée dans la protéine concernée. En d'autres termes, la nouvelle occurrence du domaine est détectée par Coin *et al.* (2003) grâce à une autre occurrence de ce même domaine. Ce genre de détection n'est pas permise par notre approche car l'information apportée par une répétition de domaines est, à notre sens, très différente de celle apportée par l'appariement de famille de domaines distinctes. Enfin, on note que seule la méthode de certification pr co-occurrence conduit à l'identification de familles de domaines inédites à cette époque chez *P. falciparum* (4 avec un FDR inférieur à 10% et 9 supplémentaires à 20%).

La comparaison avec les travaux précédents tend à montrer que notre approche bénéficie de plusieurs avantages. Tout d'abord, c'est une approche simple et intuitive qui nécessite peu de temps de calcul et peut être appliquée à n'importe quel génome. Ensuite, chaque prédiction est justifiable en exhibant le domaine validant permettant la certification. Notre méthode échappe donc aux inconvénients du temps de calcul et de l'aspect boîte noire des modèles de Markov d'ordre n. De plus, nous pouvons exploiter l'information en domaines provenant des différentes bases de domaines d'Interpro. Enfin, l'aspect le plus important est que nous accompagnons nos certifications d'une mesure de confiance, grâce à notre procédure de ré-échantillonnage.

4.8 Perspectives

4.8.1 Améliorations de la méthode

Après la publications de l'article (Terrapon *et al.*, 2009), plusieurs directions ont été envisagées pour affiner les résultats de la méthode de certification par co-occurrence. La première s'axe sur la sélection des CDP. Un article récent concernant l'architecture en domaines des protéines multidomaines révèle que 90% des architectures sont spécifiques à l'un des trois domaines du Vivant (Lee et Lee, 2009). Cette observation s'explique principalement par l'extrême spécificité des domaines eux-mêmes. Il est donc naturel de faire l'hypothèse qu'un ensemble de CDP appris spécifiquement sur des séquences eucaryotes conduirait à de meilleurs résultats de certifications chez *P. falciparum*. Une partition des séquences d'Uniprot (ensemble d'apprentissage initial) a été réalisée pour distinguer les protéines eucaryotes des protéines non-eucaryotes. Nous avons alors appris des listes de CDP alternatives qui ont été appliquées pour la certification de domaines co-occurrents chez *P. falciparum*. La figure 4.8 montre les résultats obtenus par ces différents listes de CDP. On constate une amélioration des résultats grâce à l'ensemble des CDP eucaryote-spécifiques. On observe également une nette dégradation des résultats lorsque l'ensemble des CDP est inadapté (en terme de domaine du Vivant) à l'organisme étudié. Il est donc possible d'accroître la précision de la méthode en apprenant des CDP sur un ensembles de protéines plus restreint autour du voisinage phylogénétique de l'organisme cible.

Une deuxième piste en cours d'étude, concerne la mise en exergue de la quantité d'information lors des certifications. On souhaite mettre en avant les domaines qui sont certifiés par plusieurs domaines validants par rapport aux domaines certifiés par un unique domaine validant. On différencie donc plusieurs catégories dans nos certifications en fonction du nombre de domaines validants utilisés, et on espère accéder à de meilleurs taux d'erreur lorsque les domaines sont certifiés par plus d'un domaine validant et à des taux d'erreur plus importants dans le cas contraire. La question étant de savoir si cela permet d'obtenir un plus grand nombre de domaines certifiés et si les FDR associés reflètent mieux la confiance que l'on a en ces résultats.

Enfin, la dernière amélioration envisagée consiste à passer à la version 3.0 du programme HMMER. Cette version propose notamment le calcul de scores *forward*, mais ne permet pas d'interdire la recherche de fragments de domaines (*cf.* section 2.4.4). En contournant

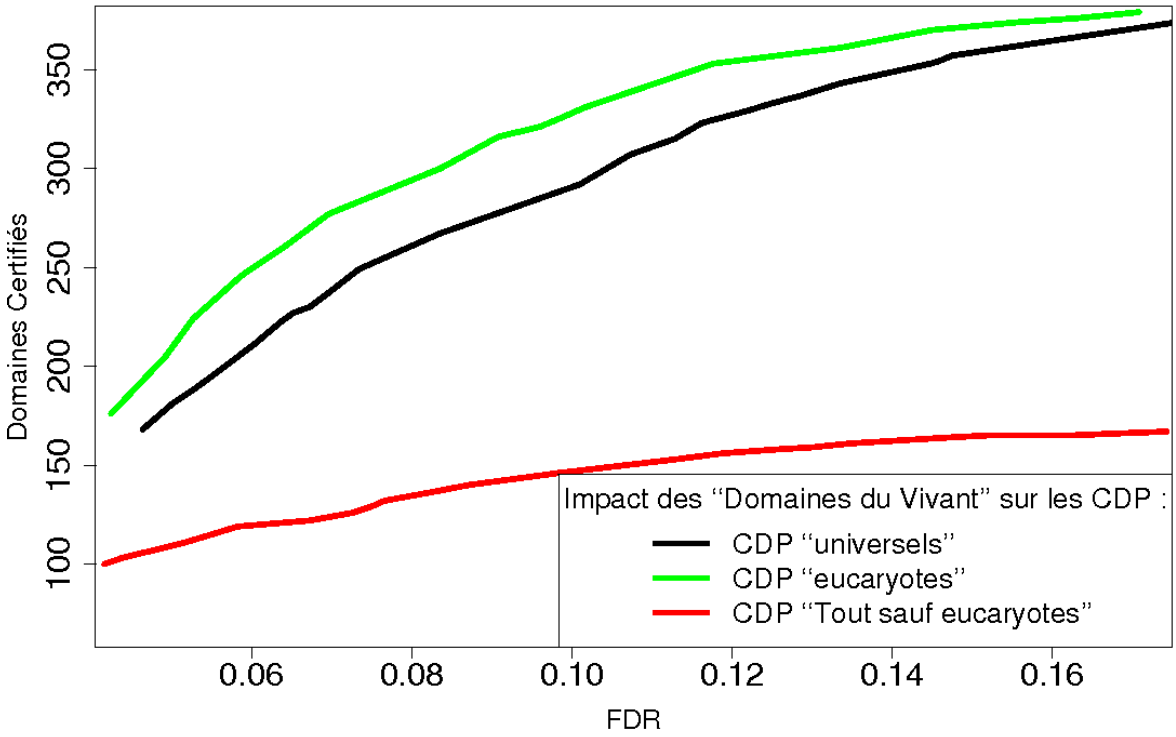


FIGURE 4.8 – **Certifications réalisées grâce à différentes listes de CDP**. La courbe noir correspond aux CDP appris sur l'ensemble des séquences d'Uniprot sans distinction. Les deux autres courbes sont obtenues en différenciant les séquences d'Uniprot en fonction de leur l'appartenance au domaine des eucaryotes (courbe verte) ou, *a contrario*, par leur appartenance à l'un des autres domaines du Vivant (courbe rouge).

ce défaut, l'approche de certification par co-occurrence pourrait bénéficier de recherches de domaines plus rapide (jusqu'à 100 fois) et d'ensemble de domaines potentiels contenant moins de faux positifs. Par conséquent, nos certifications devrait être réalisées avec des FDR plus performants.

4.8.2 Extension à d'autres organismes et présentation des résultats

Actuellement, la principale perspective concerne l'application de la méthode à un plus grand nombre d'organismes et notamment à l'ensemble des organismes modèles, ainsi que le développement d'une base de données pour contenir et interroger l'ensemble des domaines divergents identifiés dans ces espèces.

L'étude du premier organisme supplémentaire, *Arabidospsis thaliana*, fut suggérée par les membres du laboratoire de physiologie cellulaire végétale (LPCV) du CEA de Grenoble

lors d'un séminaire en 2009. Par la suite, la collaboration de l'équipe MAB avec l'institut Pasteur de Tunis nous a conduit à appliquer cette méthode aux espèces responsables de la leishmaniose : *Leishmania major*, *Leishmania infantum* et *Leishmania braziliensis* (*cf.* le site TritrypDB (Aslett *et al.*, 2010) qui s'intéresse à tous les types de trypanosomes). Après cela, nous avons étendu notre étude aux eucaryotes pathogènes chez l'Homme d'après la base de données EupathDB (Aurrecoechea *et al.*, 2007). En plus des espèces plasmodiales et des leishmanioses, la méthode a également été appliquée à deux autres apicomplexes *Toxoplasma gondii* et *Cryptosporidium parvum* (qui bénéficie chacun d'une base de données dédiée : respectivement ToxoDB (Gajria *et al.*, 2008) et CryptoDB (Heiges *et al.*, 2006)), un autre trypanosome *Trypanosoma brucei gambiense* (*cf.* TritrypDB) et un organisme appartenant un phylum distinct des précédents : *Giardia lamblia* (*cf.* GiardiaDB (Aurrecoechea *et al.*, 2009a)).

Le nombre d'organismes étudiés étant assez important, la construction d'une base de données accompagnée d'une interface Web dynamique pour remplacer le prototype (site Web statique) est devenue une nécessité. Un travail initié par des étudiants d'IUT Informatique a été récemment poursuivi par des étudiants en première année de Master. Nous avons intégré ces travaux pour obtenir une première version d'une base de données nommée EuPathDomains, actuellement disponible en ligne[3]. Un article vient d'être publié afin de présenter l'utilisation de cette base notamment dans le cadre des résultats obtenus pour les organismes précédemment cités (Ghouila *et al.*, 2010).

Depuis le portail EuPathDomains (capture d'écran en figure 4.9), on peut accéder aux résultats de la méthode de certification pour chacune des espèces précédemment citées. Les résultats peuvent être interrogées par les noms ou les identifiants des protéines, des domaines, des entrées Interpro et des termes GO. De plus, il est possible d'indiquer le seuil de FDR maximum requis pour la certification des nouveaux domaines. Enfin, on peut spécifier une espèce ou un taxon d'intérêt.

Suite à une requête concernant une protéine d'intérêt, on obtient une page similaire à celle représentée par la figure 4.10, qui correspond ici à la protéine PF11_0189 de *P. falciparum*. On trouve tout d'abord le nom de la protéine (sous la forme d'un lien vers le site PlasmoDB), son annotation actuelle, ainsi qu'une proposition de réannotation issue du *workshop* de 2008. Ensuite, nous présentons les annotations GO connues pour cette protéine et celles que nous avons déduites des combinaisons de ses domaines connus. À la suite de ces informations, deux tableaux représentent respectivement les domaines connus et certifiés de la protéine. On y trouve, pour chaque domaine :

1. son nom, qui est un lien vers la page des domaines ;
2. deux liens (sous forme d'images) respectivement vers la base de données de familles d'origine du domaine et vers Interpro ;
3. une représentation graphique linéaire de la protéine et de la localisation des différentes occurrences du domaine ;
4. les annotations GO associées à ce domaine par Interpro, et — pour les nouveaux domaines certifiés — celles que l'on a pu déduire des combinaisons avec d'autres domaines.

3. http ://www.atgc-montpellier.fr/EuPathDomains/

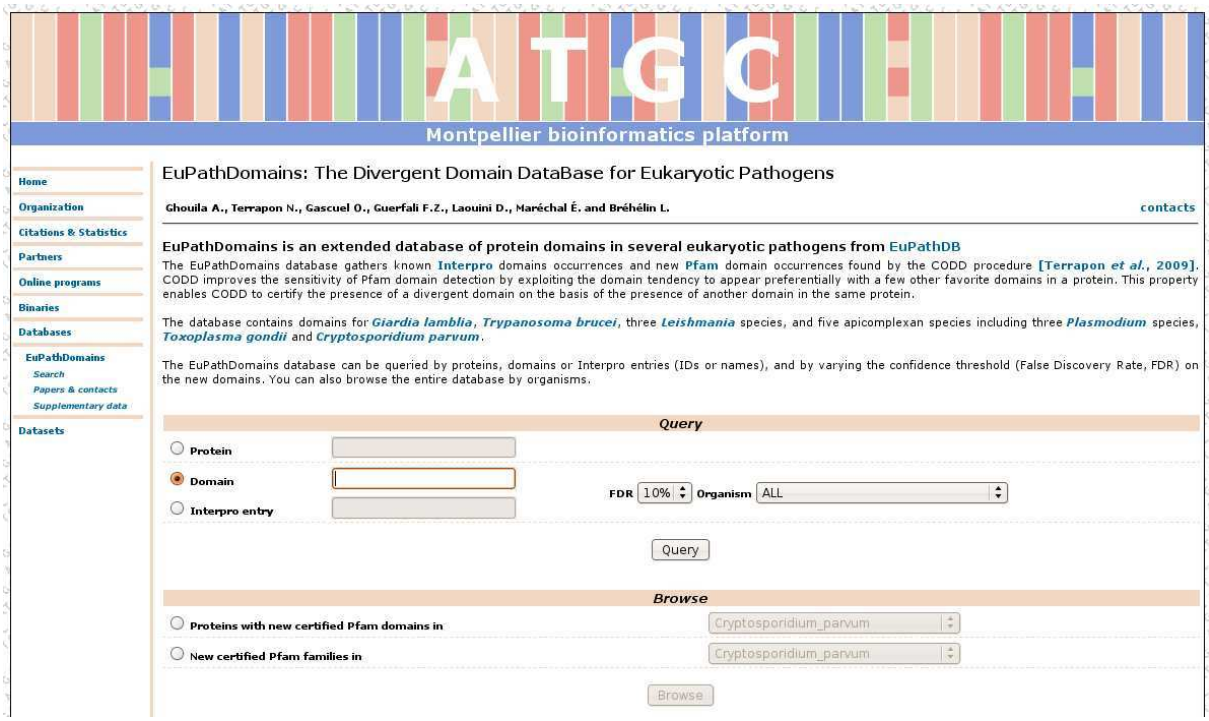


FIGURE 4.9 – **Capture d'écran du portail de la base de données EuPathDomains**.

De plus, les domaines certifiés par notre approche sont accompagnés d'autres informations relatives à la certification. Pour chaque nouvelle occurrence certifiée d'un domaine, on donne sa position, son E-valeur ainsi que le FDR associé pour chaque domaine ayant permis sa certification. Une infobulle apparait lorsque l'on survole le FDR et donne accès aux détails concernant la procédure de ré-échantillonnage : nombre total de domaines potentiels, nombre de domaines certifiés sur les données réelles et nombre de certifications attendues sous H_0. Dans cette page, les annotations GO déduites sont représentées en rouge si elles sont inédites par rapport aux annotation GO connues de la protéine, et en bleu dans le cas contraire. Chaque annotation GO est aussi un lien vers le site Web de la *Gene Ontology*, donnant accès à de plus amples informations sur la fonction décrite.

PF11_0189 (PlasmoDB link)

Annotation: hypothetical protein
PlasmoDB Workshop 2008 Reannotation: insulinase, putative

GO annotation :
GO:0004222 : metalloendopeptidase activity
GO:0046872 : metal ion binding
GO:0006508 : proteolysis
GO:0009405 : pathogenesis

No Gene Ontology annotation brought by pairs of known domains for this protein

KNOWN Interpro and Pfam domains

Domain Information	GO annotation
Pept_M16_core G3DSA:3.30.830.10 (Gene3D / InterPro) 1 — 1488	GO:0003824 : catalytic activity GO:0046872 : metal ion binding
PTHR11851:SF68 (PANTHER / InterPro) 1 — 1488	No Gene Ontology annotation for this domain.
Metalloenz metal-bd SSF63411 (InterPro) 1 — 1488	GO:0003824 : catalytic activity GO:0046872 : metal ion binding
Peptidase_M16 PF00675 (Pfam / InterPro) 1 — 1488	GO:0004222 : metalloendopeptidase activity GO:0006508 : proteolysis

NEW Pfam domains

Domain Information & Certification Details	GO annotation
M16C_assoc PF08367 (Pfam / InterPro) 1 — 1488 Localization E-value Certified by : PF00675 G3DSA:3.30.830.10 SSF63411 PF05193 688...913 0.38 — 19.5% — 20% — 20% — 19.5%	Domain itself: GO:0008237 : metallopeptidase activity GO:0008270 : zinc ion binding GO:0006508 : proteolysis In association with other domains: with PF00675 : GO:0005739 : mitochondrion with PF05193 : GO:0005739 : mitochondrion
Peptidase_M16_C PF05193 (Pfam / InterPro) 1 — 1488 Localization E-value Certified by : PF00675 G3DSA:3.30.830.10 SSF63411 PF08367 194...535 0.042 — 4.71% — — — 19.5% 1116...1262 0.044 — 4.71% — 9.66% — 9.66% — 19.5%	Domain itself: GO:0004222 : metalloendopeptidase activity GO:0008270 : zinc ion binding GO:0006508 : proteolysis In association with other domains: with PF08367 : GO:0005739 : mitochondrion with PF00675 : GO:0044464 : cell part

FIGURE 4.10 – **Résultats de la requête sur la protéine PF11_0189 de *P. falciparum*.** Sont représentés les domaines Interpro connus et les domaines Pfam certifiés. Ces derniers sont accompagnés des détails de la certification et des annotations G0 déduites.

Chapitre 5

Correction des HMM

Comme nous l'avons vu dans les chapitres précédents, les librairies classiques de HMM, telle Pfam, sont performantes pour l'annotation de protéines dites "standards" mais souffrent d'une limitation importante lorsqu'il s'agit d'identifier les domaines au sein de protéines "divergentes". Un des aspects de ce problème concerne les seuils d'identification. Ces seuils, qui minimisent le nombre de faux positifs, masquent du fait de leur rigueur les domaines les plus divergents. C'est pourquoi nous avons présenté au chapitre précédent, une méthode utilisant la co-occurrence de domaines afin de relâcher les seuils. Dans ce chapitre, nous proposons d'étudier le problème sous un angle différent en nous intéressant à l'origine des difficultés rencontrées par les librairies classiques face aux protéines les plus divergentes. Les modèles utilisés par ces librairies ont la plupart du temps été inférés sur la base de protéines issues des organismes modèles. La vocation de ces librairies est de fournir des modèles les plus généraux possibles afin d'identifier les domaines classiques dans n'importe quel génome récemment séquencé. Ce principe est contradictoire avec l'annotation de séquences où les spécificités évolutives sont nombreuses.

5.1 À quel niveau intervenir ?

Différentes approches sont envisageables pour corriger une librairie de HMM profils afin d'étudier un organisme divergent particulier. La première approche que nous avons étudiée ne constitue pas en soit une correction des modèles de domaines mais s'applique au modèle nul. L'impact du modèle nul est particulièrement important lors de l'identification des domaines. Il intervient non seulement pour le calcul du score mais aussi dans celui des E-valeurs, puisqu'il est utilisé pour générer les séquences artificielles lors du calibrages des paramètres de l'EVD. La construction d'un bon modèle nul est donc une étape importante du processus. Cela est étudié à la section 5.4.

La deuxième approche de correction que nous proposons part d'un principe assez naturel : réapprendre des modèles en utilisant des alignements de séquences graines (*cf.* section 2.4.2.c page 64) recentrés sur notre organisme cible. Nous présentons en section 5.5, une approche visant à construire des modèles "espèce-dédiés". Ces modèles sont appris à partir d'alignement-graines qui intégrent les séquences de domaines précédemment identifiées dans l'organisme

cible ou dans ses espèces les plus proches du point de vue phylogénétique. Nous verrons que cette approche produits de bons résultats, mais qu'elle a comme principal défaut d'être limitée par l'identification de domaines dans les espèces proches afin de pouvoir reconstruire un HMM. Ainsi, on ne peut améliorer l'identification que des domaines déjà connus chez *P. falciparum* ou ses relatifs.

Une troisième approche pour corriger des librairies de HMM consiste à modifier les paramètres des modèles originaux pour les adapter à l'étude de l'organisme cible. De ce point de vue, différents paramètres peuvent être corrigés. Nous avons axés nos recherches sur les probabilités de génération associées aux états *Matchs*. Ces paramètres, appris grâce aux alignements-graines, capturent l'information portée par les positions conservées au cours de l'évolution, et sont à ce titre des paramètres clefs pour la reconnaissance des domaines. Nous tâchons de définir des règles de corrections générales qui puissent être appliquées aux distributions de probabilités des différents états *Matchs*. L'idée ici est de simuler/prédire l'évolution des positions clés du domaine. Contrairement à l'approche de ré-apprentissage des modèles, qui nécessite l'identification au préalable d'occurrences du domaine dans notre espèce cible, cette méthode permet d'adapter les paramètres de l'ensemble des modèles de domaines d'une librairie. Le principe de la méthode est décrit en section 5.6. Plusieurs types de corrections ont été proposées :

- une correction numérique, appelée facteurs de correction (section 5.7) ;
- l'utilisation de matrices de substitution d'acides aminés (section 5.8) ;
- la formation de classes d'états (section 5.9) ;
- une approche de type k-plus proches voisins (section 5.10).

Les règles de corrections développées ici ne s'appliquent qu'aux états *Matchs*. Ceux-ci modélisent la majeure partie de l'information issue de l'alignement graine et sont donc les paramètres clefs des HMM. Cependant, d'autres types de correction qui n'ont pas été expérimentés dans le cadre de cette thèse sont possibles. On peut citer, par exemple, la modification de la structure des modèles, c'est à dire le nombre d'états et les transitions autorisées (ou non) entre ces états. Nous aurions également pu corriger les probabilités de génération associées aux états insertions ainsi que les probabilités de transitions entre les états. Ces approches n'ont toutefois pas été explorés en priorité car elles représentent *a priori* moins d'enjeux que les probabilités associés aux états *Matchs*.

5.2 État de l'art des méthodes de corrections de modèles

Il n'existe à notre connaissance que peu de travaux sur la correction de librairies de HMM.

En ce qui concerne l'apprentissage de modèles "espèce-dédiés", où des séquences homologues sont ajoutées à l'ensemble d'apprentissage du modèle, il s'agit là d'une approche naturelle et classique dans l'esprit de PSI-BLAST. Ce genre d'approche à été appliqué à la création d'une base de HMM profils Pfam *Fungi*-spécifique nommée FPfam (Alam *et al.*, 2007). Grâce à la disponibilité de 30 génomes de champignons, cette base propose ainsi une plus grande couverture (en terme de nombre d'occurrences de domaines par protéines et de nombre de résidus moyens impliqués dans un domaine) dans ces espèces que la librairie Pfam. Nous

proposons une approches similaire dans la section 5.5.

Il existe également quelques études publiées proposant de corriger les paramètres des modèles grâce à l'utilisation d'exemples négatifs. Ces études utilisent le cadre de familles de séquences pour lesquelles l'évolution a fait émerger plusieurs sous-familles distinctes. La correction du modèle d'une sous-famille exploite des exemples négatifs définis comme appartenant à la même famille que les séquences-graines (exemples positifs) mais classés dans une sous-famille différente. La correction des modèles permet, dans ces cas, d'obtenir des modèles plus spécifique de chaque sous-famille. Ces approches proposent :

- un nouvel algorithme d'entraînement des paramètres qui intègre l'information des exemples négatifs pour modifier les probabilités de génération (Mamitsuka, 1996) ou les probabilités de transitions entre les états (Wistrand et Sonnhammer, 2004).
- une correction des scores *a posteriori*, en isolant les positions discriminantes de chaque sous-famille grâce à une mesure d'entropie sur les probabilités de génération (Hannenhalli et Russell, 2000; Srivastava *et al.*, 2007) ou à l'estimation de sous-arbres par une phylogénie bayésienne (Brown *et al.*, 2005).

Ces méthodes s'appuient non seulement sur la divergence mais également sur la proximité des exemples négatifs pour distinguer les informations spécifiques de chaque sous-famille protéique. Ce type d'approche ne semble pas transposable à notre problématique, car nous ne cherchons pas à rendre plus spécifiques les modèles mais plutôt à accroître leur sensibilité pour détecter des séquences divergentes. De plus, la construction d'ensembles négatifs distants et proches à la fois pour une librairie d'une dizaine de milliers de familles est un processus complexe.

5.3 Évaluation des résultats des librairies corrigées

Les sections suivantes décrivent les différentes approches envisagées pour corriger les modèles de manière à les adapter à un organisme cible, par exemple *P. falciparum*. Chacune de ces approches conduit à la création d'une nouvelle librairie de modèles Pfam alternative (dans notre cas *plasmodifiée*). Ces librairies sont utilisées pour identifier les domaines de l'organisme cible, avec l'ambition de découvrir des domaines ayant échappé aux modèles de Pfam. La question qui se pose alors est de savoir comment estimer la validité des résultats obtenus, et comment les comparer à ceux de la librairie Pfam.

Une solution qui peut sembler naturelle pour cela est de se référer aux E-valeurs calculées, et de comptabiliser, pour un seuil d'E-valeur donné, le nombre de domaines découvert par l'une ou l'autre des librairies. Les meilleures librairies seraient alors celles à l'origine du plus grand nombre de domaines découverts pour un même seuil de E-valeur. La validité de cette solution, utilisée par (Brown *et al.*, 2005; Alam *et al.*, 2007), repose entièrement sur la précision des E-valeurs, et donc la pertinence du modèle nul utilisé pour la calculer. Or, nous le voyons à la section suivante, la définition d'un bon modèle nul est loin d'être une question triviale, ce qui rend cette forme de validation très discutable. Classiquement, les auteurs face à ce genre de problème se réfèrent à des familles pour lesquelles il existe une décomposition en sous-familles (Mamitsuka, 1996; Hannenhalli et Russell, 2000; Wistrand et Sonnhammer,

2004; Srivastava *et al.*, 2007). Il évaluent alors leur méthode de correction en examinant la capacité du nouveau modèle d'une sous-famille à mieux reconnaître ses propres séquences et rejeter les séquences des autres sous-familles que le modèle original. Cependant, ce type de validation ne peut être reproduite dans le cadre d'une recherche de domaines inédits au sein de protéines mal annotées.

Pour contourner ce problème, nous proposons d'utiliser notre méthode de certification par co-occurrence (*cf.* Chapitre 4) pour comparer les différentes librairies reconstruites. L'idée est que, pour un même nombre de domaines potentiels, une librairie prédisant peu de faux positifs doit permettre d'identifier un plus grand nombre de domaines certifiés par co-occurrence qu'une librairie prédisant une proportion supérieure de faux positifs. Nous allons donc évaluer et comparer les performances des librairies en utilisant les mêmes graphiques que dans le chapitre précédent, c'est à dire *via* le nombre de domaines certifiés (en ordonnées) en fonction du FDR (en abscisses). Pour éviter la multiplication des graphiques, nous nous limitons à un unique ensemble de domaines validant (*cf.* section 4.1.2 page 96) : les domaines Pfam connus.

Cette manière de procéder permet de comparer les différentes méthodes de correction de manière globale, comme le ferait des courbes ROC ou précision/rappel. Nous verrons cependant à la section 5.11, à travers une analyse plus détaillée des "meilleures" librairies, que ces différentes approches conduisent à l'identification d'ensemble de domaines parfois très différents. Bien qu'une majorité de domaines certifiés soient communs aux différentes approches, on comptabilise de nombreuses certifications uniques à chacune. En intégrant les différents résultats, on accède donc à un plus grand nombre de nouveaux domaines que la "meilleure" des approches de correction seule.

5.4 Correction du modèle nul

Le modèle nul est un des paramètres les plus important du processus d'identification des modèles. En effet, il est utilisé à deux reprises : dans la formule du score d'une séquence (*cf.* Formule 2.3 page 66), et lors du calibrage du HMM pour générer les séquences artificielles (*cf.* section 2.4.2.e page 68)

5.4.1 Le modèle nul du logiciel HMMER

Comme vu précédemment, le modèle nul d'HMMER est un HMM composé d'un seul état qui boucle sur lui même (*cf.* fig 2.6 page 67). Les probabilités de génération de cet état correspondent à la composition moyenne en acides aminés des protéines de Swiss-Prot (*cf.* figure 2.5 page 67) et la longueur moyenne des protéines générées par ce modèle (espérance du nombre de boucles) correspond à la longueur moyenne des protéines de Swiss-Prot (*cf.* Section 2.4.2.d page 66).

Nous faisons l'hypothèse que le modèle nul doit tenir compte des propriétés intrinsèques aux séquences protéiques étudiées et notamment de leur divergence. De par son paramétrage axé sur l'ensemble des protéines de Swiss-Prot, le modèle nul d'HMMER n'est donc, *a priori*, pas adapté pour l'étude d'un organisme comme *P. falciparum*. Nous proposons de le remplacer

par un modèle de structure identique mais dont la distribution de probabilités de génération est adaptée à *P. falciparum*.

5.4.2 Une distribution d'acides aminés représentative de *P. falciparum*

Le choix le plus naturel pour la distribution cible est de considérer la distribution moyenne des protéines de l'organisme, également appelée "composition globale". Cependant, pour l'étude de *P. falciparum*, il est nécessaire de prendre en compte la présence d'insertions de faible complexité au sein des protéines (*cf.* section 3.3.2 page 81). Ces insertions se caractérisent par un biais en acides aminés encore plus prononcé que dans la composition globale. Décrites comme codant des domaines non-globulaires n'affectant pas la fonction de la protéine (Pizzi et Frontali, 2001), on doit les exclure lors de l'estimation de la distribution cible. Des distributions cibles alternatives prenant en compte l'impact des zones de faible complexité doivent donc être considérées. Cependant, on dispose de peu d'informations sur ces zones dont les positions ne sont pas toujours clairement identifiées. Par conséquent, différentes approches ont été envisagées pour isoler les zones de faible complexité des protéines plasmodiales et obtenir une distribution des positions conservées. Trois solutions ont été retenues :

– **L'approche de Pizzi et Frontali (2001)** : Dans cette publication, la composition moyenne des zones de faible complexité et des zones conservées sont calculées à partir des résultats de l'algorithme SEG (Wootton et Federhen, 1993). Cet algorithme s'appuie sur la définition de *complexité compositionnelle locale*, issue de la théorie de l'information, afin de diviser les séquences d'acides aminés en zones de faible et forte complexité. Pour cela, il calcule la complexité de chaque fenêtre de lecture de longueur W puis fusionne les fenêtres recouvrantes de faible complexité. L'algorithme SEG est notamment utilisé dans le programme BLAST en prétraitement de l'alignement pour remplacer les zones de faibles complexité par l'acide aminé incertain X.

– **Utiliser les alignements de domaines Pfam connus chez *P. falciparum*** : Pour chaque domaine Pfam déjà identifié chez *P. falciparum* grâce au modèle nul original, on utilise l'algorithme de Viterbi pour extraire les positions alignées sur les états *Matchs* du HMM. Ainsi, on peut exclure les zones de faible complexité (dont les acides aminés sont alignés sur des états *Inserts*), et récupérer les positions conservées du domaine (alignées sur les états *Matchs*) à partir desquels on déduit une distribution moyenne en acides aminés.

– **Réaliser une segmentation des séquences protéiques de *P. falciparum* à l'aide d'un HMM** : Ce HMM possède deux états (*cf.* Figure 5.1). Les paramètres du modèle sont appris grâce à l'algorithme d'entraînement de Baum-Welch. On répète l'apprentissage avec une initialisation aléatoire des paramètres et on retient le résultat ayant la plus forte vraisemblance. À l'issue de l'entraînement, les distributions observées dans les deux états sont très différentes. L'une de ces distributions est fortement biaisée (plus que la composition globale), tandis que la seconde est plus proche de la distribution moyenne des protéines de Swiss-Prot. On peut donc faire l'hypothèse que le premier état a capturé les insertions et les zones de faible complexité, tandis que le second représente de façon plus précise la distribution moyenne en acides aminés des domaines de *P. falciparum*.

Nous définissons donc au total quatre compositions cibles possibles pour les domaines

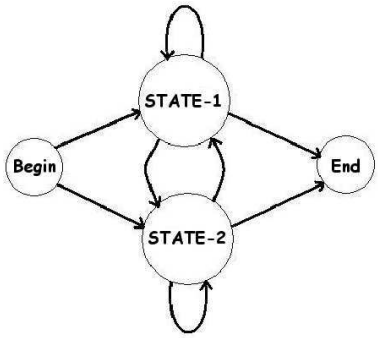


FIGURE 5.1 – **Structure du HMM à deux états utilisé pour la segmentation des protéines plasmodiales en positions conservées ou non.** Ce modèle est initié par l'état *Begin* où commence toute séquence. Les états *STATE-1* et *STATE-2*, associées à des distributions de probabilité sur les acides aminés, modélisent la génération des séquences. On espère y capturer séparément les positions conservées et les zones de faibles complexité. Enfin l'état *End* permet de clore la modélisation.

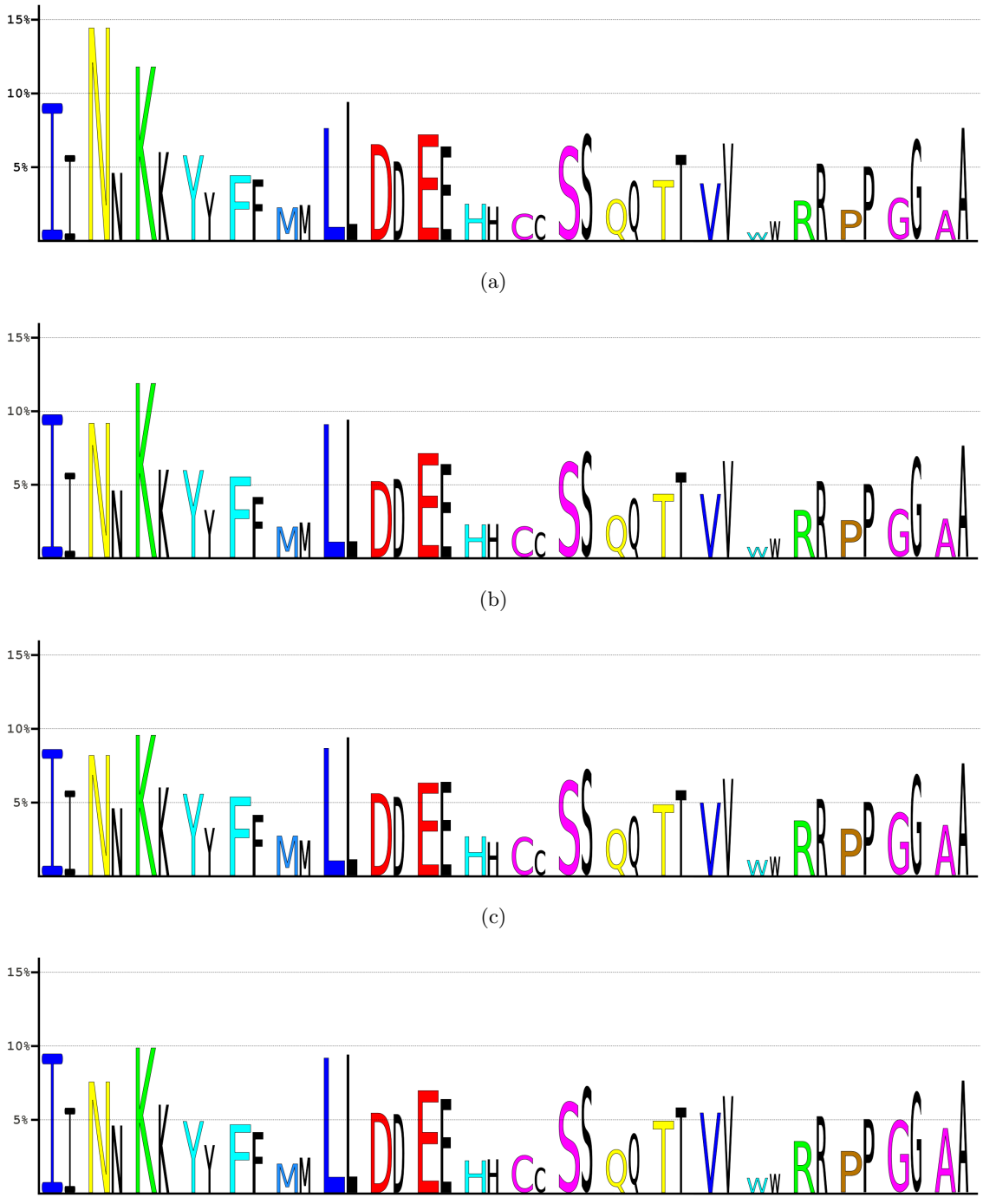
Pfam chez *P. falciparum*, représentées sur la figure 5.2. En utilisant la distance du χ^2 par rapport à la distribution de Swiss-Prot, on ordonne ces distributions de la plus proche de la composition de Swiss-Prot, et donc du modèle nul par défaut d'HMMER, à la plus biaisée :

- **la composition observée** sur les alignements des domaines Pfam connus ;
- **la composition de Pizzi** excluant les zones de faible complexité obtenues par SEG ;
- **la composition apprise** sur les protéines de *P. falciparum* par entraînement des paramètres d'un HMM à deux états ;
- **la composition globale** en acides aminés des protéines de *Plasmodium falciparum*.

5.4.3 Expérimentations

Le programme HMMER a été relancé avec chacun des modèles nuls envisagés. Notons que cela nécessite la modification de chaque HMM de la librairie (fichiers .hmm où on trouve le modèle nul des calculs de scores), ainsi que la correction du code source du programme pour la génération de séquences artificielles (calibrage des modèles). Chacun des modèles nuls a permis l'identification d'un ensemble différent de domaines potentiels, et la procédure de certification a été appliquée sur ces ensembles pour évaluer la performance de chaque librairie.

Comme l'atteste la figure 5.3, les résultats obtenus ici sont décevants au regard des résultats du modèle nul original. On constate une diminution du nombre de domaines certifiés à FDR équivalent pour les librairies ayant un modèle nul corrigé quelle que soit la nouvelle distribution. Cependant, nous verrons dans les sections suivantes que lorsque l'on modifie la composition des états du HMM, la modification conjointe du modèle nul est une étape nécessaire pour l'amélioration des performances. Cette observation semble indiquer que l'adéquation du modèle nul avec la distribution de génération globale des états des HMM profils est primordiale pour une librairie de HMM. La composition moyenne des états des HMM de



(a)

(b)

(c)

(d)

FIGURE 5.2 – **Logo des distributions en acides aminés des quatre compositions cibles envisagées et comparaison avec la distribution de Swiss-Prot**. Les fréquences des différents acides aminés sont représentées sur l'axe de droite. Dans chaque figure on trouve représentée en noir la distribution moyenne en acides aminés des protéines de Swiss-Prot, et en couleurs (*cf.* figure 2.4 66 pour le code couleur) les quatres distribution apprises ordonnées de la plus proche de Swiss-Prot à la plus biaisée (par une distance du χ^2 à la distribution de SwissProt) : la distribution observée sur les alignements de domaines Pfam connus chez *P.falciparum* (a), celle publiée par Pizzi (excluant les zones de faible complexité identifiées par SEG) (b), celle apprise grâce à un HMM à deux états par l'entraînement de Baum-Welch (c) et la distribution moyenne des protéines de *P. falciparum* (d).

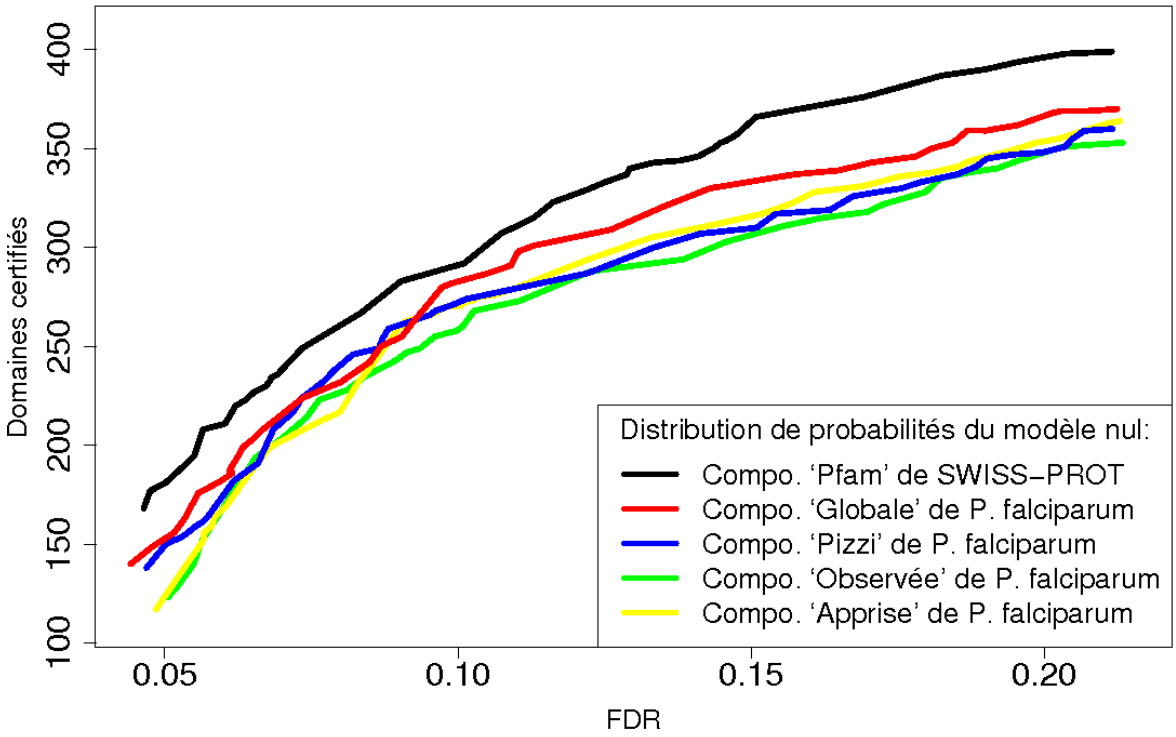


FIGURE 5.3 – **Nombre de certifications réalisées en fonction du FDR, par les quatre librairies corrigées en modifiant le modèle nul.**

Pfam étant proche de celle de Swiss-Prot, le modèle nul par défaut d'HMMER semble alors le mieux adapté. Par contre, lorsque la librairie utilisée exhibe une composition moyenne plus proche de celle de *P. falciparum*, comme c'est le cas des librairies corrigées des sections suivantes, alors la correction conjointe du modèle nul conduit souvent à de meilleurs résultats.

5.5 Réapprendre grâce aux espèces proches

Les alignements qui ont servi à l'apprentissage des HMM de Pfam sont souvent constitués de séquences éloignées du point de vue phylogénétique de l'organisme cible. Une solution naturelle pour remédier à ce problème consiste à intégrer aux alignements-graines, des séquences appartenant à l'organisme cible et à des espèces proches. L'approche que nous proposons pour cela est similaire à celle décrite dans (Alam *et al.*, 2007). Étant donnés une librairie de modèles $\mathcal{L}$ et un ensemble de séquences protéiques $\mathcal{E}$ appartenant à l'organisme cible et à des espèces proches :

1. Utiliser la librairie $\mathcal{L}$ sur l'ensemble de séquences $\mathcal{E}$ afin d'identifier tous les domaines

connus,

2. Construire de nouveaux alignements d'apprentissage grâce aux domaines identifiés,
3. Apprendre des modèles *espèce-dédiés* à partir des alignements,
4. Procéder à une nouvelle recherche de domaines dans l'organisme cible en utilisant les modèles reconstruits.

L'apprentissage de modèles espèce-dédiés présente l'avantage de construire des modèles plus sensibles grâce aux nouveaux alignements orientés vers l'organisme cible. Cependant, cette approche est limitée par le fait que l'on ne peut reconstruire que des types de domaines que l'on connaît déjà. L'utilisation des espèces proches permet de contourner le problème dans une certaine mesure, en reconstruisant des types de domaines connus dans ces espèces en dépit de leur absence dans l'organisme cible. Mais un grand nombre de domaines ne seront pas reconstruits, et notamment les domaines rares ou divergents dans le taxon étudié.

La sélection de l'ensemble $\mathcal{E}$ des protéines d'espèces proches est discutée en section 5.5.1. Puis nous voyons la construction des nouveaux modèles (section 5.5.2), avant de conclure sur les résultats de cette approche (section 5.5.3).

5.5.1 Sélection des espèces proches

La première étape pour construire des HMM profils espèce-dédiés, concerne la sélection des espèces phylogénétiquement les plus proches de notre organisme cible. La quantité et la proximité des espèces séquencées disponibles sont deux aspects primordiaux de cette approche. Si l'on dispose de suffisamment d'espèces séquencées au sein du genre il est possible de construire des modèles dont l'information (position spécifique et évolutive) est plus proche de ce que l'on s'attend à observer dans notre organisme cible. D'un autre côté, si l'on est trop proche de l'organisme cible, les domaines reconstruits correspondent majoritairement à des domaines que l'on connaît déjà. Cela n'apporte rien en terme de domaines inédits et revient à nettoyer l'annotation en domaines grâce aux protéines orthologues. Il est aussi possible de remonter au delà du genre (ordre, classe, phylum, *etc.*) afin de récolter une plus grande diversité de séquences et de types de domaines inédits dans notre cible. On conserve alors une proximité que ne possède pas les modèles de la librairie Pfam mais on perd un peu de la spécificité du genre et de l'espèce.

Nous avons donc choisi d'expérimenter deux ensembles d'espèces proches. Le premier est constitué des espèces plasmodiales complètement séquencées : *P. falciparum*, *P. vivax*, *P. yoelii*, *P. berghei*, *P. chabaudi*, et *P. knowlesi*. Les séquences protéiques de ces espèces ont été téléchargées depuis le site Web PlasmoDB.

Pour introduire plus de diversité, nous avons construit un deuxième jeu d'espèces proches. Pour cela, nous avons tout d'abord étendu le premier jeu pour englober le phylum des apicomplexes et accéder à sept génomes complets supplémentaires, comme illustré par la figure 5.4 :

– *Babesia bovis*, *Theileria annulata* et *Theileria parva*, qui font partie de la classe *Aconoidasida* comme *P. falciparum*, mais sont des *Piroplasmida* et non des *Haemasporidae*.

– *Toxoplasma gondii*, *Cryptosporidium muris*, *Cryptosporidium hominis* et *Cryptosporidium parvum*, appartenant à la classe *Coccidia*.

L'ensemble des protéines de ces espèces sont extraits du site Web ApiDB. Dans un second temps, nous avons ajouté à ces séquences l'ensemble des protéines d'*Alveolata* connues, en relâchant la contrainte concernant l'aspect "génome complet". Les *Alveolata* comprennent, en plus du phylum des apicomplexes, celui des dinoflagéllés et des ciliés (*cf.* figure 5.4). L'ensemble des séquences protéiques des espèces *Alveolata* a été obtenu *via* le site Web du NCBI. Ce deuxième jeu dispose donc d'une plus grande quantité de données mais se compose de séquences plus éloignées de *P. falciparum* que le premier, tout en restant plus proches que la plupart des séquences graines utilisées dans la librairie Pfam.

5.5.2 Reconstruction des HMM

Pour chacun des jeux d'espèces proches, on effectue une recherche des domaines Pfam avérés en utilisant les seuils recommandés. Pour chaque type de domaine, nous collectons l'ensemble des occurrences identifiées dans les espèces proches. Le HMM d'origine est utilisé pour générer deux alignements multiples de ces séquences :

– l'un contenant exclusivement les séquences des espèces proches ;
– l'autre contenant les séquences de l'alignement-graine d'origine et celles des espèces proches.

L'alignement est réalisé grâce à la fonction `hmmalign` du programme HMMER avec le paramètre `-m` pour conserver des modèles ayant le même nombre d'états *Matchs* que les originaux. Ces deux alignements sont alors utilisés pour apprendre les paramètres d'un nouveau modèle grâce à la fonction `hmmbuild` d'HMMER. Deux séries de nouveaux modèles ont donc été construits pour chaque jeu d'espèces proches.

5.5.3 Résultats

Rappelons qu'on ne peut ré-apprendre les paramètres d'un HMM que si l'on a identifié au moins une séquence correspondante du domaine dans les espèces proches. Le nombre de HMM reconstruits grâce aux espèces proches, pour les différents jeux utilisés, est respectivement de 1 635 grâce aux *Haemasporidae* et de 2 465 grâce aux *Alveolata*, sur les 10 340 modèles Pfam existants. Les librairies alternatives que nous produisons dans cette section sont donc composées des nouveaux modèles reconstruits, complétés par les modèles originaux de Pfam pour les types de domaines où aucune nouvelle séquence n'a été obtenue.

La figure 5.5 représente les résultats obtenus par les quatre nouvelles librairies et par la librairie originale. Les quatre nouvelles librairies correspondent respectivement à chaque jeux d'espèces proches — *Haemasporidae* en rouge et *Alveolata* en bleu — et, pour chaque jeu, à une reconstruction des modèles uniquement à partir de séquences proches (courbes en pointillés) ou par l'intégration de ces séquences aux alignements-graines originaux de Pfam (en trait plein).

Notons que le modèle nul par défaut de Pfam a été utilisé. Des expériences avec un modèle nul alternatif correspondant à la distribution globale des protéines plasmodiales ont montré

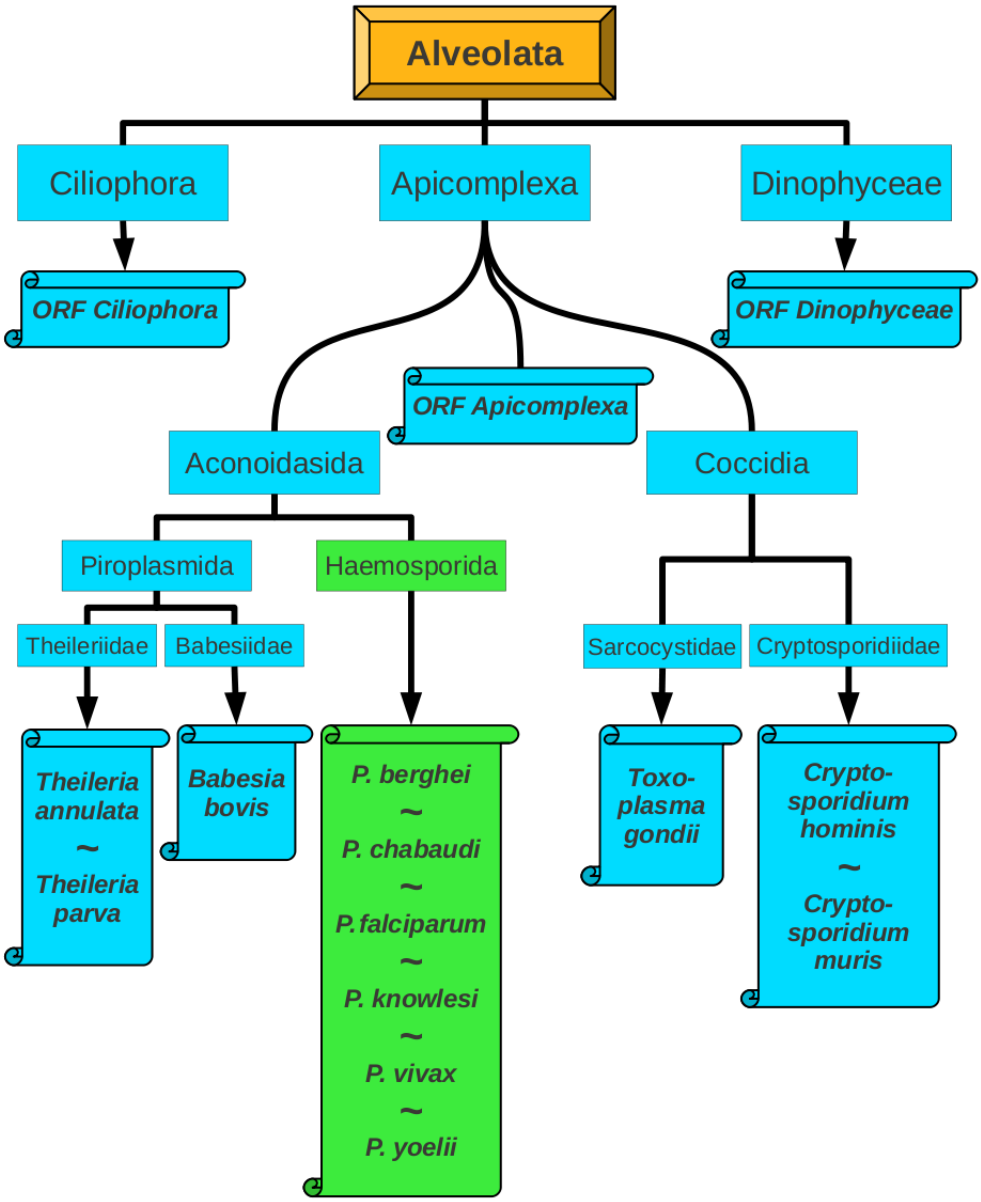


FIGURE 5.4 – **Arbre des espèces utilisées pour l'apprentissage de modèles *Plasmodium-dédiés***. Le premier jeu de séquences est construit à partir des génomes complets des sept espèces plasmodiales séquencées (partie verte de la phylogénie). Le deuxième jeu de séquences s'obtient en intégrant d'abord sept génomes complets supplémentaires d'*Apicomplexa*. Puis on complète ce second jeu par l'ensemble des ORF traduites d'Apicomplexes, de Ciliés et de Dinoflagélés, extraites de la base de données de séquences protéiques du NCBI.

une détérioration des performances (résultats non présentés).

On constate, tout d'abord, que la librairie obtenue grâce aux *Alveolata* permet de certifier

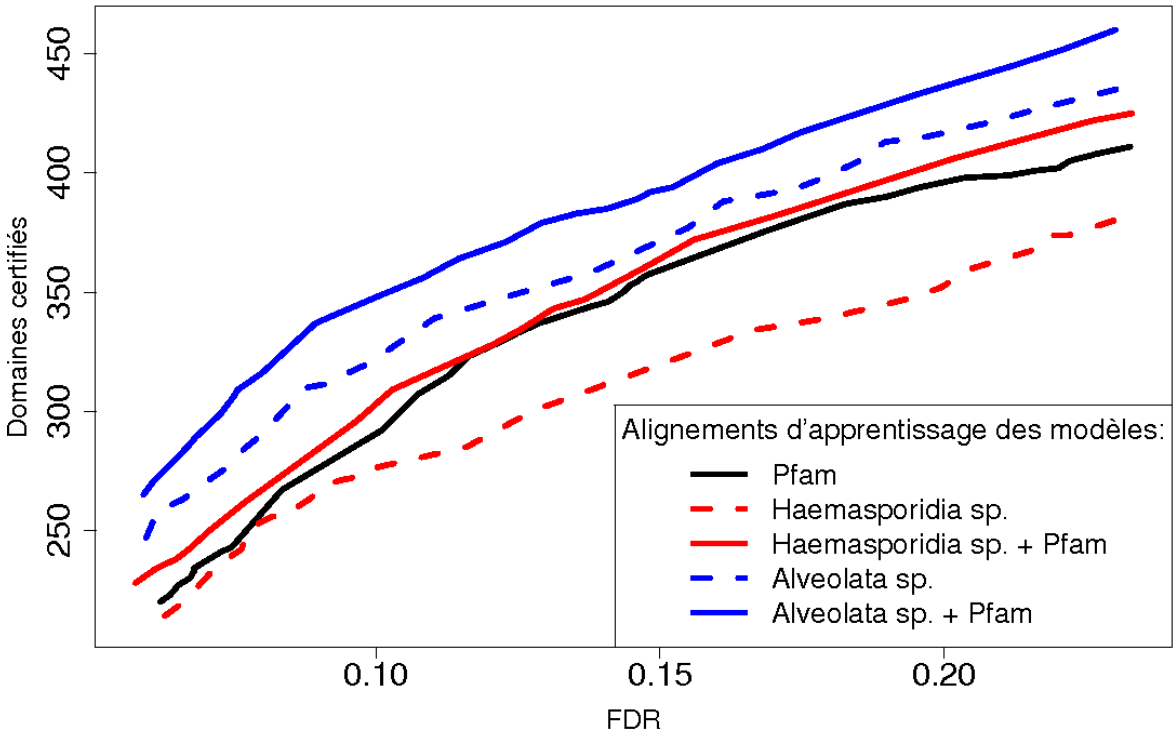


FIGURE 5.5 – **Résultats de certification par co-occurrence des librairies de HMM profils reconstruits sur des alignements contenant les séquences des espèces proches.**

un plus grand nombre de domaines que celle reconstruite uniquement sur les espèces les plus proches de *P. falciparum*. Ensuite, pour les deux jeux d'espèces proches, les modèles reconstruits à partir des séquences proches et de la graine initiale sont plus performants que ceux obtenus uniquement à partir des séquences proches.

Cette approche de correction fournit donc des résultats intéressants sur lesquels nous revenons plus en détails dans la section 5.11.

5.6 Modification des distributions associées aux états *Matchs*

Dans la suite de ce chapitre, nous proposons différentes méthodes pour modifier *a posteriori* les probabilités des états *Matchs* des modèles. Comme nous venons de le voir, l'approche de reconstruction grâce aux espèces proches ne permet de corriger que des types de domaines pour lesquels une occurrence est connue dans le jeu de séquences sélectionnées. En l'absence d'exemplaires connus d'un domaine dans l'espèce cible ou ses plus proches relatifs, il devient

nécessaire de simuler l'évolution des différentes positions du domaine pour corriger les modèles. L'objectif des méthodes développées dans ce chapitre est de proposer des règles de correction générales qui puissent être appliquées à tout état *Match*, et ainsi pouvoir corriger l'intégralité des modèles de la librairie Pfam.

Dans la suite, on considère l'ensemble des états *Matchs* de tous les HMM de Pfam, que l'on note $X = \{x_1 \dots x_N\}$. Chacun des N individus x_i est décrit par un vecteur de fréquences

$$x_i = (x_{ij})_{j\in[1..20]},$$

où x_{ij} est la probabilité de générer l'acide aminé j associée à l'état *Match* x_i. Les probabilités de génération de l'ensemble des états *Matchs* de la librairie Pfam exhibent une distribution moyenne en acides aminés $\pi = (\pi_j)_{j\in[1..20]}$, proche de celle des protéines de Swiss-Prot. Cette observation reflète le fait que les paramètres des modèles ont été entraînés sur des séquences ne présentant ni la divergence ni le biais de *P. falciparum*. Dans les sections suivantes nous proposons différentes méthodes de correction qui font tendre cette distribution moyenne vers une distribution cible $\sigma = (\sigma_j)_{j\in[1..20]}$ plus proche de celle des domaines protéiques de *P. falciparum*. Ces corrections ne se résument cependant pas à un ré-ajustement de composition globale des états *Matchs*. Il faut aussi tenir compte des spécificités de chaque position des HMM qui traduisent les contraintes physico-chimiques qui s'exercent à ces positions. Tout le problème est alors de définir l'opération à utiliser qui permette de conserver l'information position-spécifique tout en simulant une évolution divergente et biaisée comme chez *P. falciparum*. Divers solutions ont été envisagées et sont détaillées ci-après (sections 5.7 à 5.10).

5.7 Facteurs de correction

5.7.1 Principe

Une idée simple pour corriger les modèles Pfam consiste à utiliser un vecteur de *facteurs de correction* multiplicatifs, noté $(a_j)_{j\in[1..20]}$ pour transformer chaque distribution x_i associée à un état *Match* en une distribution x_i^* *plasmodifiée* en appliquant une fonction du type :

$$x_{ij}^* = f(x_{ij}) = \frac{a_j x_{ij}}{\sum_{k=1}^{20} a_k x_{ik}}, \quad \forall j \in [1..20]. \tag{5.1}$$

Le dénominateur est un terme de normalisation qui garantit que pour toute distribution x_i passée en paramètre, le résultat $x_i^* = f(x_i)$ est aussi une distribution de probabilités. L'opération doit être définie pour nous permettre de transformer la distribution globale actuelle π des états *Matchs* des HMM de Pfam en une distribution cible σ plus proche de *P. falciparum*. Cela signifie que les valeurs des $(a_j)_{j\in[1..20]}$ correspondent à la résolution de l'équation $f(\pi) = \sigma$. Ces facteurs de correction sont calculés après avoir choisi la distribution de départ π et la distribution cible σ. Puis ils sont appliqués à chaque état *Match* de l'ensemble des HMM de Pfam selon l'équation (5.1), pour obtenir une nouvelle librairie de HMM qui exhibe une composition globale de ses états *Matchs* correspondant à la distribution cible désirée.

Les $(a_j)_{j\in[1..20]}$ s'obtiennent par la résolution de l'équation suivante :

$$\forall j \in [1..20], \quad \sigma_j = \frac{a_j \pi_j}{\sum_{k=1}^{20} a_k \pi_k}.$$

En observant cette équation, on constate tout d'abord que les $(a_j)_{j\in[1..20]}$ sont définis à un facteur multiplicatif λ près, c'est à dire que si $(a_j)_{j\in[1..20]}$ est une solution alors $(\lambda a_j)_{j\in[1..20]}$ est également solution. Il est donc possible de calculer une solution pour les valeurs des facteurs de correction $(a_j)_{j\in[1..20]}$ tel que le terme de normalisation disparaisse de l'équation, c'est à dire avec la contrainte $\sum_{k=1}^{20} a_k\pi_k = 1$. On obtient ainsi une solution évidente de l'équation :

$$\forall j \in [1..20], \quad a_j = \frac{\sigma_j}{\pi_j}.$$

Dans la suite, nous discutons des différentes distributions cibles qui peuvent être envisagées (section 5.7.2) avant de présenter les résultats obtenus par les librairies corrigées (section 5.7.3)

5.7.2 Choix des distributions de départ et cible

La distribution de départ π est obtenue en moyennant les distributions de probabilités $(x_i)_{i\in[1..N]}$ associées à l'ensemble des états *Matchs* de la librairie Pfam. Comme attendu, on obtient une distribution en acides aminés très proche de celle calculée sur les protéines de Swiss-Prot (*cf.* figure 3.5 page 82).

En ce qui concerne la distribution cible, les différents choix utilisés au cours de nos expérimentations correspondent aux distributions vues précédemment (*cf.* section 5.4.2 et figure 5.2) :

- **distribution globale** en acides aminés des protéines de *Plasmodium falciparum* ;
- **distribution Pizzi** excluant les zones de faible complexité obtenues par SEG ;
- **distribution observée** sur les alignements des domaines Pfam connus ;
- **distribution apprise** par entraînement du HMM à deux états (*cf.* Figure 5.1).

Pour chaque composition cible, les facteurs de corrections appropriés sont calculés, puis appliqués aux états *Matchs* de tous les HMM originaux pour créer une nouvelle librairie.

5.7.3 Résultats

La figure 5.6 présente les résultats de la méthode de certification par co-occurrence pour les différentes librairies corrigées par facteurs de corrections. Dans la première figure (en haut) le modèle nul utilisé est celui par défaut de Pfam, tandis que dans la suivante (en bas) les probabilités de génération du modèle nul correspondent à la distribution cible σ utilisée pour le calcul des facteurs de correction. On constate que de meilleurs résultats sont obtenus avec un modèle nul corrigé. Dans ce cas, les librairies obtenues permettent de certifier un plus grand nombre de domaines que la librairie Pfam originale, à FDR équivalent. On remarque aussi que les résultats obtenus sont assez proches quelle que soit la distribution cible choisie. Toutefois, la librairie correspondant aux facteurs de correction vers la distribution globale de *P. falciparum* semble la plus performante et sera retenue pour la comparaison des meilleures librairies dans la section 5.11.

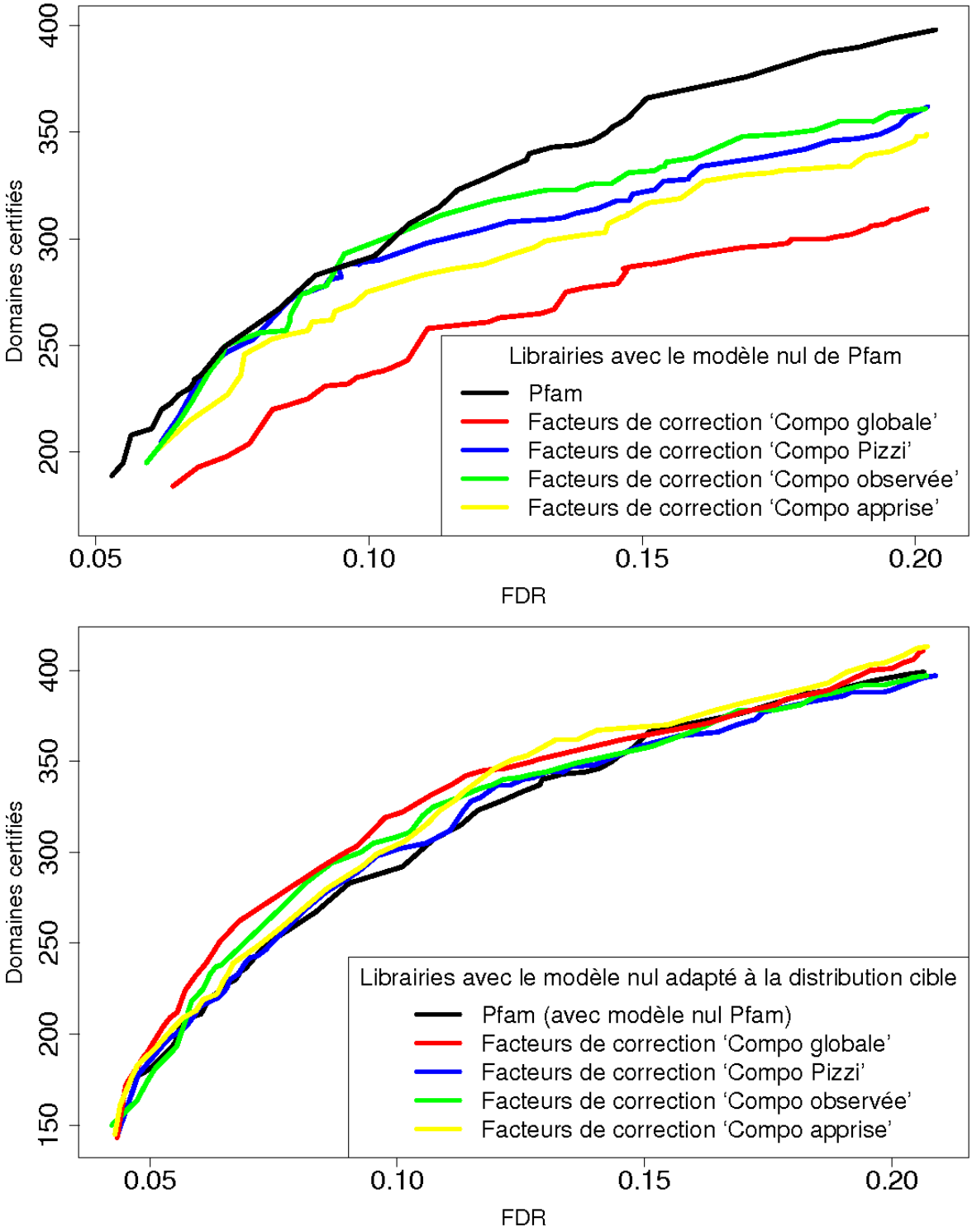


FIGURE 5.6 – **Résultats de certification par co-occurrence des librairies de HMM profils corrigés par facteurs de correction.** La figure du haut représente les résultats obtenus par des librairies corrigées et ayant un modèle identique à Pfam. La figure du bas correspond à des librairies corrigées dont le modèle nul a été adapté à la composition cible de correction.

5.8 Matrices de substitution

La deuxième méthode de correction proposée pour corriger les probabilités de génération des états *Matchs* des HMM, fait intervenir une matrice de substitution d'acides aminés. Les matrices de substitution sont utilisées en phylogénie afin de simuler l'évolution des séquences au cours du temps. L'utilisation de matrices de substitution permet d'intégrer une dimension évolutive dans nos corrections qui n'est pas prise en compte dans les facteurs de correction.

5.8.1 Probabilités de substitution entre acides aminés

Le calcul d'une matrice de substitution d'acides aminés se fait comme suit. On part d'une matrice de taux d'échange instantané notée R. Ce type de matrice représente le coût pour la conservation des propriétés physico-chimiques lors d'une substitution. Elles sont symétriques car le processus d'évolution modélisé est réversible. Pour les protéines, les plus connues sont PAM1 (Dayhoff *et al.*, 1978), JTT (Jones *et al.*, 1992b), WAG (raffinement des précédentes (Whelan et Goldman, 2001)) et LG (Le et Gascuel, 2008).

On introduit ensuite la distribution cible σ vers laquelle on souhaite tendre. On note alors Q la matrice produit de R par σ selon la formule :

$$Q_{jk} = \sigma_k R_{jk} \text{ pour } j \neq k, \quad \text{et} \quad Q_{jj} = -\sum_{j \neq k} Q_{jk}.$$

La forme normalisée de la matrice Q est obtenue par $\frac{1}{\mu}(Q_{jk})$, avec $\mu = -\sum_j \sigma_j Q_{jj}$.

Enfin, on obtient une matrice de substitution entre acide aminés notée $P(t)$ grâce à l'équation :

$$P(t) = e^{Qt}.$$

Pour construire cette matrice P, on diagonalise la matrice Q, c.-à-d. $Q = V \times U \times V^{-1}$, où V et V^{-1} sont respectivement la matrice des vecteurs propres et son inverse, et U est une matrice diagonale. Porter à l'exponentielle une matrice diagonale consiste à mettre à l'exponentielle les termes de sa diagonale. Cela permet de calculer aisément différentes matrices $P(t)$ en fonction du paramètre t par la formule :

$$P(t) = V \times e^{Ut} \times V^{-1}.$$

Pour plus de détails sur la construction d'une matrice de substitution, on pourra consulter (Guindon et Gascuel, 2007).

Les éléments $P_{ij}(t)$ représentent la probabilité qu'au cours d'un espace de temps t, l'acide aminé i se soit transformé en l'acide aminé j. La valeur de l'unité de temps t correspond à l'espérance du nombre de substitutions par position. Une propriété des matrices $P(t)$ est la suivante : quelque soit la distribution initiale π, on a : $\lim_{t \to +\infty} \pi P(t) = \sigma$. C'est pourquoi, σ est appelée *distribution stationnaire*. Cependant, pour t suffisamment petit, la multiplication d'une distribution π par $P(t)$, la transforme en une distribution π^* plus proche de la composition stationnaire. Les matrices de substitution nous permettent donc de transformer les distributions d'acides aminés $(x_i)_{i \in [1..N]}$ associés aux états *Matchs* de la librairie Pfam,

en leur faisant "faire un pas" vers une distribution cible de notre choix. De plus, l'intérêt des matrices de substitutions est la conservation, sous une contrainte évolutive, des propriétés physico-chimiques prépondérantes dans les distributions initiales. Toutefois, il faut être prudent car en choisissant t trop grand, on transforme toute distribution en un profil unique de composition σ et on perd l'information position-spécifique.

5.8.2 Matrices de substitution pour *Plasmodium falciparum*

Nous avons construit différentes matrices de substitution afin de générer de nouvelles librairies dans le cadre de l'étude de *P. falciparum*. Ces matrices de substitution s'appuient sur la matrice de taux d'échange instantané LG (Le et Gascuel, 2008). La matrice LG, téléchargeable depuis `http://atgc.lirmm.fr/LG` et intégrée au programme de phylogénie PHYML (Guindon et Gascuel, 2003), a été obtenue en raffinant l'approche d'estimation par maximum de vraisemblance de (Whelan et Goldman, 2001) en intégrant des taux variables à travers les sites et en s'appuyant sur une base de données bien plus grande. Pour le choix de la distribution stationnaire σ, nous avons envisagé les quatre distributions cibles employées dans les sections précédentes et présentées section 5.4.2. Enfin, nous avons créé des matrices de substitutions pour différentes valeurs de t, de 0.01, 0.05, 0.1 et 0.2. Les modèles corrigés sont construits en appliquant, pour toute distribution x_i associé à un état *Match*, la formule suivante :

$$x_i^* = x_i P(t).$$

5.8.3 Résultats

Différents modèles nuls ont également été testés pour ce type de correction.

Le modèle nul de Pfam semble être le mieux adapté, ce qui peut s'expliquer par le fait que la composition globale des modèles corrigés reste plus proche de celle de Swiss-Prot que de celle de *P. falciparum* (contrairement à la méthode des facteurs de correction).

La figure 5.7 présente les résultats obtenus par les librairies de modèles corrigées par des matrices de substitution pour des valeurs de t de 0.05 et 0.1 et avec le modèle nul par défaut de Pfam (les librairies corrigées avec une valeur supérieure ($t = 0.2$) et inférieure ($t = 0.01$) donnant de plus mauvais résultats, elles ne sont pas représentées). On constate que le taux t optimal semble dépendre de la distribution cible et que le nombre de domaines certifiés à FDR équivalent est très proche entre les différentes librairies. La librairie qui se détache le plus nettement de celle de Pfam est obtenue par les paramètres $t = 0.1$ et $\sigma = Compo.\ Pizzi$, on la retrouve en section 5.11 pour la comparaison des différentes méthodes de correction.

5.9 Former des classes d'états

Pour compenser l'absence de protéines plasmodiales dans les alignements utilisés lors de l'entraînement des HMM, nous avons introduit, à la section précédente, une méthode qui simule l'évolution des distributions en acides aminés associées aux états *Matchs* vers une composition cible tout en essayant de conserver les propriétés physico-chimiques des états.

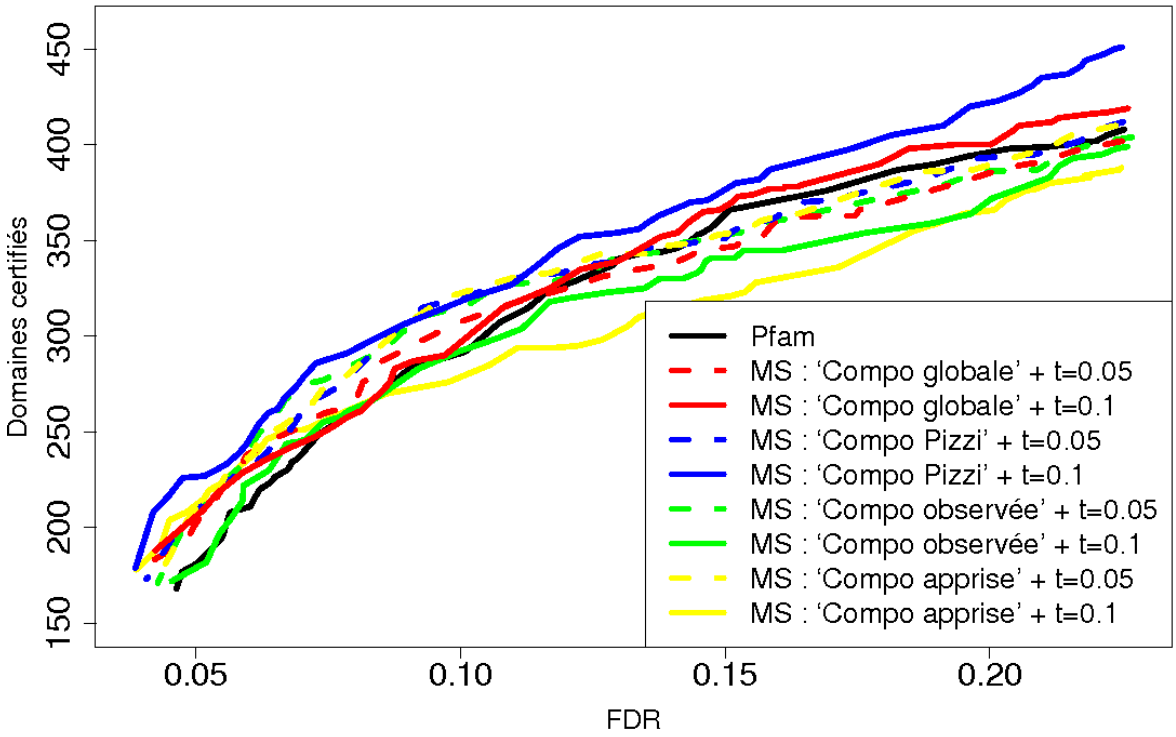


FIGURE 5.7 – **Résultats de certification par co-occurrence des librairies de HMM profils corrigés par matrices de substitution (MS) avec le modèle nul Pfam.**

Cependant, le mode d'évolution des organismes très divergents est souvent complexe et les acides aminés que l'on s'attend à observer à une position donnée sont difficilement prédictibles à l'aide d'un schéma d'évolution classique. Nous proposons dans cette section une méthode qui met à profit l'information apportée par les domaines déjà identifiés dans l'organisme cible pour estimer l'évolution de la distribution associée à chaque position, en fonction de son profil physico-chimique.

5.9.1 Principe

Les états *Matchs* des HMM profils représentent l'information des positions conservées au cours de l'évolution. Dans chaque état, la distribution d'acides aminés reflètent les contraintes physico-chimiques associées à cette position. Notre approche de correction s'appuie sur le regroupement des états ayant des distributions similaires, et donc, on l'espère, des contraintes physico-chimiques proches. Dans un premier temps, on définit différentes classes d'états à l'aide d'une procédure de classification, ou *clustering*, basée sur les probabilités de génération associés aux états *Matchs*. Une procédure classique de *clustering* est celle des *K-means* (Lloyd,

1957). Cette procédure prend en entrée le nombre de classes voulues K, et est appliquée sur tous les états *Matchs* des HMM de Pfam. À l'issue de cette procédure, chaque état d'un HMM est associé à une et une seule classe. Une fois la classification réalisée, on utilise l'ensemble des domaines déjà identifiés dans les protéines de *P. falciparum* ou de ses espèces proches, pour aligner les états des HMM sur les acides aminés qui leur correspondent dans la protéine, grâce à l'algorithme de Viterbi. Une fois ces alignements réalisés, on peut assigner à chaque classe d'état le nombre de fois où chaque acide aminé a été observé aligné sur un état membre de cette classe dans les protéines étudiées. On dispose donc d'une fonction, qui à chaque classe d'état associe une nouvelle distribution de probabilités de génération d'acides aminés. Cette fonction est alors utilisée pour modifier les distributions associées aux états des HMM de la librairie Pfam. Pour cela, on combine la distribution originale de chaque état avec celle associée à sa classe d'appartenance.

Nous allons tout d'abord présenté la procédure des *K-means* et les paramètres qui influent sur la classification : le nombre de classes et la distance choisie (section 5.9.2). Nous discutons ensuite le choix de l'ensemble de domaines connus (section 5.9.3), les proportions du mélange qui définit les nouvelles distributions (section 5.9.4), et le modèle nul utilisé avec la nouvelle librairie créée (section 5.9.5), avant de conclure dans la section 5.9.6 par les résultats obtenus par les librairies corrigées par cette combinaison d'une classification (*K-means*) et d'un apprentissage (chemin de Viterbi sur les domaines connus de *P. falciparum* et des relatifs).

5.9.2 *K-means*

a) Algorithme : Les *K-means* font partie des méthodes de *clustering* les plus populaires. Intuitivement, l'objectif d'une analyse par *clustering* est de partitionner un ensemble d'individus en différentes classes (*clusters*) de telle manière que la distance entre deux individus appartenant à une même classe tende à être plus faible que celle entre les individus de classes différentes.

Soit $(x_i)_{i\in[1..N]}$ l'ensemble de tous les états *Matchs* des HMM de Pfam, et soit K le nombre de classes voulues. La méthode des *K-means* utilise, comme son nom l'indique, K vecteurs moyens $(\mu_k)_{k\in[1..K]}$ pour définir les K classes. Chaque individu x_i est assigné à la classe $C(x_i)$ dont le vecteur moyen est le plus proche de x_i (*cf.* Equation (5.3)). L'objectif de l'algorithme des *K-means* est de trouver les K vecteurs moyens μ_k qui minimisent le critère suivant :

$$\sum_{k=1}^{K} \sum_{C(x_i)=k} dist(x_i, \mu_k), \tag{5.2}$$

avec $dist()$ une fonction de distance calculée dans un espace à 20 dimensions. Il s'agit là d'un problème NP-difficile que l'algorithme des *K-means* tente de résoudre grâce à une approche itérative appliquée à un ensemble de K vecteurs moyens initiaux :

1. On assigne chaque état à la classe dont le vecteur moyen est le plus proche :

$$C(x_i) = \underset{1\leq k\leq K}{\operatorname{argmin}} \; dist(x_i, \mu_k). \tag{5.3}$$

2. On recalcule les vecteurs moyens de chaque classe en minimisant la distance moyenne à l'ensemble des états membres de cette classe (*cf.* paragraphe suivant "Mesure de distance").

Chacune de ces deux étapes garantit une réduction du critère (5.2). On les répète donc jusqu'à convergence de l'algorithme. La convergence est assurée vers un optimum local. L'algorithme est alors appliqué un grand nombre de fois en faisant varier l'initialisation des K vecteurs moyens afin de retenir la solution amenant au meilleur optimum local.

b) Mesure de distance : Notre algorithme des *K-means* a été implémenté pour deux distances : la distance Euclidienne et la distance du χ^2. Si pour la distance Euclidienne, la formule du calcul des vecteurs moyens (étape 2 de l'algorithme) est classique, ce n'est pas le cas pour la distance du χ^2. La réduction du critère à cette étape consiste à ré-évaluer la position du vecteur moyen pour chaque classe afin de minimiser la distance moyenne entre ce vecteur et les individus membres de la classe. On résout ce problème par la minimisation du critère indépendamment pour chacune des classes : $\sum_{C(x_i)=k} dist(x_i, \mu_k),\ \forall\ \ k \in [1..K]$. Pour cela, on utilise la dérivée en 0 de cette expression par rapport au vecteur moyen.

Avec la distance Euclidienne définit par $\sum_{j=1}^{20}(x_{ij} - \mu_{kj})^2$, on obtient l'expression :

$$\frac{\delta\left(\sum_{C(x_i)=k}\sum_{j=1}^{20}(x_{ij} - \mu_{kj})^2\right)}{\delta\mu_k} = 0,\ \forall\ \ k \in [1..K].$$

On extrait la somme sur j, afin de raisonner individuellement sur chaque coordonnée (la dérivée d'une somme étant égale à la somme des dérivées) :

$$\frac{\delta\left(\sum_{C(x_i)=k}(x_{ij} - \mu_{kj})^2\right)}{\delta\mu_k} = 0,\ \forall\ \ k \in [1..K] \text{ et } \forall\ \ j \in [1..20].$$

La résolution amène aux équations suivantes :

$$\sum_{C(x_i)=k} 2(\mu_{kj} - x_{ij}) = 0 \iff (\#k)\mu_{kj} - \sum_{C(x_i)=k} x_{ij} = 0,\ \forall\ \ k \in [1..K] \text{ et } \forall\ \ j \in [1..20],$$

avec $\#k$ le cardinal de la classe k. Cela correspond à un calcul des vecteurs moyens μ_k comme la moyenne des coordonnées de l'ensemble des individus membres de la classe k :

$$\mu_{kj} = \frac{\sum_{C(x_i)=k} x_{ij}}{\#k},\ \forall\ \ j \in [1..20].$$

Pour la distance du χ^2, de formule $\sum_{j=1}^{20}\frac{(x_{ij}-\mu_{kj})^2}{\mu_{kj}}$, on suit un raisonnement identique jusqu'à l'individualisation des coordonnées. Cela conduit à l'expression :

$$\frac{\delta\left(\sum_{C(x_i)=k}\frac{(x_{ij}-\mu_{kj})^2}{\mu_{kj}}\right)}{\delta\mu_k} = 0,\ \forall\ \ k \in [1..K] \text{ et } \forall\ \ j \in [1..20],$$

dont la résolution amène aux équations suivantes :

$$\sum_{C(x_i)=k} \frac{2\mu_{kj}(\mu_{kj} - x_{ij}) - (\mu_{kj} - x_{ij})^2}{\mu_{kj}^2} = 0 \iff \sum_{C(x_i)=k} \frac{\mu_{kj}^2 - x_{ij}^2}{\mu_{kj}^2} = 0,\ \forall\ k \in [1..K] \text{ et } \forall\ j \in [1..20].$$

On obtient alors l'équation du calcul des vecteurs moyens :

$$\mu_{kj} = \sqrt{\frac{\sum_{C(x_i)=k} x_{ij}^2}{\#k}}, \ \forall \ \ j \in [1..20].$$

c) Nombre de classes : Un paramètre important de cette méthode est le nombre de classes K. Nous avons expérimenté plusieurs valeurs pour ce paramètre, et construit trois *clustering* différents en 50, 100 et 200 classes. Dans des expériences préliminaires, nous avons réalisés des *clustering* en 500 ou 1 000 classes. Cependant, ces *clustering* conduisent à la présence de nombreuses classes de cardinal très faible (voire vides) et à une détérioration des résultats (données non-présentées).

À l'issue du *clustering*, on constate que la plupart des classes apprises correspondent soit à un acide aminé unique soit à des catégories précises d'acides aminés (propriétés physico-chimiques similaires), à l'exception toutefois de classes sans profil particulier qui correspondent aux positions ne subissant, *a priori*, pas de contraintes évolutive (*cf.* figure 5.8).

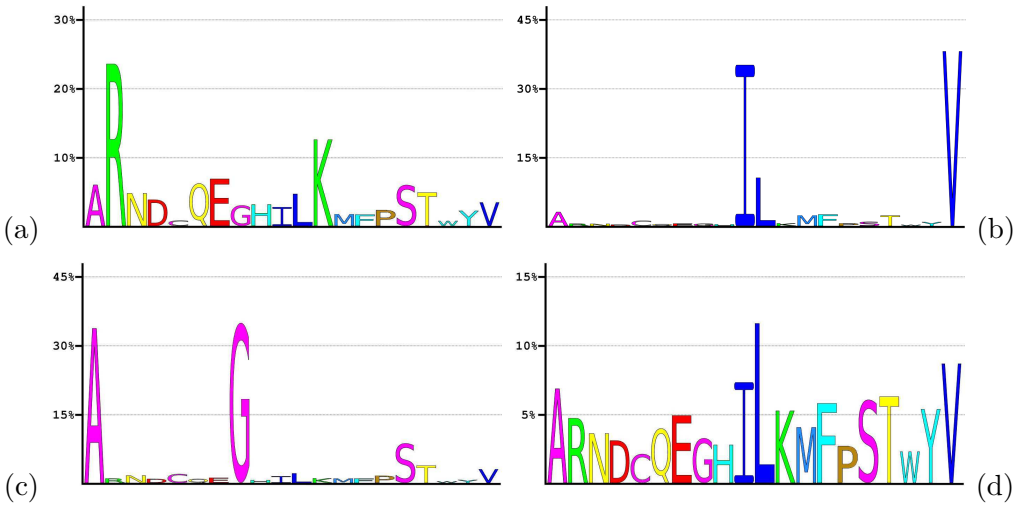


FIGURE 5.8 – **Logo des distributions en acides aminés de quatre classes obtenues par un *clustering* pour K=50**. On reconnaît, à travers ces distributions, les propriétés physico-chimiques contraintes dans ces états — acides aminés chargés positivement en (a), aliphatiques en (b) et minuscules en (c) — ainsi que l'absence de propriétés conservées en (d) (*cf.* figure 2.4 66 pour le code couleur).

5.9.3 Estimation des distributions associées aux différents classes d'états

La sélection des espèces dont on extrait les domaines connus afin d'estimer les distributions associées aux classes est également un paramètre important. Par "domaines connus", on entend les séquences détectées par HMMER grâce aux HMM profils de Pfam en respectant

les seuils de score recommandés. Deux ensembles de séquences ont été considérés pour nos expérimentations :

- les domaines connus chez *P. falciparum* ;
- les domaines connus dans l'ensemble des séquences d'alvéolés [1] (*cf.* Figure 5.4 page 135).

Les chemins de Viterbi, extraits des résultats d'HMMER, nous permettent d'associer les différents états *Matchs* des modèles avec les acides aminés qui y sont alignés dans les domaines connus des espèces sélectionnées. Pour une classe d'états, on utilise alors l'ensemble des acides aminés alignés sur les états membres de cette classe pour estimer une nouvelle distribution. On notera qu'en pratique, pour optimiser l'algorithme de *clustering* décrit dans la section précédente (5.9.2), la classification est uniquement construite à partir d'états où au moins un acide aminé a pu être aligné d'après les alignements de Viterbi de l'ensemble des domaines connus considérés.

Pour éviter la présence de probabilités nulles au sein de cette distribution, on applique un lissage grâce à la mixture de Dirichlet définit par (Sjölander *et al.*, 1996) (utilisé par défaut dans Pfam, *cf.* section 2.3.7). On obtient ainsi des distributions alliant les propriétés physico-chimiques conservées de la classe avec la divergence de *P. falciparum*.

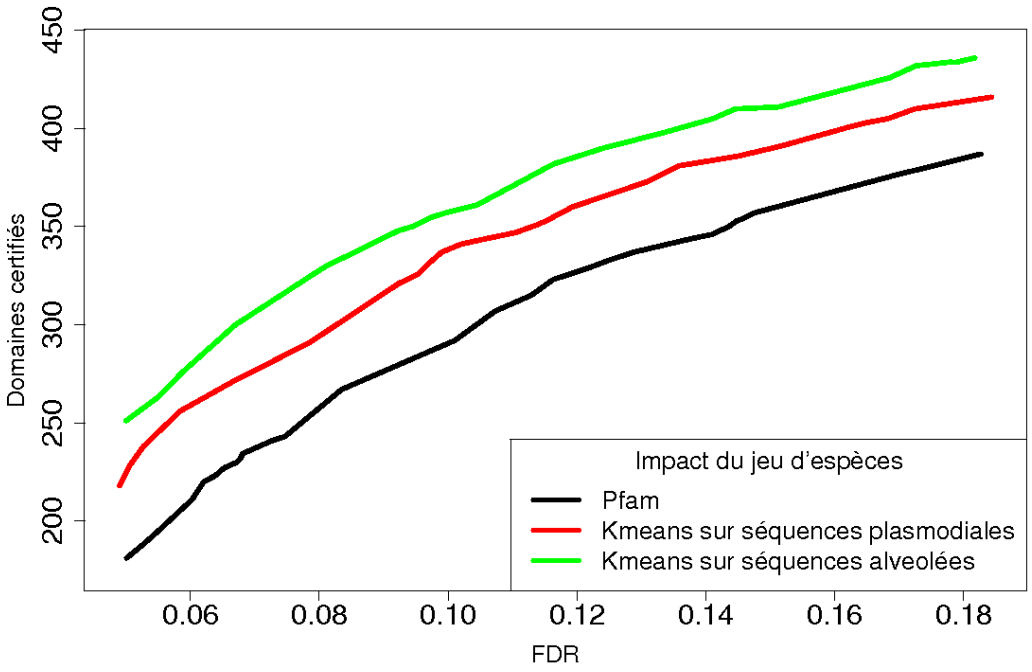
5.9.4 Correction des modèles

Enfin, on réalise la correction des modèles Pfam en modifiant successivement tous les états *Matchs* de la librairie Pfam. Pour chaque état, on détermine sa classe d'appartenance et on combine sa distribution d'acides aminés d'origine avec la distribution estimée pour cette classe grâce à un mélange de ces deux distributions. Différentes proportions ont été expérimentées pour ce mélange : 75%-25%, 50%-50% et 25%-75%. On peut alors s'interroger sur l'utilité de la mixture de Dirichlet appliquée pour lisser les distributions associées aux classes d'états (*cf.* section 5.9.3), celles-ci étant toujours mélangées avec la distribution d'origine de l'état à corriger. L'expérience a montré qu'en l'absence de ce lissage, on observe une dégradation des résultats de certification (données non présentées).

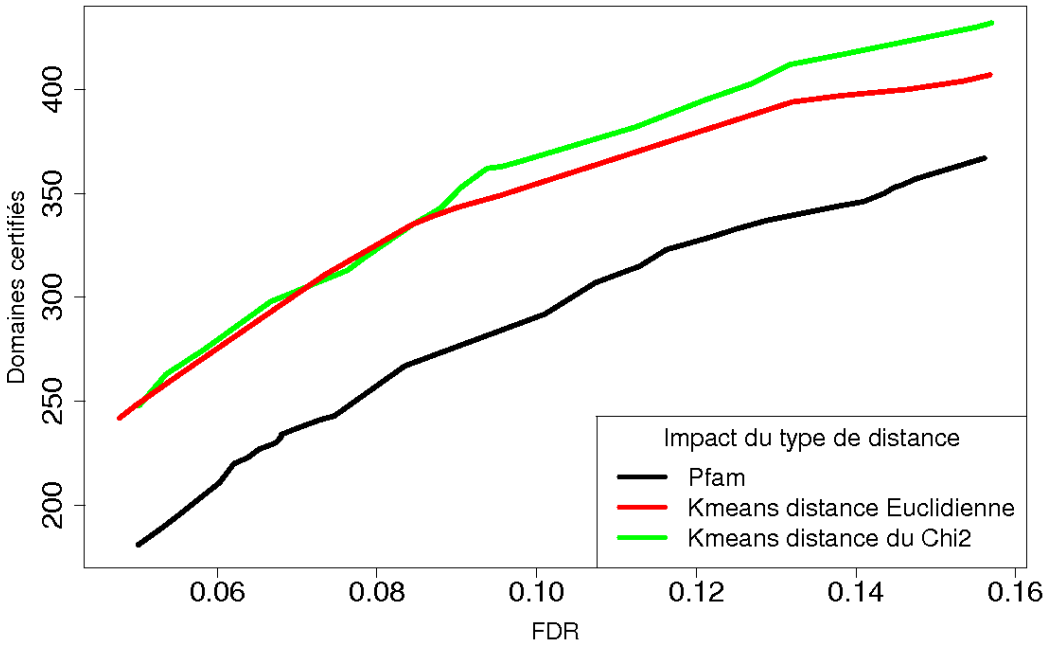
5.9.5 Modèle nul

Comme pour les approches précédentes différents modèles nuls ont été testés. Le premier est le modèle nul par défaut d'HMMER et de Pfam. Le second modèle nul possède des probabilités de générations égales à la composition globale en acides aminés des protéines de *P. falciparum*. Enfin un modèle nul intermédiaire a été utilisé dans ces séries d'expériences. La distribution associée à ce troisième modèle nul, dit pondéré, correspond au mélange des distributions des deux précédents modèles nuls. Ces distributions sont mélangées dans les mêmes proportions que lors de la correction des états *Matchs* des HMM profils, soit par exemple 75% du modèle nul de Pfam et 25% du modèle nul "Global" pour une pondération des états 75-25% (75% de l'état originel et 25% de la distribution associée à sa classe d'appartenance).

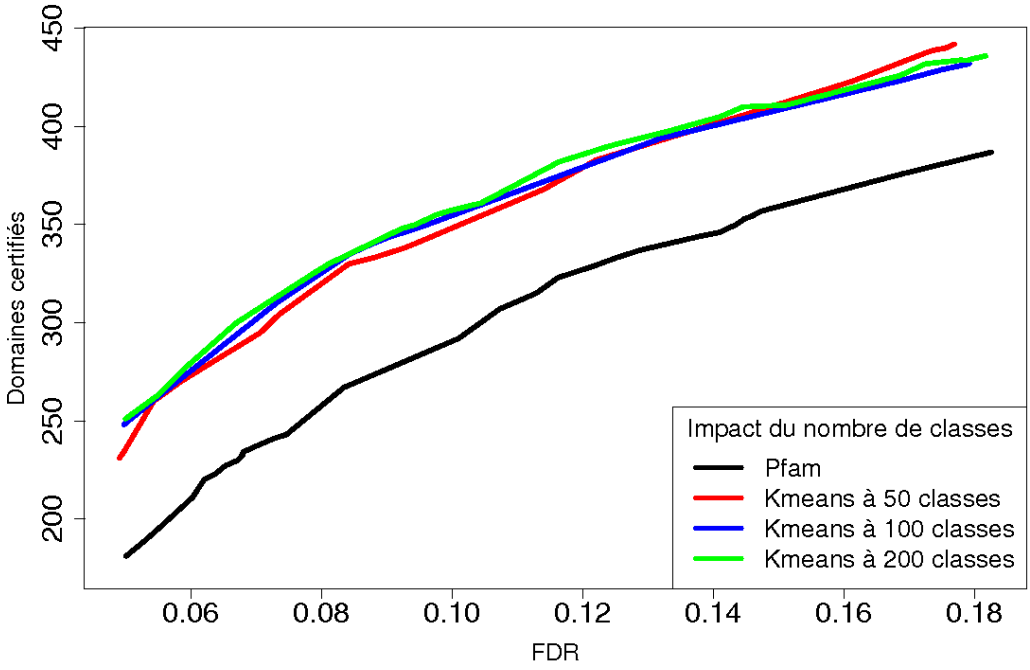
1. Les séquences protéiques non-redondantes des alvéolés ont été extraites du site Web du NCBI (Sayers *et al.*, 2009) grâce au *Taxonomy Browser* : http ://www.ncbi.nlm.nih.gov/Taxonomy/.


Domaines certifiés
200
250
300
350
400
450
0.06
0.08
0.10
0.12
0.14
0.16
0.18
FDR
Impact du jeu d'espèces
Pfam
Kmeans sur séquences plasmodiales
Kmeans sur séquences alveolées

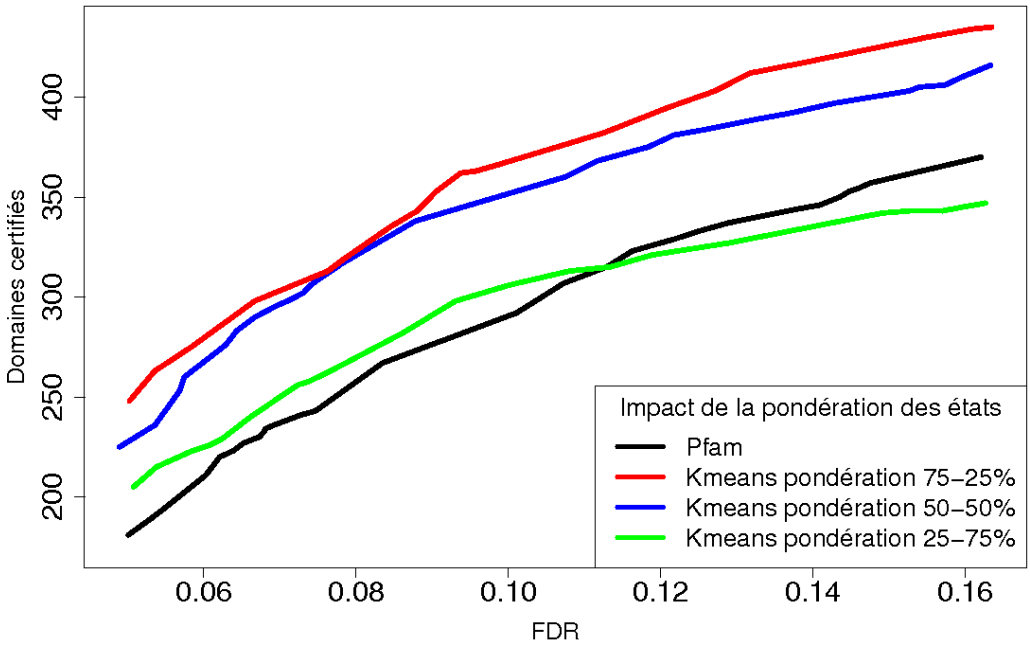
(a)


Domaines certifiés
200
250
300
350
400
0.06
0.08
0.10
0.12
0.14
0.16
FDR
Impact du type de distance
Pfam
Kmeans distance Euclidienne
Kmeans distance du Chi2

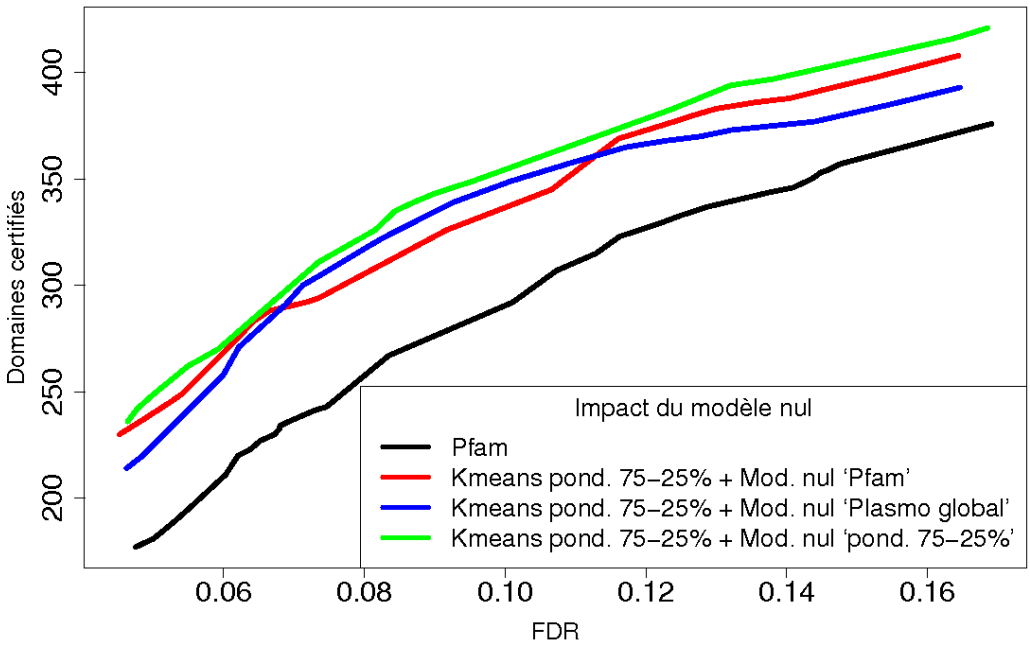
(b)


Domaines certifiés
200
250
300
350
400
450
0.06
0.08
0.10
0.12
0.14
0.16
0.18
FDR
Impact du nombre de classes
Pfam
Kmeans à 50 classes
Kmeans à 100 classes
Kmeans à 200 classes

(c)


Domaines certifiés
200
250
300
350
400
0.06
0.08
0.10
0.12
0.14
0.16
FDR
Impact de la pondération des états
Pfam
Kmeans pondération 75–25%
Kmeans pondération 50–50%
Kmeans pondération 25–75%

(d)



(e)

FIGURE 5.9 – **Résultats de certification par co-occurrence des librairies de HMM profils corrigés par *K-means*.** La figure (a) illustre l'impact du jeu d'espèces : les séquences plasmodiales de *P. falciparum* ou d'alvéolées (respectivement en rouge et en vert). La figure (b) compare distance euclidienne (en rouge) et distance du χ^2 (en vert). La figure (c) illustre l'impact du nombre de classes, c'est à dire le paramètre K pour des valeurs expérimentées de 50, 100 et 200, respectivement représentées par les courbes rouge, bleue et verte. La figure (d) permet de comparer les différentes pondérations lors du mélange des distributions originales et apprises. Ces librairies sont construites pour différentes pondérations des états *Matchs* (distribution originale-distribution observée) de 75%-25% (en rouge), 50%-50% (en bleu) et 25%-75% (en vert). La figure (e) illustre l'effet du modèle nul sur les librairies de modèles. Les trois modèles nuls comparés sont celui par défaut d'HMMER (en rouge), celui ayant pour distribution la composition moyenne des protéines de *P. falciparum* (en bleu), et le modèle nul "pondéré" (en vert) correspondant au mélange des probabilités de génération des deux précédents modèles nuls (proportions identiques à la correction des états *Matchs*). Les différentes librairies sont obtenues par un *clustering* en 200 classes pour la figure (a) et en 100 classes pour les figures (b), (d) et (e) ; en utilisant la distance euclidienne pour les figures (a), (c) et (e) et la distance du χ^2 pour la figure (c) ; sur les séquences alvéolées (Figures (b) à (e)) avec une pondération des états de 75%-25% (Figures (a), (b), (c) et (e)) et le modèle nul pondéré (Figures (a) à (d)).

5.9.6 Résultats

Les résultats obtenus sont présentés dans la figure 5.9. La première constatation est qu'en comparaison des précédentes corrections, on observe ici des librairies corrigées qui se détachent plus nettement des résultats de la librairie originale Pfam. Nous comparons tout d'abord les deux jeux d'espèces : *Plasmodium falciparum* et *Alveolata*. Les librairies obtenues grâces aux séquences alvéolées semblent être les plus performantes.

Ensuite, l'impact du type de distance choisi, χ^2 ou euclidienne, est comparé (Figure 5.9.(b)). Ce paramètre ne semble pas avoir d'impact crucial sur les résultats obtenus, avec cependant un léger avantage pour la distance du χ^2.

Concernant la valeur du paramètres K déterminant le nombre de classes, l'écart entre les valeurs expérimentées ne semble pas avoir un impact majeur sur les résultats. Les librairies construites en faisant varier ce paramètres certifient, à FDR équivalent, à peu près le même nombre de domaines, comme illustré dans la figure 5.9.(c).

Nous faisons alors varier la pondération des états corrigés (Figure 5.9.(d)). On constate que les librairies qui certifient le plus de domaines, à même FDR, sont celles issues d'une pondération 75% de l'état original et 25% de la distribution observée ; puis 50% de chacune des deux distributions ; et enfin on trouve les modèles composés à 25% de l'état original et 75% de la distribution observée. Une optimisation du mélange des états est donc une piste à suivre pour affiner cette approche.

La comparaison des différents modèles nuls suggère que le modèle nul pondéré semble le mieux adapté, certifiant un plus grand nombre de domaines à FDR équivalent (Figure 5.9.(e)).

Finalement, les paramètres optimaux de cette méthode retenus pour la suite sont : un *clustering* en 100 classes, l'utilisation du χ^2 pour les mesures de distances, l'apprentissage d'une distribution associée aux états grâce aux domaines connus chez les alvéolés et un mélange de proportion 75%-25% (distribution originale-distribution observée) pour les probabilités de génération associées aux états *Matchs* et pour le modèle nul.

5.10 Utiliser les *k*-plus proches états

Le principe de cette méthode est similaire a celui du *K-means*. Une fois encore, nous nous inspirons des séquences de domaines connus et de leurs alignements sur les états des HMM pour apprendre des distributions en vue de corriger tous les états *Matchs* de la librairie Pfam. Cependant, au lieu de réaliser une classification des états *a priori* pour estimer les nouvelles distributions, on estime une distribution différente pour chaque état *Match*. Pour réaliser cela, on se sert des états *Matchs* les plus similaires pour lesquels une occurrence d'un domaine est connue chez *P. falciparum*. Une procédure classique pour ce genre d'approche est celle d'un *k-nearest neighbor* ou k-plus proches voisins (Fix et Hodges, 1951). Nous présentons le principe de cette méthode (section 5.10.1), les paramètres qui entrent en jeu (section 5.10.2), et les résultats obtenus par les librairies corrigées (section 5.10.4).

5.10.1 Principe

L'algorithme des k-plus proches voisins fait partie des méthodes d'apprentissage supervisé (Mitchell, 1997). Notre méthode de correction utilisant les k-plus proches voisins est assez semblable à celle du *K-means* dans son principe. Elle se distingue toutefois par l'absence de construction d'un classificateur (précédemment les vecteurs moyens). Quel que soit l'individu, il est uniquement caractérisé à travers un nombre fixé k d'autres individus (les voisins). La procédure de correction par k-plus proches voisins est la suivante :

1. On collecte l'ensemble des individus d'apprentissage. Ce sont les états *Matchs* de la librairie Pfam pour lesquels on dispose d'au moins un acide aminé observé d'après les alignements de domaines connus provenant de l'organisme cible ou d'une de ses espèces proches. Soit O cet ensemble d'individus pour lesquels on garde en mémoire les acides aminés observés alignés dans cet état grâce à une fonction $f : O \rightarrow \Sigma^*$.
2. pour chaque distribution x_i associée à l'un des N états de la librairie Pfam, on détermine O^{x_i}, les k individus de l'ensemble O les plus proches en terme de distance à x_i :

$$O^{x_i} \subset O, \quad |O^{x_i}| = k, \text{ et } \forall\, o \in O^{x_i}, \nexists\, p \in \{O - O^{x_i}\} \text{ tel que } dist(p, x_i) < dist(o, x_i).$$

3. La distribution en acides aminés attendue pour x_i est approximée *via* l'ensemble des acides aminés associés aux k éléments de O^{x_i} en cumulant les ensembles mémorisés par la fonction f pour ces k éléments :

$$\widehat{f}(x_i) = \bigcup_{o \in O^{x_i}} f(o).$$

   Cet ensemble d'acides aminés est transformé en une distribution de probabilités en utilisant la mixture de Dirichlet de Pfam (Sjölander *et al.*, 1996) (*cf.* section 2.3.7) pour éviter les probabilités nulles.
4. On modifie les probabilités de génération associée à chaque état x_i, en mélangeant sa distribution d'origine avec la distribution apprise par k-plus proches voisins, de la même manière que pour la méthode de correction par *K-means*.

En répétant ce traitement à l'ensemble états *Matchs* de la librairie Pfam, on obtient une librairie de modèles corrigée par k-plus proches voisins.

5.10.2 Paramètres de la méthode

De nombreux paramètres entrent en jeu dans cette méthode :

– **Ensemble d'apprentissage :** La première étape de la méthode consiste à collecter l'ensemble des états O pour lesquels un ou plusieurs acides aminés ont pu être alignés dans un jeu de séquences protéiques. Deux jeux de séquences protéiques ont été considérés dans nos expérimentations : les alignements des domaines Pfam connus chez *P. falciparum* et ceux connus chez les alvéolés. Les ensembles qui correspondent à ces jeux de séquences protéiques, notés $O_{P.falciparum}$ et $O_{Alveolata}$, sont constitués respectivement d'environ 247 000 et 470 000 éléments.

– **Nombre de voisins :** Le principal paramètre de la méthode est le nombre k de voisins que l'on recherche pour chaque ensemble O^{x_i}. Trois valeurs différentes de k ont été testées : 2, 10 et 50.

– **Mesure de distance :** Plusieurs formules sont envisageables pour évaluer les distances entre les distributions de probabilités, comme on l'a vu pour les *K-means*. Les deux distances retenues pour nos expériences sont la distance euclidienne et la distance du χ^2.

– **Proportion du mélange :** La correction de chaque état *Match* nécessite de choisir les proportions du mélange entre sa distribution initiale et la distribution observée dans ses k plus proches voisins. Comme pour la méthode de correction précédente, trois proportions ont été expérimentés : 75%-25%, 50%-50% et 25%-75%.

– **Modèle nul :** Les trois modèles nuls testés sont identiques à ceux de la correction par *K-means* : le modèle nul par défaut de Pfam, le modèle nul ayant la composition globale de *P. falciparum* et le modèle nul pondéré correspondant à un mélange des distributions des deux précédents modèles (avec des proportions identiques au mélange des états *Matchs*).

5.10.3 Optimisation du calcul

Une procédure d'optimisation a été nécessaire pour l'implémentation de cette méthode de correction. Le calcul des ensembles O^{x_i} nécessite la comparaison de tous les états des HMM de Pfam avec chacun des états de l'ensemble O. Cette étape prend un temps considérable et rend l'approche par k-plus proches voisins extrêmement longue. Par exemple chez *P. falciparum*, il faut rechercher pour environ 2 150 000 états (nombre total d'états dans la librairie Pfam version 23.0) les k états les plus proches parmi 247 000 éléments de O, soit plus de $53 \cdot 10^{10}$ opérations de calculs de distance. Il s'agit d'un problème bien connu des k-plus proches voisins (*cf.* sections 4.5.5 de Duda *et al.* (2001) et 13.5 de Hastie *et al.* (2001)). Une solution à ce problème consiste à n'utiliser qu'un sous-ensemble de O, choisit aléatoirement (échantillon). Il est également possible de restreindre les dimensions des vecteurs à comparer, en se concentrant par exemple sur les 2 (ou 4) acides aminés de plus forte probabilité. Ce type de traitement, bien que nécessitant l'identification pour chaque état de ses composants majeurs, permettrait sur notre exemple de diviser par 10 (resp. 5) le temps de calcul en théorie. Une autre solution couramment utilisée, consiste à utiliser une partition de O préalablement calculée. Ici, on va donc utiliser le *clustering* obtenu par la procédure des *K-means* vue précédemment (section 5.9). Pour chaque état *Match*, l'identification de ses plus proches voisins est réalisée uniquement parmi les états appartenant à la même classe [2] dans un premier temps. Ensuite, on vérifie dans la deuxième classe la plus similaire, qu'il n'existe pas d'état voisin plus proche que ceux de la précédente. Si c'est le cas, on calcule alors les plus proches voisins en tenant compte des deux classes et on réitère la vérification dans la classe suivante la plus proche. Grâce à ce pré-traitement, en utilisant un *clustering* en K classes (par exemple K=100), le nombre de tests et donc le temps de calcul sont généralement divisés par $\frac{K}{2}$.

Malgré cette optimisation, la construction d'une librairie corrigée par k-plus proches voisins dure généralement plusieurs jours, là où les autres corrections de modèles ne prennent

2. La distance utilisée pour déterminer les plus proches voisins doit être la même que pour l'apprentissage du *clustering*.

que quelques minutes. Le nombre de paramètres à régler devient alors un frein important à l'exploitation de la méthode.

5.10.4 Résultats

Les résultats de la méthode de correction par k-plus proches voisins sont présentés sur la figure 5.10. Comme pour la méthode de correction par *K-means*, les différents graphiques illustrent l'impact des paramètres sur les résultats de la méthode de certification par co-occurrence. Tout d'abord, il semble que l'on certifie un nombre sensiblement plus grand de domaines à FDR équivalent lorsque l'on s'appuie sur les données d'*Alveolata*, plutôt que simplement sur les données de *P. falciparum* (Figure 5.10.(a)). Le choix du type de distance ne semble pas avoir d'impact majeur sur les modèles construits, et le modèle nul pondéré permet de certifier un plus grand nombre de nouveaux domaines à FDR équivalent, comme l'atteste la figure 5.10.(b). Ensuite, l'impact du nombre k de voisins est illustré par la figure 5.10.(c). On constate en fixant tous les autres paramètres que plus le nombre de voisins est important, plus la librairie semble performante. Cependant nous n'avons pas pu tester de valeur au-delà de 50 voisins pour des questions de temps (*cf.* section 5.10.3) et l'optimisation de ce paramètre reste à établir. Enfin, à l'instar de la correction par *K-means*, on observe que le mélange des probabilités associées aux états par une proportion 75%-25% des distributions initiale et observée conduit à des certifications plus nombreuses à FDR équivalent que les autres proportions (Figure 5.10.(d)) et que le modèle nul qui semble le mieux adapté à ce genre de méthode de correction est le modèle nul pondéré (Figures 5.10.(e)).

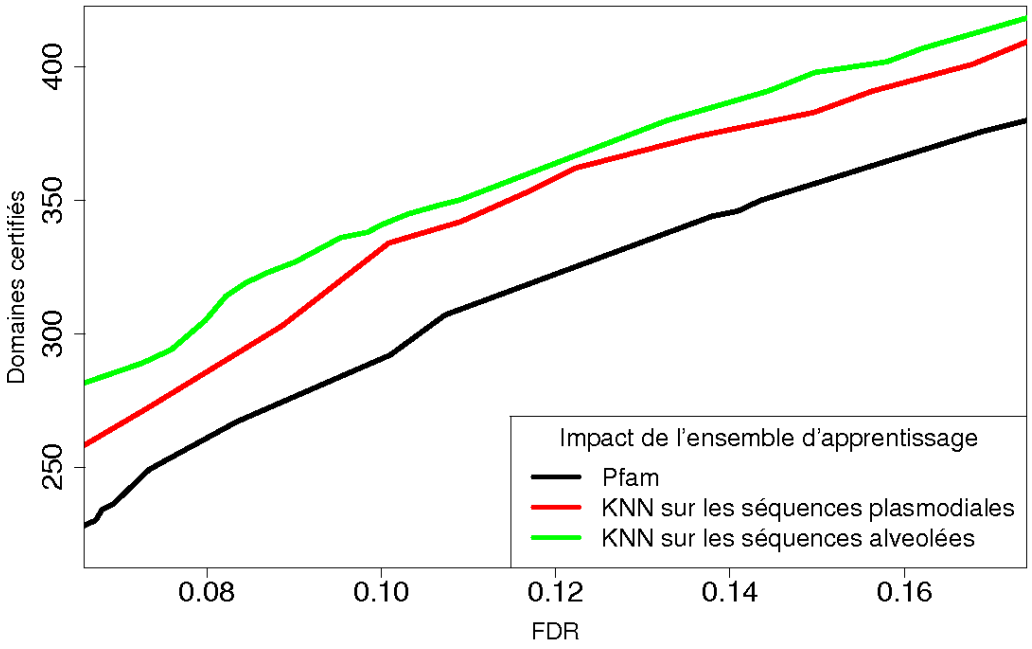
Pour conclure, on retient pour la suite les paramètres suivants comme optimaux : un nombre de 50 états voisins, l'utilisation de la distance euclidienne, l'apprentissage d'une distribution associée aux voisins grâce aux domaines connus chez les alvéolés et un mélange de proportion 75%-25% (distribution originale-distribution observée) pour les probabilités de génération associées aux états *Matchs* et pour le modèle nul.

5.11 Comparaison des différentes approches

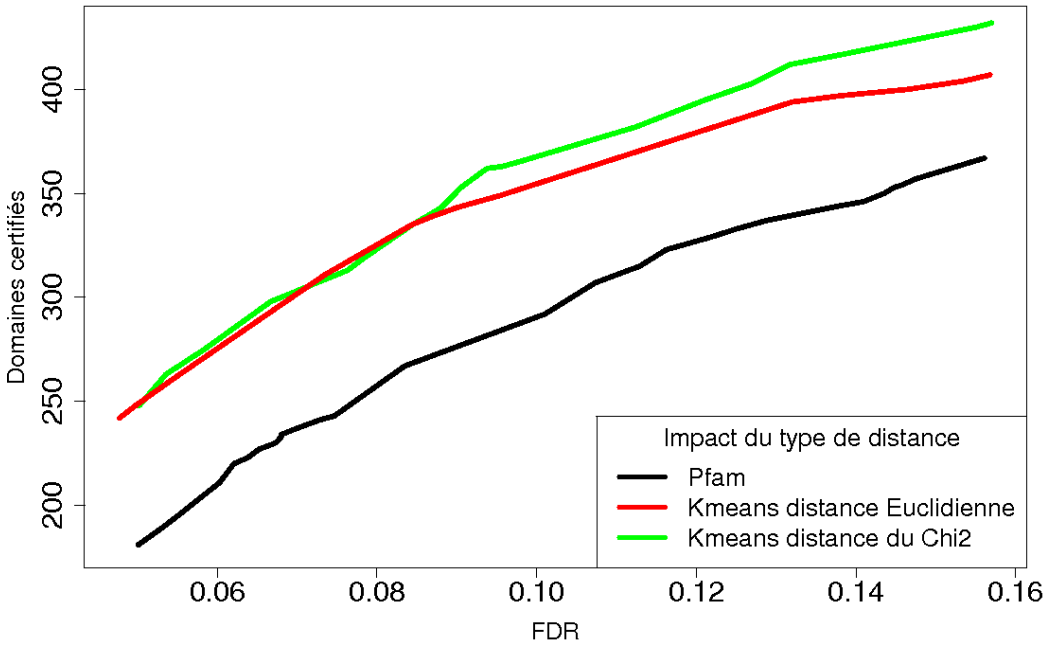
Dans les précédentes sections, nous avons proposé différentes méthodes de correction des HMM profils de la librairie Pfam. Ces approches disposent de caractéristiques différentes et, par conséquent, obtiennent des résultats également différents. Dans un premier temps, nous revenons brièvement sur ce qui distingue ces approches puis nous discutons de manière plus détaillée des résultats qu'elles obtiennent.

5.11.1 Des facultés différentes

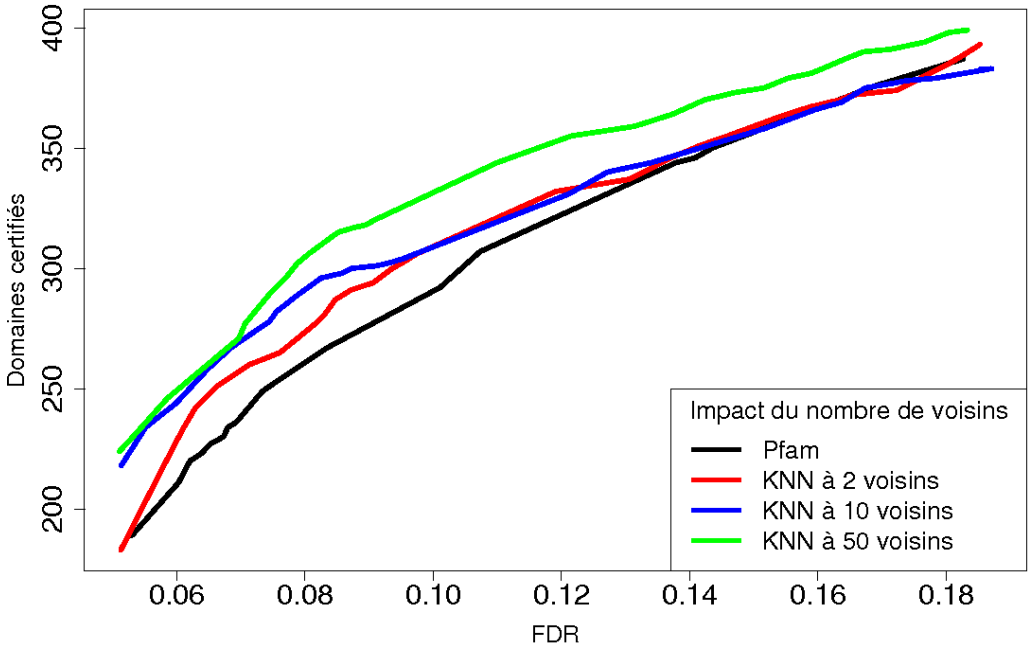
L'apprentissage de modèles de domaines *espèce-dédiés* est une solution naturelle dont les résultats sont très intéressants. Cette méthode de correction souffre cependant d'une limitation importante liée à l'identification au préalable d'occurrences des domaines dans les espèces proches de l'organisme cible. Cette approche ne peut donc être appliquée que si un certain nombre d'espèces proches de notre organisme cible sont séquencées. De plus, elle ne peut


Domaines certifiés
250
300
350
400
0.08
0.10
0.12
0.14
0.16
FDR
Impact de l'ensemble d'apprentissage
Pfam
KNN sur les séquences plasmodiales
KNN sur les séquences alveolées

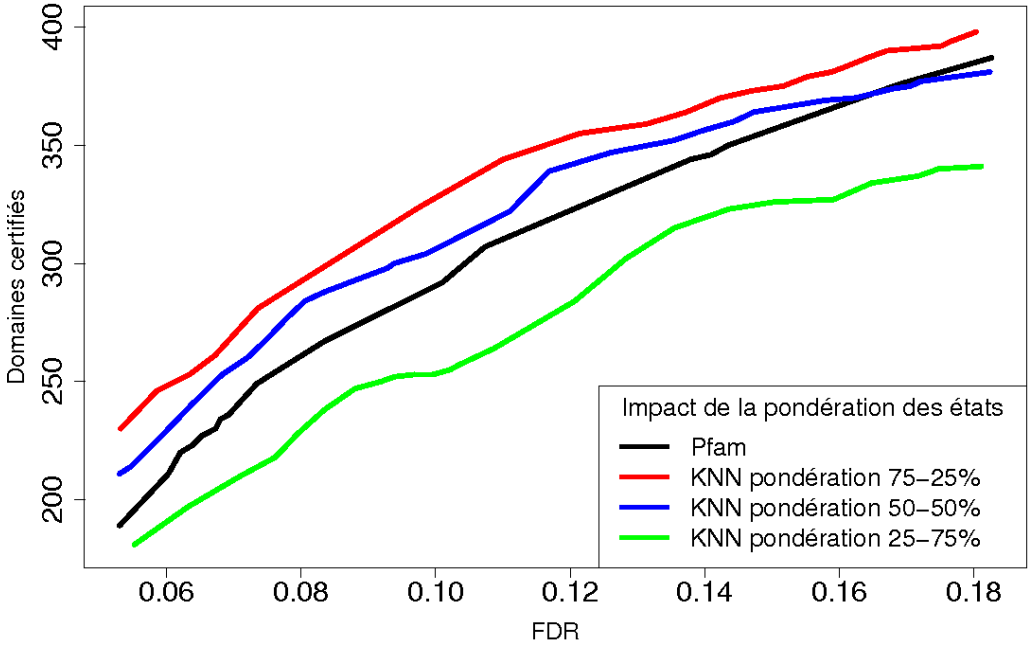
(a)


Domaines certifiés
200
250
300
350
400
0.06
0.08
0.10
0.12
0.14
0.16
FDR
Impact du type de distance
Pfam
Kmeans distance Euclidienne
Kmeans distance du Chi2

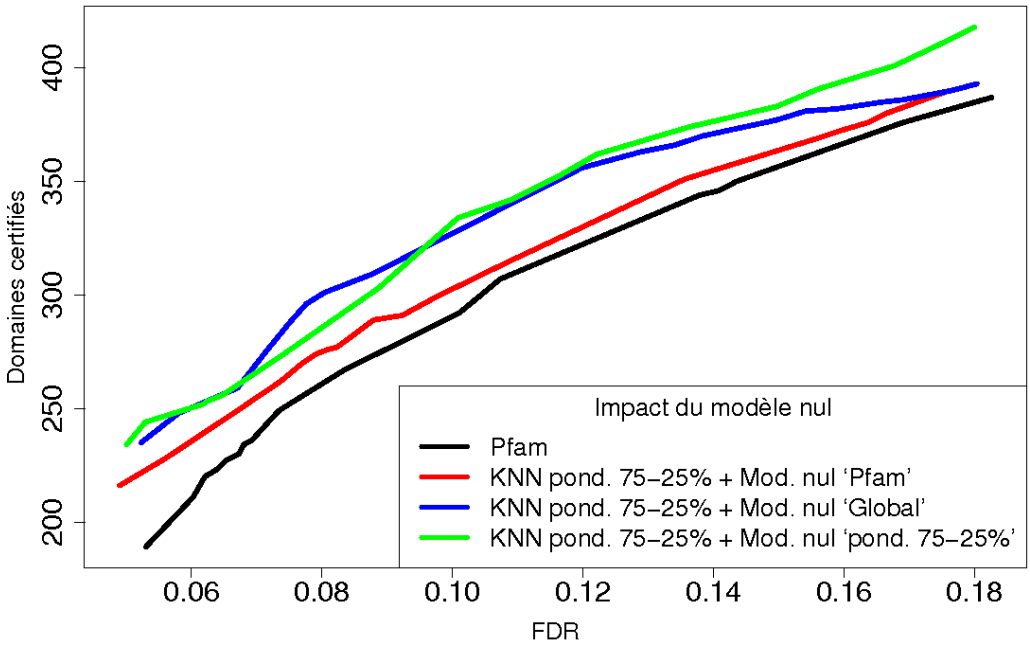
(b)


Domaines certifiés
200
250
300
350
400
0.06
0.08
0.10
0.12
0.14
0.16
0.18
FDR
Impact du nombre de voisins
Pfam
KNN à 2 voisins
KNN à 10 voisins
KNN à 50 voisins

(c)


Domaines certifiés
200
250
300
350
400
0.06
0.08
0.10
0.12
0.14
0.16
0.18
FDR
Impact de la pondération des états
Pfam
KNN pondération 75–25%
KNN pondération 50–50%
KNN pondération 25–75%

(d)



(e)

FIGURE 5.10 – **Résultats de certification par co-occurrence des librairies de HMM profils corrigés par *k*-plus proches voisins**. La figure (a) illustre l'impact de l'ensemble d'apprentissage : les séquences plasmodiales de *P. falciparum* ou d'alvéolées (respectivement en rouge et en vert). La figure (b) compare distance euclidienne (en rouge) et distance du χ^2 (en vert). La figure (c) illustre l'impact du nombre de voisins avec des valeurs expérimentées de 2, 10 et 50, respectivement représentées par les courbes rouge, bleue et verte. La figure (d) permet de comparer les différentes pondérations lors du mélange des distributions originales et apprises. Ces librairies sont construites pour différentes pondérations des états *Matchs* (distribution originale-distribution observée) de 75%-25% (en rouge), 50%-50% (en bleu) et 25%-75% (en vert). La figure (e) illustre l'effet du modèle nul sur les librairies de modèles. Les trois modèles nuls comparés sont celui de Pfam (en rouge), celui ayant pour distribution la composition moyenne des protéines de *P. falciparum* (en bleu), et le modèle nul "pondéré" (en vert). Les différentes librairies sont obtenues en fixant le nombre de voisins à 50 pour les figures (a), (b), (d) et (e) ; en utilisant la distance euclidienne pour les figures (c) et (d) et la distance du χ^2 pour les figures (a) et (e) ; sur les séquences de *P. falciparum* (Figures (b) à (e)) avec une pondération des états de 75%-25% (Figures (a), (b), (c) et (e)) et le modèle nul pondéré pour la figures (a) et le modèle nul global pour les figures (b) à (d)).

améliorer la détection de domaines dont on ne connaît aucune occurrence dans ces espèces. Pour pallier cette limitation, nous avons mis en place les autres méthodes de corrections des modèles.

Les approches de modification des distributions des états *Matchs* permettent de corriger l'ensemble des modèles d'une librairie, grâce à l'apprentissage de règles de correction générales. Si cet objectif est identique pour les différentes variantes proposées, elles se distinguent cependant par leurs fondements théoriques pour simuler l'évolution artificielle des différents positions des domaines. Tout d'abord, les facteurs de correction ne sont que de simples corrections mathématiques visant à réajuster une distribution global en acides aminés (celle de l'ensemble des états des modèles) par une autre (celle de l'organisme cible). Pour introduire un peu de biologie dans notre méthode, nous avons ensuite recouru à des matrices de substitution d'acides aminés. Cependant, les schémas d'évolution à partir desquels ont été estimés ces matrices ne sont pas forcément adaptés à des protéines aussi divergentes que celles de *P. falciparum*. Afin d'essayer de capturer les spécificités évolutives de cet organisme, nous avons finalement introduit les méthodes de corrections par *K-means* et *k*-plus proches voisins. Ces deux dernières méthodes possèdent toutefois deux inconvénients : elles font appel à des techniques d'apprentissage coûteuses en temps de calcul, et elles nécessitent l'optimisation de nombreux paramètres.

5.11.2 Des résultats différents

Comme discuté plus tôt dans ce chapitre (*cf.* section 5.3), l'évaluation des différentes librairies de modèles est une tâche difficile. Jusqu'à présent, nous avons utilisé la méthode de certification par co-occurrence (présentée au chapitre 4) afin d'optimiser les paramètres de chaque approche de corrections. Nous avons donc retenu, pour chaque méthode, le paramétrage qui permet de certifier le plus grand nombre de domaines à FDR équivalent :

- pour les espèces proches, cela correspond à la reconstruction de modèles grâce à l'intégration des séquences des domaines connus d'*Alveolata* dans les alignements-graines de Pfam.
- pour les facteurs de corrections, cela correspond à la composition globale de *P. falciparum* comme distribution cible et comme modèle nul.
- pour les matrices du substitutions, cela correspond à la composition en acides aminés issue de (Pizzi et Frontali, 2001) avec $t = 0.1$.
- pour les *K-means*, cela correspond à la librairie obtenue par un *clustering* en 100 classes (χ^2 utilisée pour la mesure de distance) des états où au moins un résidu a été observé dans les séquences d'alvéolés, avec un mélange de proportion 75%-25% (distribution originale-distribution observée) des probabilités de génération associées aux états *Matchs* et du modèle nul.
- pour les *k*-plus proches voisins, cela correspond à la librairie obtenue pour 50 états voisins (en utilisant la distance euclidienne) où au moins un résidu a été observé dans les séquences d'alvéolés, avec un mélange de proportion 75%-25% (distribution originale-distribution observée) des probabilités de génération associées aux états *Matchs* et du modèle nul.

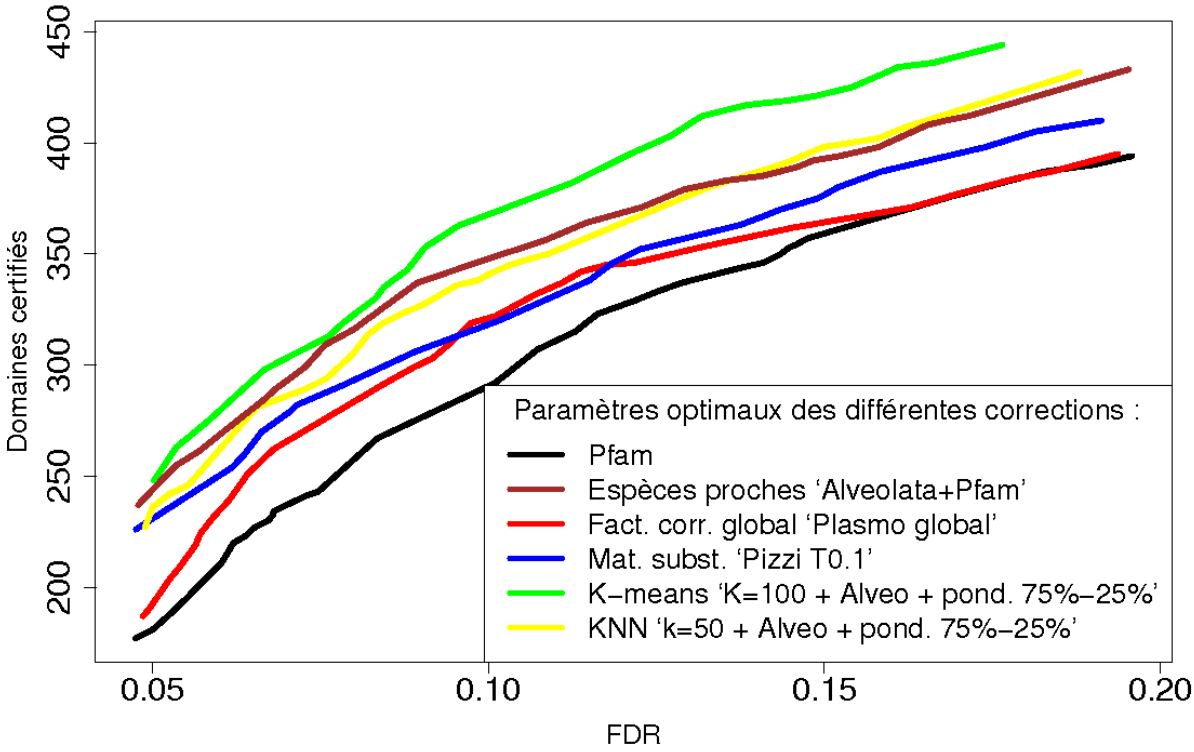


FIGURE 5.11 – **Résultats de certification par co-occurrence des librairies de HMM profils corrigés, exhibant globalement le plus grand nombre de domaines certifiés à FDR équivalent, pour les différentes méthodes de correction proposées.** Nombre de certifications réalisées en fonction du FDR, en utilisant les domaines Pfam connus comme domaines validants.

La figure 5.11 présente sur un même graphique les résultats de certification de ces différentes librairies et de la librairie Pfam originale. On constate que chaque librairie corrigée exhibe, à FDR équivalent, un nombre de domaines certifiés supérieur à celle de Pfam. La librairie conduisant au plus grand nombre de certifications, quelque soit le seuil de FDR choisi, est celle obtenue par une correction de type *K-means*. Cependant, si globalement la correction par *K-means* peut sembler la meilleure approche, les autres ne sont pas forcément à dénigrer car elles peuvent apporter des résultats complémentaires. En effet, comparer la taille des ensemble de domaines certifiés ne signifie pas que ces ensembles soient strictement inclus les uns dans les autres. Parmi tous ces résultats, y compris ceux des méthodes certifiant *a priori* le moins de nouveaux domaines, peut se trouver un domaine unique à l'une des approches et dont le type est inédit dans l'organisme cible ou recèle un intérêt biologique significatif.

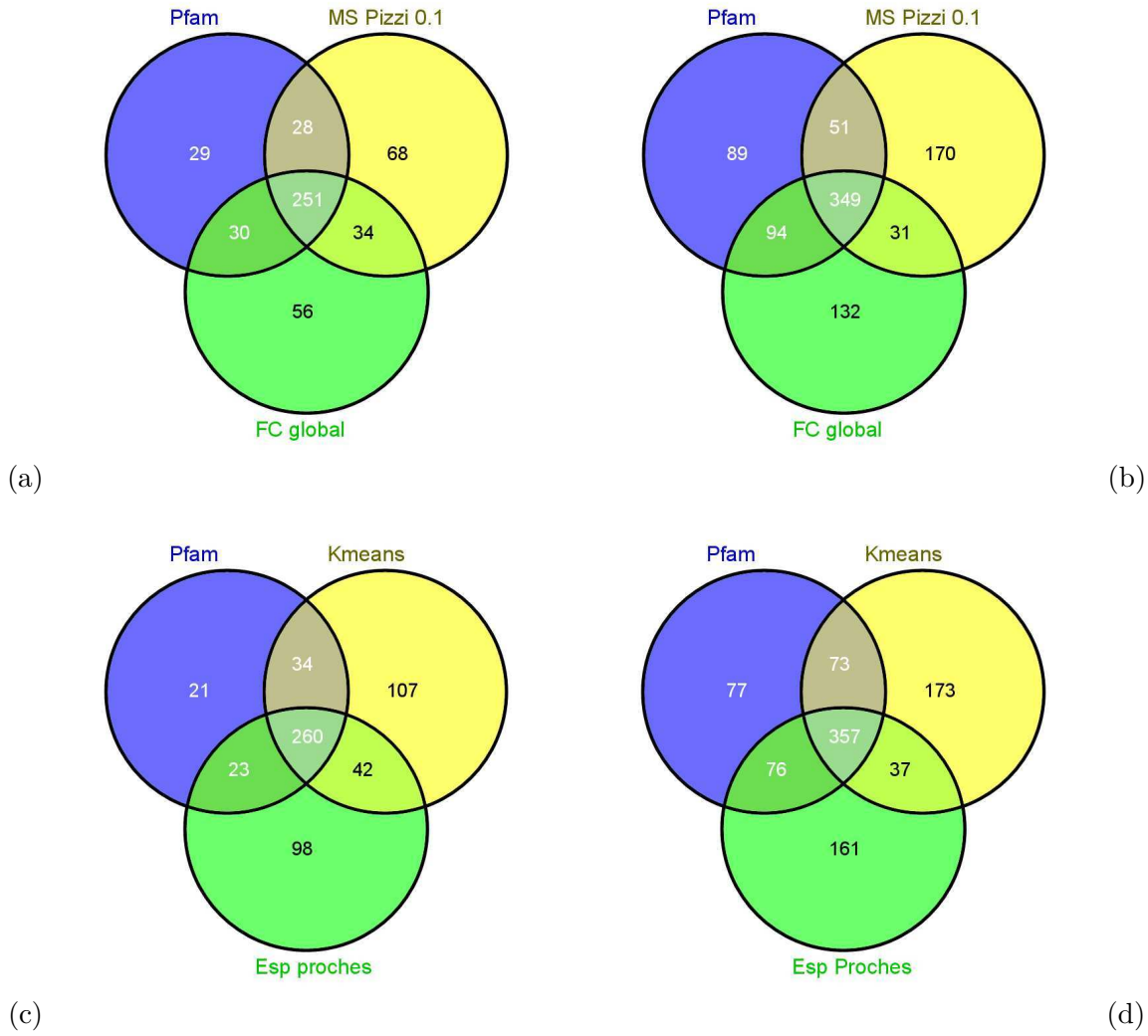


FIGURE 5.12 – **Diagramme de Venn des ensembles de domaines certifiés par différentes librairies corrigées de HMM profils.** Les diagrammes (a) et (b) représentent les ensembles de domaines certifiés, respectivement pour un FDR de 10% et 20%, par la librairie de Pfam (en bleu), celle corrigée par matrice de substitution (en jaune) et celle corrigée par facteurs de correction (en vert). De même, les diagrammes (c) et (d) correspondent aux ensembles certifiés respectivement à 10% et 20% pour les librairies Pfam (en bleu), celle corrigée par les *K-means* (en jaune) et celle obtenue en réapprenant les modèles sur les séquences d'espèces proches (en vert).

Nous nous sommes donc intéressés plus en détails aux ensembles de domaines certifiés par les différentes librairies afin d'estimer :

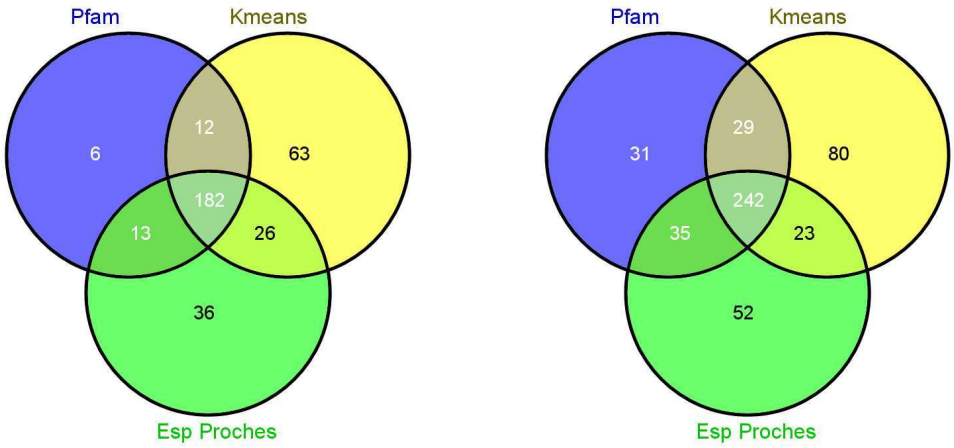
- Combien d'occurrences de domaines certifiés sont communes aux différentes librairies ?
- Combien sont spécifiques à l'une des librairies ?

La figure 5.12 représente, sous forme d'un diagramme de Venn, les occurrences certifiées par les différentes librairies (à l'exception de celle obtenue par les k-plus proches voisins dont les résultats sont très similaires à la librairie corrigée grâce aux *K-means*). On constate que, globalement, les librairies corrigées semblent inclurent une grande partie des domaines découverts par la librairie Pfam. Pour un FDR de 10%, la librairie corrigée grâce aux *K-means* certifie 87% des domaines obtenus grâce à la librairie Pfam. Ce taux avoisine les 83% pour les librairies corrigées par les espèces proches, par les facteurs de corrections et par une matrice de substitution, tandis que pour un FDR de 20%, on oscille entre 69% et 76% de domaines de la librairie Pfam également obtenus par les librairies corrigées. On remarque également que les ensembles de domaines certifiés par les différentes librairies ne sont pas identiques. À FDR équivalent, certains domaines certifiés sont communs à l'une ou l'autre des librairies corrigées mais environ un tiers semblent spécifiques : aucun ensemble n'en inclut strictement un autre. La méthode de correction par *K-means* qui exhibait la "meilleure" courbe de certification dans la figure 5.11, se distingue également ici en proposant le plus grand nombre de domaines originaux par rapport à la librairie Pfam. La principale information déduite de cette figure est que le vrai potentiel des corrections de modèles est bien plus important que celui promis par la "meilleure" courbe, bien qu'il nécessite toutefois l'intégration des résultats provenant des différentes librairies afin d'accéder à cet important ensemble de nouveaux domaines.

Nous avons ensuite approfondi l'étude des résultats pour nous intéresser aux types des nouveaux domaines certifiés. L'objectif est double :

- observer les types de domaines communs et spécifiques certifiés par les différentes librairies ;
- s'assurer que les résultats d'une librairie ne proviennent pas uniquement de la détection multiple (dans des protéines distinctes) d'un ensemble restreint de domaines.

La figure 5.13 rapporte les différents types de domaines certifiés par les librairies Pfam, corrigée par *K-means* et corrigée par ré-apprentissage sur les espèces proches, pour des FDR équivalents de 10% et 20%. Les conclusions que l'on peut en tirer sont assez proches des précédentes. Tout d'abord les résultats de la librairie Pfam sont majoritairement retrouvés par les librairies corrigées. Ensuite, la librairie corrigée par *K-means* propose une plus grande variété de types de domaines certifiés que les autres librairies. Toutefois, au final, chacune des librairies recèle des types de domaines certifiés qui lui sont spécifiques. Nous nous sommes alors intéressés aux types de domaines inédits c'est à dire considérés comme absents des protéines de *P. falciparum* jusqu'à présent. Leur identification souligne en effet le caractère inédit des résultats et donc le potentiel novateur des méthodes de correction. La figure 5.14 révèle que la correction par espèces proches permet de certifier des types de domaines inédits chez *P. falciparum* et spécifiques grâce à l'identification de ces types de domaines dans les espèces proches en dépit de leur absence dans l'organisme cible. Là encore, la librairie corrigée par *K-means* est celle qui identifie le plus de types de domaines inédits, dont près d'un tiers lui sont



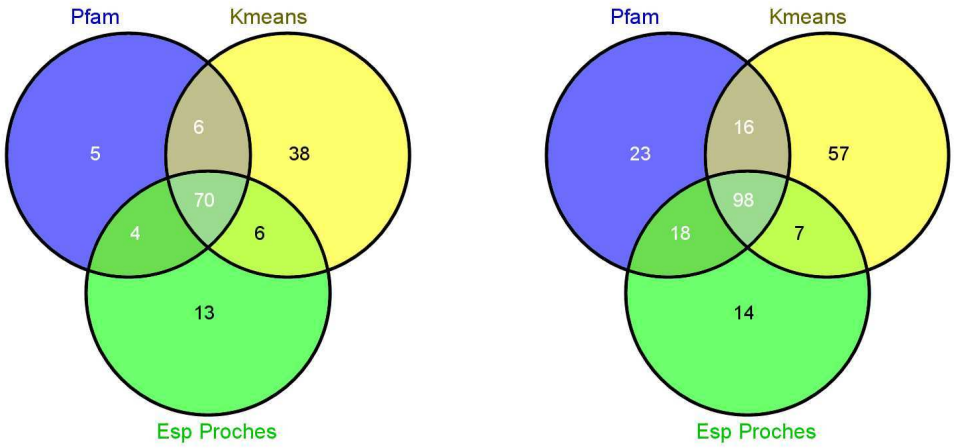
(a) (b)

FIGURE 5.13 – **Diagramme de Venn des types de domaines certifiés par différentes librairies corrigées de HMM profils.** Les diagrammes (a) et (b) correspondent aux ensembles certifiés respectivement à 10% et 20% pour les librairies Pfam (en bleu), celle corrigée par les *K-means* (en jaune) et celle obtenue en réapprenant les modèles sur les séquences d'espèces proches (en vert).

spécifiques. On observe donc des types de domaines communs mais aussi uniques à chacune des librairies. On cumule notamment un total de 143 types de domaines inédits certifiés par ces trois librairies pour un FDR de 10%, soit une augmentation de deux tiers en comparaison de la librairie Pfam seule. Il faut toutefois garder à l'esprit que le seuil de FDR choisi pour comparer les différentes méthodes est différent du taux d'erreur associé à ces cumuls. Il est vraisemblable que le taux d'erreur soit plus faible lorsque l'on regarde les domaines communs à plus d'une librairie et plus fort pour les domaines spécifiques à seule librairie, soit un taux d'erreur global des résultats cumulés plus important que le FDR fixé pour chaque méthode.

Parmi les nouveaux domaines inédits découverts avec un FDR inférieur à 10%, on peut citer par exemple :

– Grâce à la librairie corrigée par *K-means* : dans la protéine PFF0995c, annotée comme *Merozoite surface protein*, la certification du domaine Tme5_EGF_like (PF09064) par le domaine Pfam connu EGF (PF00008). Ce domaine, lié à l'interaction de la thrombomoduline avec la thrombine qui permet l'activation de la protéine-C (inhibitrice de la coagulation) (Fuentes-Prior *et al.*, 2000), permet de proposer une fonction plus précise pour ce facteur de croissance épidermique *Epidermal Growth Factor - EGF*. De plus, dans la protéine PFE0570w, annotée comme *RNA pseudouridylate synthase* putative, la certification du domaine inédit TT_ORF2 (PF02957) par le domaine potentiel de fonction inconnu DUF755 (PF05501) soulève la question de l'origine/l'évolution de cette protéine car ces deux domaines sont spécifiques de virus tel que *Torque teno* (Hino et Miyata, 2007). Notons toutefois que d'après les annotations Pfam de ces do-



(a) (b)

FIGURE 5.14 – **Diagramme de Venn des types de domaines totalement inédits dans les protéines de *P. falciparum* et certifiés par différentes librairies corrigées de HMM profils.** Les diagrammes (a) et (b) représentent les ensembles de domaines certifiés, respectivement pour un FDR de 10% et 20%, par la librairie de Pfam (en bleu), celle corrigée par les *K-means* (en jaune) et celle obtenue en réapprenant les modèles sur les séquences d'espèces proches (en vert).

maines, une occurrence de TT_ORF2 a été identifié dans un autre parasite, *Trypanosoma brucei*, tandis que DUF755 est connu chez l'amibe *Dictyostelliuim discoideum* et l'algue verte *Micromonas*.

– Grâce à la librairie reconstruite par espèces proches : dans la protéine PFB0280w, annotée comme EPSP-SK putative, le domaine EPSP_synthase (PF00275) est certifié en position N-terminal grâce au domaine potentiel SKI (PF01202). Cette protéine ne possède officiellement aucun domaine Pfam mais un unique domaine Interpro (SSF55205). Or, ces nouveaux domaines certifiés confirment l'implication de cette protéine dans le *shikimate pathway* (voie de biosynthèse d'acides aminés aromatiques à partir de chorismate absente chez les animaux, mais présents chez les apicomplexes) et représente donc une cible thérapeutique potentielle (McRobert *et al.*, 2005). De plus, la librairie Pfam originale certifie le domaine SKI grâce au domaine potentiel CM_2 (PF01817) en position C-terminal. L'observation de ce domaine également lié à la même voie métabolique, souligne la nécessité de l'intégration de tous les domaines certifiés afin de disposer d'une vision complète des architectures en domaines et donc de la fonction des protéines.

5.12 La question des domaines "non-certifiés"

5.12.1 Domaines non-certifiés et non-certifiables

La méthode de certification par co-occurrence nous a permis de comparer les différentes versions corrigées de la librairie Pfam que nous avons proposées. Cependant, on ne doit pas s'arrêter à l'étude de ces résultats pour évaluer le vrai potentiel de la correction des librairies de HMM. En effet, en se focalisant sur la certification, on laisse de côté tout un pan de résultats : les domaines *non-certifiés* et notamment les *non-certifiables*. Les domaines non-certifiés sont les domaines pour lesquels le contexte en domaines ne permet pas la certification. On peut distinguer deux cas de figure :

- soit les paires (validant, potentiel) formées n'ont pas été retenues comme significatives dans la liste des CDP. Le domaine potentiel est *certifiable mais non-certifié* ;
- soit il s'agit d'une protéine avec un seul domaine potentiel et aucun domaine validant (potentiellement une protéine monodomaine). Ce domaine potentiel est dit *non-certifiable*.

On schématise ces différentes situations comme dans la figure 5.15.

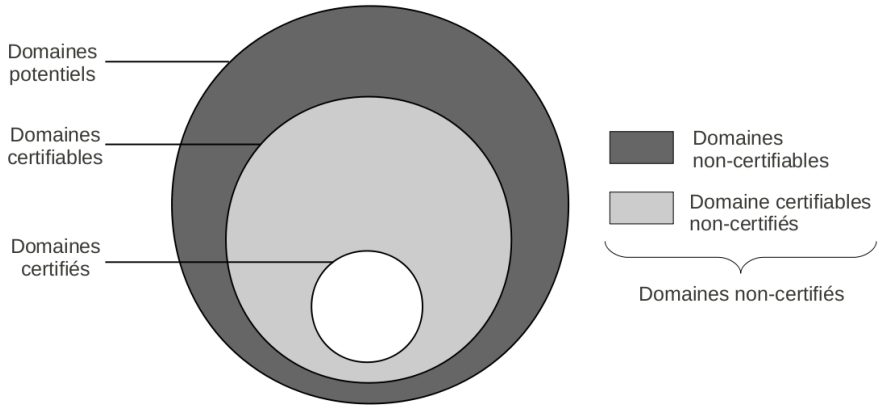


FIGURE 5.15 – **Représentation schématique des domaines potentiels accessibles ou non par la méthode de détection par co-occurrence.** L'ensemble des domaines potentiels (grand cercle) est composé de deux sous-ensembles exclusifs. On y distingue les domaines non-certifiables (en gris foncé) et les domaines certifiables (cercle moyen). De même, parmi les domaines certifiables, on différencie les domaines certifiables mais non-certifiés (en gris clair) et les domaines certifiés (petit cercle en blanc). Les domaines non-certifiés sont donc tous les domaines non-certifiables et les domaines certifiables mais non-certifiés (c'est à dire tout ce qui n'est pas en blanc).

On observe de nombreux domaines non-certifiés dans les différentes librairies de modèles corrigées. La méthode de certification par co-occurrence ne permettant pas de confirmer leur présence, ni de l'infirmer d'ailleurs. La figure 5.16 représente le nombre de domaines non-certifiés par les librairies de modèles de la figure 5.11, en fonction de l'E-valeur (utilisée pour déterminer les domaines potentiels). Comme évoqué précédemment (section 5.3), l'E-valeur

ne permet pas de comparer des librairies dont le modèle nul diffère. Il faut donc se garder de toute comparaison de performances à l'aide de cette figure. Les courbes relatives aux méthodes de *K-means* et *k*-plus proches voisins par exemple, semblent identifier moins de domaines que les autres approches à même E-valeur car elles utilisent un modèle nul plus proche de la composition de *P. falciparum*. Ce qu'il faut retenir de cette figure c'est qu'il existe plusieurs milliers de domaines non-certifiés par les différentes librairies, parmi lesquels se trouve vraisemblablement un certain nombre de domaines correctement prédits par les modèles.

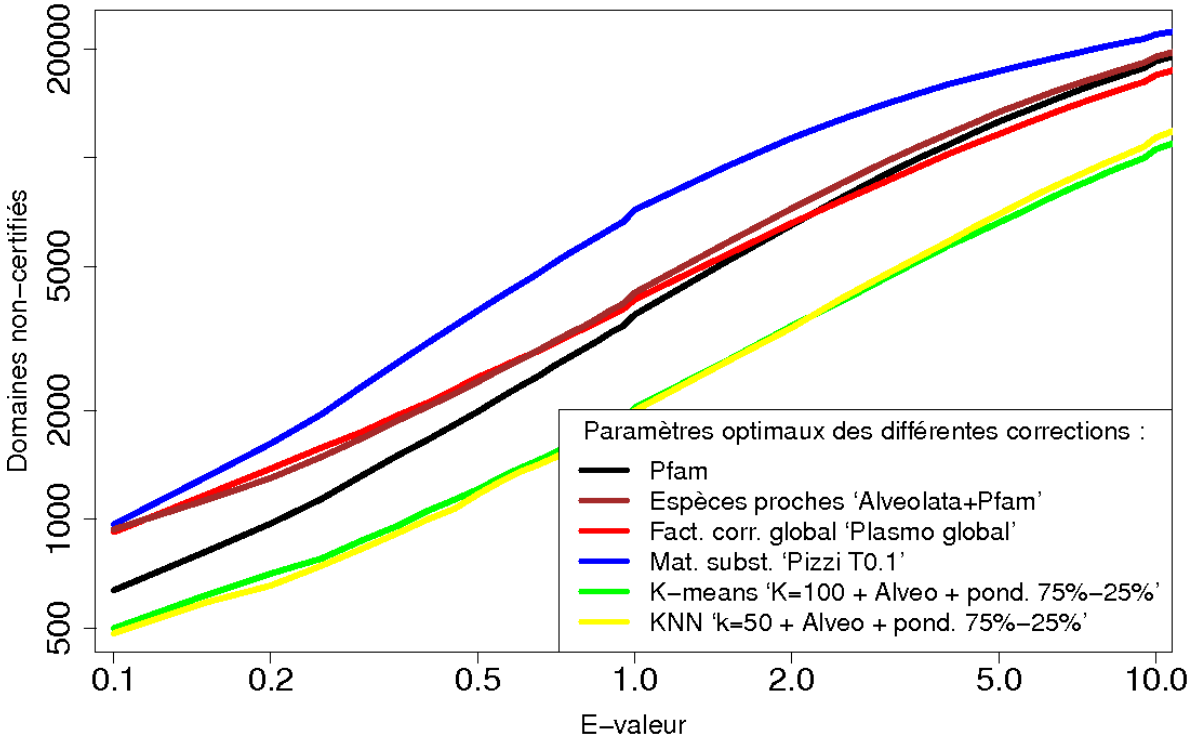


FIGURE 5.16 – **Nombre de domaines non-certifiés en fonction de l'E-valeur des domaines potentiels.** Sont représentées les librairies de HMM profils corrigés de la figure 5.11, en utilisant les domaines Pfam connus comme domaines validants pour la certification.

5.12.2 Non-certifiés/non-certifiables chez les domaines Pfam connu

Afin d'illustrer le nombre de domaines qui peuvent échapper à notre méthode de certification par co-occurrence, nous procédons à une expérimentation sur les domaines Pfam avérés. La question est de savoir combien de ces domaines serait certifiés par co-occurrence s'ils n'étaient pas considérés comme connus. Pour cela, nous prenons les 3 683 domaines Pfam

dont la présence est avérée chez *P. falciparum* et nous les considérons comme des domaines potentiels. On applique alors la procédure de certification en utilisant deux types de domaines validants : les domaines Interpro (non-Pfam) connus et les domaines potentiels eux-mêmes. On regarde alors le nombre de domaines certifiés et le nombre de domaines non-certifiables parmi ces domaines connus.

Type dom. validants	Dom. certifiés	Certifiés sur connus	Dom. certifiables	Certifiés sur certifiables	Dom. non-certifiables	Non-certifiables sur connus
Pfam potentiels	1 354	36%	1 376	98%	2 307	63%
Interp. connus	1 307	35%	1 716	76%	1 967	53%
Rés. cumulés	1 748	47%	1 982	88%	1 701	46%

TABLE 5.1 – **Tableau récapitulatif de la certification des domaines Pfam connus.** Les deux types de domaines validants utilisés sont les domaines Pfam connus (mais considérés ici comme potentiels) et les domaines Interpro connus. Les résultats cumulés de ces deux sources de domaines validants sont représentés sur la dernière ligne. Les colonnes représentent successivement : le nombre de domaines certifiés et leur ratio sur le nombre total de domaines connus, le nombre de domaines certifiables et la proportion de domaines certifiés parmi eux, et le nombre de domaines non-certifiables accompagnés du pourcentage de domaines connus que cela représente.

Les résultats obtenus sont détaillés dans le tableau 5.1. On constate que les domaines Pfam, considérés comme potentiels, permettent de certifier 1 354 d'entre eux. On retrouve ainsi 36% des domaines connus et, en cumulant avec les domaines Interpro connus comme validants, on atteint 47% de domaines retrouvés par co-occurrence. On observe également que la majorité des domaines certifiables sont effectivement certifiés par notre procédure. Par exemple, les domaines Pfam potentiels certifient 98% des domaines certifiables et les domaines Interpro connus 76%. Comme on pouvait s'y attendre, la principale faiblesse de l'approche de certification par co-occurrence se situe au niveau des domaines non-certifiables. Ces domaines non-certifiables représentent au total 46% des domaines Pfam connus chez *P. falciparum*, et 63% si l'on ne considère que les domaines Pfam comme validants. Ces résultats confirment logiquement l'existence de domaines très vraisemblables échappant à la méthode de certification par co-occurrence. Ces domaines semblent être majoritairement des domaines non-certifiables, issus vraisemblablement en grande partie de protéines monodomaines. Pour finir, notons que ces domaines, dont la présence est avérée, sont certifiés avec des taux d'erreurs respectifs de 4% et 6% en utilisant les domaines Pfam comme potentiels et les domaines Interpro connus.

5.12.3 Résultats des librairies corrigées à approfondir

Parmi les domaines potentiels non-certifiés par la librairie Pfam et les librairies obtenues par nos méthodes de correction (Figure 5.16), il est vraisemblable que certains domaines fassent partie des *véritables* compositions en domaines des protéines concernées. Cependant, seule une expertise manuelle approfondie des résultats permettrait éventuellement de distin-

guer les domaines réellement présents des faux-positifs. De plus, il est impossible d'extrapoler la proportion de faux positifs à partir des résultats obtenus sur les domaines Pfam connus (section 5.12.2). Une solution alternative consisterait à inspecter au cas par cas des domaines non-certifiés en s'appuyant sur d'autres indicateurs.

Pour illustrer cette idée, nous proposons ici d'examiner la proportion de domaines non-certifiés pour lesquels on dispose d'un domaine Interpro non-Pfam équivalent (même famille Interpro) recouvrant des positions identiques sur la séquence et dont la présence est avérée. Dans ce cas de figure, il est vraisemblable pour que le domaine potentiel, bien que non-certifié, soit réellement présent. Pour chaque librairie de modèles, on ajuste le seuil d'E-valeur de manière à récupérer environ 2 000 domaines potentiels non-certifiés (nombre de domaines non-certifiés similaire à celui obtenu sur les domaines Pfam connus). La figure 5.17 représente le nombre de domaines potentiels ayant un domaine Interpro similaire recouvrant, pour les trois sous-ensembles : domaines non-certifiables, domaines certifiables non-certifiés et domaines certifiés. À titre d'indication, ces données sont également représentées pour les résultats de certification obtenus sur les domaines Pfam connus.

On remarque pour commencer que la proportion de domaines Interpro recouvrant peut sembler étonnamment faible chez les domaines Pfam connus ($\sim 40\%$). Cela illustre en réalité la grande hétérogénéité des schémas de domaines des différentes bases qu'Interpro s'emploie à fédérer. Concernant les domaines potentiels des différentes librairies, on observe un taux important de domaines Interpro recouvrant les domaines certifiés ($\sim 30\%$). Ce taux peut suggérer une certaine redondance dans les nouvelles annotations apportées par la co-occurrence. Il est cependant important de noter que ces domaines recouvrants ne sont généralement détectés que par une minorité des neuf bases d'Interpro (et pas toujours par les mêmes). On comptabilise ensuite environ 11% des domaines non-certifiables qui sont recouverts par un domaine Interpro équivalent selon les librairies. On constate également que ce pourcentage est nettement supérieur à celui concernant les domaines certifiables non-certifiés ($<3\%$). Cela confirme ce que l'on s'attend à observer chez les domaines certifiables non-certifiés (qui n'ont donc pas passé le filtre de la co-occurrence) : une proportion de faux positifs nettement plus grande que pour les domaines non-certifiables.

Pour conclure, on retient que, pour chaque méthode de correction, la librairie de modèles générée dispose d'un potentiel d'annotation supérieur à celui suggéré initialement par la méthode de certification par co-occurrence. On trouve dans leurs résultats de nouveaux domaines vraisemblablement présents, en particulier en ce qui concerne les protéines monodomaines et leurs domaines non-certifiables. À l'inverse, il semble que les domaines certifiables non-certifiés soient souvent erronés.

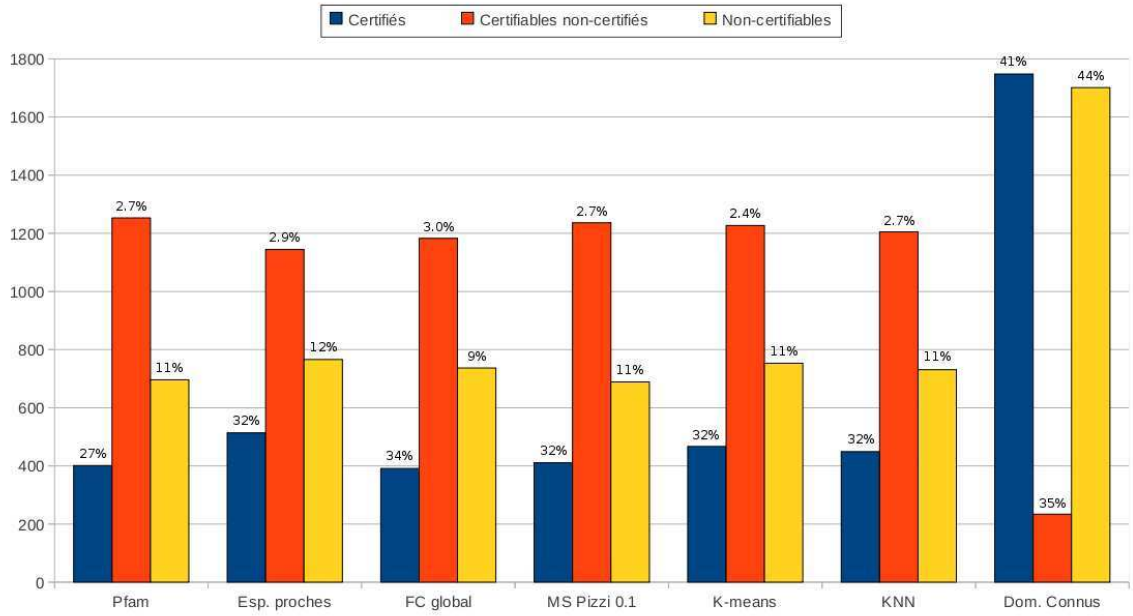


FIGURE 5.17 – **Diagramme du nombre de domaines des 3 catégories accompagné de la proportion de domaines ayant un domaine Interpro recouvrant, avéré et équivalent.** Les domaines potentiels ont été sélectionnés en ajustant le seuil d'E-valeur pour correspondre à environ 2 000 domaines non-certifiés. On distingue les trois sous-ensembles : domaines non-certifiables (en jaune), certifiables non-certifiés (en rouge) et certifiés (en bleu) pour les différentes librairies de modèles ainsi que pour les résultats de certification obtenus sur les domaines Pfam connus (expérience de la section 5.12.2). Sur chaque ensemble (bâton), on indique en chiffres le pourcentage de domaines de chaque catégorie pour lesquels il existe un domaine Interpro (non-Pfam) appartenant à la même famille et recouvrant les mêmes positions sur la séquence.

Cinquième partie

Conclusion

Dans le premier chapitre, nous avons introduit les fondements biologiques de notre problématique. Les protéines, codées par le génome (ADN), sont des molécules dont la fonction dépend de leur repliement tridimensionnel, lui même déterminé par la séquence d'acides aminés. Elles sont composées de plusieurs modules indépendants appelés domaines protéiques qui constituent une source d'information majeure pour l'annotation fonctionnelle. De nombreuses études se sont intéressées à la combinatoire des domaines protéiques. Elles ont notamment révélé que les familles de domaines ne sont généralement observées qu'avec un nombre très limité d'autres familles au sein des différentes protéines du Vivant.

Dans le deuxième chapitre, nous avons discuté des méthodes de modélisation des familles de domaines protéiques. Différents modèles ont été proposés par la communauté pour représenter les ensembles de séquences homologues. Parmi eux, le HMM profil dispose de nombreux atouts. La capacité de ce modèle à capturer l'information position-spécifique et à gérer l'incertitude liée au manque de données d'apprentissage (grâce à un cadre probabiliste) en fait un outil de choix pour l'identification de domaines protéiques. Le second chapitre détaille les HMM généraux et les HMM profils, ainsi que les logiciels classiques pour les manipuler (plus particulièrement `HMMER` utilisé dans Pfam).

Une fois posées ces bases bioinformatiques, nous avons consacré le troisième chapitre à l'organisme initiateur de ces recherches : *Plasmodium falciparum*, agent létal du paludisme humain. L'étude de ce parasite représente un enjeu majeur en terme de santé mondiale car il est responsable d'environ 2 millions de décès chaque année (principalement des enfants de moins de 5 ans habitant dans les zones d'Afrique sub-saharienne) et d'importantes conséquences socio-économiques (improductivité chronique du à l'affaiblissement des malades). Cet organisme partage avec d'autres pathogènes humains une divergence de ses séquences protéiques par rapport aux organismes modèles et exhibe de surcroît un fort biais dans la composition moyenne en acides aminés de ses protéines. Ces caractéristiques rendent particulièrement difficile son étude par les méthodes bioinformatiques classiques.

L'objectif de cette thèse était d'améliorer la sensibilité de la détection de domaines divergents à l'aide de HMM profils.

Nous avons donc proposé en premier lieu une méthode de détection de domaines protéiques basée sur la co-occurrence de paires de domaines. Lorsque la divergence des séquences empêchent l'identification du domaine grâce aux seuils de scores recommandés par les modèles, alors le contexte en domaines de la protéine est exploité. Cette méthode permet de certifier la présence de domaines jusque là incertains, en se basant sur la présence d'autres domaines (dits domaines validants) au sein de la même protéine. La méthode s'appuie sur l'identification de paires de domaines corrélés chez un grand nombre de protéines du vivant. Différentes ensembles de domaines validants ont été envisagées offrant des sources d'information différentes et donnant accès à une plus grande variété de nouveaux domaines. Une procédure de ré-échantillonage est utilisée pour estimer le taux d'erreur de nos prédictions. Nous avons vu qu'il est possible de contrôler ce taux d'erreur en jouant sur le seuil d'E-valeur déterminant les domaines potentiels. Notre approche a permis l'identification de nombreux domaines chez *Plasmodium falciparum* et ses espèces proches *Plasmodium vivax* et *Plasmodium yoelii*. Nous avons également montré qu'il est possible d'affiner la méthode, et nous

l'avons étendue à sept espèces supplémentaires d'eucaryotes pathogènes pour l'Homme. Les résultats obtenus sont intégrés dans une base de données en libre accès. On peut distinguer deux types de nouveaux domaines certifiés. D'un côté, une partie des certifications correspond à un nettoyage de l'annotation en domaines Pfam des protéines. Certains nouveaux domaines Pfam sont certifiés là où un autre domaine Interpro similaire est déjà connu. De même, on constate dans l'étude des espèces proches que les domaines certifiés sont parfois déjà connus dans les protéines homologues. D'un autre côté, de nombreuses certifications concernent la découverte de nouvelles familles de domaines dans la protéine concernée. Notre méthode permet donc de mettre l'accent sur des cibles thérapeutiques potentielles grâce à l'identification de fonctions inédites pour la protéine, voir même parfois pour l'organisme complet.

Le deuxième axe de recherche développé dans cette thèse porte sur la correction des librairies de HMM pour l'étude d'organismes biaisés. L'objectif ici est de modifier les paramètres des modèles de familles de domaines afin de les adapter à la divergence (au biais) de l'organisme étudié. Contrairement à la procédure de détection par co-occurrence, cette méthode permet également d'améliorer la détection de domaines dans les protéines monodomaines. La première correction envisagée peut s'assimiler à un réglage technique et concerne le modèle nul d'HMMER. Son impact sur les résultats s'est révélé difficilement maîtrisable mais on retiendra qu'un modèle nul doit être adapté en priorité à la composition globale des modèles de la librairie, plutôt qu'à la composition de l'organisme cible. Nous avons ensuite proposé de reconstruire les modèles grâce à des ensembles d'apprentissage composées de séquences d'espèces proches. Cette approche induit une amélioration naturelle des modèles mais ne permet pas la correction de la librairie complète de HMM, réapprendre un modèle nécessitant la disponibilité de suffisamment d'occurrences dans l'espèce cible ou ses proches relatifs. Enfin, différentes corrections des probabilités de génération associées aux états *Matchs* ont été envisagées, en s'appuyant sur des approches telles que des matrices de substitutions, les *K-means* ou les k-plus proches voisins. Ces méthodes exploitent la conservation des propriétés physico-chimiques des états et s'appliquent à l'ensemble des états des modèles, permettant de corriger la librairie complète de HMM Pfam. Elles nécessitent cependant l'ajustement de nombreux paramètres. Les différentes librairies corrigées ont été comparées *via* la méthode de certification par co-occurrence du chapitre précédent. La correction des modèles entraîne une ré-évaluation des scores des domaines ce qui influe sur la construction des ensembles de domaines potentiels et modifie donc les résultats de certification par co-occurrence. Notre analyse démontre que ces librairies alternatives permettent de découvrir de nouveaux domaines échappant aux modèles originaux de Pfam. De plus, chacune des approches certifie un certain nombre de domaines qui lui sont spécifiques. On notera que les approches dont les résultats sont les plus satisfaisants sont la reconstruction grâce aux espèces proches et la correction par *K-means* (le temps de calcul ayant été un frein à l'optimisation des paramètres de l'approche par k-plus-proches voisins). On dispose donc, grâce à ces différentes approches, d'une source d'information complémentaire à prendre en compte lors de l'annotation d'organismes divergents. Selon nous, la principale difficulté de la correction des modèles reste cependant liée à l'hétérogénéité de l'évolution des sites de profil physico-chimique similaire.

Le tableau 5.2 récapitule les résultats obtenus en appliquant notre méthode de certifica-

Librairies	Nvx dom. certifiés	Nvlles fam. InterPro	Inédites chez *Pf*	Annot. GO (dom. seul)	Annot. GO (combin. dom.)	Prot. sans GO
Pfam origin.	585 (456)	479 (363)	159 (187)	273 (122)	114 (66)	39 (169)
Esp. proches	631 (484)	519 (392)	137 (166)	281 (135)	122 (71)	42 (183)
Fact. Corr.	606 (464)	497 (368)	177 (196)	264 (128)	126 (73)	35 (176)
Mat. Subst.	601 (465)	492 (376)	168 (196)	305 (131)	123 (73)	40 (185)
K-means	640 (490)	530 (394)	178 (209)	271 (127)	118 (75)	41 (182)
KNN	619 (474)	509 (384)	165 (199)	277 (126)	140 (79)	45 (182)
Cumul	1248 (740)	1066 (647)	317 (366)	663 (291)	234 (126)	67 (358)

TABLE 5.2 – Tableau récapitulatif des résultats de certifications par co-occurrence obtenus chez *P. falciparum* par les différentes librairies de HMM (en ligne) : "Pfam origin." la librairie Pfam officielle (*cf.* Chapitre 4) et "Esp. proches", "Fact. Corr.", "Mat. Subst.", "*K-means*" et "*KNN*" les librairies corrigées respectivement grâce à des espèces proches (*cf.* Section 5.5), à des facteurs de corrections (*cf.* Section 5.7), à des matrices de substitutions (*cf.* Section 5.8), à un *clustering* par *K-means* (*cf.* Section 5.9) et à un apprentissage de type k-plus proches voisins (*cf.* Section 5.10). Pour chacune de ses librairies, "Nvx dom. certifiés" indique le nombre de domaines certifiés par co-occurrence avec un FDR<20%, en utilisant les trois types de domaines validants. Parmi ces domaines, "Nvlles fam. InterPro" correspond au nombre de domaines qui appartiennent à une famille InterPro inédite pour la protéine et "Inédites chez *Pf*" aux familles de domaines qui n'était jusque là identifiées dans aucune protéine de *P. falciparum*. On peut également déduire des annotations GO de ces domaines et des combinaisons de domaines, le nombre d'annotations GO inédites que l'on transfère aux protéines (resp. colonnes "Annot. GO (dom. seul)" et "Annot. GO (combin. dom.)"). On précise entre parenthèses le nombre de protéines concernées par ces précédictions. De plus, le nombre de protéines qui ne possédaient jusque là aucune annotation dans la *Gene Ontology* est explicité par la colonne "Prot. sans GO" et on indique entre parenthèse le nombre total de protéines que l'on annote dans la GO. Enfin, la dernière ligne du tableau ("Cumul") montre l'information apportée en cumulant les résultats des différentes librairies de modèles (ces cumuls ne peuvent bien entendu pas prétendre à un FDR inférieur à 20% comme les librairies prises individuellement).

tion par co-occurrence à *P. falciparum* (avec un FDR<20%) en utilisant la librairie Pfam originale (correspondant Chapitre 4) et aux librairies corrigées par différentes approches que nous avons vues au chapitre 5 (avec un paramétrage optimal). Ce tableau présente également un cumul du nombre de domaines certifiés et des annotations GO déduites grâce aux différentes librairies. Au total, on dénombre 1 248 nouveaux domaines dans 740 protéines, dont 1 066 correspondent à des familles de domaines InterPro jusque là inconnues chez les 647 protéines concernées. De plus, parmi ces domaines certifiés, on identifie 317 familles de domaines qui n'avaient jamais été identifiées dans une protéine de *P. falciparum* auparavant. Ces familles inédites chez *P. falciparum*, sont désormais identifiées dans 366 protéines distinctes

du parasite. Grâce à l'annotation des domaines Pfam par la *Gene Ontology*, nous avons pu déduire des domaines certifiés 663 nouveaux termes GO associés à 291 protéines, auxquels nous ajoutons 234 annotations GO déduites des combinaisons de domaines — formées par ces domaines certifiés — dans 126 protéines. Au total, 358 protéines reçoivent une annotation — par au moins un terme GO — grâce aux domaines certifiés dont 67 protéines (19%) qui ne possédaient jusqu'ici aucune annotation GO. On notera cependant que ces résultats ne reflètent qu'une partie des annotations découlant de la correction des librairies de modèles. Effectivement, parmi les domaines non-certifiés un certain nombre sont vraisemblablement justes, notamment dans les protéines monodomaines.

Les perspectives de ces travaux sont de plusieurs ordres. Concernant la méthode de détection par co-occurrence, plusieurs pistes pour optimiser son potentiel ont été évoquées. Il est tout d'abord possible (et nos premières expériences montrent que cela est judicieux) d'adapter l'apprentissage des CDP en se recentrant sur des espèces plus spécifiques de l'organisme cible (ici eucaryotes ou alvéolés), pour une meilleure appréciation des combinaisons attendues entre les domaines. La construction des ensembles de domaines potentiels peut également être améliorée. Il peut sembler arbitraire de sélectionner les domaines potentiels non-chevauchant sur un critère comme l'E-valeur quand celle-ci devient élevée car la différence de significativité n'est plus vraiment assurée. La prise en compte des protéines homologues chez les espèces proches pourrait apporter une information complémentaire robuste pour la construction des ensembles de domaines potentiels. De plus, il est vraisemblable que la détection par co-occurrence puisse bénéficier des améliorations récentes apportées par la version 3.0 du programme HMMER en matière de précision des scores et des E-valeurs grâce au passage à des scores *forward*, ainsi que par la réduction des temps de calcul. Enfin, d'autres informations pourraient être prises en compte telles que l'adjacence ou l'ordre/enchainement des domaines, ainsi que la pluralité des sources de certification (domaines validants) afin d'améliorer la confiance (FDR) associée aux nouveaux domaines. Notre approche pourrait ensuite être appliquée plus largement, notamment à l'ensemble des organismes modèles pour identifier les domaines les plus divergents. Cela permettrait la correction d'un certain nombre d'annotations incorrectes ou incomplètes, même pour les organismes considérés comme bien annotés. Au-delà des annotations fonctionnelles, de tels résultats seraient utiles à l'étude de la combinatoire des domaines, et aux mécanismes d'évolution qui y sont liés.

Concernant les méthodes de corrections de modèles, la première perspective concerne l'intégration des résultats obtenus par les différentes méthodes dans un portail unique et la publication de ces résultats. Ensuite, ces résultats nous donnent plusieurs informations utiles pour des études futures. Tout d'abord, la performance de la reconstruction par les espèces proches nous rappelle l'intérêt d'une approche itérative, dans l'esprit de PSI-BLAST, qui n'existe actuellement pas dans les bases de données de domaines. Il serait pourtant envisageable de proposer un *pipeline* qui effectuerait dans un premier temps une recherche de domaines dans une sélection des espèces phylogénétiquement proches. Les résultats obtenus servirait alors de complément/remplacement d'une partie des séquences d'apprentissage des modèles. La recherche de domaines dans l'organisme cible, réalisée dans un deuxième temps, bénéficierait alors de modèles plus adéquats et fournirait donc des résultats plus pertinents.

Ce genre d'approche n'est pas réservé à l'étude d'organismes divergents ou biaisés mais devrait bénéficier à n'importe quel génome.

Nous avons vu que les corrections par *K-means* et k-plus proches voisins pourraient également être améliorées par un choix judicieux de leur paramètres (choix vraisemblablement liés au génome de l'espèce étudiée). Une perspective pour ces méthodes est l'extension de leur application à d'autres organismes biaisés, notamment aux espèces GC-riches, afin de confirmer leur potentiel. D'un point de vue théorique, il serait également intéressant d'approfondir le problème posé par l'hétérogénéité de l'évolution des sites. Les différentes positions des séquences évoluent selon des contraintes qui leurs sont propres. Cela se traduit par des vitesses d'évolution différentes des sites de profils physico-chimiques identiques. Une solution consisterait alors à prendre en compte l'information phylogénétique afin de proposer des corrections plus adaptées, en couplant par exemple les classes de profils physico-chimiques avec des classes de vitesses d'évolution. On pourrait aussi envisager des matrices d'évolution adaptées et non-homogènes, contrairement aux matrices utilisées ici (WAG, LG, *etc.*), qui permettraient mieux modéliser la dérive des organismes présentant un fort biais compositionnel comme *P. falciparum*.

Bibliographie

Abarbanel, R., Wieneke, P., Mansfield, E., Jaffe, D. et Brutlag, D. (1984). Rapid searches for complex patterns in biological molecules. *Nucleic Acids Research*, 12(1 Pt 1):263–280. Télécharger.

Abe, N. et Warmuth, M. (1992). On the computational complexity of approximating distributions by probabilistic automata. *Machine Learning*, 9(2-3):205–260. Télécharger.

Abu-Raddad, L., Patnaik, P. et Kublin, J. (2006). Dual infection with hiv and malaria fuels the spread of both diseases in sub-saharan africa. *Science*, 314(5805):1603–1606. Télécharger.

Adams, M., Celniker, S., Holt, R., Evans, C., Gocayne, J., Amanatides, P., Scherer, S., Li, P., Hoskins, R. et *et al.*, R. G. (2000). The genome sequence of *Drosophila melanogaster*. *Science*, 287(5461):2185–2195. Télécharger.

Alam, I., Hubbard, S., Oliver, S. et Rattray, M. (2007). A kingdom-specific protein domain HMM library for improved annotation of fungal genomes. *BMC genomics*, 8:97. Télécharger.

Altschul, S., Gish, W., Miller, W., Myers, E. et Lipman, D. (1990). Basic local alignment search tool. *Journal of Molecular Biology*, 215(3):403–410. Télécharger.

Altschul, S., Madden, T., Schaffer, A., Zhang, J., Zhang, Z., Miller, W. et Lipman, D. (1997). Gapped BLAST and PSI-BLAST : a new generation of protein database search programs. *Nucleic acids research*, 25(17):3389–3402. Télécharger.

Anfinsen, C. (1973). Principles that govern the folding of protein chains. *Science*, 181(4096): 223–230. Télécharger.

Apic, G., Gough, J. et Teichmann, S. (2001). Domain combinations in archaeal, eubacterial and eukaryotic proteomes. *Journal of Molecular Biology*, 310(2):311–325. Télécharger.

Apic, G., Huber, W. et Teichmann, S. (2003). Multi-domain protein families and domain pairs : comparison with known structures and a random model of domain recombination. *Journal of Structural and Functional Genomics*, 4(2-3):67–78. Télécharger.

Apweiler, R., Attwood, T., Bairoch, A., Bateman, A., Birney, E., Biswas, M., Bucher, P., Cerutti, L., Corpet, F., Croning, M., Durbin, R., Falquet, L., Fleischmann, W., Gouzy, J., Hermjakob, H., Hulo, N., Jonassen, I., Kahn, D., Kanapin, A., Karavidopoulou, Y., Lopez, R., Marx, B., Mulder, N., Oinn, T., Pagni, M., Servant, F., Sigrist, C. et Zdobnov, E. (2001). The InterPro database, an integrated

documentation resource for protein families, domains and functional sites. *Nucleic Acids Research*, 29(1):37–40. Télécharger.

Apweiler, R., Bairoch, A., Wu, C., Barker, W., Boeckmann, B., Ferro, S., Gasteiger, E., Huang, H., Lopez, R., Magrane, M., Martin, M., Natale, D., O'Donovan, C., Redaschi, N. et Yeh, L. (2004). Uniprot : the universal protein knowledgebase. *Nucleic Acids Research*, 32(Database issue):D115–D119. Télécharger.

Argos, P. et Palau, J. (1982). Amino acid distribution in protein secondary structures. *Int J Pept Protein Res*, 19(4):380–393. Télécharger.

Ashburner, M., Ball, C., Blake, J., Botstein, D., Butler, H., Cherry, J., Davis, A., Dolinski, K., Dwight, S., Eppig, J., Harris, M., Hill, D., Issel-Tarver, L., Kasarskis, A., Lewis, S., Matese, J., Richardson, J., Ringwald, M., Rubin, G. et Sherlock, G. (2000). Gene ontology : tool for the unification of biology. the gene ontology consortium. *Nature Genetics*, 25(1):25–29. Télécharger.

Aslett, M., Aurrecoechea, C., Berriman, M., Brestelli, J., Brunk, B., Carrington, M., Depledge, D., Fischer, S., Gajria, B., Gao, X., Gardner, M., Gingle, A., Grant, G., Harb, O., Heiges, M., Hertz-Fowler, C., Houston, R., Innamorato, F., Iodice, J., Kissinger, J., Kraemer, E., Li, W., Logan, F., Miller, J., Mitra, S., Myler, P., Nayak, V., Pennington, C., Phan, I., Pinney, D., Ramasamy, G., Rogers, M., Roos, D., Ross, C., Sivam, D., Smith, D., Srinivasamoorthy, G., Stoeckert, C., Subramanian, S., Thibodeau, R., Tivey, A., Treatman, C., Velarde, G. et Wang, H. (2010). TriTrypDB : a functional genomic resource for the Trypanosomatidae. *Nucleic Acids Research*, 38(Database issue):D457. Télécharger.

Attwood, T., Beck, M., Bleasby, A. et Parry-Smith, D. (1994). PRINTS–a database of protein motif fingerprints. *Nucleic acids research*, 22(17):3590. Télécharger.

Attwood, T., Bradley, P., Flower, D., Gaulton, A., Maudling, N., Mitchell, A., Moulton, G., Nordle, A., Paine, K., Taylor, P., Uddin, A. et Zygouri, C. (2003). Prints and its automatic supplement, preprints. *Nucleic Acids Research*, 31(1):400–402. Télécharger.

Aurrecoechea, C., Brestelli, J., Brunk, B., Carlton, J., Dommer, J., Fischer, S., Gajria, B., Gao, X., Gingle, A., Grant, G., Harb, O., Heiges, M., Innamorato, F., Iodice, J., Kissinger, J., Kraemer, E., Li, W., Miller, J., Morrison, H., Nayak, V., Pennington, C., Pinney, D., Roos, D., Ross, C., Stoeckert, C. J., Sullivan, S., Treatman, C. et Wang, H. (2009a). Giardiadb and trichdb : integrated genomic resources for the eukaryotic protist pathogens *Giardia lamblia* and *Trichomonas vaginalis*. *Nucleic Acids Research*, 37(Database issue):D526–D530. Télécharger.

Aurrecoechea, C., Brestelli, J., Brunk, B., Fischer, S., Gajria, B., Gao, X., Gingle, A., Grant, G., Harb, O., Heiges, M., Innamorato, F., Iodice, J., Kissinger, J., Kraemer, E., Li, W., Miller, J., Nayak, V., Pennington, C., Pinney, D., Roos, D., Ross, C., Srinivasamoorthy, G., Stoeckert, C., Thibodeau, R., Treatman, C. et Wang, H. (2009b). EuPathDB : a portal to eukaryotic pathogen databases. *Nucleic Acids Research*, 38(Database issue):D415–D419. Télécharger.

Aurrecoechea, C., Heiges, M., Wang, H., Wang, Z., Fischer, S., Rhodes, P., Miller, J., Kraemer, E., Stoeckert, C. J., Roos, D. et *et al.* (2007). Apidb : integrated resources for the apicomplexan bioinformatics resource center. *Nucleic Acids Research*, 35(Database issue):D427–D430. Télécharger.

Bahl, A., Brunk, B., Crabtree, J., Fraunholz, M., Gajria, B., Grant, G., Ginsburg, H., Gupta, D., Kissinger, J., Labo, P., Li, L., Mailman, M., Milgram, A., Pearson, D., Roos, D., Schug, J., Stoeckert, C. J. et Whetzel, P. (2003). Plasmodb : the *Plasmodium* genome resource. a database integrating experimental and computational data. *Nucleic Acids Research*, 31(1):212–215. Télécharger.

Bailey, T., Boden, M., Buske, F., Frith, M., Grant, C., Clementi, L., Ren, J., Li, W. et Noble, W. (2009). MEME SUITE : tools for motif discovery and searching. *Nucleic Acids Research*, 37:W202–W208. Télécharger.

Bailey, T. et Elkan, C. (1995). Unsupervised learning of multiple motifs in biopolymers using expectation maximization. *Machine Learning*, 21(1-2):51–80. Télécharger.

Bairoch, A. (1991). Prosite : a dictionary of sites and patterns in proteins. *Nucleic Acids Research*, 19(Suppl):2241–2245. Télécharger.

Bairoch, A. et Apweiler, R. (1996). The swiss-prot protein sequence data bank and its new supplement trembl. *Nucleic Acids Research*, 24(1):21–25. Télécharger.

Bairoch, A., Bucher, P. et Hofmann, K. (1996). The PROSITE database, its status in 1995. *Nucleic Acids Research*, 24(1):189–196. Télécharger.

Bairoch, A. et Claverie, J. (1988). Sequence patterns in protein kinases. *Nature*, 331(6151): 22. Télécharger.

Baker, J. (1975). The dragon system-an overview. *In IEEE Transactions on Acoustics, Speech, Signal Processing*, volume 23 issue 1, pages 24–29. Télécharger.

Barret, C., Hughey, R. et Karplus, K. (1997). Scoring hidden markov models. *CABIOS*, 13(2):191–199. Télécharger.

Bastien, O., Lespinats, S., Roy, S., Métayer, K., Fertil, B., Codani, J. et Maréchal, E. (2004). Analysis of the compositional biases in *Plasmodium falciparum* genome and proteome using *Arabidopsis thaliana* as a reference. *Gene*, 336(2):163–173. Télécharger.

Bastien, O., Roy, S. et Marechal, E. (2005). Construction of non-symmetric substitution matrices derived from proteomes with biased amino acid distributions. *C R Biol*, 328(5): 445–453.

Bateman, A., Birney, E., Cerruti, L., Durbin, R., Etwiller, L., Eddy, S., Griffiths-Jones, S., Howe, K., Marshall, M. et Sonnhammer, E. (2002). The pfam protein families database. *Nucleic Acids Research*, 30(1):276–280. Télécharger.

Bateman, A., Birney, E., Durbin, R., Eddy, S., Finn, R. et Sonnhammer, E. (1999). Pfam 3.1 : 1313 multiple alignments and profile hmms match the majority of proteins. *Nucleic Acids Research*, 27(1):260–262. Télécharger.

Bateman, A., Birney, E., Durbin, R., Eddy, S., Howe, K. et Sonnhammer, E. (2000). The pfam protein families database. *Nucleic Acids Research*, 28(1):263–266. Télécharger.

Bateman, A., Coin, L., Durbin, R., Finn, R., Hollich, V., Griffiths-Jones, S., Khanna, A., Marshall, M., Moxon, S., Sonnhammer, E., Studholme, D., Yeats, C. et Eddy, S. (2004). The pfam protein families database. *Nucleic Acid Research*, 32(Database issue):D138–D141. Télécharger.

Baum, L., Petrie, T., Soules, G. et Weiss, N. (1970). A maximization technique occurring in the statistical analysis of probabilistic functions of markov chains. *The Annals of Mathematical Statistics*, 41(1):164–171. Télécharger.

Beaussart, F., 3rd Weiner, J. et Bornberg-Bauer, E. (2007). Automated improvement of domain annotations using context analysis of domain arrangements (aidan). *Bioinformatics*, 23(14):1834–1836. Télécharger.

Benjamini, Y. et Hochberg, Y. (1995). Controlling the false discovery rate : a practical and powerful approach to multiple testing. *Journal of the Royal Statistical Society*, 85:289–300. Télécharger.

Benson, D., Karsch-Mizrachi, I., Lipman, D., Ostell, J. et Sayers, E. (2009). Genbank. *Nucleic Acids Research*, 37(Database issue):D26–D31. Télécharger.

Berman, H., Westbrook, J., Feng, Z., Gilliland, G., Bhat, T., Weissig, H., Shindyalov, I. et Bourne, P. (2000). The protein data bank. *Nucleic Acids Research*, 28(1):235–242. Télécharger.

Bernal, J. (1958). General introduction structure arrangements of macromolecules. *Discussions of the Faraday Society*, 25:7–18. Télécharger.

Bertani, S., Hou
"el, E., Bourdy, G., Stien, D., Jullian, V., Landau, I. et Deharo, E. (2007). *Quassia amara* l. (simaroubaceae) leaf tea : Effect of the growing stage and desiccation status on the antimalarial activity of a traditional preparation. *Journal of Ethnopharmacology*, 111(1):40–42. Télécharger.

Betel, D., Isserlin, R. et Hogue, C. (2004). Analysis of domain correlations in yeast protein complexe. *Bioinformatics*, 20((Suppl)1):155–162. Télécharger.

Björklund, Å., Ekman, D., Light, S., Frey-Skött, J. et Elofsson, A. (2005). Domain rearrangements in protein evolution. *Journal of Molecular Biology*, 353(4):911–923. Télécharger.

Blanquart, S. et Gascuel, O. (2010). Mitochondrial genes support a common origin of rodent malaria parasites and plasmodium falciparum's relatives infecting great apes. *Pre-print*.

Blanquart, S. et Lartillot, N. (2006). A bayesian compound stochastic process for modeling nonstationary and nonhomogeneous sequence evolution. *Molecular Biology and Evolution*, 23(11):2058–2071. Télécharger.

Blanquart, S. et Lartillot, N. (2008). A site- and time-heterogeneous model of amino acid replacement. *Molecular Biology and Evolution*, 25(5):842–858. Télécharger.

Blekas, K., Fotiadis, D. et Likas, A. (2005). Motif-based protein sequence classification using neural networks. *Journal of Computational Biology*, 12(1):64–82. Télécharger.

Blout, E., de Lozé, C., Bloom, S. et Fasman, G. (1960). The dependence of the conformations of synthetic polypeptides on amino acid composition. *Journal of the American Chemical Society*, 82(14):3787–3789. Télécharger.

Bornberg-Bauer, E., Beaussart, F., Kummerfeld, S., Teichmann, S. et Weiner, J. (2005). The evolution of domain arrangements in proteins and interaction networks. *Cellular and Molecular Life Sciences (CMLS)*, 62(4):435–445. Télécharger.

Bougdour, A., Maubon, D., Baldacci, P., Ortet, P., Bastien, O., Bouillon, A., Barale, J., Pelloux, H., Ménard, R. et Hakimi, M. (2009). Drug inhibition of hdac3 and epigenetic control of differentiation in apicomplexa parasites. *The Journal of experimental medicine*, 206(4):953–966. Télécharger.

Bowie, J., Luthy et Eisenberg, D. (1991). A method to identify protein sequences that fold into a known three-dimensional structure. *Science*, 253(5016):164–170. Télécharger.

Bozdech, Z., Llinás, M., Pulliam, B., Wong, E., Zhu, J. et DeRisi, J. (2003). The transcriptome of the intraerythrocytic developmental cycle of *Plasmodium falciparum*. *PLoS Biology*, 1(1):E5. Télécharger.

Brazma, A., Jonassen, I., Eidhammer, I. et Gilbert, D. (1998). Approaches to the automatic discovery of patterns in biosequences. *Journal of Computational Biology*, 5(2):279–305. Télécharger.

Brown, D., Krishnamurthy, N., Dale, J., Christopher, W. et Sjölander, K. (2005). Subfamily hmms in functional genomics. *In Pacific Symposium on Biocomputing*, volume 10, pages 322–333. Télécharger.

Brown, M., Hughey, R., Krogh, A., Mian, I., Sjolander, K. et Haussler, D. (1993). Using Dirichlet Mixture Priors to Derive Hidden Markov Models for Protein Families. *In Proceedings of the International Conference on Intelligent Systems for Molecular Biology*, volume 1, pages 47–55. Télécharger.

Bru, C., Courcelle, E., Carrère, S., Beausse, Y., Dalmar, S. et Kahn, D. (2005). The prodom database of protein domain families : more emphasis on 3d. *Nucleic Acids Research*, 33(Database issue):D212–D215. Télécharger.

Bryant, S. et Lawrence, C. (1993). An empirical energy function for threading protein sequence through the folding motif. *Proteins : Structure, function, and genetics*, 16(1):92–112. Télécharger.

Bréhélin, L. (2001). *Modèles de Markov cachés et apprentissage par fusion d'états : algorithmes, applications, utilisations pour le test de circuits intégrés*. Thèse de doctorat, Université Montpellier 2.

Bréhélin, L., Florent, I., Gascuel, O. et Maréchal, E. (2010). Assessing functional annotation transfers with inter-species conserved coexpression : application to *Plasmodium falciparum*. *BMC genomics*, 11(1):35. Télécharger.

Buchan, D., Rison, S., Bray, J., Lee, D., Pearl, F., Thornton, J. et Orengo, C. (2003). Gene3d : Structural assignments for the biologist and bioinformaticist alike. *Nuclear Acids Research*, 31(1):469–473. Télécharger.

Bucher, P. et Bairoch, A. (1994). A generalized profile syntax for biomolecular sequence motifs and its function in automatic sequence interpretation. *Proceedings of the International Conference on Intelligent Systems for Molecular Biology*, 2:53–61. Télécharger.

Callebaut, I., Prat, K., Meurice, E., Mornon, J. et Tomavo, S. (2005). Prediction of the general transcription factors associated with rna polymerase ii in *Plasmodium falciparum* : conserved features and differences relative to other eucaryotes. *BMC Genomics*, 6:100. Télécharger.

Carlton, J., Adams, J., Silva, J., Bidwell, S., Lorenzi, H., Caler, E., Crabtree, J., Angiuoli, S., Merino, E. et *et al.*, P. A. (2008a). Comparative genomics of the neglected human malaria parasite *Plasmodium vivax*. *Nature*, 455(7214):757–763. Télécharger.

Carlton, J., Angiuoli, S., Suh, B., Kooij, T., Pertea, M., Silva, J., Ermolaeva, M., Allen, J., Selengut, J., Koo, H., Peterson, J., Pop, M., Kosack, D., Shumway, M., Bidwell, S., Shallom, S., van Aken, S., Riedmuller, S., Feldblyum, T., Cho, J., Quackenbush, J., Sedegah, M., Shoaibi, A., Cummings, L., Florens, L., Yates, J., Raine, J., Sinden, R., Harris, M., Cunningham, D., Preiser, P., Bergman, L., Vaidya, A., van Lin, L., Janse, C., Waters, A., Smith, H., White, O., Salzberg, S., Venter, J., Fraser, C., Hoffman, S., Gardner, M. et Carucci, D. (2002). Genome sequence and comparative analysis of the model rodent malaria parasite *Plasmodium yoelii yoelii*. *Nature*, 419(6906):512–519. Télécharger.

Carlton, J., Escalante, A., Neafsey, D. et Volkman, S. (2008b). Comparative evolutionary genomics of human malaria parasites. *Trends in Parasitology*, 24(12):545–550. Télécharger.

Carlton, J., Silva, J. et Hall, N. (2005). The genome of model malaria parasites, and comparative genomics. *Current issues in molecular biology*, 7(1):23–37. Télécharger.

Caspi, R., Foerster, H., Fulcher, C., Kaipa, P., Krummenacker, M., Latendresse, M., Paley, S., Rhee, S., Shearer, A., Tissier, C., Walk, T., Zhang, P. et Karp, P. (2008). The metacyc database of metabolic pathways and enzymes and the biocyc collection of pathway/genome databases. *Nucleic Acids Research*, 36(Database issue): D623–D631. Télécharger.

Cassacuberta, F. (1990). Some relations among stochastic finite state networks used in automatic speech recognition. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 12(7):691–695. Télécharger.

Chemaly, S., Chen, C. et van Zyl, R. (2007). Naturally occurring cobalamins have antimalarial activity. *Journal of Inorganic Biochemistry*, 101(5):764–773. Télécharger.

Chen, X. et Liu, M. (2005). Prediction of protein-protein interactions using random decision forest framework. *Bioinformatics*, 21(24):4394–4400. Télécharger.

Cherry, J., Adler, C., Ball, C., Chervitz, S., Dwight, S., Hester, E., Jia, Y., Juvik, G., Roe, T., Schroeder, M., Weng, S. et Botstein, D. (1998). Sgd : Saccharomyces genome database. *Nucleic Acid Research*, 26(1):73–79. Télécharger.

Chou, K. et Cai, Y. (2002). Using functional domain composition and support vector machines for prediction of protein subcellular location. *Journal of Biochemical Chemistery*, 277(48):45765–45769. Télécharger.

Churchill, G. (1989). Stochastic models for heterogeneous dna sequences. *Bulletin of Mathematical Biology*, 51(1):79–94. Télécharger.

Cohen-Gihon, I., Nussinov, R. et Sharan, R. (2007). Comprehensive analysis of co-occurring domain sets in yeast proteins. *BMC Genomics*, 8:161. Télécharger.

Coin, L., Bateman, A. et Durbin, R. (2003). Enhanced protein domain discovery by using language modeling techniques from speech recognition. *Proceedings of the National Academy of Sciences of the U.S.A.*, 100(8):4516–4520. Télécharger.

Cole, S., Brosch, R., Parkhill, J., Garnier, T., Churcher, C., Harris, D., Gordon, S., Eiglmeier, K., Gas, S., Barry, C., Badcock, F. T. K., Basham, D., Brown, D., Chillingworth, T., Connor, R., Davies, R., Devlin, K., Feltwell, T., Gentles, S., Hamlin, N., Holroyd, S., Hornsby, T., Jagels, K., McLean, A. K. J., Moule, S., Murphy, L., Oliver, K., Osborne, J., Quail, M., Rajandream, M.-A., Rogers, J., Rutter, S., Seeger, K., Skelton, J., Squares, R., Squares, S., Sulston, J., Taylor, K., Whitehead, S. et Barrell, B. (1998). Deciphering the biology of *Mycobacterium tuberculosis* from the complete genome sequence. *Nature*, 393(6685):537–544. Télécharger.

Corpet, F., Gouzy, J. et Kahn, D. (1998). The prodom database of protein domain families. *Nucleic Acids Research*, 26(1):323–326. Télécharger.

Coulson, R., Hall, N. et Ouzonis, A. (2004). Comparative genomics of transcriptional control in the human parasite *Plasmodium falciparum*. *Genome Research*, 14(8):1548–1554. Télécharger.

Cox, F. (2002). History of human parasitology. *Clinical microbiology reviews*, 15(4):595–612.

Crick, F., Barnett, L., Brenner, S. et Watts-Tobin, R. (1961). General nature of the genetic code for proteins. *Nature*, 192:1227–1232. Télécharger.

Cuff, A., Sillitoe, I., Lewis, T., Redfern, O., Garratt, R., Thornton, J. et Orengo, C. (2009). The CATH classification revisited–architectures reviewed and new ways to characterize structural divergence in superfamilies. *Nucleic Acids Research*, 37(Database issue):D310–D314. Télécharger.

Cuff, J. et Barton, G. (2000). Application of multiple sequence alignment profiles to improve protein secondary structure prediction. *Proteins : Structure, Function, and Genetics*, 40(3):502–511. Télécharger.

Dayhoff, M., Schwartz, R. et Orcutt, B. (1978). *A model of evolutionary change in proteins*, volume 5. National Biomedical Research Foundation, Washington, D.C.

Dempster, A., Laird, N. et Rubin, D. (1977). Maximum likelihood from incomplete data via the em algorithm. *Journal Royal Statistical Society. Series B (Methodological)*, 39(1):1–38. Télécharger.

Deng, M., Mehta, S., Sun, F. et Chen, T. (2002). Inferring domain-domain interactions from protein-protein interactions. *Genome research*, 12(10):1540–1548. Télécharger.

Desowitz, R. (1991). *The Malaria Capers : More Tales of Parasites and People, Research and Reality*. WW Norton & Company eds. New York.

Doolittle, R. (1986). *Of URFs and ORFs : A primer on how to analyze derived amino acid sequences*. Univ Science Books.

Duda, R., Hart, P. et Stork, D. (2001). *Pattern classification. 2nd Edition*. Wiley-Interscience.

Durbin, R., Eddy, S., Krogh, A. et Mitchison, G. (1998). *Biological sequence analysis : Probabilistic models of proteins and nucleic acids*. Cambridge University Press.

Dyson, H. et Wright, P. (2005). Intrinsically unstructured proteins and their functions. *Nature Reviews Molecular Cell Biology*, 6(3):197–208. Télécharger.

Dávalos, L. et Perkins, S. (2008). Saturation and base composition bias explain phylogenomic conflict in *Plasmodium*. *Genomics*, 91(5):433–442. Télécharger.

Ebers, G. et Stern, L. (1875). *Papyrus Ebers. Facsimile with a partial translation*.

Eddy, S. (1995). Multiple alignment using hidden markov models. *In Proceedinds of the International Conference on Intelligent System for Molecular Biology*, volume 3, pages 114–120.

Eddy, S. (1997). Maximum likelihood fitting of extreme value distributions. Technical notes, unpublished. Télécharger.

Eddy, S. (1998). Profile hidden markov models. *Bioinformatics*, 14(9):755–763. Télécharger.

Eddy, S. (2003). Hmmer user's guide version 2.3.2. Télécharger.

Eddy, S. (2008). A probabilistic model of local sequence alignment that simplifies statistical significance estimation. *PLoS Computational Biology*, 4(5). Télécharger.

Eddy, S. (2010). Hmmer user's guide version 3.0. Télécharger.

Ekman, D., Björklund, Å. et Elofsson, A. (2007). Quantification of the elevated rate of domain rearrangements in metazoa. *Journal of Molecular Biology*, 372(5):1337–1348. Télécharger.

Elofsson, A. et Sonnhammer, E. (1999). A comparison of sequence and structure protein domain families as a basis for structural genomics. *Bioinformatics*, 15(6):480–500. Télécharger.

Escalante, A. et Ayala, F. (1994). Phylogeny of the malarial genus *Plasmodium*, derived from rrna gene sequences. *Proceedings of the National Academy of Sciences of the U.S.A.*, 91(24):11373–11377. Télécharger.

Felsenstein, J. (1978). Cases in which parsimony or compatibility method will be positively misleading. *Systematic zoology*, 27(4):401–410. Télécharger.

Fields, S. et Song, O. (1989). A novel genetic system to detect protein protein interactions. *Nature*, 340(6230):245–246. Télécharger.

Finn, R., Marshall, M. et Bateman, A. (2005). iPfam : visualization of protein-protein interactions in PDB at domain and amino acid resolutions. *Bioinformatics*, 21(3):410–412. Télécharger.

Finn, R., Mistry, J., Schuster-Bockler, B., Griffiths-Jones, S., Hollich, V., Lassmann, T., Moxon, S., Marshall, M., Khanna, A., Durbin, R., Eddy, S., Sonnhammer, E. et Bateman, A. (2006). Pfam : clans, web tools and services. *Nucleic Acids Research*, 34(Database issue):D247–D251. Télécharger.

Finn, R., Mistry, J., Tate, J., Coggill, P., Heger, A., Pollington, J., Gavin, O., Gunasekaran, P., Ceric, G., Forslund, K., Holm, L., Sonnhammer, E., Eddy, S. et Bateman, A. (2010). The Pfam protein families database. *Nucleic Acids Research*, 38(Database issue):D211–D222. Télécharger.

Finn, R., Tate, J., Mistry, J., Coggill, P., Sammut, S., Hotz, H., Ceric, G., Forslund, K., Eddy, S., Sonnhammer, E. et Bateman, A. (2008). The pfam protein families database. *Nucleic Acids Research*, 36(Database issue):D281–D288. Télécharger.

Fitch, W. (2000). Homology : a personal view on some of the problems. *Trends in Genetics*, 16(5):227–231. Télécharger.

Fix, E. et Hodges, J. (1951). Discriminatory analysis. nonparametric discrimination : Consistency properties. *Technical Report 21-49-004*, 4:261–279. Télécharger.

Fleischmann, R., Adams, M., White., O., Clayton, R., Kirkness, E., Kerlavage, A., Bult, C., Tomb, J. et xand JM Merrick *et al.*, J. D. (1995). Whole-genome random sequencing and assembly of *Haemophilus influenzae Rd*. *Science*, 269(5223):496–512. Télécharger.

Florens, L., Washburn, M., Raine, J., Anthony, R., Grainger, M., Haynes, J., Moch, J., Muster, N., Sacci, J., Tabb, D., Witney, A., Wolters, D., Wu, Y., Gardner, M., Holder, A., Sinden, R., Yates, J. et Carucci, D. (2002). A proteomic view of the *Plasmodium falciparum* life cycle. *Nature*, 419(6906):520–526. Télécharger.

Florent, I., Charneau, S. et Grellier, P. (2004). *Plasmodium falciparum* genes differentially expressed during merozoite morphogenesis. *Molecular & Biochemical Parasitology*, 135(1):143–148. Télécharger.

Florent, I., Porcel, B., Guillaume, E., Silva, C. D., Artiguenave, F., Maréchal, E., Bréhélin, L., Gascuel, O., Charneau, S., Wincker, P. et Grellier, P. (2009). A *Plasmodium falciparum* fcb1-schizont-est collection providing clues to schizont specific gene structure and polymorphism. *BMC Genomics*, 10:235. Télécharger.

Forslund, K. et Sonnhammer, E. (2008). Predicting protein function from domain content. *Bioinformatics*, 24(15):1681–1687. Télécharger.

Foster, P. (2004). Modeling compositional heterogeneity. *Systematic Biology*, 53(3):485–495. Télécharger.

Frishman, D. et Argos, P. (1995). Knowledge-based protein secondary structure assignment. *Proteins*, 23(4):566–579. Télécharger.

Frishman, D. et Argos, P. (1996). Incorporation of non-local interactions in protein secondary structure prediction from the amino acid sequence. *Protein Engineering Design and Selection*, 9(2):133–142. Télécharger.

Fuentes-Prior, P., Iwanaga, Y., Huber, R., Pagila, R., Rumennik, G., Seto, M., Morser, J., Light, D. et Bode, W. (2000). Structural basis for the anticoagulant activity of the thrombin-thrombomodulin complex. *Nature*, 404(6777):518–525. Télécharger.

Gajria, B., Bahl, A., Brestelli, J., Dommer, J., Fischer, S., Gao, X., Heiges, M., Iodice, J., Kissinger, J., Mackey, A., Pinney, D., Roos, D., Stoeckert, C. J., Wang, H. et Brunk, B. (2008). ToxoDB : an integrated Toxoplasma gondii database resource. *Nucleic Acids Research*, 36(Database issue):D553.

Gajria, B., Bahl, A., Brestelli, J., Dommer, J., Fischer, S., Gao, X., Heiges, M., Iodice, J., Kissinger, J., Mackey, A. et *et al.* (2007). Toxodb : an integrated *Toxoplasma gondii* database resource. *Nucleic Acids Research*, 36(Database issue):D553–D556. Télécharger.

Galtier, N. et Gouy, M. (1998). Maximum-likelihood implementation of a nonhomogeneous model of dna sequence evolution for phylogenetic analysis. *Molecular Biology and Evolution*, 15(7):871–879. Télécharger.

Gardner, M., Hall, N., Fung, E., White, O., Berriman, M., Hyman, R., Carlton, J., Pain, A., Nelson, K., Bowman, S., Paulsen, I., James, K., Eisen, J., Rutherford, K., Salzberg, S., Craig, A., Kyes, S., Chan, M., Nene, V., Shallom, S., Suh, B., Peterson, J., Angiuoli, S., Pertea, M., Allen, J., Selengut, J., Haft, D., Mather, M., Vaidya, A., Martin, D., Fairlamb, A., Fraunholz, M., Roos, D., Ralph, S., McFadden, G., Cummings, L., Subramanian, G., Mungall, C., Venter, J., Carucci, D., Hoffman, S., Newbold, C., Davis, R., Fraser, C. et Barrell, B. (2002). Genome sequence of the human malaria parasite *Plasmodium falciparum*. *Nature*, 419(6906):498–511. Télécharger.

Garnham, P. (1966). *Malaria parasites and other haemosporidia*. Blackwell Science Ltd.

Geer, L., Domrachev, M., Lipman, D. et Bryant, S. (2002). Cdart : protein homology by domain architecture. *Genome Research*, 12(10):1619–1623. Télécharger.

Gerstein, M. et Hegyi, H. (2001). Annotation transfer for genomics : measuring functional divergence in multi-domain proteins. *Genome Research*, 11(10):1632–1640. Télécharger.

Gerstein, M., Sonnhammer, E. et Chothia, C. (1994). Volume changes in protein evolution. *Journal of Molecular Biology*, 236(4):1067–1078. Télécharger.

Ghouila, A., Terrapon, N., Gascuel, O., Guerfali, F., Laouini, D., Maréchal, E. et Bréhélin, L. (2010). Eupathdomains : the divergent domain database for eukaryotic pathogens. *Infection, Genetics and Evolution*, Sous presse(E-pub Novembre). Télécharger.

Gilbert, W. (1978). Why genes in pieces ? *Nature*, 271(5645):501.

Godzik, A., Kolinski, A. et Skolnick, J. (1992). Topology fingerprint approach to the inverse protein folding problem. *Journal of molecular biology*, 227(1):227–238. Télécharger.

Goffeau, A., Barrell, B., Bussey, H., Davis, R., Dujon, B., Feldmann, H., Galibert, F., Hoheisel, J., Jacq, C., Johnston, M., Louis, E., Mewes, H., Murakami, Y., Philippsen, P., Tettelin, H. et Oliver, S. (1996). Life with 6000 genes. *Science*, 274(5287):563–567. Télécharger.

Gomez, S. et Rzhetsky, A. (2002). Towards the prediction of complete protein–protein interaction networks. *In Pacific Symposium on Biocomputing*, pages 413–424.

Good, M. (2009). The hope but challenge for developing a vaccine that might control malaria. *European journal of immunology*, 39(4):939–943. donwload.

Gough, J. et Chothia, C. (2002). Superfamily : Hmms representing all proteins known structure. scop sequence searches, alignements and genome assignement. *Nuclear Acids Research*, 30(1):268–272. Télécharger.

Gouzy, J., Corpet, F. et Kahn, D. (1999). Whole genome protein domain analysis using a new method for domain clustering. *Computers & chemistry*, 23(3-4):333–340. Télécharger.

Gracy, J. et Argos, P. (1998a). Automated protein sequence database classification. i. integration of compositional similarity search, local similarity search, and multiple sequence alignment. *Bioinformatics*, 14(2):164–173. Télécharger.

Gracy, J. et Argos, P. (1998b). Automated protein sequence database classification. ii. delineation of domain boundaries from sequence similarities. *Bioinformatics*, 14(2):174–187. Télécharger.

Gribskov, M., McLachlan, A. et Eisenberg, D. (1987). Profile analysis : detection of distantly related proteins. *Proceedings of the National Academy of Sciences of the U.S.A.*, 84(13):4355–4358. Télécharger.

Grundy, W., Bailey, T., Elkan, C. et Baker, M. (1997). Meta-meme : Motif-based hidden markov models of protein families. *Computer Applications in the Biosciences*, 13(4):397–406. Télécharger.

Guimarães, K., Jothi, R., Zotenko, E. et Przytycka, T. (2006). Predicting domain-domain interactions using a parsimony approach. *Genome Biology*, 7(11):R104. Télécharger.

Guindon, S. et Gascuel, O. (2003). A simple, fast, and accurate algorithm to estimate large phylogenies by maximum likelihood. *Systematic biology*, 52(5):696–704. Télécharger.

Guindon, S. et Gascuel, O. (2007). *Modelling the variability of evolutionary processes*. Oxford University Press.

Gumbel, E. (1958). *Statistics of extremes*. Columbia University Press.

Hadley, C. et Jones, D. (1999). A systematic comparison of protein structure classifications : SCOP, CATH and FSSP. *Structure*, 7(9):1099–1112. Télécharger.

Haft, D., Selengut, J. et White, O. (2003). The tigrfams database of protein families. *Nucleic Acids Research*, 31(1):371–373. Télécharger.

Hagner, S., Misof, B., Maier, W. et Kampen, H. (2007). Bayesian analysis of new and old malaria parasite dna sequence data demonstrates the need for more phylogenetic signal to clarify the descent of *Plasmodium falciparum*. *Parasitology Research*, 101(3):493–503. Télécharger.

Hahne, F., Mehrle, A., Arlt, D., Poustka, A., Wiemann, S. et Beissbarth, T. (2008). Extending pathways based on gene lists using interpro domain signatures. *BMC bioinformatics*, 9:3. Télécharger.

Halabi, N., Rivoire, O., Leibler, S. et Ranganathan, R. (2009). Protein sectors : evolutionary units of three-dimensional structure. *Cell*, 138(4):774–786. Télécharger.

Hall, N., Karras, M., Raine, J., Carlton, J., Kooij, T., Berriman, M., Florens, L., Janssen, C., Pain, A., Christophides, G., James, K., Rutherford, K., Harris, B., Harris, D., Churcher, C., Quail, M., Ormond, D., Doggett, J., Trueman, H., Mendoza, J., Bidwell, S., Rajandream, M., Carucci, D., 3rd Yates, J., Kafatos, F., Janse, C., Barrell, B., Turner, C., Waters, A. et Sinden, R. (2005). A comprehensive survey of the *Plasmodium* life cycle by genomic, transcriptomic, and proteomic analyses. *Science*, 307(5706):82–86. Télécharger.

Hannenhalli, S. et Russell, R. (2000). Analysis and prediction of functional sub-types from protein sequence alignments. *Journal of Molecular Biology*, 303(1):61–76. Télécharger.

Harrison, G. (1978). *Mosquitoes, malaria and man : a history of the hostilities since 1880*. John Murray.

Hastie, T., Tibshirani, R. et Friedman, J. (2001). *The elements of statistical learning : data mining, inference and prediction. 2nd Edition*. Springer series in statistics.

Haussler, D., Krogh, A., Mian, I. et Sjölander, K. (1993). Protein modeling using hidden markov models : analysis of globins. *In 26th Hawaii International Conference on System Sciences*. Télécharger.

Hay, S., Guerraa, C., Tatema, A., Noorb, A. et Snowc, R. (2004). The global distribution and population at risk of malaria : past, present, and future. *The Lancet Infectious Diseases*, 4(6):327–336. Télécharger.

Hayakawa, T., Culleton, R., Otani, H., Horii, T. et Tanabe, K. (2008). Big bang in the evolution of extant malaria parasites. *Molecular Biology and Evolution*, 25(10):2333–2339. Télécharger.

Hayete, B. et Bienkowska, J. (2005). Gotrees : Predicting go associations from protein domain composition using decision trees. *Pacific Symposium on Biocomputing*, 10:127–138. Télécharger.

Heger, A. et Holm, L. (2003). Exhaustive enumeration of protein domain families. *Journal of molecular biology*, 328(3):749–767. Télécharger.

Heiges, M., Wang, H., Robinson, E., Aurrecoechea, C., Gao, X., Kaluskar, N., Rhodes, P., Wang, S., He, C., Su, Y., Miller, J., Kraemer, E. et Kissinger, J. (2006). CryptoDB : a Cryptosporidium bioinformatics resource update. *Nucleic Acids Research*, 34(Database issue):D419–D422. Télécharger.

Henikoff, J., Greene, E., Pietrokovski, S. et Henikoff, S. (2000). Increased coverage of protein families with the blocks database servers. *Nucleic Acids Research*, 28(1):228–230. Télécharger.

Henikoff, S. et Henikoff, J. (1992). Amino acid substitution matrices from protein blocks. *Proceedings of the National Academy of Sciences of the U.S.A.*, 89(22):10915–10919. Télécharger.

Hino, S. et Miyata, H. (2007). Torque teno virus TTV : current status. *Reviews in medical virology*, 17(1):45–57. Télécharger.

Holm, L., Kaariainen, S., Rosenstrom, P. et Schenkel, A. (2008). Searching protein structure databases with DaliLite v. 3. *Bioinformatics*, 24(23):2780–2781. Télécharger.

Holm, L. et Sander, C. (1994). The FSSP database of structurally aligned protein fold families. *Nucleic Acids Research*, 22(17):3600–3609. Télécharger.

Holm, L. et Sander, C. (1998). Touring protein fold space with Dali/FSSP. *Nucleic Acids Research*, 26(1):316–319. Télécharger.

Holt, R., Subramanian, G., Halpern, A., Sutton, G., Charlab, R., Nusskern, D., Wincker, P., Clark, A., Ribeiro, J. et *et al.*, R. W. (2002). The genome sequence of the malaria mosquito *Anopheles gambiae*. *Science*, 298(5591):129–149. Télécharger.

Hubbard, T., Aken, B., Ayling, S., Ballester, B., Beal, K., Bragin, E., Brent, S., Chen, Y., Clapham, P., Clarke, L., Coates, G., Fairley, S., Fitzgerald, S., Fernandez-Banet, J., Gordon, L., Graf, S., Haider, S., Hammond, M., Holland, R., Howe, K., Jenkinson, A., Johnson, N., Kahari, A., Keefe, D., Keenan, S., Kinsella, R., Kokocinski, F., Kulesha, E., Lawson, D., Longden, I., Megy, K., Meidl, P., Overduin, B., Parker, A., Pritchard, B., Rios, D., Schuster, M., Slater, G., Smedley, D., Spooner, W., Spudich, G., Trevanion, S., Vilella, A., Vogel, J., White, S., Wilder, S., Zadissa, A., Birney, E., Cunningham, F., Curwen, V., Durbin, R., Fernandez-Suarez, X., Herrero, J., Kasprzyk, A., Proctor, G., Smith, J., Searle, S. et Flicek, P. (2009). Ensembl 2009. *Nucleic Acids Research*, 37(Database issue):D690–D697. Télécharger.

Hughey, R. et Krogh, A. (1996). Hidden markov models for sequence analysis. extension of the basic method. *CABIOS*, 12(2):95–107. Télécharger.

Hulo, N., Bairoch, A., Bulliard, V., Cerutti, L., Castro, E. D., Langendijk-Genevaux, P., Pagni, M. et Sigrist, C. (2006). The prosite database. *Nucleic Acids Research*, 34(Database issue):D227–D230. Télécharger.

Hulo, N., Bairoch, A., Bulliard, V., Cerutti, L., Cuche, B., Castro, E. D., Lachaize, C., Langendijk-Genevaux, P. et Sigrist, C. (2008). The 20 years of prosite. *Nucleic Acids Research*, 36(Database issue):D245–249. Télécharger.

Hunter, S., Apweiler, R., Attwood, T., Bairoch, A., Bateman, A., Binns, D., Bork, P., Das, U., Daugherty, L., Duquenne, L., Finn, R., Gough, J., Haft, D., Hulo, N., Kahn, D., Kelly, E., Laugraud, A., Letunic, I., Lonsdale, D., Lopez, R., Madera, M., Maslen, J., McAnulla, C., McDowall, J., Mistry, J., Mitchell,

A., Mulder, N., Natale, D., Orengo, C., Quinn, A., Selengut, J., Sigrist, C., Thimma, M., Thomas, P., Valentin, F., Wilson, D., Wu, C. et Yeats, C. (2009). Interpro : the integrative protein signature database. *Nucleic Acid Research*, 37(Database issue):D211–215. Télécharger.

Ito, T., Chiba, T., Ozawa, R., Yoshida, M., Hattori, M. et Sakaki, Y. (2001). A comprehensive two-hybrid analysis to explore the yeast protein interactome. *Proceedings of the National Academy of Sciences*, 98(8):4569–4574. Télécharger.

Janssen, C., Phillips, R., Turner, C. et Barrett, M. (2004). *Plasmodium* interspersed repeats : the major multigene superfamily of malaria parasites. *Nucleic Acids Research*, 32(19):5712–5720. Télécharger.

Jaskiewicz, L. et Filipowicz, W. (2008). Role of dicer in posttranscriptional rna silencing. *Current Topics in Microbiology and Immunology*, 320:77–97. Télécharger.

Johnson, M., Overington, J. et Blundell, T. (1993). Alignment and searching for common protein folds using a data bank of structural templates. *Journal of molecular biology*, 231(3):735. Télécharger.

Jonassen, I., Collins, J. et Higgins, D. (1995). Finding flexible patterns in unaligned protein sequences. *Protein Science*, 4(8):1587–1595. Télécharger.

Jones, D., Taylor, W. et Thornton, J. (1992a). A new approach to protein fold recognition. *Nature*, 358(6381):86–89. Télécharger.

Jones, D., Taylor, W. et Thornton, J. (1992b). The rapid generation of mutation data matrices from protein sequences. *Bioinformatics*, 8(3):275–282. Télécharger.

Jothi, R., Cherukuri, P., Tasneem, A. et Przytycka, T. (2006). Co-evolutionary analysis of domains in interacting proteins reveals insights into domain–domain interactions mediating protein–protein interactions. *Journal of molecular biology*, 362(4):861–875. Télécharger.

Kabsch, W. et Sander, C. (1983). Dictionary of protein secondary structure : pattern recognition of hydrogen-bonded and geometrical features. *Biopolymers*, 22(12):2577–2637. Télécharger.

Kanehisa, M. et Goto, S. (2000). Kegg : Kyoto encyclopedia of genes and genomes. *Nucleic Acids Research*, 28(1):27–30. Télécharger.

Karplus, K., Barret, C. et Hughey, R. (1998). Hidden markov models for detecting remote protein homologies. *Bioinformatics*, 14(10):846–856. Télécharger.

Karplus, K., Karchin, R., Shackelford, G. et Hughey, R. (2005). Calibrating E-values for hidden Markov models using reverse-sequence null models. *Bioinformatics*, 21(22):4107–4115. Télécharger.

Kaufman, T. et Ruveda, E. (2005). The quest for quinine : those who won the battles and those who won the war. *Angewandte Chemie International Edition*, 44(6):854–885. Télécharger.

Kaur, K., Jain, M., Kaur, T. et Jain, R. (2009). Antimalarials from nature. *Bioorganic and Medicinal Chemistry*, 17(9):3229–3256. Télécharger.

Kersey, P., Bower, L., Morris, L., Horne, A., Petryszak, R., Kanz, C., Kanapin, A., Das, U., Michoud, K., Phan, I., Gattiker, A., Kulikova, T., Faruque, N., Duggan, K., Mclaren, P., Reimholz, B., Duret, L., Penel, S., Reuter, I. et Apweiler, R. (2005). Integr8 and Genome Reviews : integrated views of complete genomes and proteomes. *Nucleic acids research*, 33(Database Issue):D297–D302. Télécharger.

Kim, W., Park, J. et Suh, J. (2002). Large Scale Statistical Prediction of Protein-Protein Interaction by Potentially Interacting Domain (PID) Pair. *Genome Informatics*, 13:42–50. Télécharger.

King, R. et Sternberg, M. (1996). Identification and application of the concepts important for accurate and reliable protein secondary structure prediction. *Protein science*, 5(11): 2298–2310. Télécharger.

Kissinger, J., Souza, P., Soarest, C., Paul, R., Wahl, A., Rathore, D., McCutchan, T. et Krettli, A. (2002). Molecular phylogenetic analysis of the avian malarial parasite *Plasmodium (Novyella) juxtanucleare*. *Journal of Parasitology*, 88(4):769–773. Télécharger.

Kleene, S. (1951). *Representation of events in nerve nets and finite automata*. RAND.

Kohler, S., Delwiche, C., Denny, P., Tilney, L., Webster, P., Wilson, R., Palmer, J. et Roos, D. (1997). A plastid of probable green algal origin in apicomplexan parasites. *Science*, 275:1485–1489. Télécharger.

Kolinski, A., A, G. et Skolnick, J. (1993). A general method for the prediction of the three dimensional structure and folding pathway of globular proteins : application to designed helical proteins. *Journal of Chemical Physics*, 98:7420–7433. Télécharger.

Kolmogorov, A. (1968). Three approaches to the quantitative definition of information. *International Journal of Computer Mathematics*, 2(1):157–168. Télécharger.

Kooij, T., Carlton, J., Bidwell, S., Hall, N., Ramesar, J., Janse, C. et Waters, A. (2005). A *Plasmodium* whole-genome synteny map : indels and synteny breakpoints as foci for species-specific genes. *PLoS Pathogens*, 1(4):E44. Télécharger.

Kooij, T., Janse, C. et Waters, A. (2006). *Plasmodium* post-genomics : better the bug you know ? *Nature*, 4(5):344–357. Télécharger.

Koonin, E., Aravind, L. et Kondrashov, A. (2000). The impact of comparative genomics on our understanding of evolution. *Cell*, 101(6):573–576. Télécharger.

Krause, A., Stoye, J. et Vingron, M. (2000). The systers protein sequence cluster set. *Nucleic Acids Research*, 28(1):270–272. Télécharger.

Krogh, A., Mian, I. et Haussler, D. (1994). A hidden markov model that finds genes in e. coli dna. *Nucleic Acids Research*, 22(22):4768–4778. Télécharger.

Krogh, A. et Mitchison, G. (1995). Maximum entropy weighting of aligned sequences of proteins or DNA. *In Proc. Int. Conf. on Intelligent Systems in Molecular Biology*, volume 3, pages 215–221. Télécharger.

Kuhn, R., Karolchik, D., Zweig, A., Wang, T., Smith, K., Rosenbloom, K., Rhead, B., Raney, B., Pohl, A., Pheasant, M., Meyer, L., Hsu, F., Hinrichs, A., Harte,

R., Giardine, B., Fujita, P., Diekhans, M., Dreszer, T., Clawson, H., Barber, G., Haussler, D. et Kent, W. (2009). The UCSC genome browser database : update 2009. *Nucleic acids research*, 37(Database issue):D755–D761. Télécharger.

Kummerfeld, S. et Teichmann, S. (2005). Relative rates of gene fusion and fission in multi-domain proteins. *Trends in Genetics*, 21(1):25–30. Télécharger.

Kun, J., Hesselbach, J., Schreiber, M., Scherf, A., Gysin, J., Mattei, D., da Silva, L. P. et Müller-Hill, B. (1991). Cloning and expression of genomic dna sequences coding for putative erythrocyte membrane-associated antigens of *Plasmodium falciparum*. *Resaerch in Immunology*, 142(3):199–210. Télécharger.

Kuo, C. et Kissinger, J. (2008). Consistent and contrasting properties of lineage-specific genes in the apicomplexan parasites *Plasmodium* and *Theileria*. *BMC Evolutionary Biology*, 8:108. Télécharger.

Lathrop, R. (1994). The protein threading problem with sequence amino acid interaction preferences is NP-complete. *Protein Engineering Design and Selection*, 7(9):1059–1068. Télécharger.

Lau, A. (2009). An overview of the *Babesia*, *Plasmodium* and *Theileria* genomes : A comparative perspective. *Molecular & Biochemical Parasitology*, 164(1):1–8. Télécharger.

Laveran, A. (1880). Note sur un nouveau parasite trouvé dans le sang de plusieurs malades atteints de fièvre palustre. *Bulletin de l'Académie de médecine*, 9:1235–1236.

Le, S. et Gascuel, O. (2008). An improved general amino acid replacement matrix. *Molecular Biology and Evolution*, 25(7):1307–1320. Télécharger.

Le Roch, K., Zhou, Y., Blair, P., Grainger, M., Moch, J., Haynes, J., Vega, P. D. L., Holder, A., Batalov, S., Carucci, D. et Winzeler, E. (2003). Discovery of gene function by expression profiling of the malaria parasite life cycle. *Science*, 301(5639): 1503–1508. Télécharger.

Lee, B. et Lee, D. (2009). Protein comparison at the domain architecture level. *BMC bioinformatics*, 10(Suppl 15):S5. Télécharger.

Lempel, A. et Ziv, J. (1976). On the complexity of finite sequences. *IEEE Transactions on Information Theory*, 22(1):75–81. Télécharger.

Lesk, A. (1988). *Computational Molecular Biology : Sources and Methods for Sequence Analysis*. Oxford University Press.

Letunic, I., Doerks, T. et Bork, P. (2009). Smart 6 : recent updates and new developments. *Nucleic Acids Research*, 37(Database issue):D229–D232. Télécharger.

Li, F., Sonbuchner, L., Kyes, S., Epp, C. et Deitsh, K. (2007). Nuclear non-coding rnas are transcribed from the centromeres of *Plasmodium falciparum* and are associated with centromeric chromatin. *Journal of Biological Chemistry*, 283(9):5692–5698. Télécharger.

Li, L., Stoeckert, C. J. et Roos, D. (2003). Orthomcl : Identification of ortholog groups for eukaryotic genomes. *Genome Research*, 13(9):2178–2189. Télécharger.

Lima, T., Auchincloss, A., Coudert, E., Keller, G., Michoud, K., Rivoire, C., Bulliard, V., de Castro, E., Lachaize, C., Baratin, D., Phan, I., Bougueleret, L. et Bairoch, A. (2009). HAMAP : a database of completely sequenced microbial proteome sets and manually curated microbial protein families in UniProtKB/Swiss-Prot. *Nucleic Acids Research*, 37(Database issue):D471–D478. Télécharger.

Lindahl, E. et Elofsson, A. (2000). Identification of related protein on family, superfamily and fold level. *Journal of Molecular Biology*, 295(3):613–625. Télécharger.

Linderstrøm-Lang, K. (1952). Proteins and enzymes. *Lane Medical Lectures*, 6.

Llinás, M., Bozdech, Z., Wong, E., Adai, A. et DeRisi, J. (2006). Comparative whole genome transcriptome analysis of three *Plasmodium falciparum* strains. *Nucleic Acids Research*, 34(4):1166–1173. Télécharger.

Lloyd, S. (1957). Least squares quantization in pcm. *Technical Report*. Published in IEEE transactions on Information Theory **28** :129-137.

Lockhart, P., Howe, C., Bryant, D., Beanland, T. et Larkum, A. (1992). Substitutional bias confounds inference of cyanelle origins from sequence data. *Journal of Molecular Evolution*, 34(2):153–162. Télécharger.

Madera, M. et Gough, J. (2002). A comparison of profile hidden Markov model procedures for remote homology detection. *Nucleic acids research*, 30(19):4321–4328. Télécharger.

Maiorov, V. et Crippen, G. (1992). Contact potential that recognizes the correct folding of globular proteins. *Journal of molecular biology*, 227(3):876–888. Télécharger.

Mamitsuka, H. (1996). A learning method of hidden Markov models for sequence discrimination. *Journal of Computational Biology*, 3(3):361–373. Télécharger.

Martinsen, E., Perkins, S. et Schall, J. (2007). A three-genome phylogeny of malaria parasites (*Plasmodium* and closely related genera) : evolution of life-history traits and host switches. *Molecular Phylogenetics and Evolution*, 47(1):261–273. Télécharger.

Maréchal, E. et Cesbron-Delauw, M. (2001). The apicoplast : a new member of the plastid family. *Trends in Plant Science*, 6(5):200–205. Télécharger.

McClure, M., Smith, C. et Elton, P. (1996). Parameterization studies for the SAM and HMMER methods of hidden Markov model generation. *In Proceedings of the International Conference on Intelligent Systems for Molecular Biology*, volume 4, pages 155–164. Télécharger.

McCutchan, T., Kissinger, J., Touray, M., Rogers, M., Li, J., Sullivan, M., Braga, E., Krettli, A. et Miller, L. (1996). Comparison of circumsporozoite proteins from avian and mammalian malarias : biological and phylogenetic implications. *Proceedings of the National Academy of Sciences of the U.S.A.*, 93(21):11889–11894. Télécharger.

McGarvey, P., Huang, H., Barker, W., Orcutt, B., Garavelli, J., Srinivasarao, G., Yeh, L., Xiao, C. et Wu, C. (2000). Pir : a new resource for bioinformatics. *Bioinformatics*, 16(3):290–291. Télécharger.

McLaughlin, W., Chen, K., Hou, T. et Wang, W. (2007). On the detection of functionally coherent groups of protein domains with an extension to protein annotation. *BMC Bioinformatics*, 8:390. Télécharger.

McRobert, L., Jiang, S., Stead, A. et McConkey, G. (2005). Plasmodium falciparum : Interaction of shikimate analogues with antimalarial drugs. *Experimental parasitology*, 111(3):178–181. Télécharger.

Mi, H., Guo, N., Kejariwal, A. et Thomas, P. (2007). PANTHER version 6 : protein sequence and function evolution data with expanded representation of biological pathways. *Nucleic Acids Research*, 35(Database issue):D247–D252. Télécharger.

Mi, H., Lazareva-Ulitsky, B., Loo, R., Kejariwal, A., Vandergriff, J., Rabkin, S., Guo, N., Muruganujan, A., Doremieux, O., Campbell, M., Kitano, H. et Thomas, P. (2005). The panther database of protein families, subfamilies, functions and pathways. *Nucleic Acids Research*, 33(Database issue):D284–288. Télécharger.

Miller, J., McLachlan, A. et Klug, A. (1985). Repetitive zinc-binding domains in the protein transcription factor IIIA from Xenopus oocytes. *The EMBO journal*, 4(6):1609–1614. Télécharger.

Milosavljevic, A. et Jurka, J. (1993). Discovering simple DNA sequences by the algorithmic significance method. *CABIOS*, 9(4):407–411. Télécharger.

Mitchell, T. (1997). *Machine learning*. Mac Graw Hill.

Miyazawa, S. et Jernigan, R. (1985). Estimation of effective interresidue contact energies from protein crystal structures : quasi-chemical approximation. *Macromolecules*, 18(3): 534–552. Télécharger.

Mourier, T., Carret, C., Kyes, S., Christodoulou, Z., Gardner, P., Jeffares, D., Pinches, R., Barrell, B., Berriman, M., Griffith-Jones, S., Ivens, A., Newbold, C. et Pain, A. (2007). Genome-wide discovery and verification of novel structured rnas in *Plasmodium falciparum*. *Genome Research*, 18(2):281–292. Télécharger.

Mu, J., Duan, J., Makova, K., Joy, D., andOH Branch, C. H., Li, W. et X, S. (2002). Chromosome-wide snps reveal an ancient origin for *Plasmodium falciparum*. *Nature*, 418(6895):323–324. Télécharger.

Mulder, N., Apweiler, R., Attwood, T., Bairoch, A., Barrell, D., Bateman, A., Binns, D., Biswas, M., Bradley, P., Bork, P., Bucher, P., Copley, R., Courcelle, E., Das, U., Durbin, R., Falquet, L., Fleischmann, W., Griffiths-Jones, S., Haft, D., Harte, N., Hulo, N., Kahn, D., Kanapin, A., Krestyaninova, M., Lopez, R., Letunic, I., Lonsdale, D., Silventoinen, V., Orchard, S., Pagni, M., Peyruc, D., Ponting, C., Selengut, J., Servant, F., Sigrist, C., Vaughan, R. et Zdobnov, E. (2003). The interpro database, 2003 brings increased coverage and new features. *Nucleic Acid Research*, 31(1):315–318. Télécharger.

Mulder, N., Apweiler, R., Attwood, T., Bairoch, A., Bateman, A., Binns, D. et *et al.* (2007). New developments in the interpro database. *Nucleic Acid Research*, 35(Database issue):D224–228. Télécharger.

Murzin, A., Brenner, S., Hubbard, T. et Chothia, C. (1995). Scop : a structural classification of proteins database for the investigation of sequences and structures. *Journal of Molecular Biology*, 247(4):536–540. Télécharger.

Musto, H., Rodriguez-Maseda, H. et Bernardi, G. (1995). Compositional properties of nuclear genes from *Plasmodium falciparum*. *Gene*, 152(1):127–132. Télécharger.

Ng, S., Zhang, Z. et Tan, S. (2003). Integrative approach for computationally inferring protein domain interactions. *Bioinformatics*, 19(8):923–929. Télécharger.

Nikolskaya, A., Arighi, C., Huang, H., Barker, W. et Wu, C. (2007). Pirsf family classification system for protein functional and evolutionary analysis. *Evolutionary Bioinformatics Online*, 2:197–209. Télécharger.

Nye, T., Berzuini, C., Gilks, W., Babu, M. et Teichmann, S. (2005). Statistical analysis of domains in interacting protein pairs. *Bioinformatics*, 21(7):993–1001. Télécharger.

Orengo, C., Michie, A., Jones, S., Jones, D., Swindells, M. et Thornton, J. (1997). CATH – a hierarchic classification of protein domain structures. *Structure*, 5(8):1093–1108. Télécharger.

Orengo, C. et Taylor, W. (1996). SSAP : sequential structure alignment program for protein structure comparison. *Methods in Enzymology*, 266:617–635. Télécharger.

Pagel, P., Wong, P. et Frishman, D. (2004). A domain interaction map based on phylogenetic profiling. *Journal of molecular biology*, 344(5):1331–1346. Télécharger.

Park, J., Karplus, K., Barrett, C., Hughey, R., Haussler, D., Hubbard, T. et Chothia, C. (1998). Sequence comparisons using multiple sequences detect three times as many remote homologues as pairwise methods. *Journal of Molecular Biology*, 284(4):1201–1210. Télécharger.

Pasek, S. (2006). *Le domaine protéique, une unité d'homologie pertinente en génomique comparative*. Thèse de doctorat, Université d'Evry Val d'Essonne. Télécharger.

Pasek, S., Bergeron, A., Risler, J., Louis, A., Ollivier, E. et Raffinot, M. (2005). Identification of genomic features using microsyntenies of domains : domain teams. Télécharger.

Pasek, S., Risler, J. et Brézellec, P. (2006a). Gene fusion/fission is a major contributor to evolution of multi-domain bacterial proteins. *Bioinformatics*, 22(12):1418–1423. Télécharger.

Pasek, S., Risler, J. et Brézellec, P. (2006b). The role of domain redundancy in genetic robustness against null mutations. *Journal of molecular biology*, 362(2):184–191. Télécharger.

Pauling, L., COREY, R. et Branson, H. (1951). The structure of proteins ; two hydrogen-bonded helical configurations of the polypeptide chain. *Proceedings of the National Academy of Sciences of the U.S.A.*, 37(4):205–211. Télécharger.

Pawson, T. et Nash, P. (2003). Assembly of cell regulatory systems through protein interaction domains. *Science*, 300(5618):445–452. Télécharger.

Perkins, S. (2008). Molecular systematics of the three mitochondrial protein-coding genes of malaria parasites : Corroborative and new evidence for the origins of human malaria. *Mitochondrial DNA*, 19(6):471–478. Télécharger.

Perkins, S. et Schall, J. (2002). A molecular phylogeny of malarial parasites recovered from cytochrome b gene sequences. *Journal of Parasitology*, 88(5):972–978. Télécharger.

Pierce, S. et Miller, L. (2009). World malaria day 2009 : What malaria knows about the immune system that immunologists still do not. *The Journal of Immunology*, 182(9): 5171–5177. donwload.

Pizzi, E. et Frontali, C. (2001). Low-complexity regions in *Plasmodium falciparum* proteins. *Genome Research*, 11(2):218–229. Télécharger.

Pruitt, K., Tatusova, T. et Maglott, D. (2007). NCBI reference sequences (RefSeq) : a curated non-redundant sequence database of genomes, transcripts and proteins. *Nucleic Acids Research*, 35(Database issue):D61–D65. Télécharger.

Rabiner, L. (1989). A tutorial on hidden markov models and selected applications in speech recognition. *In Proceedings of the IEEE*, pages 257–286. Télécharger.

Raghavachari, B., Tasneem, A., Przytycka, T. et Jothi, R. (2008). DOMINE : a database of protein domain interactions. *Nucleic Acids Research*, 36(Database issue): D656–D661. Télécharger.

Rambaut, A. et Grassly, N. (1997). Seq-gen : An application for the monte carlo simulation of dna sequence evolution along phylogenetic trees. *Comput. Appl. Biosci.*, 13(3):235–238. Télécharger.

Raphael, C. (1999). Automatic segmentation of acoustic musical signals using hidden markov models. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 21(4):360–370. Télécharger.

Redfern, O., Harrison, A., Dallman, T., Pearl, F. et Orengo, C. (2007). CATHEDRAL : a fast and effective algorithm to predict folds and domain boundaries from multidomain protein structures. *PLoS Comput Biol*, 3(11):e232. Télécharger.

Rehmsmeier, M. et Vingron, M. (2001). Phylogenetic information improves homology detection. *Proteins*, 45(4):360–371. Télécharger.

Reynolds, J. et Tanford, C. (2003). *Nature's Robots : A History of Proteins*. Oxford University Press.

Riley, R., Lee, C., Sabatti, C. et Eisenberg, D. (2005). Inferring protein domain interactions from databases of interacting proteins. *Genome Biology*, 6(10):R89. Télécharger.

Ron, D., Singer, Y. et Tishby, N. (1996). The power of amnesia : learning probabilistic automata with variable memory length. *Machine Learning*, 25(2-3):117–149. Télécharger.

Rost, B. et Sander, C. (1993). Prediction of protein secondary structure at better than 70% accuracy. *Journal of Molecular Biology*, 232:584–584. Télécharger.

Roussilhon, C., Oeuvray, C., M
"uller-Graf, C., Tall, A., Rogier, C., Trape, J., Theisen, M., Balde, A., Pérignon, J. et Druilhe, P. (2007). Long-term clinical protection from *falciparum* malaria is

strongly associated with igg3 antibodies to merozoite surface protein 3. *PLoS Medicine*, 4(11):e320. Télécharger.

Roy, S. et Irimia, M. (2008). Origins of human malaria : rare genomic changes and full mitochondrial genomes confirm the relationship of *Plasmodium falciparum* to other mammalian parasites but complicate the origins of *Plasmodium vivax*. *Molecular Biology and Evolution*, 25(6):1192–1198. Télécharger.

Sachs, J. et Malaney, P. (2002). The economic and social burden of malaria. *Nature*, 415(6872):680–685. Télécharger.

Sagot, M. et Marsan, L. (2000). Algorithms for extracting structured motifs using a suffix tree with an application to promoter and regulatory site consensus identification. *Journal of Computational Biology*, 7(3-4):345–362. Télécharger.

Salamov, A. et Solovyev, V. (1995). Prediction of protein secondary structure by combining nearest-neighbor algorithms and multiple sequence alignments. *Journal of Molecular Biology*, 247(1):11–15. Télécharger.

Sander, C. et Schneider, R. (1991). Database of homology-derived protein structures and the structural meaning of sequence alignment. *Proteins*, 9(1):56–68. Télécharger.

Saunders, L. et Barber, G. (2003). The dsrna binding protein family : critical roles, diverse cellular functions. *FASEB Journal*, 17(9):961–983. Télécharger.

Sayers, E., Barrett, T., Benson, D., Bryant, S., Canese, K., Chetvernin, V., Church, D., DiCuccio, M., Edgar, R., Federhen, S., Feolo, M., Geer, L., Helmberg, W., Kapustin, Y., Landsman, D., Lipman, D., Madden, T., Maglott, D., Miller, V., Mizrachi, I., Ostell, J., Pruitt, K., Schuler, G., Sequeira, E., Sherry, S., Shumway, M., Sirotkin, K., Souvorov, A., Starchenko, G., Tatusova, T., Wagner, L., Yaschenko, E. et Ye, J. (2009). Database resources of the National Center for Biotechnology Information. *Nucleic Acids Research*, 37(Database issue):D5–D15. Télécharger.

Schellman, J. et Schellman, C. (1997). Kaj ulrik linderstrflm-lang (1896-1959). *Protein Science*, 6(5):10092–10100. Télécharger.

Schneider, A., Dessimoz, C. et Gonnet, G. (2007). Oma browser – exploring orthologous relations across 352 complete genomes. *Bioinformatics*, 23(16):2180–2182. Télécharger.

Schultz, J., Milpetz, F., Bork, P. et Ponting, C. (1998). SMART, a simple modular architecture research tool : identification of signaling domains. Télécharger.

Scott, M., Thomas, D. et Hallett, M. (2004). Predicting subcellular localization via protein motif co-occurrence. *Genome Research*, 14(10A):1957–1966. Télécharger.

Sela, M., White, F. et Anfinsen, C. (1957). Reductive cleavage of disulfide bridges in ribonuclease. *Science*, 125(3250):691–692. Télécharger.

Selengut, J., Haft, D., Davidsen, T., Ganapathy, A., Gwinn-Giglio, M., Nelson, W., Richter, A. et White, O. (2007). TIGRFAMs and Genome Properties : tools for the assignment of molecular function and biological process in prokaryotic genomes. *Nucleic Acids Research*, 35(Database issue):D260–D264. Télécharger.

Servant, F., Bru, C., Carrère, S., Courcelle, E., Gouzy, J., Peyruc, D. et Kahn, D. (2002). Prodom : Automated clustering of homologous domains. *Briefings in Bioinformatics*, 3(3):246–251. Télécharger.

Sibbald, P. et Argos, P. (1990). Weighting aligned protein or nucleic acid sequences to correct for unequal representation. *Journal of Molecular Biology*, 216(4):813–818.

Sickmeier, M., Hamilton, J., LeGall, T., Vacic, V., Cortese, M., Tantos, A., Szabo, B., Tompa, P., Chen, J., Uversky, V., Obradovic, Z. et Dunker, A. (2007). Disprot : the database of disordered proteins. *Nucleic Acids Research*, 35(Database issue):D786–D793. Télécharger.

Sigrist, C., De Castro, E., Langendijk-Genevaux, P., Le Saux, V., Bairoch, A. et Hulo, N. (2005). ProRule : a new database containing functional and structural information on PROSITE profiles. *Bioinformatics*, 21(21):4060. Télécharger.

Simpson, A. et Roger, A. (2004). The real 'kingdoms' of eukaryotes. *Current Biology*, 14(17):R693–R696. Télécharger.

Singer, G. et Hickey, D. (2000). Nucleotide bias causes a genomewide bias in the amino acid composition of proteins. *Molecular Biology and Evolution*, 17(11):1581–1588. Télécharger.

Sippl, M. (1990). Calculation of conformational ensembles from potentials of mean force. An approach to the knowledge-based prediction of local structures in globular proteins. *Journal of Molecular Biology*, 213(4):859–883. Télécharger.

Sjölander, K., Karplus, K., Brown, M., Hughey, R., Krogh, A., Mian, I. et Haussler (1996). Dirichlet mixtures : a method for improved detection of weak but significant protein sequence homology. *Bioinformatics*, 12(4):327–345. Télécharger.

Sonnhammer, E., Eddy, S., Birney, E., Bateman, A. et Durbin, R. (1998). Pfam : multiple sequence alignments and hmm-profiles of protein domains. *Nucleic Acids Research*, 26(1):320–322. Télécharger.

Sonnhammer, E., Eddy, S. et Durbin, R. (1997). Pfam : a comprehensive database of protein domain families based on seed alignments. *Proteins*, 28(3):405–420. Télécharger.

Sonnhammer, E. et Koonin, E. (2002). Orthology, paralogy and proposed classification for paralog subtypes. *Trends in Genetics*, 18(12):619–620. Télécharger.

Soriç, B. (1989). Statistical 'discoveries' and effect-size estimation. *Journal of the American Statistical Association*, 84:608–610. Télécharger.

Sprinzak, E. et Margalit, H. (2001). Correlated sequence-signatures as markers of protein-protein interaction. *Journal of molecular biology*, 311(4):681–692. Télécharger.

Srivastava, P., Desai, D., Nandi, S. et Lynn, A. (2007). HMM-ModE – Improved classification using profile hidden Markov models by optimising the discrimination threshold and modifying emission probabilities with negative training sequences. *BMC Bioinformatics*, 8(1):104. Télécharger.

Stein, A., Russell, R. et Aloy, P. (2005). 3did : interacting protein domains of known three-dimensional structure. *Nucleic acids research*, 33(Database Issue):D413–D417. Télécharger.

Stein, L. (2001). Genome annotation : from sequence to biology. *Nature Reviews Genetics*, 2(7):493–503. Télécharger.

Stolcke, A. et Omohundro, S. (1994). Inducing probabilistic grammars by Bayesian model merging. *Lecture Notes in Computer Science*, 862:106–118. Télécharger.

Swarbreck, D., Wilks, C., Lamesch, P., Berardini, T., Garcia-Hernandez, M., Foerster, H., Li, D., Meyer, T., Muller, R., Ploetz, L., Radenbaugh, A., Singh, S., Swing, V., Tissier, C., Zhang, P. et Huala, E. (2008). The Arabidopsis Information Resource (TAIR) : gene structure and function annotation. *Nucleic Acids Research*, 36(Database issue):D1009–1014. Télécharger.

Takami, J. et Sagayama, S. (1992). A successive state splitting algorithm for efficient allophonemodeling. *IEEE International Conference on Acoustics, Speech, and Signal Processing, 1992. ICASSP-92.*, 1:573–576. Télécharger.

Tatusov, R., Fedorova, N., Jackson, J., Jacobs, A., Kiryutin, B., Koonin, E., Krylov, D., Mazumder, R., Mekhedov, S., Nikolskaya, A., Rao, B., Smirnov, S., Sverdlov, A., Vasudevan, S., Wolf, Y., Yin, J. et Natale, D. (2003). The cog database : an updated version includes eukaryotes. *BMC Bioinformatics*, 4:41. Télécharger.

Tatusov, R., Koonin, E. et Lipman, D. (1997). A genomic perspective on protein families. *Science*, 278(5338):631–637. Télécharger.

Taylor, W. (1986a). The classification of amino acid conservation. *Jounal of Theorical Biology*, 119(2):205–218.

Taylor, W. (1986b). Identification of protein sequence homology by consensus template alignment. *Journal of molecular biology*, 188(2):233–258. Télécharger.

Terrapon, N., Gascuel, O., Maréchal, E. et Bréhélin, L. (2009). Detection of new protein domains using co-occurrence : application to *Plasmodium falciparum*. *Bioinformatics*, 25(23):3077–3083. Télécharger.

The Arabidopsis Genome Initiative (2000). Analysis of the genome sequence of the flowering plant *Arabidopsis thaliana*. *Nature*, 408(6814):796–815. Télécharger.

The *C. elegans* Sequencing Consortium (1998). Genome sequence of the nematode *C. elegans* : a platform for investigating biology. *Science*, 282(5396):2012–2018. Télécharger.

Thomas, P., Campbell, M., Kejariwal, A., Mi, H., Karlak, B., Daverman, R., Diemer, K., Muruganujan, A. et Narechania, A. (2003). Panther : a library of protein families and subfamilies indexed by function. *Genome Research*, 13(9):2129–2141. Télécharger.

Thompson, J., Higgins, D. et Gibson, T. (1994). Clustal w : improving the sensitivity of progressive multiple sequence alignment through sequence weighting, position-specific gap penalties and weight matrix choice. *Nucleic Acids Research*, 22(22):4673–80. Télécharger.

Tompa, P. (2002). Intrinsically unstructured proteins. *Trends in Biochemical Sciences*, 27(10):527–533. Télécharger.

Uetz, P., Giot, L., Cagney, G., Mansfield, T., Judson, R., Knight, J., Lockshon, D., Narayan, V., Srinivasan, M., Pochart, P., Qureshi-Emili, A., Li, Y., Godwin, B., Conover, D., Kalbfleisch, T., Vijayadamodar, G., Yang, M., Johnston,

M., Fields, S. et Rothberg, J. (2000). A comprehensive analysis of protein–protein interactions in *Saccharomyces cerevisiae*. *Nature*, 403(6770):623–627. Télécharger.

Valkiūnas, G. (2004). *Avian malaria parasites and other haemosporidia*. CRC Press.

Voet, D. et Voet, J. (2004). *Biochemistry, Third Edition*. John Wiley and Sons Inc.

Vogel, C., Berzuini, C., Bashton, M., Gough, J. et Teichmann, S. (2004). Supra-domains : evolutionary units larger than single protein domains. *Curr. Opin. Struct. Biol.*, 14(2):208–216. Télécharger.

Vogel, C., Teichmann, S. et Pereira-Leal, J. (2005). The relationship between domain duplication and recombination. *Journal of Molecular Biology*, 346(1):355–365. Télécharger.

Waters, A., Higgins, D. et McCutchan, T. (1991). *Plasmodium falciparum* appears to have arisen as a result of lateral transfer between avian and human hosts. *Proceedings of the National Academy of Sciences of the U.S.A.*, 88(8):3140–3144. Télécharger.

Watson, J. et Crick, F. (1953). Molecular structure of nucleic acids ; a structure for deoxyribose nucleic acid. *Nature*, 171(4356):737–738. Télécharger.

Weiner, J., Beaussart, F. et Bornberg-Bauer, E. (2006). Domain deletions and substitutions in the modular protein evolution. *FEBS Journal*, 273(9):2037–2047. Télécharger.

Weiner, J. et Bornberg-Bauer, E. (2006). Evolution of circular permutations in multidomain proteins. *Molecular biology and evolution*, 23(4):734–743. Télécharger.

Weiner, J., Moore, A. et Bornberg-Bauer, E. (2008). Just how versatile are domains ? *BMC Evolutionary Biology*, 8(1):285. Télécharger.

Weiner, J., Thomas, G. et Bornberg-Bauer, E. (2005). Rapid motif-based prediction of circular permutations in multi-domain proteins. *Bioinformatics*, 21(7):932–937. Télécharger.

Whelan, S. et Goldman, N. (2001). A general empirical model of protein evolution derived from multiple protein families using a maximum-likelihood approach. *Molecular Biology and Evolution*, 18(5):691–699. Télécharger.

Wilson, D., Pethica, R., Zhou, Y., Talbot, C., Vogel, C., Madera, M., Chothia, C. et Gough, J. (2009). Superfamily – comparative genomics, datamining and sophisticated visualisation. *Nucleic Acids Research*, 37(Database issue):D380–386. Télécharger.

Winzeler, E. (2008). Malaria research in the post-genomic era. *Nature*, 455(7214):751–756. Télécharger.

Wistrand, M. et Sonnhammer, E. (2004). Improving profile HMM discrimination by adapting transition probabilities. *Journal of Molecular Biology*, 338(4):847–854. Télécharger.

Wistrand, M. et Sonnhammer, E. (2005). Improved profile HMM performance by assessment of critical algorithmic features in SAM and HMMER. *BMC bioinformatics*, 6:99. Télécharger.

Wojcik, J. et Schächter, V. (2001). Protein-protein interaction map inference using interacting domain profile pairs. *Bioinformatics*, 17(Suppl 1):S296–S305. Télécharger.

Wong, W., Maurer-Stroh, S. et Eisenhaber, F. (2010). More Than 1,001 Problems with Protein Domain Databases : Transmembrane Regions, Signal Peptides and the Issue of Sequence Homology. *PLoS Computational Biology*, 6(7). Télécharger.

Wootton, J. et Federhen, S. (1993). Statistics of local complexity in amino acid sequences and sequence databases. *Comput. & Chem.*, 17(2):149–163. Télécharger.

Wright, P. et Dyson, H. (1999). Intrinsically unstructured proteins : re-assessing the protein structure-function paradigm. *Journal of Molecular Biology*, 293(2):321–331. Télécharger.

Wu, C., Nikolskaya, A., Huang, H., Yeh, L., Natale, D., Vinayaka, C., Hu, Z., Mazumder, R., Kumar, S., Kourtesis, P., Ledley, R., Suzek, B., Arminski, L., Chen, Y., Zhang, J., Cardenas, J., Chung, S., Castro-Alvear, J., Dinkov, G. et Barker, W. (2004). Pirsf : family classification system at the protein information resource. *Nucleic Acids Research*, 32(Database issue):D112–114. Télécharger.

Wutchy, S. et Almaas, E. (2005). Evolutionary cores of domain co-occurence networks. *BMC Evol Biol*, 5(1):24. Télécharger.

Xue, X., Zhang, Q., Huang, Y., Feng, L. et Pan, W. (2008). No mirna were found in *Plasmodium* and the ones identified in erythrocytes could not be correlated with infection. *Malaria Journal*, 7(47). Télécharger.

Yang, Z. et Roberts, D. (1995). On the use of nucleic acid sequences to infer early branchings in the tree of life. *Molecular Biology and Evolution*, 12(3):451–458. Télécharger.

Ye, Y. et Godzik, A. (2004). Comparative analysis of protein domain organization. *Genome Research*, 14(3):343–353. Télécharger.

Yeats, C., Lees, J., Reid, A., Kellam, P., Martin, N., Liu, X. et Orengo, C. (2008). Gene3d : comprehensive structural and functional annotation of genomes. *Nucleic Acids Research*, 36(Database issue):D414–418. Télécharger.

Yona, G., Linial, N. et Linial, M. (2000). Protomap : automatic classification of protein sequences and hierarchy of protein families. *Nucleic Acids Research*, 28(1):49–55. Télécharger.

Yu, Y. et Altschul, S. (2004). The construction of amino acid substitution matrices for the comparison of proteins with non-standard compositions. *Bioinformatics*, 21(7):902–911. Télécharger.

Yu, Y., Wootton, J. et Altschul, S. (2003). The compositional adjustment of amino acid substitution matrices. *Proceedings of the National Academy of Sciences of the U.S.A.*, 100(26):15688–15693. Télécharger.

Zdobnov, E. et Apweiler, R. (2001). InterProScan – an integration platform for the signature-recognition methods in InterPro. Télécharger.

Zhang, Y. (2008). Progress and challenges in protein structure prediction. *Current opinion in structural biology*, 18(3):342–348. Télécharger.

Zhang, Y., Chandonia, J., Ding, C. et Holbrook, S. (2005). Comparative mapping of sequence-based and structure-based protein domains. *BMC bioinformatics*, 6:77. Télécharger.

Zvelebil, M., Barton, G., Taylor, W. et Sternberg, M. (1987). Prediction of protein secondary structure and active sites using the alignment of homologous sequences. *Journal of Molecular Biology*, 195(4):957–961. Télécharger.

Résumé : Les modèles de Markov cachés (MMC) – par exemple ceux de la librairie Pfam – sont des outils très populaires pour l'annotation des domaines protéiques. Cependant, ils ne sont pas toujours adaptés aux protéines les plus divergentes. C'est notamment le cas avec *Plasmodium falciparum* (principal agent du paludisme chez l'Homme), où les MMC de Pfam identifient peu de familles distinctes de domaines, et couvrent moins de 50% des protéines de l'organisme. L'objectif de cette thèse est d'apporter des méthodes nouvelles pour affiner la détection de domaines dans les protéines divergentes. Le premier axe développé est une approche d'identification de domaines utilisant leurs propriétés de co-occurrence. Différentes études ont montré que la majorité des domaines apparaissent dans les protéines avec un ensemble très réduits d'autres domaines favoris. Notre méthode exploite cette propriété pour détecter des domaines trop divergents pour être identifiés par l'approche classique. Cette détection s'accompagne d'une estimation du taux d'erreur par une procédure de ré-échantillonnage. Chez *P. falciparum*, elle permet d'identifier, avec un taux d'erreur estimé inférieur à 20%, 585 nouveaux domaines – dont 159 familles étaient inédites dans cet organisme –, ce qui représente 16% du nombre de domaines connus. Le second axe de mes recherches présente plusieurs méthodes de corrections statistiques et évolutives des MMC pour l'annotation d'organismes divergents. Deux types d'approches ont été proposées. D'un côté, nous intégrons aux alignements d'apprentissage des MMC les séquences précédemment identifiés dans l'organisme cible ou ses proches relatifs. La limitation de cette solution est que seules des familles de domaines déjà connues dans le taxon peuvent ainsi être identifiées. Le deuxième type d'approches contourne cette limitation en corrigeant tous les modèles par une prise en compte de l'évolution des séquences d'apprentissage. Pour cela, nous faisons appel à des techniques classiques de la bioinformatique et de l'apprentissage statistique. Les résultats obtenus offrent un ensemble de prédictions complémentaires totalisant 663 nouveaux domaines supplémentaires – dont 504 familles inédites –, soit une augmentation de 18% à ajouter aux précédents résultats.

Mots-clés : Domaines protéiques, modèles de Markov cachés, paludisme.

Summary : Hidden Markov Models (HMMs) – from Pfam database for example – are popular tools for protein domain annotation. However, they are not well suited for studying highly divergent proteins. This is notably the case with *Plasmodium falciparum* (main causal agent of human malaria), where Pfam HMMs identify few distinct domain families and cover less than 50% of its proteins. This thesis aims at providing new methods to enhance domain detection in divergent proteins. The first axis of this work is an approach of domain identification based on domain co-occurrence. Several studies shown that a majority of domains appear in proteins with a small set of other favourite domains. Our method exploits this tendency to detect domains escaping to the classical procedure because of their divergence. Detected domains come along with an false discovery rate (FDR) estimation computed with a shuffling procedure. In *P. falciparum* proteins, this approach allows us identify, with an FDR below 20%, 585 new domains – with 159 families that were previously unseen in this organism – which account for 16% of the known domains. The second axis of my researches involves the development of statistical and evolutionary methods of HMM correction to improve the annotation of divergent organisms. Two kind of approaches are proposed. On the one hand, the sequences previously identified in the target organism and its close relatives are integrated in the learning alignments. An obvious limitation of this solution is that only new occurrences of previously known families in the taxon can be discovered. On the other hand, we evade this limitation by adjusting HMM parameters by simulating the evolution of the learning sequences. To this end, classical techniques from bioinformatics and statistical learning were used. Alternative libraries offer a complementary set of predictions summing 663 new domains – with 504 previously unseen families – corresponding to an improvement of 18% to add to the previous results.

Keywords : Protein domains, hidden Markov models, malaria.