Université d'Evry Val d'Essonne
Laboratoire Statistique et Génome

HDR

mémoire présenté en vu de l'obtention de
l'Habilitation à Diriger des Recherches

par

Claudine Landès-Devauchelle

De l'art de résumer pour tenter de comprendre en génomique évolutive

HDR soutenue le 18 octobre 2011 devant le jury composé de :

M.	Sébastien Aubourg	INRA URGV	(Rapporteur)
M.	Dominique Higuet	UPMC	(Rapporteur)
M^{me}	Frédérique Lisacek	SIB	(Examinatrice)
M.	Bernard Prum	UEVE	(Examinateur)
M.	François Rodolphe	INRA MIG	(Rapporteur)
M.	Jean-Pierre Rousset	UPS	(Examinateur)

À Marie et Alix...
À mon père...

Je remercie Sébastien Aubourg, Dominique Higuet, Frédérique Lisacek, Bernard Prum, François Rodolphe et Jean-Pierre Rousset d'avoir accepté de participer à ce jury d'habilitation à diriger des recherches. Un grand merci à tous pour avoir pris le temps de lire ce manuscrit qui tente de décrire succinctement des travaux à l'interface de plusieurs disciplines.

Je remercie Bernard Prum et Christophe Ambroise pour l'accueil au sein du laboratoire statistiques et génome, de cet "extra-terrestre biologiste" que je suis et pour avoir rendu possible l'intégration des "petits Risler".

Je remercie tous les membres du laboratoire statistiques et génome pour la bonne ambiance qu'il y régne, pour les cafés du matin, pour les repas du midi et les nombreuses discussions informelles. Un petit clin d'oeil à Gilles et Maurice pour tous les dépannages informatiques et à Julien et Catherine pour les coups de pouce en latex. Un coucou particulier à Etienne qui m'a motivé par une course à l'HDR...

Un grand merci à Carène et Yolande pour leur aide au quotidien tant en enseignement qu'en recherche et pour leur soutien constant. Un petit clin d'oeil à Michèle toujours là au bon moment.

Je remercie, Marc et Hélène pour les nombreuses discussions sur les gyrases inverses, ainsi qu'Eduardo, Ivan et Gilles avec lesquels j'ai tant apprécié de collaborer autour des outils de classification sans alignement.

Je remercie tous les membres du département de biologie pour leur accueil lors de mon intégration à l'université d'Evry. Même si je râle beaucoup, j'ai plaisir à travailler avec eux et à enseigner dans cette petite université où l'interaction entre collègues et avec les étudiants est forte.

Je remercie également Fariza, Farida, Franck, Florence, collègues du département d'informatique qui participent à l'équipe pédagogique de la filière Génie Biologique et Informatique, ainsi que les étudiants que j'ai cotoyé pendant toutes ces années d'exercice.

Je remercie Sophie, Marie-Odile, Ivan, Carène et Catherine, relecteurs des mois d'été. Je remercie également tous mes amis et anciens collègues du laboratoire Génome et Informatique, toujours là pour me soutenir dans les moments de doutes. C'est un peu beaucoup grâce à vous que ce document est arrivé à son terme.

Que mes maîtres : Jean-Loup, physicien qui m'a appris la biochimie, Alain, généticien qui m'a appris les statistiques, et Alex, physicien qui m'a appris les mathématiques, trouvent dans ce petit paragraphe l'expression de ma profonde gratitude. Sans vous ce document n'aurait jamais existé et sans vous je ferais un autre métier... J'ai eu un immense plaisir à travailler avec vous et à jouer avec les données et les concepts dans des conditions propices, sans stress excessif.

Et pour finir une dédicace particulière à mes filles, Marie et Alix, qui n'ont eu d'autres choix que de sacrifier leurs vacances pendant cette année de rédaction... Un petit clin d'oeil à mon père qui aurait fait, j'en suis sûre, le déplacement jusqu'à la capitale. Enfin un grand merci à toute ma famille pour son soutien et pour son aide malgré l'éloignement géographique.

Table des matières

Liste des figures

Introduction

Il est difficile de résumer 18 ans de recherches en quelques pages. Les différents travaux que j'ai menés jusqu'ici m'ont toujours conduite à développer conjointement des méthodes et à les appliquer sur des séquences, en collaboration étroite avec des biologistes experts du domaine étudié. Ceci faisant j'ai souvent été conduite à améliorer ou adapter les outils afin de mieux prendre en compte les spécificités des séquences traitées. C'est ce va-et-vient perpétuel entre données et méthodes qui me motive depuis plusieurs années.

Ecrire un mémoire d'habilitation à diriger des recherches présente l'avantage de *se poser* un instant pour faire le bilan de ses activités scientifiques depuis la thèse. Avec le recul, il apparaît que le fil conducteur de mes travaux de recherche est l'étude des familles de protéines en utilisant essentiellement des méthodes relevant des statistiques descriptives afin d'*observer pour mieux déduire*. L'idée étant de *faire parler les données* en introduisant le moins de paramètres possible. Bien évidemment les méthodes que j'ai été amenée à développer, modifier ou utiliser ne peuvent pas atteindre complètement cet objectif mais cette ligne de conduite est systématiquement présente dans les outils bioinformatiques que j'ai utilisés ou réalisés.

Comment décrire les données, le plus objectivement possible, comment les résumer à quelques images ou quelques chiffres sans trop les trahir et que nous apprennent-elles ? La connaissance biologique n'est injectée qu'en fin de traitement en tant que variable explicative, ou en tout début lors de la construction du jeu de données. Pourquoi avoir pris un tel point de vue alors que je suis biologiste et que ce qui m'intéresse avant tout, c'est ce que je peux apprendre sur les données ?

Tout d'abord parce que cette approche, caractéristique de l'analyse exploratoire des données, est tout à fait classique dès qu'il s'agit de décrire de gros tableaux de chiffres. Elle est appliquée sur tout type de données et l'interprétation dépend naturellement du domaine d'application : économie, sociologie, biologie, médecine, reconnaissance des formes... Les deux grandes catégories de méthodes que j'ai utilisées sont les analyses factorielles et la classification automatique. Leur utilisation en biologie moléculaire remonte aux années 80 avec les premières études sur l'usage des codons.

Ensuite parce que cette approche est devenue indispensable depuis l'avènement de la biologie à large échelle (avec l'arrivée des premiers génomes complets) et l'accroissement exponentiel du nombre de données produites. J'ai eu la chance de participer à cette aventure dès le début, puisque le Centre de Génétique Moléculaire à Gif-sur-Yvette, où j'effectuais ma thèse, était impliqué dans le séquençage du chromosome III de *S.cerevisiae*, premier génome eucaryote à avoir été entièrement séquencé en 1996.

Ce descriptif de mes travaux se présentera donc en trois parties : la première présentera quelques-uns de mes travaux antérieurs, suivie par une description de mes activités actuelles et enfin je terminerai par mon projet de recherche.

Résumé des activités passées. J'ai choisi de centrer cette partie autour de la méthode appelée ORM (*Observed Rates Matrices*) qui permet d'analyser de grandes familles de gènes ou de protéines, dans le but de tenter de comprendre leur évolution. L'idée de base est de décrire un alignement multiple de séquences en individualisant chacune des *paires* de séquences alignées (article [1] en Annexe A.2). Pour chaque paire, on calcule la matrice observée de taux de mutations. Chacune de ces matrices de taux de mutations peut être vue comme un point dans un espace à plusieurs dimensions. L'examen de la structure du nuage de points représentant l'alignement multiple permet de vérifier (ou non) que les paires de séquences étudiées *obéissent* (ou non) *à un même générateur d'évolution.*

Toujours avec l'idée de résumer l'information contenue dans une matrice, nous avons développé une nouvelle méthode pour associer une topologie d'arbre raciné à une *matrice arbitraire* de distances ou de scores. Il existe beaucoup de méthodes pour passer d'une matrice de distances à une classification. Le problème en général, c'est que de petites variations dans la matrice de distances initiale peuvent donner des topologies différentes dans les dendogrammes résultats.

L'idée principale de notre méthode est de remplacer la matrice de départ par une matrice de rangs (article [2] en Annexe A.3). Dans la grande majorité des cas, la matrice des rangs donne une classification trop pauvre (c'est à dire qu'elle ne donne correctement que les gros regroupements). Nous avons donc proposé une procédure dynamique alliée à la stratégie *diviser pour régner* pour obtenir une classification descendante hiérarchique fondée sur les rangs, plus robuste que les méthodes standards.

Résumé des activités actuelles. Mes derniers travaux de recherche appartiennent au champ de la classification sans alignement (articles [3, 4, 5] en Annexe A.4, A.5, A.6). En effet, les programmes d'alignements multiples, en dépit des nombreux développements réalisés, butent toujours sur le positionnement correct des insertions/délétions car cette information relève plus de la structure 3D de la protéine que de la séquence primaire. Or pour classer des séquences, il est inutile de les aligner car on peut utiliser les similarités locales comme descripteurs des séquences, ce qui augmente la rapidité de traitement et permet de se débarrasser du problème des insertions/délétions. Le paramètre délicat à régler alors est la taille des mots à considérer ainsi que les erreurs à autoriser. Nous avons développé une méthode originale (*MS4* pour *Multi Scale Selector of Sequence Signatures*) qui adapte automatiquement la taille des similarités locales et le nombre d'erreurs pour un jeu de séquences donné. Le seul paramètre à fixer est assez intuitif puisqu'il s'agit du nombre maximum de répétitions autorisées au sein d'une même séquence et que sa valeur par défaut donne de bons résultats dans les cas que nous avons testés.

MS4 recode chacun des segments de similarité en utilisant un alphabet

plus grand où chaque similarité locale est codée par un nouveau caractère. La simple mise en couleur des caractères identiques dans cet alphabet fait apparaître visuellement les blocs conservés.

Nous avons utilisé cette méthode pour expertiser la famille des *DNA topoisomérases IA*. Cette famille d'enzymes participe au maintien de l'ADN dans une topologie correcte lors de tous les processus cellulaires qui déplacent des machineries le long de la double hélice : réplication, transcription, recombinaison... Les DNA topoisomérases IA forment une famille multigénique qui a subi plusieurs duplications au cours de son histoire.

En collaboration avec l'équipe de Marc Nadal à l'Institut de Génétique Moléculaire d'Orsay, nous nous intéressons plus particulièrement à l'étude de la sous-fonctionnalisation des deux copies de gyrases inverses (DNA topoisomérases spécifiques des organismes thermophiles) chez les crénarchées hyperthermophiles [1].

Résumé du projet de recherche. Toujours dans le but de résumer l'information, je voudrais maintenant utiliser mes compétences en analyse de données pour m'intéresser à l'étude des gènes dupliqués en utilisant une approche intégrative. La question d'intérêt est de comprendre le devenir d'un réseau d'interactions génétiques ou protéiques après une duplication.

Les processus de duplications, totales ou partielles, de génome sont des phénomènes importants dans l'évolution des génomes eucaryotes. Nous voulons les étudier en séparant les différents types de duplications : duplication à large échelle (*WGD - Whole Genome Duplication*), à moyenne échelle (*SD - Segmental Duplication*) ou à petite échelle (*TAG - Tandemly Arrayed Genes duplications*).

Ce sujet se situe à cheval entre deux thématiques : la génomique évolutive et la biologie systémique. L'objectif est d'étudier le devenir d'un réseau biologique (génétique, métabolique ou d'interactions protéiques) après duplication de tout ou partie de ses constituants. Il est maintenant généralement admis que la duplication complète de génome est un réservoir de variabilité génétique pour l'acquisition de nouvelles fonctions (*néo-fonctionnalisation*) ou la sous-spécialisation de fonction (*sous-fonctionnalisation*) chez les eucaryotes.

Nous comptons étendre les questions posées par Ohno dans le cas de la duplication de gènes aux duplications de réseaux : quand un réseau est dupliqué, que peut signifier par exemple sa "néo-fonctionnalisation", sa "sous-fonctionnalisation" ? Par ailleurs, le maintien dans un génome d'un réseau dans son ensemble est-il lié à sa fonction, au nombre de ses constituants, à sa topologie ?

En collaboration avec deux équipes de l'Unité de Recherche en Génomique Végétale à Evry (celle de Sébastien Aubourg et celle de Boulos Chaloub), nous étudierons cette question chez les plantes, dont on connaît

1. La phylogénie des archées fait encore débat dans la communauté. On distingue quatre grands groupes taxonomiques : les Crenarcheota, les Euryarcheota, les Korarcheaota et les Thaumarcheota. Le groupe des crénarchées se caractérise par la présence de nombreux taxons vivant dans des milieux extrêmes (température, salinité, pH...).

la propension importante à la polyploïdisation [2]. Dans un premier temps chez l'arabette pour laquelle on a le génome complet pour deux espèces (*A.thaliana* et *A.lyrata*) et de nombreuses données "à large échelle" (transcriptome, interactome, métabolome, collections de mutants *knockout* [3]). Ces données sont organisées dans des banques de données dédiées et en partie expertisées.

Digression sur l'enseignement en bioinformatique. Depuis ma nomination en 1993, j'enseigne la bioinformatique dans les filières de biologie, du premier cycle à l'Ecole Doctorale. J'ai d'abord exercé à l'Université de Versailles Saint-Quentin, puis désormais à l'Université d'Evry Val d'Essonne. Il m'est par conséquent impossible de passer sous silence cette activité qui occupe officiellement la moitié de mon temps de travail, d'autant qu'elle est très enrichissante à plusieurs points de vue :

- elle élargit l'horizon intellectuel car il est rare d'enseigner dans son domaine de recherche direct,
- elle nécessite d'aller au fond des choses -*ce qui est clair s'énonce clairement*,
- elle permet de garder le contact avec des personnes jeunes,
- elle a une utilité plus immédiatement perceptible que la recherche car elle forme des jeunes et leur permet de s'intégrer dans le marché du travail
- elle m'a poussée en avant car j'ai dû assumer assez tôt des responsabilités pédagogiques et administratives [4].

Le terme bioinformatique est ambigu en français puisqu'il s'applique tout aussi bien à un biologiste faisant des analyses *in silico* avec des outils existants (*Bioinformatics* en anglais) qu'à un informaticien ou un mathématicien pour lesquel la biologie n'est qu'un terrain d'application parmi d'autres (*Computational Biology* en anglais). Je l'utilise ici car il reflète bien les deux pendants de mon activité pédagogique.

J'enseigne d'une part à un public de biologistes pour lesquels l'informatique est un outil qui va leur permettre de mener à bien des expériences *in silico*. L'objectif est double : 1) qu'ils comprennent suffisamment les algorithmes sous-jacents pour maîtriser les paramètres en entrée des programmes ; 2) qu'ils soient critiques vis-à-vis de la qualité des données biologiques et la signification des résultats en sortie.

D'autre part, j'enseigne depuis plusieurs années dans des filières formant des biologistes à la "Computational biology". L'objectif est qu'ils acquièrent une seconde compétence en informatique et en mathématiques afin de s'intégrer avec succès dans des équipes de développement ou des équipes gérant et analysant des données biologiques en masse.

Il y a une trentaine d'années, les toutes premières formations de bioinformatique s'adressaient à des biologistes car c'est cette communauté qui

2. polyploïdisation : augmentation du nombre de jeux de chromosomes par agrégation des génomes ou au cours de croisement entre espèces.

3. mutants pour lesquels on a inactivé ou délété spécifiquement, un ou plusieurs gènes.

4. mon cas est loin d'être rare dans les petites universités.

était confrontée au manque de spécialistes et de méthodes pour analyser des données que l'on pouvait déjà qualifier de "à grande échelle". Depuis une quinzaines d'années, on a vu s'ouvrir beaucoup de formations de bioinformatique s'adressant à des biologistes, à des informaticiens, ou parfois aux deux en même temps. Savoir si on doit mettre ensemble des étudiants en biologie et des étudiants en informatique ou si on doit organiser des filières spécifiques est une question naturelle et récurrente puisque la bioinformatique est par essence transdisciplinaire.

La réponse me semble dépendre du but poursuivi. Si l'objectif est de favoriser la créativité scientifique, il faut des filières spécifiques à l'intérieur des disciplines principales, car l'intuition nécessaire ne vient qu'après une longue pratique du domaine (ce qui est la définition même de discipline principale). La formation pour la seconde compétence est tournée vers l'acquisition d'outils : il faut connaître les méthodes disponibles et les comprendre assez pour les adapter au problème à traiter.

Dix ans d'étroite collaboration avec des physiciens théoriciens m'ont confortée dans ce point de vue. Il est plus facile d'écrire un code numérique raisonnablement efficace, en piochant dans les *Numerical Recipes*, que de dégager la quintessence des connaissances d'un champ disciplinaire. Cette faculté est réservée à des esprits brillants, dominant leur domaine, et capables d'expliquer simplement les notions nécessaires à la compréhension d'un concept. En d'autres termes, il faut de mon point de vue des formations spécifiques, par discipline, dès la licence. Elles peuvent se rejoindre éventuellement en Master, à condition de garder des options propres au champ disciplinaire d'origine. Ces options sont destinées à approfondir les concepts souvent complexes nécessaires en bioinformatique.

Evolution des familles multigéniques

1

Sommaire

Avant-propos

Les travaux inclus dans cette première partie ont été initiés lors de mon passage au Centre de Physique Théorique de Luminy à Marseille en 1994 et 1995, pendant deux séjours de 6 mois, dans l'équipe ondelettes dirigée par **Bruno Torrésani**. C'est à cette occasion que j'ai eu le plaisir de rencontrer **Alex Grossmann**, physicien théoricien, avec lequel je collabore toujours. Ces travaux se sont poursuivis lors de mon retour en région parisienne (à l'Université de Versailles Saint-Quentin) par l'application de notre méthode à l'étude de l'évolution des génomes mitochondriaux en collaboration avec **Monique Monnerot** du Centre de Génétique Moléculaire à Gif-sur-Yvette. N'oublions pas de citer **Alain Hénaut** et **Jean-Loup Risler** sans l'expérience desquels ce travail bioinformatique n'aurait pas eu lieu.

La section 1.1 est une présentation du modèle de Margaret Dayhoff et collaborateurs, premier modèle évolutif à avoir été décrit.

La section 1.2 introduit les notations nécessaires pour la présentation de notre méthode de description des alignements.

La section 1.3 décrit la méthode ORM - *Observed Rate Matrices*. Ces travaux ont donné lieu à deux articles :

1. un article descriptif de la méthode [1] (joint en annexe A.2),
2. un article utilisant les vecteurs de mutabilités (cf Section1.3.4) en tant que descripteurs *naturels* de quartets [1] par le biais des angles qu'ils forment entre eux [24].

Sur cette thématique, j'ai été amenée à encadrer les étudiants suivants : Cyril GRANDCOIN (9 mois en 2000), Stéphane PHILIPPE (3 mois en 2002, 6 mois en 2003), Elodie BOUCOMONT (2 mois en 2005), Guillaume CAMANES (2 mois en 2011).

La section 1.4, décrit les grandes lignes d'une méthode de classification fondée sur les rangs, que nous avons développée en collaboration avec l'équipe d'**Andréas Dress** à Bielefeld en Allemagne. Ces travaux ont donné lieu à un article [2] (inclus en annexe A.3) et à l'encadrement d'un stagiaire : Sébastien de RENTY (6 mois en 2004).

1. quartets : désignent les ensembles de 4 séquences utilisés dans certaines méthodes de reconstructions phylogénétiques

Les premières séquences à avoir été déterminées sont les protéines (séquence de l'insuline bovine en 1951). Il est donc naturel que les premières méthodes bioinformatiques développées aient travaillé sur l'alphabet à 20 lettres des acides aminés. Il est en de même pour le premier modèle d'évolution, introduit par Dayhoff et Eck en 1966 dans *Atlas of Protein Sequence and Structure* .

1.1 Le modèle d'évolution de Dayhoff

J'ai eu l'occasion de feuilleter le volume de 1972 du même *Atlas of Protein Sequence and Structure*, publié par le NBRF (National Biomedical Research Foundation), où Margaret Dayhoff et ses collaborateurs [6] publient leur modèle pour l'évolution des protéines. Ce livre, qui ne compte que 400 pages, contient cependant l'ensemble des séquences nucléiques et protéiques publiées avant janvier 1971 (ainsi que des schémas de structures 3D, des alignements multiples pour 66 familles protéiques connues et une dizaine d'articles scientifiques). Il constitue une des premières versions de la banque PIR *Protein Information Ressource* [7] - banque répertoriant toutes les séquences protéiques connues - ce qui en dit long sur la croissance spectaculaire des données de séquences en 40 ans.

Conjointement au catalogue des séquences complètes connues, il décrit un certain nombre de méthodes bioinformatiques permettant de réaliser les premières études comparatives : "Building a phylogenetic tree : cytochrome c", "Gene duplication in evolution : the globins", "Evolution of complex system : the immunoglobins"... ainsi qu'un modèle d'évolution des remplacements d'acides aminés dans les protéines : "A model of Evolutionary Change in Proteins" [6].

1.1.1 Les données

L'idée de Margaret Dayhoff était de mesurer les mutations [2] accumulées au cours du temps, pour les familles pour lesquelles on dispose d'un alignement, afin de construire un modèle évolutif. A l'époque très peu de séquences étaient connues et les méthodes d'alignements en étaient à leur début [8, 9, 10]. Par conséquent les alignements multiples publiés sont peu nombreux et construits entièrement manuellement. Les premiers algorithmes d'alignements multiples n'apparaissent qu'à la fin des années 80 ([11, 12, 13, 14]).

Dans les figures 9-1 et 9-2 de son article (reportées ici Fig. 1.1) Margaret Dayhoff illustre comment est rempli, à partir d'un alignement multiple, le tableau de comptage des *mutations observées*. L'arbre phylogénétique représentant l'histoire évolutive des séquences alignées est reconstruit. A chaque noeud la séquence ancestrale et ses variantes sont indiquées. Il suffit ensuite de compter le nombre de mutations nécessaires pour passer de la séquence mère à la séquence fille. Ces comptages sont symétrisés : par exemple dans la Fig. 1.1 : $B \rightarrow C$ est comptabilisé de même que $C \rightarrow B$ (alors que cette dernière mutation n'est pas prédite) car Dayhoff suppose explicitement que

2. Pour décrire les **mutations**, on parle de **substitutions** lorsqu'on s'intéresse à l'ADN et de **remplacements** pour les séquences protéiques.

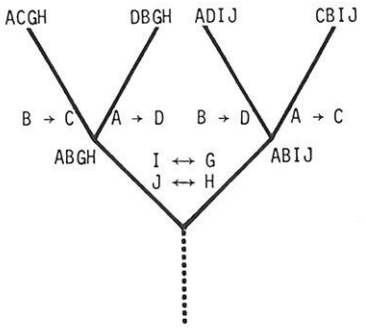


Figure 9-1. Simplified phylogenetic tree. Four "observed" proteins are shown at the top. Inferred ancestors are shown at the nodes. Amino acid exchanges are indicated along the branches.

	A	B	C	D	G	H	I	J
A			1	1				
B			1	1				
C	1	1						
D	1	1						
G							1	
H								1
I					1			
J						1		

Figure 9-2. Matrix of accepted point mutations derived from the tree of Figure 9-1.

FIGURE 1.1 – Principe du comptage des mutations à partir d'un alignement. Dans une première étape on construit l'arbre phylogénétique décrivant les séquences actuelles, en reconstruisant au niveau des noeuds les séquences ancestrales. Dans une seconde étape on compte le nombre de mutations observées entre les séquences mères et les séquences filles en symétrisant les observations. Figure extraite de [6].

la probabilité d'observer $X \to Y$ est la même que celle de $Y \to X$ pour tout X, Y et qu'elle ne dépend que de la fréquence des deux acides aminés et de leurs propriétés physico-chimiques.

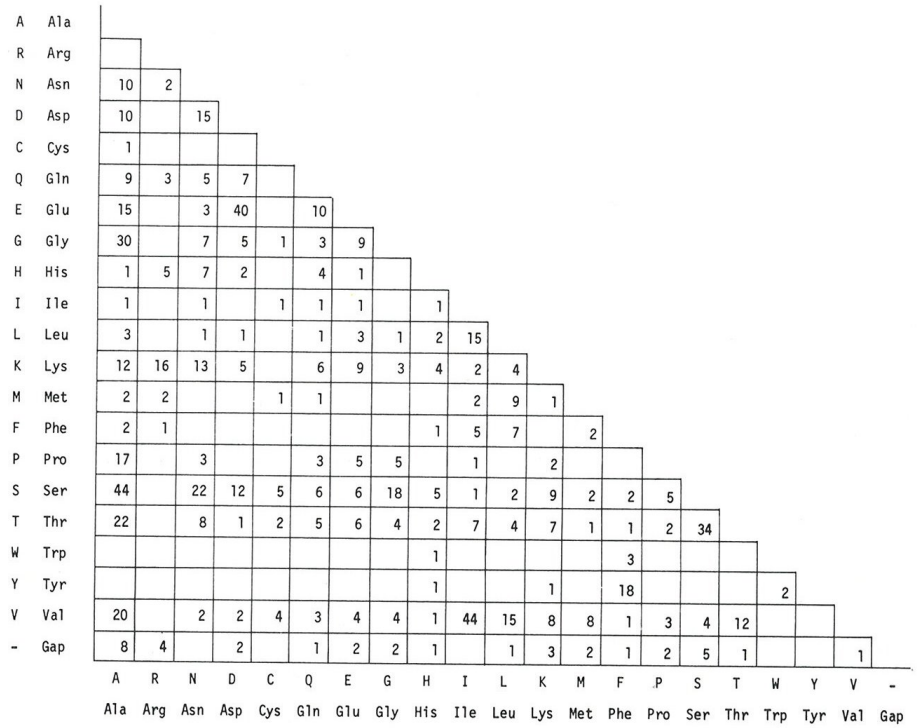
Compter les mutations en comparant les séquences actuelles aux séquences ancestrales reconstruites plutôt qu'en comparant directement les séquences actuelles entre elles, évite de compter plusieurs fois la même mutation ancestrale. Par exemple dans la Fig. 1.1, la mutation $G \longleftrightarrow I$ n'est comptée qu'une seule fois alors qu'elle l'aurait été quatre fois si on avait comparé directement les séquences actuelles ($ACGH/ADIJ$ + $ACGH/CBIJ$ + $DBGH/ADIJ$ + $DBGH/CBIJ$).

Comme à l'époque on ne connaissait pas beaucoup de familles de séquences et qu'il n'existait pas de programmes d'alignements multiples, il fallait donc que les séquences soient proches pour pouvoir aligner à la main des familles protéiques. Le pourcentage d'identité des paires de séquences étudiées par Dayhoff est supérieur à 85%. Ceci pour deux raisons :

i) l'objectif final de Dayhoff est de construire une matrice de scores entre

acides aminés pour les méthodes d'alignement. Or pour compter les remplacements, il faut préalablement aligner les séquences. Afin d'éviter tout raisonnement circulaire, elle s'est limitée aux alignements très proches, ceux pour lesquels la seule matrice identité suffit pour aligner correctement les séquences.

ii) en considérant des séquences très proches ($\leq 15\%$ de différences), elle diminue le risque de mutations multiples à un site donné. Elle peut ainsi raisonnablement espérer qu'une mutation observée correspond effectivement à un seul évènement évolutif (une mutation ponctuelle).



		A	R	N	D	C	Q	E	G	H	I	L	K	M	F	P	S	T	W	Y	V	-
A	Ala																					
R	Arg																					
N	Asn	10	2																			
D	Asp	10		15																		
C	Cys	1																				
Q	Gln	9	3	5	7																	
E	Glu	15		3	40		10															
G	Gly	30		7	5	1	3	9														
H	His	1	5	7	2		4	1														
I	Ile	1		1		1	1	1		1												
L	Leu	3		1	1		1	3	1	2	15											
K	Lys	12	16	13	5		6	9	3	4	2	4										
M	Met	2	2			1	1				2	9	1									
F	Phe	2	1							1	5	7		2								
P	Pro	17		3			3	5	5		1		2									
S	Ser	44		22	12	5	6	6	18	5	1	2	9	2	2	5						
T	Thr	22		8	1	2	5	6	4	2	7	4	7	1	1	2	34					
W	Trp									1					3							
Y	Tyr									1			1		18				2			
V	Val	20		2	2	4	3	4	4	1	44	15	8	8	1	3	4	12				
-	Gap	8	4		2		1	2	2	1		1	3	2	1	2	5	1			1	

Figure 9-3. Accepted point mutations accumulated from closely related sequences. Eight hundred fourteen changes are shown, drawn from the cytochrome c, hemoglobin α, hemoglobin β, myoglobin, virus coat protein, chymotrypsinogen, glyceraldehyde 3-phosphate dehydrogenase, clupeine, insulin A and B, and ferredoxin families in the *Atlas of Protein Sequence and Structure 1969*. The numbers in this table have been rounded off. The full precision was used in all the calculations dependent upon it.

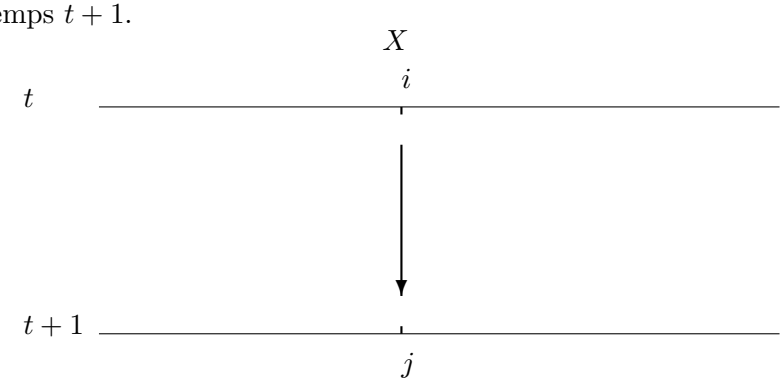
FIGURE 1.2 – *Matrice des remplacements observés par Dayhoff sur les paires de séquences alignées ayant moins de 15% de différences. Les paires de séquences ont été extraites d'alignements multiples provenant de 10 familles protéiques : cytochrome C, hémoglobine α, hémoglobine β, myoglobine, protéine de la capside du virus de la mosaïque du tabac, chymotrypsinogène, glycéraldhyde 3-phosphate déshydrognase, clupéine, insuline A et B et ferrodoxines [6].*

La matrice des mutations ponctuelles acceptées publiée par Dayhoff [6] décrit 814 remplacements observés dans 10 familles (données de 1969) (Fig. 1.2). On remarque que de nombreuses mutations ne sont jamais observées du fait du faible nombre d'alignements et de la faible probabilité de certaines mutations (C $\longleftrightarrow$ W par exemple).

C'est là qu'est toute la beauté et le caractère précurseur de ce travail, les auteurs vont utiliser un modèle, simple et de bon goût, pour simuler l'évolution des protéines et inférer ainsi les mutations non accessibles à l'observation à l'époque.

1.1.2 Le modèle

Le modèle de Dayhoff s'applique lorsque la probabilité d'une mutation ne dépend pas de son histoire, plus exactement lorsque la probabilité du remplacement d'un acide aminé à un site donné ne dépend que de la nature de cet acide aminé et non de l'état antérieur de celui-ci. En simplifiant, cela revient à dire que le modèle donne la probabilité que l'acide aminé i, occupant le site X de la protéine au temps t, mute en acide aminé j au temps $t+1$.



Ce qui s'écrit :

$$\mathbb{P}(X_t = x_j | X_{t+1} = x_i) \tag{1.1}$$

Ce temps $t+1$ est un temps relatif et correspond au temps nécessaire pour accumuler une mutation tous les 100 résidus. Cette unité de mesure de la distance évolutive introduite par Dayhoff est connue sous l'acronyme 1-PAM (*Point Accepted Mutation*).

La figure 1.3 montre la relation entre le % de différences observées entre deux séquences alignées et la distance évolutive les séparant (exprimée en PAMs) [15]. On observe que la relation entre la proportion d'acides aminés conservés et la distance évolutive n'est linaire que pour de courtes périodes de temps. Au delà de 50% de différences, les mutations s'accumulant à un même site sont nombreuses et la courbe se met à tendre vers une asymptote (ici 94%, ce qui correspond approximativement à un pourcentage d'identités de 1/20, soit la fréquence moyenne d'un acide aminé).

La courbe de la figure 1.3 permet de vérifier qu'en ne considérant que les paires de séquences alignées ayant moins de 15% de différences, les résultats de Dayhoff et des ses collaborateurs se trouvent dans la partie linéaire de la courbe, loin de la zone où les mutations multiples infléchissent la courbe.

A partir de la matrice qui cumule les remplacements observés sur 10 familles protéiques (Fig. 1.2), que les auteurs appellent la matrice $\mathcal{A}$ (de terme a_{ij} égal au nombre de paires (i,j) observées dans les alignements), les auteurs calculent la matrice $\mathcal{P}^1$ des probabilités de mutation des acides aminés en une unité de temps. Chaque terme p^1_{ij} de la matrice est calculé comme étant la probabilité conditionnelle que l'acide aminé i soit remplacé par l'aminé j sachant que le site de la protéine est dans l'état i à l'étape précédente. Ce qui s'écrit :

$$p^1_{ij} = \frac{a_{ij}}{\sum_j a_{ij}} \tag{1.2}$$

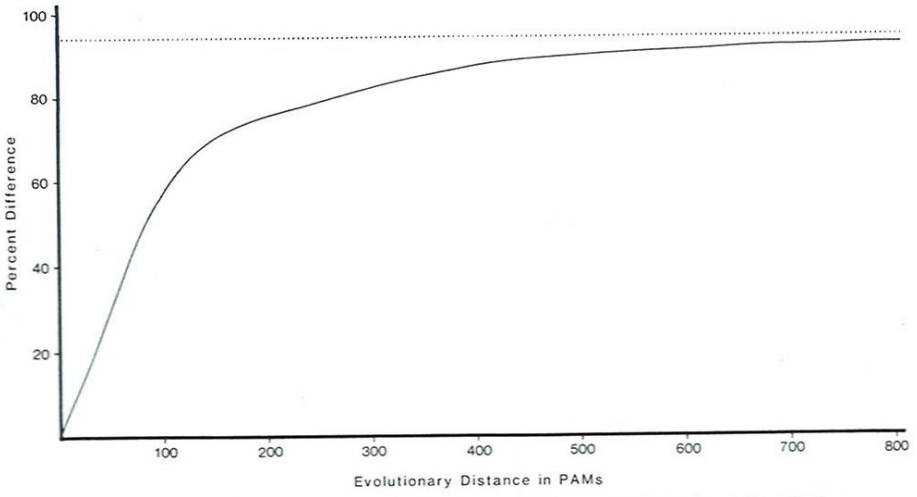


FIGURE 1.3 – *Pourcentage de différences entre deux séquences protéiques en fonction de la distance évolutive [15].*

Cette matrice s'appelle matrice de transition du modèle markovien. Elle possède des propriétés mathématiques remarquables, caractéristiques d'une matrice markovienne.

Remarque 1.1 *Chaque ligne décrit les probabilités de mutation d'un acide aminé donné et la somme de chaque ligne est égale à 1. Le terme diagonal donne la probabilité de ne pas muter de cet acide aminé.*

Remarque 1.2 *La somme des éléments diagonaux donne la probabilité qu'il n'y ait pas de mutation en une unité de temps. Dans le cas de Dayhoff la matrice de transition est normalisée de façon à ce que cette somme soit égale à 99%, ce qui correspond à une unité PAM (1-PAM).*

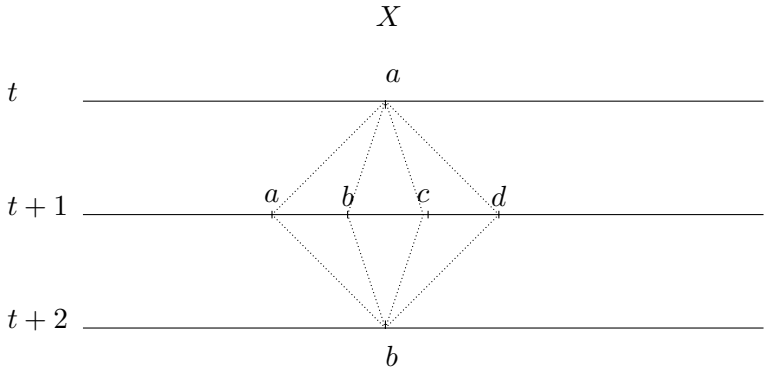
1.1.3 Inférence des changements à longue distance évolutive

A partir de la matrice $\mathcal{P}^1$ estimée, le modèle permet de calculer la probabilité que l'acide aminé i, occupant le site X de la protéine au temps t, mute en acide aminé j au temps $t+n$, pour n'importe quel n.

Afin d'illustrer le principe du calcul, considérons un alphabet à 4 lettres *abcd*. Soit la matrice $\mathcal{P}^1$ des probabilités de mutation en une unité de temps PAM (où p^1_{aa} est la probabilité pour un a de ne pas avoir muté après une étape, p^1_{ab} celle de la mutation a vers b en une étape ...) :

$$\mathcal{P}^1 = \begin{pmatrix} p^1_{aa} & p^1_{ab} & p^1_{ac} & p^1_{ad} \\ p^1_{ba} & p^1_{bb} & p^1_{bc} & p^1_{bd} \\ p^1_{ca} & p^1_{cb} & p^1_{cc} & p^1_{cd} \\ p^1_{da} & p^1_{db} & p^1_{dc} & p^1_{dd} \end{pmatrix} \tag{1.3}$$

Pour connaître la probabilité du changement $a \to b$ en 2 unités PAM, le modèle va prendre en compte toutes les possibilités pour passer de a vers b en 2 étapes (c'est à dire en passant soit par a, b, c ou d à l'étape 1).



Sachant que le modèle suppose que la mutation de a vers une autre lettre ne dépend que du fait que ce soit un a à l'étape précédente (au temps $t-1$), les différentes étapes des chemins évolutifs permettant de passer de a vers b en 2 étapes sont indépendantes les unes des autres. Ainsi la probabilité de passer de a vers b en 2-PAM (notée p^2_{ab}) se calcule :

$$p^2_{ab} = p^1_{aa}p^1_{ab} + p^1_{ab}p^1_{bb} + p^1_{ac}p^1_{cb} + p^1_{ad}p^1_{db}$$

Autrement dit, c'est le terme p^2_{ab} du produit matriciel de $\mathcal{P}^1$ par $\mathcal{P}^1$.

$$\mathcal{P}^2 = \begin{pmatrix} p^2_{aa} & p^2_{ab} & p^2_{ac} & p^2_{ad} \\ p^2_{ba} & p^2_{bb} & p^2_{bc} & p^2_{bd} \\ p^2_{ca} & p^2_{cb} & p^2_{cc} & p^2_{cd} \\ p^2_{da} & p^2_{db} & p^2_{dc} & p^2_{dd} \end{pmatrix}$$

Ainsi dans un modèle de Markov, la matrice de probabilités de mutation correspondante à de longs intervalles évolutifs est obtenue en élevant la matrice 1-PAM à la puissance souhaitée. On décline ainsi les matrices de la série PAM : PAM10, PAM20, PAM60, PAM120, PAM250 correspondant au nombre d'intervalles évolutifs inférés [3].

Remarque 1.3 *Pour de grandes distances évolutives, les valeurs de la matrice de transition d'un modèle de Markov deviennent égales à la fréquence de l'acide aminé j quelle que soit la nature de l'acide aminé initial i.*

Autrement dit, pour de grandes distances évolutives toutes les lignes de la matrice de transition sont identiques et égales à ce que l'on appelle les fréquences d'équilibre (le modèle de Markov est qualifié de *stationnaire*).

Remarque 1.4 *Les fréquences d'équilibre sont déterminées par les données sur lesquelles la matrice P^1 a été estimée.*

3. En général, la série PAM désigne plutôt les matrices de scores qui sont utilisées pour pondérer les remplacements (conservatifs ou non) dans les méthodes d'alignements de protéines. Ces matrices de score $\mathcal{R}^n$ sont calculées à partir des matrices $\mathcal{P}^n$ de probabilités de mutation (*matrice de transition*) de la manière suivante :

$$r^n_{ij} = \frac{p^n_{ij}}{f_j}$$

où f_j est la fréquence de l'acide aminé j dans les données.

Pour chaque distance évolutive, la somme des termes diagonaux [4] de la matrice de transition donne une mesure du nombre de changements attendus. La courbe pourcentage de différences en fonction de l'exposant de la matrice de Markov (Fig. 1.3) montre ce phénomène de saturation.

1.1.4 Conclusion

Le modèle décrit par Dayhoff est bien connu des mathématiciens sous le nom de modèle de Markov mais elle a été la première à l'introduire en évolution moléculaire (sans citer une seule fois le nom...). Il s'agit d'un modèle de Markov d'ordre 1 (réversible et stationnaire), c'est dire qu'il ne garde en mémoire que l'état du site à $t-1$ pour prédire l'état à t. Les modèles de Markov (également appelés chaînes de Markov) sont abondamment utilisés en biologie mais le plus souvent pour modéliser l'enchaînement des nucléotides ou des codons dans les séquences (prédiction de gènes, recherches de motifs, alignements multiples).

Bien que des matrices de mutations plus récentes aient été publiées (par exemple que BLOSUM [16] ou JTT [17] pour ne citer que les plus utilisées), les travaux de Dayhoff et collaborateurs restent d'actualité. En particulier les matrices JTT et JTT-F (largement utilisées dans les reconstructions phylogénétiques à partir de protéines) peuvent être considérées comme une réactualisation de la matrice de Dayhoff.

1.2 Modèle de Markov et évolution moléculaire

L'objectif de cette section est de rappeler les notions de modèle de Markov utilisées en phylogénie moléculaire afin d'introduire les notations que j'utiliserai par la suite. Les exemples utilisés ici sont pour des raisons didactiques des modèles sur l'alphabet à 4 lettres des nucléotides.

1.2.1 Modèle de Markov continu

Tel que je l'ai décrit jusqu'à présent, on pourrait croire que le processus modélisé n'évolue qu'avec des pas de temps discrets (t, $t+1$, $t+2$, $t+3\ldots$). Or la notion de temps que nous utilisons est une notion relative (temps pour accumuler x mutations tous les y résidus) qui varie d'une espèce à l'autre et d'une famille de gènes à l'autre. De plus l'estimation de la matrice de transition du processus est faite en cumulant des observations faites sur des paires de séquences ayant des temps de divergence variés. Pour toutes ces raisons, le modèle de Markov que l'on utilise est un modèle de Markov continu, plutôt qu'un modèle de Markov discret. Cela revient à dire que les intervalles de temps considérés sont infiniment petits. On peut même calculer le taux instantané de *disparition* de chaque acide aminé et le taux de *transformation* de cet acide aminé en un autre.

Remarque 1.5 *La matrice qui renferme ces taux de* transformation *pour un temps δt infinitésimal, s'appelle matrice de* **taux d'évolution** *dans le cas général ou bien, de manière plus précise, matrice de* **taux de remplacements** *(pour*

4. La somme des termes diagonaux d'une matrice s'appelle la trace. Ici c'est une mesure la proportion d'acides aminés restant inchangés pour cette distance évolutive.

les protéines) ou de **taux de substitutions** *(pour les gènes) ; elle est notée* $\mathcal{Q}$.

1.2.2 Taux de substitutions entre deux séquences non codantes

Le modèle nucléique le plus proche de celui utilisé par Dayhoff pour les protéines est le modèle général réversible (ou GTR pour *General Time Reversible model*) décrit dans [18]. Pour une revue complète des différents modèles évolutifs existants se reporter au chapitre 11 de l'ouvrage de Hillis *et al.*[19].

Dans le modèle GTR, chaque mutation est caractérisée par un taux de substitutions - modèle général - et la matrice de taux est symétrique - modèle réversible. C'est un modèle à 6 paramètres pour les nucléotides (Fig. 1.4).

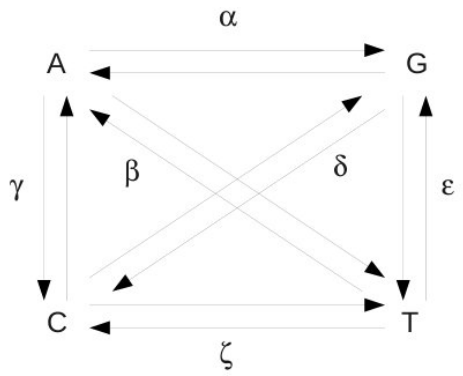


FIGURE 1.4 – *Modèle d'évolution général réversible pour l'ADN non codant.*

La matrice Q des taux de substitutions du modèle GTR est de la forme :

$$\mathcal{Q} = \begin{pmatrix} -\alpha-\beta-\gamma & \alpha & \beta & \gamma \\ \alpha & -\alpha-\epsilon-\delta & \epsilon & \delta \\ \beta & \epsilon & -\beta-\epsilon-\zeta & \zeta \\ \gamma & \delta & \zeta & -\gamma-\delta-\zeta \end{pmatrix} \tag{1.4}$$

Cette matrice Q se lit par ligne, chaque ligne caractérisant une lettre de l'alphabet. La somme marginale des lignes est égale à 0. Toute ligne de cette matrice peut être interprétée comme : *la quantité de base qui disparaît à tout instant* (terme diagonal) *et ce vers quoi elle mute* (termes extra-diagonaux).

Classiquement en évolution moléculaire, les taux de substitutions sont de l'ordre de 10^{-9} substitution par site par an (pour une discussion sur les variations des taux de substitutions observées voir [20]).

1.2.3 Calcul des probabilités de mutations

La matrice Q définie dans le paragraphe précédent (Eq. 1.4) spécifie le taux de mutation entre deux nucléotides pour un intervalle de temps δt très petit, or pour calculer le nombre de substitutions qui se sont produites entre deux séquences après une période de temps t beaucoup plus longue, il faut avoir accès aux probabilités de mutation sur une telle période.

En effet, pour connaître le nombre de mutations séparant deux séquences qui divergent depuis un temps t, il ne suffit pas de compter les différences entre les séquences actuelles car à chaque site plusieurs mutations ont pu

s'accumuler au cours du temps. Il faut donc pouvoir estimer le nombre de substitutions qui se sont produites même si les deux sites actuels sont identiques, ceci grâce à un modèle de Markov.

Pour le modèle évolutif GTR, la matrice de probabilités de substitutions de tout site au bout d'un temps t, est donnée par l'équation suivante ([21, 19]) :

$$\mathcal{P}(t) = e^{t\mathcal{Q}} \tag{1.5}$$

Le calcul de la matrice $\mathcal{P}(t)$ nécessite une décomposition de la matrice Q en valeurs propres et vecteurs propres qui se résoud aisément numériquement [19].

La matrice de probabilité $\mathcal{P}(t)$ est une matrice de Markov dont toutes les lignes se somment à 1 et qui donne pour chaque ligne, les probabilités de mutations d'un nucléotide en un autre nucléotide (terme extra-diagonaux) ou la probabilité de rester inchangé (terme diagonal).

1.2.4 Illustration pour le modèle de Jukes et Cantor

Pour le modèle évolutif de Jukes et Cantor (JC) où toutes les substitutions ont un même taux α, l'équation 1.5 a une solution analytique simple donnée ci-dessous (pour la démonstration se reporter à [19, 20]) :

$$\mathcal{Q} = \begin{pmatrix} -3\alpha & \alpha & \alpha & \alpha \\ \alpha & -3\alpha & \alpha & \alpha \\ \alpha & \alpha & -3\alpha & \alpha \\ \alpha & \alpha & \alpha & -3\alpha \end{pmatrix} \implies \mathcal{P}(t) = \begin{cases} \frac{1}{4} + \frac{3}{4}e^{-4\alpha t} \text{ si } i = j \\ \frac{1}{4} - \frac{1}{4}e^{-4\alpha t} \text{ si } i \neq j \end{cases} \tag{1.6}$$

Dans ce cas simple, on vérifie facilement que la somme marginale des lignes de $\mathcal{P}(t)$ est égale à 1 et que le système évolue, quand t augmente, vers des fréquences d'équilibre de 1/4 (modèle de Markov stationnaire). Quand le système part de l'état initial $i = j$, $p_{ij}(t)$ diminue de manière monotone vers 1/4. Quand il part de l'état initial $i \neq j$, $p_{ij}(t)$ augmente de manière monotone vers 1/4 (Fig. 1.5).

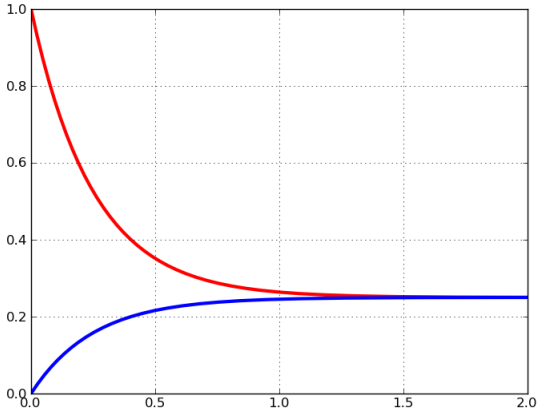


FIGURE 1.5 – *Evolution de la probabilité $p_{ij}(t)$ en fonction de t pour le modèle de JC : pour la courbe rouge lorsque $i = j$ à l'état initial, et pour la courbe bleue lorsque $i \neq j$.*

1.2.5 Prise en compte des évènements multiples

Sous le modèle de Jukes et Cantor, en considérant que les 4 bases sont équiprobables, la correction à apporter au pourcentage de différences observées D entre deux séquences x et y est donnée par l'équation suivante [20] :

$$d_{xy} = -\frac{3}{4} ln(1 - \frac{4}{3}D) \tag{1.7}$$

Cette correction permet d'obtenir la distance évolutive d_{xy} en prédisant grâce au modèle probabiliste les évènements multiples qui ne sont pas accessibles à l'observation. Dans le cas de Jukes et Cantor la formule pour prendre en compte les évènements multiples est explicite. Pour le modèle général GTR, les outils de calcul matriciel permettent de répondre à cette question en résolvant l'équation 1.5.

1.3 Matrices de taux d'évolution observées

Comme dit le chapitre introductif, notre approche consiste à décomposer un alignement multiple, décrivant une large famille multigénique, en un ensemble de matrices de taux d'évolution décrivant les paires de séquences alignées. Bien que la méthode que nous avons développée s'applique aussi bien aux séquences nucléiques que protéiques, je me suis principalement intéressée aux protéines.

L'inconvénient du modèle de Dayhoff (modèle généraliste pour les protéines), c'est qu'il comporte 190 paramètres, ce qui est beaucoup. Par conséquent, ces derniers sont estimés, même dans les versions modernes des matrices de transition [17, 22, 23], par maximum de vraisemblance à partir du comptage des mutations observées sur un ensemble hétérogène (en terme de fonctions) de protéines alignables. Cette restriction est gênante, car elle ne permet pas de prendre en compte le fait que les mécanismes évolutifs eux-mêmes sont soumis à l'évolution ; et que par conséquent, il peut exister des variations dans ces mécanismes selon les fonctions des protéines ou les phyla des espèces.

Les techniques mathématiques et informatiques permettent désormais de lever la restriction des observations aux seules divergences inférieures ou égales à 15%, divergences auxquelles Margaret Dayhoff était cantonnée à l'époque pour estimer les paramètres de son modèle. Nous allons tirer parti de ces outils pour décrire au mieux l'évolution d'une famille multigénique.

1.3.1 Des matrices de comptages aux matrices de taux observées

Le contexte général étant posé dans la section précédente (Section 1.2, voyons maintenant comment nous l'avons utilisé pour décrire l'évolution d'une famille protéique dont on dispose d'un alignement multiple. Nous allons estimer *pour chaque paire* (x, y) *de séquences alignées* extraite de l'alignement la matrice de probabilités observées P. Cette matrice est estimée par simple comptage selon l'équation suivante :

$$P^{(x,y)} = (\Pi^{(x,y)})^{-1} F^{(x,y)} \tag{1.8}$$

où $\Pi^{(x,y)}$ est la matrice diagonale des fréquences moyennes des acides aminés dans les séquences x et y, et $F^{(x,y)}$ la matrice des fréquences de remplacement des acides aminés observées dans l'alignement de la paire (x, y). L'équation est la même que celle utilisée par Dayhoff (Eq. 1.2) mais le procédé de comptage différent. En effet, nous ne reconstruisons pas les ancêtres communs, et rien ne nous oblige à faire des comptages symétriques (cf Section 1.3.3).

Par construction chaque matrice $P^{(x,y)}$ est une matrice stochastique, dont toutes les lignes se somment à 1 ; c'est la matrice de transition (au sens markovien du terme) qui donne les probabilités de mutation de chaque acide aminé pour la paire (x, y). L'idée de la méthode ORM est de calculer pour chaque paire de séquences (x, y) la matrice L, logarithme matriciel [5] de la matrice P :

$$L^{(x,y)} = logP^{(x,y)} \tag{1.9}$$

Chacune de ces matrices L peut être vue comme un point dans un espace à plusieurs dimensions. Si l'on croit que l'évolution de la famille protéique est gouvernée par une matrice $\mathcal{Q}$ unique, toutes les matrices L s'aligneront sur une droite en vertu de l'équation 1.5. En effet,

$$L^{(x,y)} = logP^{(x,y)} = log(e^{t^{(x,y)}\mathcal{Q}}) = t^{(x,y)}\mathcal{Q} \tag{1.10}$$

Pour mettre en évidence la structure naturelle d'un nuage de points dans un espace de grandes dimensions, l'outil mathématique adéquat est une *décomposition en valeurs singulières* (SVD pour *Singular Value Decomposition*). L'examen de la structure du nuage de points constitué des matrices $L^{(x,y)}$ après SVD permettra de vérifier (ou non) si les points s'alignent (aux fluctuations près) sur une même droite.

Si c'est le cas, toutes les matrices $L^{(x,y)}$ correspondent à des matrices de Markov $P^{(x,y)}$ qui suivent une loi de puissance et qui sont générées à partir d'une même matrice de taux d'évolution Q, que l'on appelle aussi *générateur d'évolution* du processus de Markov. Les matrices $L^{(x,y)}$ sont alors des matrices observées de taux d'évolution. Nous appelons notre méthode *ORM pour Observed Rate Matrices* car les objets centraux sont les matrices $L^{(x,y)}$.

5. Le logarithme matriciel de P se calcule à partir de la série suivante :

$$logP = \sum_{k=1}^{\infty} \frac{(-1)^{k+1}}{k}(P-1)^k$$

Le logarithme d'une matrice n'existe pas toujours. La série converge lorsque P n'est pas trop éloignée de la matrice identité 1. C'est à dire qu'il faut que les termes diagonaux prédominent les termes extra-diagonaux[1] :

$$\sum_i p_{ii} > \sum_i \sum_j p_{ij} \ \forall i \neq j$$

1.3.2 Calcul de la matrice de taux Q d'une famille protéique

Nous appelons les matrices $L^{(x,y)}$ des matrices observées de taux (*pseudo-rate matrices*). Ce ne sont pas toujours des "vraies" matrices de taux. En effet nous avons vu que la matrice Q génératrice d'une famille de matrices de transition de Markov a les propriétés mathématiques suivantes :

i) somme des lignes égale à 0 et

ii) tous les termes extradiagonaux positifs.

Or dans notre cas, la propriété i) est toujours vérifiée mais ii) pas toujours (bien que la somme des termes non diagonaux négatifs soit négligeable). Nous faisons donc un abus de langage en baptisant notre méthode ORM *Observed Rates Matrices* car les matrices $L^{(x,y)}$ observées sont, pour quelques unes, de putatives matrices observées de taux.

La mise en évidence de la relation linéaire entre les objets complexes que sont des matrices (20x20) se résout facilement en considérant chaque matrice $L^{(x,y)}$ comme un point dans un espace à grandes dimensions (190). Il existe de nombreux outils mathématiques pour résoudre cette question. Nous utiliserons la *décomposition en valeurs singulières* également appelée SVD (pour *Singular Value Decomposition*).

L'objet de la SVD est de faire tourner le nuage de points dans l'espace de manière à calculer un système de représentation qui maximise l'étirement du nuage. Ainsi, avec les nouvelles coordonnées, la structure naturelle du nuage de points apparaît de manière plus visible. On peut ainsi voir si toutes les matrices $L^{(x,y)}$ décrivant les paires de séquences de la famille s'alignent (ou non) le long d'une seule et même droite.

La droite le long de laquelle s'organise le nuage des matrices représente le générateur d'évolution (s'il existe) de la famille protéique étudiée (c'est à dire la matrice de taux de remplacements de la famille) et la coordonnée d'une paire (x, y) sur cette droite mesure le temps de divergence $t^{(x,y)}$ séparant les deux séquences (cf Eq. 1.10).

1.3.3 Commentaires sur le modèle

Dans de nombreux cas, les points ne s'organisent pas autour d'une droite mais dessinent plutôt des faisceaux qui partent de l'origine [1]. L'examen attentif des données, à la lumière des groupes taxonomiques ou des sous-fonctions des protéines, permet de dégager une cohérence biologique. Par exemple, dans la figure 1.6 qui correspond à l'analyse des protéines codées par le génome mitochondrial pour 123 métazoaires, on observe que les points décrivant les alignements des chordés (en violet) ont tendance à former un nuage différent de celui décrivant les alignements des arthropodes (en rouge).

Pour calculer le générateur spécifique des chordés ou des arthropodes, on recommence l'analyse sur chacun des groupes. En procédant ainsi, nous avons obtenu la matrice observée des taux de remplacements des protéines mitochondriales pour les vertébrés. Afin de comparer nos résultats avec ceux publiés par Adachi et Hasegawa [23], sur des données de même origine, nous avons généré la matrice de transition correspondant à la même divergence que la leur (PAM21). Les deux matrices, bien qu'obtenues de manière différente, sont similaires (Fig. 1.7).

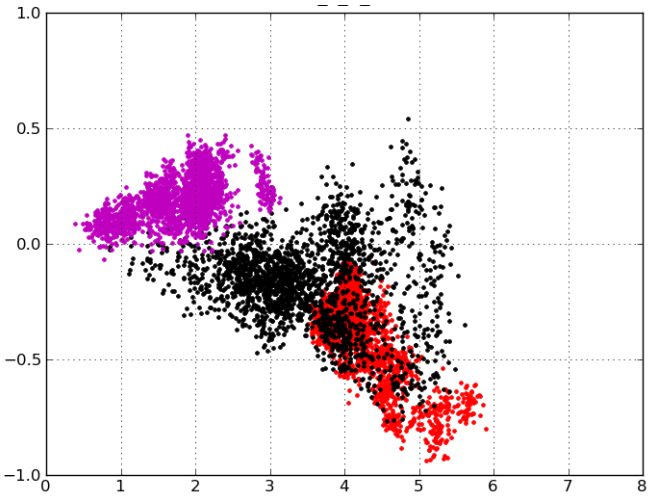


FIGURE 1.6 – *Matrice de taux des protéines mitochondriales estimée à partir d'un alignement multiple constitué de la concaténation des alignements de 12 protéines codées par l'ADNmt pour 123 métazoaires. Plan constitué des projections des matrices $L^{x,y}$ sur les deux premiers axes. En violet sont figurés les points représentant les paires de séquences de chordés, en rouge les paires de séquences d'arthropodes.*

La symétrie des comptages pour l'estimation des matrices P n'est pas une nécessité dans notre méthode. On peut ainsi associer pour toute paire (x, y) deux matrices $P^{x,y}$ et $P^{y,x}$ selon l'équation ci-dessous [6] :

$$P^{(x,y)} = (\Pi^x)^{-1} F^{(x,y)} \quad \text{et} \quad P^{(y,x)} = (\Pi^y)^{-1} F^{(y,x)} \tag{1.11}$$

où Π^x est la matrice diagonale de fréquence des acides aminés dans la séquence x et $F^{(x,y)}$ la matrice des fréquences de remplacements des acides aminés observées pour transformer la séquence x en la séquence y (et réciproquement pour le comptage de $F^{(y,x)}$).

1.3.4 Vecteurs de mutabilités et distances *LogDet*.

La plupart du temps, lorsqu'il s'agit simplement de décrire un alignement multiple selon la méthode ORM on préférera utiliser des objets plus simples et plus facilement interprétables qui sont les *vecteurs de mutabilités* $V^{(x,y)}$. Ce sont les diagonales des matrices $L^{(x,y)}$ qui "vivent" dans un espace à 20 dimensions. L'analyse du nuage de points se fait comme précédemment et permet de mettre en évidence l'individualité des acides aminés. Selon les groupes taxonomiques, certains acides aminés mutent plus facilement relativement aux autres, et ceci au sein d'une même famille protéique.

Par ailleurs, ce sont les angles formés par ces vecteurs qui ont été utilisés comme descripteurs de quartets dans l'article que nous avons écrit, en collaboration avec Jan Weyer-Menkhoff et Stephan Grünewald [24].

La distance *LogDet* que nous définissons entre x et y est égale à la trace de la matrice $L^{(x,y)}$ définie Eq.1.9 ou encore à la norme L1 du vecteur $V^{(x,y)}$. Cette distance n'est pas nécessairement symétrique puisque l'on peut estimer $V^{(x,y)}$ indépendamment de $V^{(y,x)}$. Cette distance évolutive est une variante de la distance *ldet* décrite dans la littérature comme robuste aux variations de la composition des séquences [4, 19**? ?**].

6. et par conséquent deux matrices $L^{x,y}$ et $L^{y,x}$ selon Eq. 1.9.

Vertebrates from this work (PAM 21)

	A	C	D	E	F	G	H	I	K	L	M	N	P	Q	R	S	T	V	W	Y
A	0.81717	0.00134	0.00069	0.00039	0.00429	0.01843	0.00083	0.01625	0.00125	0.01564	0.01362	0.00399	0.00578	0.00058	0.00070	0.04718	0.03896	0.01118	0.00024	0.00149
C	0.01356	0.87138	0.00010	0.00009	0.00640	0.00548	0.00068	0.00490	0.00027	0.01503	0.00967	0.00217	0.00213	0.00223	0.00004	0.02407	0.02084	0.01030	0.00233	0.00833
D	0.00273	0.00004	0.88069	0.02489	0.00101	0.00798	0.00634	0.00217	0.00325	0.00222	0.00048	0.04344	0.00342	0.00062	0.00010	0.00844	0.00430	0.00202	0.00047	0.00539
E	0.00117	0.00003	0.01832	0.91604	0.00043	0.00796	0.00386	0.00078	0.01265	0.00242	0.00080	0.00646	0.00167	0.01231	0.00012	0.00431	0.00627	0.00227	0.00098	0.00116
F	0.00526	0.00074	0.00031	0.00019	0.85324	0.00128	0.00415	0.01438	0.00051	0.05501	0.01347	0.00105	0.00088	0.00117	0.00027	0.00952	0.00775	0.00571	0.00075	0.02436
G	0.02295	0.00068	0.00252	0.00338	0.00131	0.92475	0.00020	0.00182	0.00216	0.00487	0.00184	0.00384	0.00029	0.00041	0.00052	0.01884	0.00292	0.00332	0.00240	0.00098
H	0.00232	0.00018	0.00454	0.00362	0.00945	0.00046	0.86664	0.00216	0.00620	0.00880	0.00181	0.02511	0.00589	0.01970	0.00233	0.01009	0.00789	0.00110	0.00012	0.02157
I	0.01421	0.00039	0.00049	0.00022	0.01015	0.00128	0.00068	0.73210	0.00075	0.08843	0.03413	0.00279	0.00313	0.00065	0.00015	0.00642	0.03794	0.06190	0.00072	0.00348
K	0.00384	0.00008	0.00264	0.01336	0.00130	0.00566	0.00696	0.00273	0.82503	0.00596	0.00865	0.03248	0.01003	0.02254	0.00835	0.01399	0.02684	0.00308	0.00282	0.00366
L	0.00707	0.00065	0.00025	0.00037	0.02047	0.00178	0.00144	0.04620	0.00086	0.82840	0.03881	0.00314	0.00333	0.00177	0.00074	0.00830	0.01763	0.01402	0.00241	0.00236
M	0.01888	0.00123	0.00016	0.00036	0.01513	0.00207	0.00101	0.05375	0.00376	0.11631	0.69070	0.00483	0.00303	0.00288	0.00011	0.01708	0.04139	0.02148	0.00293	0.00291
N	0.00806	0.00038	0.02137	0.00436	0.00168	0.00603	0.01762	0.00626	0.02008	0.01364	0.00688	0.78141	0.01035	0.00881	0.00292	0.04886	0.03225	0.00131	0.00104	0.00667
P	0.00779	0.00029	0.00117	0.00078	0.00098	0.00032	0.00289	0.00499	0.00434	0.01000	0.00311	0.00712	0.91735	0.00556	0.00033	0.01396	0.01597	0.00166	0.00076	0.00063
Q	0.00170	0.00069	0.00046	0.01263	0.00289	0.00099	0.02178	0.00235	0.02161	0.01212	0.00631	0.01361	0.01255	0.85146	0.00578	0.01197	0.01419	0.00117	0.00012	0.00561
R	0.00275	0.00001	0.00011	0.00016	0.00091	0.00164	0.00343	0.00061	0.01075	0.00648	0.00033	0.00590	0.00098	0.00771	0.95221	0.00177	0.00210	0.00013	0.00052	0.00149
S	0.04848	0.00235	0.00216	0.00151	0.00813	0.01565	0.00370	0.00760	0.00456	0.01869	0.01281	0.02563	0.01046	0.00411	0.00048	0.75857	0.06367	0.00566	0.00224	0.00352
T	0.03447	0.00178	0.00095	0.00191	0.00560	0.00214	0.00255	0.03882	0.00757	0.03455	0.02721	0.01462	0.01045	0.00419	0.00048	0.05509	0.73741	0.01596	0.00174	0.00252
V	0.01682	0.00149	0.00080	0.00114	0.00708	0.00407	0.00407	0.10582	0.00151	0.04611	0.02345	0.00098	0.00186	0.00057	0.00005	0.00800	0.02694	0.75088	0.00078	0.00107
W	0.00063	0.00055	0.00030	0.00085	0.00155	0.00485	0.00010	0.00209	0.00232	0.01350	0.00545	0.00138	0.00140	0.00010	0.00033	0.00540	0.00500	0.00129	0.95123	0.00167
Y	0.00340	0.00183	0.00311	0.00089	0.04617	0.00180	0.01778	0.00922	0.00263	0.01200	0.00492	0.00791	0.00109	0.00421	0.00087	0.00792	0.00650	0.00150	0.00156	0.86469
	A	C	D	E	F	G	H	I	K	L	M	N	P	Q	R	S	T	V	W	Y

Vertebrates from Adachi and Hasegawa *J. Mol. Evol. (1996),42,459-468* **(PAM 21)**

	A	C	D	E	F	G	H	I	K	L	M	N	P	Q	R	S	T	V	W	Y
A	0.81453	0.00069	0.00050	0.00047	0.00140	0.01227	0.00093	0.01721	0.00033	0.00969	0.01235	0.00288	0.00555	0.00047	0.00040	0.04306	0.06405	0.01237	0.00009	0.00057
C	0.00801	0.86116	0.00013	0.00009	0.00802	0.00377	0.00787	0.00827	0.00024	0.01249	0.00131	0.00351	0.00205	0.00192	0.00388	0.03553	0.02540	0.00066	0.00200	0.01409
D	0.00180	0.00004	0.88414	0.02591	0.00085	0.00644	0.00694	0.00134	0.00130	0.00067	0.00030	0.04992	0.00111	0.00305	0.00005	0.00927	0.00535	0.00015	0.00042	0.00096
E	0.00161	0.00002	0.02015	0.92637	0.00011	0.00234	0.00266	0.00031	0.01142	0.00031	0.00022	0.00608	0.00102	0.01432	0.00009	0.00745	0.00282	0.00162	0.00002	0.00127
F	0.00167	0.00080	0.00022	0.00004	0.87210	0.00008	0.00229	0.01149	0.00047	0.06064	0.00780	0.00094	0.00178	0.00143	0.00023	0.00793	0.00542	0.00118	0.00047	0.02298
G	0.01571	0.00042	0.00221	0.00106	0.00009	0.95382	0.00010	0.00133	0.00066	0.00075	0.00044	0.00369	0.00019	0.00045	0.00081	0.01511	0.00254	0.00036	0.00042	0.00005
H	0.00236	0.00173	0.00469	0.00223	0.00483	0.00210	0.85697	0.00274	0.00338	0.00408	0.00056	0.02850	0.00439	0.02416	0.00490	0.00846	0.00858	0.00024	0.00044	0.03639
I	0.01435	0.00059	0.00024	0.00008	0.00806	0.00090	0.00085	0.72369	0.00059	0.08175	0.03633	0.00225	0.00167	0.00080	0.00005	0.00646	0.04717	0.07219	0.00014	0.00184
K	0.00104	0.00006	0.00097	0.01180	0.00119	0.00175	0.00407	0.00230	0.87124	0.00351	0.00622	0.03107	0.00465	0.01952	0.00473	0.01223	0.01846	0.00054	0.00159	0.00306
L	0.00416	0.00042	0.00005	0.00004	0.02182	0.00032	0.00071	0.04236	0.00055	0.84097	0.04033	0.00166	0.00400	0.00178	0.00059	0.00956	0.01819	0.00814	0.00158	0.00258
M	0.01687	0.00015	0.00010	0.00010	0.00887	0.00047	0.00030	0.05954	0.00274	0.12762	0.66735	0.00300	0.00226	0.00230	0.00008	0.01473	0.06387	0.02671	0.00119	0.00189
N	0.00538	0.00052	0.02398	0.00370	0.00134	0.00516	0.02043	0.00495	0.01832	0.00710	0.00396	0.79620	0.00785	0.00615	0.00051	0.05534	0.02892	0.00119	0.00066	0.00836
P	0.00746	0.00025	0.00050	0.00050	0.00200	0.00021	0.00230	0.00274	0.00210	0.01239	0.00207	0.00564	0.91636	0.00596	0.00103	0.01869	0.01778	0.00088	0.00024	0.00092
Q	0.00120	0.00043	0.00220	0.01394	0.00328	0.00094	0.02709	0.00268	0.01822	0.01134	0.00480	0.00971	0.01306	0.85659	0.00740	0.00866	0.01415	0.00118	0.00006	0.00272
R	0.00141	0.00136	0.00005	0.00012	0.00068	0.00265	0.00744	0.00022	0.00571	0.00452	0.00019	0.00104	0.00289	0.00992	0.95770	0.00118	0.00071	0.00057	0.00123	0.00020
S	0.04360	0.00319	0.00243	0.00248	0.00679	0.01194	0.00337	0.00786	0.00406	0.02260	0.01091	0.03030	0.01420	0.03032	0.00026	0.75397	0.07287	0.00142	0.00153	0.00319
T	0.05309	0.00182	0.00112	0.00072	0.00376	0.00173	0.00273	0.04684	0.00489	0.03479	0.03866	0.01293	0.01111	0.00400	0.00025	0.05949	0.70412	0.01536	0.00083	0.00175
V	0.02067	0.00009	0.00006	0.00090	0.00168	0.00059	0.00016	0.14376	0.00020	0.03147	0.03241	0.00113	0.00114	0.00070	0.00020	0.00233	0.03074	0.73057	0.00042	0.00060
W	0.00022	0.00040	0.00041	0.00002	0.00114	0.00085	0.00044	0.00042	0.00120	0.00947	0.00236	0.00089	0.00067	0.00005	0.00082	0.00402	0.00250	0.00045	0.97225	0.00141
Y	0.00118	0.00263	0.00060	0.00085	0.04207	0.00012	0.03106	0.00473	0.00206	0.01326	0.00297	0.01006	0.00159	0.00200	0.00011	0.00679	0.00476	0.00067	0.00121	0.87069
	A	C	D	E	F	G	H	I	K	L	M	N	P	Q	R	S	T	V	W	Y

FIGURE 1.7 – *Matrice de transition PAM21 des vertébrés calculée à partir des protéines codées par ADNmt* versus *la matrice publiée par Adachi et Hasegawa [23].*

1.3.5 Conclusion

Le modèle de Dayhoff est un modèle généraliste pour les protéines qui permet de travailler directement sur l'alphabet à 20 lettres des acides aminés. Cette particularité permet d'étudier des familles multigéniques ayant divergé depuis longtemps. L'hypothèse généralement admise lorsqu'on utilise le modèle de Dayhoff est que l'évolution qui gouverne les protéines est unique. En pratique on constate que ce n'est pas vrai. Nous avons développé une méthode qui ne considère que les paires de séquences, ce qui nous permet pour une famille de protéines, ayant suffisamment de membres, d'estimer conjointement :

- le générateur d'évolution (matrice $\mathcal{Q}$) spécifique de cette famille,
- l'âge de divergence des paires de séquences (distance *ldet*)
- et surtout, d'observer les cas où le générateur d'évolution est lui-même soumis à un processus évolutif.

1.4 Classification fondée sur les rangs

En général, les biologistes veulent obtenir un arbre ou des clusters qui résument l'information contenue dans une matrice de dissimilarités. Le problème se complique rapidement car les mesures de dissimilarités sont nombreuses et variées. De plus, certaines telles que le *LogDet* décrit précédemment (Section 1.3.4) ne sont même pas symétriques. Discuter des mérites de tel ou tel indice de similarité est d'ailleurs l'objet de nombreux débats au sein de la communauté.

Nous avons pris le problème dans l'autre sens en nous demandant si l'information essentielle pour établir la topologie de l'arbre n'était pas simplement résumée par les relations d'ordre au sein d'une matrice de dissimilarités. Si c'est le cas, alors le choix de la mesure de dissimilarité perd de son importance puisque n'importe laquelle donne le même résultat sous réserve que les relations d'ordre soient conservées.

L'approche classique pour aborder un tel problème est de passer par les rangs. Cependant les matrices de rangs que nous avons obtenues donnent une classification trop pauvre. La méthode que nous proposons pour enrichir les classifications obtenues par les rangs repose sur trois étapes [2] :

– la concordance des jugements,
– l'itération,
– la détermination des clusters.

1.4.1 La concordance des jugements

Le point de départ est de distinguer les juges et les candidats. Dans une expérience de transcriptome, les juges sont les gènes et les candidats sont les conditions expérimentales. Cette distinction existe même quand les juges et les candidats sont de même nature (divergence entre séquences en phylogénie moléculaire, enfants d'un même village classant leurs camarades par ordre de préférence). Dans les exemples suivants les lignes correspondent aux juges et les colonnes aux candidats.

Le but est d'identifier des sous-groupes (des clusters) de juges qui ont une vision concordante sur l'ensemble des candidats. La concordance des

jugements est calculée à l'aide de la distance de Spearman entre les juges. Puis la matrice des distances de Spearman est elle-même ramenée à une matrice de rangs.

Nous avons une définition des rangs un peu particulière.

Définition 1.1 *Soit X la note d'un juge dans la liste des notes qu'il a attribuées, le rang de X est égal au nombre de valeurs inférieures ou égales à X.*

Par exemple, la liste des notes [0,00 ;6,66 ;6,66 ;8,23] devient la liste des rangs [1 ;3 ;3 ;4]. Nous verrons plus tard que cette définition simplifie considérablement la recherche des clusters et le calcul des distances d'arbres.

1.4.2 L'itération

L'itération est au coeur de notre méthode. La matrice de rangs initiale est transformée en matrice de distances de Spearman qui est à son tour transformée en matrice de rangs. On rappelle que la distance de Spearman entre 2 vecteurs de rangs est définie comme la somme des carrés des écarts des rangs. Par exemple, pour deux vecteurs de rangs $A = [1; 3; 3; 4]$ et $B = [3; 1; 2; 4]$,

$$d(AB) = (1-3)^2 + (3-1)^2 + (3-2)^2 + 0 = 9$$

Le processus est itéré jusqu'à ce que la matrice des rangs ne change plus entre deux itérations. C'est cette itération qui enrichit la hiérarchie de la matrice de rangs initiale. L'itération est illustrée figure 1.8 à partir d'un cas trivial.

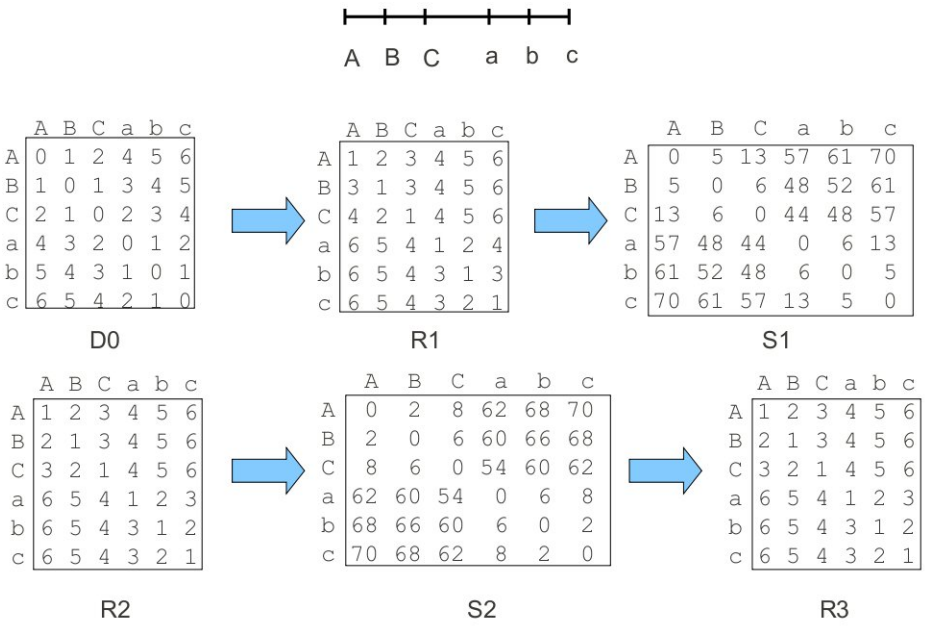


FIGURE 1.8 – *Exemple du procédé d'itération. A partir des 6 points alignés on calcule la matrice de distances initiale D0, puis on commence le procédé. R1, R2, R3 sont les matrices de rangs et S1, S2 les matrices de distances de Spearman. Les matrices R2 et R3 étant identiques le procédé s'arrête.*

Le processus ne converge pas toujours, il arrive qu'il tombe dans un cycle. De notre expérience, les différentes matrices de rangs obtenues au cours des

itérations convergent rapidement (en 4 à 5 itérations) vers une matrice qui présente une “structure” très claire quand les données ont une “histoire à raconter”. Lorsque nous avons obtenu des cycles, par simulation ou sur des données réelles, nous avons observé des cycles courts, entre 3 ou plus souvent 4 matrices de rangs, sans que l'on arrive à l'expliquer. Le procédé d'itération, qui joue un rôle essentiel ici, n'est pas vraiment expliqué mathématiquement. En ceci, il n'est pas différent de que l'on voit dans beaucoup de systèmes dynamiques [2].

1.4.3 Détermination des clusters et distance d'arbre

Un cluster C est formé par l'ensemble des objets qui ont une dissimilarité plus faible entre eux qu'avec n'importe quel objet n'appartenant pas à C. Dans notre méthode, du fait de notre définition de rang (Définition 1.1), le rang maximal d'un cluster C est la taille de C. Cette propriété simplifie considérablement la recherche des clusters. Il suffit en effet pour trouver tous les clusters de taille N (groupes de N juges ayant la même vision sur les candidats) de parcourir les lignes (qui correspondent aux juges) de la matrice de rangs finale et de rechercher toutes les notes données dont le rang est inférieur à N [2].

Notre définition des rangs facilite également le calcul de la distance entre deux objets (deux juges) pour la construction de l'arbre. Cette distance est égale au rang maximal entre les deux objets dont on soustrait 1 (dans les cas réels que nous avons traités, les juges et les candidats étaient identiques, et chaque juge se met en première position de son classement). Cette distance est une *ultramétrique* [2], c'est à dire que les deux distances les plus grandes d'un triplet sont égales entre elles, ce qui définit une hiérarchie que l'on peut représenter par un dendogramme.

Pour revenir à notre exemple illustratif, la distance ultramétrique obtenue à partir de la matrice de rangs obtenue à la convergence du processus itératif est donné figure 1.9. Elle est équivalente au dendogramme de gauche.

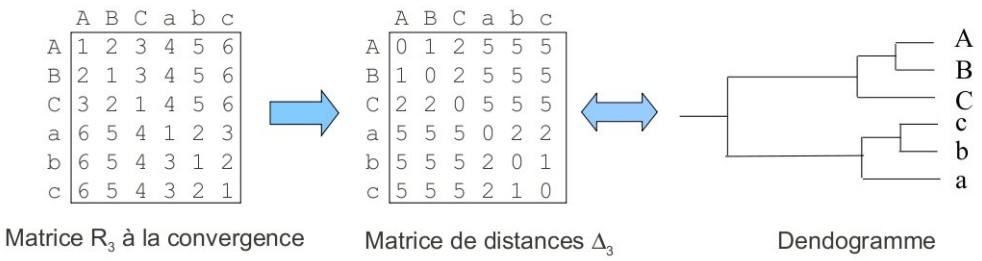


FIGURE 1.9 – *Distance ultramétrique* Δ_3 *calculée à partir de la matrice de rangs* R_3 *après convergence et dendogramme.*

1.4.4 Applications à des données réelles

Lorsque nous avons testé la méthode sur des données réelles (distances évolutives *ldet* calculées par ORM [1] entre des protéines codées par l'ADN mitochondrial) il s'est avéré que seuls les grands groupes taxonomiques étaient retrouvés (mammifères / non mammifères). Par contre à l'intérieur

de chaque groupe les sous-embranchements étaient erronés. Nous avons donc appliqué la stratégie de *diviser pour régner* pour recalculer le sous-embranchement de chaque noeud. Cette fois, les divisions en sous-arbres sont correctes et la topologie des dendogrammes sur les protéines mitochondriales est correcte. C'est à dire comparable à celles obtenues avec des méthodes telles que SplitsTree [25] ou BioNJ [26] en calculant les distances, soit avec notre distance *LogDet* [1], soit par maximum de vraisemblance avec le modèle Dayhoff-JTT [17] du programme ProtDist du progiciel Phylip [27].

Conclusion. Ce travail nous a permis de créer une méthode efficace et rapide qui conduit à une classification stable sans avoir à fixer de paramètres plus ou moins explicites. Pour mémoire, une simple classification par k-means demande de choisir la mesure de dissimilarité, le critère de classification, le nombre de classes, les conditions d'arrêt et la partition de départ.

Classification sans alignement

2

Sommaire

Avant-propos

Les travaux inclus dans cette deuxième partie ont été réalisés en collaboration avec **Gilles Didier** (qui effectuait sa thèse dans le Laboratoire de Mathématiques Discrètes de Luminy pendant que j'y effectuais mon "post-doctorat"), **Maude Pupin** (qui réalisait sa thèse au Laboratoire Génome et Informatique à Versailles), **Alex Grossmann** et **Ivan Laprevotte** du Laboratoire Génome et Informatique à Versailles.

La contribution d'**Eduardo Corel** (en post-doctorat au Laboratoire Statistiques et Génome à Evry de 2007 à 2009) a été majeure dans le développement de la méthode MS4, à laquelle il faut rajouter celle de **Florian Pitschi** (étudiant en thèse, encadré par Andréas Dress, à Shanghai) pour l'implémentation de l'application à la détection de points d'ancrage dans les alignements multiples.

Sans oublier le travail de **Mark Hoebecke**, puis celui de **Gilles Grasseau**, pour l'implémentation et la mise à disposition des applications Web et des versions distribuables, des prototypes décrits dans cette partie.

Les applications à l'étude des ADN topoisomérases ont été faites en collaboration avec de l'équipe de **Marc Nadal** à l'Institut de Génétique et Microbiologie à Orsay, dont **Marc Nadal** lui-même, **Hélène Debat**, **Ramzi El Feghali** (ATER en 2005-2007) et **Anna Bizard** (doctorante).

Articles publiés : Les travaux décrits dans cette partie ont donné lieu à trois articles (joints en annexes A.4, A.5, A.6) :

- description de la classification reposant sur la méthode NLD [3]
- généralisation à la méthode MS4 pour la classification [4]
- utilisation de MS4 pour la détection de points d'ancrage [5]

Demandes de financement ANR réalisées dans le cadre de ces travaux :

Campagne 2008, Campagne 2009, *ANR programme blanc : MASC - Multiple Analysis of Sequences by Combinatorial method* - dont j'étais le porteur principal avec comme partenaires : l'équipe de Marc Nadal (IGM, Orsay), l'équipe de Gilles Didier (IML, Marseille), l'équipe de Burkhard Morgenstern (université de Göttingen). Financement refusé.

Campagne 2011, *ANR programme blanc : ReGyMe - Study of the Reverse Gyrases Mechanisms* - dont Marc Nadal était le porteur principal du projet et dont j'étais responsable pour le partenaire 3. Financement refusé.

Encadrement de stagiaires : Pour ces recherches, j'ai été amenée à encadrer ou co-encadrer de nombreux stagiaires aussi bien sur les développements méthodologiques que sur leurs applications en biologie : Imen ELMEKKI (6 mois en 2006), Fanny GERARDIN (6 mois en 2007), Nabila KACED (3 mois en 2008), Malek GHANDOUR (3 mois en 2009), Loïc LAUREOTE (6 mois en 2009), Imen KHAROUBI (6 mois en 2009), Amyra ALLIOUAT (3 mois en 2010), Cyril DENBY WILKES (6 mois en 2010), Florence VOGLIOLO (3 mois en 2011).

2.1 Décodage N-local des séquences biologiques

A peu près toutes les méthodes pour la comparaison de séquences biologiques font des hypothèses implicites (et nécessaires) sur les types de changements qu'ont subis les séquences au cours de l'évolution. Dans ces hypothèses, des changements tels que les permutations et les inversions sont souvent ignorés, tandis que d'autres tels que les insertions et les délétions sont souvent mal évalués car ils ne suivent pas un modèle d'évolution régulier. Le résultat est que les régions qui contiennent des changements de ce type sont souvent considérées comme ambigües et éliminées des alignements avant de passer à la construction des relations évolutives. Finalement, il faut souvent recourir à un alignement manuel par un expert, en particulier lorsque les séquences ont connu une succession de duplications/délétions et permutations.

On évite ces problèmes en travaillant avec des méthodes qui comparent les séquences sans les aligner. Une façon intuitive pour le faire est de comparer leur composition, c'est à dire la fréquence des nucléotides ou des acides aminés. Deux séquences de composition très différente ne peuvent pas être apparentées. En revanche, il est tout à fait possible que des séquences non apparentées aient la même composition. Une solution pour éviter cet écueil est d'augmenter la taille de l'alphabet, par exemple en découpant les séquences en mots de N lettres consécutives (des N-mots). La méthode des N-mots apporte effectivement des informations sur les distances évolutives, mais elle manque de finesse si la mesure de dissimilitude est tout ou rien : les deux N-mots sont identiques ou différents. Ceci conduit à introduire la notion de mots *approximativement identiques*, mais il s'ensuit une explosion combinatoire car il y a beaucoup trop de façons d'être à peu près identiques.

Gilles Didier a trouvé une solution pour augmenter la taille de l'alphabet en évitant l'explosion combinatoire : la méthode du décodage local d'ordre N, encore appelée méthode du *décodage N-local* (NLD pour *N-local decoding* en anglais) [28]. Le point de départ est de considérer qu'une séquence est une suite d'états, au sens des chaînes de Markov cachées. Par exemple, une séquence d'ADN possède quatre états (A,T,G ou C). Elle peut-être ramenée à deux états : purine (A,G) ou pyrimidine (C,T) que l'on code par (Y,R) ; (Y,R) étant une projection des quatre nucléotides.

La méthode introduite par Gilles Didier consiste à rechercher l'alphabet maximal dont la séquence d'ADN découpée en N-mots est la projection. On calcule donc tous les états présents dans une séquence (ou un ensemble de séquences) en la découpant en N-mots par pas de 1, et on recherche l'alphabet de taille maximale donnant la même écriture en N-mots (*i.e.* étant une projection de cette N-écriture). L'algorithme est linéaire en temps et en espace mémoire avec la longueur de la séquence et ne dépend pas de N [29].

2.1.1 Illustration du principe de la méthode

Nous allons illustrer sur un cas concret le mécanisme de construction de l'alphabet maximal aboutissant à la même N-écriture des séquences que l'alphabet initial (ici à 4 lettres). Un exemple de séquences recodées est

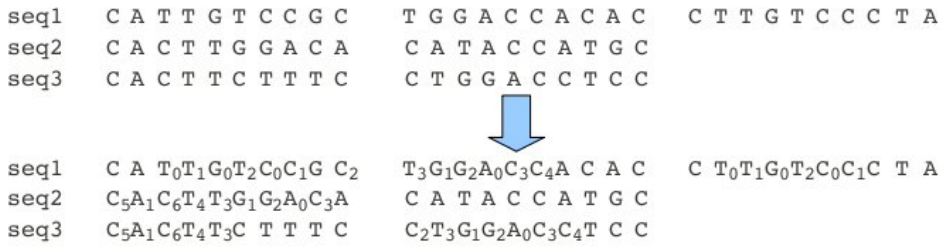


FIGURE 2.1 – *Exemple de décodage N-local obtenu sur trois séquences jouets pour N=5 (avant et après recodage). Les lettres qui ont le même indice dans le nouvel alphabet correspondent à des zones similaires. Les lettres qui n'ont pas été indicées sont des caractères uniques dans le nouvel alphabet. Extrait Fig.6 [3].*

donné figure 2.1. Le nouvel alphabet est de taille 47 : soit 17 caractères indicés,

$$\mathcal{A} = \{A_0, A_1, T_0, T_1, T_2, T_3, T_4, G_0, G_1, G_2, C_0, C_1, C_2, C_3, C_4, C_5, C_6\}$$

auxquels il faut rajouter autant de caractères non indicés qui correspondent à des caractères uniques, soit 30. Bien que cet exemple soit petit, on voit que les zones indicées (constituées de caractères présents au moins deux fois dans les séquences) correspondent à des mots similaires.

Soit un ensemble de séquences non alignées, l'algorithme va dans un premier temps examiner, par pas de 1, tous les segments chevauchants de taille $2N-1$ (que nous appellerons **voisinages** dans la suite de l'exposé). Puis il mettra dans la même classe, tous les voisinages similaires ; chaque classe sera identifiée par un nouveau symbole.

Qu'est ce que deux voisinages similaires selon le NLD ? Toute l'originalité de la méthode réside en ce point important sur lequel nous allons prendre le temps de nous arrêter.

Définition 2.1 *Deux voisinages (de taille* $2N-1$ *) sont similaires (*i.e. *appartiennent à la même classe d'équivalence) si et seulement si :*

- *ils possèdent* **à la même position** *un même mot de taille N, on parlera de* **liaison directe** *entre les voisinages,*
- *ou s'ils sont reliés par une chaîne de liaisons directes par clôture transitive ("les amis de mes amis sont mes amis"). On parlera de* **liaison indirecte**.

Pour revenir à l'exemple didactique, regardons les environnements de la classe T_3 reliés par liaison directe (Table 2.1). Chaque environnement est représenté par un couple (s, p) qui désigne le site central où s est le numéro de la séquence et p sa position dans la séquence.

(1,11)	CCGCTGGAC	(2,5)	CACTTGGAC	(1,11)	CCGCTGGAC
(2,5)	CACTTGGAC	(3,5)	CACTTCTTT	(3,12)	TTCCTGGAC
(3,12)	TTCCTGGAC				

TABLE 2.1 – *Exemple de voisinages reliés par liaison directe (classe* T_3 *Table 2.1). La similarité des voisinages en colonne repose sur le(s) mot(s) de 5 lettres indiqué(s) en bleu.*

La figure 2.2 représente les liaisons directes (traits pleins) et les liaisons indirectes (traits pointillés) reliant les environnements de la classe T_3 de notre exemple. La classe d'équivalence ainsi formée est désignée par une lettre représentant l'état du résidu central dans l'alphabet initial (ici T) suivi d'un indice, tout à fait arbitraire, correspondant au numéro de construction de la classe par l'algorithme (ici 3).

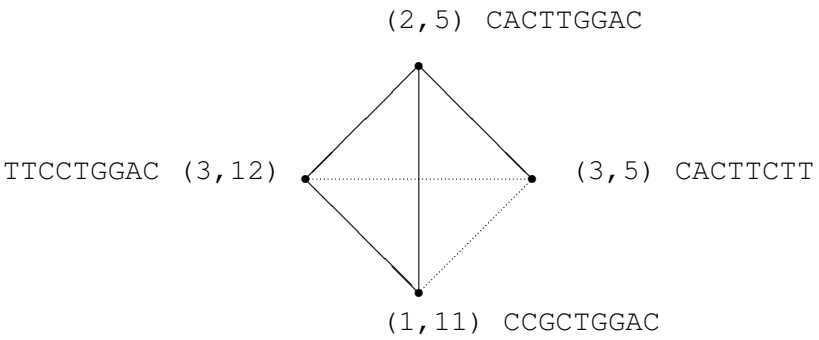


FIGURE 2.2 – *Les environnements de la classe T_3 (Table 2.1) reliés par liaison directe sont représentés en trait plein, ceux reliés par liaison indirecte en trait pointillé.*

Classification sans alignement Maintenant que les séquences ont été réécrites dans un alphabet plus grand, il suffit de comparer les compositions des séquences pour les classer, sans avoir à les aligner.

Une classe d'équivalence selon le décodage N-local correspond à un ensemble de voisinages similaires, sans que l'on ait besoin de préciser le nombre ou le type d'erreurs autorisées. Ceci grâce au critère ressemblance retenu et à la clôture transitive (*les amis de mes amis sont mes amis*). La table 2.2 montre un exemple de classe de voisinages obtenue sur un jeu de 23 séquences protéiques de gyrase inverse.

On voit clairement que les similarités prises en compte par l'algorithme du décodage N-local ont une signification biologique. Les remplacements intégrés correspondent aux propriétés physico-chimiques des chaines latérales des acides aminés sans que l'on ait jamais eu besoin de spécifier une matrice de score. En effet, l'algorithme repose uniquement sur un parcours de l'arbre des suffixes des mots de taille N présents dans les séquences, ce qui le rend particulièrement efficace (pour les détails sur l'algorithme voir [29]).

Nous utilisons la composition en caractères de l'alphabet maximal calculé par le NLD pour décrire et comparer deux séquences (sans qu'il soit nécessaire de les aligner) en utilisant la distance suivante entre deux séquences i et j :

Définition 2.2

$$d_{ij} = 1 - \frac{n_{ij}}{l}$$

l est la longueur de la séquence la plus courte,

$$n_{ij} = \sum\nolimits_c \min\{n_i(c), n_j(c)\}$$

c désigne une classe (un caractère indicé) de l'alphabet étendu, et n_i le nombre d'occurrences de c dans le séquence i (idem pour n_j).

PkoRG	89	SFSIIAPTGMGKS
PfuRG	89	SFSIIAPTGMGKS
ApeR1	101	SFSIIAPTGVGKT
PabRG	89	SFSIIAPTGMGKS
AaeR2	100	SFAIVAPTGVGKT
PhoRG	89	SFSIIAPTGMGKS
TpeRG	100	SFVILAPTGVGKT
TthR1	73	SFAMLAPTGIGKT
PaeRG	88	SFAIVAPTGSGKT
AaeR1	79	SFAMLAPTGVGKT
NeqRG	83	SFSIIAPTGMGKT
ApeR2	96	SFAIIAPTGVGKS
TmaRG	85	SFTMVAPTGVGKT
SacR1	92	SFAIIAPPGLGKT
SsoR1	93	SFAIIAPPGLGKT
StoR2	89	SFAIIAPPGLGKT
SacR2	85	SFSLSAPTGVGKT
StoR1	85	SFSLSAPTGLGKT
MjaRG	83	SFSIVVPTGVGKS
SsoR2	92	SFTMSAPTGLGKT
TteRG	84	SFTLVAPTGVGKT
AfuRG	61	SFAATAPTGVGKT
MkaRG	122	SFSILAPTGTGKT

TABLE 2.2 – *Exemple d'une classe d'équivalence calculée par le décodage N-local sur des gyrases inverse pour $N = 7$. La classe P_4 appartient au motif de fixation de l'ATP du domaine hélicase. Le numéro est la position de la proline centrale dans la séquence dont le nom est indiqué à gauche. Abréviations : Ape : Aeropyrum pernix, Pko : Pyrococcus kodakaraensis, Neq : Nanoarcheum equitans, Tpe : Thermophilus pendens, Pae : Pyrobaculum aerophilum, Aae : Aquifex aeolicus, Pab : Pyrococcus abysii, Pfu : Pyrococcus furiosus, Sto : Sulfolobus tokodaii, Sso : Sulfolobus solfataricus, Mja : Methanococcus jannaschii, Sac : Sulfolobus acidocaldarius, Tth : Thermus thermophilus, Pho : Pyrococcus horikoshii.*

2.1.2 Application au sous-typage des HIV/SIV

Nous avons utilisé cette méthode pour mesurer la matrice de dissimilarités d'un ensemble de 59 séquences de HIV (Human Immunodeficiency Virus) et 7 séquences de SIV (Simian Immunodeficiency Virus) [3]. Nous avons obtenu des résultats identiques à ceux décrits dans la banque de référence (http ://www.hiv.lanl.gov/content/index), aussi bien en analysant le génome complet qu'en procédant gène par gène et ceci sans recourir à aucun alignement, ni automatique, ni manuel. Le fait remarquable est que nous avons obtenu une phylogénie correcte en ne prenant en compte que la région non codante du LTR (Long Terminal Repeat) alors que l'alignement automatique de cette région est impossible à cause des nombreuses duplications/insertions/délétions (cf Figure 2 dans [3]).

Sur le plan théorique ce travail nous a amené à nous poser deux questions auxquelles je me suis intéressée en priorité ces dernières années.

1. Peut-on déterminer automatiquement la valeur optimale de N pour un jeu de séquences donné et pour une zone donnée ?

2. Peut-on utiliser le décodage N-local pour déterminer des points d'ancrage pour guider les programmes d'alignements multiples ?

2.2 Détection de ressemblances locales multi-échelles

Comme toutes les méthodes à base de N-mots, le paramètre délicat à fixer dans le programme NLD est la taille du mot N. Or ce paramètre est crucial dans la méthode : si N est trop petit la taille de l'alphabet final est petite, et sont considérés comme similaires des mots qui n'ont plus rien à voir les uns avec les autres, *a contrario* si N est trop grand la taille de l'alphabet final est très grande car chaque mot est unique.

Grâce à la rapidité de l'algorithme, nous calculons toutes les valeurs de N pour une gamme de valeurs (de $Nmin$ à $Nmax$ [1]). Plus N augmente, plus la taille de l'alphabet augmente car deux mots similaires à l'ordre N peuvent devenir différents à l'ordre $N+1$. Pour chaque N-écriture selon le NLD (pour une valeur donnée de N), on a une partition des *sites* [2] des séquences. Ces partitions s'emboîtent les unes dans les autres au fur et à mesure que N augmente.

N'oublions pas, comme nous l'avons vu précédemment, que les classes d'équivalence qui constituent la partition représentent les environnements de taille $2N-1$ similaires selon la définition 2.1. Ici nous désignons par site le résidu central du voisinage, selon le protocole décrit dans la section 2.1.1 une classe d'équivalence contient donc plusieurs sites (qui ont été regroupés parce qu'ils ont des environnements similaires). Selon la définition 2.1 cet environnement s'étend sur les $N-1$ résidus qui précèdent et qui succèdent le site en question. Un exemple de l'emboîtement des classes lorsque N augmente est donné figure 2.3. La classe $K10\_9$, obtenue en analysant un jeu de 23 gyrases inverse, se scinde en deux quand N passe de 10 à 11. Cette classe détecte un motif fonctionnel important dans la famille des gyrases inverses. De plus, quand N=11, la classe de scinde en deux classes qui discriminent les deux copies de gyrases inverses chez les crénarchées ayant subi une duplication du gène *topR* (voir la discussion dans la section 2.4.3).

Nous représentons l'ensemble des partitions (obtenues par le NLD) dans un arbre que nous allons élaguer de manière à ne retenir que les noeuds qui vont représenter des similarités locales "pertinentes" [4].

Qu'est-ce qu'une similarité locale pertinente et quels sont les critères retenus pour élaguer l'arbre des partitions ?

Chaque noeud de cet arbre représente un ensemble de mots qui sont des projections de N-mots de l'alphabet initial. Ces N-mots sont présents une ou plusieurs fois sur chacune des séquences. Le critère κ que nous avons fixé

1. Dans la version en ligne du programme les valeurs par défaut sont Nmin = 2 et Nmax = taille du mot répété le plus long dans les séquences. En pratique cette valeur peut être grande si le jeu de données contient des séquences quasi-identiques.

2. On appelle site d'une séquence le couple (s,p) où s désigne le numéro de la séquence et p la position du résidu dans cette séquence. Notons que l'algorithme recode tous les sites et que ce n'est que lors du post-traitement que les indices des singletons sont enlevés.

FIGURE 2.3 – *Exemple d'emboîtement de classes. Lorsque N augmente de 9 à 11, $K10\_9$ se scinde en $K23\_11$ et $K24\_11$ (à cause de la différence pour le résidu avant K). Exemple extrait du jeu des 23 gyrases inverses, voir abréviations Table 2.2.*

est égal au rapport du nombre de N-mots dans le noeud sur le nombre de séquences dans lesquelles ils ont été retrouvés. Nous élaguons l'arbre en ne conservant que les noeuds qui ont une valeur supérieure à la valeur κ fixée par l'utilisateur.

La valeur de κ est fixée par défaut à $\kappa = 1$ ce qui correspond à aucune répétition de N-mots dans une même séquence. Nous obtenons ainsi des similarités locales de taille variable selon le niveau auquel l'arbre a été

coupé. La méthode a été publiée sous le nom de MS4 (Multi-Scale Selector of Sequence Signature) [4] et l'article est donné en annexe. Le programme est en ligne sur le site du laboratoire : http ://stat.genopole.cnrs.fr/ms4/.

2.2.1 Illustration du principe de la méthode

Nous allons encore une fois illustrer la méthode sur un exemple didactique à partir de 3 séquences jouets :

```
Séquence 1      ATCGTCT
Séquence 2      ATCTCCC
Séquence 3      GTCATCGAA
```

L'algorithme procède à tous les décodages NLD pour $N = 1$ à 5 (Fig. 2.4).

N-local decoding (N=1)	N-local decoding (N=2)	N-local decoding (N=3)	N-local decoding (N=4)	N-local decoding (N=5)
> Sequence 1	> Sequence 1	> Sequence 1	> Sequence 1	> Sequence 1
A0_1 T0_1 C0_1 G0_1	A0_2 T0_2 C0_2 G0_2	A0_3 T0_3 C0_3 G0_3	A0_4 T0_4 C0_4 G0_4	A0_5 T0_5 C0_5 G0_5
T0_1 C0_1 T0_1	T0_2 C0_2 T0_2	T0_3 C0_3 T1_3	T1_4 C1_4 T2_4	T1_5 C1_5 T2_5
> Sequence 2	> Sequence 2	> Sequence 2	> Sequence 2	> Sequence 2
A0_1 T0_1 C0_1 T0_1	A0_2 T0_2 C0_2 T0_2	A0_3 T0_3 C0_3 T1_3	A1_4 T3_4 C2_4 T4_4	A1_5 T3_5 C2_5 T4_5
C0_1 C0_1 C0_1	C0_2 C0_2 C0_2	C1_3 C2_3 C3_3	C3_4 C4_4 C5_4	C3_5 C4_5 C5_5
> Sequence 3	> Sequence 3	> Sequence 3	> Sequence 3	> Sequence 3
G0_1 T0_1 C0_1 A0_1 T0_1	G0_2 T0_2 C0_2 A0_2 T0_2	G0_3 T0_3 C0_3 A0_3 T0_3	G1_4 T5_4 C6_4 A0_4 T0_4	G1_5 T5_5 C6_5 A2_5 T6_5
C0_1 G0_1 A0_1 A0_1	C0_2 G0_2 A1_2 A2_2	C0_3 G0_3 A1_3 A2_3	C0_4 G0_4 A2_4 A3_4	C7_5 G2_5 A3_5 A4_5

FIGURE 2.4 – *Exemple didactique : les décodages N-locaux de $N = 1$ à $N = 5$ sont indiqués dans une colonne pour les 3 séquences jouets. Les classes d'équivalence des sites sont désignées selon la procédé suivant : A0_1 est la classe A0 pour $N = 1$. Fig.3 [4].*

Arbre des partitions. L'emboîtement des classes lorsque N augmente est représenté dans l'arbre des partitions (Fig. 2.5). Chaque niveau de l'arbre correspond à la partition des sites pour une valeur de N.

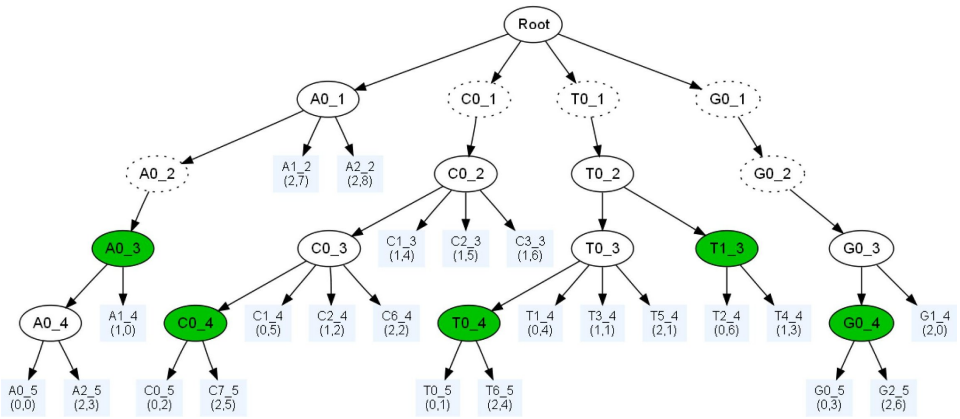


FIGURE 2.5 – *Exemple didactique : arbre des partitions. Chaque noeud est une classe de sites. Une arête relie une classe et les classes dans lesquelles elles se divisent quand N croît. Les feuilles (indiquées en bleu) sont des singletons à un seul site (s,p) où s est le numéro de la séquence et p la position dans la séquence. Les noeuds verts correspondent aux classes sélectionnées par élagage de l'arbre. Les noeuds en pointillés n'existent pas dans l'arbre compacté implémenté. Fig.3 [4].*

Si on prend le noeud $C0\_2$ de la figure 2.5 qui correspond à une classe issue du NLD à l'ordre 2 (colonne 2 de Fig. 2.4), elle se scinde en 4 classes à l'ordre 3 (C_03 et 3 singletons). Lequel $C0\_3$ se divise à son tour en $C0\_4$ et 3 singletons à l'ordre 4.

Sélection des classes multi-échelles. Les classes MS4 sont sélectionnées par élagage de l'arbre des partitions selon un paramètre κ qui est *le seul paramètre* de la méthode. *Ce paramètre mesure la quantité de répétitions tolérées au sein d'une séquence.* Il permet d'adapter la taille de l'environnement pertinent pour chaque site et se calcule de la manière suivante :

Tout noeud de l'arbre des partitions est caractérisé par deux valeurs,

- sa *taille*, c'est à dire le nombre de sites appartenant à la classe (dans notre exemple précédent 8 pour $C0\_2$, 5 pour $C0\_3$, 2 pour $C0\_4$) ;
- son *étendue*, c'est à dire le nombre de séquences dans laquelle elle apparaît (3 pour $C0\_2$, 3 pour $C0\_3$ et 2 pour $C0\_4$; cf Fig. 2.4).

Le paramètre κ est défini comme étant le rapport *taille* sur *étendue* pour ce noeud. Un noeud sera sélectionné, lorsque κ atteint la valeur fixée pour la première fois en parcourant l'arbre dans le sens descendant.

$$\kappa = \frac{taille}{étendue} \tag{2.1}$$

Par défaut κ est assigné à 1, ce qui veut dire qu'aucune répétition n'est tolérée dans aucune séquence.

Dans l'exemple Fig. 2.5, sur la branche de l'arbre qui va de la racine à $C0\_4$, le noeud sélectionné sera $C0\_4$ car c'est le premier noeud sur ce parcours qui ne contient pas deux occurrences de la classe au sein d'une même séquence. En effet, $C0\_3$ apparaît 2 fois dans la troisième séquence (site $(2, 5)$ et site $(2, 2)$) et quand $N = 5$ on n'a plus que des singletons (cf Fig. 2.4).

2.2.2 Application au sous-typage des HIV/SIV

Nous avons utilisé la méthode MS4 pour l'ensemble de 59 séquences de HIV (Human Immunodeficiency Virus) et 7 séquences de SIV (Simian Immunodeficiency Virus) étudiés avec le NLD dans [3].

Alors que le NLD avait nécessité l'examen fastidieux des arbres obtenus pour toutes les valeurs de N de 5 à 60, MS4 calcule un seul arbre pour l'ensemble de ces décodages N-locaux. En sélectionnant la taille de l'environnement le plus pertinent pour un ensemble de sites, c'est à dire la valeur de N pour laquelle on a l'ensemble de sites le plus grand qui vérifie la condition sur κ. MS4 nous permet d'obtenir une mesure de ressemblance multi-échelle que nous utilisons pour faire de la classification sans alignement selon la définition 2.2.

Nous avons obtenu des résultats identiques à ceux décrits dans [3], aussi bien en analysant le génome complet qu'en procédant gène par gène, y compris dans le cas difficile de la région non codante des LTR (*Long Terminal Repeat*) soumise à de nombreux événements de duplications/insertions/délétions (cf Fig.6 dans [4]). Nous avons également étudié

l'impact du paramètre κ sur les classifications obtenues et nous avons montré que dans le cas des HIV/SIV, les classifications obtenues sont très robustes, même pour les séquences LTR (cf Additional Files 6 et 7 de [4]).

2.3 Détection automatique de points d'ancrage

La méthode MS4 s'avère très utile pour guider les alignements multiples manuels, car la mise en couleur des lettres identiques dans le nouvel alphabet (qui correspondent à des similarités locales multi-échelles centrées autour de ces résidus) fait ressortir les portions de l'alignement conservées [4]. C'est d'ailleurs cette propriété, déjà présente dans le NLD [30], qui nous a poussé à l'utiliser pour faire de la classification sans alignement en nous fondant sur l'identité de composition de ce nouvel alphabet.

Les sorties de MS4 ne peuvent pas être utilisées telles quelles car la méthode ne fait aucun tri sur les points d'ancrage (les caractères du nouvel alphabet) alors que l'alignement multiple fait l'hypothèse que ceux-ci sont dans le même ordre dans toutes les séquences (ou absents). En d'autres termes, une condition nécessaire pour qu'un symbole MS4 soit un point d'ancrage, est qu'il n'apparaisse jamais plus d'une fois dans une séquence et ceci est vrai pour toutes les séquences. Grâce à ce critère [3], il est possible de définir des points d'ancrage valables pour une partie au moins des séquences. Ils définissent des colonnes partielles dans la représentation habituelle des alignements multiples.

Mais ce critère n'est pas suffisant car il n'exclut pas les permutations dans l'ordre des points d'ancrage. Nous cherchons donc dans un second temps les colonnes partielles qui sont *cohérentes* avec un alignement multiple global. Un tel sous-ensemble forme un graphe orienté acyclique. La recherche d'un tel graphe correspond au classique *minimum feedback arc-set problem*. Nous avons développé une solution heuristique car c'est un problème NP-dur [5].

Afin de tester nos ancres, déterminées de manière entièrement automatique et sans paramètres, nous avons analysé les résultats obtenus avec les quelques programmes qui permettent d'introduire des points d'ancrage définis par l'utilisateur : DIALIGN 2, ClustalW 2.0 avec l'option BALLAST et T-Coffee. Nous avons utilisé le benchmark BAliBASE 3[5].

L'analyse des alignements multiples montre que l'utilisation des deux types de colonnes partielles (avant et après retrait des colonnes incohérentes) améliore les performances pour ClustalW 2.0 sur BAliBASE. En revanche, nous obtenons une dégradation constante de la performance pour DIALIGN, et presque pas de variations pour T-Coffee [5].

2.4 Etude des topoisomérases IA

2.4.1 Contexte

Dès la description de la double hélice d'ADN en 1953, Watson et Crick ont soulevé une question biologique d'importance liée à cette structure : "Puisque les deux brins d'ADN sont entrelacés, il est essentiel qu'ils puissent

3. c'est-à-dire $\kappa = 1$ selon la définition donnée section 2.2

se dérouler s'ils doivent se séparer...". Ce problème topologique est réglé *in vivo* par des enzymes qui s'appellent les ADN topoisomérases. Leur rôle est de relaxer les supertours positifs ou négatifs qui s'accumulent en aval et en amont des machineries cellulaires qui progressent le long de l'ADN en séparant les deux brins.

Pour modifier la topologie de l'ADN, les topoisomérases induisent une cassure transitoire de l'ADN : cassure simple brin pour les topoisomérases de type I, cassure double brin pour les topoisomérases de type II. Toutes les ADN topoisomérases identifiées à ce jour appartiennent à quatre superfamilles qui se différencient principalement par leur mécanisme réactionnel : les topoisomérases IA et IB, les topoisomérases IIA et IIB. La distribution phylogénétique des topoisomérases connues montre que seules les *topoisomérases IA* sont présentes dans les trois royaumes du vivant [31]. Leur conservation ubiquitaire suggère un rôle crucial des topoisomérases IA dans la maintenance de la stabilité des génomes.

Dans ce contexte, les organismes hyperthermophiles sont particulièrement intéressants car avec des températures optimales de croissance supérieures à 80 °C, ils doivent faire face à des dommages spontanés de leur ADN (déamination, dépurination, coupure simple brin ou double brin). En dépit de ces conditions de vie extrêmes, ces organismes n'ont pas de taux de mutation plus élevés que les organismes mésophiles. Les études de génomique comparative montrent que le seul gène qui soit spécifique des organismes thermophiles (systématiquement présents chez eux et systématiquement absents chez les mésophiles) est celui d'une topoisomérase de type IA particulière qui s'appelle *gyrase inverse* (ou RG pour *reverse gyrase*) [32].

La gyrase inverse est une protéine constituée de deux domaines fonctionnels : un domaine topoisomérase IA et un domaine hélicase [32]. Il est à noter que cette association topoisomérase IA / hélicase (de la superfamille SF2) est ubiquitaire et est retrouvée sous forme d'interaction entre ces deux protéines chez les mésophiles. Son rôle est crucial dans la maintenance de la stabilité des génomes et un dysfonctionnement de cette interaction entraîne, chez l'homme, des maladies graves telles que le syndrome de Werner ou le syndrome de Bloom [32].

Les gyrases inverses ont un rôle particulier, elles introduisent des supertours positifs dans la molécule d'ADN en consommant de l'ATP, contrairement aux autres topoisomérases IA qui relâchent exclusivement des supertours négatifs sans consommation d'énergie. Outre ce rôle topologique, impliqué dans la protection de l'ADN contre les effets de la température, il a été montré que la gyrase inverse a un rôle dans la réparation de l'ADN [33].

Par ailleurs, le gène *topR* des gyrases inverses est dupliqué chez les *Crenarcheota* (sauf chez les *Thermoproteales*) et il a été mis en évidence que les deux copies *topR1* et *topR2* sont régulées différemment chez *Sulfolobus solfataricus* [34]. Il est tentant de penser que ces deux copies correspondent à des différences fonctionnelles. Marc Nadal et son équipe ont démontré récemment chez *S.solfataricus* que les deux gyrases inverses ont des activités topoisomérases différentes [35].

Dans ce contexte nous nous intéressons plus particulièrement aux questions suivantes :

– Peut-on reconstruire l'histoire évolutive des topoisomérases IA et retrouver la trace des duplications ancestrales de gènes ainsi que les transferts horizontaux ?
– Peut-on trouver des motifs impliqués dans la sous-fonctionnalisation des gyrases inverses chez les crénarchées ? Plus précisément, peut-on prédire des résidus qui discriminent les deux copies afin de guider les expérimentations ?

2.4.2 Classification et évolution

Préparation des données. Afin d'étudier l'évolution des topoisomérases, nous avons besoin de connaître toutes les copies qui existent au sein d'un organisme. Nous avons donc examinés seulement les organismes pour lesquels le génome complet est séquencé. Nous avons sélectionné *133 organismes (17 eucaryotes, 43 archées, 73 bactéries)*. Pour être le plus exhaustif possible, nous avons pris tous les eucaryotes correctement annotés début 2008, toutes les archées séquencées à cette même date, et nous avons sous-échantillonné les bactéries en privilégiant les bactéries extrêmophiles et les bactéries modèles, en veillant également à échantillonner au moins un individu de tous les grands taxa.

Nous avons recherché en utilisant le logiciel LASSAP [36] et le logiciel BLAST, les topoisomérases autres que celles annotées dans les génomes d'intérêt afin de trouver toutes les copies existantes dans un génome. Nous avons ainsi trouvé une autre gyrase inverse chez *Sulfolobus acidocaldarius* (ces copies sont reportées dans les annotations plus récentes du génome). Nous obtenons un total de *217 topoisomérases* réparties de la manière suivante :

	thermophiles	**non thermophiles**
bactéries	14 topoisomérases IA + 6 gyrases inverses	89 topoisomérases IA + 3 gyrases inverses*
archées	25 topoisomérases IA + 29 gyrases inverses	21 topoisomérases IA
eucaryotes		30 topoisomérases IA

FIGURE 2.6 – *Jeu de topoisomérases de type IA analysées. * les gyrases inverses de bactéries mésophiles ont été identifiées dans les génomes de* Nautilia profundicola*,* Nitratiraptor*,* Caminibacter mediatlanticus*. Ces bactéries sont en fait capables de croître à haute température [37].*

Classification avec MS4. Nous avons montré, en utilisant MS4 couplée à SplitsTree (option NeighborNet) [25], que les topoisomérases IA se divisent en cinq grands groupes (Fig. 2.7). Nous avons adopté ici la désignation des groupes proposée par Forterre [31].

- les gyrases inverses
- les TopoI bactériennes
- les TopoIII bactériennes
- les TopoIII eucaryotes
- les TopoIII archées

Notre étude exhaustive, sur un grand nombre de séquences, de manière entièrement automatique, sans paramètres et sans alignements, conforte l'organisation en cinq classes précédemment décrites [31, 38]. Cependant, il apparaît que le groupe des TopoIII archées n'est pas très solide, lorsqu'on analyse beaucoup de séquences (comme ici). Sans connaissances biologiques complémentaires, il est difficile de le considérer en tant que groupe d'après le seul réseau phylogénétique car une partie des séquences à tendance à se regrouper avec les TopoIII eucaryotes (Fig. 2.7).

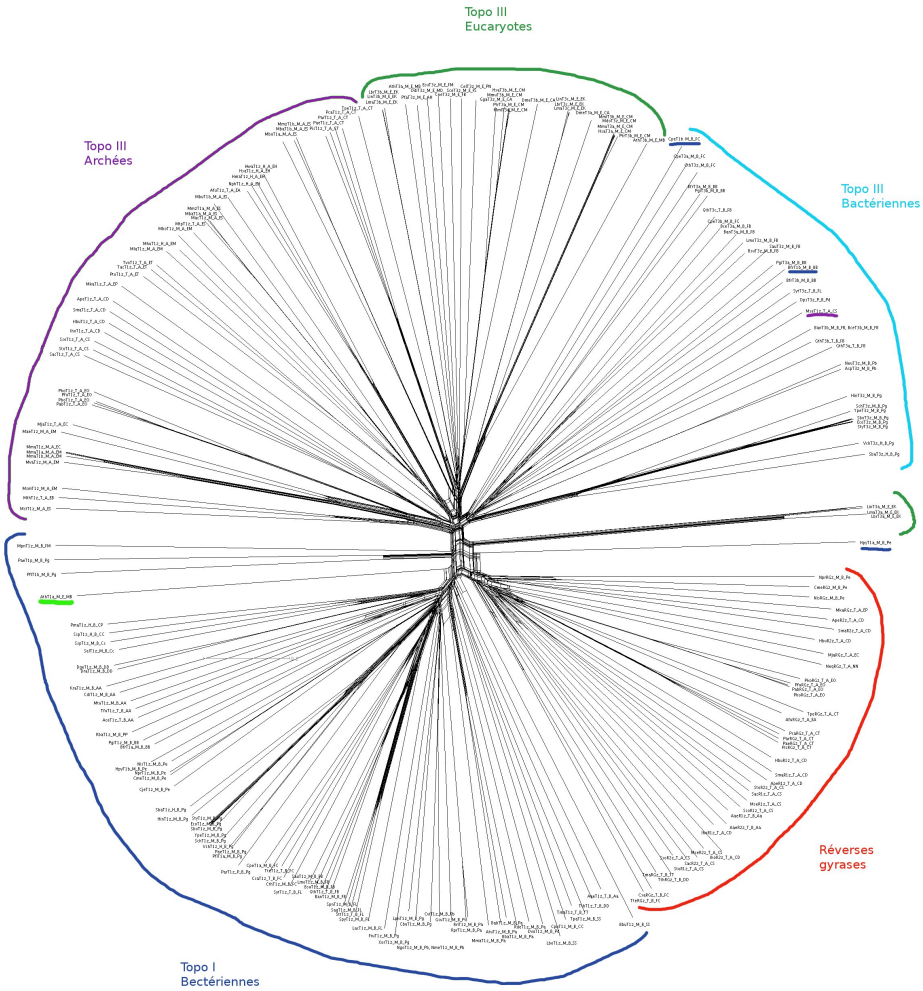


FIGURE 2.7 – *Classification selon MS4 suivie de l'option NeighborNet de SplitsTree [25] pour les 217 topoisomérases issues de 133 organismes complètement séquencés.*

MS4, couplée à SplitsTree, permet également de dégager des séquences *inclassables* dont le branchement se positionne à la base de l'arbre, ce qui veut dire qu'elles ne se rattachent à aucun des cinq grands groupes. En général cette information est corroborée par les connaissances biologiques comme par exemple pour les *Leishmania* (indiquées en vert Fig. 2.7). Il s'agit d'organismes parasites ayant trois copies de topoisomérases de type IA dont deux des copies se classent correctement aves les TopoIII eucaryotes (α et β). Il n'est pas interdit de penser que cette troisième copie est en train de diverger vers une néo-fonctionnalisation ou une pseudogénéisation.

2.4.3 Etude de la duplication des gyrases inverses

Classification "expliquée" par les motifs. Jusqu'à présent nous n'avons utilisé les similitudes locales détectées par MS4 que de manière globale pour faire de la classification, en résumant l'information à un seul chiffre pour une comparaison (le nombre de caractères identiques). Or nous disposons d'une information beaucoup plus riche, comment la résumer en utilisant l'information locale intrinsèque pour classer ?

Nous voulons coupler la détection des similarités locales avec la classification elle-même en utilisant le même outil de base, MS4, pour construire simultanément les groupes de séquences et les motifs les caractérisant. Ceci afin d'*expliquer* les groupes obtenus en termes de motifs caractéristiques ou motifs signatures.

Nous comptons adopter une approche hiérarchique descendante pour classer les séquences en fonction des classes MS4 les plus pertinentes, comme nous l'avons fait dans le passé pour classer selon les rangs (Section 1.4), pour construire les classes de séquences en fonction des motifs.

Le problème sera alors d'ordonner les classes MS4 qui sont très nombreuses de manière à utiliser les plus importantes pour bâtir la classification. Par exemple pour le jeu de 35 gyrases inverses présenté dans cette section MS4 a trouvé 5671 classes.

Nous projetons de les trier en donnant la même importance à *l'étendue* et à *l'ordre* N du décodage (cf Section 2.2.1). Par exemple, la surface de la classe MS4 : *étendue* * $2N-1$), simple à mettre en oeuvre, nous semble être un critère de tri à explorer dans un premier temps.

Les motifs pertinents seraient des candidats pour des interactions internes au sein de la protéine. Nous avons d'ores et déjà repéré des groupes de motifs qui semblent intéressants au vu de leurs positions dans la structure 3D qui devront être confirmés puis testés expérimentalement.

Etude des motifs fonctionnels de TopR1 et TopR2. Nous comptons utiliser cette nouvelle méthode de classification pour comprendre le fonctionnement des deux copies de gyrases inverses. L'équipe de Marc Nadal a montré que la copie TopR1, qui correspond au prototype des gyrases inverses, possède une activité de relaxation de l'ADN ATP indépendante en plus de son activité de surenroulement positif ATP-dépendante, activité que ne possède pas la copie TopR2. De plus, alors que TopR1 est distributive, ne générant que de l'ADN faiblement surenroulé positivement, TopR2 apparaît comme hautement processive, capable d'introduire un grand nombre de supertours positifs [35].

D'un point de vue bioinformatique le domaine topoisomérase apparaît plus conservé que le domaine hélicase (on rappelle que ce domaine n'existe pas dans les topoisomérases de type IA autres que les gyrases inverses). Il est donc tentant de rechercher dans cette partie hélicase des motifs capables d'expliquer une différence entre TopR1 et TopR2.

Une approche préliminaire a consisté à prendre les motifs les plus conservés et à rechercher des positions discriminantes entre les séquences TopR1 et TopR2. Sur le petit nombre de gyrases inverses présentes en 2 copies (8 paires de gènes dupliqués / 35 gyrases inverses au total) nous avons réussi

à trouver quelques positions discriminantes. Un exemple de classes MS4 correspondant à ces sites sont indiquées figure 2.8.

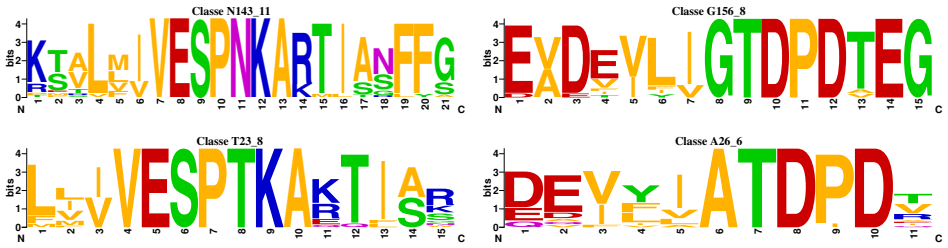


FIGURE 2.8 – *Exemples de classes MS4 qui discriminent les gyrases inverses TopR1 et TopR2. En haut, les classes discriminant les R1 (à gauche N*143_11*, à droite G*158_8*). En bas, les classes discriminant les R2 (à gauche T*23_8*, à droite A*26_6*). N*143_11 *contient 19 séquences dont les 8 TopR1, T*23_8 *contient 16 séquences dont les 8 TopR2 ; les positions discriminantes sont le N en position* 11 *et le T en position* 8*. La classe A*26_6 *contient 17 séquences dont les 8 TopR1 et G*156_8 *contient 15 séquences dont les 8 TopR2 ; les positions discriminantes sont le G en position* 8 *et le A en position 6. Les logo représentent les variations observées dans chaque classe MS4.*

L'idée du projet ANR soumis en 2011 est d'utiliser l'approche bioinformatique MS4 pour concevoir des gyrases inverses modifiées. Afin de tester les motifs impliqués dans la fonction TopR1 ou TopR2, l'équipe de Marc Nadal va construire chez *Sulfolobus solfataricus* des protéines chimériques où les motifs prédits seront échangés entre TopR1 et TopR2 afin de tester l'activité de ces nouvelles protéines : activité TopR1 ou activité TopR2 ?

Evolution des réseaux dupliqués

3

Sommaire

Avant-propos

Cette partie de mon rapport est née des séminaires “connaissons-nous”, réunions de travail du laboratoire, organisées par Etienne Birmelé, après qu'une partie de l'équipe de Jean-Loup Risler (ex-Laboratoire Génome et Informatique) composée majoritairement de biologistes ait rejoint le Laboratoire Statistique et Génome composé majoritairement de mathématiciens. Chacun a patiemment expliqué à l'autre l'essence de ce qui faisait son travail quotidien. De ces réunions est né le projet *Duplinet - Etude du devenir des réseaux biologiques après duplication* - porté par plusieurs personnes de l'équipe “réseaux biologiques et génomique évolutive” du Laboratoire Statistique et Génome :

- **Carène Rizzon** bioinformaticienne (étude des gènes dupliqués et des éléments transposables),
- **Yolande Diaz** bioinformaticienne (étude des familles multigéniques)
- **Julien Chiquet** statisticien (modélisation des réseaux biologiques),
- **Etienne Birmelé** mathématicien (théorie des graphes),
- **Catherine Matias** statisticienne (modélisation de l'évolution),
- **Justin Whalley**, doctorant sur ce projet (co-encadré par B.Prum et C.Rizzon).

L'objectif du projet *Duplinet* est d'étudier le devenir des gènes et des réseaux dans lesquels ils sont impliqués après duplication, sachant qu'il est généralement admis que la duplication de gènes est un réservoir de variabilité génétique pour l'acquisition de nouvelles fonctions ou la spécialisation en sous-fonctions (revues [39, 40, 41]).

Au cours de ma carrière, je me suis intéressée à l'étude des familles multigéniques : les amino-acides ARNt ligases [42, 43], puis plus récemment les ADN topoisomérases de type IA. Ces deux familles de protéines sont présentes dans les trois règnes du vivant et nécessitent des méthodes d'études adaptées aux séquences très divergentes. De plus, comme nous l'avons vu précédemment, la famille des ADN topoisomérases IA a subi plusieurs événements de duplication, et la sous-famille des gyrases inverses présente un cas documenté d'évolution des deux copies du gène vers une sous-fonctionnalisation [34, 35].

Mon expérience dans l'étude des familles multigéniques, couplée à celle que j'ai acquise en analyse de données, sera utile pour l'approche intégrative que nous comptons utiliser dans le projet *Duplinet*. Cette partie de mon mémoire portera donc sur ma contribution à ce projet de recherche.

3.1 Contexte

Une des surprises apportées par les projets de séquençage des génomes eucaryotes est la présence de nombreux gènes dupliqués aussi bien chez la levure (revue [41]) ou la paramécie [44], que chez les vertébrés (revue [45]) ou les plantes (revue [46]). Les mécanismes de duplications de gènes sont de diverses natures, on distingue ceux qui génèrent des duplications totales des génomes (polyploïdisation), ceux qui aboutissent à des duplications de portions de chromosomes (translocations, inversions, duplications au cours de crossing-over aberrants lors de la méïose), ceux qui ne concernent qu'un petit nombre de gènes (matériel génétique acquis par transposition ou par glissement lors de la réplication à proximité de zones répétées). La liste des mécanismes n'est pas exhaustive et il est important de noter qu'ils vont laisser des traces différentes dans les génomes :

- des duplications à large échelle : les *WGD* pour *Whole Genome Duplication*,
- des duplications à moyenne échelle : les *SD* pour *Segmental Duplication*,
- des duplications à petite échelle parmi lesquelles on distingue les duplications de gènes *en tandem* : les *TAG* pour *Tandemly Arrayed Genes*[1].

3.1.1 Duplication complète de génome : WGD

Depuis la description d'un ancêtre polyploïde chez la levure [47, 48], de nombreuses paléoploïdies ont été rapportées non seulement chez les plantes [49, 50, 51, 52], où la polyploïdie est fréquemment observée et amplement utilisée en agronomie, mais également de manière plus inattendue chez les vertébrés [53, 54, 55] et plus récemment chez la paramécie [44]. La figure 3.1 résume les polyploïdies qui ont été décrites dans la phylogénie des eucaryotes.

Chez les animaux : deux événements de paléopolyploïdie auraient eu lieu chez l'ancêtre des vertébrés (1R et 2R), suivi d'un troisième chez l'ancêtre des poissons (3R) [55]. Un lien causal entre la duplication 2R et l'émergence des vertébrés a été proposé (revue [40]). De plus, les WGD semblent avoir été suivies d'un accroissement de la complexité morphologique des espèces. Par exemple chez les vertébrés, les duplications 1R et 2R ont précédé la complexification du système nerveux, du système endocrinien et du système circulatoire ainsi que l'apparition du crâne, des vertèbres et des dents (revue [40]).

Chez les plantes à fleurs, plusieurs événements de paléoploïdie ont été décrits (Fig. 3.1) aussi bien parmi les monocotylédones (riz, maïs, blé) que parmi les dicotylédones (arabette, coton, peuplier, légumineuses). Bien qu'il soit difficile d'établir l'importance de la duplication complète de génomes chez les plantes à fleurs, on estime entre 50% et 70% la proportion des espèces ayant connu un ou plusieurs épisodes de polyploïdie au cours de leur histoire [40].

1. TAG : gènes dupliqués organisés en groupes de proximité le long du chromosome.

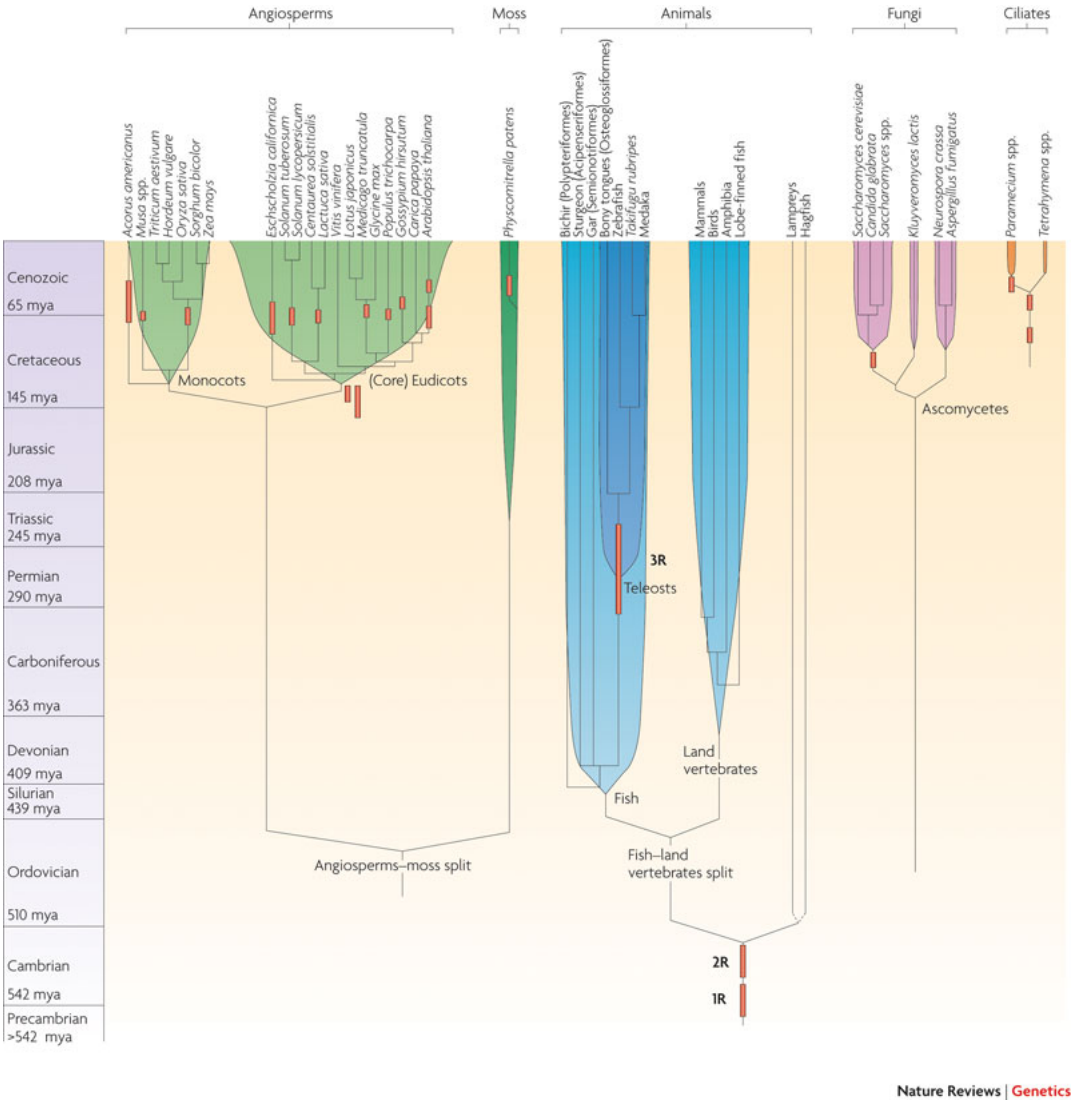


FIGURE 3.1 – *Les événements de paléopolyploïdie sont indiqués par des rectangles rouges sur les branches où ils ont eu lieu chez les plantes à fleurs, les mousses, les animaux, les champignons et les ciliés [40].*

Lors du séquençage de l'arabette, les auteurs ont observé que les chromosomes séquencés présentent de larges zones dupliquées qui vont par paires [49], et que ces zones couvrent plus de la moitié du génome, ce qui est en faveur de rares événements de polyploïdies plutôt que de multiples remaniements chromosomiques.

3.1.2 Duplications segmentales : SD

Une duplication segmentale désigne une duplication d'une partie de chromosome isolée. Bien que les mécanismes qui génèrent de telles duplications ne soient pas complètement élucidés, on peut citer les crossing over inégaux et les rétrotranspositions (revue [45]).

D'un point de vue bioinformatique la définition d'une duplication segmentale n'est pas très claire : ce sont des blocs de synténie [2] indépendants

2. synténie : groupe de gènes homologues conservés dans le "même" ordre dans deux génomes ou dans deux régions d'un même génome.

suffisamment longs. La notion d'indépendance fait allusion au fait que cette duplication ne doit pas être due à un événement global de polyploïdie. La notion de suffisamment long est pour les différencier des duplications à petite échelle touchant juste quelques gènes, en particulier les duplications en tandem.

La difficulté de la détection bioinformatique des SD vient du fait qu'il est délicat de les différencier des duplications complètes anciennes dont on n'arrive pas à retracer l'histoire du fait des remaniements chromosomiques qui suivent les évènements de polyploïdisation [54]. La plupart des études ne distinguent d'ailleurs pas les WGD des SD du fait de cette difficulté [39, 45, 52, 56].

Signalons également que dans le cas d'espèces proches on peut utiliser des techniques expérimentales pour visualiser les duplications segmentales (par exemple la peinture de chromosome [3]). Ces données sont précieuses pour calibrer les méthodes bioinformatiques de reconstruction de génomes ancestraux [57, 58].

3.1.3 Duplications en tandem : TAG

La part estimée de gènes dupliqués se situe ainsi entre [19%-45%] et entre [40%-67%] des gènes respectivement chez le riz et l'arabette [59]. Les WGD n'expliquent qu'en partie ces quantités de gènes dupliqués. Deux autres types principaux de duplications sont à prendre en compte : les duplications segmentales (ou SD, pour Segmental Duplication) qui concernent de grandes portions de chromosomes et les duplications locales encore appelées duplications de gènes en tandem dans lesquelles les gènes dupliqués sont organisés en clusters de gènes voisins sur le chromosome (TAG pour *Tandemly Arrayed Genes*).

Les TAG peuvent être formés de gènes contigus ou séparés de quelques gènes non homologues selon la définition bioinformatique que l'on prend (classiquement jusqu'à dix [59, 60]). Les gènes dupliqués en tandem peuvent correspondre jusqu'à environ un tiers des gènes dupliqués des génomes eucaryotes [59, 61].

Les méthodes bioinformatiques de détection des TAGs dépendent de plusieurs paramètres : la notion de voisinage génomique plus ou moins contrainte et la méthode de construction des familles de gènes paralogues. Certains auteurs par exemple incluent les pseudogènes dans leur définition de famille de gènes mais ils restreignent alors la notion de voisinage à une proximité immédiate [62]. De même des variations peuvent être obtenues selon les algorithmes de classification utilisés : méthode du lien simple, marche aléatoire type *Markov Cluster Algorithm* [63] pour reprendre les plus fréquemment utilisées. Pour une discussion sur la détection des TAGs se reporter à [62].

3.1.4 Modèle de rétention des gènes dupliqués

Un des mystères lié à ces paléoploïdies est le mécanisme de diploïdisation, c'est à dire le retour à l'état diploïde après un doublement complet

3. l'ADN d'une espèce est purifié, marqué par fluorescence, découpé en fragments, puis hybridé *in situ* sur les chromosomes de la seconde espèce.

du génome (revue [54]). Ces duplications complètes de génomes sont suivies d'une perte massive des gènes dupliqués soit par pseudogénéisation, soit par remaniements chromosomiques sans que les mécanismes moléculaires soient clairement identifiés.

Qu'est ce qui fait qu'un gène dupliqué est retenu ou pas au cours de l'évolution ?

La théorie actuelle considère quatre devenirs possibles pour les gènes dupliqués [64, 65] (quelques uns des scénarios sont représentés Fig. 3.2) :

- **la non-fonctionnalisation :** une des deux copies perd son activité (on parle de *knockout*).
- **la sous-fonctionnalisation :** les deux copies assurent une partie de la fonction ancestrale, soit en se partageant la fonction en deux dans le cas de protéines multifonctionnelles, soit en assurant la fonction dans deux tissus différents ou à deux stades cellulaires différents.
- **la néo-fonctionnalisation :** une des copies garde la fonction ancestrale tandis que l'autre acquiert une nouvelle fonction. Dans ce cas on a un *taux d'évolution asymétrique* une des copies accumulant plus de mutations pour acquérir une nouvelle fonction tandis que l'autre doit garder la fonction ancestrale.
- **la redondance fonctionnelle :** les deux copies continuent chacune d'assurer la fonction ancestrale. C'est *l'effet dosage* c'est à dire que l'organisme a besoin d'augmenter la quantité du produit du gène.

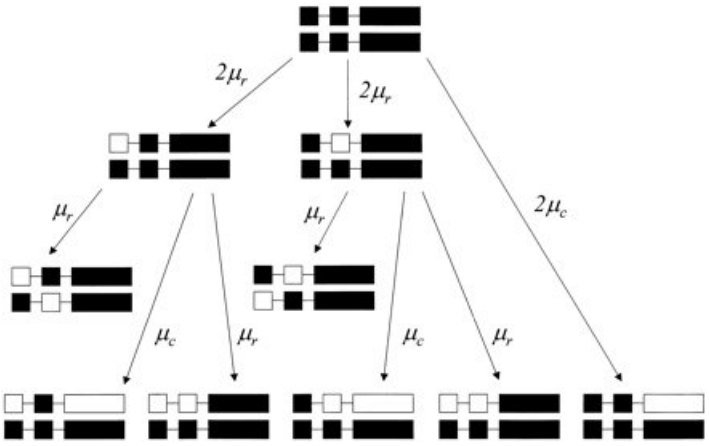


FIGURE 3.2 – *Devenir des gènes dupliqués selon le modèle de sous-fonctionnalisation. Ligne 1, le gène est dupliqué. Ligne 2, perte d'une sous-fonction d'une des copies par mutation de la région régulatrice. Ligne 3, cas de conservation des 2 copies par sous-fonctionnalisation (chacune assurant une sous-fonction de la fonction ancestrale). Ligne 4, perte d'une copie du gène par* knockout *de la région codante ou par perte des régions régulatrices.* μ_c *est le taux de* knockout *et* μ_r *le taux de sous-fonctionnalisation. [65].*

3.1.5 Evolution des gènes dupliqués chez les plantes

La mise à disposition à la communauté scientifique de données génomiques de plus en plus nombreuses offre la possibilité de vérifier les prédictions des modèles théoriques. Ainsi on a pu constater que le devenir de la

grande majorité des gènes dupliqués est la perte par délétion ou pseudogénéisation. Chez l'arabette, environ 25% des gènes dupliqués par WGD et SD perdurent dans le génome [51]. Environ 60% du génome du riz provient apparemment de duplications issues d'événements de paléopolyploïdie [66], et sur cette partie jusqu'à 50% des gènes dupliqués ont été conservés [67]. On estime que la demi-vie des gènes dupliqués chez l'arabette est de 3,2 Ma [64].

Cependant le processus de disparition de gènes n'est pas aléatoire et on a montré plusieurs biais (revues [40, 46, 52, 68]) :

Biais fonctionnels : Les gènes conservés en multiples copies après WGD ou SD sont surreprésentés pour les fonctions liées aux signaux de transduction et de transcription et pour les fonctions impliquées dans le développement [52, 56], tandis qu'ils sont sous-représentés dans les fonctions de réparation de l'ADN [52]. De plus, chez l'arabette et le riz, il a été montré que les gènes dupliqués en tandem présentent un biais fonctionnel par rapport aux autres gènes dupliqués. Ils sont surreprésentés pour les gènes codant des protéines membranaires et des fonctions impliquées dans les stress abiotiques et biotiques, tandis qu'ils sont sous-représentés dans les gènes impliqués dans la transcription et les fonctions de liaison à l'ADN ou l'ARN [59].

Biais transcriptionnels : La majorité des gènes dupliqués après WGD chez les eudicots et les monocots divergent dans leur profil d'expression (73% pour l'arabette et 88% pour le riz) ([68, 69]). De plus, Hanada *et al.* ont montré que les gènes dupliqués en tandem chez l'arabette sont surreprésentés parmi les gènes différentiellement exprimés lors des réponses aux stress [60]. Chez la levure, il semblerait que les gènes dupliqués divergent plus souvent dans leur régulation que dans leur fonction biochimique [39, 70].

Biais des taux d'évolution : Chez la levure, il a été montré que les gènes dupliqués sont biaisés en faveur des gènes évoluant lentement [71]. Chez le xénope, les gènes dupliqués évoluant lentement ont plus tendance à la sous-fonctionnalisation (*i.e.* être différentiellement exprimés dans au moins deux tissus [72]).

Biais phénotypiques : Hanada *et al.* [73] ont étudié les différences dans l'expression des gènes et dans les taux d'évolution pour 492 paires de gènes dupliqués pour lesquelles ils disposaient de mutants *knockout* (Fig. 3.3). Ils ont montré que les paires de gènes, induisant un phénotype lorsqu'une des deux copies est inactivée, ont des différences d'expression et de taux de mutation plus élevés que les paires sans phénotype associé [73].

Biais topologiques dans les réseaux d'interactions : Chez la levure, il a été montré que dans les réseaux PPI [4], les gènes dupliqués codant pour des protéines ayant beaucoup d'interactions avec d'autres protéines évoluent plus lentement que celles ayant moins d'interactants [74]. Parallèlement, une étude récente chez l'arabette portant sur 1882 paires de gènes paralogues

4. PPI : Protein Protein interaction Network

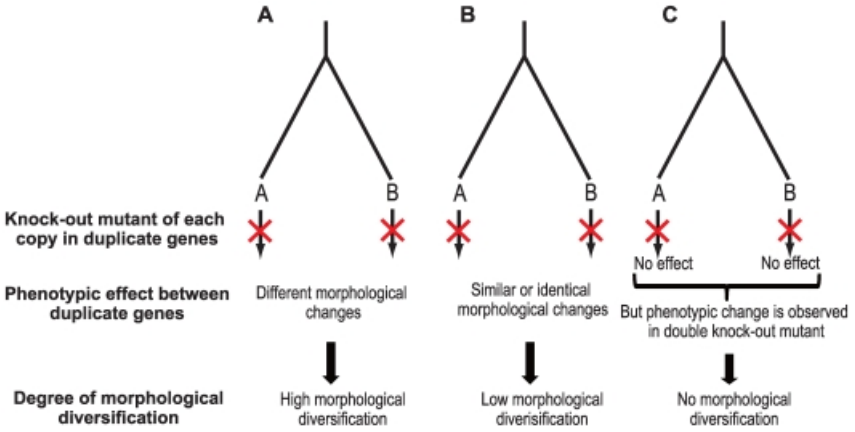


FIGURE 3.3 – *Phénotypes des gènes paralogues étudiés par Hanada* et al. *A) Paires de gènes paralogues montrant des différences phénotypiques importantes entre les deux mutants* knockout *(neo-fonctionnalisation). B) Paires de gènes paralogues montrant peu de différences phénotypiques (sous-fonctionnalisation). C) Paires de gènes paralogues nécessitant un mutant double* knockout *pour révéler une différence phénotypique (redondance fonctionnelle) [73].*

pour lesquelles on dispose de données interactomes, a révélé que les paires de gènes paralogues provenant de WGD partagent plus de protéines partenaires que les autres paires de gènes paralogues [75]. La figure 3.4 illustre les scénarios d'évolution après duplication.

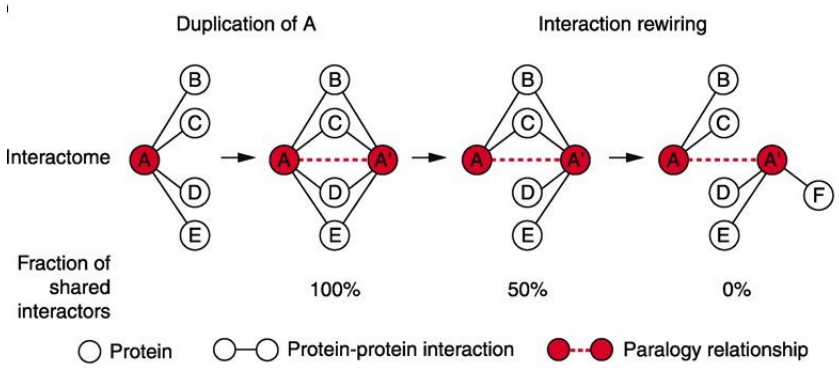


FIGURE 3.4 – *Analyse des réseaux PPI impliquant des paires de gènes paralogues. Proportion de protéines partenaires partagées par les paires de gènes paralogues. Extrait Fig.4 [75].*

3.2 Objectif du projet duplinet

En dépit des nombreuses données disponibles et des nombreuses études réalisées, il reste encore beaucoup à comprendre pour élucider les mécanismes évolutifs sous-jacents au maintien des paires de gènes dupliqués dans les génomes, en particulier chez les plantes. En effet, *chez les plantes (comme chez les levures) la polyploïdisation est une force évolutive dominante* qui a lieu bien plus souvent que chez les animaux, ce qui laisse à penser que ces organismes ont une plus grande capacité à s'adapter aux duplications complètes de génome [68].

L'objectif du projet duplinet est de participer à l'effort collectif de la compréhension des mécanismes d'acquisition de nouvelles fonctions ou sous-fonctions chez les plantes. Cela se fera en adoptant une approche intégrative pour analyser les différents niveaux d'organisation du vivant, de la cellule à l'organisme [5] à partir des données à large échelle disponibles : séquences génomiques, données transcriptomiques, données métaboliques, données d'interactomes protéiques et génétiques. Il est désormais admis que la duplication est un réservoir de diversité génétique et nous venons de voir que différents mécanismes de duplication, locaux ou globaux, participent à la création du pool de gènes dupliqués dans les génomes. Nous voulons nous attaquer à cette question importante de la *duplicabilité* [6] des gènes en traitant séparément les différents types de duplication (WGD, SD, TAG).

3.3 Déroulement du projet

Dans un premier temps nous allons nous focaliser sur l'arabette pour laquelle on dispose des génomes complets pour deux espèces (*A.thaliana* publié en 2000 [49] et *A.lyrata* publié en 2011 [76]). Nous aurons ainsi une vue complète du paysage des gènes dupliqués dans chaque génome et nous pourrons mener des études comparatives pour les deux espèces.

3.3.1 Ce qu'apporte l'analyse des séquences

L'objectif de notre étude est de comprendre l'évolution des gènes dupliqués en séparant les différents mécanismes de duplication. Or il est difficile de retrouver leurs traces dans les génomes lorsque les duplications sont anciennes car les signaux sont brouillés par les remaniements chromosomiques ultérieurs et les pertes de gènes associées à la diploïdisation (revue [54]). La première étape du travail sera donc de reconstruire l'information sur la nature de la duplication (WGD, SD ou TAG) pour les gènes dupliqués chez l'arabette : cette étape passe par la reconstruction des paires de paralogues.

Construction des familles de gènes paralogues La recherche des gènes paralogues débute par la comparaison d'un génome contre lui-même. La stratégie la plus largement utilisée consiste à comparer le protéome de

5. Cette question est également souvent abordée à l'échelle des populations. Ce sont d'ailleurs des généticiens des populations qui ont construit la plupart des modèles évolutifs sur ce sujet.

6. Duplicabilité : Propension pour un gène à être maintenu à l'état dupliqué au cours de l'évolution.

l'arabette contre lui-même à l'aide du logiciel *BLAST* qui compare 2 à 2 toutes les protéines prédites d'*A.thaliana*.

Certaines études ne considèrent que les BBH (*Best Bidirectionnal Hit*) pour définir les paires de gènes paralogues. Cette stratégie consiste à considérer que A et B sont homologues si et seulement si B obtient le meilleur score lorsque la comparaison contre le protéome est faite à partir de A (et réciproquement). Cette stratégie nous semble critiquable dans le cas des plantes qui présentent des familles multigéniques pouvant comprendre un grand nombre de membres.

Nous considérerons donc l'ensemble des protéines homologues lors d'une requête, la partie critique étant alors l'établissement des seuils d'homologie qui dépendent essentiellement de 2 paramètres : le score d'alignement (ou bien la *E-value*) et la taille de la ressemblance (souvent un certain pourcentage de la longueur de la protéine la plus courte). Nous reprendrons dans un premier temps le protocole publié par Rizzon *et al.* [59] qui allie ces 2 critères pour détecter tous les paralogues d'un gène tout en minimisant le risque de faux positifs (voir [77] pour une discussion sur ce dernier point).

Cependant le protocole décrit dans [59] comporte une faiblesse quant à la construction des groupes de séquences paralogues car il utilise l'algorithme du lien simple qui, bien que rapide d'où son utilisation pour de grands jeux de données, présente l'inconvénient d'être sensible à l'effet de chaîne. Nous nous proposons d'utiliser d'autres algorithmes de classification notamment ceux du type *Markov clustering* [63]. Ces algorithmes dépendent également de paramètres. De la même manière que précédemment, nous testerons plusieurs paramètres pour étudier leur incidence sur le nombre et la taille des familles multigéniques générées.

Construction des différents types de duplication. Chez l'arabette on dispose des génomes complets pour deux espèces séparées depuis environ 5 millions d'années, *A.thaliana* et *A.lyrata*. Grâce à cette connaissance, la recherche de synténies sera facilitée pour détecter les duplications segmentales.

Nous proposons d'utiliser les reconstructions de génomes ancestraux de plantes, prochainement disponibles sur la plateforme *Genomicus* [78] ou bien les données de Salse *et al.* [68] pour déterminer les synténies dues à des WGD.

Pour la détection des gènes dupliqués en tandem (TAG), nous utiliserons le protocole utilisé par Rizzon *et al.* en 2006 car il présente l'avantage de distinguer deux définitions de la proximité : une qui est stricte et l'autre qui tolère des remaniements locaux postérieurs à la duplication [59].

Le troisième groupe de gènes dupliqués, les duplications segmentales (SG), sera défini par défaut comme n'étant ni WGD, ni TAG. Il regroupera des synténies rejetées pour des raisons d'inconsistence lors de l'algorithme de reconstruction des génomes ancestraux ou parce qu'elles sont trop courtes. Nous les identifierons par soustraction après comparaison des génomes d'*A.thaliana* et d'*A.lyrata*.

Distinction entre néo-fonctionnalisation, sous-fonctionnalisation et redondance fonctionnelle. L'estimation de la divergence entre les copies de gènes par la méthode ORM décrite dans la première partie, soit avec le modèle acide aminé, soit avec le modèle codon permettra de vérifier dans les cas calculables [7] si les matrices d'évolution observées suivent un même générateur ou non, ce qui devrait permettre de distinguer trois types d'évolution des gènes dupliqués :

- la néo-fonctionnalisation devrait se traduire par un changement du générateur de remplacements des acides aminés,
- la sous-fonctionnalisation sera caractérisée par des changements dans les générateurs des codons au sein d'une famille multigénique. En effet, l'usage des codons varie selon les compartiments cellulaires et/ou le niveau d'expression des gènes,
- la redondance fonctionnelle gardera le générateur inchangé.

Estimation des temps de divergence des protéines La méthode la plus largement utilisée dans la littérature pour mesurer les taux d'évolution des protéines est de loin la méthode du rapport K_A/K_S (taux de mutations non synonymes / taux de mutations synonymes) selon les modèles de codons implémentés dans le logiciel PAML (pour des détails sur les modèles de codons se reporter au chapitre 14 de [79] ou au chapitre 3 de [80]).

Le rapport K_A/K_S présente deux défauts :

1. La précision du rapport dépend de deux contraintes contradictoires : i) une estimation précise du K_A nécessite un nombre suffisant de changement d'acides aminés ; ii) K_s doit être assez faible pour que l'estimation ne soit pas faussée par des évènements multiples (sinon le rapport K_A/K_S se ramène à l'estimation du pourcentage d'identité des acides aminés).
2. La mesure du K_A ne prend en compte que le code génétique pour le taux de remplacement et pas la nature des acides aminés interchangés. Or on sait depuis les travaux précurseurs de Dayhoff [6] que les propriétés physico-chimiques des chaînes latérales jouent un rôle primordial dans la probabilité de remplacement des acides aminés. Certaines initiatives tentent de corriger ce dernier point comme le rapport K_R/K_C (taux de remplacement radical / taux de remplacement conservatif) développé par Hanada *et al.* [81].

Nous disposons d'un outil permettant d'estimer les taux d'évolution au niveau protéique qui utilise mieux l'information disponible : la méthode ORM *Observed Rates Matrices* (article [1] en annexe A.2). Cette méthode permet d'estimer les matrices de taux de remplacement des acides aminés à partir d'alignements fournis par l'utilisateur. Contrairement aux modèles protéiques implémentés dans PAML qui font appel à des matrices fournies dans la littérature (matrices Dayhoff [**?**], JTT [17], WAG [**?**], mtMam [**?**] et mtREV [23]), les matrices de taux estimées sur les données à partir d'alignements fournis par l'utilisateur par la méthode ORM (cf Section1.3) sont

7. un des intérêts de la méthode est qu'elle ne peut pas calculer de taux lorsque le signal est saturé, évitant ainsi les sur-interprétations

spécifiques de chaque famille protéique (sous réserve de disposer d'un alignement multiple décrivant la famille). Ainsi nous estimerons la divergence des paires de gènes dupliqués de manière plus fine (car adaptée à chaque famille) et pour des paires ayant divergé depuis plus longtemps que ne le permet la méthode K_S.

Changement de générateurs de codons Il est désormais acquis que les génomes eucaryotes sont organisés en îlots riches en GC (et en contrepartie d'autres riches en AT). Grâce à l'usage du code propre à chaque gène, nous pouvons distinguer ceux qui se trouvent depuis longtemps dans une région riche en AT ou en GC des autres. En effet, les premiers ont eu le temps d'adapter leur usage du code à la particularité de ces régions, contrairement aux gènes transférés récemment dans de telles régions.

Les codons utilisés préférentiellement dans les régions riches en AT sont ceux terminant par A ou T, et réciproquement pour les régions riches en GC. Par contre pour mesurer uniquement le biais en faveur des codons se terminant par G ou C (corr. A ou T), il faut corriger l'effet acide aminé et l'effet nombre de codons pour cet acide aminé (cf. protocole de Delorme et Hénaut [82]). De ce fait nous mesurerons uniquement le biais dans l'usage du code génétique pour les codons correspondant aux cases colorées sur la figure 3.1(acides aminés fréquents codés par 2, 4 ou 6 codons).

L'organisation des chromosomes en région riche en AT ou GC devrait faire apparaître 2 générateurs de codons : un pour les régions riches en AT, l'autre pour les régions riches en GC. Dans le cadre d'une redondance fonctionnelle, les paires de gènes dupliqués devraient suivre l'un ou l'autre de ces générateurs. Les paires pour lesquelles on observera un autre générateur seront celles potentiellement sujettes à néo- ou sous-fonctionnalisation.

Première base	Deuxième base T		Deuxième base C		Deuxième base A		Deuxième base G		Troisième base
T	TTT	Phe	TCT	Ser	TAT	Tyr	TGT	Cys	T
	TTC	Phe	TCC	Ser	TAC	Tyr	TGC	Cys	C
	TTA	Leu	TCA	Ser	TAA	Stop	TGA	Stop	A
	TTG	Leu	TCG	Ser	TAG	Stop	TGG	Trp	G
C	CTT	Leu	CCT	Pro	CAT	His	CGT	Arg	T
	CTC	Leu	CCC	Pro	CAC	His	CGC	Arg	C
	CTA	Leu	CCA	Pro	CAA	Gln	CGA	Arg	A
	CTG	Leu	CCG	Pro	CAG	Gln	CGG	Arg	G
A	ATT	Ile	ACT	Thr	AAT	Asn	AGT	Ser	T
	ATC	Ile	ACC	Thr	AAC	Asn	AGC	Ser	C
	ATA	Ile	ACA	Thr	AAA	Lys	AGA	Arg	A
	ATG	Met	ACG	Thr	AAG	Lys	AGG	Arg	G
G	GTT	Val	GCT	Ala	GAT	Asp	GGT	Gly	T
	GTC	Val	GCC	Ala	GAC	Asp	GGC	Gly	C
	GTA	Val	GCA	Ala	GAA	Glu	GGA	Gly	A
	GTG	Val	GCG	Ala	GAG	Glu	GGG	Gly	G

TABLE 3.1 – *Codons considérés pour la mise en évidence d'un biais de l'utilisation des codons se terminant par A ou T versus ceux se terminant par G ou C. On se limite aux acides aminés les plus fréquents : Ala, Asn, His, Ile, Leu, Phe, Pro, Ser, Thr, Tyr, Val [82]*

Prédiction du compartiment cellulaire Une dernière information d'importance va nous être apportée par les séquences en analysant le biais dans l'usage des codons synonymes. Il a en effet été démontré que l'analyse factorielle des correspondances de la fréquence des codons synonymes permet de prédire la localisation cellulaire [83] chez certains organismes présentant un biais marqué dans l'usage du code génétique comme c'est le cas pour l'arabette [84]. Cette information sera précieuse pour détecter une sous-fonctionnalisation et d'identifier les réseaux auxquels se rattachent potentiellement les différentes copies.

3.3.2 Ce qu'apporte l'analyse du transcriptome

Les catégories de devenir des gènes après duplication à savoir néo-fonctionnalisation, sous-fonctionnalisation, redondance fonctionnelle, telles qu'elles auront été déduites de l'analyse des séquences devraient être confirmées par les données du transcriptome qui vont également nous permettre de distinguer trois cas de figure :

- si les copies des gènes dupliqués s'expriment toujours ensemble, leur expressions seront toujours corrélées quelles que soient les conditions (redondance fonctionnelle).
- si les copies dupliquées s'expriment dans des conditions physiologiques différentes, chacune étant caractéristique d'une condition, leur expressions seront systématiquement anti-corrélées (sous-fonctionnalisation).
- si les copies sont impliquées dans des processus indépendants alors on n'observera pas de tendances générales (néo-fonctionnalisation ou nouvelle régulation).

Pour arriver à mettre en évidence ces corrélations il faut regarder l'expression des gènes dans un grand nombre de conditions, les plus variées possible. L'arabette est un modèle de choix pour ces analyses du fait du grand nombre de données disponibles. Nous comptons utiliser les données accessibles dans les banques publiques telle que FlagDB développée à l'URGV à Evry [85]. Cette base de données présente l'avantage d'intégrer dans un seul et même environnement des données spécifiques aux plantes modèles (*A.thaliana*, *Oryza sativa*, *Populus trichocarpa* et *Vitis vinifera*) et de nature diverses (séquences, informations génomiques, groupes de gènes paralogues, structures 2D et 3D, transcriptome...).

Les données de transcriptome sont affectées de biais systématiques qui sont variables d'une plate-forme à l'autre et d'une technologie à l'autre. Ces biais peuvent être contournés par un pré-traitement des données ou en choisissant une statistique robuste. Par exemple, il a été montré chez les bactéries que le τ de Kendall, basé sur les statistiques de signe, permet de tirer une information biologiquement pertinente d'un ensemble de données très hétérogène [86, 87]. C'est cette dernière solution que nous allons explorer dans un premier temps.

3.3.3 Ce qu'apporte l'analyse des réseaux d'interactions

Dans un troisième temps nous examinerons les paires de gènes paralogues au sein des réseaux d'interactions. Plusieurs types de réseaux sont accessibles à l'étude : les réseaux métaboliques, les réseaux d'interactions protéiques, les réseaux d'interactions génétiques inférés à partir des données de transcriptome, et chez la levure les réseaux d'interactions génétiques synthétiques permettant de mesurer expérimentalement à large échelle le phénoptype induit par des mutants double *knockout* (revue [88]).

Chez l'arabette, à notre connaissance, seuls les réseaux d'interaction PPI [75] et les réseaux métaboliques [89] ont été étudiés. Nous comptons analyser ceux-ci en prenant soin de séparer les paires de gènes dupliqués en catégories de duplication (WGD, SD et TAG) et nous nous proposons d'inférer les réseaux génétiques grâce aux méthodes développées au laboratoire [90] pour croiser toutes les approches.

3.4 En conclusion

A l'issue de ce projet nous devrions disposer de deux types d'informations sur les gènes dupliqués : i) le mécanisme d'obtention (WGD, SD et TAG) et l'âge de ces duplications et ii) leur devenir fonctionnel selon Ohno (néo-fonctionnalisation, sous-fonctionnalisation ou redondance fonctionnelle). Pour ce dernier point on a inféré la catégorie selon deux sources d'informations complètement indépendantes : les données de séquences et les données de transcriptome. Nous serons ainsi en mesure d'analyser s'il existe un biais dans les familles de protéines selon les mécanismes de duplication ou selon l'évolution fonctionnelle définie par Ohno.

Nous pourrons comparer nos résultats avec ceux obtenus sur la levure organisme modèle concernant l'évolution des gènes dupliqués. Cette comparaison est intéressante car beaucoup d'études ont déjà été réalisées sur cet organisme et parce qu'on dispose de données d'interactions génétiques expérimentales qui vont permettre de valider les réseaux prédits d'après les données transcriptomes.

Nous espérons qu'en combinant toutes ces informations, qui correspondent chacune à un angle de vision particulier de l'évolution des gènes dupliqués, nous arriverons à reconstruire une image globale et cohérente chez l'arabette des différents mécanismes de maintien des duplications. Nous contribuerons ainsi un peu à l'avancée de la connaissance des mécanismes évolutifs qui génèrent la diversité et l'émergence de nouvelles fonctions chez les eucaryotes.

Bibliographie

[1] C Devauchelle, A Grossmann, A Hénaut, M Holscneider, M Monnerot, J-L Risler, and B Torrésani. Rate matrices for analysing large families of proteins sequences. *Journal of Computational Biology*, 8 :381–399, 2004. (Cité pages 2, 8, 19, 20, 25, 26 et 53.)

[2] C Devauchelle, A Dress, A Grossmann, S Grünewald, and A Hénaut. Constructing hierarchical set systems. *Annals of Combinatorics*, 8 :441–456, 2004. (Cité pages 2, 8, 23 et 25.)

[3] G Didier, L Debomy, M Pupin, Zhang M., A. Grossmann, C Devauchelle, and I Laprevotte. Comparing sequences without using alignments : application to HIV/SIV subtyping. *BMC Bioinformatics*, 8 :1, 2007. (Cité pages 2, 28, 30, 32 et 36.)

[4] E Corel, F Pitschi, I Laprevotte, G Grasseau, G Didier, and C Devauchelle. MS4 - multi-scale selector of sequence signatures : An alignment-free method for classification of biological sequences. *BMC Bioinformatics*, 11 :406, 2010. (Cité pages 2, 21, 28, 33, 35, 36 et 37.)

[5] F Pitschi, C Devauchelle, and E Corel. Automatic detection of anchor points for multiple sequence alignment. *BMC Bioinformatics*, 11 :445, 2010. (Cité pages 2, 28 et 37.)

[6] M O Dayhoff, R V Eck, and C M Park. A model of evolutionary change in proteins. *Atlas of Protein Sequence and Structure*, 5 :89–99, 1972. (Cité pages 9, 10, 11 et 53.)

[7] University of Delaware and Georgetown University Medical Center. Protein information ressource : integrated protein informatics resource gor genomic, proteomic and systems biology research, 2009. University of Delaware and Georgetown University Medical Center. (Cité page 9.)

[8] S B Needleman and C D Wunsch. A general method applicable to the search for similarities in the amino acid sequence of two proteins. *Journal of Molecular Biology*, 48 :443–453, 1970. (Cité page 9.)

[9] P H Sellers. On the theory and computation of evolutionary distances. *SIAM Journal of Applied Mathematics*, 26 :787–793, 1974. (Cité page 9.)

[10] T F Smith and M S Waterman. Identification of common molecular subsequences. *Journal of Molecular Biology*, 147 :195–197, 1981. (Cité page 9.)

[11] G J Barton and M J Strenberg. A strategy for the rapid multiple alignment of protein sequences. *Journal of Molecular Biology*, 198 :327–337, 1987. (Cité page 9.)

[12] D-F Feng and R F Doolittle. Progressive sequence alignment as a prerequisite to correct phylogenetic trees. *Journal of Molecular Evolution*, 25 :351–360, 1987. (Cité page 9.)

[13] F Corpet. Multiple sequence alignment with hierarchical clustering. *Nucleic Acids Research*, 16 :10881–10890, 1988. (Cité page 9.)

[14] M Vingron and P Argos. A fast and sensitive multiple sequence alignment algorithm. *Computer Applications in BIOSciences*, 5 :115–121, 1989. (Cité page 9.)

[15] D G George, W C Barker, and L T Hunt. Mutation data matrix and its uses. *Method in enzymology*, 183 :333–345, 1972. (Cité pages 12 et 13.)

[16] S Henikoff and J G Henikoff. Amino acid substitution matrices from protein blocks. *Proceedings of National Academy of Sciences USA*, 89 :10915–10919, 1992. (Cité page 15.)

[17] T F Jones, W R Taylor, and J M Thornton. The rapid generation of mutation data matrices from protein sequences. *Computer Applications in BIOSciences*, 8 :275–282, 1992. (Cité pages 15, 18, 26 et 53.)

[18] C Lanave, C Preparata, C. Saccone, and G. Serio. A new method for calculating evolutionary substitution rates. *Journal of Molecular Evolution*, 20 :86–93, 1984. (Cité page 16.)

[19] DL Swofford, Y Olsen GJ, PJ Waddell, and DM Hillis. *Molecular systematics (2nd edition) : Phylogenetic inference*. Hillis, DM and Moritz, C and Mable, BK, 1996. (Cité pages 16, 17 et 21.)

[20] WH Li and Graur D. *Fundamentals of molecular evolution : Evolutionary change in nucleotide sequences*. Sinauer associates inc., 1991. (Cité pages 16, 17 et 18.)

[21] M Hasegawa, H Kishino, and T Yano. Dating of the human-ape splitting by a molecular clock of mitochondrial dna. *Journal of Molecular Evolution*, 21 :160–174, 1985. (Cité page 17.)

[22] T F Jones, W R Taylor, and J M Thornton. The rapid generation of mutation data matrices from protein sequences. *FEBS Letters*, 339 :269–275, 1994. (Cité page 18.)

[23] J Adachi and M Hasegawa. Model of amino acid substitution in proteins encoded by mitochondrial DNA. *Journal of Molecular Evolution*, 42 :459–468, 1996. (Cité pages 18, 20, 22 et 53.)

[24] J Weyer-Menkhoff, C Devauchelle, A Grossmann, and S Grünewald. Integer linear programming as a tool for constructing trees from quartet data. *Computer Biology and Chemistry*, 295 :196–203, 2005. (Cité pages 8 et 21.)

[25] D H Huson and D Bryant. Application of phylogenetics networks in evolutionary studies. *Molecular Biology and Evolution*, 23 :254–267, 2006. (Cité pages 26, 39 et 40.)

[26] O Gascuel. BIONJ : an improved version of the NJ algorithm based on a simple model of sequence data. *Molecular Biology and Evolution*, 14 :685–695, 1997. (Cité page 26.)

[27] F Felsenstein. PHYLIP : Phylogeny inference package (version 3.2). *Cladistics*, 5 :164–166, 1989. (Cité page 26.)

[28] G Didier. Caractérisation des N-écritures et application à l'étude des suites de complexité ultimement n+cstes. *Theoretical Computer Science*, 215 :31–49, 1999. (Cité page 29.)

[29] G Didier, I Laprevotte, M Pupin, and Hénaut A. Local decoding of sequences and alignment-free comparison. *Journal of Computational Biology*, 13 :1465–1476, 2006. (Cité pages 29 et 31.)

[30] I Laprevotte, M Pupin, E. Coward, G Didier, C Terzian, C Devauchelle, and Hénaut A. HIV-1 and HIV-2 nucleotide sequences : assessment of the alignment by N-block presentation, "retroviral signatures" of over-repeated oligonucleotides, and probable important role of scrambled stepwise duplications/deletions in molecular evolutionlocal decoding of sequences and alignment-free comparison. *Molecular Biology and Evolution*, 18 :1231–1245, 2001. (Cité page 37.)

[31] P Forterre, S Gribaldo, D Gadelle, and M C Serre. Origin and evolution of DNA topoisomerases. *Biochimie*, 89 :427–446, 2007. (Cité pages 38, 39 et 40.)

[32] M Nadal. Reverse gyrase : an insight into the role of DNA topoisomerases. *Biochimie*, 89 :447–455, 2007. (Cité page 38.)

[33] A Napoli, A Valenti, M Salerno, M Nadal, F Garnier, M Rossi, and M Ciaramella. Reverse gyrase recruitment to DNA after UV light irradiation in *Sulfolobus solfataricus*. *Journal of Biology and Chemistry*, 279 :33192–8, 2001. (Cité page 38.)

[34] F Garnier and M Nadal. Transcriptional analysis of the two reverse gyrase encoding genes of *Sulfolobus solfataricus* in relation to the growth phases and temperature conditions. *Extremophiles*, 12 :799–809, 2008. (Cité pages 38 et 44.)

[35] A Bizard, F Garnier, and M Nadal. Topr2, the second reverse gyrase of *Sulfolobus solfataricus*, exhibits unusual properties. *Journal of Molecular Biology*, to appear, 2011. (Cité pages 38, 41 et 44.)

[36] E Glémet and Codani J-J. Lassap, a large scale sequence comparison package. *Computer Applications in BIOSciences*, 13 :137–143, 1997. (Cité page 39.)

[37] B J Campbell, J L Smith, T E Hanson, M G Klotz, L Y Stein, K C Lee, D Wu, J M Robinson, H M Khouri, J A Eisen, and S C Cary.

Adaptations to submarine hydrothermal environments exemplified by the genome of *Nautilia profundicola*. *PLoS Genetics*, 5 :e10000362, 2009. (Cité page 39.)

[38] M Duguet, M C Serre, and C Bouthier de La Tour. A universal type IA topoisomerase fold. *Journal of Molecular Biology*, 359 :805–812, 2006. (Cité page 40.)

[39] G C Conan and K H Wolfe. Turning a hobby into a job : How duplicated genes find new functions. *Nature Reviews Genetics*, 9 :938–950, 2008. (Cité pages 44, 47 et 49.)

[40] Y Ven de Peer, S Maere, and A Meyer. The evolutionary significance of ancient genome duplications. *Nature Reviews Genetics*, 10 :725–732, 2009. (Cité pages 44, 45, 46 et 49.)

[41] G Liti and E J Louis. Yeast evolution and comparative genomics. *Reviews of Microbiology*, 59 :135–153, 2005. (Cité pages 44 et 45.)

[42] C Landès, J Perona, S Brunie, M Rould, C Zelwer, T Steitz, and J-L Risler. A structure-based multiple sequence alignment of all class I aminoacyl-tRNA synthetases. *Biochimie*, 77 :194–203, 1995. (Cité page 44.)

[43] Y Diaz-Lascoz, J-C Aude, P Nitschké, H Chiapello, C Landès-Devauchelle, and J-L Risler. Evolution of genes, evolution of species : The case of aminoacyl-tRNA synthetases. *Molecular Biology and Evolution*, 15 :1548–1561, 1998. (Cité page 44.)

[44] J-M Aury, O Jaillon, L Duret, ..., and P Winker. Global trends of whole-genome duplications revealed by the ciliate *Paramecium tetraurelia*. *Nature*, 444 :171–178, 2006. (Cité page 45.)

[45] T Marques-Bonet, S Girirajan, and E E Eichler. The origins and impact of primate segmental duplications. *Trends in Genetics*, 25 :443–453, 2009. (Cité pages 45, 46 et 47.)

[46] J F Wendel. Genome evolution in polyploids. *Plant Molecular Biology*, 42 :225–249, 2000. (Cité pages 45 et 49.)

[47] K H Wolfe and D C Shields. Molecular evidence for an ancient duplication of the entire yeast genome. *Nature*, 387 :708–713, 1997. (Cité page 45.)

[48] C Seoighe and K H Wolfe. Updated map of duplicated regions in the yeast genome. *Gene*, 238 :253–261, 1999. (Cité page 45.)

[49] The Arabidopsis Genome Initiative. Analysis of the genome sequence of the flowering plant *Arabidopsis thaliana*. *Nature*, 408 :796–815, 2000. (Cité pages 45, 46 et 51.)

[50] C Simillion, K Vandepoele, M Van Montagu, and Y Van de Peer. The hidden duplication past of *Arabidopsis thaliana*. *Proceedings of National Academy of Sciences USA*, 99 :13627–13632, 2002. (Cité page 45.)

[51] G Blanc, K Hokamp, and K H Wolfe. A recent polyploidy superimposed on older large-scale duplications in the Arabidopsis genome. *Genome research*, 13 :137–144, 2003. (Cité pages 45 et 49.)

[52] G Blanc and K H Wolfe. Widespread paleopolyploidy in model plant species inferred from age distributions of duplicate genes. *Plant Cell*, 16 :1667–1678, 2004. (Cité pages 45, 47 et 49.)

[53] L Skrabanek and K H Wolfe. Eukaryote genome duplication - where's the evidence ? *Current Opinion in Genetics and Development*, 8 :694–700, 1998. (Cité page 45.)

[54] K H Wolfe. Yesterday's polyploids and the mystery of diploidization. *Nature Reviews Genetics*, 2 :333–341, 2001. (Cité pages 45, 47, 48 et 51.)

[55] O Jaillon, J-M Aury, F Brunet, ..., and H Roest Crollius. Genome duplication in the teleost fish *Tetraodon nigroviridis* reveals the early vertebrate proto-karyotype. *Nature*, 431 :946–957, 2004. (Cité page 45.)

[56] S Maere, S de Bodt, J Raes, Casneuf T, and M Van Montagu. Modeling gene and genome duplications in eukaryotes. *Proceedings of National Academy of Sciences USA*, 102 :5454–5459, 2005. (Cité pages 47 et 49.)

[57] J Wienberg, A Jauch, R Stanyon, and T Cremer. Molecular cytotaxonomy of primates by chromosomal *in situ* suppression hybridization. *Genomics*, 8 :347–350, 1990. (Cité page 47.)

[58] M Muffato. *Reconstruction des génomes ancestraux chez les vertébrés*. PhD thesis, Université d'Evry Val d'Essonne, 2010. (Cité page 47.)

[59] C Rizzon, L Ponger, and B S Gaut. Striking similarities in the genomic distribution of tandemly arrayed genes in Arabidopsis and rice. *PLoS Computational Biology*, 2 :e115, 2006. (Cité pages 47, 49 et 52.)

[60] K Hanada, C Zou, MD Lehti-Shiu, K Shinozaki, and S-H Shiu. Importance of plant tandem duplicates in the adaptative response to environmental *stimuli*. *Plant Physiology*, 148 :993–1003, 2008. (Cité pages 47 et 49.)

[61] V Shoja, T M Murali, and L Zhang. Expression divergence of tandemly arrayed genes in human and mouse. *Computational Functional Genomics*, 60964, 2007. (Cité page 47.)

[62] L Despons, P V Baret, L Frangeul, V Leh Louis, P Durrens, and J-L Souciet. Genome-wide computational prediction of tandem gene arrays : application in yeast. *BMC Genomics*, 11 :56, 2010. (Cité page 47.)

[63] A J Enright, S Van Dongen, and C A Ouzounis. An efficient algorithm for large-scale detection of protein families. *Nucleic Acids Research*, 30 :1575–1584, 2002. (Cité pages 47 et 52.)

[64] M Lynch and J Conery. The evolutionary fate and consequence of duplicate genes. *Science*, 290 :1151–1155, 2000. (Cité pages 48 et 49.)

[65] M Lynch and A Force. The probability of duplicate gene preservation by subfunctionalization. *Genetics*, 154 :459–473, 2000. (Cité page 48.)

[66] A H Paterson, J E Bowers, and B A Chapman. Ancient polyploidization predating divergence of the cereals, and its consequences for comparative genomics. *Proceedings of National Academy of Sciences USA*, 101 :9903–9908, 2004. (Cité page 49.)

[67] X Wang, X Shi, B Hao, S Ge, and J Luo. Duplication and DNA segmental loss in the rice genome : Implications for diploidization. *New Phytologist*, 165 :937–946, 2007. (Cité page 49.)

[68] M Abrouk, F Murat, C Pont, J Messing, S Jackson, T Faraut, E Tannier, C Plomion, R Cooke, C Feuillet, and J Salse. Paleogenomics of plants : synteny-based modelling of extinct ancestors. *Trends in Plant Science*, 15 :479–487, 2010. (Cité pages 49, 51 et 52.)

[69] G Blanc and K H Wolfe. Functional divergence of duplicated genes formed by polyploidy during Arabidopsis evolution. *The Plant Cell*, 16 :1679–1691, 2004. (Cité page 49.)

[70] I Wapinski, A Pfeffer, N Friedman, and A Regev. Natural history and evolutionary principles of gene duplication in fungi. *Nature*, 449 :55–61, 2007. (Cité page 49.)

[71] J C Davis and A Petrov. Preferential duplication of conserved proteins in eukaryotic genomes. *PLoS Biology*, 2 :318–326, 2004. (Cité page 49.)

[72] M Sémon and K H Wolfe. Preferential subfunctionlization of slow-evolving genes after allopolyploidization in *Xenopus laevis*. *Proceedings of National Academy of Sciences USA*, 105 :8333–8338, 2008. (Cité page 49.)

[73] K Hanada, T Kuromori, F Myouga, T Toyoda, and K Shinozaki. Importance of plant tandem duplicates in the adaptative response to environmental *stimuli*. *PLoS Genetics*, 5 :e1000781, 2009. (Cité pages 49 et 50.)

[74] T Makino, Y Suzuki, and T Gojobori. Differential evolutionary rates of duplicated genes in protein interaction network. *Gene*, 385 :57–63, 2006. (Cité page 49.)

[75] *Arabidopsis* Interactome Mapping Consortium. Evidence for network evolution in an Arabidopsis interactome map. *Science*, 333 :601–607, 2011. (Cité pages 50 et 56.)

[76] T T Hu, P Pattyn, E G Bakker, J Cao, J-F Cheng, R M Clark, N Fahlgren, J A Fawcett, J Grimwood, H Gundlach, G Haberer, J D Hollister, S Ossowski, R P Ottilar, A A Salamov, K Schneeberger, M Spannagl, X Wang, L Yang, M E Nasrallah, J Bergelson, J C Carrington, B S Gaut, J Schmutz, K F X Mayer, P Van de Peer, I V Grigoriev, M Nordborg, D Weige, and Y-L Guo. The *Arabidopsis lyrata* genome sequence and the basis of rapid genome size change. *Nature Genetics*, 43 :476–481, 2011. (Cité page 51.)

[77] J Salse, M Abrouk, F Murat, U M Quraishi, and C Feuillet. Improved criteria and comparative genomics tool provide new insights into grass paleogenomics. *Briefings in bioinformatics*, 10 :619–630, 2009. (Cité page 52.)

[78] M Muffato, A Louis, C-E Poisnel, and H Roest Crollius. Genomicus : a database and a browser to study gene synteny in modern and ancestral genomes. *Bioinformatics*, 28 :1119–1121, 2010. (Cité page 52.)

[79] J Felsenstein. *Inferring Phylogenies*. Sinauer Associates, Inc, 2004. (Cité page 53.)

[80] O Gascuel and M Steel. *Reconstructing Evolution. New mathematical and computational advances*. Oxford, 2007. (Cité page 53.)

[81] K Hanada, S-H Shiu, and W-H Li. The nonsynonymous/synonymous substitution rate ratio versus the radical/conservative replacement rate ratio in the evolution of mammalian genes. *Molecular Biology and Evolution*, 24 :2235–2241, 2007. (Cité page 53.)

[82] M-O Delorme and A Hénaut. Codon usage is imposed by the gene location in the transcription unit. *Current Genetics*, 20 :353–358, 1991. (Cité page 54.)

[83] H Chiapello, B Torrésani, A Grossmann, and A Hénaut. Codon usage as a tool to predict the cellular location of eukaryotic ribosomal proteins and aminoacyl-tRNA synthetases. *Nucleic Acids Research*, 27 :2848–2851, 1999. (Cité page 55.)

[84] H Chiapello, F Lisacek, M Caboche, and A Hénaut. Codon usage and gene function are related in sequences of *Arabidopsis thaliana*. *Gene*, 209 :2848–2851, 1998. (Cité page 55.)

[85] S Dèrozier, F Samson, J-P Tamby, C Guichard, V Brunaud, P Grevet, S Gagnot, P Label, J-C Leplé, A Lecharny, and S Aubourg. Exploration of plant genomes in the FLAGdb++ environment. *Plant methods*, 7 :8, 2011. (Cité page 55.)

[86] A-S Carpentier, B Torrésani, A Grossmann, and A Hénaut. Decoding the nucleoid organisation of *Bacillus subtilis* and *Escherichia coli* through gene expression data. *BMC Genomics*, 6 :84, 2005. (Cité page 55.)

[87] A Riva, A-S Carpentier, F Barloy-Hubler, A Chéron, and A Hénaut. Analyzing stochastic transcription to elucidate the nucleoid's organization. *BMC Genomics*, 9 :125, 2008. (Cité page 55.)

[88] A Baryshnikova, M Costanzo, S Dixon, F J Vizeacoumar, C L Myers, B Andrews, and C Boone. Synthetic Genetic Array (SGS). analysis in *Saccharomyces cerevisiae* and *Schizosaccharomyces pombe*. *Methods in enzymology*, 470 :470, 2010. (Cité page 56.)

[89] K Hanada, Y Sawada, T Kuromori, R Klausnitzer, K Saito, T Toyoda, K Shinozaki, W-H Li, and M Y Hirai. Functional compensation of

primary and secondary metabolites by duplicate genes in *Arabidopsis thaliana*. *Molecular Biology and Evolution*, 28 :377–382, 2011. (Cité page 56.)

[90] J Chiquet, A Smith, G Grasseau, C Matias, and C Ambroise. SIMoNe : Statistical Inference for MOdular NEtworks. *Bioinformatics*, 25 :417–418, 2009. (Cité page 56.)

Annexes

A

Sommaire