Académie de Montpellier
Université Montpellier II
— Sciences et Techniques du Languedoc —

Thèse

présentée au Laboratoire d'Informatique de Robotique
et de Microélectronique de Montpellier

Spécialité	:	**Informatique**
Formation Doctorale	:	**Informatique**
École Doctorale	:	**Information, Structures, Systèmes**

Méthodes pour l'identification de domaines protéiques divergents
Soutenue le 16 Décembre 2013

par

Amel Ghouila

Jury composé de:

Andrey Kajava	Rapporteur	Directeur de Recherche	CRBM, Université de Montpellier 2, Montpellier
Engelbert Mephu Nguifo	Rapporteur	Professeur	LIMOS, Université Blaise Pascal, Clermont-Ferrand
Mourad Elloumi	Rapporteur	Professeur	FESMT, Université de Tunis El Manar, Tunis
Isabelle Florent	Examinateur	Professeur	Museum d'histoire Naturelle, Paris
Yahia Slimani,	Examinateur	Professeur	Université de la Manouba, Tunis
Laurent Bréhélin	Directeur	Chargé de Recherche	LIRMM, Montpellier
Olivier Gascuel	Directeur	Directeur de Recherche	LIRMM, Montpellier
Sadok Ben Yahia	Directeur	Maître de conférences	FST, Tunis

Table des matières

Résumé 1

I. Introduction 2

Introduction 4

II. État de l'art 7

1. Du gène à l'annotation fonctionnelle 9
1.1. Du gène à la protéine . . . 10
1.2. Les protéines . . . 12
1.2.1. Étymologie . . . 12
1.2.2. Niveaux de structure d'une protéine . . . 12
1.2.3. Les protéines et leurs fonctions . . . 17
1.2.4. Annotation fonctionnelle des protéines . . . 17
1.3. Les domaines protéiques . . . 18
1.3.1. Le domaine une unité d'évolution à part entière . . . 19
1.3.2. Combinaisons de domaines . . . 19
1.3.3. Interactions des domaines . . . 20
1.3.4. Du domaine à la fonction . . . 20
1.4. Modélisation des familles de domaines protéiques . . . 20
1.4.1. Familles et banques de domaines . . . 22
1.4.2. La méta-base Interpro . . . 25

2. Modélisation des familles de domaines à l'aide de modèles de Markov cachés 27
2.1. Les modèles de Markov cachés . . . 28
2.1.1. Présentation . . . 28
2.1.2. Domaines d'applications des HMM . . . 28
2.1.3. Probabilité de génération d'une séquence par un HMM . . . 29
2.1.4. Apprentissage des HMM . . . 32
2.2. Les HMM profils . . . 33
2.2.1. Définition . . . 33
2.2.2. HMMER et SAM . . . 34
2.2.3. Construction d'un HMM profil . . . 35
2.2.4. Recherche de domaines . . . 36

2.3. La banque de domaines Pfam 38
2.3.1. Pfam-A 39
2.3.2. Pfam-B 40

3. Pathogènes Eukaryotes : présentation et problèmes d'annotation 41
3.1. Deux pathogènes emblématiques : *Leishmania* et *Plasmodium* 42
3.1.1. Leishmanioses et *Leishmania* 42
3.1.2. Le parasite *Leishmania* 45
3.1.3. Paludisme et *Plasmodium* 48
3.1.4. Le parasite *Plasmodium* 50
3.2. Annotation fonctionnelle des pathogènes Eukaryotes 51
3.2.1. Annotation fonctionnelle de *Leishmania* et *Plasmodium* 51
3.2.2. Détection des domaines sur les protéines divergentes 52

III. Contributions 55

4. EuPathDomains : la base de données des domaines divergents chez les pathogènes Eukaryotes 57
4.1. Introduction 58
4.2. CODD : la stratégie de détection de domaines par Co-occurrence . . . 58
4.2.1. Construction de la CDP 60
4.2.2. Identification des domaines connus et recherche des domaines potentiels 60
4.2.3. Processus de certification par co-occurrence 61
4.2.4. Estimation du taux d'erreur 62
4.2.5. Inférences des annotations Gene Ontology 62
4.3. Pathogènes étudiés 64
4.4. La base de données EupathDomains 65
4.4.1. Mise en place 65
4.4.2. Interface et contenu 65
4.4.3. Étude des caractéristiques des nouvelles prédictions 69
4.4.4. *Leishmania* comme étude de cas 72
4.4.5. Les prédictions communes à tous les pathogènes étudiés 73
4.5. Nouvelle version 2.0 de EupathDomains 74
4.6. Conclusion 76

5. Stratégies d'amélioration de la prédiction par co-occurence 77
5.1. Traitement des chevauchements de domaines potentiels 78
5.1.1. Recherche du meilleur chemin sur le graphe de domaines 78
5.1.2. Études expérimentales 79
5.2. Utilisation du contexte négatif pour l'amélioration de la qualité des prédictions 81
5.2.1. Certification par les domaines connus 82

5.2.2. Certification par les domaines potentiels 84
5.2.3. Étude expérimentale 84
5.3. Études des combinaisons de domaines par les règles d'association . . . 88
5.3.1. Les règles d'association en bref 88
5.3.2. Extraction de règles fréquentes : Limites de la métrique *minsup* sur les bases de données biologiques 89
5.3.3. Extraction des associations entre domaines 90
5.3.4. Certification par les règles d'association 90
5.4. Utilisation des espèces proches pour l'amélioration des reconnaissances de domaines 92

6. Une nouvelle approche basée sur les comparaisons profil-profil 94
6.1. Introduction 95
6.2. HHPRED à l'échelle d'un génome 95
6.3. Les méthodes profil-profil manquent de sensibilité 97
6.4. Combiner les comparaisons HMM-HMM et la détection par co-occurrence 99
6.5. Amélioration de la politique d'inférence des annotations GO 100
6.6. Analyse des nouveaux domaines prédits 103
6.6.1. Domaines découverts chez *P. falciparum* 106
6.6.2. Domaines découverts chez *L. major* 107

IV. Conclusion 108

Conclusion 110

List of Figures 113

List of Tables 115

Bibliography 116

Résumé

Résumé L'étude de la composition des protéines en domaines est une étape clé pour la détermination de ses fonctions. Pfam est l'une des banques de domaines les plus répandues où chaque domaine est représenté par un HMM profil construit à partir d'un alignement multiple de protéines contenant le domaine. La méthode classique de recherche des domaines Pfam consiste à comparer la séquence cible à la librairie complète des HMM profils pour mesurer sa ressemblance aux différents modèles. Cependant, appliquée aux protéines d'organismes divergents, cette méthode manque de sensibilité. L'objectif de cette thèse est d'apporter de nouvelles méthodes pour améliorer le processus de prédictions des domaines plus adaptées à l'étude des protéines divergentes. Les premiers travaux ont consisté en l'adaptation et application de la méthode CODD, récemment proposée, à l'ensemble des pathogènes de la base de données EuPathDB. Une base de données nommée EupathDomains recensant l'ensemble des domaines connus et ceux nouvellement prédits chez ces pathogènes a été mise en place à l'issue de ces travaux. Nous proposons ensuite diverses améliorations de la méthode CODD. Nous proposons un algorithme "CODD_exclusive" qui utilise des informations d'incompatibilité de domaines pour améliorer la précision des prédictions. Nous proposons également une autre stratégie basée sur l'utilisation de règles d'association pour la détermination des co-occurrences de domaines utilisées dans le processus de certification. La dernière partie de cette thèse s'intéresse à l'utilisation des méthodes profil/profil pour annoter un génome entier. Couplée à la procédure d'annotation par co-occurrence, cette approche permet une amélioration notable en termes de nombre de domaines certifiés et également en termes de précision.

Abstract The determination of protein domain composition provides strong clues for the protein function prediction. One of the most widely used domain scheme is the Pfam database in which each family is represented by a multiple sequence alignment and a profile Hidden Markov Model (profile HMM). When analyzing a new sequence, each Pfam HMM is used to compute a score measuring the similarity between the sequence and the domain. However, applied to divergent proteins, this strategy may miss several domains. This is the case for all eukaryotic pathogens, where no Pfam domains are detected in half or even more of their proteins. The main objective of this thesis is to develop methods to improve the sensitivity of Pfam domain detection in divergent proteins. We first adapted the recently proposed CODD method to the whole set of pathogens in EupathDB and buit EupathDomains database that gathers known and new detected domains along with the associated GO annotations.

We then proposed other methods to further improve domain detection in these organisms. We proposed "CODD_exclusive" algorithm that integrates domain exclusion information to prune false positive domains that are in conflict with other domains of the protein. We also suggested the use of association rules to determine the correlations between domains and used these informations in the certification process. In the last part of this thesis, we focused in the use of profile/profile methods to predict protein domains in a whole genome. Combined with the co-occurrence informations, it achieved high sensitivity and accuracy in predicting domains.

Partie I.

Introduction

Introduction

Ces dernières années ont été marquées par une forte augmentation du nombre de génomes séquencés du fait de l'évolution des techniques de séquençage. Le nombre d'espèces, appartenant aux différents règnes du vivant, entièrement séquencées a atteint 3699 à la fin de l'année 2012 contre moins d'une dizaine au début des années 2000 (Pagani et al., 2012). La figure 0.1 illustre cette évolution. Ces données ont considérablement amélioré nos connaissances sur la diversité

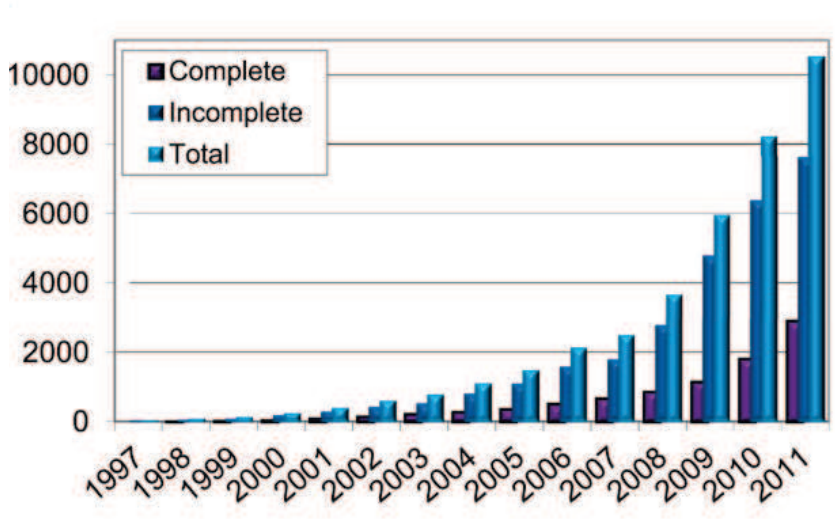


FIGURE 0.1.: Évolution du nombre des projets génomes (en cours de séquençage ou complètement séquencées) contenus dans la base de données Genomes On Line (GOLD) (Pagani et al., 2012).

des organismes et présentent une source d'informations importante pour la compréhension de leur biologie. Cependant, après l'étape de séquençage, un travail important reste à réaliser pour transformer la séquence génomique brute en connaissances biologiques utiles. Un des premiers travaux consiste à localiser l'ensemble des séquences codantes (gènes) du génome. Une fois cela réalisé, il faut caractériser la fonction des protéines et des ARN pour lesquels codent ces gènes. Ce travail d'annotation constitue un des défis majeurs de la bioinformatique et est une étape primordiale pour comprendre la biologie d'un organisme. Dans le cadre des organismes pathogènes par exemple, c'est une étape clef pour la recherche de nouvelles cibles thérapeutiques et le développement de nouveaux médicaments. Les domaines protéiques constituent une des principales annotations qui peuvent être associées à une protéine. Ces domaines constituent les sous-unités structurales et fonctionnelles des protéines et sont donc considérés comme une unité pertinente d'évolution (Yang and Bourne, 2009; Zhang et al., 2012). Ils se combinent de différentes manières pour donner lieu à un répertoire de protéines varié. La détermination des domaines d'une protéine est une étape clé pour la détermination de ses fonctions (Fors-

lund and Sonnhammer, 2008). Pfam est l'une des banques de domaines les plus utilisées et les plus riches et couvre une grande proportion des protéines de la base Swiss-Prot (Bateman et al., 2000). Chaque domaine dans Pfam est représenté par un modèle probabiliste particulier appelé Modèle de Markov Caché (Hidden Markov Model en anglais ou HMM), construit à partir d'un alignement multiple de séquences protéiques contenant le domaine. Lorsque l'on cherche à déterminer les domaines Pfam d'une séquence protéique cible, la méthode classique consiste à comparer cette séquence à la librairie complète des HMM de Pfam afin de mesurer sa ressemblance aux différents modèles. Pour chaque comparaison, un score reflétant le degré de similarité de la séquence au modèle est calculé. Ce score est comparé à un seuil de référence fourni par Pfam, au dessus duquel la présence du domaine dans la protéine peut être considérée avec un risque minimal d'erreur. Cette approche est d'autant plus efficace que le nouveau génome est phylogénétiquement proche des espèces modèles classiques ayant servi à la construction des différents HMM. Dans le cas inverse, la sensibilité de l'approche peut être faible (Ghouila et al., 2010). C'est le cas par exemple pour les pathogènes eukaryotes, qui sont dans leur grande majorité phylogénétiquement très éloignés des espèces modèles classiques. Alors que la proportion de protéines ayant au moins un domaine Pfam connu est souvent bien au-dessus de 70% chez les Unikonts (groupe regroupant les levures et les animaux) et les plantes, il est compris entre 50% et 60% chez les pathogènes eukaryotes. Pour remédier à ces limites, une méthode nommée CODD (Terrapon et al., 2009) a été proposée récemment dans le laboratoire. Cette méthode procède en deux étapes. CODD relâche dans un premier temps les seuils de scores stringents proposés par Pfam pour augmenter la sensibilité de détection des HMM. Dans un second temps, il utilise les informations de co-occurences de domaines afin de supprimer les faux positifs parmi les nouveaux domaines découverts. La procédure est basée sur une liste de paires de domaines dont la co-occurrence dans les protéines a été attestée par un grand nombre de fois et sur différents organismes.

L'objectif de cette thèse est d'améliorer les méthodes de prédiction de domaines pour l'annotation fonctionnelle des pathogènes eukaryotes. Deux organismes ont été plus particulièrement étudiés dans ces travaux : les espèces *Plasmodium* et *Leishmania*, agents causant respectivement la malaria et la leishmaniose. Les premiers travaux de cette thèse ont consisté à adapter et appliquer la méthode CODD à l'ensemble des pathogènes de la base de données EuPathDB. Dans ce travail, une attention particulière a été portée aux annotations fonctionnelles qui peuvent être déduites des nouveaux domaines prédits. Une base de données nommée EuPathDomains (http ://www.atgc-montpellier.fr/EuPathDomains/) recensant les domaines connus et ceux nouvellement prédits chez ces pathogènes, ainsi que les annotations fonctionnelles qui peuvent en être déduites a été mise en place à l'issue de ces travaux. EuPathDomains a permis d'améliorer le taux de couverture en domaines Pfam de 17% à 41% suivant les organismes. Ces travaux ont fait l'objet de plusieurs présentations dans des conférences de bioinformatique (ISCB Afica ASBCB Joint conference on Bioinformatics of Infectious Diseases 2009, 15th International Congress on Infectious Diseases 2012) et ont été publiés dans la revue *Infection, Genetics and Evolution*.

Nous nous sommes ensuite attachés à proposer diverses améliorations de la procédure d'annotation générale par les HMM, et de la procédure CODD en particulier. :

- Recherche d'une méthode exacte pour résoudre les conflits en cas de chevauchements de plusieurs domaines.
- Prise en compte du contexte négatif dans la procédure CODD. Dans la procédure originale on ne se sert que du contexte positif d'apparition des domaines. Nous avons donc proposé une stratégie pour intégrer le contexte négatif à la procédure. Cela nous permet de réduire de manière significative le nombre de faux positifs parmi les nouveaux domaines certifiés.

Ce travail a été présenté au workshop *BIOKDD'12*, et a été publié dans les actes de la conférence.

- Intégration de règles d'association en lieu et place de la liste des paires de domaines co-occurrents utilisée dans la procédure CODD originale. Comme pour le contexte négatif, cette approche permet de réduire de manière significative le taux d'erreur associé aux prédictions.
- Utilisation des séquences protéiques homologues d'autres espèces afin d'améliorer la sensibilité des HMM profils. Il s'agit d'une extension de l'approche classique dans laquelle on ne compare plus une séquence à un HMM mais deux HMM entre eux. Couplée à la procédure d'annotation par co-occurrence, cette approche permet une amélioration notable en termes de nombre de domaines certifiés mais aussi en termes de précision. Ces travaux ont fait l'objet d'un article en cours de révision.

Ce manuscrit de thèse comporte deux parties. La première partie dresse un état de l'art du domaine, tandis que la deuxième présente l'ensemble de nos contributions :

1. Première partie :
   - Le premier chapitre aborde les notions de biologie nécessaires pour appréhender les données manipulées au cours de cette thèse ainsi que les différentes bases de données de domaines protéiques.
   - Le second chapitre présente les HMM et les HMM profils dédiés au traitement des séquences biologiques. Il présente également la procédure classique de recherches de domaines dans ce formalisme.
   - Le troisième chapitre présente les deux espèces d'intérêt *Leishmania* et *Plasmodium* et discute les limites des méthodes classiques de détection de domaines protéiques.
2. Deuxième partie :
   - Le quatrième chapitre présente la base de données EuPathDomains recensant l'ensemble des prédictions obtenues sur l'ensemble des pathogènes Eukaryotes de la base EuPathDB.
   - Le cinquième chapitre présente les différentes stratégies que nous avons proposées afin d'améliorer la procédure d'annotation générale par HMM, et la procédure CODD en particulier.
   - Le sixième et dernier chapitre présente la nouvelle approche basée sur les comparaisons entre HMM combinées avec l'utilisation des informations de co-occurrences de domaines.

Partie II.

État de l'art

1. Du gène à l'annotation fonctionnelle

Il s'agit dans ce premier chapitre de situer le contexte scientifique de ce manuscrit et de présenter les notions de base nécessaires à la compréhension de ces travaux de thèse. Nous présentons dans une première partie les éléments clés de la biologie moléculaire. Nous discutons ensuite le problème de l'annotation fonctionnelle et les techniques bioinformatiques liées aux problèmes d'annotation. Nous présentons dans la dernière partie les domaines protéiques et quelques caractéristiques qui leur sont particulières.

Sommaire

1.1. Du gène à la protéine 10
1.2. Les protéines 12
1.2.1. Étymologie 12
1.2.2. Niveaux de structure d'une protéine 12
1.2.3. Les protéines et leurs fonctions 17
1.2.4. Annotation fonctionnelle des protéines 17
1.3. Les domaines protéiques 18
1.3.1. Le domaine une unité d'évolution à part entière 19
1.3.2. Combinaisons de domaines 19
1.3.3. Interactions des domaines 20
1.3.4. Du domaine à la fonction 20
1.4. Modélisation des familles de domaines protéiques 20
1.4.1. Familles et banques de domaines 22
1.4.2. La méta-base Interpro 25

1.1. Du gène à la protéine

Les organismes des différentes espèces sont composés d'une ou plusieurs cellules. Les organismes bactériens, par exemple, sont unicellulaires tandis que les êtres humains sont multicellulaires, leur organisme est composé de plusieurs milliards de cellules. Chacune de ces cellules possède un ou plusieurs chromosomes. Ces derniers sont transmis des cellules mères aux cellules filles lors des divisions cellulaires. Un chromosome est composé d'une longue molécule d'ADN (Acide DesoxyriboNucléique). En 1953, Watson et Crick avaient découvert la structure en double brins de cette grande molécule d'ADN. Le nombre de chromosomes est l'une des caractéristiques de chaque espèce, l'humain possède par exemple 46 chromosomes (23 paires).

L'ADN : La molécule d'ADN est porteuse du patrimoine génétique. L'ADN est une suite de paires de nucléotides qui sont des molécules chimiques ayant la particularité de s'unir deux à deux par complémentarité. Quatre types de nucléotides sont utilisées pour le codage des génomes : l'adénine (A), la cytosine (C), la guanine (G) et la thymine (T). La taille d'un génome est souvent mesurée en nombre de paires de bases nucléotidiques. Ces tailles varient de quelques kilobases chez les virus à des centaines de milliers de mégabases chez les eukaryotes. Certaines régions d'un génome sont dites codantes : elles correspondent aux gènes, qui portent les instructions de la production d'une protéine.

En 1958, une formulation des relations entre ADN, ARN et protéine, couramment connue sous le nom du "dogme central de la biologie moléculaire", a été introduite par Francis Crick (Crick, 1970). Ce processus est résumé dans la figure 1.1.

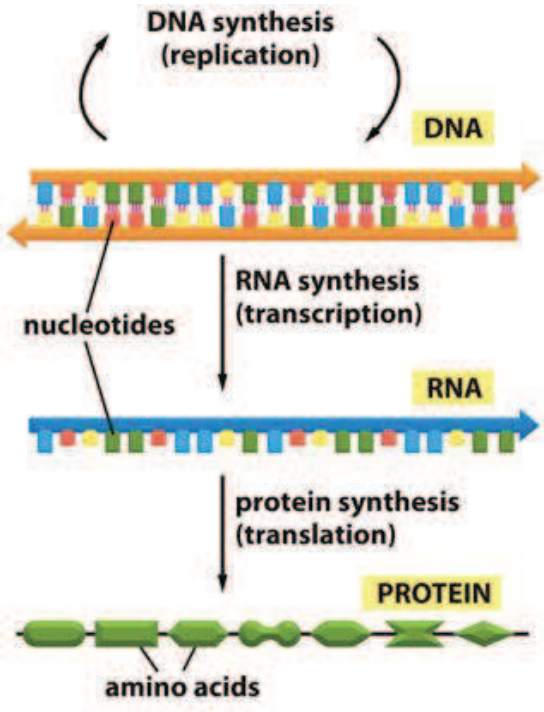
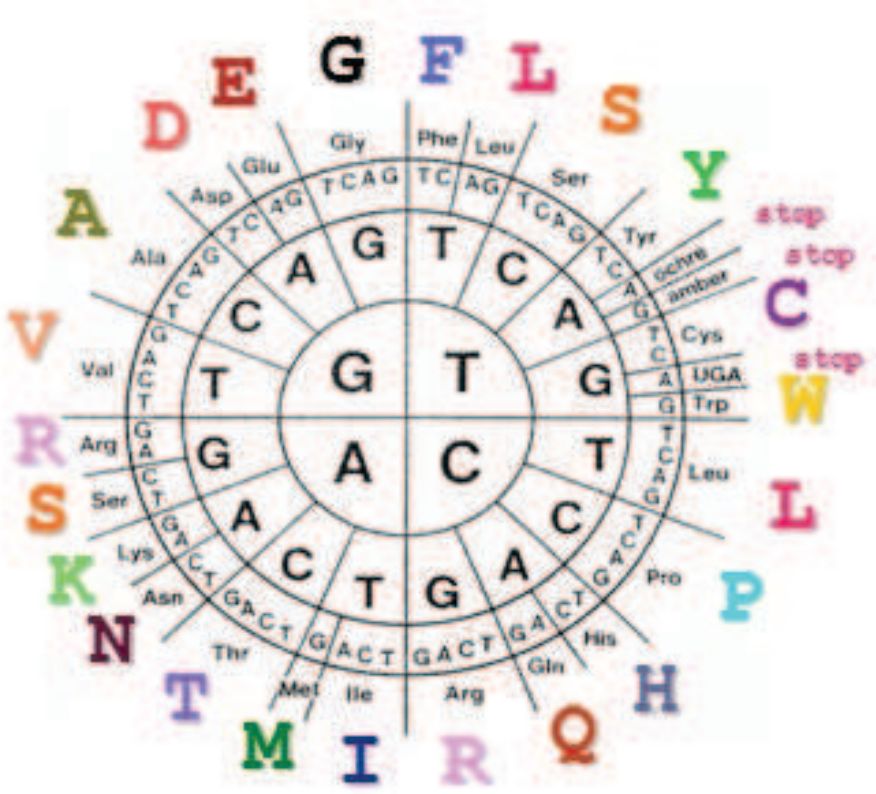


FIGURE 1.1.: Le dogme central de la biologie moléculaire : le processus de la synthèse des protéines à partir de l'ADN introduit par Francis Crick en 1958


G
F
L
S
Y
C
W
L
P
H
Q
R
I
M
T
N
K
S
R
V
A
D
E
Gly
Phe
Leu
Ser
Tyr
Cys
Trp
Leu
Pro
His
Gln
Arg
Ile
Met
Thr
Asn
Lys
Ser
Arg
Val
Ala
Asp
Glu
stop
stop
stop
ochre
amber
UGA
T C A G
T C A G
T C A G

Modifications co-traductionnelles et post-traductionnelles des protéines : Certaines protéines peuvent subir diverses modifications au cours de leur biosynthèse : il s'agit de modifications dites co-traductionnelles pouvant modifier un ou plusieurs acides aminés. Après leur synthèse, les protéines peuvent aussi être sujettes à des modifications chimiques souvent réalisées par des enzymes. Ces modifications, dites post-traductionnelles, peuvent altérer les fonctions accomplies par les protéines ou modifier leur localisation cellulaire (Kia-Ki and Martinage, 1992).

1.2. Les protéines

1.2.1. Étymologie

D'un point de vue étymologique, le terme protéine vient du grec ancien : "prôtos" qui signifie premier, essentiel. Cette appellation est probablement liée au fait que les protéines sont indispensables à la vie et qu'elles constituent souvent la part majoritaire (plus de 50%) du poids sec des cellules. D'autres théories font le lien entre le mot protéine et protéiforme qui fait référence au dieu grec "Protée" dont la caractéristique essentielle était de pouvoir apparaître sous des formes très diverses (forme animale ou encore eau ou feu). Les protéines peuvent en effet avoir plusieurs formes ou structures et assurent de multiples fonctions, mais ceci ne fut découvert qu'au cours du 20ème siècle. Les protéines furent découvertes en 1835 par le chimiste Gerardus Johannes Mulder aux Pays-Bas sous le nom de *Wortelstof*. Le nom protéine fut suggéré trois ans plus tard par le suédois Jöns Jacob Berzelius. À la fin des années 90, et conjointement à la naissance de la génomique, les termes "protéome" et "protéomique" d'origine anglosaxone virent le jour. Le protéome est pour les protéines ce qu'est le génome pour les gènes. Il désigne l'ensemble des protéines exprimées dans une cellule, ou un groupe de cellules dans des conditions données et à un moment donné. À la différence du génome, qui est en général stable dans les différentes cellules d'un organisme, le protéome est variable dans le temps et en fonction de sa localisation cellulaire.

1.2.2. Niveaux de structure d'une protéine

Les protéines peuvent être considérées selon différents niveaux de structure représentés dans la figure 1.3. Nous distinguons communément trois niveaux de structure auxquels s'ajoute un quatrième dans certains cas. Toutes les protéines peuvent être considérées du point de vue de leur structure primaire définie par la chaine d'acides aminés qui la composent. Les acides aminés composant les protéines sont au nombre de 22. La sélénocystéine (U) et la pyrrolysine (O), deux de ces acides aminés codés par des codons STOP, sont beaucoup moins fréquents que les autres. De plus, la pyrrolysine n'apparait que chez certains groupes du vivant. La liste des 20 acides aminés est donnée par la table 1.1.

Dans l'espace, la séquence d'acides aminés n'est pas une simple chaîne linéaire dépourvue de structure. Les liaisons hydrogènes entre les acides aminés tendent à former des repliements. Ainsi, la protéine peut aussi être décrite par un enchaînement des repliements locaux de la chaîne d'acides aminés. C'est ce qu'on appelle la structure secondaire. Malgré la grande diversité possible des enchaînements d'acides aminés, leur structure secondaire n'est formée que par deux principales formes : les hélices α, formés par des spiralisations locales, et la forme de nappe plane dite feuillet β. D'autres structures plus difficilement reconnaissables et moins régulières que les deux premières existent : les boucles et les coudes. La prédiction de la structure secondaire à partir d'une séquence d'acides aminés a fait l'objet de nombreuses études et de nombreux algorithmes ont ainsi vu le jour et sont aujourd'hui utilisés. Ces derniers sont

Acide aminé	Abréviation	code
Alanine	Ala	A
Cystéine	Cys	C
Acide aspartique	Asp	D
Acide glutamique	Glu	E
Phénylalanine	Phe	F
Glycine	Gly	G
Histidine	His	H
Isoleucine	Ile	I
Lysine	Lys	K
Leucine	Leu	L
Méthionine	Met	M
Asparagine	Asn	N
Proline	Pro	P
Glutamine	Gln	Q
Arginine	Arg	R
Sérine	Ser	S
Thréonine	Thr	T
Valine	Val	V
Tryptophane	Trp	W
Tyrosine	Tyr	Y

TABLE 1.1.: Les 20 acides aminés du vivant : noms, trigrammes et codes standards

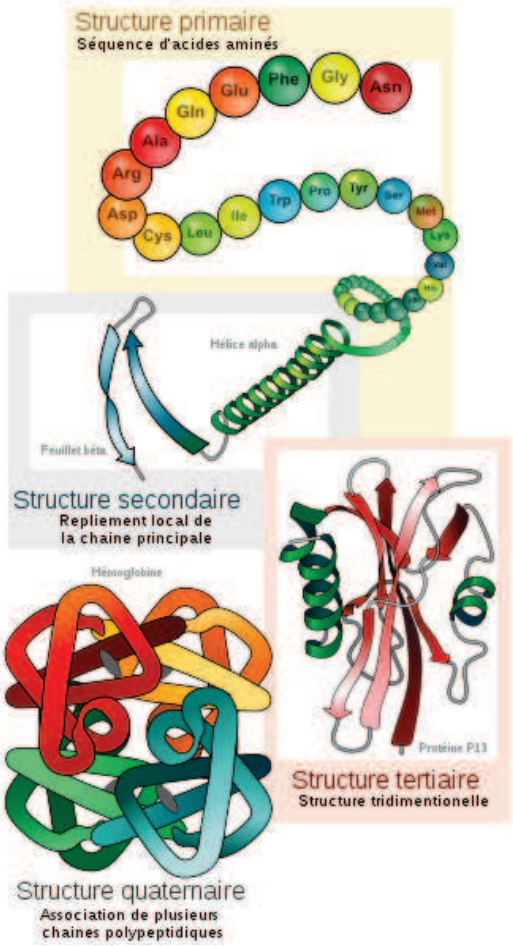


FIGURE 1.3.: Les différents niveaux de structure d'une protéine : de la structure primaire à la structure quaternaire

basés sur différentes techniques issues de l'apprentissage tels que les réseaux de neurones, les modèles de Markov cachés ou les machines à vecteurs de support (Lin et al., 2005; Pirovano and Heringa, 2010). Des méthodes plus récentes sont basées sur l'utilisation de l'information évolutive (Rost, 2001; Pirovano and Heringa, 2010). De plus, les méthodes récemment développées fournissent des scores de confiance associés à chacune des positions des prédictions.

La réunion des différents éléments de la structure secondaire dans l'espace conduit à la formation de la structure tertiaire des protéines couramment appelée structure 3D ou structure tridimensionnelle. Cette structure est constituée de plusieurs zones "actives" qui permettent de réaliser effectivement les différentes fonctions, comme les sites actifs de fixation et de reconnaissance du substrat chez les enzymes par exemple. La structure tridimensionnelle d'une protéine est intimement liée à sa structure primaire. Ainsi, les protéines ayant des séquences très similaires partageront également de fortes similarités de structures 3D. Plusieurs méthodes expérimentales telles que la cristallographie par rayons X ou encore la spectroscopie par résonance magnétique nucléaire permettent de connaître la structure tridimensionnelle des protéines. Cependant, le processus de détermination de la structure d'une protéine reste

un processus lent, complexe et coûteux. Grâce au développement de la bioinformatique, de nombreuses méthodes efficaces de prédiction automatique, plus rapide et beaucoup moins coûteuses, ont été développées.

Il existe des protéines complexes formées de plusieurs chaînes polypeptidiques appelées protéines multimériques ou encore complexes protéiques. La structure quaternaire constitue dans ce cas l'agencement des différentes sous unités (présentant chacune une structure tertiaire) les unes par rapport aux autres. Un exemple bien connu de protéine multimérique de structure quaternaire est l'hémoglobine humaine. Cette dernière est composée de quatre chaînes identiques deux à deux : deux chaînes α de 141 acides aminés chacune et de deux chaînes β de 146 acides aminés chacune. Sa principale fonction est le transport du dioxygène dans l'organisme.

Banques de données structurales

PDB Une des banques de données structurales la plus utilisée est la Protein Data Bank (PDB) (Berman et al., 2000) qui répertorie les structures 3D ainsi que les séquences d'acides aminés des protéines. Les structures 3D contenues dans cette banque sont généralement déterminées d'une manière expérimentale par cristallographie aux rayons X ou par spectroscopie. En Juin 2013, la PDB regroupait 91.359 structures. Le site web de la PDB offre la possibilité d'utiliser plusieurs outils d'analyse de structures. En 2003, la World Wide Protein Data Bank (WWPDB) [1] (Berman et al., 2003), une ressource mondiale dont le but est de regrouper et archiver les données en provenance de diverses PDB : RCSB, PDB et BMRB aux États unis, PDBe en Europe, PDBj au Japon a vu le jour. SCOP (Andreeva et al., 2004) et CATH sont deux autres banques de données de structures. Elles peuvent être considérées comme des extensions de la PDB du fait qu'elles contiennent une classification des protéines de cette dernière.

SCOP Créée en 1994, SCOP [2] (Structural Classification of Proteins) (Murzin et al., 1995; Andreeva et al., 2004) est une base de données de classification des protéines basée sur la similarité des structures et des séquences d'acides aminés. Un des principaux objectifs de cette base est la détermination des relations évolutives entre les différentes protéines. Les protéines ayant une forte similarité au niveau de leurs structures et partageant également une forte similarité de séquences sont placées dans la même famille et sont considérées comme ayant un proche ancêtre commun. La classification dans SCOP est basée sur les structures issues de la PDB. La classification est d'abord générée automatiquement par des comparaisons automatiques des séquences les plus proches. Cette classification est vérifiée manuellement à l'aide d'outils de visualisation, notamment dans le cas des séquences les plus distantes. On distingue différents niveaux de classification dans SCOP :

- La classes, qui désigne les différents types de repliements : tout alpha, tout beta, alpha et beta, alpha plus beta, protéines multi-domaines, protéines membranaires, petites protéines, "coiled coil proteins", structures de protéines de faibles résolution, peptides et protéines "artificiels".
- Le repliement ("fold"), qui désigne les différentes formes des domaines au sein d'une même classe. Il regroupe les protéines partageant les mêmes structures secondaires majeures.
- La super-famille, qui réunit les protéines ayant une similarité au niveau de la structure mais des similarités plus faibles au niveau séquences. Ces protéines possèdent au moins un ancêtre structural commun distant.

1. http ://www.wwpdb.org/
2. http ://scop.mrc-lmb.cam.ac.uk/scop/index.html

- La famille qui contient les éléments partageant une forte similarité de séquences primaires (la similarité est généralement supérieure à 30%). Les éléments au sein d'une super-famille sont regroupés en familles.

Des arbres phylogénétiques sont associés à chaque superfamille. La construction de ces arbres est basée sur les alignements multiples ainsi que sur les distances structurelles.

CATH Mise en place en 1997, la base de données CATH [3] (Class, Architecture, Topology, Homology) (Orengo et al., 1997; Cuff et al., 2009) regroupe une classification hiérarchique de domaines protéiques. On distingue quatre principaux niveaux de classification dans CATH à savoir (Cuff et al., 2009) :

- La classe : à ce niveau (C-level), les domaines sont regroupés sur la base de la similarité de leurs structures secondaires. On distingue quatre classes : la première concerne les domaines formés essentiellement par des hélices α, la seconde est celle contenant ceux formés essentiellement par des feuillets β. La troisième classe regroupe les éléments formés par α et β. La dernière est celle des domaines ayant très peu de structures secondaires.
- L'architecture (A-level) : la classification à ce niveau est basée sur la forme globale de la structure déterminée par l'orientation spatiale des structures secondaires en ignorant leurs interconnexions.
- La topologie (T-level) : chaque classe du A-level est divisée en un ou plusieurs groupes de repliement. À ce niveau, les connexions entre les différentes structures secondaires sont prises en considération.
- L'homologie (H-level) : les domaines sont classés en super-familles d'homologues selon leur similarité de séquences, fonctions et structures.

La classification dans CATH est basée dans une première étape sur l'identification des domaines protéiques. Ces classifications sont semi-automatiques : des méthodes automatiques génèrent les classifications qui sont ensuite vérifiées manuellement. Quelques classifications sont assignées manuellement en se basant sur la littérature ou des outils de prédiction. CATH est dotée d'une interface Web permettant de faire des parcours et/ou des recherches en utilisant les codes de la PDB.

Gene3D La base de données Gene3D [4] est une collection de domaines CATH et leurs assignations aux protéines Uniprot, RefSeq et à plusieurs génomes dans ENSEMBL (Buchan et al., 2002; Lees et al., 2010). Plusieurs autres informations fonctionnelles accompagnent les assignations des domaines telles que les termes de la Gene Ontology, les voies métaboliques (KEGG) ou encore des liens vers des données de transcriptome. D'autres types d'informations telles que les informations structurales (polymorphismes, régions non structurées) et des liens vers les différentes ressources de domaines sont aussi intégrées. Gene3D offre également un outil permettant la visualisation graphique interactive de la couverture en domaines des génomes (Lees et al., 2010). Le point de départ du processus d'annotation dans Gene3D est l'ensemble des HMM représentatifs des familles CATH. Ces derniers sont scannés contre les séquences en provenance des différentes banques de données. Ainsi, de nouvelles annotations structurales sont transférées à un grand nombre de séquences. La procédure d'annotation dans Gene3D est similaire à celle de la base de données SCOP basée sur les domaines de SUPERFAMILY. L'algorithme DomainFinder3 [5] (Yeats et al., 2010) est finalement utilisé pour filtrer les annotations qui sont en conflit.

3. http ://www.biochem.ucl.ac.uk/bsm/cath
4. http ://gene3d.biochem.ucl.ac.uk/Gene3D/
5. ftp ://ftp.biochem.ucl.ac.uk/pub/$gene3d_data$/DomainFinder3/

Protéines non structurées

La fonction de la protéine est altérée lorsque sa structure est cassée par l'utilisation d'un agent dénaturant ou dans certaines conditions particulières (pH, température élevée, *etc.*) (SELA et al., 1957). Cependant, certaines protéines sont dépourvues de structure tertiaire stable sauf quand elles interagissent avec d'autres facteurs. Ces protéines sont dites "non structurées". Elles jouent des rôles parfois très important dans les processus cellulaires. On estime que leur proportion dans les protéines Eukaryotes est entre 25% et 30% (Uversky et al., 2007). La base de données Disprot [6] (Database of Protein Disorder) répertorie un ensemble de protéines contenant au moins une région non structurée caractérisée expérimentalement (Sickmeier et al., 2007).

1.2.3. Les protéines et leurs fonctions

Les protéines assurent différents types de fonctions au sein de la cellule. Ces fonctions peuvent être regroupées en quelques grandes classes principales :

- Fonctions de structures permettant à la cellule de maintenir son organisation dans l'espace et résister aux contraintes physiques. Ce type de fonctions est assuré par des protéines telles que la kératine ou le collagène.
- Fonctions de signalisation qui captent les signaux extérieurs et assurent leur transmission intracellulaire et intercellulaire : les protéines kinases, les récepteurs d'hormones, *etc.*
- Fonctions de motricité qui sont à l'origine des mouvements cellulaires, ainsi que de la mitose assurées par les myosine et les kinésine.
- Fonctions de transport : l'hémoglobine assure par exemple le transport du dioxygène et la transferrine assure celui du fer.
- Fonctions catalytique : beaucoup de protéines jouent le rôle de catalyseurs de réactions chimiques, elles permettent en effet à des réactions chimiques de se dérouler rapidement dans les conditions de température et de pression conformes à la vie. Ces protéines sont appelées enzymes.
- Fonctions de la régulation de l'expression des gènes joués par les protéines qui sont des facteurs de transcription.

Les protéines sont donc les éléments essentiels de la vie de la cellule. Déterminer les fonctions des protéines est une tâche cruciale qui présente une des problématiques les plus étudiées en bioinformatique.

1.2.4. Annotation fonctionnelle des protéines

L'annotation fonctionnelle des protéines consiste à identifier ou à prédire les fonctions auxquelles, elles contribuent, ainsi que leurs localisations. Ces propriétés sont fréquemment décrites à l'aide de vocabulaires structurés ou ontologies telles que la Gene Ontology (GO) (Friedberg, 2006).

La Gene Ontology (GO)

La GO décrit les gènes et leurs produits via un vocabulaire bien défini, unifié, commun à toutes les espèces et structuré dans le cadre d'une ontologie (Consortium, 2004). Les propriétés des produits géniques sont décrites selon trois axes indépendants dans la GO (Consortium, 2001). Ces trois axes regroupant différentes catégories de termes concernent :

6. http ://www.disprot.org/

- Fonction moléculaire : cet axe regroupe les termes décrivant les fonctionnalités des protéines au niveau biochimique. “enzyme“, “transporter“, *etc.* sont des exemples de termes ;
- Processus biologique : les termes dans cet axe décrivent les fonctions biologiques auxquelles contribue une protéine. Par exemple, la croissance cellulaire et la transduction des signaux sont décrites par les termes : ”cell growth” et ”signal transduction“ ;
- Composante cellulaire : cet axe décrit l'endroit de la cellule où les produits des gènes sont actifs. Par exemple : “nuclear membrane”, “ribosome”, *etc.* .

Chaque axe de la GO est structuré sous la forme d'un graphe dirigé acyclique (DAG) où chaque noeud est un terme et les arcs représentent les relations de généralisation/spécialisation. Un terme GO peut être relié à un ou plusieurs autres termes. Il peut en effet posséder un ou plusieurs parents et un ou plusieurs fils. Chaque terme de la GO, à l'exception des trois racines, possède au moins une relation de type “is a” indiquant qu'il est un sous type d'un autre terme.

Méthodes d'annotation fonctionnelle

Les méthodes classiques d'annotations des protéines reposent sur le principe du transfert d'annotations par homologie de séquences (Punta and Ofran, 2008), c'est à dire que l'on transfère les connaissances acquises sur une protéine aux autres protéines homologues. Ce type d'approches est basé sur le fait que deux séquences homologues, *i.e.* ayant une histoire évolutive commune, ont souvent les mêmes fonctions et partagent une grande similarité de séquence. D'autres approches moins communes sont basées sur le principe Guilt By Association (GBA) et utilisent des données autres que les séquences telles que les données de transcription, les données d'interactions entre protéines, *etc.* (Bréhélin et al., 2008; Quackenbush, 2001).

1.3. Les domaines protéiques

Dans les années 1970, il a été observé que certains motifs ou séquences d'acides aminés ont tendance à apparaître dans plusieurs protéines (Ingolfsson and Yona, 2008). Il a été également observé que la grande majorité des protéines s'organisent en plusieurs modules structuraux. Depuis, plusieurs définitions de ces “domaines“ ont vu le jour. En effet, le ”domaine” joue aujourd'hui un rôle crucial, que ce soit en biologie structurale, en génétique, en évolution, ou encore en biochimie. En biologie structurale, un domaine est vu comme une sous-séquence d'acides aminés capable de se replier en une structure stable indépendamment du reste de la protéine. En biochimie, un domaine est vu comme une région de la protéine ayant sa propre fonction. Du point de vue évolutionnaire, les domaines sont considérés comme des unités élémentaires qui forment des protéines multi-domaines par des événements de combinaison, de duplication, de suppression, *etc.*En résumé, les domaines protéiques sont des motifs séquentiels et structurels qui sont trouvés indépendamment dans différentes protéines et dans des combinaisons différentes et semblent, en tant que tels, être les sous-unités fonctionnelles de protéines au-dessus du niveau des acides aminés (Richardson, 1981). La notion de domaine est souvent confondue avec celle de motif. Les motifs sont généralement plus courts que les domaines, leur longueur ne dépasse pas les 30 acides aminés. Ils correspondent souvent à des sites de liaison, des sites actifs d'enzymes et sont en général associés à des interactions bien précises ou encore à des sites de modification post-traductionnelles (Schwartz and Gygi, 2005). Le motif « doigt de zinc », impliqué dans des interactions spécifiques avec l'ADN, fût parmi les premiers motifs protéiques décrits dans la littérature. Un motif peut être impliqué dans un domaine structural. Un domaine peut être alors composé de plusieurs motifs protéiques (Copley et al., 2002). Un domaine protéique peut composer à lui seul une protéine mono-domaine

ou peut s'associer avec d'autres domaines partenaires au sein d'une protéine multi-domaines (Vogel et al., 2004). Bien que beaucoup de protéines connues soient aujourd'hui des protéines contenant un seul domaine, la majorité de protéines contiennent plusieurs domaines et sont dites multi-domaines. Ces dernières sont particulièrement abondantes chez les eukaryotes, où 80% des protéines eukaryotes sont multi-domaines, et elles représentent 65% des protéines procaryotes (Apic, 2001). Le nombre de domaines dans ces protéines multi-domaines s'étend de deux à plusieurs centaines de domaines (c'est le cas de la protéine humaine $Q10466_HUMAN$ (Ingolfsson and Yona, 2008)). Selon les évolutionnistes, les organismes les plus complexes ont un nombre élevé de protéines multi-domaines (Bornberg-Bauer et al., 2005).

1.3.1. Le domaine une unité d'évolution à part entière

Le gène est l'unité la plus utilisée pour étudier les phénomènes évolutifs. Les différents mécanismes que subissent les gènes tels que mutations, duplications, inversions, transpositions, délétions, insertions, sont étudiés pour retracer les phylogénies des espèces. Cependant, les différentes régions d'une protéine ne partagent pas systématiquement la même histoire évolutive. Comme les gènes, les domaines ont aussi leur propre histoire évolutive. Ils peuvent faire l'objet de mutations ponctuelles, de délétions ou encore d'insertions donnant lieu à la formation de nouvelles familles de domaines (Yang and Bourne, 2009; Zhang et al., 2012). Ainsi, une protéine peut être formée par la combinaison de plusieurs domaines en provenance de plusieurs autres protéines (Fitch, 2002). Différentes études ont suggéré l'utilisation de la composition en domaines protéiques pour l'étude des phénomènes évolutifs (Fitch, 2002). Ils présentent des unités d'homologie pertinentes en génomique comparative et pour les études d'évolution moléculaire à une échelle plus petite que celle des gènes entiers (Velthuis and Bagowski, 2008; Zhang et al., 2012).

1.3.2. Combinaisons de domaines

Les domaines se combinent de différentes manières pour donner lieu à un répertoire de protéines varié et un large éventail de fonctions biologiques. Les études sur les combinaisons de domaines au sein des protéines ont conduit à plusieurs constats :

- Certains domaines ont tendance à apparaître préférentiellement avec d'autres domaines favoris dans les architectures multi-domaines. Les combinaisons de domaines observées ne représentent qu'une petite partie des combinaisons possibles et quelques études suggèrent que la formation de ces combinaisons est probablement soumise à une pression de sélection naturelle permettant ainsi de favoriser certaines familles de domaines plus que d'autres (Apic, 2001). Les couples ou triplets de domaines ayant une sur-représentation statistique d'apparition ensemble sont souvent appelés "supra-domaines" (Vogel et al., 2004).
- 90% des domaines apparaissent toujours dans le même ordre : l'orientation N-C terminale des combinaisons de domaines est très conservée (Kummerfeld and Teichmann, 2009).
- Certaines combinaisons de domaines sont spécifiques à certains royaumes ou encore à certains clades d'espèces aboutissant ainsi à la diversification fonctionnelle (Basu et al., 2008).
- Plusieurs domaines ont tendance à se combiner avec plusieurs types de domaines partenaires. Ces domaines sont dits *versatiles* et ont des activités dites auxiliaires. Ils sont souvent impliqués dans les interactions protéine-protéine (*cf.* la section suivante 1.3.3) et apparaissent notamment dans les protéines impliqués dans la transduction de signaux (Forslund and Sonnhammer, 2012).

1.3.3. Interactions des domaines

Plusieurs études récentes ont révélé l'intérêt de l'étude des interactions entre domaines protéiques pour l'étude des interactions entre protéines. Les interactions entre domaines sont souvent déduites *in silico* à partir de la composition en domaines des protéines connues pour être en interaction (Ta and Holm, 2009; Kim et al., 2002; Lee et al., 2006; Nye et al., 2006). Les interactions de domaines prédites peuvent servir à :

- la prédiction de nouvelles interactions de protéines à partir des connaissances des domaines qui les composent (NG et al., 2003).
- l'identification des domaines qui interagissent dans une interaction entre deux protéines données.
- la filtration des grandes quantités de données d'interactomes résultant d'expérimentations à grande échelle (Betel et al., 2004; Shoemaker and Panchenko, 2007).

Les interactions domaine-domaine connues sont stockées dans des bases de données telles que iPfam [7] (Finn et al., 2005), 3DID [8] (Stein et al., 2011) et Domine [9] (Yellaboina et al., 2011). Cette dernière regroupe des interactions de domaines extraites de la PDB ainsi qu'un ensemble de nouvelles prédictions.

1.3.4. Du domaine à la fonction

À partir de l'étude de la composition en domaines d'une protéine, il est souvent possible de déduire une annotation fonctionnelle. En effet, deux tiers des protéines mono-domaines ayant le même domaine ont aussi les mêmes fonctions. De même, 35% des protéines multi-domaines ayant un domaine commun présentent des fonctions semblables, tandis que ce taux s'élève à 80% quand elles partagent au moins deux domaines communs (Hegyi and Gerstein, 2001). Le consortium Interpro qui regroupe les différentes bases de données de domaines (voir ci dessous), annote les familles de domaines par des termes de la Gene Ontology (GO). Cette annotation est faite manuellement, mais le principe général est d'associer une annotation à un domaine si toutes les protéines (annotées) qui ont ce domaine possèdent cette annotation (Mulder and al., 2003).

1.4. Modélisation des familles de domaines protéiques

La modélisation d'un domaine protéique vise à représenter l'ensemble des informations conservées sur un alignement de séquences contenant le domaine protéique en question. Le modèle doit donc représenter non seulement les positions conservées et les substitutions, mais également les insertions et délétions communes dans la famille (Durbin, 1998). Les modèles construits servent ensuite à l'identification et la reconnaissance des domaines sur des nouvelles séquences protéiques. Plusieurs types de modélisation ont été proposés dans la littérature :

- les expressions régulières ;
- les profils ou Position-Specific Scoring Matrices (PSSM) (Gribskov et al., 1987; Gribskov, 1994) ;
- les modèles de Markov cachés et particulièrement les HMM-profils (Eddy, 1996; Eddy, 1998; Sonnhammer et al., 1998) ;
- D'autres représentations sont utilisées telles que : les arbres de suffixes probabilistes (Sagot and Marsan, 2000) et les réseaux de neurones (Blekas et al., 2005) mais nous nous

7. http ://ipfam.sanger.ac.uk/
8. http ://3did.irbbarcelona.org/
9. http ://domine.utdallas.edu/cgi-bin/Domine

contenterons de présenter les trois premiers types de modélisation qui sont ceux que l'on rencontre le plus souvent pour la modélisation des domaines protéiques dans les banques de données.

Les expressions régulières

Les expressions régulières sont un formalisme issu des théories mathématiques des langages formels. Elles permettent de représenter, selon une syntaxe bien déterminée, un ensemble de chaînes de caractères possibles par un ensemble de caractères reliés, souvent appelé "pattern". Dans les années 80, ce formalise a été utilisé pour la modélisation de familles de séquences. Il s'agit de représenter la séquence consensus de l'alignement d'une famille à modéliser selon une syntaxe adaptée au traitement des séquences biologiques (Bairoch, 1988). Ceci a conduit à la mise en place de la banque Prosite (*cf.* section 1.4.1) (Sigrist et al., 2010) regroupant un ensemble de motifs représentés chacun par une expression régulière. Bien que ces modèles servent à marquer les zones conservées d'une manière simple permettant des recherches assez rapides, ils présentent plusieurs limites notamment dans le cas de reconnaissance de longues séquences et sont plutôt adaptés à la reconnaissance de motifs ou de domaines de petites tailles.

Les profils

Initialement introduits par Gribskov en 1987 (Gribskov et al., 1987) et généralisés ultérieurement pour la modélisation des motifs protéiques en 1994 (Gribskov et al., 1994), les profils, souvent appelés Position-Specific Scoring Matrices (PSSM), permettent de représenter un ensemble de séquences et de mesurer la ressemblance de n'importe quelle séquence donnée à cet ensemble. La création d'un profil relatif à une famille de protéines donnée requiert la génération d'un alignement multiple des différentes séquences de la famille en premier lieu. Les informations issues de cet alignement sont ensuite modélisées par une matrice $\mathcal{M}$ de scores spécifiques aux différentes positions sur l'alignement. Chaque score M_{ij} de la matrice reflète la probabilité d'observation du résidu i à la position j dans une séquence de la famille. Ces scores sont calculés via des matrices de substitution (telles que Dayhoff, PAM, BLOSSUM). Une séquence est proche (similaire) au modèle quand elle a tendance à avoir des scores élevés à chaque position. Le score total calculé est la somme des scores des différentes positions du profil. La comparaison du score obtenu par rapport au score seuil fixé par les experts permet d'assigner ou non la séquence en question à la famille. Plusieurs banques de données sont basées sur la représentation par profils des familles telles que Prosite (*cf.* section 1.4.1). Les profils permettent de modéliser plus finement que les expressions régulières la fréquence des acides aminés à chaque position. Ils présentent néanmoins quelques inconvénients principalement reliés au fait que les insertions et délétions ne peuvent pas être modélisées, contrairement aux HMMs.

Les modèles de Markov cachés : HMM

Un modèle de Markov caché (Hidden Markov Model : HMM) est un modèle probabiliste permettant de modéliser des séquences. Les HMM ont été introduits par *Leonard E. Baum* (Baum et al., 1970) à la fin des années 60 et ont été utilisés depuis le début des années 70 pour la reconnaissance de la parole ainsi que pour le traitement automatique des langages naturels (Rabiner, 1989). Depuis les années 90, ces modèles sont devenus un outil populaire pour la modélisation des séquences d'ADN et protéiques depuis les travaux de *Krogh* et *Hausslery* (Krogh et al., 1996). Ces modèles ont été adaptés pour le traitement spécifique des familles de séquences biologiques, d'où la naissance des "HMM-profils". Plusieurs banques de données

de domaines sont basées sur les HMM notamment PFAM (Sonnhammer et al., 1998). Nous consacrons le chapitre 2 pour une présentation plus détaillée de cette base et des HMMs.

1.4.1. Familles et banques de domaines

Parallèlement au développement des méthodes de modélisation et d'algorithmes de recherche des domaines protéiques, de nombreuses banques recensant les domaines protéiques ont été mises en place depuis les années 90. Ces banques diffèrent les unes des autres sur quelques critères : la définition du domaine, le type de représentation formelle du domaine et la méthode d'alimentation de la base. Certaines de ces banques sont alimentées d'une manière complètement automatique, d'autres sont sujettes à des vérifications manuelles. Les alimentations de ProDom (Corpet et al., 1998) et BLOCKS (Henikoff et al., 2000) sont entièrement automatiques par exemple, alors que Pfam et PROSITE (Sigrist et al., 2010) sont plutôt basées sur des approches hybrides, des phases de validation manuelles suivies d'une recherche automatique d'alignements complets. Ces banques ne couvrent pas toutes les mêmes séquences protéiques et le nombre de séquences couvertes varie d'une banque à une autre. Nous présentons dans ce qui suit les banques basées principalement sur les similarités des séquences primaires et qui font partie de la méta-base Interpro (voir dernière section de ce chapitre).

Pfam

Pfam [10] (Finn, 2008; Punta et al., 2012) est l'une des bases de données de domaines les plus utilisées. C'est une base d'alignements multiples représentant des domaines protéiques et des familles de protéines (Sonnhammer et al., 1998). C'est la première banque de domaines à avoir utilisé les HMM-profils pour représenter les différentes familles. Nous distinguons deux sous-ensembles dans la banque Pfam. Il s'agit des sections Pfam-A et Pfam-B. Pfam-A regroupe un ensemble domaines protéiques représentés par des HMM-profils générés à partir d'alignement multiples vérifiés manuellement. Pfam-A dans sa version 23.0 regroupe 10340 domaines protéiques différents. Elle en regroupe 13672 dans la version 26.0 (Punta et al., 2012). Pfam-B est la version entièrement automatisée. Une description plus détaillée de cette banque est donnée dans le chapitre suivant à la section 2.3 .

SMART

Créée en 1998, SMART [11] (Simple Modular Architecture Research Tool) (Schultz et al., 1998) regroupe un ensemble d'alignements multiples et les HMM-profils qui leurs sont associés. SMART fournit avec chaque domaine diverses informations telles que : les voies métaboliques, des données d'interactions ainsi que des informations taxonomiques (Letunic et al., 2012). Dans sa version 7, SMART contient 1009 domaines protéiques expertisés manuellement couvrant 1133 espèces différentes (Letunic et al., 2012). SMART offre une interface permettant la recherche de domaines sur les séquences protéiques données en entrée.

TigrFam

TigrFam [12] (Haft et al., 2003), développée par The Institute of Genome Research (TIGR), est une collection de familles de protéines mise en place pour automatiser l'annotation fonctionnelle

10. http ://pfam.sanger.ac.uk/
11. http ://smart.embl-heidelberg.de/
12. www.tigr.org/TIGRFAMs

de nouvelles séquences procaryotes. Elle est basée essentiellement sur la notion d'"équivalogues" pour les différentes familles. Les membres d'une famille équivalogue partagent les mêmes fonctions moléculaires et ont un ancêtre commun. À la différence des groupes orthologues, un équivalogue peut contenir des paralogues ou encore des gènes ayant subi des transferts horizontaux à condition qu'ils aient les mêmes fonctions (Selengut et al., 2007). La formation d'équivalogues est souvent expertisée manuellement en se basant sur les annotations attribuées à certains membres du groupe sur l'examen d'arbres phylogénétiques afin de vérifier que les différents membres possèdent le même ancêtre. Chaque famille de TigrFam est représentée par un HMM construit à partir de l'alignement multiple des séquences représentatives de la famille.

Prosite

Créée en 1988 par Amos Bairoch, Prosite [13] fait partie des outils disponibles sur la plateforme Bioinformatique d'analyse génomique et protéomique Expasy [14]. Elle fût la première base de données contenant la description des familles de protéines accompagnées de leurs caractérisations fonctionnelles. Prosite contient également les domaines et les sites fonctionnels. Toutes ces données sont générées et vérifiées manuellement par l'institut Suisse de bioinformatique et sont directement intégrées aux annotations Swiss-Prot (Sigrist et al., 2010). Des signatures et des profils sont associés à ces différentes informations. Les domaines et motifs de Prosite étaient à l'origine modélisés par des expressions régulières mais sont maintenant modélisés par des profils construits manuellement à partir d'alignements multiples. De plus, un ensemble de règles ProRules (Sigrist et al., 2005) décrivant des informations biologiquement significatives sur les différents résidus est associé à chaque domaine permettant de régir la reconnaissance des domaines, et améliorant ainsi la pertinence de leur détection. Les profils Prosite peuvent être détectés sur les séquences protéiques via l'outil scanProsite [15] (Castro et al., 2006). Prosite dans sa version 20.85 (Aout 2012) recense 1039 profiles et 1041 ProRules.

Prints

Prints [16] (Attwood et al., 2012) est aussi l'une des plus anciennes bases de données regroupant des empreintes de protéines. Une empreinte est un groupe de motifs conservés, souvent utilisée pour caractériser une famille de protéines. Les empreintes sont générées par des recherches itératives des banques SWISS-PROT/TrEMBL et la modélisation des différentes empreintes est basée sur l'utilisation des expression régulières. Les motifs dans Prints ne sont généralement pas chevauchants sur la séquence mais peuvent être contigus sur la structure 3D. Les empreintes permettent un codage de la structure et des fonctionnalités plus puissant et plus flexible que les motifs en tirant profit de l'information contextuelle fournie pour le voisinage des motifs. Dans sa version 42.0 (Février 2012), Prints inclut 2156 empreintes et 12444 motifs individuels couvrant un grand nombre de protéines de différentes fonctions.

ProDom

La base de données de familles de domaines protéiques ProDom [17] (Corpet et al., 1998) a été développée en 1993 à Toulouse en collaboration entre le Laboratoire de Génétique Cellulaire

13. http ://prosite.expasy.org/
14. http ://www.expasy.org/
15. http ://www.expasy.org/tools/scanprosite/
16. http ://www.bioinf.man.ac.uk/dbbrowser/PRINTS/index.php
17. http ://prodom.prabi.fr/prodom/current/html/home.php

(INRA) et Laboratoire d'Interactions Plantes-Microorganismes (CNRS). PRABI (le centre Bioinformatique Rhone-Alpes) se charge actuellement de sa maintenance. ProDom est un ensemble de familles de domaines construites automatiquement par regroupement et alignement des fragments homologues de protéines à partir des bases Swiss-Prot et TrEMBL. La procédure MKDOM2 de remplissage de ProDom est basée sur des recherches PSI-BLAST récursives sur les séquences d'Uniprot. Contrairement aux autres bases de domaines, aucune modélisation n'est associée aux familles de domaines. L'identification de la famille ProDom d'une nouvelle séquence se fait à l'aide d'un BLAST. ProDom possède des liens vers Pfam ainsi que vers Prosite et Interpro. ProDom fait également le lien entre ces familles et les structures 3D de la PDB et de SCOP (BRU et al., 2005).

SUPERFAMILY

La base SUPERFAMILY [18] (Gougha and Chothia, 2002; De Lima Morais et al., 2011) est une librairie de HMM-profils représentant l'ensemble des protéines de structure connue. Cette librairie est basée sur la classification des familles dans SCOP : chaque modèle correspond à un domaine SCOP et représente la super-famille à laquelle appartient le domaine. Un serveur en ligne permet d'analyser et d'annoter de nouvelles séquences. Étant donnée une séquence protéique ou nucléotidique à analyser, SUPERFAMILY retourne une prédiction de son architecture en domaines ainsi que sa classification dans SCOP. Un alignement entre la séquence requête et les séquences ayant des structures connues est aussi fourni. Le serveur de SUPERFAMILY est utilisé pour l'annotation de génomes, la génomique structurelle, la prédiction de gènes et des études génomiques à base de domaine (De Lima Morais et al., 2011). Il a été utilisé pour identifier les structures 3D de l'ensemble des génomes séquencés des protéines d'Uniprot.

PANTHER

La base de données PANTHER [19] (Protein ANalysis THrough Evolutionary Relationships) (Thomas et al., 2003; Mi et al., 2013) a pour principal objectif la classification fonctionnelle des protéines. Une famille représente un groupe de protéines ayant des relations évolutives. Une famille peut être divisée en plusieurs sous-familles permettant ainsi de raffiner la classification et de modéliser les divergences au sein d'une même famille. Des arbres phylogénétiques représentent chaque famille dans PANTHER. Ces derniers sont divisés en sous-arbres pour représenter les sous-familles. Les fonctions moléculaires, biologiques et les voies métaboliques dans lesquelles les protéines sont impliquées sont prises en considération dans le processus de classification. La classification des familles dans PANTHER est faite via des méthodes bioinformatiques et est sujette à des vérifications manuelles par des experts (Thomas et al., 2003). Chacune des familles et sous-familles est modélisée par un HMM généré à partir d'un alignement multiple. PANTHER inclut également des associations à des ontologies de voies métaboliques dans PANTHER *pathway* qui peuvent être visualisées d'une manière interactive. L'association des voies métaboliques est déduite par des experts humains en se basant sur la littérature. Dans sa version 8.0 (Novembre 2012), PANTHER regroupe 7729 familles de protéines regroupées en 69465 sous-familles fonctionnellement distinctes (Mi et al., 2013).

18. http ://supfam.cs.bris.ac.uk/SUPERFAMILY/
19. http ://www.pantherdb.org/

PIRSF

PIRSF [20] (Nikolskaya et al., 2007; Wu et al., 2004) est une banque de données fournissant une classification hiérarchique non-recouvrante de l'ensemble des protéines Uniprot, reflétant ainsi leurs relations évolutives. La classification est plutôt basée sur les séquences protéiques complètes et non pas sur leurs compositions en domaines. Cette base permet d'apporter des annotations génériques et spécifiques aussi bien qu'une classification des protéines sans domaines bien définis. Les familles dans PIRSF sont sujettes à des vérifications basées sur la littérature, l'étude des similarités de séquences et de structures ainsi que le contexte génomique et phylogénique. On distingue trois niveaux hiérarchiques dans PIRSF :

- Un niveau contenant les familles homéomorphiques dont les membres sont homologues (ils ont un ancêtre commun) et homéomorphes (ils partagent une grande similarité de séquences ainsi qu'une composition similaire en domaines). Chaque famille homéomorphique est représentée par un HMM-profil construit à partir de l'alignement multiple d'un sous-ensemble de séquences représentatives de la famille.
- Un niveau contenant des super-familles au dessus du niveau homéomorphique. Il permet de connecter les familles similaires ou encore les protéines "orphelines" en se basant sur les similarités de compositions en domaines.
- Un niveau de sous-familles, au dessous du niveau homéomorphique. On y trouve les sous-familles représentant des spécialisations par rapport à la famille.

PIRSF est représenté sous la forme d'un réseau permettant ainsi de visualiser les liens d'appartenance aux différentes super-familles.

1.4.2. La méta-base Interpro

Interpro [21] (Hunter et al., 2009; Mulder and al., 2003) est la méta-base qui fédère un grand nombre de banques de données de domaines protéiques, de motifs ou encore de sites fonctionnels. Interpro a été créée en 1999 et est maintenue par l'European Bioinformatics Institute (EBI). Interpro fait appel à onze différentes bases de données : PROSITE, PRINTS, Pfam, ProDom, SMART, TIGRFAMs, PIRSF, SUPERFAMILY, Gene3D, HAMAP et PANTHER. Chaque famille (appelée aussi entrée) dans Interpro a un nom et numéro d'accession uniques et peut contenir une ou plusieurs signatures d'un domaine donné en provenance des différentes bases de données (Mulder and Apweiler, 2002). Les liens évolutifs et fonctionnels entre les différentes entrées sont présentés dans Interpro par des relations hiérarchiques. Des statistiques concernant les recouvrements (en terme de protéines et d'acides aminés) entre les différentes entrées sont également proposées. À chaque entrée, sont associés une description de la famille, des références bibliographiques ainsi que des liens vers les banques du consortium mais aussi vers des bases de structure, d'interactions, *etc.* (Hunter et al., 2009). Le consortium Interpro annote, comme expliqué précédemment (*cf* section 1.3.4), les différentes familles de domaines par des termes de la Gene Ontology. Ces annotations sont propagées aux familles des différentes bases.

Le logiciel *interproscan* [22] (Quevillon et al., 2005) permet d'identifier les domaines protéiques sur une séquence donnée. Interproscan combine les méthodes de recherche des différentes banques et fournit les résultats en provenance des différentes banques d'une manière uniforme.

20. http ://pir.georgetown.edu/pirsf/
21. http ://www.ebi.ac.uk/interpro/
22. http ://www.ebi.ac.uk/Tools/pfa/iprscan/

Notons qu'à côté d'Interpro il existe plusieurs autres banques de données regroupant les informations en provenance de plusieurs banques de domaines (Ingolfsson and Yona, 2008) :

- La banque CDD [23] (Marchler-Bauer, 2005) (Conserved Domain Database) maintenue par le National Center for Biotechnology Information (NCBI). Elle regroupe dans un premier volet les domaines en provenance de deux des banques populaires de domaines : Pfam et SMART. Le deuxième volet consiste en un ensemble de données générées par les experts du NCBI.
- La banque SBASE [24] (Vlahovicek et al., 2005) regroupant un ensemble de domaines protéiques annotés en provenance de l'une des bases de données de domaines : Swiss-Prot, TrEMBL, PIR, Pfam, SMART et PRINTS. La base est également alimentée par des descriptions de domaines extraites de la littérature. Les séquences de domaines sont regroupées sur la base de scores de similarité calculés avec BLAST.

23. http ://www.ncbi.nlm.nih.gov/cdd
24. http ://hydra.icgeb.trieste.it/sbase/

2. Modélisation des familles de domaines à l'aide de modèles de Markov cachés

Ce chapitre est consacré à la présentation des modèles de Markov cachés (Hidden Markov Models ou HMM en anglais). La première partie expose les algorithmes et méthodes utilisés classiquement quand on manipule des HMM. La deuxième partie est dédiée à la présentation des HMM profils utilisés pour la modélisation des séquences biologiques ainsi qu'à la procédure classique de recherches de domaines. Dans la troisième et dernière partie, nous présentons brièvement Pfam, la banque de données de domaines basée sur les HMM profils.

Sommaire

2.1. Les modèles de Markov cachés 28
2.1.1. Présentation 28
2.1.2. Domaines d'applications des HMM 28
2.1.3. Probabilité de génération d'une séquence par un HMM 29
2.1.4. Apprentissage des HMM 32
2.2. Les HMM profils 33
2.2.1. Définition 33
2.2.2. HMMER et SAM 34
2.2.3. Construction d'un HMM profil 35
2.2.4. Recherche de domaines 36
2.3. La banque de domaines Pfam 38
2.3.1. Pfam-A 39
2.3.2. Pfam-B 40

2.1. Les modèles de Markov cachés

2.1.1. Présentation

Un modèle de Markov Caché (Hidden Markov Model) est une structure composée d'un ensemble d'états finis et de transitions qui les relient. Deux distributions de probabilités sont associées à chacun des états. La première est reliée aux transitions sortantes, l'autre est une distribution de probabilités sur les différents symboles possibles d'un alphabet. À la différence des automates probabilistes, où chaque transition permet la génération d'un symbole, les symboles sont ici générés par les états.

HMM : Un HMM H peut être vu comme un quadruplet (ϱ, Σ, T, G) :
- ϱ : un ensemble de N 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*, et servent respectivement à débuter et conclure une séquence.
- Σ : un alphabet fini de symboles (par exemple les 20 acides aminés).
- T : une matrice regroupant l'ensemble des probabilités de transitions entre les différents états. On note $P\ (q \rightarrow q')$ la probabilité de transition de l'état q vers l'état q'. La somme des probabilités de l'état q vers les différents états possible $q' \in \varrho$ est égale à 1.
- G : la matrice des probabilités de génération des symboles de Σ par chacun des états de ϱ. On note $P(s|q)$ la probabilité de générer le symbole s dans l'état q. La somme des probabilités associées à la génération des différents symboles dans chaque état est égale à 1.

Cette définition est celle habituellement utilisée dans la plupart des applications. Elle diffère de la définition originale introduite par Baum (Baum et al., 1970) par l'introduction des états *Begin* et *End*. La figure 2.1 représente un exemple de HMM avec sept états et onze transitions. La génération d'une séquence S de symboles par un HMM consiste à se déplacer d'un état à l'autre en partant de l'état *Begin*. Sur chaque état, un symbole est généré selon la distribution de probabilité associée. Le passage de l'état courant au suivant se fait par une des transitions sortantes via les probabilités de transitions T. Le processus s'arrête lorsque l'on atteint l'état final *End*. La séquence S= *atcct* par exemple, peut être générée par le HMM de la figure 2.1 en partant de l'état *Begin*, puis en allant vers l'état 1 puis sur l'état 3, puis l'état 5 sur lequel on boucle pour générer deux fois le symbole *c*. On va ensuite vers l'état 2 et on sort par l'état *End*. Mais la génération de cette même séquence peut se faire également en empruntant d'autres chemins. Elle peut par exemple être générée en suivant le chemin *Begin*- 1- 4- 5- 5- 2- *End*, ou *Begin*- 2- 4- 5- 5- 2- *End*. Les HMM définissent donc un processus stochastique non déterministe. Le processus de passage d'un état à l'autre est markovien : la probabilité de l'état n dépend uniquement de l'état $n-1$. Ces modèles sont dits "cachés" car étant donné une séquence de symboles générée par le HMM, il est impossible de connaître le processus.

2.1.2. Domaines d'applications des HMM

Le premier domaine dans lequel les HMMs furent utilisés est le traitement de la parole au début des années 70. Ces modèles furent ensuite utilisés dans des domaines variés tels que : le traitement des langages naturels, la reconnaissance d'images et de textes, le traitement des séquences biologiques. Les HMM sont d'une manière générale utilisés pour deux grandes classes de problèmes :

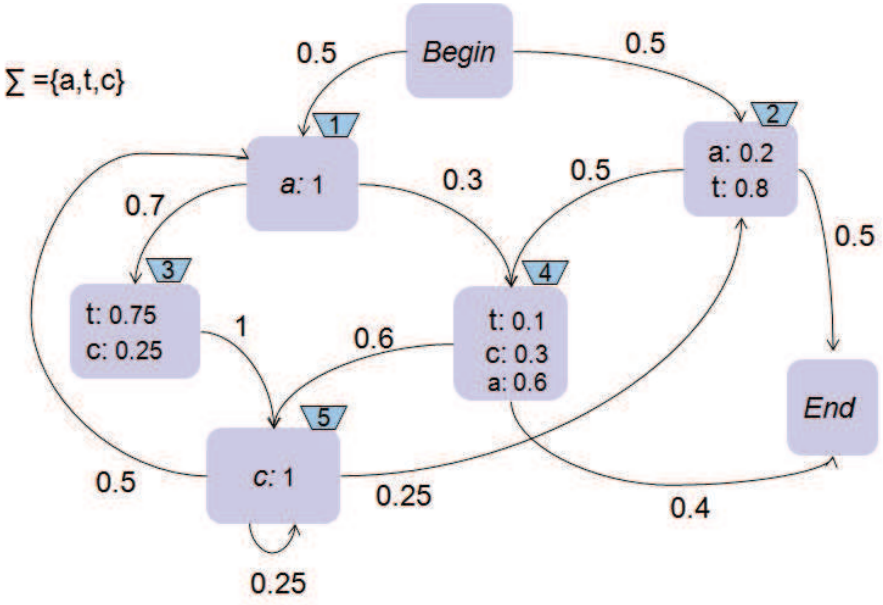


FIGURE 2.1.: Un exemple de HMM

- Les problèmes de classification : dans cette catégorie de problèmes, les différentes classes (ou familles dans le cas des séquences protéiques) sont représentées par des HMM. Étant donnée une nouvelle séquence, le problème est de prédire la classe d'appartenance de cette séquence. Pour cela, on calcule la probabilité de génération de la séquence par chacun des HMM. La séquence est assignée à la classe la plus probable. Des exemples typiques de cette classe de problèmes sont : la classification d'une séquence protéique, la reconnaissance d'un mot parmi un ensemble de mots à partir d'un signal audio, *etc.*
- Les problèmes de segmentation : dans ce cas, il s'agit de découper une séquence en plusieurs sous-séquences ou de chercher des sous-séquences spécifiques. On trouve dans cette catégorie des applications de découpage de signaux musicaux en notes ou la recherche de régions d'intérêt dans les séquences biologiques (régions codantes, régions fonctionnelles, *etc.*). L'idée est de calculer sur le HMM le chemin ayant la probabilité maximale de générer une séquence d'intérêt et de réaliser un découpage de la séquence en fonction des états utilisés dans ce chemin.

2.1.3. Probabilité de génération d'une séquence par un HMM

Soit $S = s_1\ s_2\ s_3\ ...\ s_L$, une séquence de symboles de longueur L. S peut être générée par un HMM H suivant une séquence d'états de H (ou chemin). La probabilité $P(S, C)$ de génération de S selon un chemin $C = q_0\ q_1\ q_2 ...\ q_L\ q_{L+1}$ est donnée par la formule 2.1. C'est la probabilité de transition de l'état q_0 (correspondant à l'état *Begin* de H) à l'état q_1, multiplié par la probabilité de générer le premier symbole s_1 de S sur cet état, multiplié par la probabilité de transition vers l'état q_2 et ainsi de suite. Ce processus est répété jusqu'à l'arrivée à l'état q_{L+1}

correspondant à l'état *End* de H.

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

Or, pour un HMM H de taille N, il peut exister jusqu'à N^L chemins possibles pour la génération de S. La probabilité de génération d'une séquence S par un HMM H est donnée par la formule 2.2. C'est la somme des probabilités de génération de la séquence par chacun des chemins possibles du HMM. L'ensemble des chemins possibles est noté C^S.

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

Si l'on veut calculer la probabilité de S en utilisant cette formule, il faut donc énumérer l'ensemble de ces chemins. Le calcul de la probabilité suivant un unique chemin (formule 2.1), demande de l'ordre de L opérations. La complexité du calcul de $P(S,H)$ en utilisant la formule 2.2 est alors de l'ordre de $O(LN^L)$. La recherche d'une séquence représentant un domaine protéique composé de 200 acides aminés (la taille moyenne d'un domaine protéique) dans une séquence de longueur 200 nécessiterait par exemple 10^{461} opérations, ce qui n'est clairement pas acceptable. Heureusement, il existe une procédure bien plus efficace permettant le calcul de $P(S|H)$: l'algorithme $forward - backward$ basé sur la programmation dynamique.

Algorithmes $forward - backward$

Introduit initialement par Baum en 1972, il est basé sur l'utilisation d'une variable $forward$ définie comme suit :

$$\alpha_k(q) = P(s_1 ... s_k, q_k = q|H), \tag{2.3}$$

C'est la probabilité d'obtenir les k premières observations d'une séquence S en terminant dans l'état q du HMM. Cette variable est calculée de manière récursive dans l'algorithme de programmation dynamique $forward$ (algorithme 1).

Algorithm 1: Algorithme $forward$

Inputs : Séquence S= s_1, s_2, s_3 ... s_L, un HMM H= (ϱ, Σ,T, G)
Output : Calcul de $P(S|H)$
/* ***Étape 1 : Initialisation*** */
foreach $q \in \varrho$ **do**
    $\alpha_1(q) = P(Begin \rightarrow q) P(s_1|q)$
/* ***Étape 2 : Récursion*** */
foreach *k de 2 à L* **do**
    foreach $q \in \varrho$ **do**
        $\alpha_k(q) = (\sum_{q' \in \varrho} \alpha_{k-1}(q') P(q' \rightarrow q)) P(s_k|q)$
Return $P(S|H) = \sum_{q \in \varrho} \alpha_L(q) P(q \rightarrow End)$

Cet algorithme est appelé $forward$ car l'induction est réalisée en avant. Le calcul de la probabilité de génération de la séquence par cet algorithme est fait en deux étapes. La première

consiste à calculer la probabilité de génération du premier symbole de la séquence par chaque état du HMM. Ensuite, à chaque étape, on rajoute un symbole et on réitère la procédure jusqu'à atteindre le dernier symbole de la séquence.

Une autre version de cet algorithme consiste à effectuer le calcul à l'envers. Elle utilise la variable *backward* définie comme suit :

$$\beta_k(q) = P(s_{k+1}...s_L, q_k = q|H). \tag{2.4}$$

$\beta_k(q)$ exprime la probabilité de génération de la séquence $s_{k+1}...s_L$ en partant de l'état q et en arrivant à l'état *End*. Cette variable est calculée par une procédure récursive décrite par l'algorithme 2.

Algorithm 2: Algorithme *backward*

Inputs : Séquence S= s_1, s_2, s_3 ... s_L, un HMM H= (ϱ, Σ,T, G)
Output : Calcul de $P(S|H)$
/* ***Étape 1 : Initialisation*** */
$\forall q \in \varrho, \beta_L(q) = P(q \rightarrow End)$
/* ***Étape 2 : Récursion*** */
foreach *k de L* − 1 *à 1* **do**
    $\beta_k(q) = \sum_{q' \in \varrho} P(q \rightarrow q')P(s_{k+1}|q')\beta_{k+1}(q')$
Return $P(S|H) = \sum_{q \in \varrho} P(Begin \rightarrow q)P(s_1|q)\beta_1(q)$

La complexité de ces deux algorithmes dans le pire des cas est similaire, elle est de l'ordre de $O(LN^2)$ (Eddy, 1998).

Algorithme de viterbi

Dans le cadre des problèmes de segmentation notamment, le but est de déterminer parmi tous les chemins possibles, celui permettant de générer une séquence S avec la probabilité maximale. Ce chemin est couramment appelé chemin de $Viterbi$ et nous le notons C^*. L'algorithme de programmation dynamique de $Viterbi$ (Forney, 1973), introduit par Andrew Viterbi en 1973, permet de déterminer ce chemin de manière efficace. Très similaire à $forward$, il est basé sur l'utilisation d'une variable $\delta_k(q)$ définie comme suit :

$$\delta_k(q) = \max_{q_0 \cdots q_{k-1}} P(s_0, s_1, ...s_k, q_k = q|H). \tag{2.5}$$

$\delta_k(q)$ exprime la probabilité de génération de la séquence $s_1...s_k$ suivant un seul chemin de probabilité maximale, partant de l'état *Begin* et arrivant à l'état q pour générer le $k^{ème}$ symbole. Comme pour la variable $\alpha_k(q)$, $\delta_k(q)$ est calculée de manière itérative et permet, à la fin de la procédure, le calcul de la probabilité de génération de S suivant un unique chemin de probabilité maximale. Or, dans ce problème, c'est le chemin lui même qui nous importe. Une deuxième variable $\psi_k(q)$ est alors utilisée afin de mémoriser pour chaque étape k et pour chaque état q, l'état q' précédent dans le chemin. Une procédure de rétro-propagation permettant de retracer le chemin C^* de Viterbi en partant de l'état *End* est ensuite utilisée (*cf.* algorithme 3).

Comme les algorithmes $forward$ et *backward*, la complexité de l'algorithme de Viterbi est de l'ordre de $O(LN^2)$. À l'exception de la phase de rétro-propagation spécifique à Viterbi,

Algorithm 3: Algorithme de Viterbi

Inputs : Séquence S= s_1, s_2, s_3 ... s_L, un HMM H= (ϱ, Σ,T, G)
Output : Le chemin C^* pour générer S par H avec la probabilité maximale
/* ***Étape 1 : Initialisation*** */
foreach $(q \in \varrho)$ **do**
    $\delta_1(q)$= $P(Begin \rightarrow q)P(s_1|q)$
    $\psi_1(q) = Begin$
/* ***Étape 2 : Récursion*** */
foreach k *de 2 à* L **do**
    foreach $q \in \varrho$ **do**
        $\delta_k(q)$= $\max_{q' \in \varrho}(\delta_{k-1}(q')P(q' \rightarrow q))P(s_k|q)$
        $\psi_k(q) = \arg\max_{q' \in \varrho}(\delta_{k-1}(q')P(q' \rightarrow q))$
/* ***Étape 3 : Rétro-propagation*** */
$q_L^* = \arg\max_{q \in \varrho}(\delta_L(q)P(q \rightarrow End))$
foreach k *de* $L-1$ *à* 0 **do**
    $q_k^* = \psi_k(q_{t+1}^*)$
Return $C^* = Begin, q_1^*, ..., q_L^*, End$

ce dernier est très similaire à l'algorithme $forward$. La différence essentielle tient au fait que Viterbi cherche à maximiser les probabilités des états précédents alors que $forward$ calcule la somme de ces probabilités.

2.1.4. Apprentissage des HMM

Avant de pouvoir être utilisé pour faire de la classification ou de la segmentation, un HMM doit d'abord être construit et paramétré (Durbin, 1998). La manière dont est construit le modèle dépend des connaissances dont on dispose. Dans le cas de la modélisation d'une famille de protéines par exemple, nous connaissons un certain nombre d'exemples de protéines de la famille. Il faut donc construire un modèle à partir de ces exemples, et on se trouve donc dans le cadre d'une problématique d'apprentissage. Deux cas de figure se présentent :

- La structure du HMM est connue : l'apprentissage consiste à paramétrer le modèle, il s'agit alors d'un entrainement du modèle pour apprendre les différentes probabilités (les probabilités d'émission des états et les probabilités de transitions entre états). Deux méthodes basées sur des processus itératifs sont communément utilisées pour l'estimation des paramètres du modèle (Durbin, 1998; Baum et al., 1970) :
  - L'entrainement de Baum-Welch : c'est un algorithme de type Expectation-Maximisation (EM). Il s'agit d'un algorithme itératif de ré-estimation qui cherche à optimiser les paramètres du modèle de manière à maximiser la probabilités des séquences d'apprentissage (principe du maximum de vraisemblance).
  - L'entrainement de Viterbi : une alternative à la procédure de Baum-Welch. Dans cet algorithme, on ne cherche pas à maximiser la probabilité des séquences, mais la probabilité de chaque séquence suivant le chemin de probabilité maximale.

– La structure n'est pas connue : dans ce cas de figure, c'est à la fois la structure (états et transitions) du modèle et l'ensemble des probabilités qui doivent être déterminés à partir de l'ensemble d'apprentissage.

Pour des détails sur ces problématiques d'apprentissage, on pourra se référer à (Bréhélin, 2001).

2.2. Les HMM profils

2.2.1. Définition

Les modèles de markov cachés sont utilisés pour le traitement des données biologiques depuis 1989 (Durbin, 1998). Une adaptation des HMM pour la modélisation des séquences d'ADN (Krogh et al., 1994) et des alignement de séquences protéiques inspirée des profils de Gribskov (*cf.* section 1.4) a été introduite par Krogh et al. en 1994 (Eddy, 1998; Krogh et al., 1996). Ces modèles, appelés HMM profils, font aujourd'hui partie de la palette des modèles les plus utilisés en Bioinformatique (Durbin, 1998). Dans les modèles HMM classiques, un seul type d'état est possible. Une des particularités des HMM profils réside dans l'existence de trois différents types d'états. Un état Match accompagné de deux autres états : Insertion et Délétion. Une autre particularité est l'absence de cycles. Un exemple d'un HMM profil à 4 états est donné par la Figure 2.2. Les états Match représentent les positions conservées. À chaque état Match

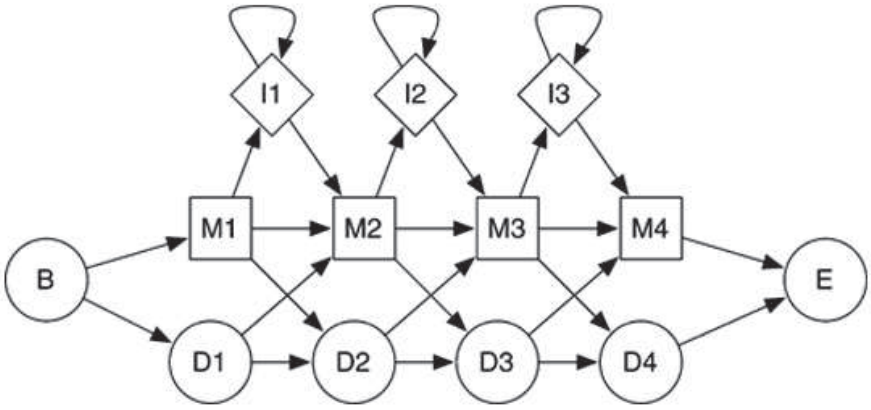


FIGURE 2.2.: Structure d'un HMM profil. L'état *Begin* (B), d'où commence tout(e) chemin (séquence). Ensuite, on suit la succession linéaire des états Matchs (M). Chaque état Match est accompagné des états Insert (I) et Délétion (D). Enfin, on finit toujours sur l'état *End* (E)

M_i est associée la distribution de probabilité des acides aminés (nucléotides dans le cas de traitement de séquences d'ADN) attendus à la $i^{ème}$ position. L'état d'insertion I_i modélise une éventuelle insertion d'un ou plusieurs acides aminés, après la $i^{ème}$ position. Comme les états Match, les états Insert sont associés à une distribution de probabilité sur les acides aminés.

Au contraire, les états de délétion (Delete) ne permettent pas de générer des caractères et ne possèdent donc pas de distribution de probabilités de génération d'acides aminés. Leur rôle est de contourner les états Match afin de pouvoir représenter les séquences ayant subi une délétion d'un acide aminé à une position donnée. Les probabilités associées aux transitions permettent de modéliser d'une manière fine les coûts associés aux événements d'insertion et de délétion. De fait, la probabilité de transition de l'état M_i à M_{i+1} est plus forte que celle de passage de M_i à D_{i+1}. Il est possible de modéliser des délétions successives en passant d'un état Delete D_i à l'état Delete suivant D_{i+1}. Ces délétions successives sont pénalisées. Ces pénalités s'expriment par une plus forte probabilité de passage de D_i à M_{i+1} que de D_i à D_{i+1} (Durbin, 1998). Les HMM profils sont des modèles puissants pour la représentation des familles de séquences protéiques et l'identification de séquences homologues car ils permettent de modéliser finement les informations spécifiques à chaque position d'un alignement ainsi que les éventuels événements d'insertion et de suppression qui peuvent affecter certaines séquences de la famille. De fait, de nombreuses banques de données ont adopté ces modèles pour la représentation des familles de protéines et des domaines protéiques telles que les banques SMART et Pfam (introduites dans la section 1.4.1).

2.2.2. HMMER et SAM

Deux principaux programmes permettent de manipuler les HMM profils pour l'analyse des séquences biologiques : HMMER [1] (Eddy, 1996) et SAM (Hughey,). La banque Superfamily (section 1.4.1) est basée sur l'utilisation de SAM tandis que les profils des domaines dans la base de données Pfam sont conçus pour être manipulés par le logiciel HMMER initialement développé par Sean Eddy (Eddy, 1996). Comme nous nous intéressons dans cette thèse à la recherche des domaines Pfam, nous nous limitons dans ce chapitre au type de modèles utilisés par HMMER ainsi qu'aux méthodes de manipulation des HMM profils qui lui sont spécifiques. La première version publique (1.8) d'HMMER date de 1995. Depuis cette date, plusieurs changements ont été apportés à HMMER pour donner naissance à la version 2 qui montre une nette amélioration de ses fonctionnalités. C'est à la fin de l'année 2003 qu'est apparue la version 2.3.2 (Eddy, 2003) à laquelle on fait référence dans ce manuscrit. Certaines expériences (*cf.* la section 6.4 du chapitre 6) ont été réalisées avec la version 3.0 qui a vu le jour en Mars 2010 (Finn et al., 2011). HMMER regroupe un ensemble de méthodes permettant la manipulation des HMM profils. Ces méthodes vont de la construction et calibrage des HMM profils à la reconnaissance des familles en se basant sur des comparaisons entre profils et séquences. Les HMM profils manipulés par HMMER sont formatés sous un format (.hmm) bien particulier. Les fichiers .hmm sont composés de deux parties :

- Un entête comportant un ensemble d'informations (nom et identifiants) sur le modèle et la famille de protéines modélisée. Cet entête contient également des informations sur l'ensemble des séquences de l'alignement d'apprentissage (nombre de séquences, *etc.*). Les différents seuils de scores pour la recherche des homologues sont également donnés (*cf.* section 2.2.4).
- Le corps contenant les détails du modèle : les probabilités de génération des différents acides aminés par chacune des positions ainsi que les probabilités de transition entre les différents états Match et Insert. Le corps du modèle commence par une ligne décrivant les transitions de sortie de l'état *Begin*. Ensuite, chaque état Match du modèle est représenté par 3 lignes successives :
  - La distribution de probabilités de génération des acides aminés sur l'état Match ;
  - La distribution de probabilités de génération de l'état Insert ;

1. http ://hmmer.janelia.org/

– Les probabilités de transition entre les différents états de la position i à ceux de la position suivante $i+1$.

Enfin, le symbole "//" marque la fin du modèle.

Les HMM profils utilisés par HMMER 2 suivent une architecture canonique dite *Plan*7. Un exemple d'architecture Plan7 est donné par la figure 2.3. Comme les HMM profils classiques, dont un exemple est donné à la figure 2.2, les états Match sont représentés par des carrés. Chacun de ces états est accompagné de deux états : *Insert* (représenté par un losange) et un autre *Delete* (représenté par un cercle). On y retrouve également les états *Begin* (B) et *End* (E). Par contre, trois nouveaux états N, C et J sont ajoutés à ce modèle. Les états N et C permettent de modéliser respectivement la partie de séquence en amont du N-ter et celle en aval du C-ter du domaine. L'état J permet de chercher plusieurs occurences du domaine dans une séquence. En effet, les transitions non nulles $E \rightarrow J$, $J \rightarrow J$ et $J \rightarrow B$ permettent de boucler sur le modèle et d'identifier ainsi plusieurs occurences. Les flèches indiquent les transitions possibles entre les différents états.

2.2.3. Construction d'un HMM profil

La première phase de la construction d'un HMM profil consiste à sélectionner un ensemble de séquences, appelées séquences "seed", qui soit le plus représentatif possible de la diversité d'une famille à modéliser. Il est important que les séquences choisies soient de bonne qualité pour pouvoir créer des modèles capables de reconnaître les membres de la famille avec un faible taux de faux positifs. D'une manière générale, ces ensembles de séquences "seed" contiennent peu de séquences de protéines divergentes. Ils sont majoritairement formés par des séquences d'organismes modèles classiques partageant de fortes similarités. Des algorithmes de pondération des séquences et de lissage des probabilités sont alors utilisés lors du paramétrage du modèle pour minimiser cet effet. Nous verrons dans le chapitre 3.2 qu'en dépit de cette phase d'adaptation des paramètres, les modèles classiques restent peu adaptés à l'étude de protéines divergentes. La seconde phase consiste à générer un HMM profil à partir d'un alignement multiple des séquences "seed" (Durbin, 1998). La méthode *hmmbuild* d'HMMER permet de le

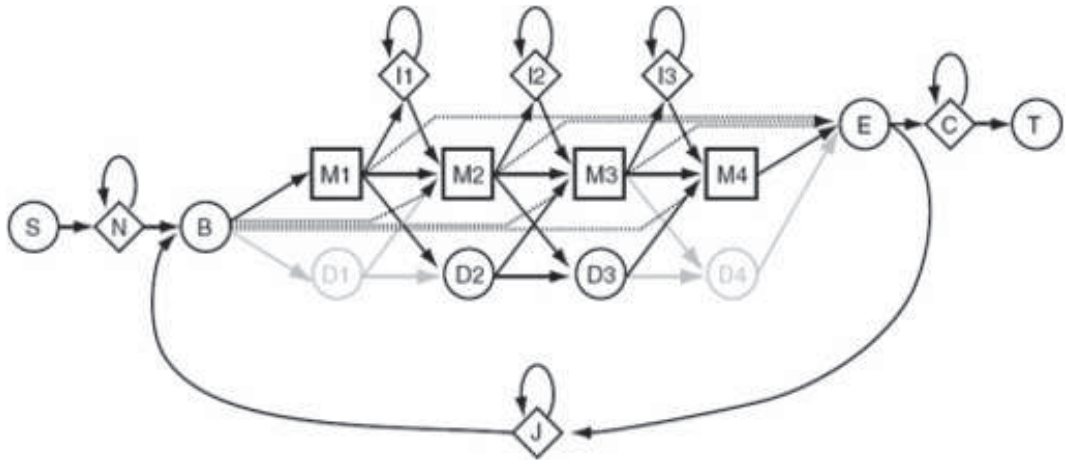


FIGURE 2.3.: Architecture *Plan*7 utilisée pour modéliser les HMM profils dans HMMER

faire en procédant en plusieurs étapes décrites ci dessous :

- Pondération relative des séquences les unes par rapport aux autres afin de limiter l'effet de présence de séquences identiques et avantager les séquences particulières (peu représentées) dans l'ensemble d'apprentissage. Elle est basée par défaut sur l'arbre de Gestein/-Sonnhammer/Chotia (GSC) dans HMMER 2. L'utilisation d'autres méthodes telles que l'entropie maximale ou encore l'algorithme de Voronoi est possible (Eddy, 2003).
- Identification des positions les plus conservées (*i.e.* présentes dans une majorité de séquences sans gaps) de l'alignement qui vont être modélisées par des états Match. Plusieurs règles peuvent être adoptées à ce niveau. Une approche classique est par exemple de considérer une colonne lorsqu'elle comporte une proportion de gaps inférieure à un seuil donné.
- Apprentissage des probabilités d'émission associées aux états matchs. Deux algorithmes sont communément utilisés dans cette étape :
  - l'algorithme de pondération des séquences permettant d'attribuer aux acides aminés qui ne sont jamais observés à une position donnée une probabilité non nulle. Cette phase est indispensable pour éviter les séquences non reconnues. Les matrices de substitution entre acides aminés et les mélanges de Dirichlet sont utilisés dans cette phase (Sjolander et al., 1996).
  - l'algorithme de lissage des probabilités permettant d'associer aux acides aminés non observés à une position donnée des probabilités non nulles.
- Apprentissage des probabilités de transitions et d'émissions des états Insert. Cet apprentissage ne se fait pas uniquement sur la base de l'alignement mais aussi en utilisant des connaissances *a priori* ou par des processus itératifs de recherche de paramètres optimaux.

Il y a beaucoup d'options possibles pour *hmmbuild*. Les algorithmes utilisés par défaut dans les différentes étapes décrites ci-dessus peuvent être modifiés. Une des autres options possible est par exemple de choisir entre un alignement local ou global. C'est à dire qu'on choisit d'adapter le modèle construit à la recherche de fragments de domaines ou la recherche globale du modèle dans la séquence cible (Eddy, 2003). Ce point est plus amplement détaillé dans la section 2.2.4.

2.2.4. Recherche de domaines

Une des tâches pour laquelle les HMM profils sont les plus utilisés est l'identification des domaines qui composent une séquence protéique. Pour ce problème, on dispose d'une base de données de familles de domaines (par exemple la base Pfam *cf.* section 2.3), dans laquelle chaque famille est modélisée par un HMM différent. Étant donnée une séquence protéique, chaque HMM est utilisé pour calculer un score, appelé aussi *log-odds* score, reflétant la similarité d'une partie de la séquence au domaine modélisé par le HMM (*cf.* section 2.1.3). À l'issue de cette étape, une décision est prise pour prédire les domaines composant la protéine sur la base des scores obtenus. Pour affirmer la présence du domaine sur la protéine, deux métriques peuvent être considérées :

- Comparaison du score obtenu à un seuil de référence au dessus duquel on peut affirmer la présence du domaine dans la protéine. Chaque modèle HMM de Pfam est accompagné d'un seuil de score appelé "gathering threshold" (Eddy, 2003). Cependant, les seuils suggérés par Pfam sont très restrictifs notamment dans le cas de protéines divergentes (*cf.* chapitre 3).
- Estimation de la E-valeur associée au score (voir section 2.2.4), et comparaison à un seuil de référence. Un seuil égal à 0.1 est considéré comme un seuil de confiance au dessus duquel des traitements supplémentaires sont recommandés pour éviter les faux positifs (Wong et al., 2011). Ce seuil est égal à 10 par défaut dans HMMER (Eddy, 2003) et peut

être paramétré par l'utilisateur.

Calcul des scores

Le score d'un HMM H sur une séquence protéique S est basé sur la probabilité $P(S|H)$ de générer la séquence S par le HMM H. Le score est calculé comme suit :

$$score(S|H) = \log \frac{P(S|H)}{P(S|H_0)}, \tag{2.6}$$

avec $P(S|H_0)$ la probabilité de S par le modèle nul H_0. Le modèle nul est un HMM très simple dont la structure est composée d'un seul état G qui boucle sur lui même (*cf.* figure 2.4). Les distributions de probabilités dans cet état correspondent à la composition moyenne en acides aminés des protéines de la banque Swiss-Prot (Eddy, 2003). La transition $G \rightarrow G$ permet de boucler plusieurs fois sur l'état G et de générer ainsi des séquences aléatoires de longueurs différentes (Barret et al., 1997).

Calcul de la E-valeur

À partir du score d'une séquence pour un HMM donné, une E-valeur peut également être calculée afin d'évaluer la significativité statistique du score. La E-valeur représente l'espérance du nombre de séquences pouvant obtenir un aussi bon score que la séquence requête, dans une base de données de séquences aléatoires de taille M, où M est le nombre de séquences (Eddy, 2003). Dans notre cas, M est égal au nombre de protéines d'un organisme. Une E-valeur de 1 par exemple, indique qu'en espérance, on peut s'attendre à obtenir un hit aussi bon dans un ensemble de séquences aléatoires de même taille. Ainsi, plus la E-valeur est faible, plus la ressemblance de la séquence au HMM est significative (Eddy, 2010). La E-valeur peut être calculée par deux méthodes différentes (Wong et al., 2011) :

– Estimation basée sur la distribution des scores de Viterbi. Cette distribution est supposée suivre une EVD (extreme-value distribution) (Wong et al., 2011). L'EVD a deux paramètres μ et λ qui sont estimés lors de la phase de calibrage du HMM. Pour cela, un ensemble de séquences aléatoires est généré, le score du HMM pour chacune des séquences est calculé, et on trace l'histogramme de la distribution des scores. Ensuite les paramètres μ et λ qui permettent de faire correspondre au mieux l'EVD à l'histogramme sont estimés. À l'issue de la phase de calibrage du HMM par la fonction *hmmcalibrate* de HMMER, les valeurs estimées de ces deux paramètres sont associées au fichier *.hmm*. La E-valeur

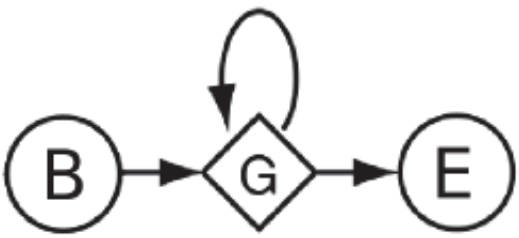


FIGURE 2.4.: La structure du modèle nul utilisée par HMMER

est ensuite calculée en fonction de μ et λ par la formule suivante :

$$E - valeur(S|H) = M(1 - Z^{-Z^{-\lambda(score(S|H)-\mu)}}) \tag{2.7}$$

Z étant la base logarithmique utilisée dans le calcul du score.

– Estimation brute selon une fonction logistique consistant à estimer une borne supérieure de la E-valeur tel que proposé dans (Barret et al., 1997). Cette borne est estimée par la formule suivante :

$$E - valeur(S|H) = M_z(1 - Z^{-score(S|H)}) \tag{2.8}$$

L'utilisation de la deuxième manière est plus rapide mais beaucoup moins précise. Cette méthode a d'ailleurs été abandonnée dans HMMER 3 et seule l'estimation basée sur une EVD est encore utilisée (Wong et al., 2011; Eddy, 2010).

Recherches locales et globales

Les recherches s'effectuent par défaut en mode dit "glocal" permettant d'identifier la correspondance entre la totalité du modèle (recherche globale du côté du modèle) avec une partie de la séquence cible (locale du côté de la séquence). Or dans certains cas, seule une partie du modèle partage des similarités avec la séquence cible. Dans ces cas, on peut donc souhaiter effectuer une recherche locale (seule une partie du HMM est alignée contre la séquence).

La recherche en mode global ou local dans HMMER ne dépend pas des algorithmes de recherche mais plutôt des modèles utilisés dans la représentation des profils (Eddy, 1998). La présence des transitions non nulles en pointillés sur l'architecture *Plan*7 permet d'avoir une deuxième version du modèle adaptée aux recherches locales. Les flèches pointillées sortantes de l'état *Begin* permettent de "sauter" le début du domaine, tandis que celles entrantes dans l'état *End* permettent de "sauter" la fin du domaine. Dans sa version 23.0, Pfam propose pour chacun des domaines deux versions du HMM dites fs et ls selon qu'on souhaite faire une recherche de fragments de domaines ou une recherche globale sur le modèle. Dans la dernière version HMMER 3.0, l'architecture des HMM profils a changé et seules les recherches locales sont autorisées (Eddy, 2010; Finn et al., 2011). Le principal atout apporté par cette version est lié à la réduction considérable du temps de calcul. Quelques unes des modifications apportées au niveau d'HMMER 3.0 sont citées ci dessous (Eddy, 2010) :

– Ajout d'une étape de filtration des séquences données en entrée pour *hmmsearch*. La filtration progressive est basée sur des heuristiques qui mesurent la divergence des séquences. Ces heuristiques sont semblables à celles utilisées dans BLAST. Ceci permet d'obtenir un gain non négligeable en termes de temps de calcul.
– Une modification majeure réside dans le remplacement de l'algorithme de Viterbi par l'algorithme $forward$ pour l'estimation des scores.
– L'architecture Plan 7 (*cf.* figure 2.3) n'est plus utilisée et seules les recherches locales sur ces modèles sont désormais autorisées.

2.3. La banque de domaines Pfam

Pfam [2] est l'une des bases de données de domaines les plus utilisées. Elle regroupe la plus grande collection de domaines protéiques, représentés chacun par un alignement multiple et un HMM profil. On distingue deux sous-ensembles dans la banque Pfam : Pfam-A et Pfam-B (Punta et al., 2012).

2. http ://pfam.sanger.ac.uk/

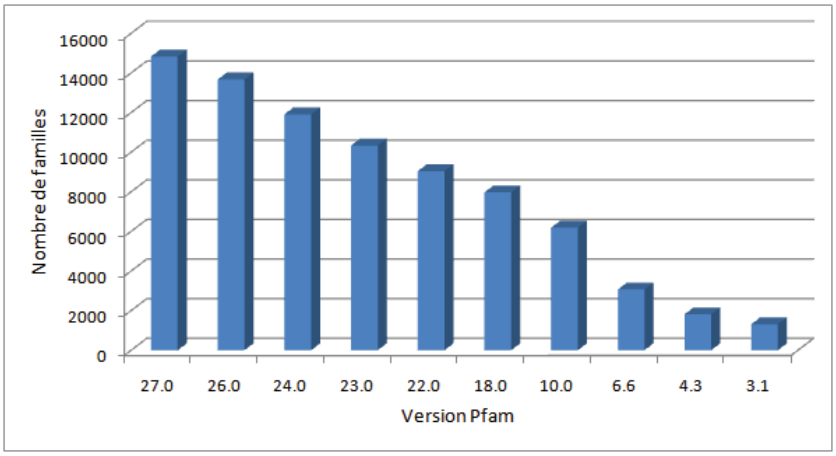


FIGURE 2.5.: Les versions sont représentées en abscisses et le nombre de domaines en ordonnée. Ces informations sont extraites à partir des publications décrivant les mises à jour apportées à Pfam lors des passages d'une version à une autre (Sonnhammer et al., 1997; Sonnhammer et al., 1998; Bateman et al., 1999; Bateman et al., 2000; Bateman et al., a; Bateman et al., b; Finn et al., a; Finn, 2008; Sammut et al., 2008; Finn et al., b; Punta et al., 2012)

2.3.1. Pfam-A

La section Pfam-A fournit une importante librairie de domaines protéiques. Chaque domaine est modélisé par un HMM profil construit à partir d'un alignement multiple de séquences de protéines d'organismes modèles contenant le domaine. Dans Pfam 23.0, deux modèles HMM adaptés respectivement aux recherches globales et locales sont présents. L'ajout d'une nouvelle famille dans la base Pfam se fait suivant quatre grandes étapes (Punta et al., 2012) :

1. Génération d'un alignement multiple de bonne qualité (appelé "seed alignment"). Cet alignement est vérifié manuellement par des experts. Les séquences choisies doivent être le plus représentatives possible de la diversité de la famille pour permettre d'identifier les nouveaux membres d'une famille tout en minimisant le nombre de faux positifs (Sonnhammer et al., 1997). Ces séquences proviennent le plus souvent d'organismes modèles classiques tels que l'humain, la levure et la drosophile. Les protéines des organismes moins étudiés étant moins bien annotées, elles sont plus rarement présentes dans les alignements.
2. Construction d'un HMM profil à partir de cet alignement. Le logiciel HMMER est utilisé pour la construction des modèles. Les HMM profils sont accompagnés de seuils de score permettant de limiter le nombre de faux positifs lors de l'identification de nouveaux membres de la famille (Finn, 2008). Ces seuils sont spécifiques à chaque famille et sont attribués manuellement par les experts.
3. Comparaison du HMM profil à l'ensemble des séquences de la base Uniprot. Les séquences ayant un score supérieur au seuil associé au modèle sont identifiées.
4. Génération de l'alignement de la famille à partir des séquences identifiées dans l'étape précédente.

Le nombre des familles de domaines représentées dans Pfam ne cesse d'augmenter couvrant ainsi de plus en plus de régions non annotées des protéines Uniprot (Sammut et al., 2008). Dans sa version 26.0, Pfam offre une collection de 13.672 domaines (Punta et al., 2012). La version 23.0 n'en contenait que 10.340 (Finn, 2008). La nouvelle version 27.0 datant de Mars 2013 en contient 14.831. La figure 2.5 illustre cette évolution.

2.3.2. Pfam-B

Comme pour les domaines Pfam-A, les domaines Pfam-B sont représentés par un alignement de séquences. Cependant, contrairement aux premiers qui sont annotés et vérifiés manuellement, ces alignements sont générés de manière automatique sans l'intervention d'experts humains (Sonnhammer et al., 1998). Les alignements sont obtenus à partir de clusters formés par la méthode ADDA (Heger et al., 2005; Heger and Holm, 2003) et sont ensuite filtrés pour éliminer les domaines chevauchant les domaines Pfam-A. Une autre différence réside dans le fait que toutes ces familles ne sont pas représentées par des HMM. De moins bonne qualité que Pfam-A, Pfam-B regroupe cependant un nombre beaucoup plus important de domaines protéiques. Pfam 26.0 regroupe par exemple 142.303 domaines Pfam-B (Punta et al., 2012). Pfam B est utilisé pour identifier les familles intéressantes qui seront intégrées et modélisées dans une prochaine version de Pfam-A (Sammut et al., 2008; Finn, 2008).

3. Pathogènes Eukaryotes : présentation et problèmes d'annotation

Les pathogènes Eukaryotes appartiennent à divers clades phylogénétiques et présentent une source majeure de morbidité et de mortalité dans le monde entier. En Afrique, un des continents les plus impactés, ils causent des millions de morts et constituent un fardeau économique immense. Ils appartiennent à diverses origines phylogénétiques et causent une large variété d'infections chez les organismes hôtes (Haldar et al., 2006). Ces espèces sont souvent très mal annotées. Nous présentons dans une première partie de ce chapitre les espèces *Plasmodium* et *Leishmania*, agents responsables respectivement de la malaria et la leishmaniose qui figurent parmi les maladies les plus fréquentes et les plus dangereuses en terme de mortalité. Dans la seconde partie, nous discutons de la difficulté d'annotation de ces deux espèces en particulier et de celles des autres espèces Eukaryotes pathogènes en général.

Sommaire

3.1. Deux pathogènes emblématiques : *Leishmania* et *Plasmodium* 42
3.1.1. Leishmanioses et *Leishmania* 42
3.1.2. Le parasite *Leishmania* 45
3.1.3. Paludisme et *Plasmodium* 48
3.1.4. Le parasite *Plasmodium* 50
3.2. Annotation fonctionnelle des pathogènes Eukaryotes . . . 51
3.2.1. Annotation fonctionnelle de *Leishmania* et *Plasmodium* . . . 51
3.2.2. Détection des domaines sur les protéines divergentes 52

3.1. Deux pathogènes emblématiques : *Leishmania* et *Plasmodium*

3.1.1. Leishmanioses et *Leishmania*

La leishmaniose est une maladie vectorielle, transmise par la piqûre de différents arthropodes hématophages (Herwaldta, 1999). Elle figure parmi les maladies les plus négligées (Alvar et al., 2012), même si on estime qu'environ 350 millions de personnes habitent dans des zones à risques et que plus de 2 millions de cas sont déclarés chaque année selon les rapports de l'OMS. Le nombre global de personnes touchées est estimé à 12 millions dans le monde entier (Raymond et al., 2012). De nombreux cas sont non diagnostiqués ou non signalés, particulièrement dans les zones où les patients n'ont pas accès aux soins ou encore quand les capacités de diagnostic sont limitées, voire absentes.

Histoire et répartition géographique

Connues de l'Ancien Monde sous le nom d'ulcère oriental, les leishmanioses étaient déjà décrites dans des textes du VII siècle. D'autres textes dateraient d'entre 1500 à 2500 av. J-C. (Cox, 2002). Avicenne et d'autres scientifiques arabo-musulmans avaient décrit la maladie au nord de l'Afghanistan, à Baghdad et à Jéricho. Ces formes avaient des désignations variées utilisées encore aujourd'hui tels que "Boutôn d' Alep", "Ulcère de Baghdad" ou "Ulcère de Delhi" (Zink et al., 2006). Des analyses récentes d'ADN, ont identifié des traces de leishmaniose viscérale dans la moelle osseuse de momies égyptiennes et celles de chrétiens venues de Nubie (région du nord Soudan et du sud de l'Égypte, longeant le Nil) datant de plus de 4000 ans (Zink et al., 2006). L'infection des égyptiens de cette époque peut être expliquée par leurs relations commerciales pendant le moyen âge. C'est justement au Soudan qu'on imagine le foyer originel des leishmanioses viscérales, qui se seraient étendues au sous-continent Indien puis au nouveau monde (Pratlong et al., 2001). La leishmaniose a attiré pour la première fois l'attention des médecins occidentaux en 1824 à Jessore, en Inde, où on a cru au départ avoir affaire à une forme de malaria. L'Inde a donné au "kala-azar" son nom traditionnel, qui signifie en hindi "la fièvre noire", ainsi nommée à cause de la coloration noire de la peau qui apparaît sur les extrémités et l'abdomen et qui est l'un des symptômes de la forme indienne de la maladie (Roberts and Janovy, 2008). Le parasite *Leishmania* a été isolé pour la première fois par Sir William Leishman et Charles Donovan en 1900 par une biopsie de la rate d'un soldat mort de fièvre à Dum-Dum en Inde. Le parasite fut alors nommé *Leishmania donovani* (Roberts and Janovy, 2008). C'est donc à la fin du XIXème siècle que les scientifiques purent étudier les origines de la leishmaiose et son mode de transmission. Jusque là, on avait recueilli dans plusieurs pays de nombreux témoignages et signes de passage de la maladie. L'infection chez l'homme a été mise en évidence pour la première fois par Charles Nicolle à l'Institut Pasteur de Tunis en 1908. Dans le Nouveau Monde, des marques de lésions défigurantes causées par la leishmaniose cutanée et muco-cutanée apparaissent sur des sculptures datant du Vème siècle. Dans des écritures du XVIème siècle, des missionnaires espagnols décrivaient aussi une maladie défigurante. On pensait au début qu'il s'agissait de la même forme que dans l'ancien monde, mais, en 1911, Gaspar Vianna a constaté que les parasites en Amérique du Sud diffèrent de ceux d'Inde et d'Afrique. La nouvelle espèce pris le nom de *Leishmania braziliensis*. Depuis, de nombreuses autres espèces, trouvées dans le nouveau monde uniquement, ont été découvertes (*cf.* section 3.3).

La leishmaniose met en jeu la santé d'un très grand nombre d'individus sur la planète (Alvar et al., 2012). La répartition géographique de la maladie est très étendue, elle touche plus

de 88 pays (dont 72 pays en cours de développement) notamment dans les régions tropicales et sub-tropicales. La leishmaniose sous ses différentes formes touche 4 continents comme le montrent les cartes des figures 3.1 et 3.2. On distingue deux grandes situations géographiques, l'Ancien monde (sud de l'Europe, Afrique, Proche-orient et Asie) et le Nouveau monde (Amériques du Nord, du Sud et Centrale). La maladie se concentre dans les zones de sécheresse, de famine, et à densité de population élevée (Alvar et al., 2006). Les principaux foyers endémiques sont : l'Inde, le Bangladesh, le Népal, le Soudan et le Brésil (Desjeux, 2004; Alvar et al., 2012). On retrouve 90% des cas de la leishmaniose viscérale au Bangladesh, Brésil, Inde,

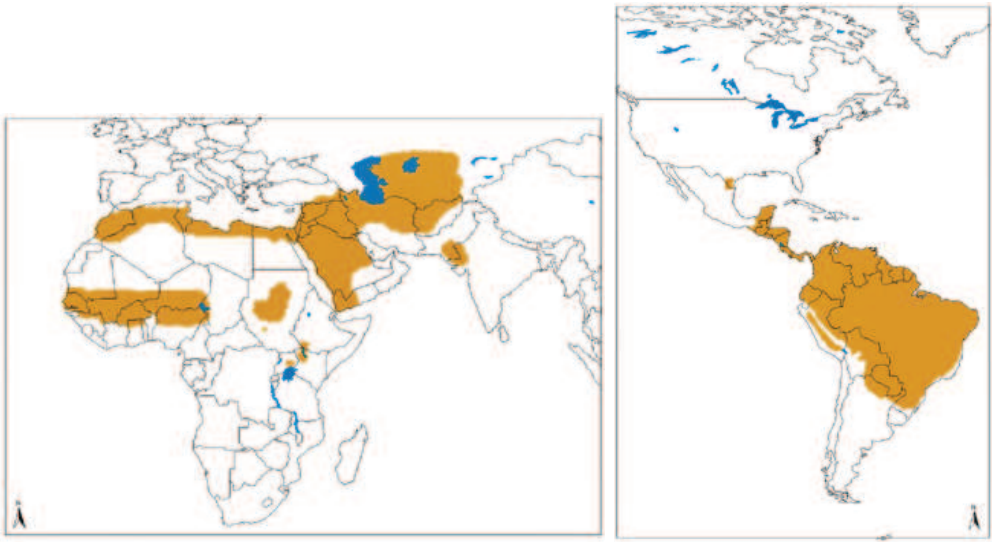


FIGURE 3.1.: Répartition géographique des leishmanioses cutanées et muco-cutanées à travers l'ancien monde (gauche) et dans le nouveau monde (droite) (*http : //www.who.int/leishmaniasis/leishmaniasis_maps/en/index.html*)

Népal et Soudan. C'est au Brésil, Pérou et aussi en Afghanistan, Iran, Arabie Saoudite et en Syrie que l'on retrouve la grande majorité des cas de leishmaniose cutanée (autour de 90% des cas) (Alvar et al., 2012). La forme cutanée diffuse est très répandue en Éthiopie et au Kenya (WHO, 2012). Les leishmanioses sont de plus en plus fréquemment observées chez des individus immunodéprimés, notamment chez les individus infectés par le VIH (Desjeux, 2004). On compte aujourd'hui des centaines de milliers de cas (surtout des adultes entre 30 et 45 ans) de co-infection VIH/leishmaniose (viscérale notamment) dans les pays en voie de développement. Plusieurs cas de co-infection sont également observés au sud de l'Europe (Espagne, France, Italie et Portugal) (Alvar et al., 2008; Desjeux and Alvar, 2003).

Formes et traitements de la leishmaniose

L'infection par le parasite *Leishmania* est à l'origine de diverses manifestations cliniques dont les caractéristiques symptomatiques se manifestent sous la forme d'ulcérations du derme pouvant s'étendre aux muqueuses et aux viscères. Les différentes formes cliniques de la leishmaniose sont communément classées en trois formes principales. Nous distinguons alors la leishmaniose sous la forme cutanée, muco-cutanée et viscérale. Une autre sous classe de leishmaniose cutanée dite non localisée est parfois distinguée (Convit et al., 1972). Les différentes

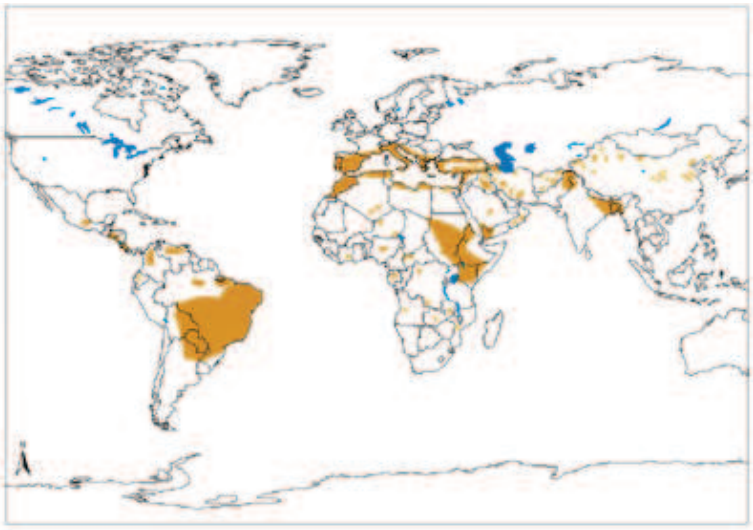


FIGURE 3.2.: Répartition géographique de leishmaniose viscérale et muco-cutannée dans le monde
(*http : //www.who.int/leishmaniasis/leishmaniasis_maps/en/index.html*)

formes cliniques de la leishmaniose diffèrent par leur gravité et leur répercussion sur la santé (Murray et al., 2005).

- **La leishmaniose cutanée** : C'est une forme de la maladie pour laquelle des ulcérations dermiques se forment et s'infiltrent en profondeur. Les lésions cutanées siègent en général sur le lieu de la piqûre, sur des zones exposées comme le visage, les bras et les jambes. Ces lésions guérissent généralement de manière spontanée en quelques mois, mais laissent souvent des cicatrices permanentes.
- **La leishmaniose muco-cutanée** : Sous cette forme, les ulcérations sont plus profondes. Elles s'étendent aux muqueuses de la bouche, de la gorge et du nez et causent des dommages tissulaires défigurants. Cette forme de leishmaniose cause parfois la mort par des infections secondaires telles que les pneumonies.
- **La leishmaniose viscérale** : cette dernière forme, également connue sous le nom de *Kala Azar*, est caractérisée par une forte fièvre, la perte de poids substantielle et l'anémie. Contrairement aux autres formes de leishmanioses, cette dernière ne se présente pas par des ulcérations cutanées. Les parasites migrent vers les organes lymphoïdes comme le foie, la rate et la moelle osseuse. La leishmaniose viscérale est souvent fatale en l'absence de traitement. Elle occupe la seconde place après la malaria dans le classement des problèmes de santé publique DALYs (indicateur utilisé par l'OMS pour mesurer le niveau de santé général) [1]. Cette forme peut engendrer des infections (notamment dans le cas de traitements inadéquats) connues sous le nom de *post Kala Azar* et caractérisées par l'apparition de lésions dermiques.

Les vaccins contre la leishmaniose n'existent pas, et les rares médicaments disponibles deviennent de plus en plus inefficaces à cause des phénomènes de résistance (Croft et al., 2006). Les deux molécules fréquemment utilisées pour le traitement des différentes formes de leish-

1. http ://www.who.int/healthinfo/global_burden_disease/metrics_daly/en/

maniose sont l'antimoine et l'antimoniate qui sont administrées par injections intramusculaires ou intraveineuses et nécessitent une longue période d'hospitalisation (Murray et al., 2005). À côté de ces contraintes, le phénomène de résistance est le problème majeur, notamment pour la leishmaniose viscérale dans les zones endémiques telles que l'Inde et le Soudan. Plusieurs cas de résistance de leishmanioses cutanées et muco-cutanées sont également signalés dans le nouveau monde. L'échec thérapeutique peut parfois être attribué à d'autres facteurs, tels que la co-infection avec le VIH ou la tuberculose.

3.1.2. Le parasite *Leishmania*

Le genre *Leishmania* regroupe les différentes espèces responsables des leishmanioses. Plusieurs études ont établi le lien entre les différentes formes cliniques de la leishmaniose et l'espèce *Leishmania* infectante parmi les nombreuses aujourd'hui répertoriées. Ainsi, la manifestation d'une forme ou l'autre est étroitement liée à l'espèce ayant causé l'infection. Le genre *Leishmania* regroupe des eucaryotes unicellulaires appartenant au sous-phylum *Kinetoplasta* du phylum *Euglenozoa*. Il fait partie de l'ordre *Kinetoplastidae*, qui sont des protozoaires flagellés, une des branches les plus anciennes des eucaryotes, et plus spécifiquement de la famille *Trypanosomatidae* (Ivens et al., 2005; Roberts and Janovy, 2008). Cette famille comprend entre autres *Trypanosoma brucei*, agent causatif de la maladie du sommeil, et *Trypanosoma cruzi* agent responsable de la maladie de Chagas. Plusieurs espèces de *Leishmania* sont aujourd'hui répertoriées. On dénombre une trentaine d'espèces qui infectent les mammifères (SHAW and Jeffrey, 1994) et près d'une vingtaine qui infectent les reptiles (Stevens et al., 1998). Les espèces infectant les mammifères sont divisées en deux sous-genres, *Leishmania* et *Viannia*, selon certaines caractéristiques de leur développement chez l'insecte telles que la localisation dans l'intestin du vecteur, tandis que celles infectant les reptiles appartiennent au sous-genre *Sauroleishmania* (Stevens et al., 1998). Les sous-genres *Leishmania* et *Viannia* sont divisés en complexes qui sont eux-mêmes divisés en espèces. Le sous-genre *Leishmania* comprend des espèces du Nouveau Monde (comme *L. chagasi* et *L. mexicana*) et d'autres espèces de l'Ancien Monde telles que *L. infantum* et *L. major*. Le sous-genre *Viannia* comprend uniquement des espèces du Nouveau Monde comme *L. braziliensis* et *L. guyanensis*. La figure 3.3 donne une classification de certaines espèces de *Leishmania*. L'étude de ces parasites est compliquée par la diversité des formes cliniques de leishmanioses. La transmission de la leishmaniose se fait par piqûre d'insecte ailé infecté. Le cycle de vie du parasite est partagé entre l'insecte et le mammifère cible. Le parasite se retrouve sous deux formes principales au cours de son cycle parasitaire (*cf.*, figure 3.4) : la forme promastigote, qui se multiplie dans le système digestif de l'insecte (stade extra-cellulaire) et la forme amastigote qui se multiplie à l'intérieur des macrophages de l'hôte infecté (stade intracellulaire). Chez le vecteur, le parasite se différencie de la forme promastigote procyclique en la forme métacyclique infectieuse transmise à l'hôte. Le vecteur est un insecte diptère hématophage qui vit dans les zones tropicales et subtropicales du globe. Seules les phlébotomes femelles sont capables de transmettre la maladie en prenant un repas sanguin. Ces phlébotomes sont aussi vecteurs d'autres pathogènes : *Bartonella*, *Phlebovirus* et *Vesiculovirus*. Plusieurs espèces de phlébotomes sont capables de transmettre les différents parasites *Leishmania* chez les mammifères. La diversité de vecteurs et de parasites entraîne de grandes variations épidémiologiques et cliniques dans les zones infectées.

Leishmania à l'ère de la génomique

Le premier génome qui fut séquencé est le génome de la souche Friedlin de *L. major* en 2005. Son génome d'environ 35 Mb est distribué sur 36 chromosomes (Ivens et al., 2005). Les informations relatives aux génomes des différents pathogènes *Trypanosomatidae* séquencés sont

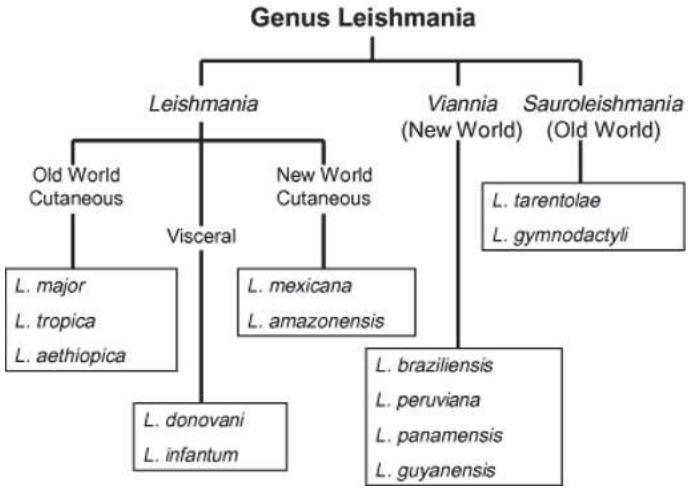


FIGURE 3.3.: Classification de quelques espèces du parasite *Leishmania* selon leur appartenance aux 3 sous-genres connus (Bates, 2007). Seules quelques unes des espèces de *Leishmania* connues sont indiquées ici.

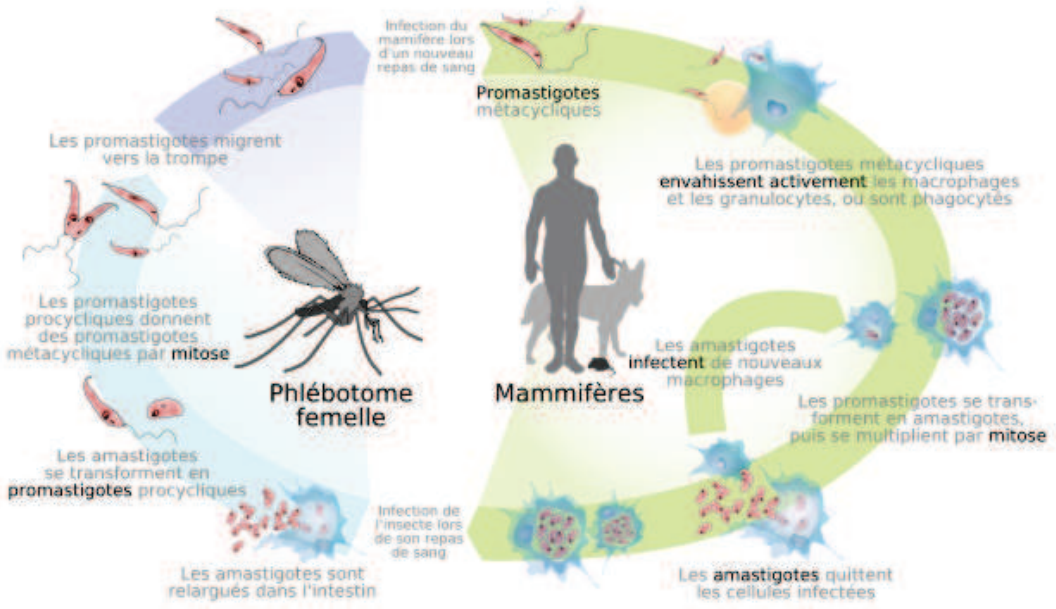


FIGURE 3.4.: Cycle de vie du parasite Leishmania

	L. major	*L. infantum*	*L. braziliensis*	*L. mexicana*	*L. tarentolae*
Nombre de chromosomes	36	36	35	34	36
Taille (bp)	32 816 678	32 134 935	32 005 207	32	30 440 719
Nombre de protéines	8412	8154	8153	8250	8452
Contenance en G+C	59.7%	59.3%	57.76%	59%	57.2%

TABLE 3.1.: Caractéristiques des différents génomes de *Leishmania* séquencés.

regroupées dans la banque de données TritrypDB [2]. Pour *Leishmania*, cinq espèces différentes ont été séquencées : *Leishmania major*, *infantum*, *braziliensis*, *mexicana* et *tarentolae* (une espèce modèle infectant les reptiles) (Raymond et al., 2012). Elles révèlent chacune un peu plus de 8000 gènes codant pour des protéines. *L. donovani* (Downing et al., 2012) a été également séquencée, mais les séquences ne sont pas toutes encore rassemblées sur TritrypDB. La table 3.1 résume les caractéristiques des différents génomes (Peacock et al., 2007; Raymond et al., 2012).

Ayant divergé très tôt au cours de l'évolution, les *Leishmania* ainsi que les autres *Kinetoplastidae* présentent de nombreuses particularités structurales et métaboliques. Par exemple, on compte dans chaque cellule une seule mitochondrie, contenant un organelle appelée *kinétoplaste* où se trouve l'ADN mitochondrial. La séquence des ARNms mitochondriaux est modifiée de façon post-transcriptionnelle par l'addition et/ou la délétion d'une grande quantité de résidus uridines, un processus nommé "édition des ARNs". Une autre particularité remarquable des parasites *Leishmania* réside dans le fait que leurs gènes soient organisés de façon très particulière. Ils font partie de larges unités polycistroniques (Ivens et al., 2005; Peacock et al., 2007; Myler and Fasel, 2007) et sont donc transcrits par bloc. Néanmoins, les délimitations de quelques polycistrons seulement sont connus. D'une manière générale, les mécanismes de transcription chez ces parasites restent peu connus.

Des études comparatives entre les différentes espèces de *Leishmania* ont montré que leurs génomes partagent de grandes similarités de contenu en gènes (Peacock et al., 2007). Une très grande conservation de la synthénie et de l'architecture est également observée. Le degré de conservation des acides aminés et des nucléotides est aussi très élevé (supérieur à 70%) notamment entre *L. major* and *L. infantum* où il s'élève à plus de 90% (Peacock et al., 2007). Bien qu'elle infecte uniquement les reptiles et ne cause aucune infection chez les humains, *L. tarentolae* partage plus de 90% de ses gènes avec les espèces infectant l'homme (Raymond et al., 2012). La grande majorité des gènes d'une espèce *Leishmania* ont des orthologues chez les autres espèces, à l'exception de quelques dizaines de gènes qui seraient spécifiques à chacune de ces espèces (Peacock et al., 2007). La majorité des gènes ont également des orthologues chez les autres espèces *Trypanosomatidae* (El-Sayed, 2005). Seul un ensemble formé par approximativement 1000 gènes ne possèdent pas d'orthologues dans ce clade et seraient uniques aux parasites *Leishmania* (Myler and Fasel, 2007; Peacock et al., 2007).

2. http ://tritrypdb.org/tritrypdb/

3.1.3. Paludisme et *Plasmodium*

Histoire et répartition géographique

Le paludisme ou malaria aurait infecté l'homme depuis le début de l'histoire de notre espèce. Plusieurs études ont établi le fait que la malaria a surgi chez nos premiers ancêtres en Afrique et qu'elle s'est étendue avec les migrations humaines dans les zones tropicales, les zones sub-tropicales et les régions tempérées de l'Ancien Monde. Elle s'est ensuite propagée au auau Monde avec les mouvements des missionnaires et des esclaves. On trouve des références au paludisme dans des textes indiens et chinois datant de 2000 ans avant J.C ainsi que dans des textes égyptiens datant de 1600 avant J.C. Au 5ème siècle avant J.C, les textes d'Hippocrate font référence au paludisme en Grèce. On trouve également des références au paludisme en Italie et partout dans l'Empire romain (où il est connu sous le nom de "fièvre romaine") (Cox, 2002). À cette époque, la maladie était associée aux marais, ce qui a donné le nom de paludisme (du grec "paludis"). En 1717, Giovanni Lancisi prouve que la maladie est transmise par les moustiques et introduit l'appellation malaria (mal-aria), tirée de l'Italien, qui signifie "mauvais air". La compréhension et l'étude de la cause de la malaria, n'a commencé qu'à la fin du 19ème siècle et la découverte de l'origine parasitaire de la malaria par Alphonse Levaran fut parmi les événements les plus marquants de l'histoire des maladies infectieuses. En 1897, Ronald Ross émit l'hypothèse de la transmission du parasite par piqûre d'insecte infecté du genre *Anophèle*. De nombreuses manifestations cliniques sont communes à tous les types de malaria mais leur intensité diffère selon le parasite infectant. Une fièvre croissante atteignant les 40 degrés marque le début de l'infection. Cette fièvre est accompagnée par des céphalées, des douleurs abdominales, des vomissements et des diarrhées dans certains cas. Chez l'enfant, elle provoque souvent de fortes anémies. La malaria causée par *P. falciparum* est très dangereuse et souvent mortelle si elle n'est pas prise en charge rapidement. Dans certains cas, les symptômes peuvent apparaître plusieurs semaines voire plusieurs mois après la piqûre, même si le sujet a quitté la zone impaludée. Plusieurs facteurs rentrent en jeu et sont décisifs quand à l'intensité et la gravité de la maladie : le type de parasite transmis, l'anophèle ayant piqué, le fond génétique et le système immunitaire de l'hôte ainsi que certaines conditions environnementales. La malaria demeure la maladie infectieuse la plus répandue dans le monde (Kooij et al., 2005). Du fait de son fort impact socio-économique et de son taux de mortalité important, elle est classée par l'OMS en tête des maladies à combattre prioritairement (Bousema and Drakeley, 2011). On dénombre approximativement 216 millions de cas de malaria, et 655.000 personnes en moyenne décèdent chaque année, selon les estimations de l'OMS de 2012. Selon les rapports de l'OMS, un peu moins de la moitié de la population mondiale vit sous le risque de contracter la maladie. Elle affecte le continent Asiatique, l'Amérique latine et certains pays européens comme l'illustre la carte de la figure 3.5. La zone la plus touchée est l'Afrique sub-saharienne qui est aussi fortement touchée par le VIH (Grobusch and Van Vugt, 2010). De nombreux cas sont recensés en Europe suite à des voyages en zones impaludées ou à la suite de piqûres de moustiques transportés dans les soutes des avions en provenance de zones impaludées (ce dernier cas est couramment appelée "Paludisme d'aéroport"). La plupart des morts causées par la malaria (90% à peu près) sont des enfants en Afrique. L'OMS estime qu'un enfant meurt chaque minute de la malaria (Rapport de l'OMS 2011). Plusieurs facteurs font de l'Afrique le continent le plus touché (85 à 90% des morts sont originaires d'Afrique subsaharienne). Ceci est relié à la pauvreté, la malnutrition, les conditions d'hygiène, la faiblesse des systèmes de surveillance et de prévention de la maladie, le fait que cette zone soit également fortement touchée par le VIH (l'infection par le VIH affaibli le système immunitaire et augmente donc le risque d'infection malarique (Abu-Raddad et al., 2006)), les conditions environnementales qui favorisent la multiplication des moustiques et donc le risque de transmission, ou encore la

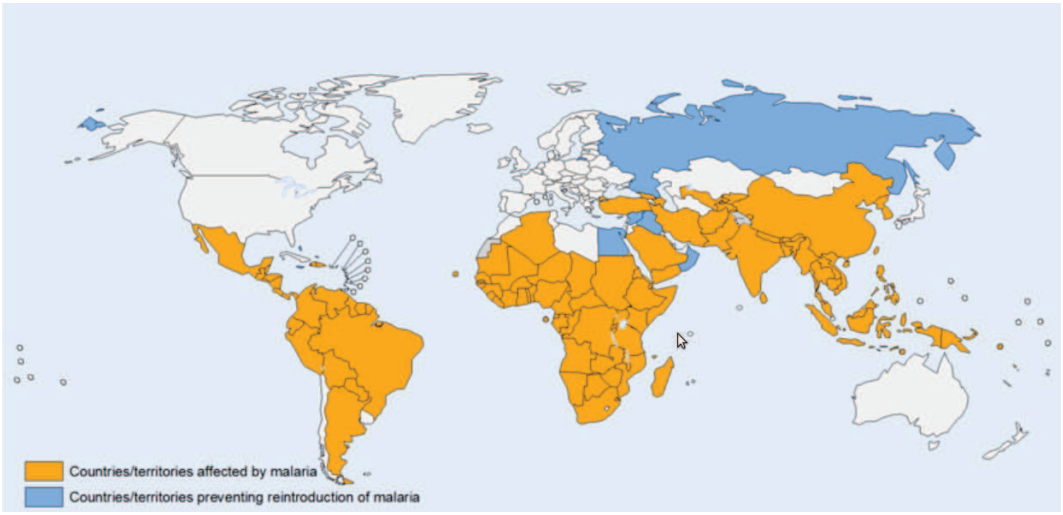


FIGURE 3.5.: Répartition géographique du paludisme dans le monde (http ://www.who.int/malaria/data/en/)

taille des anophèles trouvées en Afrique qui sont plus grandes et permettent un développement plus rapide du parasite.

Traitements et prévention de la malaria

Aucun vaccin n'est aujourd'hui disponible. Il existe cependant plusieurs molécules antipaludiques qui peuvent être utilisées pour prévenir la maladie lors de déplacements en zones endémiques. Ces médicaments à base d'artémisinine (ACT : Artemisinin-based Combination Therapy) sont également administrés pour traiter la malaria après infection. Le traitement de la malaria varie d'un endroit à l'autre en raison de l'émergence de résistances aux antipaludiques (Grobusch and Van Vugt, 2010). Le médicament le moins cher et le plus courant est la chloroquine. Cependant, les phénomènes de résistance à cette molécule se généralise en Afrique. La sulfadoxine-pyriméthamine (SP), médicament de deuxième recours, efficace mais plus cher, est lui aussi contré par l'émergence de résistances dans les zones fortement endémiques d'Afrique. Avant d'entamer un traitement, une analyse visant à identifier le parasite infectant est primordiale. Ce test n'étant pas toujours possible, notamment en Afrique, des traitements sont souvent administrés sans diagnostic complet.

Dans les régions sèches, les périodes de paludisme peuvent être prédites sans trop d'erreurs en utilisant les cartes de précipitations. L'OMS recommande plusieurs mesures préventives à prendre pour minimiser le risque d'infection en évitant notamment les piqûres de moustiques vecteurs (Reyburn, 2010) :

- utilisation de moustiquaires ;
- pulvérisation d'insecticides ;
- port de vêtements couvrant après le coucher du soleil, les moustiques vecteurs du paludisme piquent habituellement entre le coucher et le lever du soleil ;
- prise d'antipaludique lors des déplacements dans les zones endémiques.

3.1.4. Le parasite *Plasmodium*

Les espèces plasmodiales sont des Eukaryotes unicellulaires appartenant au sous règne des *Apicomplexes* regroupant plusieurs autres pathogènes. Les *Apicomplexes* font partie de la branche des *Chromalveolata* et plus précisément de celle de la sous-classe *Alveolata*. Plus de deux cents espèces plasmodiales sont répertoriées aujourd'hui. Au moins cinq d'entre elles causent des infections chez l'homme : *P. falciparum*, *P. vivax*, *P. malariae*, *P. ovale* et *P. knowlesi*. L'infection par cette dernière espèce déjà connue chez le singe, a été découverte récemment chez l'homme en Asie (White, 2008). Les infections par *P. falciparum* et *P. vivax* sont les plus fréquentes. *P. falciparum* cause la forme la plus dangereuse de la malaria et engendre la quasi totalité des morts liées à la maladie (Hay et al., 2004). Les autres espèces causent des formes moins graves, mais fortement persistantes. *P. ovale* n'est pas très fréquent et n'est trouvé que dans certaines zones d'Afrique inter-tropicale. La figure 3.6 donne une classification de différentes espèces du genre *Plasmodium* selon les espèces qu'elles infectent. Le cycle de vie de *Plasmodium* est très complexe (*cf.* la figure 3.7 pour *P. falciparum*) ce qui complique davantage la compréhension de la malaria et freine le développement de vaccins. Comme pour *Leishmania*, le cycle de vie fait intervenir deux acteurs : un hôte vertébré et un vecteur. Le cycle de vie est ainsi divisé en deux grandes étapes : la phase asexuée (schizogonie) chez l'hôte vertébré et celle sexuée (sporogonie) chez le vecteur (Liu et al., 2010; Bousema and Drakeley, 2011). *P. falciparum* est transmis principalement par la femelle hématophage d'un moustique du genre anophèle *Anopheles gambiae*.

Plasmodium à l'ère de la génomique

Le premier génome de *Plasmodium* fut entièrement séquencé en 2002 (Gardner et al., 2002). Le génome de *P. falciparum* 3D7 est composé de 22.8 megabases (Mb) distribuées sur 14 chromosomes. Pendant la même année, le génome du vecteur *Anopheles gambiae* a été également séquencé (Holt, 2002). Le séquençage de ces deux génomes marque un tournant dans la lutte contre la malaria. Les informations relatives aux génomes des différentes espèces de *Plasmodium* séquencées sont accessibles à partir de la base de données dédiée **PlasmoDB** [3] (Kissinger, 2002). Cette base regroupe l'ensemble des séquences génomiques et protéiques ainsi que diverses informations en rapport avec la fonction des protéines, leur structure, etc. Les génomes des différentes espèces de *Plasmodium* contiennent autours de 5000 gènes codant pour des protéines. Les génomes des parasites *Plasmodium* présentent plusieurs particularités. Ces génomes, et plus particulièrement *P. falciparum*, sont très riches en AT. Les A et T représentent 80% du génome *P. falciparum* par exemple et atteignent 90% par régions (Ross, 2001; Gardner et al., 2002). Ce taux est nettement supérieur à celui observé d'habitude chez la plupart des autres espèces. Cette atypicité impacte également la composition en acides aminés de ses protéines. Par exemple, 3 acides aminés (I,K,N) codent pour plus du tiers des séquences protéiques. De plus, les protéines de *Plasmodium* sont en moyenne 20% plus longues que les protéines homologues connues chez les autres organismes (Pizzi and Frontali, 2001). Quand un alignement multiple est possible, cette différence de taille semble provenir de la présence de longues zones d'insertion sur les protéines de *Plasmodium* (Wootton and Federhen, 1993). Ces particularités compliquent notablement la reconnaissance d'homologie et ainsi le transfert d'annotations.

3. http ://plasmodb.org

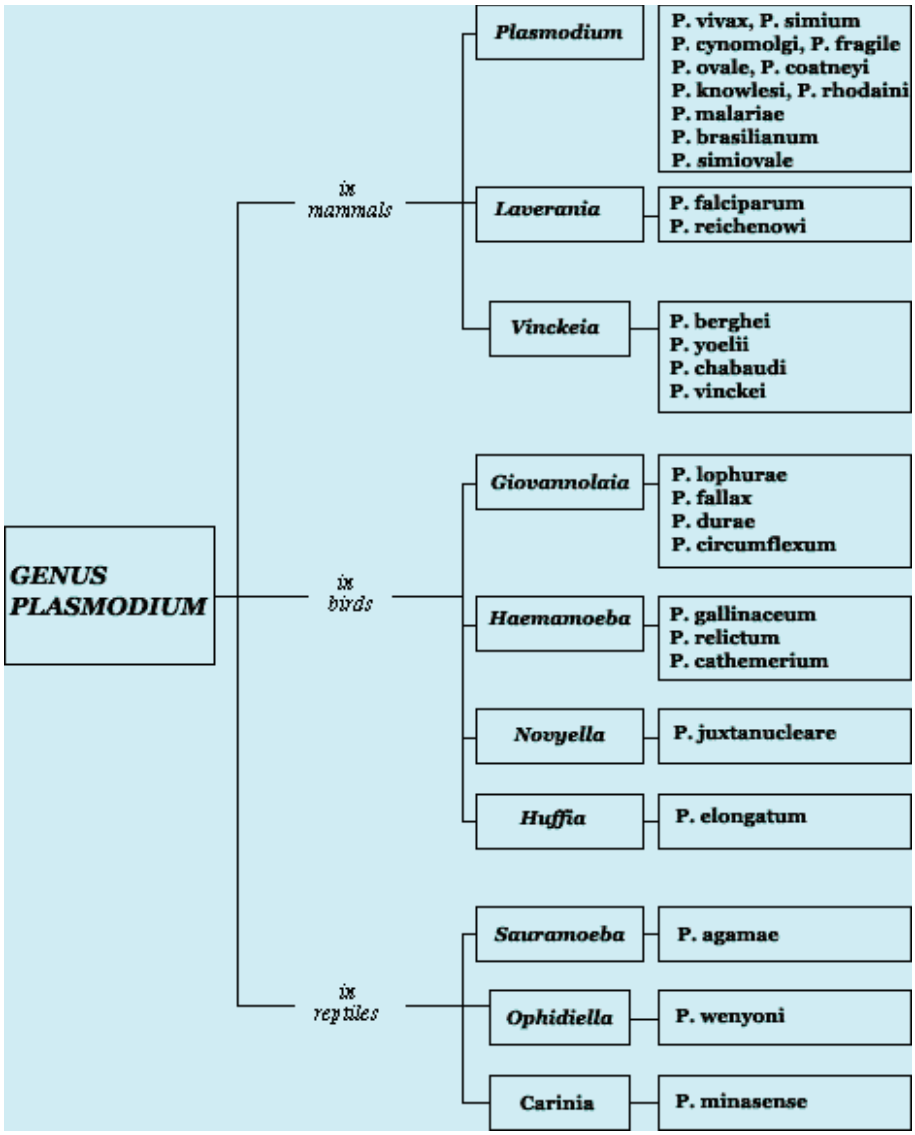


FIGURE 3.6.: Classification des différentes espèces de *Plasmodium* selon les espèces qu'elles infectent. Cette figure est extraite de http ://www.impact-malaria.com

3.2. Annotation fonctionnelle des pathogènes Eukaryotes

3.2.1. Annotation fonctionnelle de *Leishmania* et *Plasmodium*

L'annotation fonctionnelle des protéines de *Leishmania* et *Plasmodium* en particulier et celles des autres pathogènes Eukaryotes d'une manière générale est loin d'être triviale. Les méthodes classiques basées sur le transfert d'annotations par recherche d'homologie présentent

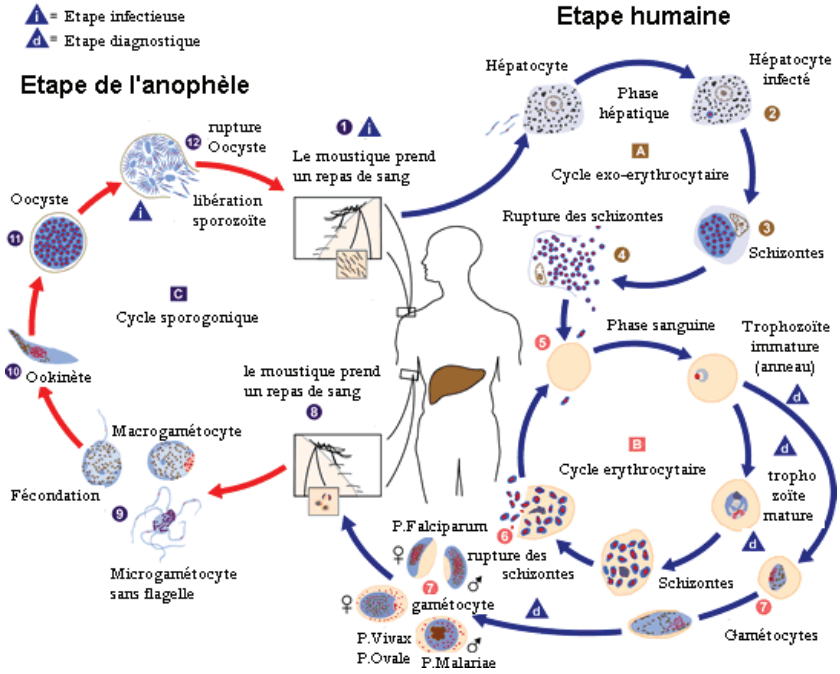


FIGURE 3.7.: Cycle de vie du parasite *Plasmodium falciparum* (http ://www.revuedesante.com/Article/paludisme-113.html)

plusieurs limites, pour l'étude de ces génomes divergents (Bréhélin et al., 2010; Ghouila et al., 2010). A titre d'exemple, une recherche des annotations des différentes espèces *Leishmania* et *Plasmodium* disponibles sur TritryDB [4] et PlasmoDB [5] montre que ces espèces demeurent mal annotées :

- 5335 protéines, représentant 63% des protéines de *L. major*, sont des protéines hypothétiques dont le rôle n'est pas encore connu. La majorité des ces protéines sont conservées entre les différentes espèces *Leishmania*.
- 2455 protéines, représentant 45% des protéines de *P. falciparum*, ne sont pas caractérisées (elles sont annotées "unknown function"). La majorité (90%) de ces protéines sont conservées chez les autres espèces plasmodiales.
- Plus de 50% des protéines *L. major* et *P. falciparum* ne possèdent pas d'orthologues chez les espèces modèles.
- Un peu moins de la moitié des protéines de ces deux espèces ne possèdent aucun domaine protéique connu.

3.2.2. Détection des domaines sur les protéines divergentes

Une interrogation de la base de données Uniprot indique que les informations relatives aux différentes espèces Eukaryotes ne sont pas équilibrées sur l'arbre des espèces (voir Figure 3.8) :

4. http ://tritrypdb.org/tritrypdb/
5. http ://plasmodb.org

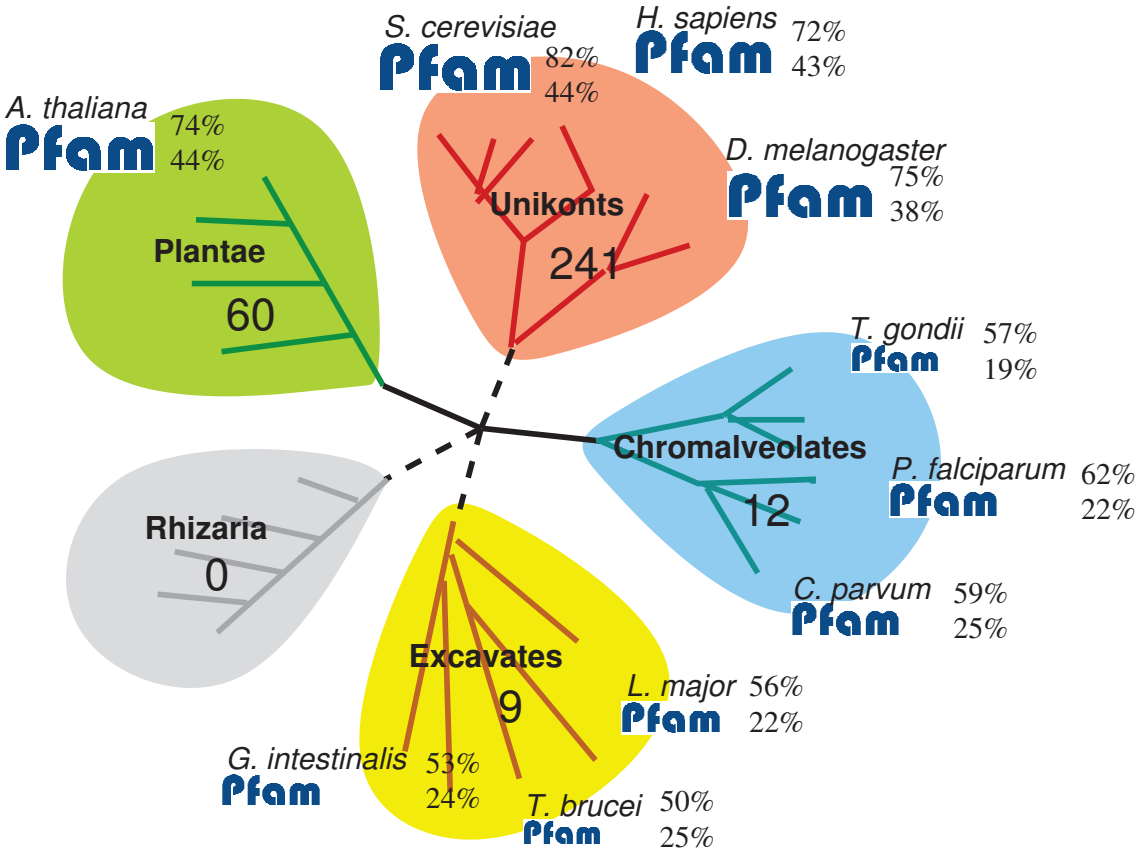


FIGURE 3.8.: Nombre de génomes séquencés et couverture en domaines dans l'arbre des Eukaryotes. La figure rapporte le nombre de génomes entièrement séquencés dans chacun des 5 super-groupes de l'arbre des Eukaryotes (Keeling et al., 2005). Pour chaque groupe sont indiqués quelques exemples de génomes séquencés, avec deux statiques concernant les domaines Pfam (version 26.0) : la proportion de protéines où au moins un domaine Pfam a été identifié (sur la base des seuils recommandés par PFAM) (en haut) et la proportion d'acides aminés couverte par un domaine Pfam (en bas).

- la plupart des organismes séquencés appartiennent aux super-groupes des Unikonts (241) et des Plantes (60)
- la couverture en domaines des espèces pathogènes est nettement plus faible que celle de la plupart des Eukaryotes séquencés appartenant au groupe des plantes ou au super-groupe des Unikont. La proportion de protéines ayant au moins un domaine Pfam connu est souvent au-dessus de 70% chez les Unikonts et les plantes, alors que ce taux se trouve entre 50% et 60% dans les autres groupes. Il en va de même pour la proportion d'acides aminés couverts par un domaine Pfam. Elle est souvent au-dessus de 40% chez les plantes et les Unikonts mais autour de 22% dans les autres super groupes.

Deux raisons peuvent expliquer cela :

- l'existence de gènes et de domaines qui seraient spécifiques aux parasites et à leurs modes de vie. Plusieurs études ont montré que ces parasites, apparus tôt sur l'échelle évolutive, ont développé une biologie complexe due à des différenciations successives au cours de leurs cycles de vie respectifs. Ces différenciations sont intimement liées à l'interaction et la transmission des parasites de l'hôte cellulaire au vecteur, et inversement du vecteur à l'hôte (Julius, 2002; Ochman and Moran, 2001).
- la grande distance phylogénétique entre ces espèces et les espèces eukaryotes modèles classiques (*cf.* figure 3.8) qui rend difficile la détection d'homologie (Baldauf, 2003). Durant l'évolution, les gènes de ces parasites ont en effet subi plusieurs changements. Ils ont pu par ailleurs adopter de nouvelles fonctions en réponse à une pression de sélection. Ces changements subis engendrent des altérations des séquences génomiques et protéiques des parasites par rapport à celles des autres organismes classiques. Les séquences de leurs protéines ayant divergé loin des séquences des organismes modèles ayant servi à la construction des HMM profils, il est difficile d'obtenir des scores au dessus des seuils de scores recommandés par Pfam sur ces protéines.

Nous nous intéressons dans cette thèse à l'amélioration de la stratégie de reconnaissance des domaines protéiques chez les organismes divergents d'une manière générale pour pallier aux limites des méthodes classiques.

Partie III.

Contributions

4. EuPathDomains : la base de données des domaines divergents chez les pathogènes Eukaryotes

Comme nous l'avons discuté au chapitre précédent, les génomes des pathogènes Eukaryotes sont souvent très mal annotés. La moitié de leurs protéines, à peu près, ne possèdent aucune annotation GO et il en va de même pour les domaines protéiques. On s'est intéressés dans le cadre de cette thèse à enrichir ces annotations. Développée à l'origine pour *Plasmodium falciparum*, la méthode CODD a été appliquée à la majorité des pathogènes eukaryotes séquencés. Ce travail a mené à la création d'une base de données de domaines protéiques dédiée à ces organismes, la base EuPathdomains. EuPathDomains améliore la couverture en domaines de l'ensemble des organismes étudiés et associe de nouvelles annotations GO à de nombreuses protéines.

Sommaire

4.1. Introduction 58
4.2. CODD : la stratégie de détection de domaines par Co-occurrence 58
4.2.1. Construction de la *CDP* 60
4.2.2. Identification des domaines connus et recherche des domaines potentiels 60
4.2.3. Processus de certification par co-occurrence 61
4.2.4. Estimation du taux d'erreur 62
4.2.5. Inférences des annotations Gene Ontology 62
4.3. Pathogènes étudiés 64
4.4. La base de données EupathDomains 65
4.4.1. Mise en place 65
4.4.2. Interface et contenu 65
4.4.3. Étude des caractéristiques des nouvelles prédictions 69
4.4.4. *Leishmania* comme étude de cas 72
4.4.5. Les prédictions communes à tous les pathogènes étudiés . . . 73
4.5. Nouvelle version 2.0 de EupathDomains 74
4.6. Conclusion 76

4.1. Introduction

Comme on l'a vu au chapitre précédent (chapitre 3), les protéines des pathogènes Eukaryotes sont généralement mal annotées en comparaison à d'autres génomes comme la levure, la drosophile, l'homme, *etc.* Tandis que la couverture en domaines Pfam de ces organismes tourne autour de 73% en moyenne (*cf.* table 4.1 et figure 3.8), cette proportion varie pour les pathogènes eukaryotes de 38% (pour *Trypanosoma brucei*), à 53% (pour *Plasmodium falciparum*). Bien que l'on puisse expliquer cette situation par l'existence de gènes uniques à ces organismes ou spécifiques à un style de vie parasitaire, ceci est renforcé par la grande distance évolutive entre ces organismes et les organismes utilisés pour la construction des modèles Pfam, qui rend plus difficile la détection d'homologie (Wardl et al., 2004). Pour essayer de remédier à ces limites, une méthode nommée CODD (Terrapon et al., 2009) a été proposée dans le cadre de la thèse de Nicolas Terrapon. Elle relâche dans un premier temps les seuils de scores stringent proposés par Pfam pour augmenter la sensibilité de détection de domaines par les HMMs. Cette méthode a été initialement développée et testée pour la recherche des domaines protéiques chez les espèces *Plasmodium*. Les premières expérimentations ont mis en exergue l'intérêt de l'utilisation de cette approche pour l'amélioration de la détection des domaines protéiques sur l'ensemble des protéines de cette espèce (Terrapon et al., 2009). L'objectif de mes premiers travaux de thèse est d'adapter cette approche à l'ensemble des pathogènes de la base de données EuPathDB [1] (Aurrecoechea, 2010) (trois espèces de *Leishmania*, trois espèces de *Plasmodium*, *Toxoplasma gondii*, *Cryptospordium parvum*, *Giardia lamblia* et *T. brucei*) et de créer une base de données de domaines qui leur soit dédiée.

Nous commençons ce chapitre par une présentation de la méthode CODD et de ses principales étapes. Nous présentons dans la deuxième partie EuPathDomains, la base de données étendue de domaines protéiques connus et nouvellement identifiés par CODD sur l'ensemble des pathogènes Eukayrotes étudiés. Enfin, une étude des prédictions effectuées chez les espèces *Leishmania* est présentée. En moyenne, EuPathDomains augmente de 10% le nombre de domaines de chaque espèce (avec un FDR $\leq$ 20%). Nous avons découvert plusieurs nouveaux types de domaines protéiques, permettant un enrichissement de 10% à 16% du nombre total de familles de domaines connues chez chacune des espèces. Certains de ces domaines sont prédits sur des protéines qui n'avaient aucun domaine connu. De plus, les domaines nouvellement prédits permettent d'associer de nouvelles annotations GO pour environ 15% des protéines dans chacun des organismes étudiés. Ces annotations sont également intégrées dans EuPathDomains. La première version de la base EuPathDomains était basée sur la version 23 de Pfam et la version 2 du logiciel HMMER. En août 2013, est sortie une seconde version d'EuPathDomains basée sur la version 26 de Pfam et HMMER 3. La section 4.5 présente cette nouvelle version, et donne quelques statistiques relatives à son contenu.

4.2. CODD : la stratégie de détection de domaines par Co-occurrence

Le principe de la méthode CODD est le suivant. Dans un premier temps CODD relâche les seuils de score stringent utilisés par Pfam pour augmenter la sensibilité des HMMs et permettre la découverte de nouveaux domaines potentiels. Ceci conduit effectivement à la détection d'un grand nombre de domaines, mais au prix d'une proportion élevée de faux positifs. Afin de limiter le nombre de faux positifs parmi les domaines ainsi détectés, cette

1. http ://eupathdb.org/eupathdb/

Organisme	Nombre de protéines	Familles de domaines Pfam	Couverture Pfam
Organismes eukaryotes classiques			
A. gambiae	12347	2991	74%
A. thaliana	34517	3125	74%
C. elegans	22637	2953	65%
D. melanogaster	16224	3129	72%
D. rerio	31884	3384	84%
H. sapiens	40252	3914	68%
S. cerevisae	5862	2369	76%
Pathogènes Eukaryotes étudiés dans EuPathDomains			
P. falciparum	5460	1429	53%
P. vivax	5432	1415	50%
P. yoelii	7724	1313	42%
L. major	8406	1607	49%
L. infantum	8216	1607	49%
L. brazelinsis	8310	1587	48%
T. gondii	8102	1689	43%
G. lamblia	4889	845	49%
C. parvum	3805	1123	51%
T. brucei	9895	1554	38%

TABLE 4.1.: Nombre de familles de domaines Pfam et couverture des protéines en domaines chez plusieurs Eukaryote. La colonne "familles de domaines Pfam" donne le nombre de familles de domaines Pfam différentes connues pour un organisme donné. La colonne "Couverture Pfam" donne le pourcentage de protéines pour lesquelles au moins un domaine Pfam est connu.

première phase est suivie d'une étape de filtrage des résultats. Pour cela, CODD utilise la tendance (observée dans les trois royaumes du vivant) de beaucoup de familles de domaines d'apparaître préférentiellement avec quelques autres familles bien particulières (Apic, 2001; Beaussart et al., 2007). La présence d'un domaine dans une protéine peut donc être un indice fort de la présence d'un autre domaine. CODD utilise ces informations pour confirmer la présence de certains domaines au dessous des seuils recommandés par Pfam. Cinq étapes principales illustrées par la figure 4.1 constituent la procédure :

- Identification des paires de domaines qui ont tendance à apparaître préférentiellement ensemble. Ces paires forment une liste appelée CDP (pour Conditionaly Dependent Pairs). Elle est déterminée à partir des paires de domaines observées dans l'ensemble des protéines UniProt de tous les organismes.
- Identification des domaines potentiels sur les protéines de l'organisme cible. Ces domaines sont déterminés par les HMM de Pfam en relâchant les seuils de scores recommandés par Pfam.
- Utilisation de la CDP pour certifier la présence de domaines potentiels en fonction de la présence d'autres domaines sur la même protéine.
- Estimation du taux de fausses découvertes (FDR) associé aux nouvelles prédictions permettant d'évaluer le taux de faux positifs par une procédure de "shuffling".
- Identification des annotations GO apportées par les nouveaux domaines prédits.

Ces cinq étapes sont plus amplement détaillées dans les sections suivantes.

4.2.1. Construction de la CDP

La première étape de cette approche consiste à identifier les paires de domaines souvent co-occurrentes. Ces paires de domaines doivent révéler des dépendances conditionnelles entre un domaine Pfam et un autre domaine Interpro (Pfam ou non-Pfam), *i.e.*, la présence du domaine Interpro doit être un indice fort de la présence du domaine Pfam. Les paires de domaines conditionnellement dépendantes sont identifiées grâce à un test statistique à partir de l'ensemble des paires de domaines observées sur l'ensemble des protéines Uniprot (tout organismes confondus). Pour effectuer ce test, toutes les paires possibles de domaines sont énumérées. Pour chacune, nous comptons le nombre de protéines où les deux domaines sont présents et celles où au moins un des deux domaines est présent. Un test exact de Fisher est alors utilisé pour évaluer si la co-occurrence des deux domaines est significativement plus élevée qu'attendue par hasard. Toutes les paires ayant une P-valeur $< 1\%$ sont considérées comme fortement corrélées et sont ajoutées à la CDP. La liste calculée ainsi est utilisée comme une liste de référence pour confirmer les prédictions de domaines dans la suite du processus.

4.2.2. Identification des domaines connus et recherche des domaines potentiels

Les séquences protéiques des organismes étudiés sont comparées à la librairie complète des HMM profils de Pfam à l'aide de la méthode *hmmsearch* de l'outil HMMER. Comme nous avons vu dans le chapitre précédent (*cf.* section 2.2.4), cette recherche renvoie, pour chaque séquence, les différentes occurences de domaines détectées, accompagnées de leur score. Les occurences au dessus du seuil recommandé par Pfam sont considérées dans la suite comme les domaines "connus" des protéines. Un nouveau seuil de score moins stringent est ensuite choisi, et les occurences dont le score est compris entre ce nouveau seuil et le seuil de Pfam sont considérées comme les domaines "potentiels" de la protéine. Certains de ces domaines potentiels chevauchent des domaines connus : ces domaines potentiels sont supprimés. Lorsque deux

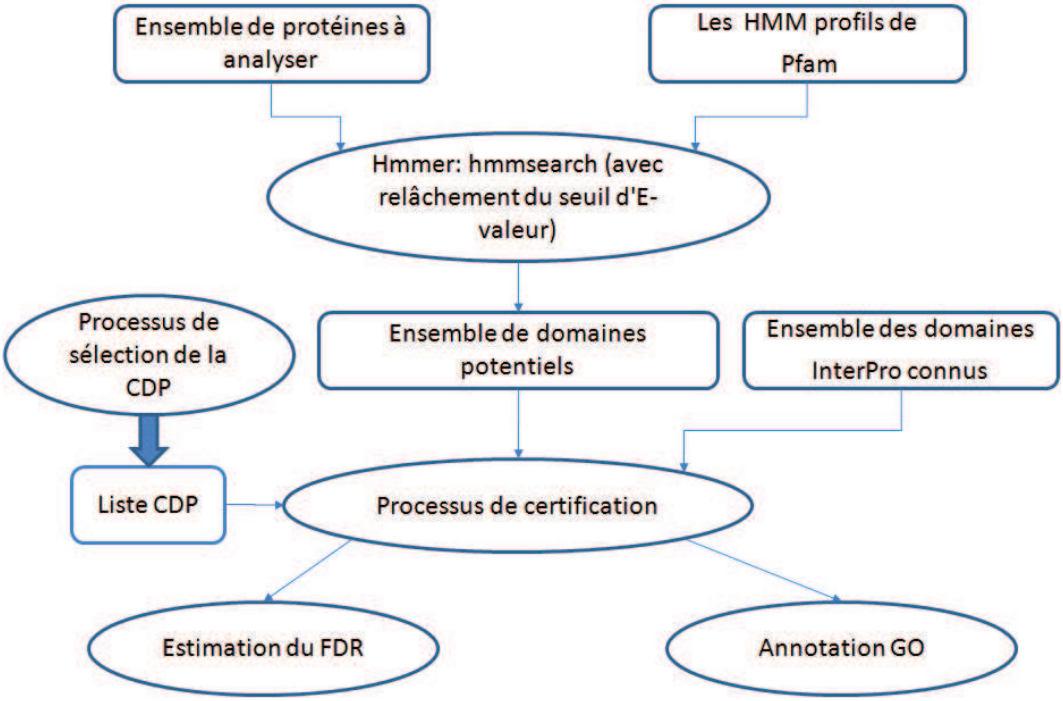


FIGURE 4.1.: Les différentes étapes de prédictions de domaines et annotations GO par la méthode CODD

domaines potentiels sont chevauchants, celui avec le score le plus bas (le moins vraisemblable) est lui aussi supprimé. À coté des domaines Pfam, nous considérons également l'ensemble des domaines Interpro non-Pfam (*cf.* section 1.4.2) connus sur chaque protéine. Ces domaines sont tirés des bases de données spécifiques à chaque organisme. Contrairement aux domaines Pfam connus, ils ne sont pas utilisés pour éliminer d'éventuels domaines potentiels chevauchants, mais seulement pour certifier les domaines potentiels.

4.2.3. Processus de certification par co-occurrence

Afin de sélectionner parmi l'ensemble des domaines potentiels détectés ceux dont la présence est la plus vraisemblable, chaque domaine potentiel est évalué au regard des autres domaines de la protéine. Trois types de certifications ont été considérés. Dans le premier type, les domaines Pfam connus de la protéine sont énumérés, et si le domaine potentiel forme avec un de ces domaines une paire identifiée dans la CDP, il est considéré comme "certifié". Une solution alternative consiste à procéder à la certification en utilisant les domaines connus non Pfam. Ceci nous permet d'augmenter le nombre de certifications. Cependant, en raison de l'hétérogénéité de la base Interpro, les certifications réalisées de cette façon sont parfois de moins bonne qualité que celles réalisées par les domaines Pfam. Les deux premières solutions ne permettent de certifier des domaines que pour les protéines ayant au moins un domaine connu. Une troisième solution consiste à certifier un domaine potentiel par un autre domaine potentiel de la protéine. Ainsi, toutes les paires de domaines potentiels sont énumérées, si la paire appartient à la CDP, les deux domaines sont certifiés. Cette approche mène évidemment à un plus grand nombre de faux positifs, mais nous allons voir qu'il existe une méthode efficace pour contrôler ce nombre

d'erreur.

4.2.4. Estimation du taux d'erreur

La procédure de certification permet de certifier un certain nombre de nouveaux domaines parmi tous les domaines potentiels identifiés sur les différentes protéines de l'organisme étudié. Une estimation de la proportion de faux positifs parmi ces nouveaux domaines prédits est nécessaire. Pour cela, on estime l'espérance du nombre de nouveaux domaines que cette approche certifierait sous l'hypothèse H_0 où tous les domaines potentiels sont prédits de manière aléatoire. Une procédure de ré-échantillonnage ("shuffling" en anglais) est mise en place. Elle considère l'ensemble des domaines potentiels détectés sur toutes les protéines et les redistribue aléatoirement sur les différentes protéines. La procédure de certification est appliquée aux domaines aléatoires et le nombre de domaines certifiés est calculé. La procédure est répétée plusieurs fois (typiquement 1000 fois) pour obtenir une bonne estimation de l'espérance. Le nombre de domaines certifiés "par hasard" est alors utilisé pour estimer la proportion de faux positifs parmi les domaines certifiés grâce à la formule :

$$FDR = \frac{\texttt{Espérance du nombre de certification sous H_0}}{\texttt{Nombre de domaines certifiés réellement}} \tag{4.1}$$

Plus le seuil de E-valeur utilisé pour identifier les domaines potentiels est haut, plus le nombre de domaines certifiés est important mais au prix de taux d'erreur également plus élevés. Cela est illustré sur la Figure 4.2. Nous constatons également que plus on augmente le seuil de E-valeur, plus la taille de l'ensemble des domaines potentiels croît, plus le nombre de certifications sur les données réelles et sous H_0 augmente.

4.2.5. Inférences des annotations Gene Ontology

À partir des nouvelles prédictions, de nouvelles annotations de la Gene Ontology (GO) peuvent être déduites. Comme nous l'avons introduit dans le chapitre 1, il s'agit d'un vocabulaire contrôlé et structuré décrivant trois aspects fonctionnels : fonction moléculaire, processus biologique et localisation cellulaire. Le consortium Interpro a établi des correspondances entre domaines et termes GO. Cette association est effectuée en grande partie manuellement, mais la politique générale est d'associer un terme à un domaine si l'ensemble des protéines annotées dans lequel on connaît ce domaine partage le terme (Hunter et al., 2009). Grâce à ces associations, lorsqu'un domaine est identifié sur une protéine, tous les termes associés à ce dernier peuvent être transférés à la protéine. Cependant, les fonctions des protéines résultent souvent de la combinaison de plusieurs domaines (Forslund and Sonnhammer, 2008). Sur le même principe, nous avons donc créé une liste d'associations entre combinaisons de domaines et termes GO. Pour cela, nous avons énuméré toutes les combinaisons de domaines observées dans l'ensemble des protéines de Swiss-Prot. Nous considérons ensuite les combinaisons présentes sur au moins dix protéines. Si les protéines exhibant une telle combinaison ont un ou plusieurs termes GO en commun, ce(s) terme(s) est (sont) transféré(s) à la combinaison de domaines. Avec les domaines Pfam 23, 2235 combinaisons de domaines Pfam ont pu être associées avec au moins un terme GO : 2115 paires, 119 triplets de domaines et 1 quadruplet. Ces différentes associations sont disponibles en ligne [2].

2. http ://www.atgc-montpellier.fr/EuPathDomains/dddb/pfamCombinations2go.html

Organisme		FDR ≤ 10%			
	Type certif.	Pfam	Interp.	Pot.	Tous
L. major	Certif. dom.	231	260	165	423 (8%)
	Nouv. interp.	187	200	143	341
	Nouv. types dom.	76	80	48	118 (7%)
P. falciparum	Certif. dom.	250	145	50	320 (9%)
	Nouv. interp.	192	106	40	241
	Nouv. types dom.	64	36	14	80 (6%)
T. gondii	Certif. dom.	340	169	126	466 (8%)
	Nouv. interp.	284	138	107	392
	Nouv. types dom.	90	49	39	119 (7%)
C. parvum	Certif. dom.	292	246	123	417 (21%)
	Nouv. interp.	251	187	103	342
	Nouv. types dom.	109	78	42	135 (12%)
G. lamblia	Certif. dom.	126	63	18	144 (5%)
	Nouv. interp.	108	54	16	126
	Nouv. types dom.	44	23	13	52 (6%)
T. brucei	Certif. dom.	348	285	120	515 (9%)
	Nouv. interp.	285	201	109	418
	Nouv. types dom.	73	48	29	99 (6%)

TABLE 4.2.: Prédictions sous un FDR à 10% chez les différents organismes. “Type certif.” indique le type de certification : par les domaines Pfam connus (“Pfam”), par les domaines InterPro (non-Pfam) connus (“Interp.”), ou par les domaines potentiels (“Pot.”) ; la colonne ”Tous” indique le nombre total de domaines certifiés en cumulant les différents types de certification. “Certif. dom.” indique le nombre de nouveaux domaines certifiés ; “New Interp.” indique le nombre de certifications permettant d’identifier une nouvelle entrée InterPro pour la protéine en question ; “Nouv. types dom.” indique le nombre de nouveaux types de domaines qui n’ont jamais été identifiés chez les protéines du génome. Les chiffres entre parenthèses indiquent les proportions de nouveaux domaines ou nouveaux types de domaines par rapport à ceux déjà connus.

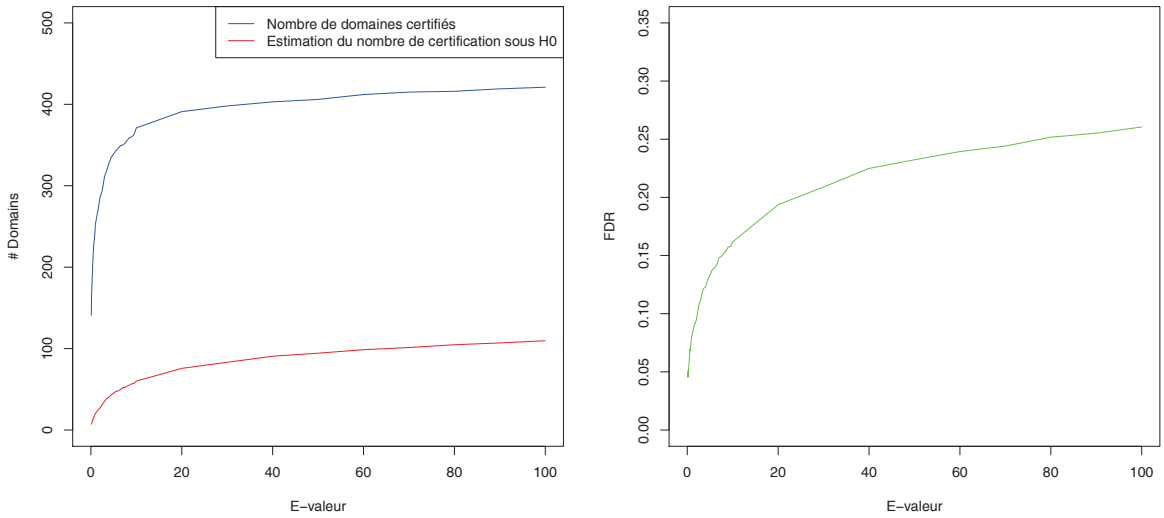


FIGURE 4.2.: Évolution du nombre de certification et de l'estimation du nombre de domaines certifiés sous H_0 sur les protéines de *P. falciparum* en fonction de la E-valeur (à gauche). Les FDR correspondants calculés par la formule 4.1 sont représentés en vert (à droite)

4.3. Pathogènes étudiés

Nous avons sélectionné pour cette étude un ensemble de pathogènes ayant comme principal hôte l'humain en choisissant un organisme dans chacun des différents phylums représentés dans la base de données EupathDB (Aurrecoechea, 2010). À côté de *L. major* et *P. falciparum*, présentés en détail au chapitre précédent (*cf.* section 3.1), nous avons inclus dans l'étude deux autres *apicomplexes* (*T. gondii* et *C. parvum*), un autre *trypanosome* (*T. brucei*) et un organisme appartenant à un phylum distinct des précédents (*G. lamblia*). Afin de pouvoir effectuer des études comparatives entre espèces proches, nous avons appliqué cette méthode sur deux autres espèces *Leishmanies* (*L. infantum* et *L. braziliensis*) et deux autres espèces plasmodiales (*P. vivax* et *P. yoeli*).

Toxoplasma gondii *Toxoplasma gondii* est un parasite intracellulaire unicellulaire qui cause une maladie mondialement répandue et connue sous le nom de toxoplasmose. Le chat est l'hôte principal de *T. gondii* mais le parasite peut être porté par beaucoup d'animaux à sang chaud y compris l'être humain. La toxoplasmose présente un risque sérieux chez les femmes enceintes et les sujets ayant un système de défense immunitaire affaibli. Elle est bénigne, voire asymptomatique dans la majorité des cas pour les sujets immunocompétents (Blader and Saeij, 2009). Le génome de *T. gondii* compte 8102 gènes codant pour des protéines, environ 57% d'entre eux n'ont aucun domaine Pfam connu.

Cryptosporidium parvum Le parasite *Cryptosporidium parvum* est un protozoaire qui provoque une maladie diarrhéique appelée cryptosporidiose affectant le tube digestif de certains vertébrés dont l'homme. Il est transmis à l'hôte via la voie fécale ou orale. Sur les sujets immunodéprimés (malades du sida, personnes greffées, *etc.*), la cryptosporidiose peut avoir des effets mortels. La cryptosporidiose est observée partout dans le monde et notamment dans les pays en voie de développement, $8-19\%$ des maladies diarrhéiques sont attribuées à *Cryptosporidium* (Gatei et al., 2006). Son génome est de petite taille (9.1 Mo) avec une organisation simple contenant seulement 3805 gènes codant pour des protéines (Abrahamsen and Templeton, 2004). Au total, 1123 domaines Pfam différents sont connus sur ces protéines couvrant uniquement 51% du génome entier.

Trypanosoma brucei *Trypanosoma brucei* est un parasite du groupe des kinétoplastidés infectant les mammifères. Il est transmis par des morsures d'une mouche généralement appelée mouche "tsé-tsé". Il cause la trypanosomiase africaine, également connue sous le nom de "maladie du sommeil". Son adaptation aux différents hôtes (une stratégie commune de tous les parasites) engendre des changements complexes pendant son cycle de vie. Le génome de *T. brucei* est composé de 11 chromosomes et contient presque 9800 gènes codant pour des protéines. Environ 62% de ses gènes n'ont aucun domaine Pfam connu.

4.4. La base de données EupathDomains

4.4.1. Mise en place

Dans une première étape, les séquences protéiques de chaque espèce étudiée ont été comparées à l'ensemble des HMM profils de la base Pfam. Deux seuils de E-valeur (typiquement 1 et 50) ont été utilisés pour identifier les domaines potentiels. La procédure de certification a été ensuite appliquée sur les deux ensembles de domaines avec les trois types de certification (Pfam connus, Interpro non-Pfam connus et Pfam potentiels). Les informations sur les domaines Interpro connus (Pfam et non Pfam) ont été téléchargées des bases de données spécifiques de chaque organisme. Les FDR associés aux différentes certifications réalisées ont été estimés. Toutes les certifications avec des FDR au-dessous de 20% ont été intégrées dans la base de données EuPathDomains. Pour chaque domaine nouvellement certifié, les annotations GO qui peuvent être déduites de ce domaine, seul ou en association avec d'autres domaines connus ou nouvellement certifiés, ont été aussi intégrées dans la base de données.

4.4.2. Interface et contenu

EuPathDomains (Ghouila et al., 2010), la base de données des domaines divergents chez les pathogènes eukaryotes, est librement accessible à l'adresse suivante : `http://www.atgc-montpellier.fr/EuPathDomains/`. Elle est dotée d'une interface simple permettant une navigation facile. La base de données peut être interrogée selon plusieurs critères : par noms de protéines, identifiants de domaines Pfam ou Interpro, ou termes GO. On a également la possibilité de limiter la recherche à un organisme d'intérêt ou un taxon (Leishmanies, Apicomplexe,...). Il est aussi possible de préciser un seuil de FDR souhaité. Les résultats d'une requête incluent les domaines déjà connus et ceux nouvellement prédits avec les FDRs qui leurs sont associés.

La figure 4.3 illustre un exemple de résultat quand on interroge la base pour la protéine *LmjF.36.4590* de *L. major*. Le résultat contient les domaines connus de la protéine ainsi que les nouveaux domaines prédits, accompagnés des domaines ayant servi à les certifier. Les termes GO associés à la protéine apparaissent également. Ici, quelques termes sont déjà connus, tel

Organisme		FDR $\leq$ 20%			
	Type certif.	Pfam	Interp.	Pot.	Tous
L. major	Certif. dom.	385	417	316	719 (14%)
	Nouv. interp.	326	333	279	607
	Nouv. types dom.	104	106	98	184 (11%)
P. falciparum	Certif. dom.	348	363	130	555 (15%)
	Nouv. interp.	284	288	114	451
	Nouv. types dom.	97	92	36	144 (10%)
T. gondii	Certif. dom.	436	552	255	776 (13%)
	Nouv. interp.	379	475	228	671
	Nouv. types dom.	118	148	67	198 (12%)
C. parvum	Certif. dom.	344	379	169	554 (28%)
	Nouv. interp.	302	313	144	473
	Nouv. types dom.	131	127	60	182 (16%)
G. lamblia	Certif. dom.	303	279	74	386 (13%)
	Nouv. interp.	279	146	67	357
	Nouv. types dom.	75	42	25	90 (10%)
T. brucei	Certif. dom.	513	616	270	908 (16%)
	Nouv. interp.	442	501	243	768
	Nouv. types dom.	111	118	60	176 (11%)

TABLE 4.3.: Prédictions sous un FDR de 20% chez les organismes étudiés. "Type certif." indique le type de certification : par les domaines Pfam connus ("Pfam"), les Interpro (non-Pfam) connus ("Interp."), ou par les domaines potentiels ("Pot.") ; la colonne "Tous" indique le nombre total de domaines certifiés en cumulant les différents types de certification. "Dom. Certif." indique le nombre de nouveaux domaines certifiés ; "Nouv. interp." indique le nombre de certifications permettant d'attribuer une nouvelle entrée Interpro à la protéine ; "Nouv. types dom." indique le nombre de types de domaines qui n'étaient pas connus chez les différentes protéines.

que : ”zinc ion binding” et ”intracellular”. D'autres annotations sont nouvelles telles que les termes ”regulation of transcription” et ”DNA dependent”.

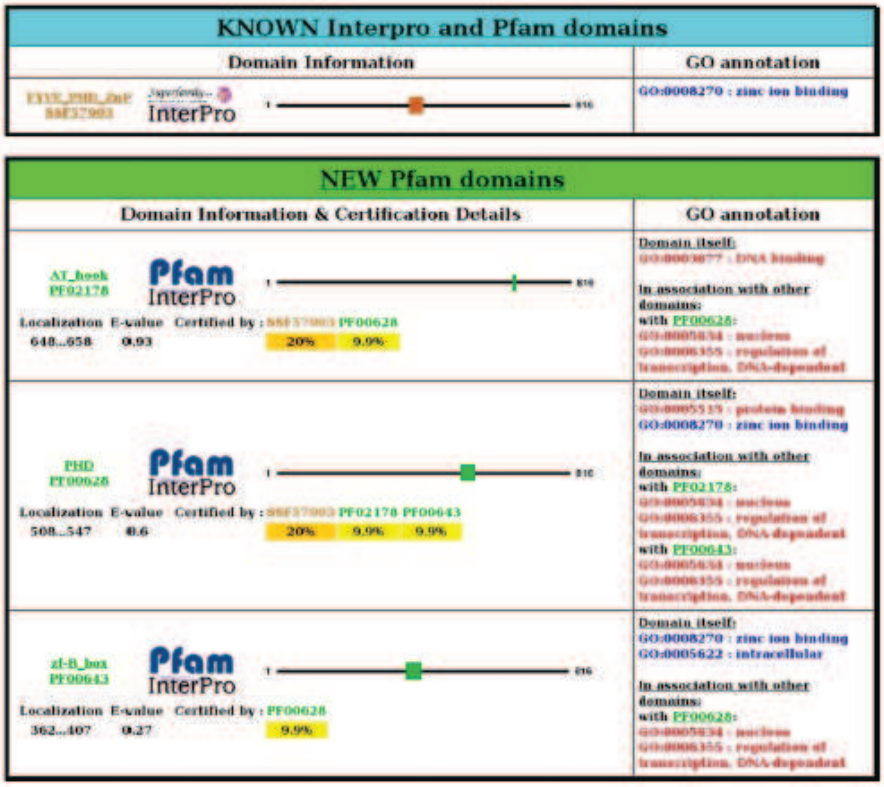


FIGURE 4.3.: Domaines connus et nouveaux domaines prédits sur la protéine *LmjF36.4590* de *L. major*. Un seul domaine y est connu : SSF57903, sa localisation et les annotations GO connus associés à la protéine sont données. Trois nouvelles occurences de domaines Pfam sont prédites : PF02178, PF00628 et PF00643 avec des E-valeur respectives de 0.93, 0.6 et 0.27. Ces domaines ont été certifiés par plusieurs domaines avec différentes valeurs de FDR. Le deuxième domaine a été par exemple certifié avec le domaine Interpro connu SSF57903 avec un FDR égal à 20%. Il a été également certifié par les deux domaines Pfam potentiels PF02178 et PF00643 avec des valeurs de FDR inférieures à 10%. Deux annotations GO ont été déduites du domaine PF00628 : “zinc ion binding” et “protein binding”. En combinaison avec d'autres domaines prédits, le domaine PF00628 apporte deux autres annotations : “nucleus” et “regulation of transcription, DNA related“

La base de données regroupe plusieurs centaines de nouveaux domaines Pfam pour chacune des espèces. Globalement, les prédictions obtenues ont permis d'enrichir l'ensemble de domaines connus chez tous les organismes étudiés. À un FDR au dessous de 20%, une extension de

Organisme	FDR	Domaines seuls	Domaines combinés	Prot. non annotées
L. major	$\leq 10\%$	158	108	28
	$\leq 20\%$	302	164	53
P. falciparum	$\leq 10\%$	117	77	20
	$\leq 20\%$	250	111	39
T. gondii	$\leq 10\%$	226	129	38
	$\leq 20\%$	412	165	71
C. parvum	$\leq 10\%$	176	114	44
	$\leq 20\%$	258	141	56
G. lamblia	$\leq 10\%$	51	37	14
	$\leq 20\%$	203	99	59
T. brucei	$\leq 10\%$	246*	175	37*
	$\leq 20\%$	402	282	79

TABLE 4.4.: Nouvelles annotations GO sur les protéines des différentes espèces. "Domaines seuls" est le nombre de termes GO que l'on peut transférer à la protéine par la présence d'un nouveau domaine certifié seul ; "Domaines combinés" est le nombre de termes GO supplémentaires (différents de ceux comptés dans la colonne précédente) déduits à partir des combinaisons de nouveaux domaines entre eux ou avec les domaines déjà connus ; "Prot. non annotées" est le nombre de protéines qui n'ont aucune annotation GO connue et pour lesquelles de nouvelles annotations sont proposées.* Résultats obtenus avec un FDR = 12% ; aucune certification avec les domaines Interpro non-Pfam n'est obtenue chez *T. brucei* avec un FDR $\leq 10\%$.

l'ensemble des domaines de toutes les espèces a été observée. Cette extension va de 13% chez *G. lamblia* à 28% chez *C. parvum*. Certains domaines sont complètement nouveaux chez ces espèces, c'est à dire qu'aucune occurrence n'était connue sur l'ensemble des protéines de l'espèce. Ainsi un enrichissement de l'ensemble des familles de domaines connus dans les différents organismes a été observée, cette hausse varie entre 10% pour *P. falciparum* à 16% pour *C. parvum*.

De plus, les nouveaux domaines prédits permettent d'associer de nouvelles annotations fonctionnelles aux protéines des différentes espèces. À un FDR au dessous de 20%, le nombre de nouveaux termes GO ajoutés varie entre 302 pour *G. lamblia* et 684 dans *T. Brucei* (*cf.*, table 4.4). Les tables 4.2 et 4.3 résument les nombres de domaines trouvés dans chaque espèce avec les trois types de certification à des seuils de FDR au dessous de 10% et 20% respectivement. Le tableau 4.4 rapporte le nombre d'annotations GO apportées par les nouvelles prédictions.

Pour *L. major*, par exemple, EuPathDomains contient un total de 719 nouveaux domaines avec un $FDR \leq 20\%$ (*cf.* table 4.3). Ceci présente une augmentation de 14% en comparaison avec les 5216 domaines Pfam déjà connus dans cet organisme (nous considérons ici une seule occurrence par protéine de chaque domaine connu ou prédit). Parmi les 719 domaines, 607 domaines associent une nouvelle famille Interpro à la protéine, c'est à dire qu'aucune occurrence de la famille du domaine n'était connue sur la protéine en question. Les domaines Pfam connus permettent de certifier 385 des 719 nouveaux domaines, les domaines connus non-Pfam en cer-

tifient 417, et 316 domaines sont certifiés par les domaines potentiels eux-mêmes (plusieurs nouveaux domaines sont certifiés par deux ou les trois types de certifications). Comme on peut s'y attendre, les domaines potentiels permettent la certification de moins de domaines que les domaines connus à même FDR, puisque les domaines certifiants sont eux-mêmes potentiellement faux. De fait, seuls des seuils de E-valeur très bas permettent d'avoir des prédictions avec de bas FDRs pour ce type de certification, ce qui implique la sélection d'un nombre restreint de domaines potentiels. De plus, nous avons découvert 184 nouveaux types de domaines qui n'avaient jamais été observés précédemment sur les protéines de *L. major*, ce qui représente une augmentation de 11% du nombre total de types de domaines connus chez ce parasite. Avec un $FDR \leq 20\%$, les domaines nouvellement découverts ont permis d'identifier un total de 466 nouveaux termes GO (*cf.* Table 4.4), représentant 6% du total des termes GO déjà connus dans les protéines de *L. major*. Parmi les protéines ayant reçu des annotations, 53 n'avaient aucune annotation connue. Des résultats similaires ont été obtenus sur les deux autres espèces de *Leishmania* (*L. infantum* et *L. braziliensis*) comme le montre la table 4.5.

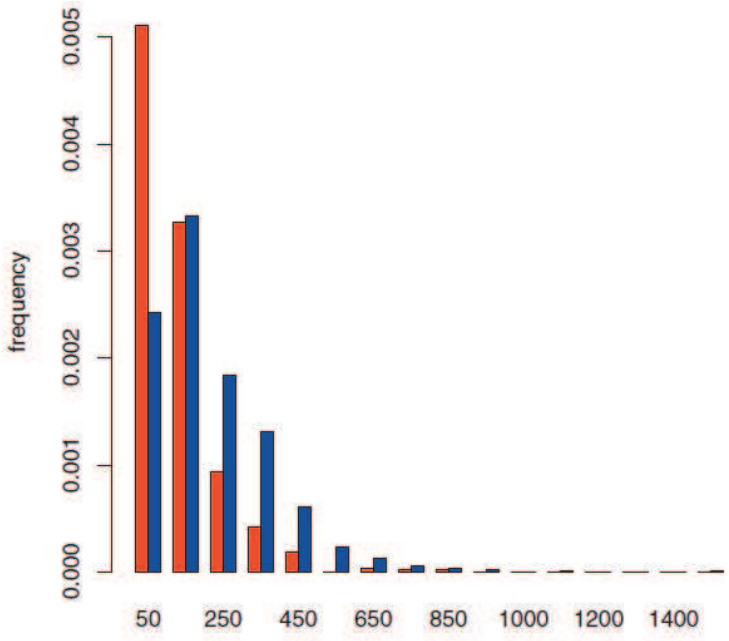


FIGURE 4.4.: Comparaison de la fréquence des domaines connus (bleue) et des nouveaux domaines certifiés (rouge) chez *L. major* en fonctions de leurs longueurs

4.4.3. Étude des caractéristiques des nouvelles prédictions

Nous avons ensuite étudié les spécificités des domaines nouvellement découverts, comparés à ceux déjà connus dans chaque espèce. Une des premières particularités constatées est liée à la longueur des domaines. La longueur des domaines protéiques en général varie entre 25 à 500 acides aminés. Une comparaison entre les domaines Pfam connus et ceux nouvellement prédits montre qu'en moyenne, les domaines nouvellement certifiés sont plus courts que

L. major	*L. Infantum*		
	Domaines connus	Nouv dom.% FDR≤ 10%	Nouv dom. FDR ≤ 20%
Domaines connus	5114/5279 (97%)	31/5279 (1%)	36/5279 (1%)
Nouv. dom. FDR ≤ 10%	18/448 (4%)	390/448 (87%)	392/448 (88%)
Nouv. dom. FDR ≤ 20%	19/776 (2%)	535/776 (69%)	604/776 (78%)

TABLE 4.5.: Comparaison des prédictions effectuées dans deux espèces : *L. major* et *L. infantum* en termes de proportion de domaines Pfam (nouveaux et déjà connus) sur les protéines *L. major* qui sont aussi présents parmi les domaines Pfam ((nouveaux et déjà connus)) chez leurs homologues chez *L. infantum*. 5279 des domaines Pfam connus chez *L. major*, le sont sur des protéines ayant un homologue connu chez *L. infantum*. Parmi ces domaines, 5114 sont aussi trouvés chez les homologues *L. infantum*. Il en va de même pour les nouveaux domaines : 448 nouveaux domaines chez *L. major* (avec un FDR < 10%) sont sur des protéines ayant des homologues chez *L. infantum*. 18 de ces derniers sont déjà connus chez les homologues *L. infantum* et 390 sont détectés également sur les homologues *L. infantum* avec un FDR < 10%.

ceux déjà connus comme (*cf.* figure 4.4). Cette différence des longueurs de domaines n'est pas surprenante. Elle peut être expliquée par le fait que les domaines courts portent moins d'informations que les plus longs et sont plus difficiles à détecter, particulièrement sur des séquences de protéines divergentes. C'est donc tout naturellement qu'une approche comme la nôtre va découvrir ces domaines là où les approches classiques échouent. Nous avons ensuite étudié les annotations GO apportées par les domaines nouvellement prédits. Nous avons considéré aussi bien les annotations déduites de la présence d'un domaine seul que celles déduites par la combinaison de plusieurs domaines. Pour des protéines déjà annotées, on constate que les nouveaux domaines permettent généralement la découverte de fonctions plus précises confirmant les annotations déjà connues. Par exemple, le terme "RNA polymerase II transcription factor activity", qui catalyse la transcription d'ADN pour synthétiser les précurseurs d'ARNm et la plupart des snRNA et micro ARN, est assigné à différentes protéines de *Leishmania* en combinant des nouveaux domaines avec des domaines connus. Pour des protéines hypothétiques ou très mal annotées, deux cas de figures sont observés suivant le niveau d'annotation des domaines nouvellement détectés :

- Les nouveaux domaines n'ont aucun terme GO qui leur est associé ou sont associés à des termes GO peu informatifs. Dans ce cas, aucune annotation fonctionnelle précise ne peut être déduite, mais la catégorisation structurelle des protéines est raffinée, fournissant des indices pour des inférences fonctionnelles futures ;
- Les domaines nouvellement détectés sont suffisamment informatifs et ont des termes GO précis, qui peuvent alors être transférés à la protéine hypothétique ou mal annotée.

Nous avons ensuite utilisé l'outil GOSTAT [3] (Beissbarth and Speed, 2004) pour comparer les annotations GO associées aux domaines connus et ceux nouvellement certifiés selon les trois axes de la Gene Ontology : fonction moléculaire, processus biologique et localisation cel-

3. http ://gostat.wehi.edu.au/

GO term	L. m.	P. f.	T. g.	G. l.	C. p.	T. b.
Molecular function						
Signal transducer activity	X		X		X	
Protein_kinase activity	X				X	X
Transmembrane receptor protein kinase activity	X					
Zinc ion binding		X				
Purine ribonucleotide binding		X				X
ATP binding		X				
Structure specific DNA binding			X		X	
Enzyme inhibitor activity			X		X	
GTP binding						X
Biological process						
Post-translational protein modification	X					X
Cell communication	X					
Cell surface receptor linked signaling pathway	X		X			
Signal transduction	X					
DNA replication, synthesis of RNA primer	X		X			
Glucose catabolic process process						X
Cellular component						
Nucleus	X		X			
Receptor complex	X		X			
Intracellular membrane-bounded organelle		X				
Clathrin adaptor complex		X				
Phosphopyruvate hydratase complex						X

TABLE 4.6.: Termes GO sur-représentés parmi les annotations apportées par les domaines prédits en comparaison des annotations associées aux domaines connus. L. m. pour *L. major* ; P. f. pour *P. falciparum* ; T. g. pour *T. gondii* ; G. l. pour *G. lamblia* ; C. p. pour *C.parvum* et T. b. pour *T. brucei*.

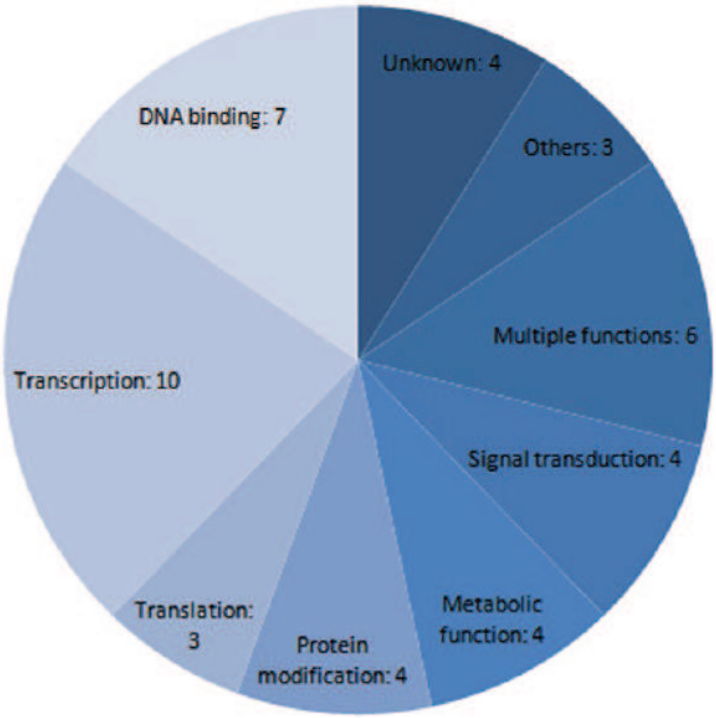


FIGURE 4.5.: Annotations GO des nouvelles familles de domaines prédites chez *L. major*, *L. infantum* et *L. braziliensis*. Les chiffres indiquent le nombre de domaines associés à chaque fonction

lulaire. Pour chaque espèce, nous avons téléchargé dans GOSTAT l'ensemble des familles de domaine nouvellement découvertes, l'ensemble des familles de domaines connus et les associations pfam2go. Nous avons déterminé avec GOSTAT les termes GO qui sont sur-représentés dans les nouvelles familles de domaine en comparaison avec celles déjà connues. La table 4.6 liste les résultats réalisés dans chaque espèce pour l'ensemble des domaines certifiés avec un $FDR \leq 10\%$. À l'exception de *G. lamblia*, plusieurs termes semblent être sur-représentés dans les différents organismes. Par exemple, plusieurs termes sur-représentés dans les espèces plasmodiales sont reliés à la fixation de l'ADN ou de l'ARN. Cette découverte pourrait être utile dans la compréhension et décodage du mode de transcription qui demeure mal connu dans ces organismes (Essien and Stoeckert, 2010; Bischoff and Vaquero, 2010).

4.4.4. *Leishmania* comme étude de cas

Nous avons poussé plus loin nos analyses sur les trois espèces du genre *Leishmania* incluses dans la base de données EuPathDomains. À première vue, nous remarquons que beaucoup de termes liés à l'ADN, la stabilité de l'ARN et la transcription sont trouvés dans les nouveaux domaines découverts (*cf.* la table 4.6 et la figure 4.5) dans ces espèces, en comparaison avec les domaines connus. Ceci devrait aider à identifier de nouveaux éléments clés impliqués dans les mécanismes complexes de la transcription polycistronique chez *Leishmania*. L'analyse des

prédictions réalisées sur l'ensemble de ces trois espèces montre qu'un grand nombre de domaines nouvellement découverts est partagé entre eux (comme le montre la table 4.5). Ceci augmente notre confiance en la précision de l'approche. En effet, la majorité (87% avec un $FDR \leq 10\%$) des nouvelles familles de domaines qu'on découvre chez *L. major* sont aussi prédites chez *L. infantum*. De plus, 18 des domaines restants sont en fait déjà connus sur les protéines de *L. infantum*. Une étude des nouvelles familles de domaines qui n'étaient pas connus chez le genre *Leishmania* nous montre que 45 (40%) domaines prédits avec un FDR $\leq$ 10% sont prédits chez les trois espèces étudiées. Parmi ces domaines, figure le domaine "RQC" (PF09382) relié à l'activité de fixation de l'ADN. Ce domaine est spécifique à la famille RecQ des ADN hélicases qui fixe et dévide l'ADN-G4. Une implication de l'ADN-G4 dans le processus de régulation et la maintenance de la stabilité génomique a été décrite récemment (Hanakahi and Sun, 1999; Parkinson et al., 2002; Maizels, 2006). Il participe en effet au processus de transcription en maintenant l'ADN dans une conformation ouverte (Du et al., 2008). De plus, le domaine "RQC" est doté d'une structure helix-turn-helix (HTH) trouvée dans la majorité des protéines impliquées dans la régulation des gènes et fixe l'ADN de manière spécifique (Struhl, 1989; Mathee and Narasimhan, 2003). Ainsi, le domaine "RQC" peut représenter une composante importante du processus de régulation de gènes chez *Leishmania*. Un domaine "Response reg" (PF00072), relié à un autre niveau de la régulation a été également prédit. Chez d'autres organismes, ce dernier est connu par son association au système "two-component signal transduction system" (TCS) qui est impliqué dans les événements de transduction de signaux (Krauss, 2001; Pao and Saier, 1995). La voie métabolique TCS a été initialement décrite chez les bactéries et récemment chez les Eukaryotes. Elle est impliquée dans la réponse environnementale face au stress ou aux conditions de développement (Whitworth and Cock, 2009; Skerker et al., 2005). Parmi les activités dans lesquelles la voie métabolique TCS est impliquée figurent la résistance aux médicaments, l'osmorégulation et la motilité. On peut alors suggérer l'implication du domaine "Response reg" associé à cette voie dans l'adaptation du statut physiologique des pathogènes aux différents stimuli.

Nous avons classé les 45 nouvelles familles de domaines en neuf groupes fonctionnels en nous basant sur les associations pfam2go et la littérature. Les deux familles de fonctions sur-représentées sont reliées à la transcription et à la fixation de l'ADN comme le montre la figure 4.5. Mis à part ces domaines prédits en commun chez les trois espèces, une petite proportion (16% chez *L. major* par exemple) des nouvelles familles prédites semble être spécifique à chaque espèce, même si on ne peut pas exclure l'existence de faux positifs parmi ces prédictions.

4.4.5. Les prédictions communes à tous les pathogènes étudiés

Nous avons énuméré pour les différentes espèces les nouvelles familles de domaines prédites *i.e.* qui n'étaient pas connues chez l'espèce. Huit de ces familles sont communes à tous les pathogènes. La présence de ces domaines en commun pose la question de leur relation à un mode de vie parasitaire commun à l'ensemble des pathogènes malgré la diversité des infections associées. Ceci comprendrait des mécanismes infectieux et des stratégies de survie communes par exemple (Hybiske and Stephens, 2008). Trois des familles de domaines communes (Utp12, Utp21 et PWP2) sont associées à des protéines impliquées dans le SSU-processome pour la biogenèse des sous-unités ribosomiques ARNr (Dragon, 2002). Quelques études suggèrent l'implication du SSU-processome dans le développement du pre-ARNr et dans l'assemblage des ribosomes à travers son association avec des protéines ribosomales (Bernstein, 2004). La quatrième nouvelle famille de domaines prédite est la PAM2 connue chez plusieurs protéines des Eukaryotes (Albrecht and Lengauer, 2004). Ce domaine est connu pour son interaction avec une famille particulière d'ARN hélicases et participe à la reconnaissance des régions 3' des

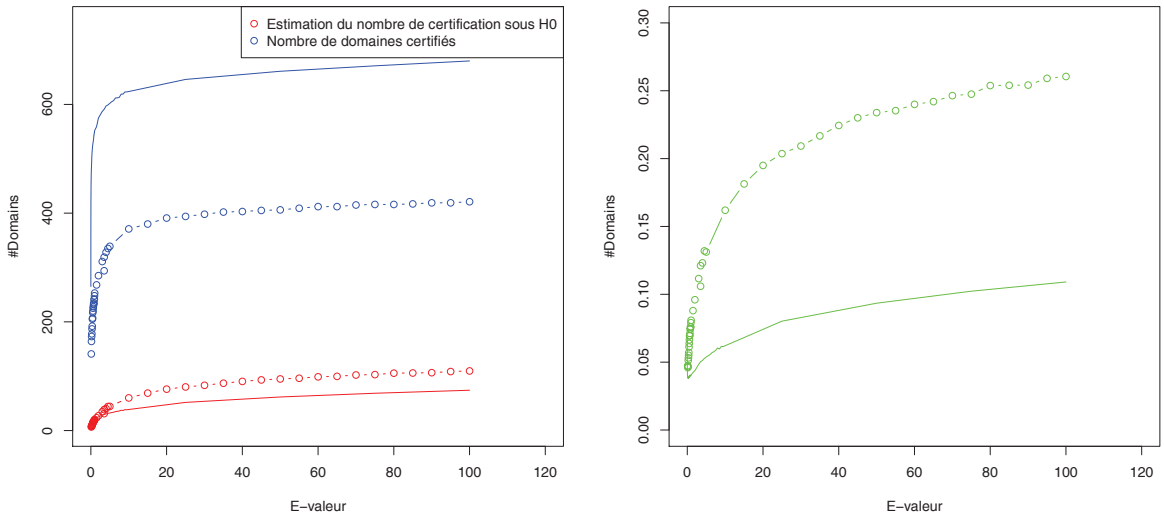


FIGURE 4.6.: Comparaison du nombre de certification et de l'estimation du nombre de domaines certifiés sous H_0 sur les protéines de *P. falciparum* en fonction de la E-valeur (à gauche) entre HMMER 2 + Pfam 23 (en pointillé) et HMMER 3 + Pfam 26 (ligne pleine). Les FDR correspondant calculés par la formule 4.1 sont représentés en vert (à droite)

ARNm. Il est également impliqué dans l'initiation de la traduction et la stabilisation et dégradation de l'ARNm (Nonhoff, 2007). Le cinquième domaine "DBP10CT" (PF08147) de cet ensemble est fréquemment trouvé sur des protéines hypothétiques annotées "RNA helicase". Ceci peut refléter un besoin de synthèse rapide de molécules favorisant la survie de n'importe quel organisme et sa résistance au stress environnemental, commun à tous les pathogènes. Les trois derniers domaines de cet ensemble ont des fonctions distinctes. Le domaine "MutS_II" est trouvé par exemple sur des protéines impliquées dans les "mismatch repair". Le deuxième PF08953 n'a pas de fonctions connues qui lui sont associées. Enfin, le domaine "EFhand-like" (PF09279) a été prédit chez les différents pathogènes sauf sur les espèces plasmodiales où il est déjà connu. Ce domaine est relié à l'activité phospholipase qui participe à la dégradation des membranes contenant les pathogènes en diminuant le nombre de phospholipides causant ainsi l'affaiblissement de l'intégrité de ces membranes (Cummings et al., 2010).

En conclusion, nous faisons l'hypothèse qu'une partie de ces nouvelles familles détectées chez l'ensemble des pathogènes pourraient être révélateurs d'un mode de vie parasitaire associé à l'hôte humain. Néanmoins cette analyse doit être complétée par une étude de leur occurrence chez les espèces non parasites.

4.5. Nouvelle version 2.0 de EupathDomains

La première version de la base et l'étude présentée ci-dessus ont été effectuées durant les années 2009-2010. Elle est basée sur l'utilisation de la version 23.0 de Pfam et de HMMER

Organisme		FDR $\leq$ 10%			
	Type certif.	Pfam	Interp.	Pot.	Tous
L. major	Certif. dom.	1266	664	1394	2196 (41%)
	Nouv. types dom.	351	203	342	537 (33%)
L. infantum	Certif. dom.	931	356	1075	1611 (34%)
	Nouv. types dom.	282	141	276	436 (28%)
L. braziliensis	Certif. dom.	971	734	772	1416 (27%)
	Nouv. types dom.	300	227	205	413 (26%)
P. falciparum	Certif. dom.	703	979	783	1328 (32%)
	Nouv. types dom.	247	301	251	420 (25%)
P. vivax	Certif. dom.	640	560	567	922 (23%)
	Nouv. types dom.	224	185	192	338 (20%)
P. yoeli	Certif. dom.	523	396	432	738 (17%)
	Nouv. types dom.	191	143	152	270 (18%)
T. gondii	Certif. dom.	911	354	816	989 (22%)
	Nouv. types dom.	334	156	270	467 (27%)
C. parvum	Certif. dom.	553	457	428	705 (27%)
	Nouv. types dom.	258	212	149	359 (32%)
G. lamblia	Certif. dom.	323	288	531	660 (22%)
	Nouv. types dom.	107	72	163	211 (23%)
T. brucei	Certif. dom.	1017	429	844	1464 (30%)
	Nouv. types dom.	296	159	241	421 (27%)

TABLE 4.7.: Utilisation de HMMER 3 et Pfam 26.0 : Prédictions sous un FDR de 10% chez les organismes étudiés "Type certif." indique le type de certification : par les domaines Pfam connus ("Pfam"), les Interpro (non-Pfam) connus ("Interp."), ou par les domaines potentiels ("Pot.") ; la colonne "Tous" indique le nombre total de domaines certifiés en cumulant les différents types de certification. "Dom. Certif." indique le nombre de nouveaux domaines certifiés ; "Nouv. types dom." indique le nombre de types de domaines qui n'étaient pas connus chez les différentes protéines. Les chiffres entre parenthèses indiquent les proportions de nouveaux domaines ou nouveaux types de domaines par rapport à ceux déjà connus.

2, et les recherches de domaines ont été faites en mode global. Depuis cette date, la nouvelle version 3 de HMMER, beaucoup plus rapide, a vu le jour. Cette dernière n'autorise que les recherches locales sur les HMM profils (*cf.* section 2.2.4). Les occurrences de domaines peuvent alors correspondre à des domaines complets ou des fragments de domaines. En août 2013, nous avons mis à jour la base EuPathDomains en nous basant sur les domaines Pfam 26.0. HMMER 3 a été utilisé pour identifier les domaines potentiels sur les protéines des différentes espèces, et la même procédure de certification a été appliquée. Une nette amélioration des certifications a été observée chez toutes les espèces étudiées comme le montre la table 4.7. Les résultats de la certification par les domaines Pfam connus chez *P. falciparum* sont donnés par les courbes de la figure 4.6. Comme on s'y attendait, le nombre de domaines certifiés augmente en raison de la présence de fragments de domaines. Un gain considérable est également observé au niveau des FDR associés aux différentes prédictions. Une explication à cela peut être que l'utilisation des scores forward dans HMMER 3 à la place des scores de Viterbi (*cf.* section 2.1.3) permet de réduire le nombre de faux positifs dans l'ensemble de domaines potentiels. Cette amélioration peut également être expliquée par l'évolution du nombre de domaines de la base Pfam (26 vs. 23). La même tendance est observée pour les deux autres types de certification par les domaines Interpro (non-Pfam) et par les domaines potentiels.

4.6. Conclusion

Nous avons présenté dans ce chapitre une base de données de domaines dédiée aux pathogènes Eukaryotes. Elle regroupe les domaines protéiques déjà connus et de nouvelles prédictions obtenues par la méthode CODD. Dans l'ensemble, EuPathDomains améliore significativement la couverture en domaines des différents organismes étudiés. Avec un FDR au dessous de 10%, la version 2.0 de EupathDomains apporte une amélioration de 17% à 41% suivant les espèces, en terme de nombre d'occurrences de domaines. Un enrichissement des répertoires des familles de domaines est également observé chez toutes les espèces, il varie entre 18% et 33%.

La méthode CODD permet donc d'améliorer de manière significative l'annotation en domaines des protéines de tous les pathogènes eukaryotes, et vraisemblablement de toutes les espèces mal annotées. Certains aspects de la méthode pourraient néanmoins être améliorés. Ces améliorations font l'objet des chapitres suivants. Parmi elles ont peut distinguer :

- La validation des domaines potentiels dans l'approche CODD ne prend en considération que le contexte positif d'apparition des domaines. Or, certaines paires de domaines ont peu de chance de se retrouver sur la même protéine. L'utilisation de ce contexte négatif est abordée dans le chapitre 5.
- Le choix entre domaines chevauchants est un problème courant lorsque l'on cherche à identifier la composition en domaines d'une nouvelle protéine. La stratégie utilisée est de sélectionner pour chaque conflit le domaine ayant la E-valeur la plus basse. D'autres stratégies sont envisageables. Nous discutons une approche alternative dans le chapitre 5.
- La sélection des domaines co-occurrents ne prend pas en considération l'ordre d'apparition des domaines sur la protéine ni le sens de l'implication : les paires de la CDP sont utilisées aussi bien pour certifier la présence du domaine A par le domaine B que du domaine B par le domaine A. Or suivant les probabilités respectives des deux domaines, ces certifications peuvent être plus incertaines dans un sens que dans un autre, voire faibles l'une et l'autre même si la corrélation est très significative. Ce point est également abordé au chapitre 5.
- L'identification des domaines potentiels est basée sur des comparaisons entre séquences et profils de domaines via l'outil HMMER. Là aussi d'autres approches avec un niveau de sensibilité plus élevé sont possibles. Dans le chapitre 6 nous présentons une nouvelle approche de prédiction de domaines basée sur les comparaisons de HMM-profils.

5. Stratégies d'amélioration de la prédiction par co-occurence

DANS le but de raffiner le processus de prédiction de domaines protéiques par co-occurrence, nous avons testé plusieurs stratégies que nous présentons dans ce chapitre. Nous commençons par présenter une étude pour le traitement des chevauchement de domaines. Nous présentons ensuite l'algorithme *CODD_exclusive* basé sur l'utilisation d'informations sur les exclusions de domaines et l'adaptation apportée à la procédure CODD originale avec les différents types de certification. Nous présentons enfin une autre stratégie basée sur l'utilisation de règles d'association pour la détermination des co-occurences de domaines utilisées dans le processus de certification. Nous illustrons les améliorations apportées par ces nouvelles stratégies par une série d'expériences sur les protéines de *Plasmodium* et *Leishmania*.

Sommaire

5.1. Traitement des chevauchements de domaines potentiels . 78
5.1.1. Recherche du meilleur chemin sur le graphe de domaines . . . 78
5.1.2. Études expérimentales 79
5.2. Utilisation du contexte négatif pour l'amélioration de la qualité des prédictions 81
5.2.1. Certification par les domaines connus 82
5.2.2. Certification par les domaines potentiels 84
5.2.3. Étude expérimentale 84
5.3. Études des combinaisons de domaines par les règles d'association 88
5.3.1. Les règles d'association en bref 88
5.3.2. Extraction de règles fréquentes : Limites de la métrique *minsup* sur les bases de données biologiques 89
5.3.3. Extraction des associations entre domaines 90
5.3.4. Certification par les règles d'association 90
5.4. Utilisation des espèces proches pour l'amélioration des reconnaissances de domaines 92

5.1. Traitement des chevauchements de domaines potentiels

Dans la procédure CODD, l'étape de recherche des domaines potentiels par relâchement des seuils de scores requis mène souvent à l'identification de domaines recouvrants (*cf.* section 4.2.2 du chapitre 4). Une phase de filtration est alors nécessaire afin de garder un ensemble de domaines non chevauchants. Les domaines chevauchants les domaines connus de la protéine sont écartés en premier. Ensuite on procède à la filtration des domaines potentiels restants pour garder un ensemble non recouvrant de domaines. Dans le chapitre précédent, nous avons utilisé une heuristique simple qui consiste à garder en cas de chevauchement le domaine qui a la meilleure E-valeur. Dans ce chapitre, nous étudions d'autres stratégies pour améliorer ce processus. Ce travail était en partie motivée par la publication de l'article de Yeats et al. (Yeats et al., 2010). Dans cet article, les auteurs montrent que lorsqu'on cherche à prédire les domaines qui composent une protéine, en cas de chevauchement, la stratégie visant à éliminer le domaine le moins vraisemblable (celui avec la e-valeur la plus haute) n'est pas forcément la stratégie la meilleure. Ils proposent une nouvelle approche basée sur la sélection de la clique maximale à partir d'un graphe pondéré représentant les différents domaines potentiels d'une protéine. Chaque occurrence de domaine est modélisée par un noeud pondéré par -log(E-valeur). Un arc relie deux noeuds si les domaines associés sont non chevauchants. Afin de sélectionner l'ensemble non recouvrant de domaines ayant le poids maximal, l'algorithme Cliquer (Ostergard, 2002) est utilisé pour déterminer la clique ayant le poids maximal. Cette approche tolère un certain seuil de chevauchements entre domaines dans une première phase. Ce seuil toléré peut être fixe (*i.e*, un certain nombre d'acides aminés) ou exprimé en pourcentage des longueurs des domaines. Comparé à la stratégie du choix du meilleur parmi les chevauchements, les auteurs montrent que cette approche permet de reconnaître plus (une amélioration de 15%) de domaines CATH sur trois benchmarks de séquences protéiques. Cependant, la recherche de la clique maximale dans un graphe est un problème NP difficile en théorie des graphes. Nous montrons dans la section suivante que ce problème peut être reformulé en un problème de recherche de chemin de poids maximal qui admet une solution exacte en temps polynomial.

5.1.1. Recherche du meilleur chemin sur le graphe de domaines

Comme expliqué ci-dessus, l'ensemble des occurrences de domaines Pfam détectées sur chaque protéine est représenté par un graphe. Chaque occurrence de domaine est représentée par un noeud pondéré par -log(P-valeur). La P-valeur est calculée (*cf.* formule 5.1) à partir de la E-valeur rapportée au nombre de séquences étudiées, *i.e*, la taille d'un génome dans notre cas :

$$P - valeur = \frac{E - valeur}{Nombredeséquencesétudiées} \tag{5.1}$$

Les domaines non chevauchants sont connectés. Contrairement à l'approche de Yeats et al., nous représentons également sur ce graphe l'ordre d'apparition des domaines sur la protéines en orientant le graphe : si le domaine X est positionné avant le domaine non chevauchant Y, il y a un arc de X vers Y. La figure 5.1 montre un exemple de protéine et le graphe orienté associé. De cette manière, chercher le sous-ensemble de domaines non chevauchant de poids maximal revient à déterminer le chemin ayant le poids maximal. Sur l'exemple de la figure 5.1, les domaines potentiels donnés par le chemin de poids maximal sont : C, D et E. Ceci peut être résolu d'une manière efficace en adaptant (en pondérant les noeuds au lieu de pondérer les arcs) l'algorithme de programmation dynamique de recherche du chemin le plus long (Longest Path Problem) (Cormen et al., 2001). La complexité de cet algorithme est de l'ordre de $O(|V|+|E|)$ où $|E|$ représente le nombre d'arcs et $|V|$ le nombre de noeuds.

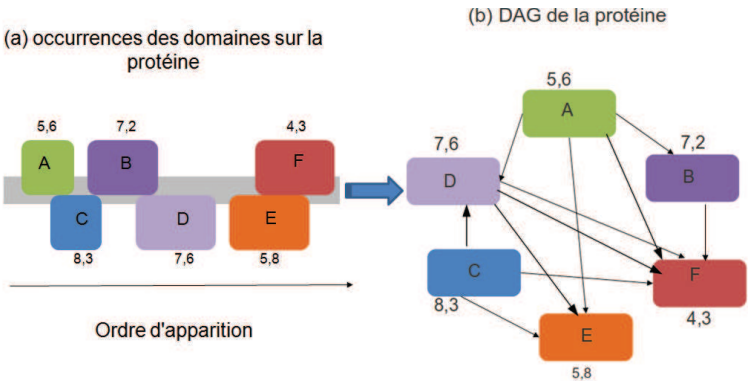


FIGURE 5.1.: Exemple de modélisation des occurrences de domaine sur une protéine par un graphe acyclique orienté et pondéré

5.1.2. Études expérimentales

Comme dans le chapitre précédent, nous avons utilisé différents seuils de score pour définir l'ensemble des domaines potentiels. Pour chaque seuil, les chevauchements ont été résolus en utilisant la politique de la meilleure E-valeur, ou la procédure décrite ci-dessus. La procédure CODD a été appliquée aux ensembles de domaines potentiels ainsi obtenus, et le FDR a été estimé en utilisant la procédure décrite à la section 4.2.4. Nous avons commencé par tester cette approche sur l'ensemble des protéines de *L. major*. Du point de vue du nombre de domaines potentiels par protéine, on observe comme on s'y attendait, un nombre plus important de domaines comparés à la procédure de sélection basée sur la meilleure E-valeur. Cependant, à même taux d'erreur, le nombre de domaines certifiés en utilisant la procédure CODD est moins élevé que lorsque les chevauchements sont résolus sur la base de la meilleure E-valeur comme le montrent les courbes de la figure 5.2.

Ces résultats négatifs remettent en cause la fonction de score utilisée. Le score d'une solution est égal à la somme des poids -(log(P-valeur)) des noeuds. Lorsque l'on analyse en détails les résultats obtenus, on s'aperçoit que cette approche a tendance à trop privilégier des enchaînements de domaines avec des P-valeur moyennes au détriment d'un domaine unique avec une bonne P-valeur. Prenons l'exemple de la protéine *LmjF01.0300* : le domaine "ABC2_membrane" (-log(P-valeur)= 10.68) est chevauchant avec les trois domaines "Ldl_recept_b" (-log(P-valeur)= 6.19), "NapE" (-log(P-valeur)= 5.39) et "DUF1145" (-log(P-valeur)= 4.75) qui figurent sur le meilleur chemin avec une somme des -log(P-valeur)= 16.34. La procédure de certification ne certifie aucun de ces trois domaines. Par contre, le domaine "ABC2_membrane" sélectionné par la stratégie du meilleur domaine est certifié grâce à la présence du domaine "ABC_tran" connu sur la protéine. Une solution envisageable à ce problème est de re-pondérer les noeuds du graphe en favorisant les scores élevés. On peut par exemple utiliser une fonction puissance pour cela. Chacun noeud est pondéré par sa P-valeur à la puissance N ($N \geq 2$). Plusieurs valeurs de N ont été testées (de 2 à 20). Aucune n'a permis d'améliorer les résultats et seul le passage à la puissance 10 a permis d'atteindre des performances similaires à la stratégie basée sur le choix de la meilleure E-valeur. Une autre solution consiste à filtrer les domaines chevauchants en fonction de leurs P-valeurs afin d'éviter qu'un domaine avec une petite P-valeur soit remplacé par plusieurs domaines avec des P-valeurs élevées. Un domaine est écarté du graphe si sa P-valeur $> X\times$ P-valeur d'un autre domaine

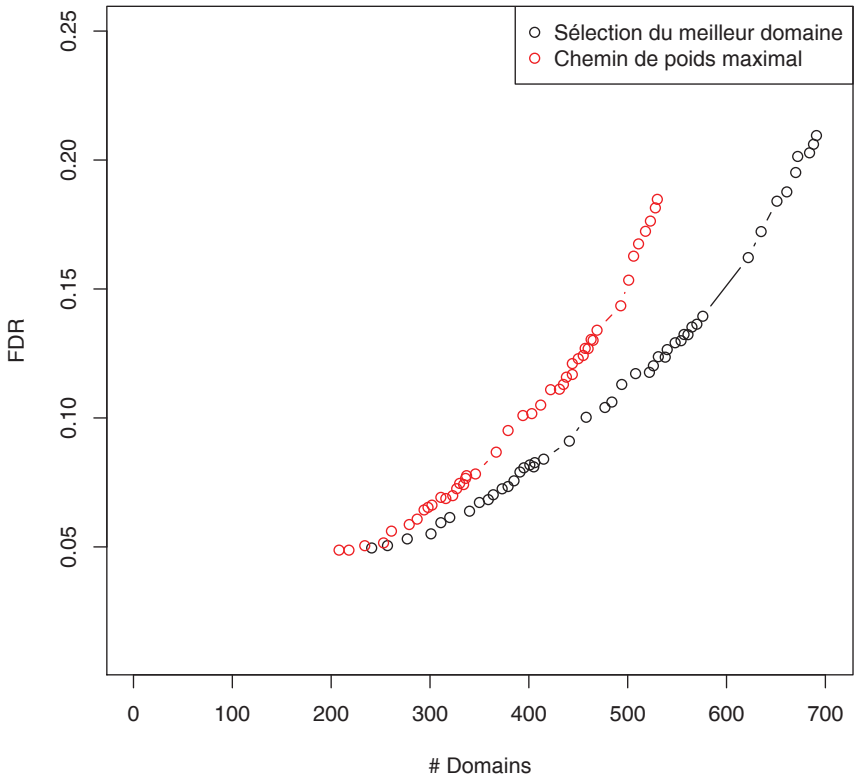


FIGURE 5.2.: Comparaison des certifications obtenues par la sélection du meilleur domaine à la recherche du chemin de poids maximal sur les protéines de *L. major*

qu'il chevauche. Plusieurs valeurs de X ont été testées. On ordonne les domaines par ordre croissant de P-valeur. On parcourt cette liste et pour chaque domaine, on retire de la liste tous les domaines le chevauchant ayant une P-valeur $> X\times$ sa P-valeur. De légères augmentations en termes de nombre de domaines certifiés ont été observées pour quelques valeurs de X (2, 3 et 4) comme l'illustre la figure 5.4. Cependant, cette approche n'est pas complètement satisfaisante, comme elle intervient en amont de la recherche du chemin de poids maximal, elle mène parfois à écarter prématurément certains domaines. Ce problème est illustré par la figure 5.3. Le domaine C a été supprimé car il est chevauchant avec le domaine E (la P-valeur de C $> 2\times$ P-valeur de E). Les domaines potentiels donnés par le chemin de poids maximal sont A, B et D. Le domaine C, non chevauchant avec ces derniers, aurait donc pu lui aussi figurer sur ce chemin.

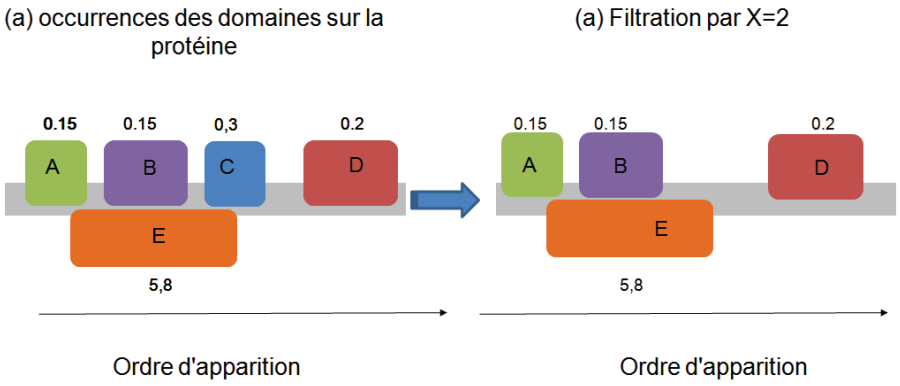


FIGURE 5.3.: Exemple de filtration de domaines par X=2

5.2. Utilisation du contexte négatif pour l'amélioration de la qualité des prédictions

Comme nous l'avons présenté dans le chapitre précédent, la méthode CODD augmente la sensibilité de détection de domaines Pfam dans les protéines divergentes, tout en contrôlant le taux de fausses prédictions. CODD utilise pour cela la tendance qu'ont la plupart des domaines protéiques d'apparaître préférentiellement avec quelques autres domaines préférés dans les protéines (Apic, 2001; Beaussart et al., 2007) (*cf.* section 4.2).

Le test statistique de Fisher est utilisé pour identifier les paires de domaines fortement co-occurrentes dans l'ensemble des protéines de la base de données Uniprot (*cf.* section 4.2.1). Cette liste de paires de domaines (appelée liste des CDP) est ensuite utilisée pour valider la présence de domaines potentiels grâce à la présence d'un autre domaine sur la même protéine (Terrapon et al., 2009). Cependant, cette méthode ne se sert que du contexte positif d'apparition de domaines et ignore le fait que certaines paires de domaines ont peu de chance de se retrouver sur la même protéine. Or, cette information est d'une grande importance car elle peut être utilisée pour rejeter un domaine erroné et abaisser le taux d'erreur. Nous proposons dans ce chapitre une extension de l'algorithme CODD qui utilise le contexte négatif pour élaguer les domaines en contradiction avec les autres domaines de la même protéine.

Étant donnée une séquence protéique à analyser et $\mathcal{K}$ l'ensemble des domaines Pfam connus de cette protéine, on note $\mathcal{P}$ l'ensemble des domaines potentiels identifiés avec un seuil de score relâché. Comme dans la procédure CODD originale, seuls les domaines qui ne chevauchent aucun des domaines connus sont inclus dans $\mathcal{P}$. De plus, en cas de chevauchement de deux domaines potentiels, celui avec le score le plus bas est écarté. Finalement, $\mathcal{DP}$ dénote la liste des paires de domaines observées ensemble dans au moins une des protéines de la base de données Uniprot. Nous considérons dans cette nouvelle approche que toutes les paires de domaines qui ne figurent pas dans la liste $\mathcal{DP}$ sont incompatibles : elles ne peuvent pas apparaître ensemble sur la même protéine. Cette règle peut sembler restrictive, mais notons que la base de données Uniprot rassemble toutes les séquences protéiques de tous les organismes séquencés à ce jour. On peut donc raisonnablement penser que l'essentiel des co-occurences compatibles sont effectivement présentes dans la liste $\mathcal{DP}$. Comme pour la CDP, ni les positions ni l'ordre des domaines sur les protéines ne sont pris en considération. On a donc $CDP \subset \mathcal{DP}$. Les

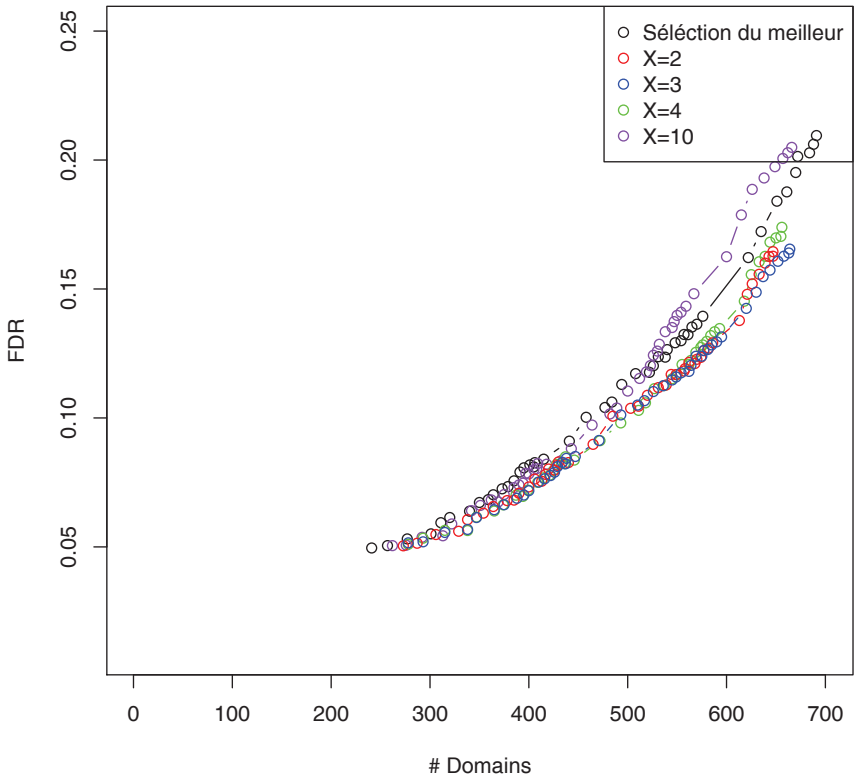


FIGURE 5.4.: Comparaison des certifications obtenues par la sélection du meilleur domaine v.s. la recherche du chemin de poids maximal avec différents seuils de filtration par X=2, X=3, X=4 et X=10

différentes étapes impliquées dans cette approche sont résumées par la figure 5.5. L'intégration de l'exclusion des domaines dans la procédure CODD exige la distinction entre les différents types de certifications : celles basées sur les domaines déjà connus et celles faites sur la base d'un autre domaine potentiel de la protéine. Ces deux nouvelles procédures sont détaillées ci-dessous.

5.2.1. Certification par les domaines connus

L'algorithme 4 décrit l'adaptation de la procédure CODD avec l'intégration des informations d'incompatibilité de domaines dans la procédure de certification par les domaines déjà connus. L'algorithme comporte deux principales étapes. La première consiste à filtrer l'ensemble des domaines potentiels. Pour chaque domaine potentiel p, nous vérifions si ce dernier présente un conflit avec les domaines Pfam connus de la protéine. Si p n'a jamais été observé avec un (ou plusieurs) domaines connus de la protéine —*i.e.* $\exists\ k \in \mathcal{K}$ tel que $(p,k) \notin \mathcal{DP}$—il est supprimé de $\mathcal{P}$ (lignes 4-8). Les domaines potentiels qui passent ce filtre sont alors testés pour

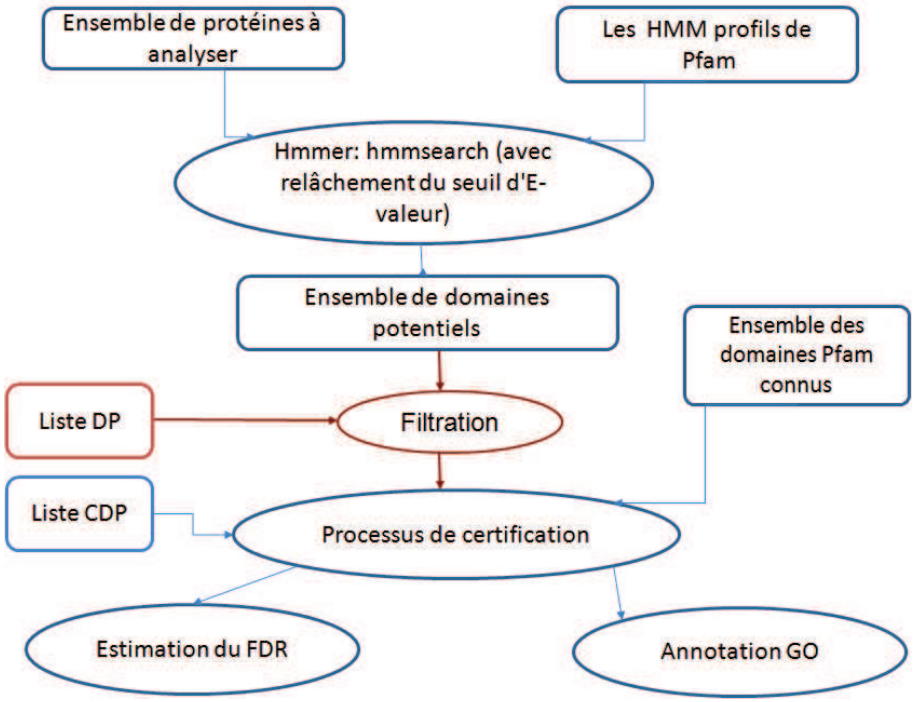


FIGURE 5.5.: Celle en rouge n'était pas considérée dans l'algorithme CODD original. Elle est reliée à l'utilisation de la liste $\mathcal{DP}$ pour le filtrage de l'ensemble des domaines potentiels.

la certification avec les domaines Pfam connus de la protéine comme dans la procédure CODD originale, *i.e.* tous les domaines dans $\mathcal{P}$ qui forment des paires appartenant à la CDP avec un des domaines Pfam connus de la protéine sont considérés comme certifiés (lignes 11-15).

Algorithm 4: *CODD_exclusive* pour les domaines connus

```
1  Inputs : P, K, CDP, DP
2  Output : C l'ensemble des domaines certifiés
3  /* Step 1 : Étape de filtration */
4  foreach p ∈ P do
5      foreach k ∈ K do
6          If (p,k) ∉ DP
7          then supprimer p de P
8  /*Step 2 : Étape de certification */
9  foreach p ∈ P do
10     foreach k ∈ K do
11         If (p,k) ∈ CDP
12         then ajouter p dans C
13 Return C
```

5.2.2. Certification par les domaines potentiels

L'algorithme 5 décrit l'adaptation de la procédure CODD pour l'intégration des informations d'incompatibilité dans la procédure de certification par les domaines potentiels. Si la protéine en question a des domaines connus, la première étape consiste à filtrer les domaines potentiels qui présentent des conflits avec les domaines Pfam connus de la protéine (lignes 4-9). Ensuite, les domaines potentiels restants sont ordonnés en fonction de leurs scores. Les domaines les plus vraisemblables (*i.e.* ayant les scores les plus élevés) sont considérés en premier. Pour chaque domaine potentiel p, nous énumérons tous les autres potentiels p' ayant un score meilleur que celui de p. Si la paire (p, p') figure dans la $\mathcal{CDP}$, les deux domaines sont certifiés (lignes 12-15). Une fois certifiés, les domaines p et p' sont utilisés dans l'élagage des domaines potentiels restants : chaque domaine restant en conflit avec p ou p' est enlevé de $\mathcal{P}$ (lignes 16-19).

5.2.3. Étude expérimentale

Pour évaluer ces nouvelles approches, nous avons mené plusieurs tests sur l'ensemble des protéines de : *Leishmania major* et *Plasmodium falciparum*. En se basant sur les seuils de score recommandés par Pfam, 1607 et 1429 domaines connus ont été identifiés respectivement chez *L. major* et *P. falciparum*. Nous avons ensuite relâché les seuils stringents proposés par Pfam et identifié les domaines potentiels sur les différentes protéines. Les procédures de certification avec les domaines connus et celle avec les domaines potentiels ont été appliquées avec et sans l'utilisation des informations d'exclusion de domaines. Pour chaque expérience, le nombre de domaines nouvellement certifiés a été calculé et le taux d'erreur associé aux certifications a été estimé en utilisant la procédure de ré-échantillonnage décrite dans la section 4.2.4 du chapitre 4. Différents seuils de scores ont été utilisés dans ces expérimentations pour définir l'ensemble

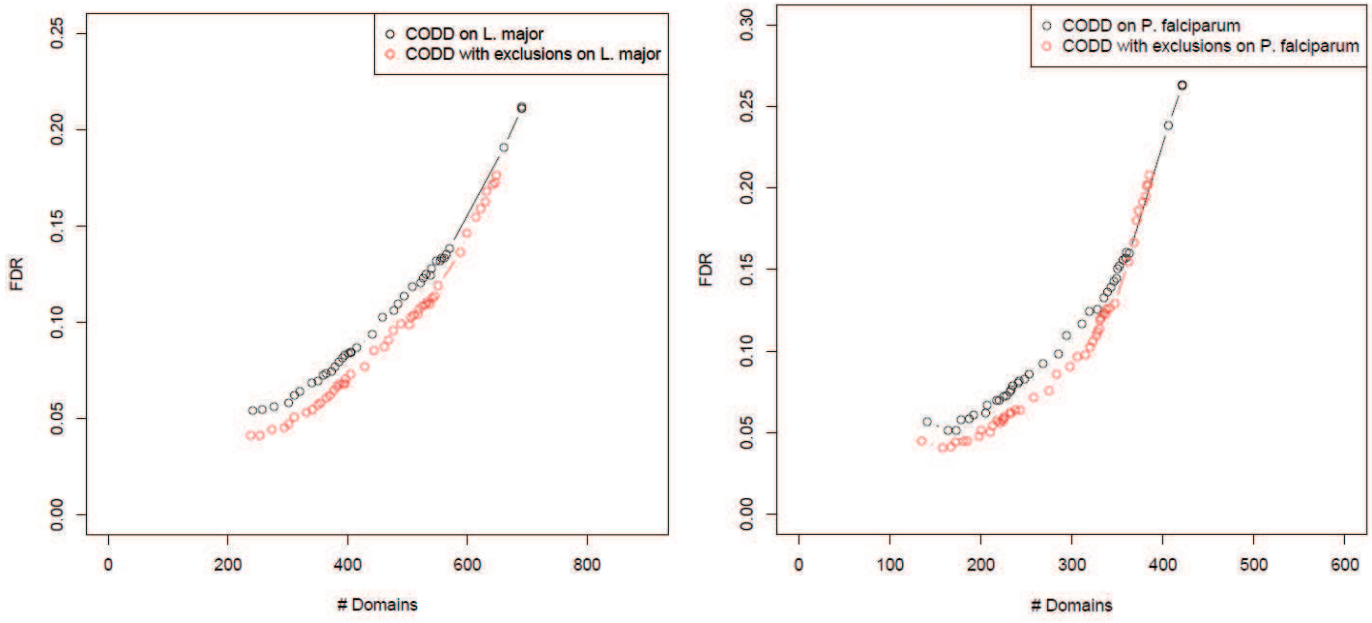


FIGURE 5.6.: Nombre de nouveaux domaines certifiés avec les différents seuils de FDR par les domaines Pfam connus chez *P.falciparum* (droite) et *L. major* (gauche) avec (points rouges) et sans (points noirs) utilisation des exclusions de domaines

Algorithm 5: *CODD_exclusive* pour les domaines potentiels

Inputs : $\mathcal{P}$ trié par par scores décroissants, $\mathcal{K}$, $\mathcal{CDP}$, $\mathcal{DP}$
Output : $\mathcal{C}$ l'ensemble des domaines certifiés
/* ***Étape de filtration*** */
foreach $p \in \mathcal{P}$ **do**
    foreach $k \in \mathcal{K}$ **do**
        If $(p,k) \notin \mathcal{DP}$
        then remove p from $\mathcal{P}$
        Endif
/* ***Étape de certification*** */
foreach $p \in \mathcal{P}$ *triés par scores décroissants* **do**
    foreach $p' \in \mathcal{P}$ $(p' \neq p)$ **do**
        If $(p,p') \in CDP$ **then**
        add p and p' to $\mathcal{C}$
        foreach $p'' \in \mathcal{P}$ **do**
            If $((p'',p) \notin \mathcal{DP})$ **then** supprimer p''
            If $((p'',p') \notin \mathcal{DP})$ **then** supprimer p''
        Endif
Return $\mathcal{C}$

des domaines potentiels $\mathcal{P}$. La figure 5.6 récapitule les résultats des certifications réalisés avec les domaines Pfam connus. Comme dans les figures précédentes, chaque point correspond à un seuil de score particulier, les points les plus à droite correspondent aux scores les plus faibles (les E-valeurs les plus élevées). L'axe des abscisses représente le nombre de domaines nouvellement découverts, tandis que l'axe des ordonnées représente les FDR associés à ces prédictions.

Dans l'ensemble, nous pouvons conclure qu'à même nombre de domaines prédits, la nouvelle procédure nous permet de réaliser des prédictions avec des taux d'erreur (FDR) un peu plus bas. Chez *L. major*, par exemple, la procédure CODD classique permet de prédire au total 600 domaines avec un FDR de 23% sur l'ensemble de tous les domaines potentiels, avec un seuil de E-valeur égal à 100. Avec cette même E-valeur, l'algorithme *CODD_exclusive* certifie à peu prés 50 domaines en moins. Cependant, une amélioration significative du taux d'erreur est observée, ce dernier passe de 23% à 16%. Ainsi, une amélioration de la qualité des prédictions (une baisse des FDR) est observée compensant la légère perte en terme de nombre de domaines prédits.

La même tendance est observée quand les domaines potentiels sont utilisés pour la certification (*cf.* figure 5.7). À même seuil d'E-valeur, les FDR associés aux différents ensembles de domaines potentiels sont à la baisse au prix d'une légère perte en termes de nombre de domaines prédits. Sur l'ensemble des domaines potentiels détectés chez *L. major* avec un seuil de E-valeur égal à 10 par exemple, la procédure CODD originale permet de certifier 494 domaines avec un FDR de 23%. Avec la procédure d'exclusion, 421 domaines sont certifiés sur ce même ensemble et le FDR chute à 12%. Ainsi, à un même FDR, on certifie un peu plus de domaines avec la procédure d'exclusion qu'avec l'algorithme CODD original.

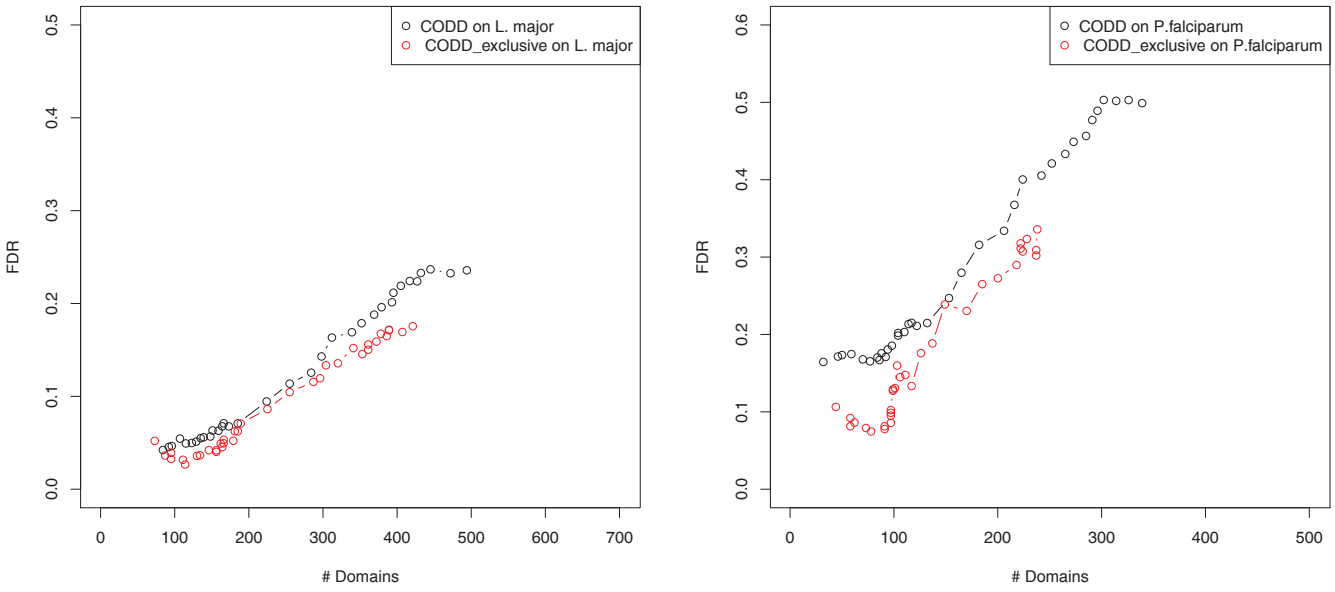


FIGURE 5.7.: Nombre de nouveaux domaines certifiés avec les différents seuils de FDR par les domaines potentiels chez *P.falciparum* (droite) et *L. major* (gauche) avec (points rouges) et sans (points noirs) utilisation des exclusions de domaines

5.3. Études des combinaisons de domaines par les règles d'association

Comme on l'a vu, dans la méthode CODD originale, les certifications sont basées sur les paires de domaines co-occurrentes identifiées grâce au test exact de Fisher dans les protéines d'Uniprot. Cette modélisation simple de la co-occurrence produit de bons résultats, mais elle pourrait sans doute être améliorée sur certains points. Un point semble particulièrement important qui concerne le sens d'implication des certifications. Dans la version actuelle, un couple (X,Y) de la CDP peut servir à certifier la présence d'un domaine X grâce à la présence d'un domaine Y ou le contraire. Il y a une symétrie des règles de certification. Or cette symétrie n'est pas forcément pertinente. On peut en effet imaginer que le domaine X soit toujours accompagné du domaine Y mais que l'inverse ne soit pas forcément vrai. Dans ce cas, certifier Y avec X est moins risqué que certifier X avec Y. Nous avons donc cherché une solution à ce problème, et il nous a semblé qu'une approche basée sur les règles d'association pouvait répondre à cette question.

5.3.1. Les règles d'association en bref

Les règles d'association sont des implications conditionnelles qui permettent de découvrir des relations/corrélations intéressantes entre des attributs (appelés aussi "items") (Han and Kamber, 2001). Dans notre cas, ces attributs sont les domaines. Les règles d'association classiques (positives) sont de la forme $X \rightarrow Y$ où X et Y sont des ensembles d'attributs. X et Y sont appelés respectivement prémisse et conclusion de la règle. Elles peuvent donc mettre en évidence des associations entre deux domaines ou plus.

Une règle d'association peut être définie formellement comme :
- I= $\{i_1, i_2, ..., i_m\}$ un ensemble d'items
- T= $\{t_1, t_2, ..., t_n\}$ un ensemble de transactions telles que t_i soit un sous-ensemble de I ($t_i \subseteq I$)
- Une règle d'association R s'exprime sous la forme :
  $R = X \rightarrow Y$, où $X \in T$, $Y \in T$, et $X \cap Y = \emptyset$

Deux métriques sont associées aux règles et permettent de vérifier leur validité :
- Le "support" mesure la fréquence d'apparition de X et Y ensemble sur l'ensemble des données. Le support de R est égal au support de $X \cup Y$.
- La "confiance" qui mesure la force d'une règle. La confiance de la règle R, donnée par la formule 5.2, est calculée à partir du support de $X \cup Y$ (le nombre d'apparitions de X et Y ensemble) rapporté au support de X (le nombre d'apparitions de X). Une règle ayant une confiance égale à 1 est appelée "règle exacte". Si non, elle est appelée "règle approximative" (Pasquier, 2000).

$$confiance(X \rightarrow Y) = \frac{Supp(X \cup Y)}{Supp(X)} \tag{5.2}$$

L'extraction de règles d'association requiert dans une première étape le choix d'un seuil minimal de support ($minsup$). Les algorithmes classiques de règles d'association permettent d'étudier les corrélations entre les objets qui sont fréquents (ayant un $support \geq minsup$). Dans la littérature, le problème d'extraction des règles d'association est décomposé en deux sous-problèmes qui sont (Han and Kamber, 2001; Agrawal et al., 1993) :
- la recherche des ensembles fréquents d'items ;

– la génération des règles d'association à partir de ces ensembles.

Plusieurs algorithmes permettant l'extraction de ce genre de règles ont été développés durant les dernières années. L'algorithme *APriori* (Agrawal and Skirant, 1994), introduit en 1994 par Agrawal, demeure l'un des algorithmes les plus utilisés. Il sert à reconnaitre des propriétés qui reviennent fréquemment dans un ensemble de données et d'en déduire une catégorisation. L'algorithme *APriori* se déroule en deux principales étapes :

– détermination de l'ensemble IF de tous les itemsets fréquents :
$IF = \{X_i \subseteq T | supp(X_i) \geq minsupp, i = 1, 2, ..n\}$
– extraction des règles d'associations ayant une confiance supérieure au seuil $minconf$ entre itemsets fréquents :
$\{X_i, Y_j \subseteq IF | Conf(X_i \rightarrow Y_j) \geq minconf; i = 1, 2, ..p \ j = 1, 2, ..q\}$

Comme pour l'indice du support minimum $minsupp$, le seuil minimal de confiance $minconf$ doit être donné en entrée comme paramètre de l'algorithme.

5.3.2. Extraction de règles fréquentes : Limites de la métrique $minsup$ sur les bases de données biologiques

Nous nous sommes intéressés dans une première étape à l'extraction des règles d'association entre domaines fréquents à partir de la composition en domaines de l'ensemble des protéines Uniprot, tout organismes confondus (le même ensemble utilisé pour l'extraction de la CDP (*cf.* section 4.2.1)). Cet ensemble contient la composition en domaines d'environ 10 millions de protéines différentes. Plusieurs valeurs de $minsupp$ ont été testées. Le support le plus bas qu'on ait pu atteindre avec les algorithmes d'extraction de règles était de 0.005%. Pour être fréquent, un couple de domaines doit apparaître dans au moins 500 protéines sur les 10 millions. Cette valeur de $minsupp$ nous a permis d'extraire $\simeq$ 1187 règles mettant en relation deux domaines et ayant un $minconf \geq 60\%$. Nous avons alors lancé une procédure de certification sur les données de *L. major* en substituant à la CDP les règles d'association extraites. Malheureusement, le nombre de certifications obtenues par cette approche est loin des résultats obtenus grâce à la CDP. Au lieu des 691 domaines certifiés grâce à la CDP, seuls 144 domaines peuvent être certifiés par cette approche, pour un FDR de 6%. Cette perte en termes de nombre de domaines certifiés provient du fait que la CDP contient un nombre beaucoup plus important de couples que les 1187 règles extraites grâce aux algorithmes d'extraction classiques (la CDP contient 18804 paires de domaines lorsque l'on utilise une P-valeur de 0.01 dans le test de Fisher). Un examen minutieux de l'ensemble des couples figurant dans la CDP nous montre que la plupart ($\simeq$ 93%) sont non fréquents (présents sur moins de 500 protéines). Afin d'enrichir l'ensemble des corrélations apportées par les règles, nous avons également extrait des règles d'association mettant en jeu plus de deux domaines, c'est à dire les règles de la forme $X, Y \rightarrow Z$ ou encore $X, Y, Q, P \rightarrow Z$. Celles-ci ont permis de certifier un petit nombre de domaines supplémentaires, mais sans atteindre le nombre de certifications réalisées par la CDP. Pour *L. major* par exemple, ces règles ont permis d'ajouter 5 domaines au total (avec une E-value seuil égale à 100) sur l'ensemble des protéines.

L'extraction des associations fréquentes à partir d'une base de données contenant des millions de protéines variées n'est pas triviale. La stratégie des méthodes d'extraction classiques basée sur les ensembles d'items fréquents se révèle donc mal adaptée pour ce genre de base de données très denses. Ainsi, seules les associations très fréquentes (présentes sur au moins 500 protéines différentes) ont pu être extraites. Une examination manuelle rapide de ces règles a montré qu'elles concernent généralement des domaines abondants et qui sont communs à plusieurs règnes du vivant tels que les domaines impliqués dans les interactions entre protéines. Pour pouvoir extraire des informations pertinentes dans les bases de données de domaines,

le support doit être beaucoup plus faible, ce qui pose un problème de temps calcul aux programmes d'extraction classiques.

5.3.3. Extraction des associations entre domaines

Pour remédier à la limite évoquée ci-dessus, nous avons décidé d'ignorer la métrique *minsup* et d'extraire les règles d'association en partant des paires de domaines identifiées par le test de Fisher. L'approche que nous avons suivie est la suivante : à partir d'une paire (d, d') de la CDP, deux règles différentes peuvent être générées : $d \rightarrow d'$ et $d' \rightarrow d$. On calcule alors la confiance $conf$ associée à chaque règle. Si $conf$ est supérieure à un certain seuil fixé de confiance, la règle est adoptée. Les différentes étapes empruntées pour générer l'ensemble des règles d'association entre domaines (noté ER) sont décrites par l'algorithme 6. Ce dernier prend en paramètre une valeur seuil $Spval$ qui désigne le seuil de P-valeur, une valeur seuil de la confiance $minconf$ ainsi que l'ensemble $DOMs$ de tous les domaines Pfam. La première étape consiste à créer la CDP en utilisant la valeur seuil $Spval$ dans le test de Fisher. Seules les paires de domaines figurant dans la CDP sont utilisées dans la deuxième étape dans laquelle on calcule la confiance $conf$ associées aux différents règles possibles. Si $conf$ est supérieure à la confiance minimale $minconf$ donnée en entrée, alors la règle est adoptée.

Algorithm 6: Extraction des RA entre domaines

Inputs : composition en domaines de l'ensemble des protéines Uniprot, $minconf$, $Spval$, $DOMs$
Output : ensemble ER des règles d'associations
/* ***Calcul de la*** CDP ***par le test de*** $Fisher$ */
foreach $d \in DOMs$ **do**
    foreach $d' \in DOMs$ **do**
        P-value$(d, d') \leftarrow Fisher(d, d')$
        If (P-value$(d, d') \leq Spval$) **then** ajouter (d, d') à CDP
/* ***Estimation des confiances sur la*** CDP */
foreach $(d, d') \in CDP$ **do**
    Calculer $conf$ $(d \rightarrow d')= \frac{Supp(d \cup d')}{Supp(d)}$
    calculer $conf$ $(d' \rightarrow d)= \frac{Supp(d \cup d')}{Supp(d')}$
    If $(conf(d \rightarrow d') \geq minconf)$ **then** ajouter $(d \rightarrow d')$ à ER
    If $(conf(d' \rightarrow d) \geq minconf)$ **then** ajouter $(d' \rightarrow d)$ à ER
    Endif
Return ER

5.3.4. Certification par les règles d'association

Une fois appris, l'ensemble ER des règles générées est utilisé à la place de la CDP dans la phase de certification. Un domaine X peut alors être certifié par un domaine Y si la règle $Y \rightarrow X$ est présente dans notre ensemble de règles. Afin d'évaluer l'apport de l'utilisation de ces règles, nous avons effectué des tests sur l'ensemble des protéines des deux pathogènes d'intérêt : *Leishmania major* et *Plasmodium falciparum*. Pour la création de la CDP, nous avons fixé

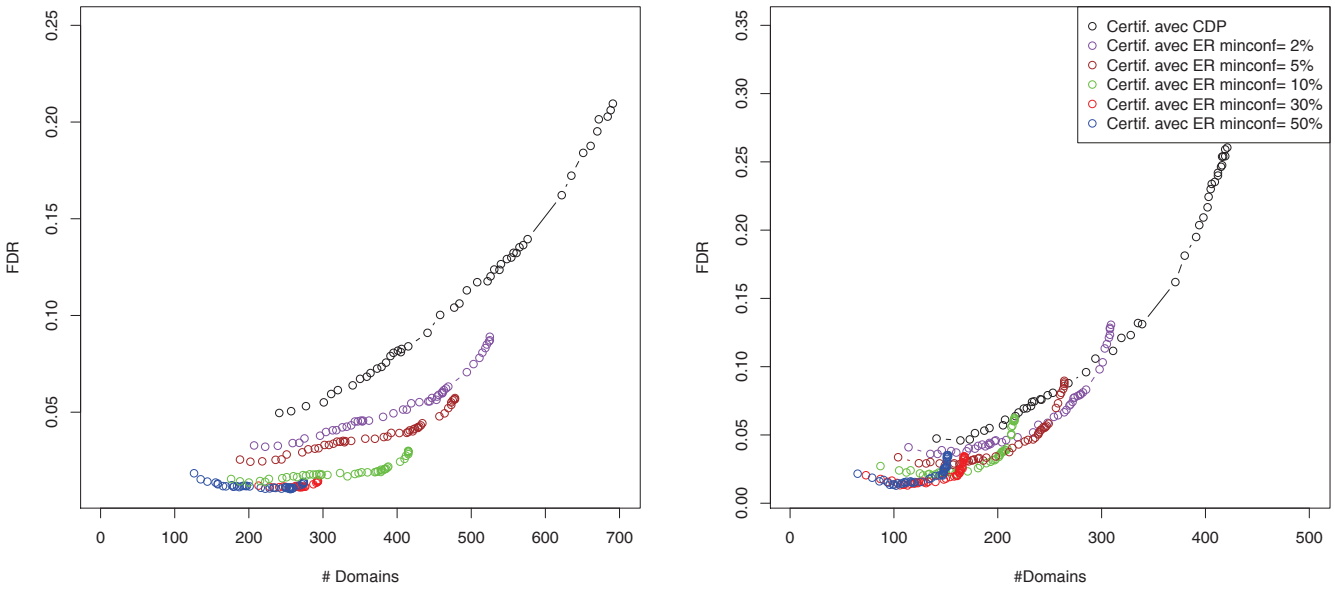


FIGURE 5.8.: Nombre de nouveaux domaines certifiés avec les domaines Pfam connus en utilisant les trois ensembles de règles de certification chez *P. falciparum* (droite) et *L. major* (gauche) en comparaison avec l'utilisation de la CDP (en noir)

la valeur seuil *Spval* à 0.01. Nous avons fait varier le seuil $minconf$ suivant cinq valeurs (2%, 5%, 10%, 30% et 50%) pour former cinq ensembles différents de règles de certification. Comme dans toutes les expérimentations précédentes, nous avons fait varier le seuil de scores afin de former plusieurs ensembles de domaines potentiels de différentes tailles et pour chaque ensemble, le FDR associé aux certifications a été estimé. La figure 5.8 récapitule les résultats de certification réalisés avec les domaines Pfam connus en utilisant les cinq ensembles de règles. L'axe des abscisses représente le nombre de domaines nouvellement découverts, tandis que l'axe des ordonnées représente les FDR associés à ces prédictions. Comme nous pouvons le remarquer sur ces différentes courbes, un gain considérable en termes de FDR est observé avec l'utilisation des règles ER pour la certification par rapport à l'utilisation de la CDP. Plus la valeur de $minconf$ augmente et plus le nombre de règles dans ER diminue. Ainsi, le nombre de domaines certifiés décroît avec l'augmentation de $minconf$. Cependant, les différences au niveau des FDRs obtenus sont légères notamment pour les trois dernières valeurs seuils (10%, 30% et 50%).

Nous pouvons remarquer qu'avec les deux plus faibles valeurs de $minconf$ (5% et 2%), la courbe se rapproche des certifications obtenues par la CDP ce qui est complètement attendu. Le paramètre $minconf$ semble donc être un paramètre très intéressant pour ajuster le niveau de précision requis par l'utilisateur.

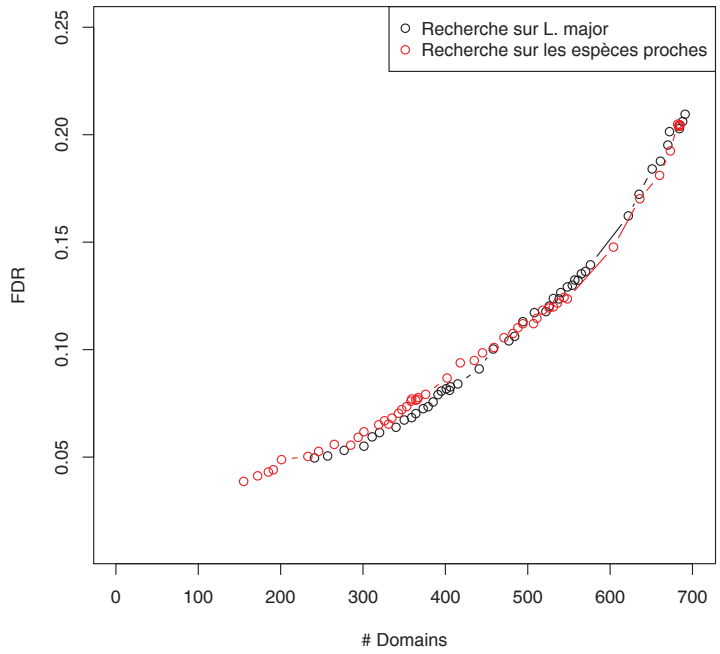


FIGURE 5.9.: Comparaison des certifications obtenues avec la recherche sur les espèces proches *Leishmania* avec celles obtenues sur avec *L.major* seul

5.4. Utilisation des espèces proches pour l'amélioration des reconnaissances de domaines

Nous terminons ce chapitre par la description d'une tentative d'amélioration de la procédure basée sur l'utilisation d'espèces proches lors de la phase d'identification des domaines potentiels. Ces expériences ont été menées sur les protéines de *Leishmania*. L'idée consiste à effectuer une recherche parallèle sur les protéines orthologues provenant des trois espèces *L. major*, *L. infantum* et *L. braziliensis*. Bien qu'elles causent diverses formes d'infections, ces espèces partagent plus de 90% de leurs gènes et présentent une très forte homologie entre leurs séquences génomiques et protéiques (*cf.* section 3.1.2) (Peacock et al., 2007). De ce fait, l'observation conjointe des protéines orthologues peut être une stratégie efficace lors de l'identification des domaines potentiels. L'approche que nous avons suivie est la suivante :

1. Identification des homologues des protéines de *L. major* dans les deux autres espèces de *Leishmania*.
2. Création d'un alignement multiple pour chaque protéine de *L. major*. 91% des protéines ont des orthologues connus chez *L. infantun* et *L. braziliensis*.
3. Exécution d'HMMER sur les protéines des trois espèces avec une E-valeur seuil égale à 100. Chaque séquence protéique est comparée à l'ensemble des HMM profils de la banque Pfam.
4. Identification des hits communs sur les protéines homologues : mêmes domaines au même

endroit (avec +/-10% de décalage).

5. Attribution à chaque domaine identifié chez *L. major* un score égal au produit des P-valeurs obtenues sur les différentes protéines homologues. Si un domaine n'est pas identifié sur une de ces protéines, il est considéré comme ayant une P-valeur de 1.
6. Détermination d'un ensemble non chevauchant de domaines potentiels en choisissant dans chaque conflit celui ayant le meilleur score.

Nous avons ensuite certifié par co-occurrence avec les domaines Pfam connus les domaines potentiels ainsi identifiés. Les résultats obtenus sont illustrés par la figure 5.9. Comme on peut le voir sur cette figure, notre approche pour essayer de tirer partie des protéines homologues chez les espèces proches ne permet pas d'améliorer la détection des domaines de manière significative. Une raison possible à cela est que, dans notre approche, les domaines sont cherchés de manière indépendantes sur chaque protéine homologue. C'est le résultat de ces recherches qui est ensuite combiné à travers le produit des P-valeurs. Une manière d'aller plus loin serait donc d'utiliser une méthode de recherche de domaines qui puissent utiliser directement l'alignement multiple des protéines homologues. Cette approche fait l'objet du chapitre suivant, et, comme nous allons le voir, elle permet effectivement des gains importants en terme de nombre de domaines certifiés et de FDR.

6. Une nouvelle approche basée sur les comparaisons profil-profil

QUAND un nouveau génome a été séquencé et qu'il faut annoter ses protéines, les méthodes classiques basées sur les comparaisons séquence-profil sont généralement utilisées. Ces dernières années, un nouveau type d'approches basées sur des comparaisons profil-profil a été développé. Ces approches, souvent utilisées pour la prédiction de la structure 3D des protéines ne sont jamais utilisées à l'échelle d'un génome. Nous montrons dans ce chapitre que ces méthodes ont pourtant une meilleure sensibilité que les méthodes séquence-profil. Cependant, elles ne sont pas suffisamment précises pour être utilisées telles quelles. Nous proposons donc une nouvelle approche basée sur la combinaison des algorithmes de comparaison profil-profil avec les informations de co-occurences de domaines.

Sommaire

6.1. Introduction **95**
6.2. HHPRED à l'échelle d'un génome **95**
6.3. Les méthodes profil-profil manquent de sensibilité **97**
6.4. Combiner les comparaisons HMM-HMM et la détection par co-occurrence **99**
6.5. Amélioration de la politique d'inférence des annotations GO **100**
6.6. Analyse des nouveaux domaines prédits **103**
6.6.1. Domaines découverts chez *P. falciparum* 106
6.6.2. Domaines découverts chez *L. major* 107

6.1. Introduction

Comme on l'a vu dans les chapitres précédents, les HMM sont utilisés la plupart du temps dans des approches de type "séquence-profil" : lorsqu'on analyse une nouvelle séquence protéique, chaque HMM profil de Pfam est utilisé pour mesurer la ressemblance entre la séquence et le domaine protéique via le logiciel HMMER. Si le score est au dessus du seuil de score recommandé par Pfam, la présence du domaine dans la protéine est confirmée (*cf.* section 2.2.4). Cependant, quand elle est appliquée sur des protéines divergentes, cette stratégie manque de sensibilité et de nombreux domaines ne sont pas reconnus. C'est le cas pour la majorité des pathogènes Eukaryotes pour lesquels $\simeq 80\%$ des acides aminés (toutes protéines confondues) ne sont couverts par aucun domaine. Une amélioration significative des approches séquence-profil a été le développement de méthodes profil-profil (Soding and Remmert, 2011; Dlakic, 2009). Dans ce genre d'approches, les profils ne modélisent pas uniquement les familles de domaines mais aussi la séquence protéique cible. Il a été démontré sur certains cas que les approches profil-profil sont plus efficaces et arrivent à détecter des homologies lointaines que les méthodes séquence-profil n'arrivent pas à reconnaître (Soding, 2005). En effet, un profil construit à partir d'un alignement de protéines homologues permet de pondérer les informations apportées par chaque position de la séquence protéique cible, en distinguant les positions conservées de celles qui ne le sont pas (Dlakic, 2009). HHPRED est une des méthodes proposées récemment pour la comparaison de HMM profils (Soding, 2005). Elle permet de comparer un HMM profil construit à partir d'un alignement de protéines à une base de HMM comme Pfam. Bien que cette approche soit largement utilisée pour la prédiction de structure, et spécifiquement pour l'identification d'homologues lointains avec une structure 3D connue (Soding and Remmert, 2011; Soding, 2005), ce type d'approche n'est pas utilisé pour l'identification de domaines protéiques sur un génome entier. La plus-value que peuvent apporter ces méthodes pour ce genre de prédictions n'a jamais été étudiée, et pour cette tâche les méthodes séquences-profil demeurent un standard.

Nous étudions dans ce chapitre ce que ce genre d'approche peut apporter par rapport aux méthodes séquence-profil classiques. Plus précisément, notre but est d'utiliser HHPRED pour identifier de nouvelles occurences de domaines non identifiées par HMMER. Nous montrons que HHPRED est plus sensible que HMMER pour identifier des occurrences divergentes, mais que la méthode n'est pas suffisamment précise pour être utilisée telle qu'elle en complément de HMMER. Ainsi, un post-traitement est nécessaire pour pouvoir distinguer entre les vrais et les faux positifs. Nous proposons alors d'utiliser les informations de co-occurences de domaines pour cela.

6.2. HHPRED à l'échelle d'un génome

Dans la suite, tous les domaines Pfam identifiés par HMMER en se basant sur les seuils standard de détection sont considérés comme "connus" et notre but est d'identifier grâce à HHPRED de nouvelles occurrences de domaines. Pour cela, la première étape consiste à construire à partir de chaque séquence requête un HMM profil. Chaque séquence protéique est alors utilisée pour construire un alignement multiple de séquences. Pour ceci, deux approches ont été utilisées : la première consiste à utiliser seulement les protéines homologues de certaines espèces proches, alors que la deuxième utilise les homologues séquencés de toutes espèces confondues, à travers la méthode HHBLITS (Remmert et al., 2012).

Approche spécifique aux Phylums Pour *L. major*, six espèces sont incluses : quatre espèces *Leishmania* (*L. major*, *L. infantum*, *L. braziliensis* et *L. mexicana*) et deux espèces

Trypanosoma (*T. cruzi* et *T. brucei*) (http ://tritrypdb.org/tritrypdb/). Pour *P. falciparum*, six espèces sont analysées : *P. falciparum*, *P. vivax*, *P. yoelii*, *P. berghei*, *P. chabaudi* et *P. knowlesi*. Pour chaque protéine de *L. major* et de *P. falciparum*, nous déterminons ses homologues chez les espèces proches à partir de la base OrthoMCL (Chen et al., 2006). 99.4% des protéines de *L. major* et 99.2% des protéines *P. falciparum* ont au moins un homologue dans les autres espèces du phylum. La majorité des protéines de *L. major* (95%) ont un orthologue dans les trois autres espèces de *Leishmania*. Pour *P. falciparum*, 72% des protéines ont des homologues dans les cinq autres espèces de *Plasmodium*. Quand une protéine possède des paralogues dans la même espèce, (16% des protéines *P. falciparum* et 15% des protéines *L. major*), ces paralogues sont également utilisés. La méthode Muscle (Edgar, 2004) est ensuite utilisée pour créer un alignement multiple à partir des protéines sélectionnées en utilisant le paramétrage par défaut.

Approche non-spécifique aux Phylums (HHblits) HHblits est un outil du package HH-suite permettant de construire un alignement multiple à partir d'une séquence cible. Il procède de manière analogue à Psi-blast en alignant itérativement de nouvelles séquences homologues à la séquence cible. HHblits a été exécuté avec les valeurs des paramètres par défaut : 3 itérations, mode d'alignement local, et seuil de E-valeur égal à $1.e^{-3}$. Nous avons par contre fixé le seuil minimum d'identité de séquences avec la séquence requête à 20% au lieu de la valeur par défaut de 0% pour s'assurer que seuls les vrais homologues sont inclus dans l'alignement.

Les alignements générés dans la première étape sont transformés en HMM profils en utilisant la procédure *hhmake* de HHPRED. *hhmake* a été exécutée avec les paramètres par défaut, sauf pour la valeur maximale d'identité qui a été fixée à 100% pour prendre en considération la grande similarité des séquences homologues, notamment chez les espèces *Leishmania*. Chaque HMM a été ensuite comparé à l'ensemble des HMM de la base de données Pfam en utilisant la procédure *hhsearch*. Cette procédure calcule un score pour chaque paire de HMM (Soding, 2005) qui est une adaptation du *log-odds* score utilisé dans les comparaisons séquence-profil (Eddy, 1996). Il maximise la probabilité de co-émission, *i.e* la probabilité que les deux HMM émettent la même séquence de résidus (Soding, 2005). Les alignements sont faits dans les deux modes : global (*i.e.* l'alignement s'étend du début à la fin du HMM profil qui représente le domaine Pfam) et local (*i.e.* les alignements de fragments de domaines sont autorisés). Comme pour la majorité des programmes de recherche de similarités, la significativité du score est estimée via une E-valeur représentant l'espérance du nombre de séquences pouvant obtenir un aussi bon score que la séquence requête parmi un ensemble de séquences aléatoires de même taille. Toutes les occurrences avec une E-valeur au dessous d'un seuil fixé sont considérées dans la suite. Chacune de ces occurrences correspond à un alignement entre une partie du HMM de la protéine et une partie (en mode local) ou la totalité (en mode global) du HMM du domaine. Nous déterminons à partir de cet alignement de deux HMM, l'alignement entre la séquence cible et le HMM du domaine, et nous en déduisons la position du domaine. Ensuite la procédure classique de traitement des chevauchements de domaines (*cf.* chapitre 4) est appliquée : toutes les occurences chevauchants un domaine connu de la protéine sont écartées en premier. Quand deux nouveaux domaines se chevauchent, le domaine qui a la E-valeur la plus élevée (*i.e.* celui qui est le moins probable) est supprimé. Les domaines restants sont considérés comme les domaines "potentiels" de la protéine.

6.3. Les méthodes profil-profil manquent de sensibilité

Nous avons dans une première étape cherché à estimer la qualité des prédictions obtenues par HHPRED en utilisant différentes valeurs seuils de E-valeur comprises entre 0.001 et 50. Nous avons pour cela utilisé une procédure simple proposée dans (Terrapon et al., 2012) pour estimer le taux de faux positifs dans un ensemble de domaines prédits. Comme la procédure de certification utilisée pour la création de la base EuPathDomains, cette méthode (détaillée ci-dessous) est basée sur la tendance des domaines protéiques d'apparaître préférentiellement avec certains autres domaines favoris (Cohen et al., 2007).

Estimation du FDR Soit $\mathcal{L}$ un ensemble de nouvelles occurrences de domaines (identifiées par HMMER ou HHPRED). Nous souhaitons estimer le FDR associé à $\mathcal{L}$ *i.e.* la probabilité $Pr(d = \text{faux})$ pour $d \in \mathcal{L}$. La première étape consiste à identifier les paires de domaines fortement co-occurrentes à partir de la composition en domaines des protéines Uniprot comme décrit dans la section 4.2.1 du chapitre 4. Les paires ayant une P-valeur inférieure à 0.01 sont sélectionnées pour former l'ensemble C des paires de domaines co-occurrentes (CDP). Ensuite, à partir des protéines de l'espèce cible contenant à la fois des domaines connus et des nouveaux domaines prédits, une liste L de couples (connus-prédits) est formée. Ces couples sont formés aléatoirement en associant à chaque nouveau domaine un autre domaine connu de la même protéine. Notons (d_k, d_n) une paire de domaines (connu, nouveau) de la liste L. Cette liste L est utilisée pour estimer le FDR de $\mathcal{L}$. Nous supposons que la proportion de faux positifs parmi les nouveaux domaines d_n de L est la même que pour tous les domaines de $\mathcal{L}$. Ceci suppose en particulier que la proportion de faux positifs parmi les domaines des protéines multi-domaines (ceux qui sont dans L) est la même que sur les protéines mono-domaines (qui ne sont pas dans L). Même si les domaines des protéines mono-domaines diffèrent souvent des protéines multi-domaines, ils partagent d'une manière générale la même composition en acides aminés et il n'y a pas de raison de penser que les HMM sont plus sensibles aux erreurs pour l'un ou l'autre de ces deux types de domaines. Soit $|L|$ le nombre de paires dans L. Nous notons $\mathcal{T}$, la probabilité qu'une paire de L appartienne à l'ensemble CDP, dans l'hypothèse où le domaine potentiel est un vrai positif. De la même manière, $\mathcal{F}$ est la probabilité qu'une paire de L appartienne à C, dans l'hypothèse où le domaine potentiel est un faux positif. L'espérance du nombre de paires de L qui appartiennent à C peut être estimé de la manière suivante :

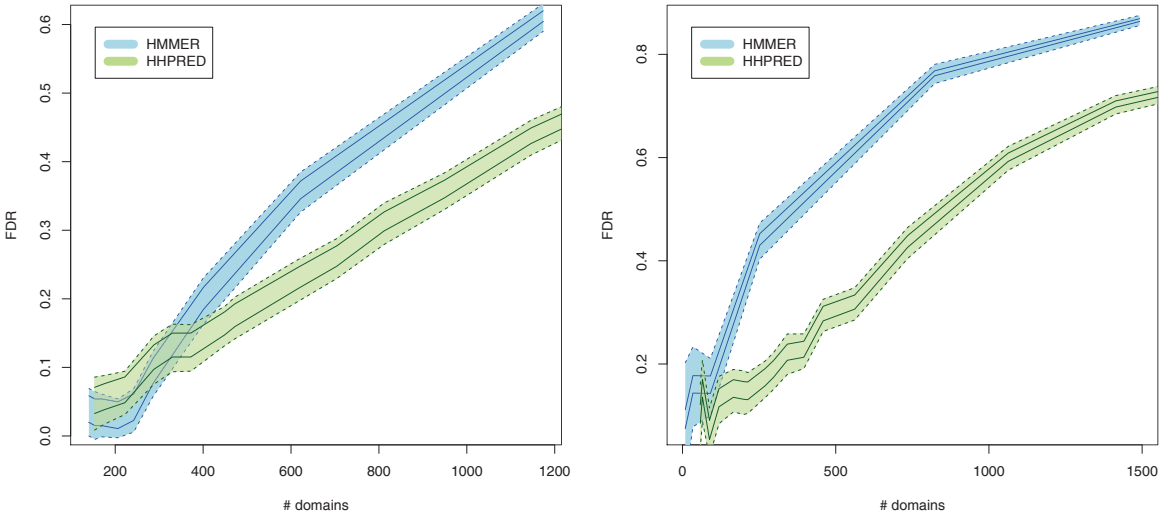
$$\begin{aligned} E[|L_C|] &= |L| \cdot Pr((d_n, d_k) \in C) \\ &= |L| \cdot \Big(Pr((d_n, d_k) \in C | d_n = \text{true}) \cdot Pr(d_n = \text{true}) + \\ &\qquad Pr((d_n, d_k) \in C | d_n = \text{false}) \cdot Pr(d_n = \text{false}) \Big) \\ &= |L| \cdot \Big(\mathcal{T} \cdot (1 - \text{FDR}) + \mathcal{F} \cdot \text{FDR} \Big). \end{aligned}$$

Nous avons alors :

$$\text{FDR} = 1 - \frac{\frac{E[|L_C|]}{|L|} - \mathcal{F}}{\mathcal{T} - \mathcal{F}}. \tag{6.1}$$

$|L|$ est connu et $E[|L_C|]$ est estimé par le nombre de paires de L qui appartiennent à C. Pour $\mathcal{F}$, une liste L' est créée en permutant aléatoirement les nouveaux domaines des paires de L. Ceci équivaut à la permutation aléatoire des nouveaux domaines sur les protéines de l'espèce cible et crée ainsi une situation où la majorité des domaines sont probablement des faux positifs. $\mathcal{F}$ est estimé par la proportion de paires de L' qui sont dans C. La procédure est répétée plusieurs fois et une moyenne est calculée pour obtenir une meilleure estimation. Pour $\mathcal{T}$, nous utilisons

P. falciparum



L. major

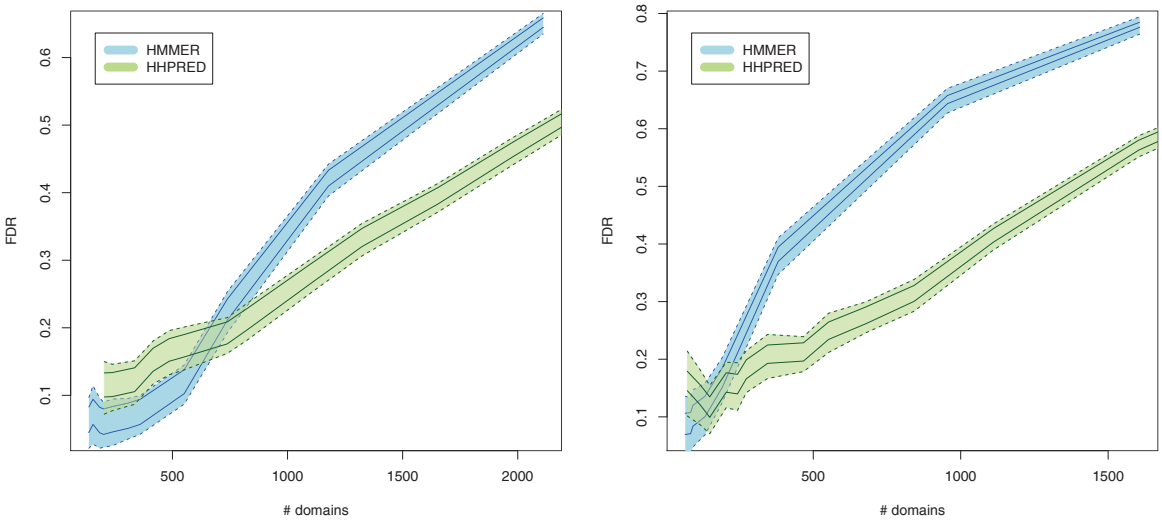


FIGURE 6.1.: **Comparaison de la sensibilité et de la précision de HHPRED et de HMMER sur les génomes de *P. falciparum* et *L. major*.** L'axe des abscisses représente le nombre de nouveaux domaines identifiés par HHPRED (vert) et HMMER (bleu) en mode local (gauche) et global (droite) pour différentes valeurs de FDR (axe des ordonnés). Pour chaque approches, les deux lignes continues représentent les valeurs minimales et maximales estimées pour le FDR. Les lignes discontinues représentent l'erreur standard associée à ces deux estimations.

les occurrences connues des domaines. Une liste L'' est crée à partir des paires de domaines (connu, connu) observées sur les protéines ayant au moins deux domaines connus. Ceci permet de simuler une situation où tous les domaines sont des vrais positifs et $\mathcal{T}$ est estimée par la proportion de paires L'' qui sont dans C. $\mathcal{F}$ et $\mathcal{T}$ ont des valeurs estimées très différentes. La valeur de $\mathcal{F}$ est comprise entre 1% et 2%, indépendamment de la méthode et du seuil de E-valeur utilisé pour les prédictions. Pour $\mathcal{T}$, nous avons généré plusieurs listes de paires de domaines (connu, connu) et observé que les valeurs de $\mathcal{T}$ obtenues varient entre 96% et 99%. Bien que ces deux valeurs soient relativement proches, elles peuvent conduire à des estimations très différentes du FDR, en particulier lorsque celui-ci est faible. Nous proposons donc deux estimations du FDR : une calculée avec $\mathcal{T} = 96\%$ et l'autre avec $\mathcal{T} = 99\%$.

Cette procédure d'estimation du FDR a été utilisée sur l'ensemble des domaines potentiels prédits par HHPRED à différents seuils de E-valeurs. Pour comparaison, nous avons également utilisé HMMER avec différents seuils de E-valeurs, et la même procédure de résolution des chevauchements a été appliquée. Comme discuté au chapitre 2 (section 2.2.4), la dernière version de HMMER (HMMER3) ne gère que les alignements locaux. Pour les alignements globaux nous avons donc utilisé la version précédente du programme. Cependant, comme cette version n'est pas compatible avec les HMM des dernières versions de Pfam, les comparaisons en mode global ont été faites sur la version 23 de Pfam. . La figure 6.1 résume le nombre de nouveaux domaines identifiés à un seuil de FDR donné par HHPRED et HMMER avec les deux modes d'alignement et les deux méthodes de construction des HMM de protéines (spécifique et non-spécifique aux Phylums) sur les deux espèces. Comme on s'y attendait, les deux approches permettent de découvrir de plus en plus de domaines à mesure que le seuil de E-valeur augmente mais au prix d'une augmentation du FDR. Pour les seuils faibles de E-valeur, le nombre de nouveaux domaines n'est pas assez élevé pour donner une estimation précise du FDR et il est difficile d'évaluer précisément les deux approches pour ce genre de valeurs. Cependant, sur les protéines de *L. major* les prédictions de HHPRED ne semble pas passer sous le seuil de 10%. Notons que cette observation concerne un cas bien particulier : tous les domaines bien conservés sont déjà connus, et la méthode est mise au défi de reconnaître les domaines divergents uniquement. Ceci n'implique donc pas que HHPRED manque de précision dans le cas général. Nous remarquons également à partir les courbes de la figure 6.1 que pour les seuils modérés et élevés de E-valeurs, les résultats des deux approches sont très différents et qu'à même seuil de FDR, HHPRED détecte plus de domaines que HMMER. Bien que HHPRED manque de précision sur les petites E-valeurs, il semble donc plus sensible que HMMER pour les seuils de E-valeurs modérés et élevés. À condition d'être capable de filtrer les faux positifs, on peut donc espérer identifier de nombreux nouveaux domaines avec cette procédure. À l'échelle d'un génome, une telle procédure de filtre doit être complètement automatique, et nous proposons donc d'utiliser la procédure CODD pour cela.

6.4. Combiner les comparaisons HMM-HMM et la détection par co-occurrence

Comme nous l'avons montré, bien que HHPRED semble être plus sensible que HMMER pour détecter les domaines divergents, il n'est pas suffisamment précis pour être utilisé tel quel comme un outil d'annotation complémentaire. Un post-traitement est donc nécessaire pour distinguer les vrais et faux positifs. Nous proposons pour cela d'utiliser les propriétés de co-occurrence des domaines. Comme nous l'avons présenté dans les chapitres précédents, CODD (Terrapon et al., 2009) est une approche algorithmique qui permet de sélectionner les

occurrences les plus probables parmi un ensemble d'occurrences de domaine potentiels, tout en contrôlant le taux de fausses découvertes associé aux prédictions (*cf.* section 4.2.4).

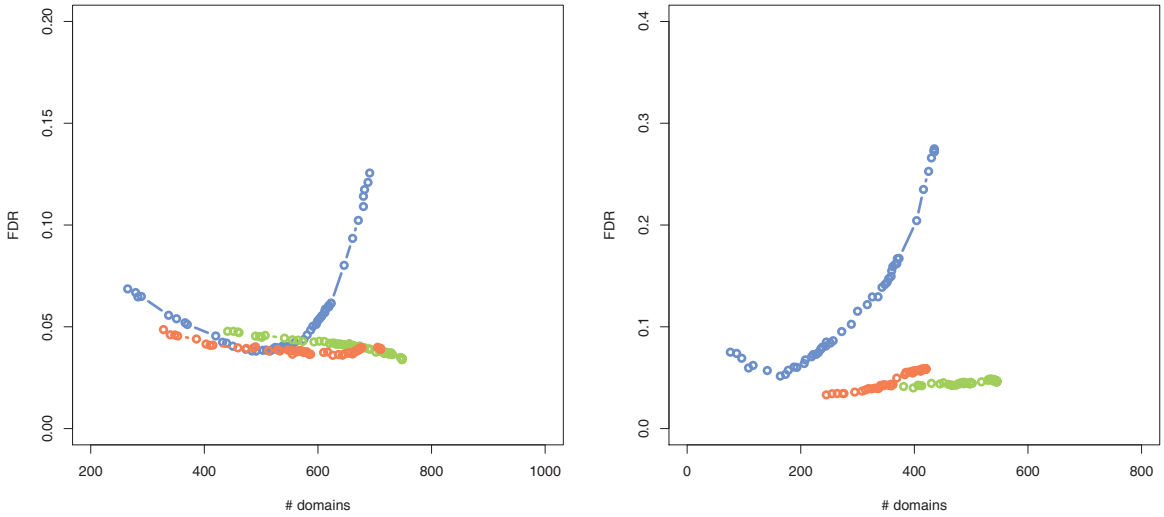
Dans une première série d'expériences, la procédure CODD a été utilisée sur l'ensemble des domaines prédits par HHPRED (pour les deux approches : spécifique aux Phylums et non-spécifique aux Phylums) en utilisant les domaines Pfam déjà connus pour la certification. Ceci a été effectué pour les deux types d'alignement (mode local et mode global). La figure 6.2 résume les résultats obtenus sur *L. major* et *P. falciparum* avec les deux modes et en variant les seuils de E-valeur. Pour chaque seuil de E-valeur, nous avons calculé le nombre de domaines certifiés et estimé les FDR en utilisant la procédure décrite à la section 4.2.4. Ces résultats ont été comparés à ceux obtenus par l'application de CODD sur les domaines détectés par HMMER2 (mode global) et HMMER3 (mode local) pour différents seuils de E-valeur. Comme nous pouvons le constater sur la figure 6.2, quand il est couplé avec CODD, HHPRED fait beaucoup mieux que HMMER. De plus, la procédure CODD améliore la précision des résultats de HHPRED, et des FDR inférieurs à 5% sont maintenant associés aux prédictions. Les approches spécifique au phylum et non spécifique au phylums donnent des résultats similaires en terme de précision (*i.e.* des FDR similaires). Cependant, l'approche non-spécifique aux Phylums fait mieux en terme de sensibilité (elle arrive à détecter un plus grand nombre de domaines). Les mêmes tendances sont observées pour les deux autres types de certification (Interpro non-Pfam et potentiels comme le montrent les figures 6.3 et 6.4), mis à part que les FDR associés aux prédictions dépassent les 5% dans certains cas.

Le tableau 6.1 récapitule les résultats obtenus avec les différents types de certifications sur les deux espèces étudiées. Au total, 1119 et 1266 nouveaux domaines non redondants sont prédits avec un FDR de 5% sur les deux organismes. "Non redondant" signifie ici qu'une seule occurrence d'une famille de domaine est comptée sur chaque protéine. Les occurences de domaines déjà connus sur la protéines ne sont pas comptées et les occurences multiples d'une même famille sont comptées une seule fois. En comparaison avec les 5978 et 7002 domaines Pfam connus non-redondant sur ces organismes, ces chiffres correspondent à une amélioration de 27.6% et 23.8% respectivement pour *P. falciparum* et *L. major*. Nous avons constaté que la majorité (90% environ) des domaines certifiés par HMMER le sont aussi par HHPRED. Il est intéressant de noter que plusieurs domaines prédits n'étaient pas connus chez l'espèce étudiée (351 pour *L. major* et 350 pour *P. falciparum*), ce qui correspond à un enrichissement autour de 21% du catalogue de domaines chez les deux organismes.

6.5. Amélioration de la politique d'inférence des annotations GO

Les nouveaux domaines prédits permettent d'apporter de nouvelles annotations GO aux protéines. Certains domaines sont associés à des termes GO par le consortium d'Interpro. Comme on l'a vu au chapitre 4, la première version de la base EuPathDomains propose une extension de la politique d'Interpro pour associer des annotations GO à des combinaison de domaines (*cf.* section 4.2.5). La politique suivie est d'associer à un domaine (ou une combinaison de domaines) les annotations GO partagées par toutes les protéines annotées qui possèdent ce domaine (cette combinaison). Cependant, cette politique stringente peut manquer certaines associations domaine-annotation. En effet, une seule protéine qui possèderait un domaine D mais qui n'aurait par erreur pas été annotée par le terme A empêche l'association $D - A$. Nous avons alors choisi ici de relâcher le seuil de 100% et avons considéré toutes les associations $D - A$ où au moins 95% des protéines annotées ayant le domaine D partagent le terme A. Nous avons pu associer des annotations GO à 2466 domaines de cette manière et nous avons

P. falciparum



L. major

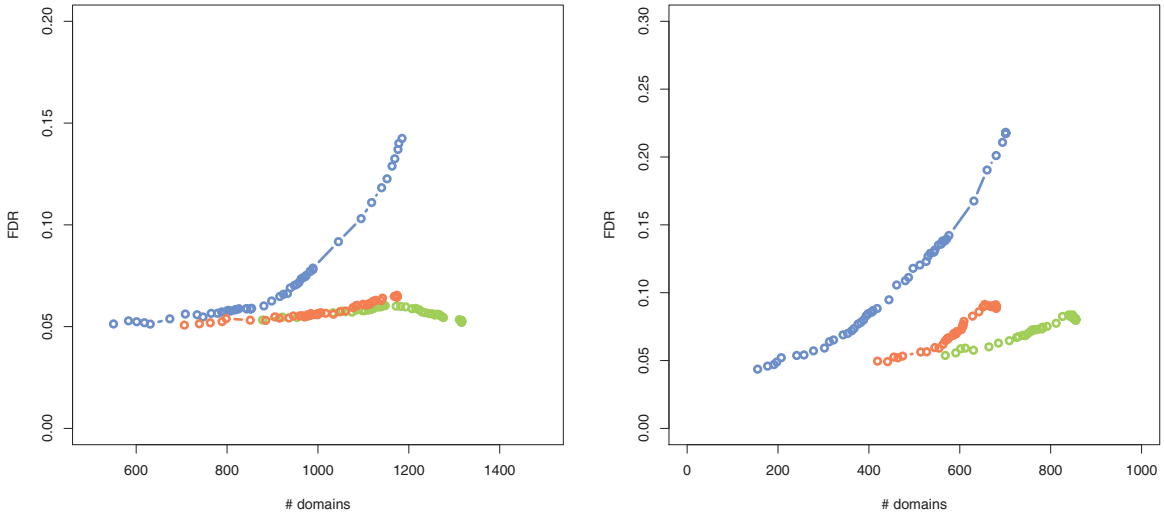


FIGURE 6.2.: **Comparaison de la sensibilité et de la précision de HHPRED + CODD et HMMER+CODD en utilisant les domaines Pfam connus pour la certification.** La figure rapporte le nombre de nouveaux domaines (axe des X) certifiés par HHPRED + CODD (en orange et vert respectivement pour l'approche spécifique au phylum et l'approche non-spécifique au phylum) et HMMER+CODD (bleue) en mode local (gauche) et global (droite) pour différentes valeurs seuils de FDR (axe des Y).

		Pfam	Interpro	Pot.	Tous
P. f	# dom.	703 / 703	- / 979	695 / 783	1119 (27.6%) / 1328 (32%)
	# fam.	247 / 247	- / 301	220 / 251	350 (21.4%) / 420 (25%)
L. m	# dom.	1266 / 1266	- / 664*	- / 1394	1266 (23.8%) / 2196 (41%)
	# fam.	351 / 351	- / 203*	- / 342	351 (22%) / 537 (33%)

TABLE 6.1.: **Nouveaux domaines Pfam (Pfam 26) identifiés à 5% et 10% de FDR chez les deux espèces *P. falciparum* (*P. f*) et *L. major* (*L. m*).** La table rapporte le nombre de nouveaux domaines identifiés par HHPRED (en mode local, approche non-spécifique aux Phylums) et CODD pour les trois types de certification : domaines Pfam connus ("Pfam"), domaines InterPro non-Pfam ("Interp."), domaines potentiels ("Pot."). "Tous" : résultats obtenus en combinant les trois types. "# dom." :nombre de nouveaux domaines identifiés. "# fam." : nombre de nouvelles familles qui n'avaient pas été détectées auparavant sur les différentes protéines de l'organisme en question. Dans chaque case, les chiffres à gauche et à droite indiquent respectivement les résultats obtenus à 5% et 10% de FDR. Entre parenthèses sont indiqués la proportion de domaines ou de familles déjà connus que cela représente. * Pour les certifications avec les domaines Interpro chez *L. major* ce nombre représente les domaines identifiés à 14% de FDR. Aucune certification n'est obtenue avec des FDR moins de 10% avec ce type de certification.

	# GO connus	# GO dom. connus	# GO nouv. dom.
P. falciparum	15661	3127	1152 (5869)
L. major	11958	6200	1817 (7205)

TABLE 6.2.: **Nouvelles annotations GO obtenues à 5% de FDR.** "# GO connus" est le nombre d'annotations GO connus à partir de la base EuPathDB ; "# GO dom. connus" est le nombre d'annotations GO déduites à partir des domaines connus ; "#GO nouv. dom." est le nombre d'annotations GO déduites à partir des nouveaux domaines. Les nombres entre parenthèses rapportent le nombre d'annotations GO qui confirment une annotation connue ou qui sont apportées par les domaines connues.

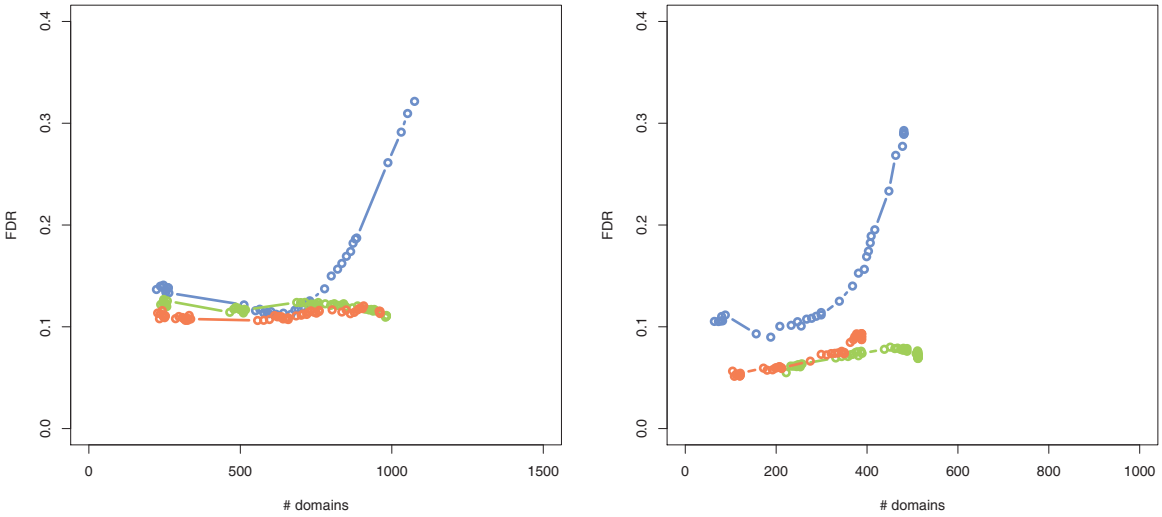
étendu cette politique aux combinaisons de deux domaines. Nous avons pour cela énuméré toutes les paires de domaines Pfam présentes sur les protéines Swiss-Prot et identifié pour chaque combinaison les termes GO partagés par au moins 95% des protéines où la paire de domaines est présente. Seules les paires observées sur au moins 5 protéines annotées sont prises en considération. Nous avons ainsi pu associer des annotations GO à 3460 paires de domaines Pfam.

Nous avons ensuite déterminé les annotations GO qui peuvent être déduites à partir des nouveaux domaines prédits en se basant sur ces associations. La table 6.2 résume pour chaque espèce le nombre d'annotations connues, celles qui peuvent être déduites à partir des domaines connus en utilisant les nouvelles associations, ainsi que celles apportées par les nouveaux domaines (5% de FDR), seuls et en combinaison avec les autres (connus ou nouveaux). Parmi ces dernières, nous distinguons les annotations qui sont déjà connues ou qui peuvent être déduites à partir des domaines connus, de celles apportées par les nouveaux domaines et qui sont complètement nouvelles. Au total les nouveaux domaines apportent respectivement 7021 et 9022 annotations à *P. falciparum* et *L. major*. Parmi ces dernières, 5869 (83%) et 7205 (80%) confirment des annotations connues ou qui peuvent être déduites des domaines connus. 1152 (16%) et 1817 (20%) sont complètement nouvelles. Toutes les nouvelles prédictions ainsi que les annotations GO qui en découlent ont été intégrées à la base de données EuPathDomains (Ghouila et al., 2010).

6.6. Analyse des nouveaux domaines prédits

Une investigation des prédictions apportées par la combinaison de HHPRED et de CODD peut être un point de départ intéressant pour le développement de nouvelles hypothèses concernant la biologie complexe de ces parasites. Nous avons donc procédé à une examination manuelle des nouvelles familles de domaines prédites chez les deux espèces. Avec un $FDR \leq 10\%$, 1328 et 2196 domaines ont été identifiés respectivement sur *P. falciparum* et *L. major* (*cf.* table 6.1) en combinant HHPRED (en mode local) et CODD. Dans cette analyse, nous nous sommes plutôt intéressés aux domaines identifiés en mode global et dont aucune occurrence n'était connue sur les protéines de l'organisme cible puisque ces derniers sont les plus susceptibles d'apporter de nouvelles informations. Afin de réduire cet ensemble des domaines à interpréter manuellement, nous nous sommes focalisés sur les domaines identifiés en utilisant

P. falciparum



L. major

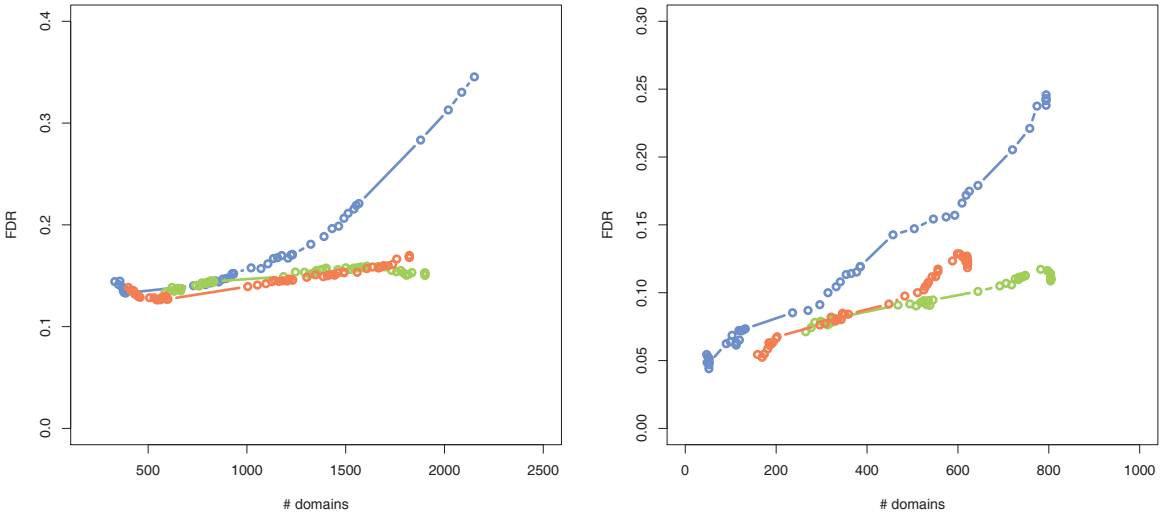
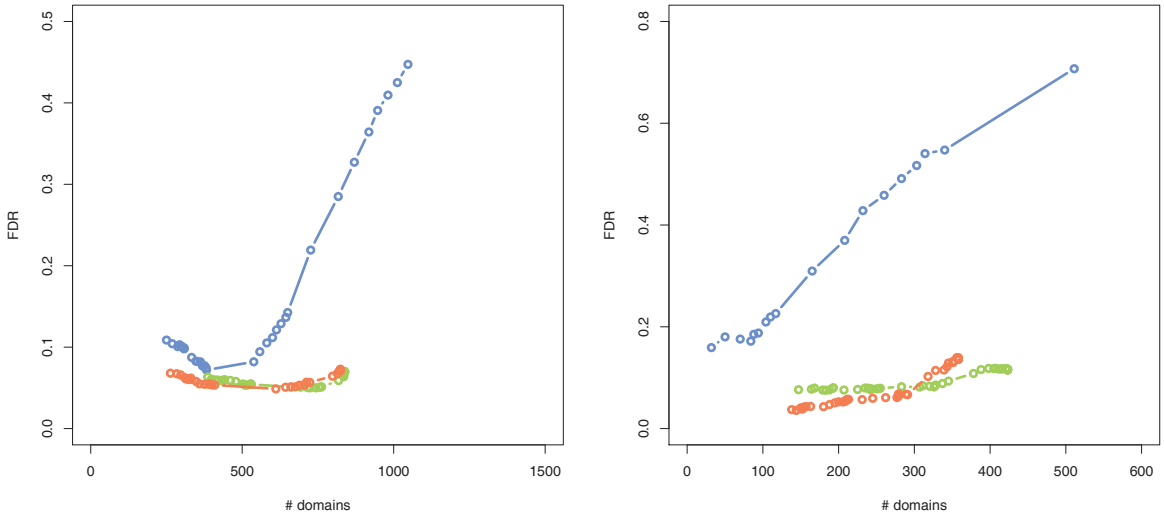


FIGURE 6.3.: **Comparaison de la sensibilité et précision de HHPRED + CODD et HMMER + CODD en utilisant les domaines InterPro connus pour la certification.** La figure rapporte le nombre de nouveaux domaines (axe des X) certifiés par HHPRED + CODD (en orange et vert respectivement pour l'approche spécifique au phylum et l'approche non spécifique au phylum) et HMMER + CODD (bleue) en mode local (gauche) et global (droite) pour différentes valeurs seuils de FDR (axe des Y).

P. falciparum



L. major

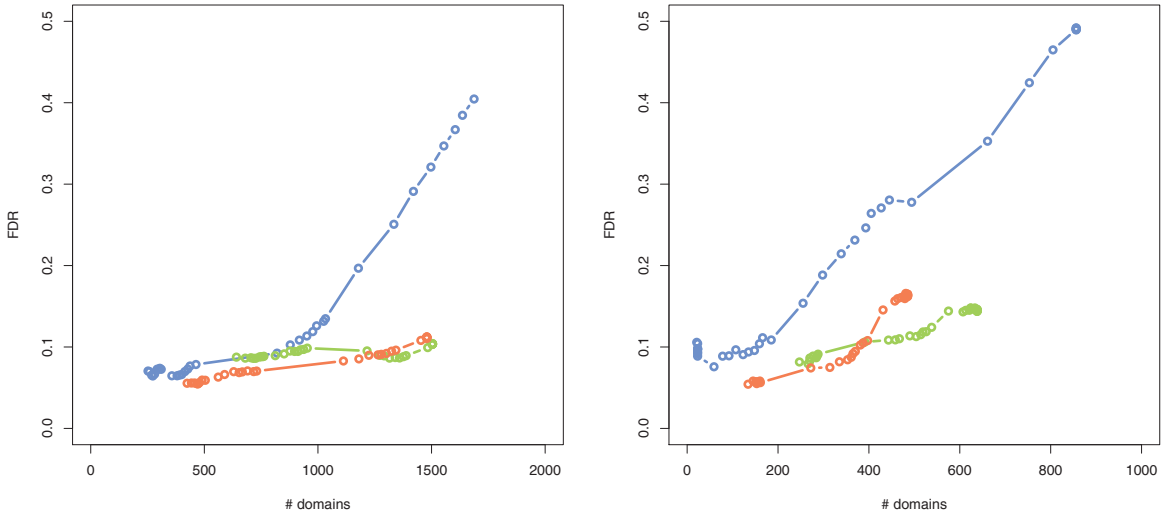


FIGURE 6.4.: **Comparaison de la sensibilité et précision de HHPRED + CODD et HMMER + CODD en utilisant les domaines potentiels pour la certification.** La figure rapporte le nombre de nouveaux domaines (axe des X) certifiés par HHPRED + CODD (en orange et vert respectivement pour l'approche spécifique au phylum et l'approche non spécifique au phylum) et HMMER + CODD (bleue) en mode local (gauche) et global (droite) pour différentes valeurs seuils de FDR (axe des Y)).

l'approche spécifique aux Phylums à 10% de FDR et qui n'ont pas été détectés par l'approche HMMER+CODD. Deux ensembles de 36 et 37 domaines respectivement pour *P. falciparum* et *L. major* ont été ainsi formés. Dans cette analyse, nous avons pris en considération les fonctions connues de la protéine sur laquelle le domaine en question est identifié. Nous avons également utilisé l'information sur la position du nouveau domaine sur la séquence protéique. La description du domaine ainsi que les termes qui lui sont associés sont aussi considérés. Nous avons ensuite essayé d'étudier la relation fonctionnelle entre le domaine et les fonctions connues de la protéine. Pour quelques unes des prédictions, nous avons pu trouvé un support dans la littérature. Nous avons aussi essayé de considérer les autres espèces sur lesquelles le domaine est connu afin de chercher les points communs et des liens qui pourraient expliquer l'association du domaine à la protéine. Dans le cas des protéines hypothétiques, ces prédictions de domaines peuvent suggérer l'attribution de nouvelles fonctions. Quand un domaine prédit est spécifique à un stade particulier du développement, nous avons étudié les profils transcriptomiques de la protéine. Nous présentons ici quelques exemples de domaines prédits chez les deux espèces.

6.6.1. Domaines découverts chez *P. falciparum*

À première vue, la majorité ($\simeq$ 90%) des 36 domaines considérés sont en accord avec les fonctions connues des protéines sur lesquelles ils sont prédits ou compatibles avec les domaines ou annotations GO connues des protéines. Plusieurs nouveaux domaines permettent de raffiner les annotations connus de la protéine sur laquelle ils sont prédits. Nous avons par exemple prédit le domaine WH2, connu pour son interaction avec les actines, de la position 2863 à 2880 sur la protéine *PFL0925w* appelée "formin 2, putative" dans PlasmoDB. Cette protéine possède déjà des termes GO reliés à l'actine tels que : GO :0003779 "actin bindin" et GO :0030036 "actin cytoskeleton organization". D'autres domaines proposent de nouvelles fonctions pour des protéines qui n'ont ni domaines Pfam connus ni annotations GO associées ou pour lesquelles les domaines qui y sont connus sont très peu informatifs. Nous avons noté que plusieurs de ces domaines (tels que "DENN" et "TFIID_90k") sont associés au domaine WD40. Les domaines WD40 sont connus pour être impliqués dans une variété de fonctions telles que la transduction de signaux, la maturation de l'ARN, les dynamiques des chromatines *etc.* (Stirnimann et al., 2010). Les domaines "WD40 repeat" sont parmi les domains les plus fréquents chez les protéines eukaryotes. Ainsi, l'identification d'un second domaine avec le WD40 nous permet de préciser les fonctions putatives associées à cette variété de protéines WD40. Une des découvertes les plus intéressantes concerne la prédiction du domaine "DHQ_synthase" sur la protéine *PFB0280w* de la position 1 à 171. DHQ signifie 3-dehydroquinate et l'enzyme codant pour cette "3-dehydroquinate synthase activity" impliquée dans les premières étapes de la voie métabolique shikimate n'est pas encore identifiée chez *Plasmodium* (McConkey et al., 2004). La voie métabolique shikimate permet la synthèse des acides aminés aromatiques. Elle est présente chez les plantes et les micro-organismes. Il a été démontré que cette voie est active chez *Plasmodium* et elle est considérée comme une potentielle cible thérapeutique intéressante car elle est absente des cellules des mammifères (McConkey et al., 2004; Richards et al., 2006). Cependant les protéines codant pour les quatre premières étapes de cette voie formée par sept étapes (DHQ_synthase correspond à la seconde étape) ne sont pas encore bien connues chez *Plasmodium* (McConkey et al., 2004). La protéine *PFB0280w* est annotée "conserved Plasmodium protein, unknown function" dans PlasmoDB mais une fouille dans la littérature a révélé que cette protéine est considérée comme une enzyme bi-fonctionnelle codant pour la cinquième et sixième étapes de la voie shikimate. D'autre part, le domaine "DHQ_synthase" est trouvé sur deux types distincts d'enzymes : il est trouvé sur diverses protéines "3-dehydroquinate syn-

thases" bactériales ou comme une partie d'un enzyme penta-fonctionnelle comme la protéine *arom* chez les *fungi*. Il est à noter que ce domaine a été également trouvé chez *T. gondii* sur une enzyme penta-fonctionnelle homologue à la protéine *arom* des *fungi*. La protéine *PFB0280w* contient déjà le domaine "EPSP" aux positions 287-1386 (voir PlasmoDB) et le domaine "DHQ_synthase" découvert a été certifié par le domaine "SKI" (position 1322-1600) qui est aussi impliqué dans la voie shikimate. L'organisation des domaines ainsi proposée pour la protéine *PFB0280w* se rapproche de la structure penta-fonctionnelle de la protéine arôme des *fungi* et de *T. gondii*. Cette découverte est alors très intéressante que ce soit pour la conception de médicaments anti-malaria ou pour l'étude de l'évolution des eukaryotes (McConkey et al., 2004).

6.6.2. Domaines découverts chez *L. major*

Comme pour *P. falciparum* plusieurs domains prédits chez *L. major* sont complémentaires aux connaissances déjà connues des protéines sur lesquelles ils sont prédits et confirment ainsi les annotations qui leurs sont associées. Nous avons par exemple, prédit le domaine "Ku-c" dans la partie C-terminal de la protéine *LmjF29.1050* qui possède déjà les domaines "Ku-N" et "Ku70/Ku80". Plusieurs autres domaines proposent de nouvelles fonctions à des protéines non encore annotées ou ne possédant que des fonctions très générales. Comme pour *P. falciparum*, un grand nombre de domaines de cette catégorie ont été prédits en association avec des domaines de la famille "WD40". Parmi les prédictions les plus intéressantes, nous pouvons citer par exemple le domaine "Rhomboid" prédit de la position 21 à 223 sur la protéine *LmjF24.1580*. Les domaines "Rhomboid" appartiennent à la grande famille des "intra-membrane serine proteases". Leur conservation chez la majorité des espèces de toutes les branches de l'arbre de vie suggère leur implication dans des événements biologiques clés tels que : le déclenchement et la signalisation d'évènements chez la *Drosophile*, l'association avec des pathogénies chez les parasites protozoaires et la prolifération parasitaires chez *T. gondii* (Santos et al., 2012). Aucun domaine Pfam "Rhomboid" n'a été décrit auparavant chez *L. major*. Des occurrences partielles de "Rhomboid" ont été par contre décrites dans d'autres espèces *Leishmania*. La présence d'une séquence "mitochondrial like" dans la partie N-terminale de la protéine *LmjF24.1580* suggère une fonction mitochondriale putative à cette dernière (Besteiro et al., 2007). En dépit de leur conservation, les "Rhomboid" semblent jouer une grande variété de fonctions sur les différents organismes, il est donc peu probable qu'une fonction unique soit conservée sur les différentes espèces. Plusieurs exemples de fonctions mitochondriales associées à la voie de signalisation du stress oxydatif ont été décrits chez la levure ou la maladie de Parkinson chez l'humain. Chez les pathogènes extra-cellulaires, les "Rhomboid" conservés sont associés à l'invasion des cellules hôtes ou à l'évasion immunitaire chez l'amibe *Entamoeba histolytica* (Baxt et al., 2008).

Partie IV.

Conclusion

Conclusion

Dans le premier chapitre, nous avons introduit les notions biologiques nécessaires à la compréhension de notre problématique. Les protéines, codées par le génome (ADN), sont des éléments essentiels qui remplissent des fonctions très diverses au sein des cellules d'un organisme. 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 l'étude des caractéristiques de ces unités fonctionnelles et ont révélé certaines propriétés sur leur combinatoire. Ceci est notamment relié au fait que la plus part des familles de domaines ont tendance à apparaître préférentiellement avec certaines autres familles au sein des différentes protéines du Vivant. A l'inverse, certains domaines ont par contre très peu de chance de se retrouver ensemble sur la même protéine. Dans le deuxième chapitre, nous avons présenté les HMM profils qui demeurent les modèles les plus utilisés pour la modélisation des familles de domaines protéiques. La force de ces modèles réside dans le fait qu'il sont capables de modéliser l'information spécifique à chaque position. De plus le cadre probabiliste dans lequel ils s'inscrivent permet de gérer l'incertitude liée au manque de données d'apprentissage. Nous avons également présenté les algorithmes classiques pour les manipuler (plus particulièrement HMMER utilisé dans Pfam). La seconde partie de ce chapitre discute la difficulté de détection des domaines protéiques en utilisant ces modèles et expose la problématique principale de cette thèse : la détection de domaines chez des organismes phylogénétiquement très éloignés des organismes modèles classiques. Le troisième chapitre est consacré à la présentation des deux organismes auxquels nous nous intéressons plus particulièrement : *Leishmania* et *Plasmodium*, causant respectivement la leishmaniose et la malaria. Tous les deux présentent des enjeux majeurs en terme de santé mondiale et causent un grand nombre de décès notamment en Afrique. Sur ces organismes, les méthodes bioinformatiques standards manquent de sensibilité, et les HMM ne font pas exception. Les chapitres 4, 5 et 6 exposent les travaux effectués dans cette thèse pour améliorer les méthodes de détection de domaines et apporter de nouvelles annotations aux protéines de ces organismes.

Un des premiers travaux de cette thèse (chapitre 4) a consisté à adapter et étendre la procédure de certification par co-occurrence CODD à un grand ensemble de pathogènes eukaryotes ayant l'homme comme hôte principal. Ceci nous a permis de prédire un grand nombre de domaines protéiques chez ces différentes espèces et d'améliorer le taux de couverture en domaines Pfam de leurs protéines. Cette amélioration varie entre 17% (chez *P. yoeli*) et 41% (chez *L. major*). Il est à noter que certaines protéines de ces génomes n'avaient aucun domaine Pfam connu. Ces nouvelles prédictions ont permis d'enrichir le nombre de familles de domaines connus chez les différentes espèces étudiées de 25% en moyenne. Nous avons également proposé de nouvelles associations entre domaines protéiques et annotations de la Gene Ontology qui ont permis de transférer de nouvelles annotations à un grand nombre de protéines dont certaines n'avaient aucune annotation. Toutes ces prédictions sont regroupées dans une base de données appelée EuPathDomains (http ://www.atgc-montpellier.fr/EuPathDomains/). Ces travaux ont fait l'objet d'une publication dans le journal Infection, Genetics and Evolution.

Nous avons présenté dans le chapitre 5 diverses méthodes pour améliorer la procédure d'identification des domaines qui composent une protéine. Nous commençons ce chapitre par la présentation d'une stratégie basée sur l'utilisation de graphes pour résoudre les conflits dans la

première phase de traitement des chevauchements. Nous décrivons ensuite une approche pour intégrer à la méthode CODD les informations d'incompatibilité de domaines (les domaines qui ne figurent jamais ou rarement ensemble sur la même protéine). Cette approche nous a permis de réduire de manière significative le nombre de faux positifs parmi les nouveaux domaines certifiés. Ce travail a été présenté au workshop BIOKDD'12 qui s'est tenu à Vienne en Septembre 2012, et a été publié dans les actes de la conférence. Dans la deuxième partie de ce chapitre, nous avons présenté une nouvelle procédure basée sur l'extraction de règles d'association qui permet de découvrir des relations/corrélations intéressantes entre domaines en prenant en considération le sens de l'implication. Les règles extraites sont utilisées pour remplacer la CDP dans le processus de certification. Des expérimentations préalables nous ont permis d'établir que les algorithmes classiques d'extraction de règles d'associations ne sont pas adaptés aux domaines protéiques car ils ne permettent d'extraire que les associations entre domaines fréquents. Nous nous sommes alors concentré sur les règles simples qui peuvent être extraites des paires de domaines contenues dans la CDP, et nous avons pu vérifier que cette approche permet d'abaisser de manière très significative le taux d'erreur associé aux prédiction, par rapport à la procédure CODD originale.

Finalement, un dernier volet de cette thèse, présenté dans le sixième et dernier chapitre, s'intéresse au bénéfice que l'on peut tirer des séquences protéiques homologues dans la recherche de domaines. Dans cette approche, les protéines cible sont sont elles mêmes représentées par un HMM profil construit à partir de l'alignement de protéines homologues. La recherche de domaines ne se fait alors plus sur la base de comparaisons séquence-HMM mais grâce à des comparaisons HMM-HMM. L'intérêt de cette approche est que l'on peut de cette manière modéliser également les positions conservées sur la protéine. Bien que ce type de méthodes soit souvent utilisé pour la reconnaissance d'homologues lointains dans le cadre de prédiction de structures 3D, il n'a jamais utilisé pour l'annotation fonctionnelle à l'échelle d'un génome entier. Les premiers tests effectués nous ont permis de montrer que cette stratégie améliore effectivement la sensibilité mais affecte également la précision des résultats, et qu'un post-traitement est nécessaire pour que l'approche soit intéressante. Nous avons alors proposé une nouvelle méthode qui consiste à coupler cette approche avec les informations de co-occurences de domaines. Appliquée à *Leishmania* et *Plasmodium*, cette nouvelle approche nous a permis de prédire un grand nombre de domaines supplémentaires pour les protéines des deux espèces, permettant ainsi d'améliorer la couverture des protéines en domaines et d'enrichir leurs annotations GO. Une analyse manuelle des prédictions relatives aux familles de domaines dont aucune occurrence n'était connue chez ces espèces nous a permis d'enrichir le catalogue de domaines et de proposer de nouvelles hypothèse pour certains mécanismes mal compris chez ces espèces.

Perspectives

Les différentes méthodes proposées dans le cadre de cette thèse ont apporté des améliorations notables dans la prédiction des domaines protéiques sur les protéines divergentes. L'analyse biologique des résultats obtenus sur les deux organismes d'intérêt a montré que les prédictions obtenues sont pertinentes et peuvent aider dans la compréhension des mécanismes biologiques complexes de ces organismes. Toutefois, de nombreuses améliorations peuvent être envisagées pour améliorer encore la détection des domaines divergents :

– Prise en considération de l'ordre d'apparition des domaines sur les protéines pour affiner la procédure de certification. Il a été en effet démontré que 90% des domaines apparaissent toujours dans le même ordre : l'orientation N-C terminale des combinaisons de domaines est très conservée (Kummerfeld and Teichmann, 2009). La prise en compte de l'ordre

d'apparition pourrait donc contribuer à la minimisation du nombre de faux positifs parmi les domaines potentiels certifiés.

- Amélioration de la procédure d'extraction des règles d'association proposée dans le chapitre 5. Les perspectives autour des règles d'associations sont nombreuses. Nous souhaitons par exemple extraire des associations mettant en relation plus de deux domaines. Des règles faisant intervenir deux domaines ou plus dans la partie prémisse permettrait de modéliser des corrélations plus complexes ou la présence d'un domaine pourrait être attestée par la présence conjointe de plusieurs autres domaines. De même, un autre type de règles potentiellement très intéressant concerne des règles incluant une négation dans le prémisse. Ce type de règles pourrait par exemple exprimer le fait qu'un domaine A est très probable si le domaine B est présent mais que le domaine C est absent.
- Utilisation de règles de classification pour étudier les correspondances entre domaines protéiques et annotations de la Gene Ontology. Ces règles permettraient d'associer à un domaine ou une combinaison de domaines des termes GO , avec là encore la possibilité de faire entrer les négations dans les règles, ce que ne permet pas l'approche que nous avons utilisée dans cette thèse.
- Utilisation d'autres types d'information dans la procédure CODD. Pour l'instant, la procédure n'est basée que sur la co-occurrence de domaines. Or ce principe pourrait être étendu à d'autres types d'informations dont on dispose sur les séquences, comme la présence de séquences non-structurées, de répétitions en tandem, de séquences de faible complexité, de peptide signaux, *etc.*.

Table des figures

0.1. Évolution du nombre des projets génomes (en cours de séquençage ou complètement séquencées) contenus dans la base de données Genomes On Line (GOLD) (Pagani et al., 2012). . . . 4

1.1. Le dogme central de la biologie moléculaire : le processus de la synthèse des protéines à partir de l'ADN introduit par Francis Crick en 1958 . . . 10
1.2. Le code génétique "universel" qui établit les correspondances entre les codons de l'ARN messager et les acides aminés qui forment les protéines pour la grande majorité des espèces vivantes. . . . 11
1.3. Les différents niveaux de structure d'une protéine : de la structure primaire à la structure quaternaire . . . 14

2.1. Un exemple de HMM . . . 29
2.2. Structure d'un HMM profil. L'état *Begin* (B), d'où commence tout(e) chemin (séquence). Ensuite, on suit la succession linéaire des états Matchs (M). Chaque état Match est accompagné des états Insert (I) et Délétion (D). Enfin, on finit toujours sur l'état *End* (E) . . . 33
2.3. Architecture *Plan*7 utilisée pour modéliser les HMM profils dans HMMER . . . 35
2.4. La structure du modèle nul utilisée par HMMER . . . 37
2.5. Évolution du nombre de familles Pfam-A dans les différentes versions Pfam. . . . 39

3.1. Répartition géographique des leishmanioses cutanées et muco-cutanées à travers l'ancien monde (gauche) et dans le nouveau monde (droite) . . . 43
3.2. Répartition géographique de leishmaniose viscérale et muco-cutannée dans le monde . . . 44
3.3. Classification de quelques espèces du parasite *Leishmania* selon leur appartenance aux 3 sous-genres connus . . . 46
3.4. Cycle de vie du parasite Leishmania . . . 46
3.5. Répartition géographique du paludisme dans le monde . . . 49
3.6. Classification des différentes espèces de *Plasmodium* selon les espèces qu'elles infectent. . . . 51
3.7. Cycle de vie du parasite *Plasmodium falciparum* . . . 52
3.8. Nombre de génomes séquencés et couverture en domaines dans l'arbre des Eukaryotes. . . . 53

4.1. Les différentes étapes de prédictions de domaines et annotations GO par la méthode CODD . . . 61
4.2. Évolution du nombre de certification et de l'estimation du nombre de domaines certifiés sous H_0 sur les protéines de *P. falciparum* en fonction de la E-valeur (à gauche) . . . 64
4.3. Domaines connus et nouveaux domaines prédits sur la protéine *LmjF36.4590* de *L. major*. . . . 67

4.4. Comparaison de la fréquence des domaines connus (bleue) et des nouveaux domaines certifiés (rouge) chez *L. major* en fonctions de leurs longueurs 69
4.5. Annotations GO des nouvelles familles de domaines prédites chez *L. major*, *L. infantum* et *L. braziliensis*. 72
4.6. Comparaison du nombre de certification et de l'estimation du nombre de domaines certifiés sous H_0 sur les protéines de *P. falciparum* en fonction de la E-valeur (à gauche) entre HMMER 2 + Pfam 23 (en pointillé) et HMMER 3 + Pfam 26 (ligne pleine). 74

5.1. Exemple de modélisation des occurrences de domaine sur une protéine par un graphe acyclique orienté et pondéré 79
5.2. Comparaison des certifications obtenues par la sélection du meilleur domaine à la recherche du chemin de poids maximal sur les protéines de *L. major* 80
5.3. Exemple de filtration de domaines par X=2 81
5.4. Comparaison des certifications obtenues par la sélection du meilleur domaine v.s. la recherche du chemin de poids maximal avec différents seuils de filtration par X=2, X=3, X=4 et X=10 82
5.5. Les différentes étapes de la procédure de certification *CODD_exclusive*. . . . 83
5.6. Nombre de nouveaux domaines certifiés avec les différents seuils de FDR par les domaines Pfam connus chez *P.falciparum* (droite) et *L. major* (gauche) avec (points rouges) et sans (points noirs) utilisation des exclusions de domaines . . 85
5.7. Nombre de nouveaux domaines certifiés avec les différents seuils de FDR par les domaines potentiels chez *P.falciparum* (droite) et *L. major* (gauche) avec (points rouges) et sans (points noirs) utilisation des exclusions de domaines . . 87
5.8. Nombre de nouveaux domaines certifiés avec les domaines Pfam connus en utilisant les trois ensembles de règles de certification 91
5.9. Comparaison des certifications obtenues avec la recherche sur les espèces proches *Leishmania* avec celles obtenues sur avec *L.major* seul 92

6.1. Comparaison de la sensibilité et de la précision de HHPRED et de HMMER sur les génomes de *P. falciparum* et *L. major*. 98
6.2. Comparaison de la sensibilité et de la précision de HHPRED + CODD et HMMER+CODD en utilisant les domaines Pfam connus pour la certification. . . . 101
6.3. Comparaison de la sensibilité et précision de HHPRED + CODD et HMMER + CODD en utilisant les domaines InterPro connus pour la certification. 104
6.4. Comparaison de la sensibilité et précision de HHPRED + CODD et HMMER + CODD en utilisant les domaines potentiels pour la certification. 105

Liste des tableaux

1.1. Les 20 acides aminés du vivant : noms, trigrammes et codes standards 13

3.1. Caractéristiques des différents génomes de *Leishmania* séquencés. 47

4.1. Nombre de familles de domaines Pfam et couverture des protéines en domaines chez plusieurs Eukaryotes 59
4.2. Prédictions sous un FDR à 10% chez les différents organismes. 63
4.3. Prédictions sous un FDR de 20% chez les organismes étudiés. 66
4.4. Nouvelles annotations GO sur les protéines des différentes espèces. 68
4.5. Comparaison des prédictions effectuées dans deux espèces : *L. major* et *L. infantum* 70
4.6. Termes GO sur-représentés parmi les annotations apportées par les domaines prédits en comparaison des annotations associées aux domaines connus. 71
4.7. Utilisation de HMMER 3 et Pfam 26.0 : Prédictions sous un FDR de 10% chez les organismes étudiés. 75

6.1. Nouveaux domaines Pfam (Pfam 26) identifiés à 5% et 10% de FDR chez les deux espèces *P. falciparum* (*P. f*) et *L. major* (*L. m*). 102
6.2. Nouvelles annotations GO obtenues à 5% de FDR 103

Bibliographie

Abrahamsen, M. and Templeton, T.J., e. a. (2004). Complete genome sequence of the apicomplexan, cryptosporidium parvum. *Science*, 304(5669) :441–445.

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

Agrawal, R., Imielinski, T., and Swami, A. (1993). Mining association rules between sets of items in large databases. In *Proceedings of the ACM SIGMOD Intl. Conference on Management of Data, Washington, USA*, pages 207–216.

Agrawal, R. and Skirant, R. (1994). Fast algorithms for mining association rules. In *Proceedings of the 20th Intl. Conference on Very Large Databases, Santiago, Chile*, pages 478–499.

Albrecht, M. and Lengauer, T. (2004). Survey on the pabc recognition motif pam2. *Biochem. Biophys. Res. Commun.*, 316(1) :129–138.

Alvar, J., Aparicio, P., Aseffa, A., Den Boer, M., Cañavate, C., Dedet, J. P., Gradoni, L., Horst, R. T., López-Vélez, R., and Moreno, J. (2008). The relationship between leishmaniasis and aids : the second 10 years. *Clin. Microbiol. Rev.*, 21(2) :334–359.

Alvar, J., Vélez, I., Bern, C., Herrero, M., Desjeux, P., Cano, J., Jannin, J., Boer, M., and the WHO Leishmaniasis Control Team (2012). Leishmaniasis worldwide and global estimates of its incidence. *PLoS ONE*, 7(5) :e35671.

Alvar, J., Yactayo, S., and Bern, C. (2006). Leishmaniasis and poverty. *Trends in Parasitology*, 22(12) :552–557.

Andreeva, A., Howorth, D., Brenner, S., Hubbard, T., Chothia, C., and Murzin, A. (2004). Scop database in 2004 : refinements integrate structure and sequence family data. *Nucleic Acids Research*, 32 :226–229.

Apic, G. (2001). Domain combinations in archeal, eubacterial and eukaryotic proteomes. *Journal of Molecular Biology*, 310 :311–325.

Attwood, T. K., Coletta, A., Muirhead, G., Pavlopoulou, A., Philippou, P. B., Popov, I., Romá-Mateo, C., Theodosiou, A., and Mitchell, A. L. (2012). The prints database : a fine-grained protein sequence annotation and analysis resource–its status in 2012. *Database (Oxford)*, 15.

Aurrecoechea, C. e. a. (2010). Eupathdb : a portal to eukaryotic pathogen databases. *Nucleic Acids Research, database issue*, 38 :D415–D419.

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

Baldauf, S. (2003). The deep roots of eukaryotes. *Science*, 300(5626) :1703–1706.

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

Basu, M. K., Carmel, L., Rogozin, I., and Koonin, E. V. (2008). Evolution of protein domain promiscuity in eukaryotes. *Genome Res.*, 18 :449–461.

Bateman, A., Birney, E., Cerruti, L., Durbin, R., Etwiller, L., Eddy, S. R., Griffiths-Jones, S., Howe, K. L., Marshall, M., and Sonnhammer, E. L.

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

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

Bateman, A., Coin, L., Durbin, R., Finn, R. D., Hollich, V., Griffiths-Jones, S., Khanna, A., Marshall, M., Moxon, S., Sonnhammer, E. L. L., and al.

Bates, P. A. (2007). Transmission of leishmania metacyclic promastigotes by phlebotomine sand flies. *International Journal for Parasitology*, 37(10) :1097–1106.

Baum, L., Petrie, T., Soules, G., and 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.

Baxt, L. A., Baker, R. P., Singh, U., and Urban, S. (2008). An entamoeba histolytica rhomboid protease with atypical specificity cleaves a surface lectin involved in phagocytosis and immune evasion. *GENES and DEVELOPMENT*, 22 :1636–1646.

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

Beissbarth, T. and Speed, T. P. (2004). Gostat : Find statistically overrepresented gene ontologies within a group of genes. *Bioinformatics*, 20(9) :1464–1465.

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

Berman, H. M., Henrick, K., and Nakamura, H. (2003). Announcing the worldwide protein data bank. *Nat Struct Biol*, 10 :980.

Bernstein, K. e. a. (2004). The small-subunit processome is a ribosome assembly intermediate. *Eukaryot. Cell*, 3 :1619–1626.

Besteiro, S., Williams, R. A., Coombs, G. H., and Mottram, J. C. (2007). Protein turnover and differentiation in leishmania. *International Journal for Parasitology*, 37 :1063–1075.

Betel, D., Isserlin, R., and Hogue, C. W. (2004). Analysis of domain correlations in yeast protein complexes. *Bioinformatics*, 4(20) :i55–62.

Bischoff, E. and Vaquero, C. (2010). n silico and biological survey of transcriptionassociated proteins implicated in the transcriptional machinery during the erythrocytic development of plasmodium falciparum. *BMC Genomics*, 11(34).

Blader, J. and Saeij, J. (2009). Communication between toxoplasma gondii and its host : impact on parasite growth, development, immune evasion, and virulence. *APMIS*, 117 :458–476.

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

Bornberg-Bauer, E., Beausart, F., Kummerfeld, S. K., Teichmann, S. A., and Weiner, J. (2005). The evolution of domain arrangements in proteins and interaction networks. *Cell. Mol. Life Sci.*, 62 :435–445.

Bousema, T. and Drakeley, C. (2011). Epidemiology and infectivity of plasmodium falciparum and plasmodium vivax gametocytes in relation to malaria control and elimination. *Clin Microbiol Rev.*, 24(2) :377–410.

BRU, C., COURCELLE, E., CARRÈRE, S., BEAUSSE, Y., DALMAR, S., and KAHN, D. (2005). The prodom database of protein domain families : more emphasis on 3d. *Nucleic Acids Research*, 33(Database issue) :D212–D215.

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

Bréhélin, L., Dufayard, J., and Gascuel, O. (2008). Plasmodraft : a database of plasmodium falciparum gene function predictions based on postgenomic data. *BMC Bioinformatics*, 16(9) :440.

Bréhélin, L., Florent, I., Gascuel, O., and Maréchal, E. (2010). Assessing functional annotation transfers with inter-species conserved coexpression : application to plasmodium falciparum. *BMC Genomics*, 11(35).

Buchan, D. W., Shepherd, A. J., Lee, D., Pearl, F. M., Rison, S. C., Thornton, J. M., and Orengo, C. A. (2002). Gene3d : structural assignment for whole genes and genomes using the cath domain structure database. *Genome Res.*, 12(3) :503–514.

Castro, E., Sigrist, C. J., Gattiker, A., Bulliard, A., Langendijk-Genevaux, P. S., Gasteiger, E., Bairoch, A., and Hulo, N. (2006). Scanprosite : detection of prosite signature matches and prorule-associated functional and structural residues in proteins. *Nucleic Acids Res.*, 34 :W362–W365.

Chen, F., Mackey, A., Stoeckert, C. J., and Roos, D. S. (2006). Orthomcl-db : querying a comprehensive multi-species collection of ortholog groups. *Nucleic Acids Research*, 34 :D363–D368.

Cohen, G., Nussinov, R., and Sharan, R. (2007). Comprehensive analysis of co-occurring domain sets in yeast proteins. *BMC Genomics*, 8.

Consortium, G. O. (2001). Creating the gene ontology resource : Design and implementation. *Genome Res.*, 11 :1425–1433.

Consortium, G. O. (2004). The gene ontology (go) database and informatics resource. *Nucleic Acids Research*, 32(Database issue) :D258–D261.

Convit, J., Pinardi, M. E., and Rondón, A. J. (1972). Diffuse cutaneous leishmaniasis : A disease due to an immunological defect of the host. *Transactions of the Royal Society of Tropical Medicine and Hygiene*, 66(4) :603–610.

Copley, R. R., Ponting, C. P., Schultz, J., and Bork, P. (2002). Sequence analysis of multidomain proteins : past perspectives and future directions. *Adv Protein Chem*, 61 :75–98.

Cormen, T. H., Leiserson, C. E., Rivest, R. L., and Stein, C. (2001). *ntroduction to Algorithms, Second Edition*.

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

Cox, F. E. G. (2002). History of human parasitology. *Clin. Microbiol. Rev.*, 15(4) :595–612.

Crick, F. (1970). Central dogma of molecular biology. *Nature*, 227 :561–563.

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

Croft, S. L., Sundar, S., and Fairlamb, A. H. (2006). Drug resistance in leishmaniasis. *Clin. Microbiol. Rev.*, 19(1) :111–126.

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

Cummings, B. S., McHowat, J., and Schnellmann, R. G. (2010). Phospholipase a2s in cell injury and death. *J. Pharmacol. Exp. Ther.*, 294 :793–799.

De Lima Morais, D. A., Fang, H., Rackham, O. J. L., Wilson, D., Pethica, R., Chothia, C., and Gough, J. (2011). Superfamily 1.75 including a domain-centric gene ontology method. *Nucleic Acids Research*, 39(Database issue) :D427–D434.

Desjeux, P. (2004). Leishmaniasis : current situation and new perspectives. *Comparative Immunology, Microbiology and Infectious Diseases*, 27(5) :305–318.

Desjeux, P. and Alvar, J. (2003). Leishmania hiv co-infections : epidemiology in europe. *Annals of Tropical Medicine and Parasitology*, 97(1) :S3–S15.

Dlakic, M. (2009). Hhsvm : fast and accurate classification of profile–profile matches identified by hhsearch. *Bioinformatics*, 25 :3071–3076.

Downing, T., Imamura, H., Decuypere, S., Clark, T. G., Coombs, G. H., Cotton, J. A., Hilley, J. D., Doncker, S., Maes, I., Mottram, J. C., Quail, M. A., Rijal, S., Sanders, M., Schönian, G., Stark, O., Sundar, S., Vanaerschot, M., Hertz-Fowler, C., Dujardin, J. C., and Berriman, M. (2012). Whole genome sequencing of multiple leishmania donovani clinical isolates provides insights into population structure and mechanisms of drug resistance. *Genome Res.*, 21(12) :2143–2156.

Dragon, F. e. a. (2002). A large nucleolar u3 ribonucleoprotein required for 18s ribosomal rna biogenesis. *Nature*, 417 :967–970.

Du, Z., Zhao, Y., and Li, N. (2008). Genome-wide analysis reveals regulatory role of g4 dna in gene transcription. *Genome Res.*, 18 :233–241.

Durbin, R. (1998). *Biological sequence analysis : probabilistic models of proteins and nucleic acids*. Cambridge University Press, Cambridge, UK.

Eddy, S. R. (1996). Hidden markov models. *Current Opinion in Structural Biology*, 6(3) :361–365.

Eddy, S. R. (1998). Profile hidden markov models. *Bioinformatics*, 14(9) :755–763.

Eddy, S. R. (2003). Hmmer user's guide version 2.3.2.

Eddy, S. R. (2010). Hmmer user's guide version 3.0.

Edgar, R. (2004). Muscle : multiple sequence alignment with high accuracy and high throughput. *Nucleic Acids Research*, 32(5) :1792–1797.

El-Sayed, N. M. e. a. (2005). Comparative genomics of trypanosomatid parasitic protozoa. *Science*, 09(5733) :404–409.

Essien, K. and Stoeckert, C. (2010). Conservation and divergence of known apicomplexan transcriptional regulons. *BMC Genomics*, 11(147).

Finn, R. (2008). The pfam protein families database. *Nucleic Acids Research*, 36 :D281–D288.

Finn, R. D., Clements, J., and Eddy, S. R. (2011). Hmmer web server : interactive sequence similarity searching. *Nucleic Acids Res.*, 39(Web Server issue) :W29–W37.

Finn, R. D., Marshall, M., and Bateman, A. (2005). ipfam : visualization of protein–protein interactions in pdb at domain and amino acid resolutions. *Bioinformatics*, 21(3) :410–412.

Finn, R. D., Mistry, J., Schuster-Bo, B., Griffiths-Jones, S., Hollich, V., Lassmann, T., Moxon, S., Marshall, M., Khanna, A., Durbin, R., Eddy, S. R., , Sonnhammer, E., and Bateman, A.

Finn, R. D., Mistry, J., Tate, J., Coggill, P. C., Heger, A., Pollington, J. E., Gavin, O. L., Gunasekaran, P., Ceric, G., Forslund, K., Holm, L., Sonnhammer, E. L., Eddy, S. R., and Bateman, A.

Fitch, W. (2002). Homology : a personal view on some of the problems. *Trends in Genetics*, 16(5) :227–231.

Forney, G. (1973). The viterbi algorithm. *Proceedings of the IEEE*, 61(3) :268–278.

Forslund, K. and Sonnhammer, E. (2008). Predicting protein function from domain content. *Bioinformatics*, 24 :1681–1687.

Forslund, K. and Sonnhammer, E. L. (2012). Evolution of protein domain architectures. *Methods Mol Biol.*, 856 :187–216.

Friedberg, I. (2006). Automated protein function prediction—the genomic challenge. *Brief Bioinform*, 7(3) :225–242.

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., and BARRELL, B. (2002). Genome sequence of the human malaria parasite plasmodium falciparum. *Nature*, 419(6906) :498–511.

Gatei, W., Wamae, C., and Mbae, C. e. a. (2006). Cryptosporidiosis : prevalence, genotype analysis, and symptoms associated with infections in children in kenya. *Am. J. Trop. Med. Hyg.*, 75(1) :78–82.

Ghouila, A., Terrapon, N., Gascuel, O., Guerfali, F., Laouini, D., Maréchal, E., and Bréhélin, L. (2010). Eupathdomains : The divergent domain database for eukaryotic pathogens. *Infect Genet Evol*, 11(4) :698–707.

Gougha, J. and Chothia, C. (2002). Superfamily : Hmms representing all proteins of known structure. scop sequence searches, alignments and genome assignments. *Nucleic Acids Research*, 30(1) :268–272.

Gribskov, M. (1994). Profile analysis. *Methods in Molecular Biology*, 25 :247–266.

Gribskov, M., Mclachlan, A., and 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.

Gribskov, M., Mclachlan, A., and Eisenberg, D. (1994). A generalized profile syntax for biomolecular sequence motifs and its function in automatic sequence interpretation. *Proc Int Conf Intell Syst Mol Biol.*, 2 :53–61.

Grobusch, M. P. and Van Vugt, M. (2010). Malaria therapy : where do we stand, what next ? *Future Microbiology*, 5(10) :1447–1449.

Haft, D. H., Selengut, J. D., and White, O. (2003). The tigrfams database of protein families. *Nucleic Acids Res.*, 31 :371–373.

Haldar, K., Kamoun, S., Hiller, N., Bhattacharje, S., and Ooij, C. (2006). Common infection strategies of pathogenic eukaryotes. *Nat. Rev. Microbiol*, 4 :922–931.

Han, J. and Kamber, M. (2001). *Data Mining : Concepts and Techniques*. Morgan Kaufmann Publishers, San Francisco.

Hanakahi, L. A. and Sun, H. (1999). High affinity interactions of nucleolin with g-gpaired rdna. *J. Biol. Chem.*, 274 :15908–15912.

Hay, S., Guerra, C. A., Tatem, A. J., Noor, A. M., and Snow, R. W. (2004). The global distribution and population at risk of malaria : past, present, and future. *Lancet Infect Dis.*, 4(6) :327–336.

Heger, A. and Holm, L. (2003). Exhaustive enumeration of protein domain families. *J. Mol. Biol.*, 328 :749–7671.

Heger, A., Wilton, A., Sivakumar, A., and Holm, L. (2005). Adda : a domain database with global coverage of the protein universe. *Nucleic Acids Res.*, 33 :188–191.

Hegyi, H. and Gerstein, M. (2001). Annotation transfer for genomics : measuring functional divergence in multi-domain proteins. *Genome Res.*, 11 :1632–1640.

Henikoff, J., Greene, E., Pietrokovski, S., and Henikoff, S. (2000). Increased coverage of protein families with the blocks database servers. *Nucleic Acids Research*, 28(1) :228–230.

Herwaldta, M. B. L. (1999). Leishmaniasis. *Lancet*, 354(9185) :1191–1199.

Holt, K. A. e. a. (2002). The genome sequence of the malaria mosquito anopheles gambiae. *Science*, 298(5591) :129–149.

Hughey, R. Krogh, A. Hidden markov models for sequence analysis : extension and analysis of the basic method.

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., Mitchell, A., Mulder, N., Natale, D., Orengo, C., Quinn, A., Selengut, J., Sigrist, C., Thimma, M., Thomas, P., Valentin, F., Wilson, D., Wu, C., and Yeats, C. (2009). Interpro : the integrative protein signature database. *Nucleic Acids Research*, 3.

Hybiske, K. and Stephens, R. S. (2008). Exit strategies of intracellular pathogens. *Nat. Rev. Microbiol.*, 6 :99–110.

Ingolfsson, H. and Yona, G. (2008). *Structural Proteomics*, volume 426, chapter Protein Domain Prediction, pages 117–143. Humana Press.

Ivens, C., Peacock, S., Worthey, E. A., and al. (2005). The genome of the kinetoplastid parasite, leishmania major. *Science*, 309(5733) :436–442.

Julius, P. (2002). *Infection, Resistance and Immunity, second Edition.*

Keeling, P., Burger, G., Durnford, D., Lang, B., Lee, R., Pearlman, R., Roger, A., and Gray, M. (2005). The tree of eukaryotes. *Trends Ecol Evol*, 20(12) :670–676.

Kia-Ki, H. and Martinage, A. (1992). Post-translational chemical modification(s) of proteins. *International Journal of Biochemistry*, 24(1) :19–28.

Kim, W. K., Park, J., and Suh, J. K. (2002). Large scale statistical prediction of protein-protein interaction by potentially interacting domain (pid) pair. *Genome Inform.*, 13 :42–50.

Kissinger, J. C. e. a. (2002). The plasmodium genome database. *Nature*, 419 :490–492.

Kooij, T., Carlton, J., Bidwell, S., Hall, N., and Ramesar, J. e. a. (2005). A plasmodium whole-genome synteny map : indels and synteny breakpoints as foci for species-specific genes. *PLoS Pathog.*, 1(4) :349–361.

Krauss, G. (2001). *GBiochemistry of signal transduction and regulation. In : Other Receptor Classes*, chapter 12, pages 377–384. Wiley-VCH.

Krogh, A., Brown, M., Mian, S., Sjolander, K., and Hausslery, D. (1996). Hidden markov models in computational biology : Applications to protein modeling. *Journal Molecular Biology*, 235 :1501–1531.

Krogh, A., Mian, S., and Hausslery, D. (1994). A hidden markov model that finds genes in e. coli dna. *Nucleic Acids Research*, 22 :4768–4778.

Kummerfeld, S. K. and Teichmann, S. A. (2009). Protein domain organisation : adding order. *BMC Bioinformatics*, 10(39).

Lee, H., Deng, M., Sun, F., and Chen, T. (2006). An integrated approach to the prediction of domain-domain interactions. *BMC Bioinformatics*, 25(7) :269.

Lees, J., Yeats, C., Redfern, O., Clegg, A., and Orengo, C. (2010). Gene3d : merging structure and function for a thousand genomes. *Nucleic Acids Res.*, 38(Database issue) :D296–300.

Letunic, I., Doerks, T., and Bork, P. (2012). Smart 7 : recent updates to the protein domain annotation resource. *Nucleic Acids Res.*, 40(D1) :D302–D305.

Lin, K., Simossis, V. A., Taylor, W. R., and Heringa, J. (2005). A simple and fast secondary structure prediction method using hidden neural networks. *Bioinformatics*, 21(2) :152–159.

Liu, W., Li, L., Learn, G. H., and al. (2010). Origin of the human malaria parasite plasmodium falciparum in gorillas. *Nature*, 467 :420–425.

Maizels, N. (2006). Dynamic roles for g4 dna in the biology of eukaryotic cells. *Struct. Mol. Biol.*, 13 :1055–1059.

Marchler-Bauer, A. e. a. (2005). Cdd : a conserved domain database for protein classification. *Nucleic Acids Res.*, 33 :D192–196.

Mathee, K. and Narasimhan, G. (2003). Detection of dna-binding helix-turn-helix motifs in proteins using the pattern dictionary method. *Methods Enzymol.*, 370 :250–264.

McConkey, G. A., Pinney, J., Westhead, D., Plueckhahn, K., Fitzpatrick, T., Macheroux, P., and Kappes, B. (2004). Annotating the plasmodium genome and the enigma of the shikimate pathway. *Trends in Parasitology*, 20(2) :60–65.

Mi, H., Muruganujan, A., and Thomas, P. D. (2013). Panther in 2013 : modeling the evolution of gene function, and other gene attributes, in the context of phylogenetic trees. *Nucleic Acids Research*, 41(D1) :D377–D386.

Mulder, N. J. and al. (2003). The interpro database, 2003 brings increased coverage and new features. *Nucleic Acids Research*, 31(1) :315–318.

Mulder, N. J. and Apweiler, R. (2002). Tools and resources for identifying protein families, domains and motifs. *Genome Biol.*, 3(1).

Murray, H. W., Berman, J. D., Davies, C. R., and Saravia, N. G. (2005). Advances in leishmaniasis. *Lancet*, 366(9496) :1561–1577.

Murzin, A. G., Brenner, S. E., Hubbard, T., and Chothia, C. (1995). Scop : a structural classification of proteins database for the investigation of sequences and structures. *J Mol Biol.*, 247(4) :536–540.

Myler, P. J. and Fasel, N. (2007). *Leishmania – After the Genome*. Caister Academic Press.

NG, S., ZHANG, Z., and TAN, S. (2003). Integrative approach for computationally inferring protein domain interactions. *Bioinformatics*, 19(8) :923–929.

Nikolskaya, A. N., Arighi, C. N., Huang, H., Barker, W. C., and Wu, C. H. (2007). Pirsf family classification system for protein functional and evolutionary analysis. *Evol Bioinform Online*, 10(2) :197–209.

Nonhoff, e. a. (2007). Ataxin-2 interacts with the dead/h-box rna helicase ddx6 and interferes with p-bodies and stress granules. *Mol. Cell Biol.*, 18 :1385–1396.

Nye, T. M., Berzuini, C., Gilks, W. R., Babu, M. M., and Teichmann, S. A. (2006). Predicting the strongest domain-domain contact in interacting protein pairs. *Stat Appl Genet Mol Biol.*, 24(5).

Ochman, H. and Moran, N. (2001). Genes lost and genes found : evolution of bacterial pathogenesis and symbiosis. *Science*, 952(5519) :1096–1099.

Orengo, C., Michie, A. D., Jones, S., Jones, D. T., Swindells, M. B., and Thornton, J. M. (1997). Cath—a hierarchical classification of protein domain structures. *Structure*, 5 :1093–1108.

Ostergard, P. R. J. (2002). A fast algorithm for the maximum clique problem. *Discrete Appl. Math*, 120 :179–207.

Pagani, I., Liolios, K., Jansson, J., Chen, I. M., Smirnova, T., Nosrat, B., Markowitz, V. M., and Kyrpides1, N. (2012). The genomes online database (gold) v.4 : status of genomic and metagenomic projects and their associated metadata. *Nucleic Acids Res.*, 40(D1) :D571–D579.

Pao, G. and Saier, M. (1995). Response regulators of bacterial signal transduction systems : selective domain shuffling during evolution. *J. Mol. Evol.*, 40(2) :136–154.

Parkinson, G., Lee, M., and Neidle, S. (2002). Crystal structure of parallel quadruplexes from human telomeric dna. *Nature*, 417 :876–880.

Pasquier, N. (2000). *Data Mining : algorithmes d'extraction et de réduction des règles d'association dans les bases de données*. PhD thesis, Université de Clermont-Ferrand II, France.

Peacock, C., Seeger, K., Harris, D., and al. (2007). Comparative genomic analysis of three leishmania species that cause diverse human disease. *Nature Genet.*, 39(7) :839–847.

Pirovano, W. and Heringa, J. (2010). Protein secondary structure prediction. *Methods in Molecular Biology*, 609 :327–348.

Pizzi, E. and Frontali, C. (2001). Low-complexity regions in plasmodium falciparum proteins. *Genome Research*, 11(2) :218–229.

Pratlong, F., Dereure, J., Bucheton, B., El-Saf, S., Dessein, A., Lanotte, G., and Dedet, J. P. (2001). Sudan : the possible original focus of visceral leishmaniasis. *Parasitology*, 12(10) :599–605.

Punta, M., Coggill, P. C., Eberhardt, R. Y., Mistry, J., Tate, J., Boursnell, C., Pang, N., Forslund, K., Ceric, G., Clements, J., Heger, A., Holm, L., Sonnhammer, E. L., Eddy, S. R., Bateman, A., and Finn, R. D. (2012). The pfam protein families database. *Nucleic Acids Research*, 40(D1) :D290–D301.

Punta, M. and Ofran, Y. (2008). The rough guide to in silico function prediction, or how to use sequence and structure information to predict protein function. *PLoS Comput Biol*, 4(10) :e1000160.

Quackenbush, J. (2001). Computational analysis of cdna microarray data. *Nature Reviews Genetics*, 6 :418–428.

Quevillon, E., Silventoinen, V., Pillai, S., Harte, N., Mulder, N., Apweiler, R., and Lopez, R. (2005). Interproscan : protein domains identifier. *Nucleic Acids Res.*, 1(33) :W116–120.

Rabiner, L. R. (1989). A tutorial on hidden markov models and selected applications in speech recognition. *Bioinformatics*, 77(22) :257–286.

Raymond, F., Boisvert, S., Roy, G., Ritt, J. F., Légaré, D., Isnard, A., Stanke, M., Olivier, M., Tremblay, M. J., Papadopoulou, B., Ouellette, M., and Corbeil, J. (2012). Genome sequencing of the lizard parasite leishmania tarentolae reveals loss of genes associated to the intracellular stage of human pathogenic species. *Nucleic Acids Research*, 40(3) :1131–1147.

Remmert, M., Biegert, A., Hauser, A., and Soding, J. (2012). Hhblits : lightning-fast iterative protein sequence searching by hmm-hmm alignment. *Nat Methods*, 9(2) :173–5.

Reyburn, H. (2010). New who guidelines for the treatment of malaria. *BMJ*, 28(340).

Richards, T. A., Dacks, J. B., Campbell, S. A., Blanchard, J. L., Foster, P. G., McLeod, R., and Roberts, C. W. (2006). Evolutionary origins of the eukaryotic shikimate pathway : Gene fusions, horizontal gene transfer, and endosymbiotic replacements. *EUKARYOTIC CELL*, 5(9) :1517–1531.

Richardson, J. (1981). The anatomy and taxonomy of protein structure. *Adv. Protein Chem*, 34 :167–339.

Roberts, L. and Janovy, J. (2008). *Gerald D. Schmidt and Larry S. Roberts' foundations of parasitology*, chapter Kinetoplasta : Trypanosomes and their Kin. McGraw-Hill Higher Education, Boston.

Ross, L. (2001). Bioinformatics and the malaria genome : facilitating access and exploitation of sequence information. *Molecular and Biochemical Parasitology*, 118(2) :139–145.

Rost, B. (2001). Review : protein secondary structure prediction continues to rise. *J Struct Biol*, 134 :204–218.

Sagot, M. and 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.

Sammut, S. J., Finn, R. D., and Bateman, A. (2008). Pfam10 years on : 10 000 families and still growing. *BRIEFINGS IN BIOINFORMATICS*, 9(3) :210–219.

Santos, J. M., Graindorge, A., and Soldati-Favre, D. (2012). New insights into parasite rhomboid proteases. *Molecular and Biochemical Parasitology*, 182 :27–36.

Schultz, J., Milpetz, F., Bork, P., and Ponting, C. P. (1998). Smart, a simple modular architecture research tool : identification of signaling domains. *Proc. Natl. Acad. Sci. U.S.A.*, 95(11) :5857–5864.

Schwartz, D. and Gygi, S. (2005). An iterative statistical approach to the identification of protein phosphorylation motifs from large-scale data sets. *Nature Biotechnol.*, 23 :1391–1398.

SELA, M., WHITE, F., and ANFINSEN, C. (1957). Reductive cleavage of disulfide bridges in ribonuclease. *Science*, 125 :691–692.

Selengut, J. D., Haft, D. H., Davidsen, T., Ganapathy, A., Gwinn-Giglio, M., Nelson, M. C., Richter, A. R., and White, O. (2007). Tigrfams and genome properties : tools for the assignment of molecular function and biological process in prokaryotic genomes. *Nucleic Acids Res.*, 35(Database Issue) :D260–D264.

SHAW and Jeffrey, J. (1994). Taxonomy of the genus leishmania : present and future trends and their implications. *Mem. Inst. Oswaldo Cruz*, 89(3) :471–478.

Shoemaker, B. A. and Panchenko, A. R. (2007). Deciphering protein-protein interactions. part ii. computational methods to predict protein and domain interaction partners. *PLoS Comput Biol.*, 3(4) :e43.

Sickmeier, M., Hamilton, J. A., LeGall, T., Vacic, V., Cortese, M. S., Tantos, A., Szabo, B., Tompa, P., Chen, J., Uversky, V. N., Obradovic, Z., and Dunker, A. K. (2007). Disprot : the database of disordered proteins. *Nucleic Acids Res.*, 35 :D786–D793.

Sigrist, C. J., Cerutti, L., De Castro, E., Langendijk-Genevaux, P. S., Bulliard, V., Bairoch, A., and Hulo, N. (2010). Prosite, a protein domain database for functional characterization and annotation. *Nucleic Acids Res.*, 38 :D161–166.

Sigrist, C. J., De Castro, E., Langendijk-Genevaux, P. S., Le Saux, V., Bairoch, A., and Hulo, N. (2005). Prorule : a new database containing functional and structural information on prosite profiles. *Bioinformatics*, 21(21) :4060–4066.

Sjolander, K., Karplus, K., Brown, M., Hughey, R., Krogh, A., Mian, I., and Haussler (1996). Dirichlet mixtures : a method for improved detection of weak but significant protein sequence homology. *Bioinformatics*, 12(4) :327–345.

Skerker, J. M., Prasol, M. S., Perchuk, B. S., Biondi, E. G., and Laub, M. T. (2005). Two-component signal transduction pathways regulating growth and cell cycle progression in a bacterium : a system-level analysis. *PLoS Biol.*, 3(10) :e334.

Soding, J. (2005). Protein homology detection by hmm-hmm comparison. *Bioinformatics*, 21(7) :951–960.

Soding, J. and Remmert, M. (2011). Protein sequence comparison and fold recognition : progress and good-practice benchmarking. *Current opinion in Structural Biology*, 21 :404–411.

Sonnhammer, E., Eddy, S. R., Birney, E., Bateman, A., and Durbin, R. (1998). Pfam : multiple sequence alignments and hmm-profiles of protein domains. *Nucleic Acids Research*, 26 :320–322.

Sonnhammer, E., Eddy, S. R., and Durbin, R. (1997). Pfam : a comprehensive database of protein domain families based on seed alignments. *Proteins*, 28 :405–420.

Stein, A., Ceol, A., and Aloy, P. (2011). 3did : identification and classification of domain-based interactions of known three-dimensional structure. *Nucleic Acids Res.*, 39 :D718–723.

Stevens, J., Noyes, H., and Gibson, W. (1998). The evolution of trypanosomes infecting humans and primates. *Mem. Inst. Oswaldo Cruz*, 93(5) :669–676.

Stirnimann, C., Petsalaki, E., Russell, R., and Müller, C. (2010). Wd40 proteins propel cellular networks. *Trends Biochem Sci.*, 35(10) :565–574.

Struhl, K. (1989). Helix-turn-helix, zinc-finger, and leucine-zipper motifs for eukaryotic transcriptional regulatory proteins. *Trends Biochem. Sci.*, 14(4) :137–140.

Ta, H. X. and Holm, L. (2009). Evaluation of different domain-based methods in protein interaction prediction. *Biochem Biophys Res Commun.*, 390(3) :347–362.

Terrapon, N., Gascuel, O., Maréchal, E., and Bréhélin, L. (2009). Detection of new protein domains using co-occurrence : application to plasmodium falciparum. *Bioinformatics*, 25 :3077–3083.

Terrapon, N., Gascuel, O., Maréchal, E., and Bréhélin, L. (2012). Fitting hidden markov models of protein domains to a target species : application to plasmodium falciparum. *BMC Bioinformatics*, 13 :67.

Thomas, P. D., Campbell, M. J., Kejariwal, A., Mi, H., Karlak, B., Daverman, R., Diemer, K., Muruganujan, A., and Narechania, A. (2003). Panther : a library of protein families and subfamilies indexed by function. *Genome Res.*, 13(9) :2129–2141.

Uversky, V., Oldfield, C., and Dunker, A. (2007). Ieee 7th bibe invited tutorial lecture : Intrinsically disordered proteins in human diseases. In *Bioinformatics and Bioengineering, 2007. BIBE 2007. Proceedings of the 7th IEEE International Conference on*, pages 12–12.

Velthuis, A. J. W. and Bagowski, C. P. (2008). Linking fold, function and phylogeny : A comparative genomics view on protein (domain) evolution. *Curr Genomics*, 9(2) :88–96.

Vlahovicek, K., Kajan, L., Agoston, V., and Pongor, S. (2005). The sbase domain sequence resource, release 12 : prediction of protein domain-architecture using support vector machines. *Nucleic Acids Res.*, 33 :D223–225.

Vogel, C., Berzuini, C., Bashton, M., Gough, J., and Teichmann, S. A. (2004). Supra-domains : Evolutionary units larger than single protein domains. *J Mol Biol*, 336 :809–823.

Wardl, P., Equinet, L., Packer, J., and Doerig, C. (2004). Protein kinases of the human malaria parasite plasmodium falciparum : the kinome of a divergent eukaryote. *BMC Genomics*, 5 :79.

White, N. J. (2008). Plasmodium knowlesi : The fifth human malaria parasite. *Clin Infect Dis.*, 46(2) :172–173.

Whitworth, D. and Cock, P. (2009). Evolution of prokaryotic two-component systems : insights from comparative genomics. *Amino Acids*, 37(3) :459–466.

Wong, W. C., Maurer-Stroh, S., and Eisenhaber, F. (2011). The janus-faced e-values of hmmer2 : extreme value distribution or logistic function ? *J Bioinform Comput Biol.*, 9(1) :179–206.

Wootton, J. and Federhen, S. (1993). Statistics of local complexity in amino acid sequences and sequence databases. *Comput. and Chem.*, 17(2) :149–163.

Wu, C. H., Nikolskaya, A., Huang, H., Yeh, L. L., Natale, D. A., Vinayaka, C. R., Hu, Z. Z., Mazumder, R., Kumar, S., Kourtesis, P., Ledley, R. S., Suzek, B. E., Arminski, L., Chen, Y., Zhang, J., Cardenas, J. L., Chung, S., Castro-Alvear, J., Dinkov, G., and Barker, W. C. (2004). Pirsf : family classification system at the protein information resource. *Nucleic Acids Research*, 32(Database issue) :D112–D114.

Yang, S. and Bourne, P. E. (2009). The evolutionary history of protein domains viewed by species phylogeny. *PLoS ONE*, 4(12) :e8378.

Yeats, C., Redfern, O., and Orengo, C. (2010). A fast and automated solution for accurately resolving protein domain architectures. *Bioinformatics*, 15(26) :745–751.

Yellaboina, S., Tasneem, A., Zaykin, D. V., Raghavachari, B., and Jothi, R. (2011). Domine : A comprehensive collection of known and predicted domain-domain interactions. *Nucleic Acids Research*, 39(Database Issue) :D730–735.

Zhang, X. C., Wang, Z., Zhang, X., Le, M. H., Sun, J., Xu, D., Cheng, J., and Stacey, G. (2012). Evolutionary dynamics of protein domain architecture in plants. *BMC Evolutionary Biology*, 12(6).

Zink, A. R., Spigelman, M., Schraut, B., Greenblatt, C. L., Nerlich, A. G., and Donoghue, H. D. (2006). Leishmaniasis in ancient egypt and upper nubia. *Emerg Infect Dis.*, 12(10) :1616–1617.