Évolution in silico des protéines monomériques et dimériques.

Josselin Noirel

École doctorale de l'École polytechnique

Mémoire présenté en vue de l'obtention du
titre de docteur de l'École polytechnique

Évolution *in silico* des protéines monomériques et dimériques

In silico evolution of monomeric and dimeric proteins

Josselin Noirel

Denis	Couvet	Président
Richard	Lavery	Rapporteur
Marc	Delarue	Rapporteur
Nicolas	Lartillot	
Yves-Henri	Sanejouand	
Thomas	Simonson	Directeur de thèse

23 octobre 2006

If one asks, where does this [genetic] information come from and what is its primary semantics, the answer is: information generates itself in feedback loops via replication and selection, the objective being 'to be or not to be.'

—Eigen

Résumé

La simulation *in silico* des gènes codant pour des ARN de transfert et des protéines a connu un développement considérable ces dernières années car elle permet de déduire de modèles simples des comportements inattendus qui découlent de la structure des génotypes dans l'espace des séquences et de la correspondance génotype-phénotype. Le modèle d'évolution le plus élémentaire, la théorie dite « de l'évolution neutre » conçue et défendue par Kimura principalement, donne lieu à un phénomène à présent bien documenté : les génotypes robustes aux mutations et exprimant des protéines se repliant efficacement, sont surreprésentés en comparaison des génotypes « fragiles ».

De nombreuses questions restent en suspens notamment en ce qui concerne l'incidence que peut avoir une modélisation plus réaliste de la fonctionnalité d'une protéine sur le schéma tracé à partir de considérations purement structurales et cinétiques. Pour cela, nous avons développé un modèle incluant une contrainte sélective imposant une dimérisation spécifique minimale de deux protéines codées par deux gènes pour qu'un individu puisse survivre. Nous démontrons que les réseaux neutres construits d'après des critères structuraux sont grandement plastiques et peuvent s'adapter à une fonction vitale sans souffrir de baisse de stabilité. La surreprésentation des génotypes robustes est maintenue, elle est même amplifiée par l'interaction épistatique existant entre les deux gènes. On observe que cela s'accompagne d'une augmentation en moyenne de la fonctionnalité résultant de l'émergence d'un *super-funnel* fonctionnel dans l'espace des séquences. Cette propriété remarquable pourrait avoir d'importantes implications dans l'explication de l'émergence de nouvelles fonctions biologiques.

Une autre question concerne les simplifications impliquées par le choix des modèles protéiques. Puisque les simulations évolutives supposent un coût de calcul important, les protéines sur réseau ont eu la préférence de nombreux modélisateurs. Dans ce mémoire, nous proposons un modèle de protéine hors réseau possédant des cartes de contacts plus complexes que les protéines sur réseau. Il confirme les conclusions tirées des simulations sur les protéines sur réseau.

Abstract

The *in silico* evolution simulation of protein- or tRNAs-encoding genes has recently been paid much attention because it offers the possibility to deduce from simple assumptions striking and unexpected behaviours deriving from the genotype structure in sequence space and the genotype-phenotype mapping. The most elementary evolutionary model, known as the 'neutral theory of molecular evolution' developed and advocated by Kimura from the late 60s on, has given rise to a now well documented phenomenon: genotypes that are robust to mutations and encode fast-folding proteins as evidenced and measured by the folding temperature, turn out to be overrepresented as compared to fragile ones.

Many questions still remain unsettled, amongst which whether a realistic model of functionality could change the picture drawn from models mainly based on structure preservation and/or foldability. We have therefore developed a model including a selective constraint imposing that the proteins encoded by two genes must specifically and efficiently dimerise in order to make an individual survive. We shall demonstrate that the neutral networks built upon structural considerations are plastic enough to accommodate the functional requirement without any noticeable impact on stability. The overrepresentation of robust genotypes still holds and appears to be magnified by the epistatic interactions between genes. Is is of interest that such a phenomenon is as well accompanied by an improvement in average functionality resulting from a functional superfunnel organisation in sequence space. This conclusion could have important implications in the way to explain the emergence of new functions.

Another issue regards the simplifications involved by the proteins models. Simulations of evolution entail large computations, that have been tackled using lattice protein models. In this dissertation we propose an off-lattice protein model exhibiting more complex contact maps–which supports the conclusions drawn from the lattice model.

Table des matières

I Introduction — 15

I.1 Théories de l'évolution 18
I.2 Tempo de l'évolution 24
I.3 Contraintes fonctionnelles et sélection négative 36
I.4 Évolution à l'échelle moléculaire 38
I.5 Champ du travail 45

II Modèle et méthodes — 47

II.1 Introduction 47
II.2 Définition et propriétés des graphes. 48
II.3 Modèle évolutif. 53
II.4 Modèles de protéine 62

III Évolution des protéines monomériques — 87

III.1 Introduction 87
III.2 Résultats 88
III.3 Discussion 126
III.4 Annexes 145

IV MODÈLE ET MÉTHODES (protéines dimériques) — 163

IV.1 Adaptation du modèle évolutif. 163
IV.2 Modélisation de la dimérisation 168

V ÉVOLUTION DES PROTÉINES DIMÉRIQUES — 177

V.1 Introduction. 177
V.2 Résultats 179
V.3 Discussion 207
V.4 Annexes 216

VI ARTICLE — 219

CONCLUSION — 247

REMERCIEMENTS — 253

BIBLIOGRAPHIE — 257

Table des matières détaillée

I Introduction — 15

I.1 Théories de l'évolution 18
I.1.1 Théorie déterministe *18*
I.1.2 Théorie de l'évolution neutre *21*
I.1.3 Théorie de l'évolution quasi neutre *23*
I.2 Tempo de l'évolution 24
I.2.1 Paradoxe d'Haldane *24*
I.2.1.1 Charge génétique. *24*
I.2.1.2 Charge associée à la substitution d'un allèle. *26*
I.2.2 Dérive génétique *27*
I.2.3 Horloge moléculaire. *31*
I.2.4 Faiblesses de l'évolution neutre. *33*
I.2.4.1 Révision du paradoxe d'Haldane. *33*
I.2.4.2 Dispersion de l'horloge moléculaire. *34*
I.2.4.3 Polymorphisme protéique *35*
I.3 Contraintes fonctionnelles et sélection négative 36
I.3.1 Mutations synonymes et non-synonymes. *36*
I.3.2 Pseudogènes. *36*
I.3.3 Conservation au niveau moléculaire *37*
I.4 Évolution à l'échelle moléculaire 38
I.4.1 Du génotype au fitness. *38*
I.4.1.1 Correspondance génotype-phénotype-fitness *38*
I.4.1.2 Modèles de correspondance génotype-phénotype-fitness . . *40*

I.4.2 Façonnage des séquences et des structures *43*
I.4.2.1 Évolution de modèles d'ARN *43*
I.4.2.2 Évolution de modèles de protéine. *44*
I.5 Champ du travail 45

II MODÈLE ET MÉTHODES — 47

II.1 Introduction 47
II.2 Définition et propriétés des graphes. 48
II.2.1 Définition *48*
II.2.2 Spectre d'un graphe *50*
II.2.3 Méthode de la puissance *50*
II.2.4 Théorème de Frobenius-Perron *51*
II.2.5 Composantes connexes. *52*
II.2.6 Plus court chemin reliant deux points *52*
II.3 Modèle évolutif. 53
II.3.1 Construction du réseau neutre *53*
II.3.2 Équation d'évolution *54*
II.3.3 Caractérisation de l'état stationnaire *55*
II.3.4 Convergence de la récurrence *57*
II.3.4.1 Réseau non bipartite *57*
II.3.4.2 Réseau bipartite *59*
II.3.5 Autres modèles proches *59*
II.3.6 Polymorphisme, monomorphisme à l'état stationnaire *61*
II.4 Modèles de protéine 62
II.4.1 Protéines sur réseau *63*
II.4.1.1 Conformations *63*
II.4.1.2 Séquences et alphabet HP *66*
II.4.1.3 Énergie et propriétés associées *67*
II.4.1.4 Matrices d'interaction *68*
II.4.1.5 Séquences viables, réseaux neutres *70*
II.4.1.6 Stockage des séquences, calcul des connexions *71*
II.4.1.7 Définition de la séquence prototype *72*
II.4.1.8 Remarques concernant le modèle. *72*
II.4.2 Protéines tridimensionnelles. *73*
II.4.2.1 Conformations et séquences natives. *74*
II.4.2.2 États dénaturés *75*
II.4.2.3 Enfilage des séquences et calcul de l'énergie *77*

II.4.2.4 Classes d'acides aminés, matrice d'interaction 78
II.4.2.5 Trajectoires 80
II.4.2.6 Préoptimisation 82
II.4.2.7 Profils, réseaux neutres 83
II.4.2.8 Schéma global 84

III Évolution des protéines monomériques — 87

III.1 Introduction 87
III.2 Résultats 88
III.2.1 Protéines sur réseaux 88
III.2.1.1 Statistiques de repliement, réseaux neutres 88
III.2.1.2 Propriétés thermodynamiques 92
III.2.1.3 Propriétés topologiques 100
III.2.1.4 Propriétés évolutives 102
III.2.2 Protéines tridimensionnelles. 106
III.2.2.1 Trajectoires réalisées 107
III.2.2.2 Exemple de préoptimisation 107
III.2.2.3 Propriétés topologiques 110
III.2.2.4 Propriétés évolutives 117
III.2.3 Effet de la taille de la population 121
III.2.3.1 Du régime uniforme à l'état stationnaire 121
III.2.3.2 Influence de la taille du réseau neutre 123
III.2.3.3 Effet du modèle évolutif : modèle de Wright-Fisher 124
III.3 Discussion 126
III.3.1 Structure des réseaux neutres 128
III.3.1.1 Organisation de superfunnel 128
III.3.1.2 Mutations compensatrices dans les petits réseaux 129
III.3.1.3 Distribution de la robustesse mutationnelle 130
III.3.2 Modifiabilité 132
III.3.2.1 Designability principle 132
III.3.2.2 Modifiabilité et taux d'évolution 132
III.3.2.3 Modifiabilité et structures secondaires 133
III.3.3 Évolution neutre adaptative 135
III.3.3.1 Évolution adaptative vers la séquence prototype 135
III.3.3.2 Presqu'îles neutres et scénarios évolutifs 135
III.3.3.3 L'hétérozygotie à l'origine de l'adaptation 137
III.3.4 Qualité du modèle structural 140

III.4 Annexes 145
III.4.1 Statistiques *145*
III.4.2 Propriétés moyennes des réseaux neutres *147*
III.4.3 Distribution de la robustesse mutationnelle *150*
III.4.3.1 Distribution universelle *150*
III.4.3.2 Origine de la forme des distributions *151*
III.4.4 Modifiabilité *153*
III.4.5 Réseaux neutres de la matrice Ising *154*
III.4.6 Résumé des propriétés des réseaux neutres *157*
III.4.7 Alphabets plus complets *158*
III.4.8 Modèle tous atomes *160*
III.4.9 Distance fondée sur les carbones Cβ *161*
III.4.10 Randomisation du graphe *162*

IV MODÈLE ET MÉTHODES (protéines dimériques) — 163

IV.1 Adaptation du modèle évolutif. 163
IV.1.1 Construction des hyper-réseaux neutres *163*
IV.1.2 Équation d'évolution et état stationnaire *165*
IV.1.3 État stationnaire sans contrainte *166*
IV.2 Modélisation de la dimérisation 168
IV.2.1 Protéines sur réseau *168*
IV.2.1.1 Modèle d'interaction *168*
IV.2.1.2 Équilibre chimique *169*
IV.2.1.3 Température de repliement *172*
IV.2.1.4 Critère sélectif *172*
IV.2.2 Protéines tridimensionnelles. *172*
IV.2.2.1 Principe de l'étude *172*
IV.2.2.2 Séquences de référence. *174*
IV.2.2.3 Formation des fausses structures *174*

V ÉVOLUTION DES PROTÉINES DIMÉRIQUES — 177

V.1 Introduction. 177
V.2 Résultats 179
V.2.1 Protéines sur réseau *179*
V.2.1.1 Concentration en dimère fonctionnel *180*
V.2.1.2 Composantes connexes de l'hyper-réseau neutre *182*

V.2.1.3 Localisation des génotypes viables . . . 182
V.2.1.4 Filtrage . . . 184
V.2.1.5 Propriétés à l'état stationnaire . . . 186
V.2.1.6 Effet de la dynamique de population . . . 190
V.2.1.7 Superfunnel *fonctionnel . . . 193*
V.2.2 Protéines tridimensionnelles. . . . 195
V.2.2.1 Statistiques de viabilité . . . 195
V.2.2.2 Propriétés à l'état stationnaire . . . 195
V.2.2.3 Localisation des génotypes viables . . . 200
V.2.2.4 Superfunnel *fonctionnel . . . 202*
V.2.2.5 Conservation des résidus . . . 202
V.3 Discussion . . . 207
V.3.1 Dimérisation comme modèle de fonctionnalité. . . . 207
V.3.1.1 Une fonctionnalité « non structurale » . . . 207
V.3.1.2 Importance de la sélection négative . . . 208
V.3.2 Propriétés des génotypes viables . . . 209
V.3.2.1 Diversité et connectivité des génotypes viables . . . 209
V.3.2.2 Stabilité des séquences fonctionnelles . . . 209
V.3.2.3 Organisation en superfunnel *fonctionnel . . . 210*
V.3.3 Évolution neutre de la fonctionnalité . . . 210
V.3.4 Émergence d'une fonction . . . 211
V.3.5 Conservation des résidus . . . 213
V.3.6 Taux d'évolution . . . 214
V.4 Annexes . . . 216
V.4.1 Composition hydrophobe expérimentales . . . 216

VI Article — 219

Conclusion — 247

Remerciements — 253

Bibliographie — 257

Chapitre I

Introduction

À l'échelle de la population, par conséquent, la mutation n'est nullement un phénomène d'exception : c'est la règle.

Jacques Monod, *Le Hasard et la nécessité*

Dans son *Évolution créatrice* [10], Bergson distingue « corps inorganisés » et « corps organisés » en ces termes :

> Plus généralement, les corps inorganisés, qui sont ceux dont nous avons besoin pour agir et sur lesquels nous avons modelé notre façon de penser, sont régis par cette loi simple : « le présent ne contient rien de plus que le passé, et ce qu'on trouve dans le corps était déjà dans sa cause ». Mais supposons que le corps organisé ait pour trait distinctif de croître et de se modifier sans cesse, comme en témoigne d'ailleurs l'observation la plus superficielle, il n'y aurait rien d'étonnant à ce qu'il fût *un* d'abord et *plusieurs* ensuite [10, p. 14].

Selon Bergson, c'est une évolution qui caractérise les corps organisés, l'évolution *créatrice* qui fait que l'organisé présent est supérieur à ce qu'il fut, s'enrichissant du temps qui passe ; ce qu'il soutient avec plus de lyrisme quelques pages plus loin : « Partout où quelque chose vit, il y a, ouvert quelque part, un registre où le temps s'inscrit ». Ce registre est cette chimère dont parle Monod dans la préface du *Hasard et la nécessité*.

> Cependant, si assuré qu'on fût dès le XIX^e^ siècle de sa validité phénoménologique, la théorie de l'Évolution, tout en dominant la biologie entière, demeurait comme suspendue tant que n'était pas élaborée une théorie *physique* de l'héré-

> dité. L'espoir d'y parvenir bientôt paraissait presque chimérique il y a trente ans, malgré les succès de la génétique classique [137, p. 12].

Si la découverte de la « théorie du code génétique » ôta une partie du voile de mystère sur le registre dans lequel l'évolution et le temps s'inscrivent, elle n'était pas suffisante pour identifier ni même décrire les principes et les mécanismes impliqués dans l'évolution. L'outil qui a permis de faire les premiers pas dans cette direction est la génétique des populations (cf. références [50, 53]). Reprenant les concepts développés par Mendel, elle a apporté un formalisme et les concepts qui survivent encore aujourd'hui.

Une théorie essentiellement déterministe de l'évolution fut développée en premier lieu. Elle régna longtemps en maître, car aucune donnée moléculaire ne venait la contredire. Quand les progrès de la biochimie et de la biologie moléculaire permirent enfin d'élucider la séquence des protéines et des gènes, on se rendit compte que la théorie et la réalité étaient incompatibles sur un certain nombre de points. C'est sur ces bases, que se développa la théorie de l'évolution neutre de Kimura.

La capacité d'un individu à peupler de sa descendance la génération qui le suit, le « *fitness* »[1], est déterminée par le phénotype qui est, lui-même, l'expression du génotype. Du génotype au *fitness*, on a une double correspondance

$$\text{génotype} \longrightarrow \text{phénotype} \longrightarrow \text{capacité reproductrice}. \tag{1}$$

À l'échelle du gène, le phénotype est le produit d'expression, l'ARNt, l'ARNr, la protéine. Sous les paramètres phénoménologiques de la théorie mathématique de l'évolution que nous allons présenter, se cachent des contraintes structurales et fonctionnelles au niveau moléculaire. Une compréhension microscopique de l'évolution doit tenir compte de ces considérations structurales et fonctionnelles. Comme l'écrit Feynman :

> Certainly no subject or field is making more progress on so many fronts at the present moment, than biology, and if we were to name the most powerful assumption of all, which leads one on and on in an attempt to understand life, it is that *all things are made of atoms*, and that everything that living things do can be understood in terms of the jigglings and wigglings of atoms [54].

La théorie de l'évolution neutre postule que la plus grande majorité des mutations qui jalonnent l'histoire d'un gène sont neutres, c'est-à-dire qu'elles ne furent pas motivées par un avantage sélectif. L'occurrence de ces mutations et leur incorporation dans l'his-

1. Le français ne dispose pas, à ma connaissance, de terme qui soit approprié pour traduire *fitness*. Les expressions comme « taux différentiel de reproduction », « potentiel reproductif », etc., étant lourdes, nous nous résignerons à recourir au terme anglais.

toire d'un gène dans une espèce repose sur un phénomène appelé « dérive génétique » : la destinée d'un gène pour lequel existe une multitude d'allèles possibles est foncièrement aléatoire.

L'enjeu des simulations d'évolution *in silico* est multiple. Il s'agit de justifier l'hypothèse fondatrice de la théorie neutre : comment est-il possible de légitimer, par des arguments moléculaires, qu'un grand nombre de séquences soit compatible avec la fonction d'un gène (exprimant une enzyme, par exemple) ? Il s'agit également de quantifier comment s'articulent l'évolution neutre et l'évolution adaptative : peut-on évaluer si les mutations adaptatives résultant d'un modèle structural et fonctionnel sont effectivement rares et quels sont les mécanismes par lesquels l'évolution peut les réaliser malgré tout ?

Si l'on accepte la théorie de l'évolution neutre, d'autres questions sont soulevées. Les formules classiques de Kimura sur le polymorphisme génétique d'une population (cf. l'équation 16, p. 30) ou la vitesse d'évolution d'une protéine (cf. l'équation 19′, p. 32) découlent d'hypothèses simplistes : par exemple, le modèle des allèles infinis suppose que tous les allèles possèdent le même taux de mutation neutre. Cela est équivalent à la supposition que les mutations procèdent indépendamment les unes des autres. Ces simplications sont nécessaires au traitement mathématique mais les données expérimentales structurales et fonctionnelles sur les protéines prouvent qu'elles ne sont pas valides. Une mutation d'un acide aminé donné est soumis à certaines conditions affectant les acides aminés à proximité. Cette observation s'est avérée pour les structures secondaires d'ARNt aussi. La dépendance entre les différents sites d'une séquence est comparée à l'épistase[2] par Fontana dans son article *Modelling 'evo-devo' with RNA*. Qu'une mutation A → G soit neutre ou pas est conditionné par l'ensemble d'une séquence, par des chemins que détermine essentiellement la structure [58].

Il est donc nécessaire de réexaminer les simplifications classiques de l'évolution à la lumière de considérations structurales et fonctionnelles modernes. Notre objectif est de modéliser l'évolution neutre des interactions protéine-proteine, par l'étude de protéines dont la fonction repose sur la formation d'un dimère.

Notre modèle s'appuiera sur la théorie de l'évolution neutre. Au préalable, l'introduction présentera donc les théories de l'évolution (section I.1). Nous introduirons les concepts de la théorie classique telle qu'elle fut échafaudée par Fisher au début du vingtième siècle. Puis, nous décrirons la théorie de l'évolution neutre et la théorie de l'évo-

2. L'« épistase » est la dépendance *entre les gènes* dans un génome. Elle survient, par exemple, quand l'effet d'une mutation dans un gène est conditionnée par l'allèle présent à un autre *locus*. Ainsi, une double mutation $x^- y^-$ peut être létale alors qu'aucune des mutations simples, $x^- y^+$ et $x^+ y^-$, ne l'est.

lution quasi neutre qui essayèrent de rendre compte des données moléculaires rendues disponibles dans les années 1960.

Nous examinerons ensuite les vitesses d'évolution prédites par ces différentes théories, leur adéquation à l'horloge moléculaire (section I.2). Le fonctionnement d'une protéine œuvrant de concert avec d'autres partenaires dans un complexe étant une contrainte supplémentaire, la relation qu'entretiennent l'évolution et la contrainte structurale fonctionnelle fera l'objet de la section I.3.

Les résultats importants dans la compréhension de l'évolution par des approches computationnelles seront explorés dans la section I.4. Enfin, la section I.5 resituera, plus précisément, l'importance des interactions protéine-protéine dans notre travail et les travaux futurs.

I.1 Théories de l'évolution

La définition du modèle que nous allons mettre en œuvre, reposant sur l'évolution neutre, requiert de présenter les révisions qui ont été apportées à la théorie de l'évolution développée dans la première moitié du vingtième siècle. La théorie déterministe est influencée par le darwinisme selon lequel l'évolution n'opère que par sélection d'individus « supérieurs ». La théorie de l'évolution neutre (et la théorie de l'évolution quasi neutre) reconnaît un rôle prépondérant au hasard parce qu'un grand nombre de génotypes est compatible avec la survie d'un organisme.

I.1.1 Théorie déterministe

Nous nommons « théorie déterministe » la conception qui émergea des travaux de Fisher et se concrétisa sous le nom de « néodarwinisme » selon laquelle l'évolution d'une population est essentiellement déterministe et gouvernée par une sélection naturelle toute-puissante.

La génétique des populations introduisit le concept de *fitness* qui mesure l'aptitude d'un individu à peupler la génération suivante d'individus fertiles qui lui soient identiques ou non. Deux composantes interviennent dans cette grandeur. La « viabilité » v indique la probabilité qu'un individu porteur d'un certain génotype parvienne à l'âge de la reproduction. La « fertilité » f dénote la capacité d'un individu à se reproduire. Le *fitness* absolu est noté traditionnellement W et vaut $v\,f$. On introduit également le *fitness* relatif $w = \mathrm{W}/\mathrm{W}_0$ où W_0 est un *fitness* absolu de référence, celui de la souche sauvage par exemple. Les mo-

dèles de génétique de population les plus simples font l'hypothèse de générations discrètes, non chevauchantes, s'accouplant au hasard[3] et considèrent seulement deux allèles.

Soit, chez un organisme diploïde, un gène pour lequel existent deux allèles, l'un sauvage, A, et l'un mutant, a, en proportions p et $q = 1 - p$ dans la population. Le *fitness* et les proportions génotypiques à l'instant t sont indiqués dans le tableau I.1.

Génotype	AA	Aa	aa
Fitness relatif	$w_{AA} = 1$	$w_{Aa} = 1 + h\,s$	$w_{aa} = 1 + s$
Proportion (t)	p^2	$2\,p\,q$	q^2
Quantité de zygotes	p^2	$2\,(1 + h\,s)\,p\,q$	$(1 + s)\,q^2$

Tab. I.1 : Tableau récapitulant à l'instant t, les fréquences et le *fitness* des génotypes considérés. La variable s représente l'avantage sélectif de l'allèle mutant a relativement à l'allèle résident A, h le degré de dominance (une valeur nulle fait de l'allèle mutant un allèle récessif, et au contraire un allèle dominant lorsque h vaut un). Les variables p et $q = p - 1$ sont les fréquences d'occurrence des allèles A et a respectivement, et les proportions p^2, $2\,p\,q$ et q^2 les proportions des différents génotypes diploïdes au temps t.

Le paramètre s est l'avantage sélectif du mutant et h est le degré de dominance. Dans la suite du développement, s est supposé strictement positif (mutant avantageux) ; h est soumis aux mêmes conditions. Les cas particuliers $h = 0$, $h = 1$, $0 < h < 1$ et $h > 1$ correspondent respectivement à un mutant récessif, à un mutant dominant, à un mutant partiellement dominant et à la surdominance (voir plus bas). Dans ce qui suit, AA', Aa' et aa' désignent le nombre d'individus portant un certain génotype au temps $t+1$. Puisqu'un individu de génotype AA est porteur de deux allèles A et qu'un individu de génotype Aa en possède seulement un exemplaire, la proportion p' de l'allèle sauvage au temps $t+1$ est donnée par

$$p' = \frac{2\,AA' + Aa'}{2\,(AA' + Aa' + aa')}.$$

En posant les tables de croisement de Mendel, on peut montrer que (cf. [53, p. 44], pour un développement complet et quelque peu fastidieux)

$$p' = \frac{p^2\,w_{AA} + p\,q\,w_{Aa}}{p^2\,w_{AA} + 2\,p\,q\,w_{Aa} + q^2\,w_{aa}}.$$

On réécrit :

$$p' = \frac{p^2 + (1 + h\,s)\,p\,q}{p^2 + 2\,(1 + h\,s)\,p\,q + (1 + s)\,q^2} = p\,\frac{\langle w\rangle_A}{\langle w\rangle}, \tag{2}$$

3. *Randomly mating population with non-overlapping discrete generations* : population sexuée se reproduisant de manière synchrone aux temps entiers $t = 0, 1, 2$, etc. Le choix des partenaires sexuels est supposé complètement aléatoire ; la probabilité des accouplements entre deux individus de génotypes dont les fréquences sont x et y est, par conséquent, $x\,y$.

où $\langle w \rangle$ est le *fitness* relatif moyen de la population à l'instant t, et $\langle w \rangle_A$ est celui de l'allèle A :

$$\langle w \rangle_A = p\, w_{AA} + q\, w_{Aa}\,.$$

La condition d'équilibre $\Delta p = p' - p = 0$ correspond aux solutions

$$p_1 = 0 \tag{3a}$$

$$p_2 = 1 \tag{3b}$$

$$p_3 = 1 + \frac{h}{1 - 2\,h}\,. \tag{3c}$$

L'expression donnant p_3 peut être écartée pour des valeurs de $0 \leq h \leq 1$. En effet, elle donne soit des valeurs en dehors de l'intervalle $[0, 1]$, soit des valeurs redondantes avec p_1 et p_2. La stabilité des équilibres peut être étudiée en dérivant l'expression de p' par rapport à p :

$$\left(\frac{\mathrm{d}p'}{\mathrm{d}p}\right)_{p_1} = \frac{1 + h\,s}{1 + s}\,, \tag{4a}$$

$$\left(\frac{\mathrm{d}p'}{\mathrm{d}p}\right)_{p_2} = 1 + h\,s\,. \tag{4b}$$

Par conséquent, pour une mutation partiellement dominante ($0 \leq h < 1$), l'équilibre est stable en p_1 ($\mathrm{d}p'/\mathrm{d}p < 1$) et instable en p_2 ($\mathrm{d}p'/\mathrm{d}p > 1$).

À présent, si $h > 1$, les équilibres en p_1 et en p_2 deviennent tous deux instables. L'équilibre en p_3 prend alors non seulement une valeur physiquement acceptable dans l'intervalle $]0, 1[$ mais en outre devient stable :

$$\left(\frac{\mathrm{d}p'}{\mathrm{d}p}\right)_{p_3} = \frac{h\,(2 + s) - 1}{h\,(2 + h\,s) - 1} < 1\,. \tag{4c}$$

Ce dernier cas, quand $h > 1$ et $s > 0$, donne à l'hétérozygote un phénotype plus marqué qu'aux homozygotes. Il porte le nom de « surdominance » ; nous y reviendrons plus tard au moment de définir la « charge génétique ».

La figure I.1 représente la dynamique des populations dans trois cas : mutant récessif ($h = 0$), mutant dominant ($h = 1$) et surdominance. Si l'on fait abstraction du cas de surdominance, la dynamique tend vers l'élimination de l'allèle le moins favorable et l'envahissement de la population par le plus favorable. On parle de la « fixation » de l'allèle a. On parle également de la « substitution » de l'allèle A par l'allèle a lorsque l'on s'intéresse à l'alternance des allèles à un *locus* (un gène). La vitesse à laquelle se succèdent les substitutions est appelée « taux de substitution » et caractérise la rapidité d'évolution d'un gène.

La figure I.1 suggère que les allèles délétères récessifs ($h = 1$ puisqu'ici l'allèle délétère est le résident) sont plus difficilement éliminés que ceux dominants. En effet, quand un allèle devient rare $p \approx 0$, il est porté principalement par les hétérozygotes (en proportion $2\,p\,q \approx 2\,p$, à comparer à la proportion p^2 des homozygotes) ; or la sélection agit sur les homozygotes.

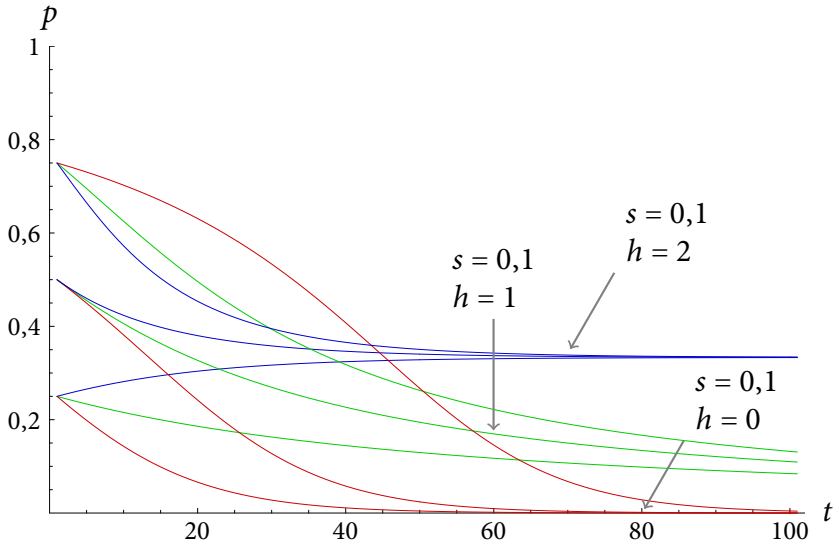


Fig. I.1 : Évolution de la proportion p de l'allèle A en fonction du temps (générations) avec les valeurs $s = 0{,}1$ et $h = 0$ (courbes rouges), $s = 0{,}1$ et $h = 1$ (courbes vertes), $s = 0{,}1$ et $h = 2$ (courbes bleues). Les simulations partent de trois états initiaux $p_0 = 1/4$, $p_0 = 1/2$ et $p_0 = 3/4$.

Les dynamiques décrites précédemment sont fondées sur des arguments mendéléens. Elles rendent compte de deux caractéristiques majeures des populations biologiques : leur capacité à évoluer et celle à maintenir une diversité (surdominance).

I.1.2 Théorie de l'évolution neutre

Comparée à la théorie déterministe, la « théorie de l'évolution neutre », qui est l'œuvre principalement de Kimura, interprète l'évolution comme un phénomène avant tout stochastique. D'après cette théorie, la fixation d'un allèle plutôt qu'un autre est due aux fluctuations dans la taille de la progéniture. L'effet de ces fluctuations avait déjà été étudié par Fisher [56, 57] et Wright [199] mais ces auteurs ont conclu que dans les populations biologiques il devait être largement surpassé par la sélection naturelle.

Plus précisément, la théorie neutre de l'évolution est résumée par le principe (v) de *The neutral theory of molecular evolution*.

> (v) Selective elimination of definitely deleterious mutants and random fixation of selectively neutral or very slightly deleterious mutants occur far more fre-

> quently in evolution than positive Darwinian selection of definitely advantageous mutants. […] This leads us to an important principle for the neutral theory stating that *'the neutral mutants' are not the limit of selectively advantageous mutants but the limit of deleterious mutants when the effect of mutation on fitness becomes indefinitely small* [94, p. 113].

En d'autres termes, la théorie de l'évolution neutre, contredit le néodarwinisme sur l'importance de la sélection positive : le néodarwinisme en fait la règle quand la théorie de l'évolution neutre en fait l'exception. La théorie de l'évolution neutre postule qu'il existe essentiellement deux types de mutations :

— les mutations délétères qui sont purgées par la sélection naturelle (« sélection négative ») ;

— les mutations neutres qui se fixent ou sont éliminées par hasard.

Selon la théorie de l'évolution neutre, le schéma 1 (p. 16) doit être modifié en

$$\begin{matrix}\text{génotype}\searrow\\ \text{génotype}\longrightarrow\end{matrix}\ \text{phénotype}\longrightarrow\text{capacité reproductrice},\qquad(5)$$
$$\text{génotype}\nearrow$$

où les trois occurrences de « génotype » représentent un nombre considérable de génotypes compatibles avec un unique phénotype.

L'idée de la possible fixation de mutants neutres existe déjà dans l'article de Sueoka qui bâtit une théorie décrivant la variation de la composition AT-GC des génomes bactériens [168]. Même plus tôt, en 1932, Morgan dans son livre *The scientific basis of evolution* reconnaît :

> If the new mutant is neither more advantageous than the old character, nor less so, it may or may not replace the old character, depending partly on chance; but if the same mutation recurs again and again, it will most probably replace the original character [138, p. 132].

Et bien sûr les travaux de Kimura et Crow montrent également que la substitution par des mutants neutres a été envisagée avant 1968 [98] ; le modèle des allèles infinis (*infinite allele model*, 1964) est curieusement antérieur à la théorie neutre. Dans leur article, Kimura et Crow notent en introduction :

> It has sometimes been suggested that the wild-type allele is not a single entity, but rather a population of different isoalleles that are indistinguishable by any ordinary procedure. […] It is known that a single nucleotide substitution can have the most dratic consequences, but there are also mutations with very minute effects and there is the possibility that many are so small as to be undetectable.

Enfin, le taux de mutation évalué par Kimura en appliquant le principe de charge minimale[4] est beaucoup plus faible que celui déduit des données de séquençage [93]. Takahata remarque que la différence aurait pu suggérer que les substitutions ne sont pas régies par la sélection positive [97]. Finalement, ce sont les données expérimentales qui manquaient pour convertir ces présomptions en réelle théorie.

À l'époque, le contexte n'était pas favorable à la réception de l'hypothèse de Kimura, comme le montre la citation suivante, reprise par King et Jukes [100] :

> The consensus is that completely neutral genes or alleles must be very rare if they exist at all. To an evolutionary biologist, it therefore seems highly improbable that proteins, supposedly fully determined by genes, should have nonfunctional parts, that dormant genes should exist over periods of generations, or that molecules should change in a regular but nonadaptive way… [natural selection] is the composer of the genetic message, and DNA, RNA, enzymes, and other molecules in the system are successively its messengers [167].

King et Jukes remarquent très justement « Natural selection is the editor, rather than the composer, of the genetic message » : les modifications de l'ADN sont imposées par les processus physico-chimiques et biologiques, la sélection naturelle n'arrivant qu'ensuite pour donner ou non son assentiment [100]. Autrement dit, l'existence ou non de mutations neutres n'est pas liée à la possibilité pour la sélection naturelle de les produire mais bien à celle d'en ressentir les effets lorsqu'elles apparaissent.

I.1.3 Théorie de l'évolution quasi neutre

La théorie de l'évolution neutre rencontre quelques difficultés à rendre compte de certaines observations sur lesquelles nous reviendrons plus bas : la constance de l'horloge moléculaire et les données de polymorphisme. L'une des théories les plus sophistiquées pour expliquer ces observations est celle des « mutations légèrement délétères » (*slightly deleterious mutation theory*) développée par Ohta en 1973 et 1974 (voir les références [147–149]) et renommée plus tard en théorie de l'évolution quasi neutre (*nearly neutral evolution theory*). Il s'agit d'une version de la théorie neutre augmentée de mutations légèrement délétères (voir figure I.2).

4. Nous décrivons le concept de « charge génétique » plus bas.

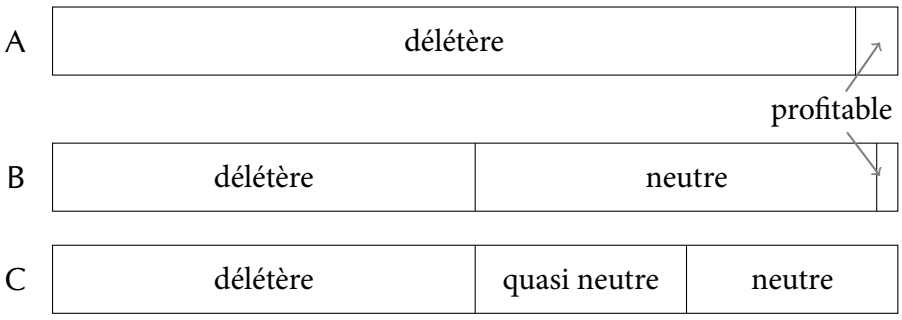


Fig. I.2 : Proportion des mutations délétères, neutres, quasi neutres et favorables selon le point de vue darwiniste ou pansélectionniste (A), neutraliste (B), ou enfin selon la thérie des mutations quasi neutres (légèrement délétères) d'Ohta (C).

I.2 Tempo de l'évolution

Les théories de l'évolution que nous venons de présenter assignent une cause différente à l'évolution des gènes : d'une part, la sélection positive et, d'autre part, le hasard (la « dérive génétique »). Ces causes ne prédisent pas les mêmes vitesses d'évolution. La découverte de l'« horloge moléculaire », une vitesse d'évolution constante pour chaque protéine, est très simplement expliquée par la théorie de l'évolution neutre. Nous verrons dans notre thèse comment l'intégration d'une information structurale peut également affecter le rythme auquel se succèdent les mutations. Cette section discute donc du tempo de l'évolution. Elle introduit les notions de « charge génétique », de « dérive génétique » et d'« horloge moléculaire ». Ces notions sont décrites dans un certain détail, même si elles sont peu exploitées dans la suite de la thèse. Enfin, nous considérerons quelques faiblesses de la théorie neutre.

I.2.1 Paradoxe d'Haldane

Les taux d'évolution prédits par la théorie déterministe de l'évolution sont limités par la « charge génétique » associée aux substitutions. Ils sont nettement moins élevés que ce que les premières données expérimentales au milieu des années 1960 laissaient penser. Cette incohérence a pris le nom de « paradoxe d'Haldane ». Historiquement, c'est pour résoudre ce paradoxe que Kimura avança sa théorie de l'évolution neutre.

I.2.1.1 Charge génétique

L'évolution et le maintien de la diversité d'une espèce que nous avons mis en évidence dans la section « Théorie déterministe » ont un prix qui est quantifié par la « charge génétique ».

La surdominance est illustrée par le cas d'école de la chaîne β de l'hémoglobine humaine et de la mutation incomplètement récessive de l'acide glutamique (allèle Hb^A) en valine (allèle Hb^S) en position six. Un porteur homozygote Hb^S Hb^S souffre d'anémie falciforme. Dans un environnement paludéen, le porteur hétérozygote Hb^A Hb^S est cependant protégé du paludisme tandis que l'homozygote Hb^A Hb^A ne l'est pas. L'avantage sélectif de l'hétérozygote maintient dans un environnement paludéen un équilibre entre les deux allèles Hb^A et Hb^S. À chaque génération, une certaine proportion d'homozygotes est nécessairement produite, notamment par le croisement des hétérozygotes, ce qui cause un excès de mortalité due d'un côté à l'anémie falciforme et de l'autre à la susceptibilité au paludisme.

La « charge génétique » (*genetic load*, [199]) tente de quantifier l'impact de cet excès de mortalité. On la définit par

$$\mathrm{L} = \frac{w_{\mathrm{opt}} - \langle w \rangle}{w_{\mathrm{opt}}}, \tag{6}$$

où w_{opt} est le *fitness* optimal et où l'opérateur $\langle \cdot \rangle$ est la valeur moyennée sur la population. On interprète la charge génétique comme l'effort reproducteur supplémentaire à fournir par une espèce pour subsister, c'est-à-dire au « prix » que doit payer une espèce pour la survivance des individus inférieurs en son sein.

Plusieurs catégories de charge génétique ont été isolées : la charge mutationnelle (*mutational load*), résultant de l'élimination permanente des mutants délétères ; la charge substitutionnelle (*substitutional load*), provenant de la substitution du génotype résident par un nouveau génotype plus avantageux ; la charge ségrégationnelle (*segregational load*), intervenant dans le cas de la surdominance. La charge totale est alors définie comme la somme de ces composantes : $\mathrm{L} = \mathrm{L}_{\mathrm{mut}} + \mathrm{L}_{\mathrm{subst}} + \cdots$. Dans notre modèle la charge mutationnelle tirera son origine des mutations létales ne préservant pas la structure native ou une habilité minimale d'interaction entre deux protéines.

Si nous revenons au cas de la surdominance, avec les notations du tableau I.2, nous pouvons vérifier que la proportion d'équilibre p_{eq} vaut $t/(s+t)$. Dans ce cas, la charge génétique vaut $\mathrm{L} = s\,t/(s+t)$ qui est minimale en p_{eq} :

$$\left(\frac{\mathrm{d}\mathrm{L}}{\mathrm{d}p}\right)_{p_{\mathrm{eq}}} = 0 \tag{7a}$$

$$\left(\frac{\mathrm{d}^2\mathrm{L}}{\mathrm{d}p^2}\right)_{p_{\mathrm{eq}}} = 2\,(s+t)\,. \tag{7b}$$

En tout état de cause, la charge est non nulle ; pour $s = t = 0{,}1$, elle s'élève à $\mathrm{L} = 0{,}05$.

La coïncidence que la charge est minimisée à l'état stationnaire dans le cas de la surdominance a pu suggérer que, mécaniquement, la sélection naturelle œuvrerait de façon à diminuer la charge génétique de l'espèce. C'est l'hypothèse que Kimura baptisa « principe de charge minimale ». Il mit en œuvre ce principe pour calculer un taux de mutation optimal qui minimisât la charge découlant de la mutation permanente et de la nécessité de s'adapter à un environnement changeant [93]. Cependant, ce principe n'est pas vrai dans l'absolu. Dans la controverse les opposant aux neutralistes, les darwinistes ont proposé des modèles dans lesquels le *fitness* dépend des fréquences alléliques ; dans ces modèles, la charge à l'équilibre n'est pas minimale.

Génotype	AA	Aa	aa
Fitness relatif	$1-s$	1	$1-t$
Proportion (τ)	p^2	$2\,p\,q$	q^2
Quantité de zygotes	$(1-s)\,p^2$	$2\,p\,q$	$(1-t)\,q^2$

Tab. I.2 : Tableau récapitulant à l'instant τ, les fréquences et le *fitness* des génotypes considérés dans le cas de surdominance ($s > 0$, $t > 0$).

I.2.1.2 Charge associée à la substitution d'un allèle

Haldane publia en 1957 le calcul de la charge substitutionnelle occasionnée par la substitution d'un allèle par un autre, plus avantageux [75]. Le cas des organismes haploïdes est le plus simple, c'est celui que nous allons traiter. Les autres situations envisagées par Haldane ne sont que des variations sur ce même thème. Considérons un *locus* présentant deux allèles A et a de *fitness* respectifs 1 et $1-s$. Leurs fréquences sont notées p et q. Nous souhaitons calculer la charge découlant de la substitution de l'allèle résident a par l'allèle A. On suppose que la fréquence initiale p_0 de l'allèle A est faible. Dans un modèle de générations discrètes et non chevauchantes (cf. note 3, p. 19), l'évolution de la fraction q s'écrit : $q' = q\,w_a/\langle w\rangle$. D'où

$$q' - q = \Delta q = -\frac{s\,q\,(1-q)}{1-s\,q}.$$

Dans un modèle continu de générations chevauchantes, où la reproduction des individus n'est plus synchronisée, l'évolution de q est donnée par la formule analogue

$$\frac{\mathrm{d}q}{\mathrm{d}t} = -\frac{s\,q\,(1-q)}{1-s\,q}. \tag{8}$$

La charge génétique dL calculée entre les instants t et $t + \mathrm{d}t$ est donnée par

$$\mathrm{dL} = s\,q\,\mathrm{d}t. \tag{9}$$

En intégrant et en posant $q_0 = 1 - p_0$, nous arrivons à

$$\mathrm{L} = \int_0^\infty s\, q\, \mathrm{d}t = \int_0^{q_0} \frac{1 - s\, q}{1 - q}\, \mathrm{d}q = s\, q_0 - (1 - s)\, \log(1 - q_0)\,.$$

Lorsque s est petit, on a donc

$$\mathrm{L} \approx -\log p_0\,. \tag{10}$$

La propriété la plus frappante de l'équation 10 est que la charge substitutionnelle, dans ce cas, ne dépend pas de s (avec une erreur relative au plus de l'ordre de s).

Autrement dit, indépendamment de l'avantage sélectif du mutant par rapport au sauvage, la substitution se fait en payant un forfait qui vaut $-\log p_0$. Ce forfait est le « coût de la substitution ». Cette indépendance par rapport à s s'explique par la compensation de deux effets contraires. Si s diminue, la charge est certes affaiblie par la faible valeur de s (puisque la charge dL est proportionelle à s, équation 9). Mais le processus de substitution se fait d'autant plus lent (puisque la vitesse de disparition de l'allèle le moins favorable dq est proportionnelle à s, équation 8).

Haldane a démontré que cette indépendance valait pour d'autres modes reproductifs et d'autres ploïdies. Des estimations supposées raisonnables de p_0 et du coût que pouvait tolérer une espèce ont mené au « paradoxe d'Haldane » : dans l'espèce humaine, par exemple, une substitution d'allèlle ne pourrait intervenir que toutes les trois cents générations. Au moment de la parution de l'article d'Haldane, le paradoxe n'en était pas encore un, même s'il ne manquait pas de surprendre par la lenteur de l'évolution qu'il impliquait. Les contradictions sont apparues dans les années 1960 avec les techniques de séparation électrophorétique des macromolécules biologiques qui ont mis au jour un polymorphisme des protéines et une vitesse d'évolution incompatibles avec la limite proclamée par Haldane. C'est sur cette contradiction que se fonde la théorie de l'évolution neutre.

I.2.2 Dérive génétique

La « dérive génétique » (*random genetic drift*) est une conséquence de l'affaiblissement de l'une des hypothèses les plus courantes en génétique des populations : la population n'est plus supposée être infinie. Nous allons voir que, sous cette hypothèse, le taux de substitution est déterminé par le taux de mutation.

Si l'on considère une population de N individus au lieu d'une population infinie, les fréquences alléliques ne forment plus un continuum entre zéro et un, mais un ensemble discret de valeurs. En outre, les fluctuations stochastiques des fréquences alléliques qui s'évanouissaient pour des populations infinies écartent nécessairement le système de l'équilibre. Cette déviation occasionne d'ailleurs une charge génétique comme remarqué par

Kimura [94, p. 133] qui définit la « charge de dérive » par :

$$\mathrm{L}_{\mathrm{drift}} = \frac{w_\infty - \langle w_\mathrm{N} \rangle}{w_\infty},$$

où w_∞ est le *fitness* moyen atteint dans la limite $\mathrm{N} \to +\infty$ et $\langle w_\mathrm{N} \rangle$ est l'espérance du *fitness* dans une population de N individus.

Kimura procède à un calcul éloquent[5] démontrant que le destin d'un mutant n'est pas déterministe. La chaîne α de l'hémoglobine possède un taux de substitution de l'ordre de un acide aminé tous les milliards d'années par résidu. Étant donnée la longueur de la chaîne peptidique $\mathrm{L} = 141$, une substitution est réalisée tous les sept millions d'années. Dans cet intervalle, des mutations ont le temps d'apparaître. Supposons que le taux de mutation soit de 10^{-6} par génération et par position, en se limitant aux mutations non-synonymes. Supposons, en outre, qu'une génération s'étende sur deux années et que la taille de la population soit de 500 000. Alors le nombre de mutations apparues au niveau individuel est $3{,}5 \cdot 10^6$. Parmi ces mutations seule une parvient à se fixer. Compte tenu du nombre de mutations possibles dans le gène, on peut estimer qu'une mutation doit être produite environ $3{,}5 \cdot 10^3$ fois avant de se fixer [94, p. 101, 102]. D'ailleurs, cela implique que si la fixation est déterminée par un avantage sélectif, cet avantage doit être infime pour qu'un tel nombre d'occurrences soit nécessaire pour que la fixation arrive à terme.

Les fluctuations dans la taille de la descendance peuvent être formalisées par le « modèle de Wright-Fisher ». Dans ce modèle, on conçoit le peuplement de la génération $t + 1$ par un tirage aléatoire avec remise des individus de la génération t. Dans une population haploïde de taille N, un individu laisse un nombre de descendants n obéissant à une loi binomiale :

$$\Pr(n) = \binom{\mathrm{N}}{n} p^n \, (1-p)^{\mathrm{N}-n},$$

avec $p = 1/\mathrm{N}$. Lorsque la population devient suffisamment grande, la loi binomiale tend vers une loi de Poisson dont le paramètre est $p\,\mathrm{N}$ et vaut donc 1. La figure I.3 présente plusieurs trajectoires d'une population soumise à la dérive génétique.

Supposons que dans cette population haploïde, les N individus soient porteurs chacun d'un allèle différent A_1, A_2, etc., aucun n'étant avantageux par rapport aux autres. Si l'on néglige l'effet des mutations, il paraît intuitif que la variabilité génétique diminue : en effet, si un allèle est perdu parce qu'il ne laisse aucun descendant, il ne saurait réapparaître. Les états où un allèle est fixé, c'est-à-dire occupe l'entièreté de la population, et où un allèle disparaît sont des états dits « absorbants », ce sont des points de non retour. Notons G_t

5. Quoique l'on puisse lui reprocher une surestimation du taux de mutation.

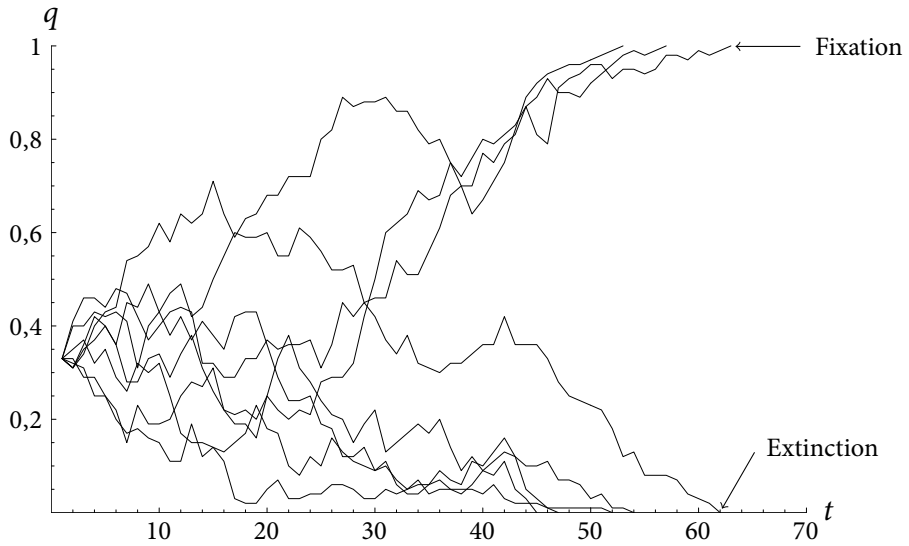


Fig. I.3 : Une population haploïde composée de N = 100 individus est soumise à la dérive génétique par implémentation du modèle de Wright-Fisher. Un gène existe sous la forme de deux allèles neutres A et *a*, initialement en proportion 2/3 et 1/3 respectivement. Nous représentons la fréquence q de l'allèle *a* en fonction du temps (en générations). Trois trajectoires mènent à la fixation de l'allèle le plus rare et cinq autres mènent à son extinction.

la probabilité que deux individus choisis au hasard dans la génération t aient un ancêtre commun dans l'une des générations 0, 1, …, $t-1$.

Avec une probabilité 1/N, deux individus ont un ancêtre commun à la génération $t-1$ (voir les filiations bleues dans la figure I.4). La probabilité que deux individus aient un ancêtre commun dans l'une des générations 0, 1, …, $t-2$ est $(1-1/\mathrm{N})\,\mathrm{G}_{t-1}$ (voir les filiations rouges dans la figure I.4). On a donc la récurrence :

$$\mathrm{G}_t = 1/\mathrm{N} + (1-1/\mathrm{N})\,\mathrm{G}_{t-1}, \tag{11}$$

ce qui mène à

$$1-\mathrm{G}_t = (1-1/\mathrm{N})^t\,(1-\mathrm{G}_0). \tag{12}$$

G_t tend vers un quand t tend vers l'infini. Autrement dit, la probabilité que deux individus ne descendent pas d'un même individu diminue d'un facteur 1/N à chaque génération et après un nombre suffisant de générations, la population est génétiquement homogène.

Puisque le processus est parfaitement symétrique du point de vue de tous les allèles A_i, la fixation d'un allèle est aussi vraisemblable que celle de tout autre et se produit avec une probabilité 1/N. L'envahissement d'une population résidente par un mutant sélectivement neutre peut se concevoir avec $A_1 \neq A_2 = \cdots = A_\mathrm{N}$.

Supposons maintenant que des mutations soient produites à un taux μ dans un gène. Nous supposons, en outre, que toute mutation génère un nouvel allèle (*infinite allele model*,

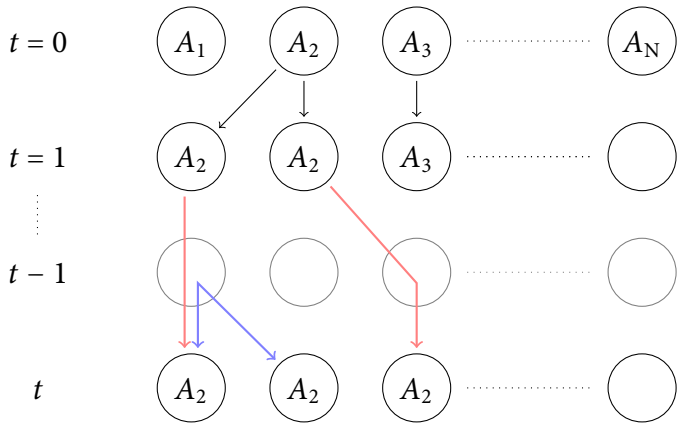


Fig. I.4 : Modèle de peuplement par tirage aléatoire avec remise (modèle de Wright-Fisher). Le lien de parenté entre la génération t et $t+1$ est obtenu en tirant au hasard pour chaque individu de la génération $t+1$ un parent dans la génération t.

[98]). Cela modifie l'équation 11 qui devient

$$\begin{aligned} G_t &= (1-\mu)^2 \left(1/N + (1-1/N)\, G_{t-1}\right), \\ &\approx (1-2\,\mu) \left(1/N + (1-1/N)\, G_{t-1}\right). \end{aligned} \tag{13}$$

Le terme $(1-\mu)^2$ vient de la nécessité qu'aucune mutation ne doive avoir affecté les deux individus choisis au hasard. À l'équilibre, $\hat{G} = G_t = G_{t-1}$, donc

$$\hat{G} \approx \frac{1}{1+2\,\mu\,N}. \tag{14}$$

Dans le cas diploïde, des propriétés analogues peuvent être déduites. La probabilité qu'un individu tiré au hasard soit homozygote, G'_t, tend vers un. La vitesse de convergence de la suite G'_t vaut $1/(2\,N)$. L'introduction d'un taux de mutation μ mène à

$$\hat{G}' \approx \frac{1}{1+4\,\mu\,N}. \tag{15}$$

La probabilité d'être hétérozygote se nomme l'« hétérozygotie ». Elle vaut $\hat{H}' = 1 - \hat{G}'$:

$$\hat{H}' \approx \frac{4\,\mu\,N}{1+4\,\mu\,N}. \tag{16}$$

Une population soumise à la dérive génétique devient presque certainement homogène du point de vue génétique. Tout individu de la génération $t = 0$ a autant de chance qu'un autre de fixer l'un de ses deux allèles (dans un organisme diploïde). Le nombre de mutants produits par génération vaut $2\,\mu\,N$, car chaque individu dispose de deux copies de chaque gène. Chaque mutant a une probabilité de se fixer $1/(2\,N)$. Nous trouvons le

taux de substitution k :

$$k = 2\,\mu\,N/(2\,N) = \mu\,.$$

Si μ ne dépend pas de la taille de la population, le taux de substitution ne dépend pas de la population.

Le taux de substitution d'une protéine étant égale à son taux de mutation neutre, la vitesse d'évolution d'une protéine est déterminée à l'échelle microscopique par sa capacité à tolérer des mutations. Cette capacité dépend principalement de la stabilité, des contraintes structurales et fonctionnelles régissant une protéine. Autrement dit, les détails microscopiques ont une répercussion macroscopique.

I.2.3 Horloge moléculaire

Les premières données des taux d'évolution des protéines ont donné naissance à l'hypothèse de l'« horloge moléculaire », à l'instigation de Zuckerkandl et Pauling, qui stipule que le nombre de substitutions de résidu d'acide aminé affectant une protéine est proportionnel au temps de divergence entre deux espèces [41, 119, 210] (voir les données recueillies dans la table I.3).

Protéine	*Taux de substitution*
Fibrinopeptides	8,3
Insuline C	2,4
Ribonucléase	2,1
Lysozyme	2,0
Hémoglobine	1,0
Myoglobine	0,9
Insuline A et B	0,4
Cytochrome C	0,3
Histone H4	0,01

Tab. I.3 : Taux de substitution (par milliard d'années et par site) de quelques protéines classées par vitesse de substitution (d'après [122]).

Ces données ne sont pas compatibles avec l'estimation de la vitesse de l'évolution faite par Haldane. En effet, l'étude comparative de diverses protéines comme les hémoglobines, le cytochrome *c*, etc., amène à considérer que dans une protéine de cent acides aminés, un acide aminé est substitué en gros tous les dix millions d'années chez les mammifères. Ce taux extrapolé au génome entier implique la substitution d'un nucléotide tous les deux ans environ, une période bien inférieure aux trois cents générations préconisées par

Haldane. En utilisant un modèle approché, Kimura tira les formules suivantes dans une population diploïde *finie* de taille N, pour un avantage sélectif s petit ($|\mathrm{N}\,s| \ll 1$)

$$\mathrm{L} = -4\,\mathrm{N}\,s \log p_0\,, \tag{17a}$$

$$u = p_0 + 2\,\mathrm{N}\,s\,p_0\,(1 - p_0)\,, \tag{17b}$$

où L est la charge de la substition, u la probabilité de fixation de l'allèle le plus favorable.

Notamment, quand l'avantage sélectif s tend vers zéro, la charge peut devenir infiniment petite et ce faisant la probabilité de fixation devient égale à p_0. La caractérisation mathématique des mutations neutres est donc

$$|\mathrm{N}\,s| \ll 1\,. \tag{18}$$

Si μ désigne le taux de mutation neutre d'un gène et N la taille de la population d'une espèce diploïde, le nombre de mutants neutres produits par génération est $2\,\mu\,\mathrm{N}$, car chaque individu possède deux copies de chaque gène. D'après l'équation 17b, chacun de ces mutants a une probabilité $p_0 = 1/(2\,\mathrm{N})$ de se fixer. Donc le taux de substitution des mutants dans une population, c'est-à-dire la vitesse à laquelle les fixations se succèdent, vaut $2\,\mu\,\mathrm{N}/(2\,\mathrm{N}) = \mu$. *Le taux de substitution k des mutants dans la population est égal au taux de mutation neutre* μ :

$$k = \mu\,. \tag{19}$$

Bien qu'elle soit nulle par définition d'une mutation neutre, la charge occasionnée par la substitution d'un allèle neutre ne peut être calculée dans le modèle d'Haldane. Ce dernier fait l'hypothèse d'une population infinie. Par conséquent, la fixation ne peut jamais être atteinte : un allèle neutre dans une population infinie a une fréquence constante d'après l'équation 2 (p. 19). Dans une population finie, en revanche, les fluctuations stochastiques dans la production de descendants permettent de mener à la fixation, par dérive génétique.

L'interprétation neutraliste de l'horloge est donc extrêmement simple : le taux de mutation neutre est constant pour une protéine et quantifie l'intensité de la sélection négative. On peut réécrire l'équation 19 comme

$$k = f_0\,\mu_\mathrm{T}\,. \tag{19′}$$

où μ_T est le taux de mutation total et f_0 la fraction des mutations qui sont neutres. La théorie de l'évolution postule que la fraction f_0 ne dépend que de la contrainte fonctionnelle agissant sur une protéine.

Des méthodes approchées, appelées « modèles diffusifs » ont été employées pour connaître la dynamique de l'évolution d'une population. Les temps de fixation et d'élimination des mutants sont les quantités biologiquement les plus importantes. Une fixation nécessite en moyenne $4\,N$ générations pour aboutir, l'extinction en revanche est un processus sensiblement plus rapide puisqu'il en requiert $2\log(2\,N)$ [99]. L'ensemble de ces données permet de dresser le tableau suivant : les subtitutions de mutants se succèdent avec une période de $1/\mu$ et chaque substitution requiert de l'ordre de $4\,N$ générations pour arriver à son terme (cf. figure I.5).

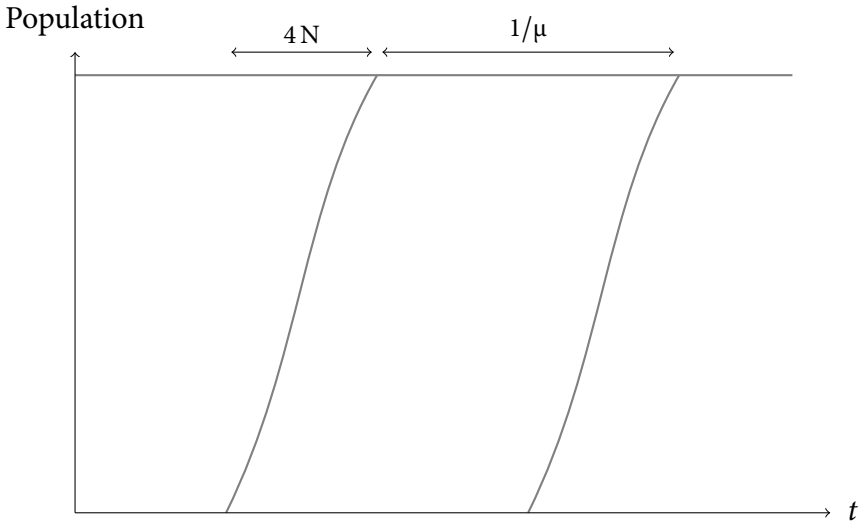


Fig. I.5 : Schéma de la dynamique de substitution des allèles dans une population en fonction du temps (générations). La fixation d'un nouvel allèle se produit en moyenne une fois par intervalle de temps égal à $1/\mu$, où μ est le taux de mutation neutre. La fixation elle-même nécessite $4\,N$ générations pour arriver à terme.

I.2.4 Faiblesses de l'évolution neutre

La simplicité était la qualité qu'appréciait le plus Kimura dans sa théorie de l'évolution neutre : la théorie de l'évolution neutre explique beaucoup tout en restant concise. Si elle a suscité une controverse soutenue, c'est premièrement que le néodarwinisme était profondément ancré dans les esprits, deuxièmement parce qu'elle souffrait de quelques faiblesses.

I.2.4.1 Révision du paradoxe d'Haldane

La première faiblesse est la révision du paradoxe d'Haldane, lequel repose sur la *charge génétique* dont la définition est arbitraire. Wright, celui-là même qui avait introduit le concept, reconnut

> There seems to be no general biological meaning of load as defined above other than the percentage difference of the mean selective values of the population in question from that of some reference genotype [201].

En particulier, le génotype de référence peut ou peut ne pas exister ; si sa probabilité d'occurrence est infinitésimale, calculer la charge par rapport à lui est illusoire, ou pour le moins risqué. Ewens en considérant un génotype de référence ayant une probabilité raisonnable d'apparaître dans une population finie arriva à concilier six substitutions par génération pour une charge substitutionnelle de 0,5. Le même calcul autorise une substitution toutes les seize générations sans dépasser une charge de 0,1 [50, p. 83, 84].

La deuxième faille du raisonnement de Kimura dans son article de 1968 est qu'il extrapola les substitutions à l'intégralité du génome alors que seule une fraction du génome est effectivement exprimée : par exemple, les exons représentent 1,1 % du génome de l'Homme [187] (1,5 % d'après [105]). Ce n'est donc pas l'argument original de Kimura qui assura la pérennité de sa théorie. D'autres données furent nécessaires à son établissement.

I.2.4.2 Dispersion de l'horloge moléculaire

La seconde difficulté rencontrée par l'évolution neutre est la capacité de la théorie neutre à expliquer l'horloge moléculaire. D'après cette théorie, l'accumulation de mutations devrait obéir à une loi de Poisson. Mais l'horloge est plus erratique que cela. La déviance par rapport à la loi de Poisson est mesurée par l'index de dispersion R, le rapport de la variance observée et la moyenne (cf. table I.4).

Protéine	*Nombre d'espèces*	*Nombre de substitutions moyen* M	*Variance* S^2	$R = S^2/M$
Hémoglobine α	6	13,15	18,30	1,39
Hémoglobine β	6	15,61	54,19	3,47*
Myoglobine	6	12,77	23,83	1,87
Cytochrome *c*	4	8,55	30,92	3,62*
Ribonucléase	4	21,99	62,68	2,85

Tab. I.4 : Taux de substitution moyen chez les mammifères (adapté de la référence [122, p. 151], d'après Gillespie [63] et Kimura [94]). L'index de dispersion R vaut un pour un processus parfaitement poissonien. Les astérisques indiquent une valeur significativement supérieure à un.

Un certain nombre de théories ont été avancées pour contredire ou compléter la théorie neutre : Gillespie développa un modèle d'« horloge épisodique » [64], Takahata envisagea que certains paramètres pussent varier avec le temps, la contrainte sélective par

exemple [169]. Ce qui est remis en cause ce sont finalement les traits caricaturaux de la théorie de l'évolution. Gillespie envisage lui aussi les complications prises en considération par Takahata :

> A complication of this model involves a fundamental issue as to whether the rates of evolution that remain nearly constant for long periods of time are better represented as random quantities chosen from a common probability distribution […] [64].

C'est dans cette brèche que la description au niveau moléculaire peut devenir cruciale [6, 7].

Un autre aspect de l'horloge moléculaire sur lequel la théorie de l'évolution neutre n'est pas pleinement satisfaisante, est que la constance de l'horloge moléculaire est plus manifeste lorsque le taux de subtitution est calculé par année que lorsqu'il l'est par génération. La théorie de l'évolution quasi neutre a tenté de répondre à cette défaillance [147,148]. Mais c'est peut-être le taux de mutation qu'il faut reconsidérer comme l'envisagea Kimura

> On the other hand, it is likely that errors in DNA replication and repair are the main causes of DNA changes that are responsible for molecular evolution. Thus, the mutation rate for nucleotide substitutions may depend on the number of cell divisions in the germ lines, particularly in the male line, and this will make the molecular mutation rate roughly proportional to years. Experimental studies on this subject are much needed [96].

Kumar et Subramanian qui proposent eux que des processus *indépendants de la division cellulaire* (méthylation et réparation de l'ADN, recombinaison) puissent être impliqués dans la majeure partie des mutations observées [104]. Il semble néanmoins qu'un consensus ne soit pas encore prêt d'émerger à ce sujet [192].

I.2.4.3 Polymorphisme protéique

L'évolution neutre et le modèle des allèles infinis prédisent que l'hétérozygotie présentée par une population dépend de sa taille (cf. équation 16). Les espèces nombreuses devraient présenter une diversité allélique à un *locus* plus élevée que les espèces peu nombreuses. Or les taux observés expérimentalement vont de 6 % à 18 % et ce plus ou moins indépendamment de la taille de la population. Par la théorie de l'évolution quasi neutre, Ohta tenta de rendre compte de ces valeurs. Dans les populations faibles, la dérive génétique pourrait permettre d'entretenir une proportion importante de mutations légèrement délètères. Dans les populations vastes, la sélection serait plus stringente et éliminerait ces mêmes mutations [148]. Une autre explication est proposée par Kimura [94] et Nei [142] selon laquelle l'hétérozygotie pourrait être un vestige des périodes où les espèces auraient été moins peuplées.

I.3 Contraintes fonctionnelles et sélection négative

Comme le dit Kimura dans la citation de la page 21, les mutations neutres sont la limite des mutations délétères lorsque leur effet devient infime et imperceptible. La vitesse d'évolution reflète donc, d'après la théorie de l'évolution neutre, le degré de *non-incidence* des mutations. Les acides aminés à la surface d'une protéine, par exemple, sont généralement moins conservés que ceux du cœur [2]. L'application du *protein design* aboutit à la même conclusion [85]. Nous faisons un inventaire de quelques observations confortant la théorie de l'évolution neutre et son acceptation comme hypothèse nulle dans les tests statistiques.

I.3.1 Mutations synonymes et non-synonymes

Malgré les imperfections que nous venons de mentionner, la théorie de l'évolution neutre semble essentiellement correcte. L'association de taux de substitutions et de l'absence de contrainte est l'idée qui a reçu le plus de confirmations expérimentales, à commencer par l'article *Non-darwinian evolution* de King et Jukes [100]. D'une part, une fois découvert le code génétique, on a pu observer que les mutations synonymes affectant la troisième position des codons sont nettement plus fréquentes que les mutations non-synonymes (cf. table I.5).

	r_1	r_2	r_3	b
Souris α3	0,69	0,69	3,32	5,0
Souris α1	0,74	0,67	2,51	5,1
Lapin β2	0,71	0,51	2,09	3,6
Moyenne	0,71	0,62	2,64	4,6

Tab. I.5 : Taux de substitution nucléotidique des première (r_1), seconde (r_2) et troisième (r_3) positions des codons de trois globines de mammifères, ainsi que le taux de substitution de leur pseudogène correspondant b (d'après la référence [116]). Les taux sont donnés par milliard d'années.

I.3.2 Pseudogènes

La découverte des pseudogènes fut d'une grande importance dans l'affirmation de la théorie de l'évolution neutre [116, 134]. Ces parties du génome ne sont soumises à aucune contrainte et évoluent effectivement à une vitesse supérieure à celle des gènes fonctionnels. Les pseudogènes possèdent des taux de substitution égaux aux trois positions des codons

avec pour valeur typique $k = 5 \cdot 10^{-9}$. Cette valeur est environ deux fois le taux de substitution synonyme d'un gène de globine (cf. table I.5)[6].

I.3.3 Conservation au niveau moléculaire

Au niveau moléculaire, des données plus fonctionnelles ont également permis de conforter la relation entre les taux d'évolution et la contrainte. Une illustration intéressante est donnée par le segment intermédiaire C de la proinsuline. Ce segment est en effet clivé pour donner naissance aux peptides A et B de l'insuline active. Il est probablement moins important pour le fonctionnement de l'hormone que ne le sont les autres segments A et B. Le taux de substitution du segment C s'élève à $2{,}4 \cdot 10^{-9}$ substitution par résidu et par année, tandis que ceux des segments A et B valent $0{,}4 \cdot 10^{-9}$ substitution par résidu et par année [89]. La poche de l'hémoglobine liant l'hème donne lieu à la conclusion complémentaire : les mutations y sont nettement moins fréquentes [153].

Différents auteurs ont développé et utilisé des mesures de similitude ou de dissimilitude biochimique entre acides aminés pour déchiffrer les mécanismes de substitution : Epstein en 1967, Clarke en 1970, Miyata *et al.* en 1979 [133]. Kimura réutilisa la distance de Miyata pour mettre en évidence que la substitution d'un acide aminé par un autre est d'autant plus fréquente que ces deux acides aminés sont semblables [94, p. 152]. En d'autres termes, une substitution est d'autant plus susceptible d'avoir lieu qu'elle implique peu de modifications biochimiques. En vertu de l'égalité 19 (p. 32), l'interprétation neutraliste reformule cela en : une *mutation* est d'autant plus susceptible *d'être neutre du point de vue de la sélection naturelle* qu'elle implique peu de modifications biochimiques.

Les querelles suscitées par la proposition de la théorie de l'évolution neutre a donné lieu à l'invention de nombreux tests statistiques dont on trouvera une revue dans l'ouvrage de Kimura [94] et qui ont dans l'ensemble confirmé l'exactitude de la théorie neutre (cf. référence [142] et références citées). La théorie de l'évolution neutre constitue à présent l'hypothèse nulle de ces tests dont la fonction est de mettre en évidence une sélection positive. Ces indicateurs de sélection positive reposent le plus souvent sur la comparaison des taux de mutation synonyme et non-synonyme (voir par exemple [83]).

6. Puisque le taux de substitution synonyme n'est pas égal aux taux de substitution maximal, il semblerait que les mutations synonymes ne soient pas strictement neutres. Leur incidence sur le *fitness* peut être due à l'utilisation préférentielle des codons, aux mécanismes d'épissage ou encore à la stabilité des ARNm [27].

I.4 Évolution à l'échelle moléculaire

Dans notre introduction, nous avons souligné les atouts autant que les faiblesses de la théorie de l'évolution neutre. La grande qualité de l'évolution neutre est, d'après Kimura, sa faculté à expliquer beaucoup en étant extrêmement simple. Kimura reconnaissait l'importance de la sélection positive dans l'évolution des phénotypes. Mais sa théorie est incapable de l'expliquer : la théorie de l'évolution est fondamentalement une description de l'évolution moléculaire à l'échelle du génotype. Ses faiblesses montrent, en outre, que des études sont encore nécessaires pour percer les mécanismes de l'évolution. Ces études peuvent être expérimentales (par exemple la mesure des taux de mutation par gamète pour expliquer la constance de l'horloge moléculaire par année plutôt que par génération), théoriques ou computationnelles. Cette dernière approche est justifiée par une connaissance sans cesse croissante des déterminants de la structure et de la fonction des macromolécules biologiques.

Des modèles simples comme ceux que nous allons introduire dans les paragraphes qui suivent, permettent d'élucider les traits caractéristiques de modèles d'évolution. Le faible nombre de paramètres garantit une interprétation évidente des résultats. Les données qu'ils fournissent sont aussi complètes et dépourvues d'ambiguïté. Une application naturelle de ce constat simple est la validation de méthodes d'analyse rétrospective : voir les références [76, 77] et également l'anecdote rapportée par Kimura [94, p. 88].

I.4.1 Du génotype au *fitness*

I.4.1.1 Correspondance génotype-phénotype-*fitness*

L'idée d'« espace des génotypes » prend corps la première fois avec le *fitness landscape* de Wright en 1932 [200]. Il imagina que l'évolution pouvait être vue comme un parcours tendant vers les cimes sur un paysage vallonné dont les altitudes représentent les potentiels reproductifs de chaque génotype. Ici, le génotype n'est pas limité à un seul *locus* mais est formé de l'ensemble de la composition allélique du génome. L'évolution et en particulier le jeu de l'évolution adaptative et de l'évolution neutre dépend de manière décisive de l'aspect du paysage de *fitness*.

1. Il est donc important de connaître comment sa rugosité affecte le parcours évolutif. La famille des modèles N K de Kauffman a été une tentative de formaliser l'aspect de ce paysage dans le cadre de l'interaction épistatique de K gènes parmi N existant sous deux formes alléliques [91]. Ce modèle décline toute une variété de paysages de *fitness*, de parfaitement lisse à complètement aléatoire.

2. Il est également essentiel de connaître quel est le paysage de *fitness* qui émerge de modèles structuraux et fonctionnels des macromolécules.

Maynard Smith introduisit en 1970 le concept voisin d'« espace des protéines » [121]. Il défend l'idée que l'évolution peut être efficace en dépit de l'immensité de l'ensemble des séquences protéiques possibles (l'espace des protéines). D'après lui, il suffit que les séquences fonctionnelles soient *connectées*. La connexion entre deux séquences est assurée par les mutations ponctuelles et les autres événements mutationnels de l'ADN (insertions, délétions, recombinaisons, etc.). L'ensemble de ces séquences interconnectées forme un réseau. Ce réseau prend le nom de « réseau neutre » en l'absence de sélection positive.

Un réseau neutre est donc un ensemble de séquences compatibles avec un phénotype donné dans lequel une population évolue en empruntant les arêtes reliant les séquences. La validation de la théorie proposée par Maynard Smith repose sur la mise en évidence de ces réseaux neutres. Même si Maynard Smith n'était pas un fervent partisan de la théorie de l'évolution neutre, la neutralité est explicitement admise dans son article [121]. Expérimentalement, les connexions d'un réseau neutre peuvent être observées. Par exemple, la mutation de la glycine 210 du site actif de la tryptophane-synthétase A d'*E. coli* en sérine ou alanine est pleinement fonctionnelle. Néanmoins, l'observation plus complète de réseaux neutres requiert des simulations *in silico*, plus particulièrement, s'il s'agit de connaître leurs traits communs.

Maynard Smith pose la question : est-il possible de passer d'une séquence de protéine fonctionnelle à une autre par des mutations ponctuelles ? Il compare cette éventualité à la possibilité de ponter des mots anglais entre eux. Par exemple, peut-on trouver une suite de mots anglais permettant de passer de « word » à « gene » en ne modifiant qu'une seule lettre à la fois ? Une solution possible à ce casse-tête est :

$$\text{word} \to \text{wore} \to \text{gore} \to \text{gone} \to \text{gene}.$$

Lorsque les données sont complètes, on peut répondre à des questions de plus grande envergure. Nous osons livrer quelques résultats de nos investigations. Combien de mots peuvent être atteints à partir de « word » ? Réponse : 1 617 (90 %) parmi les 1 778 mots de quatre lettres répertoriés dans le dictionnaire (une composante géante dans le vocabulaire de Schuster). Quels sont les mots accessibles en deux étapes (ou « mutations ») ? Réponse : soixante-dix mots[7]. En moyenne, combien de mots peuvent être atteints par modification d'une lettre ? Réponse : 7,3, l'écart type valant 4,8 (soit 7,3 % des $4\times(26-1) = 100$

7. La liste complète est la suivante : « bard », « bold », « bore », « born », « card », « coed », « cold », « cord », « core », « cork », « corn », « curd », « fold », « fond », « food », « ford », « fore », « fork », « form », « fort », « gold », « good », « gore », « hard », « hold », « hood », « horn », « lard », « load », « lord », « lore », « loud », « mold », « mood », « more », « morn », « norm », « pore », « pork », « rood », « sold », « sore », « told », « tore », « torn », « wand », « ward », « ware », « warm », « warn », « warp », « wart », « wary », « weld », « were », « wild », « wire »,

mutations possibles). Enfin, quels sont les mots les plus connectés ? Réponse : « bale » et « ware » qui tolèrent 23 mutations chacun[8]. Si la pertinence de ce jeu dans le champ de la linguistique reste à prouver, dans le champ de l'évolution des protéines, il ne fait aucun doute que des réponses de ce type sont cruciales dans l'analyse rétrospective (comment l'organisation et la régulation cellulaire ont-elles pu être façonnées [62] ?) comme dans la conception prospective (*protein design*).

La citation de Gillespie (p. 35) sous-tend également l'importance de la correspondance génotype-phénotype. Le taux de mutation neutre doit-il être supposé constant ? Et le cas échéant, quelle cause est à l'origine des fluctuations ? Faut-il recourir à des changements environnementaux ou existe-t-il une cause intrinsèque liée à la correspondance génotype-phénotype ? Supposons que le jeu des mots de quatre lettres de Maynard Smith possède quelque ressemblance avec les réseaux neutres biologiques. Alors toutes les séquences ne sont pas équivalentes du point de vue de leur fraction de mutations neutres f_0. Nous avons remarqué que la fraction de mutations neutres moyenne vaut 7,3 % avec un écart type de 4,8 % et que certaines séquences (« bale » et « ware ») tolèrent environ un quart de toutes les mutations ponctuelles.

I.4.1.2 Modèles de correspondance génotype-phénotype-*fitness*

Correspondance génotype-phénotype La correspondance génotype-phénotype que nous avons vue plus haut (équation 1, p. 16) est singulièrement modifiée par la théorie de l'évolution neutre : le nombre allèles compatibles avec une certaine fonction, enzymatique par exemple, est fantastiquement grand. On a une *dégénérescence* de la correspondance phénotype-*fitness* :

$$\begin{matrix}\text{génotype} \searrow \\ \text{génotype} \longrightarrow \text{phénotype} \longrightarrow \text{capacité reproductrice}. \\ \text{génotype} \nearrow\end{matrix} \tag{5}$$

Dans un cadre purement darwiniste, cette correspondance serait, en principe, moins cruciale : « To hold that selectively neutral isoalleles cannot occur is equivalent to maintaining that there is one and only one optimal form for every gene at any point in evolutionary time » [100].

« woke », « wold », « wolf », « wood », « wool », « word », « wore », « work », « worm », « worn », « wove », « yard » et « yore ».

8. « bale », par exemple, est connecté à « babe », « bade », « bake », « bald », « balk », « ball », « balm », « bane », « bare », « base », « bate », « bile », « bole », « dale », « gale », « hale », « kale », « male », « pale », « sale », « tale », « vale » et « wale ». On observe que les lettres *a* et *e* en deuxième et quatrième positions sont très conservées.

Dans une perspective de sélection centrée sur les gènes, l'espace des génotypes est un espace gigantesque mais relativement simple puisqu'il est constitué des combinaisons de $n = 4$ lettres (pour les acides nucléiques) ou $n = 20$ lettres (pour les séquences peptidiques). L'espace des phénotypes est plus réduit mais de description nettement plus complexe : il peut être l'ensemble des repliements secondaires de ARN, ou des repliements protéiques possédant une activité catalytique.

Modèles d'ARN Par rapport aux études structurales ou fonctionnelles classiques, la prise en compte de l'espace des génotypes (ou espace des séquences) ajoute un niveau de complexité. Pour explorer cet espace *in silico*, le recours à des modèles simplifiés des macromolécules biologiques devient nécessaire. Les modèles d'évolution des ARNt, par exemple, identifient le phénotype, ou la fonction, d'une molécule à sa structure secondaire [84, 162, 184]. En effet, une partie substantielle de l'énergie libre de la structure de l'ARN est contenue dans la structure secondaire. De plus, les structures secondaires sont conservées au cours de l'évolution. D'un point de vue pratique, le calcul de la structure secondaire peut se faire en temps polynomial par programmation dynamique [146].

Modèles de protéine La modélisation des protéines fait abondamment usage des modèles dits sur réseau. Les résidus d'une chaîne peptidique sont des perles placées sur une grille bidimensionnelle ou tridimensionnelle. Le phénotype est identifié à une ou plusieurs propriétés structurales (stabilité, etc.). Des modèles tridimensionnels hors-réseau ont été occasionnellement développés (cf., par exemple, [7, 40]). Aussi, présenterons-nous dans ce travail deux modèles : un modèle de protéine sur réseau et un modèle tridimensionnel hors-réseau. Cette stratégie permettra d'explorer des propriétés nécessitant des calculs exhaustifs et de se prémunir des particularités des modèles sur réseau.

L'évaluation de la capacité d'une protéine à se replier est un problème NP-complet [9]. Il est donc nécessaire de se résoudre à des approximations. Les modèles sur réseau permettent d'explorer intensivement voire exhaustivement l'espace des séquences et l'espace des structures [204]. Pour les modèles de protéines tridimensionnelles hors-réseau, plus réalistes, l'échantillonnage est plus limité. Des méthodes heuristiques ont été développées pour estimer rapidement la stabilité structurale. En particulier, les matrices d'interaction entre paires de résidus ont connu un grand développement [5, 101] et sont extrêmement utiles dans les simulations d'évolution [7] ou dans la prédiction d'arbres phylogénétiques par exemple [157].

Correspondance phénotype-*fitness* L'espace des *fitness* est simplement la demi-droite réelle $[0, +\infty[$. La manière dont se projette l'espace des phénotypes dans l'espace des *fitness* n'est certainement pas aisée à déterminer et donne lieu à des choix plus ou moins arbitraires dans les simulations menées jusqu'à présent. Williams *et al.* proposèrent un

fitness fondé sur la vitesse de catalyse d'une réaction michaelienne [197] ou simplement sur la probabilité de fixation d'un ligand [196]. Ce choix est assez proche de celui fait par Bloom *et al.* [18], tandis que Blackburne et Hirst identifièrent le *fitness* au nombre de résidus hydrophobes formant une poche capable d'accueillir un ligand. Fontana et Schuster [59] définirent le *fitness* par la proximité à une structure secondaire d'ARNt.

L'application extrême de la théorie neutraliste (*pan-neutralist theory*) cependant permet de faire tomber la difficulté tout en évitant l'introduction d'une fonction arbitraire : à un phénotype particulier est associé un *fitness* relatif w, à tous les autres, zéro. Ce faisant, la correspondance génotype-*fitness* est une fonction indicatrice dont nous souhaitons déterminer l'ensemble sous-jacent. Les mutations neutres transformant un élément de cet ensemble en un autre confèrent une structure de graphe : le « réseau neutre ».

Structure des réseaux neutres Quelles que soient les approximations utilisées dans la modélisation des protéines, une vue unifiée de la topologie des réseaux neutres a pris forme. En particulier, ils s'organisent autour d'une « séquence prototype » qui coïncide essentiellement avec la séquence consensus construite par alignement des séquences compatibles avec un repliement [20]. Elle est caractérisée dans l'article de Bornberg-Bauer par sa robustesse élevée aux mutations[9].

Autour de la séquence prototype, le réseau neutre s'organise en couches successives : un petit groupe de séquences interconnectées, qui sont également très robustes aux mutations, forme une première couche, puis à mesure que l'on s'éloigne de la séquence prototype, le réseau devient plus épars, moins connecté, mais aussi plus nombreux. Cette organisation a été observée par Bornberg-Bauer par un calcul exhaustif sur des polymères HP de longueur dix-huit [20] et par Govindarajan et Goldstein qui observèrent que des contraintes plus stringentes ont tendance à générer des ensembles de séquences plus ramassés [68].

Cette structure en couches se reflète aussi dans la distribution de la stabilité des différentes séquences [21] (voir aussi l'étude de Xia et Levitt [202]). Les couches les plus proches de la séquence prototypes bénéficient d'une stabilité plus grande. Ce résultat a donné naissance au terme *superfunnel* parce que le réseau neutre a une forme d'entonnoir[10] qui potentiellement peut guider l'évolution, soit adaptative soit neutre comme nous allons le voir, vers les régions qui associent forte tolérance aux mutations et stabilité thermodynamique.

9. Nous adopterons dans cet ouvrage une définition quelque peu différente, plus appropriée, nous semble-t-il, aux simulations évolutives.

10. Le terme *superfunnel* a été proposé car les réseaux neutres possèdent une topologie évocatrice de la structure de *funnel* (entonnoir) mise en évidence dans l'étude du repliement des protéines.

I.4.2 Façonnage des séquences et des structures

La double correspondance génotype-phénotype-*fitness* ne constitue qu'un « support » sur lequel s'appuie l'évolution. En effet, elle caractérise le rapport entre les génotypes, la fonction et la capacité reproductrice, mais pas l'évolution elle-même. *L'évolution est l'apanage des populations.* La compréhension des comportements évolutifs, qu'ils soient neutres ou adaptatifs, impliquent de comprendre comment une population évolue, une fois définie la correspondance génotype-phénotype-*fitness*. Il s'agit donc de définir les mécanismes évolutifs élémentaires, microscopiques, à partir desquels s'élabore l'évolution macroscopique et se façonnent les séquences et les structures que nous observons.

Avant d'aborder les résultats spécifiques aux ARN et aux protéines, nous souhaitons mentionner un travail élégant ne reposant sur aucun modèle structural. Sella et Hirsh ont trouvé que la généralisation du processus de Moran à un graphe adaptatif de séquences peut être appréhendée par le formalisme de la physique statistique indépendamment de la topologie du graphe et donc indépendamment d'un modèle moléculaire [163]. Dans des situations plus complexes, cependant l'évolution dépend de la topologie des réseaux de séquences et requiert donc un modèle explicite de correspondance génotype-phénotype.

I.4.2.1 Évolution de modèles d'ARN

La représentation dynamique des populations sur un graphe fut utilisée en premier par Eigen. Il proposa une théorie d'évolution adaptative prébiotique expliquant l'émergence de quasi-espèces au sein d'une population d'ARN autocatalytiques [44]. Cette théorie est fondée sur la capacité des ARN à s'autorépliquer et dans une certaine mesure à générer des nouvelles séquences *via* des erreurs de réplication. La fraction de séquence s_i varie selon

$$\frac{\mathrm{d}s_i}{\mathrm{d}t} = (\mathrm{A}_i - \mathrm{D}_i)\, s_i + \sum_j \phi_{ij}\, s_j,$$

où A_i est la constante cinétique d'autoréplication, D est la constante cinétique de dégradation et ϕ_{ij} est la constante cinétique d'autoréplication fautive de la séquence j qui donne la séquence i. On peut donc caractériser la cinétique du système par un graphe orienté pondéré décrit par la matrice $\mathcal{A} = (a_{ij})$:

$$a_{ij} = \begin{cases} \mathrm{A}_i - \mathrm{D}_i & \text{si } i = j, \\ \phi_{ij} & \text{sinon.} \end{cases}$$

Le système converge vers un état où seuls quelques nuages, regroupés et isolés, de séquences existent, ce sont les « quasi-espèces ». Les quasi-espèces démontrent l'importance

du modèle évolutif. Premièrement, elles sont un phénomène émergeant de la dynamique et non seulement du graphe. Deuxièmement, elles enseignent que la représentation d'une séquence dépend fortement des propriétés de ses voisines dans le graphe. Ces deux propriétés apparaîtront également dans nos résultats.

Grâce aux modèles moléculaires, il est également possible de décrire comment évoluent les phénotypes. Les études de Fontana et Schuster ont apporté un éclaircissement important sur le jeu de l'évolution neutre et l'évolution adaptative dans la progression vers les cimes du paysage de *fitness* [59]. En utilisant un modèle d'ARN et en définissant le *fitness* comme une fonction décroissante de la distance structurale à une structure secondaire cible, ils ont montré que les phases d'évolution adaptative n'étaient que de courtes transitions entre de plus longues périodes d'exploration neutre. Conformément à ce que pensait Kimura, les mutations adaptatives sont rares et le passage de l'une à l'autre ne peut se faire que si une évolution neutre les relie. Le résultat de Fontana et Schuster offre un fondement moléculaire à l'hypothèse de l'« équilibre ponctué », introduit dans les années 1970 [46,67]. L'observation analogue existe pour les protéines, voir référence [196].

La capacité de l'évolution neutre à découvrir de nouveaux phénotypes a été étudiée plus profondément par Huynen [84] : il a mesuré le nombre de phénotypes qui pouvaient être découverts au contact d'un réseau neutre. Il a montré que le ce nombre croissait linéairement avec le temps d'évolution et n'était que faiblement réduit (de moitié environ) par rapport à une évolution non confinée à l'intérieur d'un réseau neutre.

I.4.2.2 Évolution de modèles de protéine

Modèles à structure contrainte Bastolla et ses collaborateurs ont pu mettre en évidence une cause intrinsèque de la dispersion de l'horloge moléculaire. Ils calculèrent, à l'aide d'un modèle tridimensionnel de protéine et un potentiel d'énergie simplifié, une correspondance génotype-phénotype fondée sur la stabilité structurale. Ils purent observer des disparités dans la connectivité du réseau neutre. Cette disparité conduit les populations à être gelées dans certaines régions du réseau neutre et à évoluer rapidement dans les régions densément connectées [7] conférant un caractère épisodique à l'horloge moléculaire : des phases d'évolution rapide alternent avec des périodes d'évolution plus lente.

L'une des découvertes les plus surprenantes est que l'évolution au sein d'un réseau neutre peut présenter un caractère adaptatif. Ce phénomène a été constaté sous l'effet de mutations ponctuelles seules [21, 173] ou de mutations ponctuelles associées à des recombinaisons [203]. Taverna et Golstein ont comparé l'évolution d'un unique individu errant dans un réseau neutre sous l'effet de mutations ponctuelles et celle d'une *population*. Ces auteurs ont découverts qu'une localisation préférentielle résulte d'un effet de dynamique de la population. La population se concentre au centre des réseaux neutres, là où

les génotypes tolèrent le plus de mutations. Cela démontre l'importance quantitative de la robustesse aux mutations dans le phénomène évolutif. Cet effet a été nommé *survival of the flattest*, par opposition au classique *survival of the fittest*, par Wilke [193]. Il engendre aussi un comportement adaptatif que Wilke a appelé *neutral staircase* [194].

Évolution des structures Il est également possible d'envisager l'évolution des structures. Un déterminant essentiel semble être la « modifiabilité » qui est le nombre de séquences compatibles avec une certaine conformation [180]. L'estimation de la modifiabilité présente un enjeu de taille dans le *protein design* [204] : une structure très modifiable est plus robuste aux mutations en moyenne et plus stable. England et Shakhnovich ont pu caractériser structuralement la modifiabilité [49]. Par cette méthode, il purent prouver que les structures hautement modifiables étaient plus fréquentes dans les organismes thermophiles [48] (une forte sélection des séquences est observée expérimentalement [90]).

I.5 Champ du travail

Notre objectif est de concevoir un modèle d'évolution incluant une contrainte fonctionnelle fondée sur l'interaction protéine-protéine. La prise en compte d'un modèle plus satisfaisant de la fonctionnalité est une sophistication récente. Mais les méthodes proposées jusqu'à présent ne nous semblent pas suffisantes. Blackburne et Hirst ont travaillé sur les poches hydrophobes d'un modèle HP sur réseau carré et tétraédrique [12, 82]. Ils ont construit des « paysages fonctionnels ». De manière assez similaire, des modèles de liaison à des ligands ont vu le jour [196]. Ces ajouts ne modifient pas singulièrement le schéma qui émerge des critères de stabilité structurale ou de l'efficacité du repliement. En particulier, la disparité des modifiabilités [196] et les structures de *superfunnel* persistent dans ces modèles [14].

L'intégration d'un critère fonctionnel n'est cependant pas complètement résolue. Par exemple, l'*apparition* d'une fonction n'est pas traitée. L'interaction protéine-protéine est une approche qui permet de tenter de répondre à cette question. Bien que l'interaction entre protéines soit peu éloignée des modèles de ligand, dans notre travail, le ligand est une protéine qui est, elle aussi, capable d'évoluer.

L'interaction protéine-protéine n'est pas une seule préoccupation de l'évolution *in silico*, à cause de son rôle ubiquitaire dans le fonctionnement de la cellule [62]. En évolution classique, l'émergence de systèmes d'interaction est également une question complexe en suspens. Dans les perspectives de son article *Selectionism and neutralism in molecular evolution*, Nei reconnaît l'importance de ces sujets d'investigation :

> However, the interaction of transcription factors and protein-coding genes or protein-protein interaction is quite complex. Therefore, the study of evolution

> of complex phenotypic characters would not be easy. Yet, this is one of the most important problems in evolutionary biology at present [142].

Dans leur revue, Chan et Bornberg-Bauer discutent cette approche, d'un point de vue computationnel :

> The most widely used among the criteria are based either on (1) native structure, (2) foldability or (3) native thermodynamic stability. More complex fitness criteria that might, in a sense, be more relevant ot biological function would require more complicated setups, such as taking into account multiple chain interactions. Consideration of such criteria would likely lead to big increases in the number of model parameters and computational intensity [32, p. 127].

Contrairement au contenu de la citation précédente, la prise en compte de l'interaction protéine-protéine ne demande pas l'addition de nombreux paramètres. Certes, des approximations sont nécessaires. Nous nous limitons aux dimères à trois états : les partenaires du complexe doivent se replier avant d'entreprendre de s'associer. Par conséquent, l'étude que nous allons proposer repose sur un travail préalable des protéines monomériques. Les avantages de cette méthode sont, premièrement, un coût de calcul compatible avec les ressources des ordinateurs disponibles et deuxièmement qu'il est possible de présenter dans ce modèle comment une structure s'accomode d'une contrainte structurale.

Nous proposons un modèle de protéine sur réseau et un autre de protéine tridimensionnelle hors-réseau. L'étude des protéines monomériques tridimensionnelles permet de confirmer l'existence de la structure de *superfunnel* en utilisant un modèle plus réaliste, sur la route *from lattice to all-atom models* [132].

Le nombre de paramètres n'est pas rédhibitoire, comme l'ont pensé initialement Chan et Borberg-Bauer. Nous verrons que seule une constante γ, rendant compte de la pénalité entropique qui incombe aux protéines interagissant, est requise pour modéliser la dimérisation de protéines sur réseau. Un modèle tridimensionnel est proposé plus tard. Dans ce cas, c'est le rang du dimère natif, par ordre d'énergie croissante, qui suffit à estimer l'efficacité de l'association.

Par l'étude de l'interaction protéine-protéine, nous introduisons un autre ingrédient largement étudié en génétique des populations mais jusqu'alors ignoré dans ce type de simulation : les relations épistatiques entre gènes. Pour l'instant, il n'existe pas de modèle physique de l'évolution de tels systèmes d'interaction. Nous verrons que l'interaction épistatique accentue les effets de dynamique des populations.

Le réseau neutre résultant de la contrainte protéine-protéine maintient une structure de *superfunnel*. Nous trouvons que la capacité à dimériser est augmentée sous l'action de l'évolution neutre. Cette observation est symptomatique d'une structure de « *superfunnel* fonctionnel » : les séquences dimérisant efficacement sont au centre du réseau neutre dans la région fortement connectée du réseau.

Chapitre II

Modèle et méthodes

The ways in which mutations become advantageous are so opportunistic that it is not easy to find simple rules to describe them.

Kimura, *The neutral theory of molecular evolution*

II.1 Introduction

Le modèle d'interaction entre deux protéines que nous développerons dans des chapitres ultérieurs seront bâtis sur les réseaux neutres *monomériques* de ces protéines. Ce chapitre introduit les modèles et les outils nécessaires à l'étude des monomères. Comme cela a été dit dans l'introduction, les simulations d'évolution *in silico* nécessitent un modèle structural et fonctionnel qui fournit le support de l'évolution, le réseau neutre, et un modèle évolutif qui décrit la dynamique d'une population évoluant sur ce réseau.

L'évolution est supposée neutre et s'applique, à une population haploïde : une séquence doit se replier dans une conformation donnée pour assurer la survie de l'individu qui en est porteur. Nous dirons que la séquence est « viable ». L'ensemble des séquences viables connectées par des mutations ponctuelles forment un réseau neutre. Nous définirons précisément les règles décrivant itérativement comment évolue une population infinie sur ce réseau. Ces règles sont inspirées de travaux antérieurs [21, 44, 184, 203]. Nous démontrerons que ces règles assurent une convergence vers un état stationnaire déterminé unique-

ment par la topologie du réseau neutre. La preuve de la convergence requiert l'introduction préliminaire de quelques outils issus de la théorie des graphes.

Les réseaux neutres seront construits à partir de deux modèles de protéine. Dans les deux modèles, les séquences viables sont celles qui se replient dans une conformation cible avec une stabilité suffisante.

Nous présenterons un modèle de protéine sur réseau bidimensionnel utilisé à diverses reprises par différents auteurs [112, 172, 173, 195, 203]. Nous utiliserons plusieurs modèles inspirés du modèle HP [38]. Afin d'atteindre des complexités raisonnables, le modèle HP considère seulement deux catégories d'acides aminés : les acides aminés hydrophobes (H) et les acides aminés hydrophiles (P).

Nous détaillerons également un modèle de protéine tridimensionnelle très similaire à celui de Bastolla *et al.* [7]. À cet effet, le squelette peptidique de trois protéines sera utilisé. Les trois protéines en question sont le trp-cage (1L2Y), le domaine SH3 de Vav et le domaine SH3 de Grb2 (chaînes B et C de 1GCQ). La résolution de ce modèle est à l'échelle du résidu d'acide aminé et le potentiel énergétique est un potentiel d'interaction entre les paires d'acides aminés en contact. La capacité d'une séquence à se replier dans la conformation cible sera évaluée à l'aide d'un jeu de conformations mimant les états dénaturés de la séquence. Les séquences viables seront ensuite converties en « profils » à partir desquels seront bâtis les réseaux neutres.

II.2 Définition et propriétés des graphes

Les réseaux neutres sont des graphes de séquences ou de paires de séquences. Nous définissons plus rigoureusement un graphe et énonçons quelques propriétés qui nous seront utiles par la suite.

II.2.1 Définition

Un « graphe orienté non pondéré » $\mathcal{G}$ est, en mathématique et en informatique, une structure définie par les nœuds qui le constituent $\mathbf{N}$ et les arêtes $\mathbf{E}$ qui relient les nœuds entre eux :

$$\mathcal{G} = (\mathbf{N}, \mathbf{E}).$$

$\mathbf{E}$ est un sous-ensemble de $\mathbf{N}^2$. Par exemple, le graphe $(\{1,2,3,4\}, \{(1,2),(2,3),(2,4)\})$ définit la structure présentée dans la figure II.1A.

Fig. II.1 : A — Graphe orienté simple. Une flèche $i \to j$ figure l'arête (i, j). B — Graphe non orienté.

En pratique, il existe deux manières de représenter un graphe de N nœuds. La première d'entre elles est la « liste d'adjacence » qui est simplement la donnée de l'ensemble **E** selon le format

({⟨voisins du nœud 1⟩},
{⟨voisins du nœud 2⟩},
etc.
{⟨voisins du nœud N⟩})

La liste d'adjacence du graphe de la figure II.1A est donc $(\{2\}, \{3, 4\}, \varnothing, \varnothing)$. La seconde façon de formaliser un graphe est sa « matrice d'adjacence ». La matrice d'adjacence $\mathcal{A} = (a_{ij})$ est une matrice N × N telle que $a_{ij} = 1$ si (i, j) est une arête du graphe et $a_{ij} = 0$ sinon. Le graphe de la figure II.1A est décrit par la matrice d'adjacence

$$\mathcal{A} = \begin{pmatrix} 0 & 1 & 0 & 0 \\ 0 & 0 & 1 & 1 \\ 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 \end{pmatrix}.$$

Un graphe non orienté est un graphe tel que si (i, j) fait partie de ses arêtes, (j, i) en fait partie aussi. Le graphe de la figure II.1B en est un exemple. La matrice d'adjacence de tels graphes est symétrique : $a_{ij} = a_{ji}$; par exemple, dans le cas du graphe de la figure II.1B :

$$\mathcal{A} = \begin{pmatrix} 0 & 1 & 0 & 0 \\ 1 & 0 & 1 & 1 \\ 0 & 1 & 0 & 0 \\ 0 & 1 & 0 & 0 \end{pmatrix}.$$

Dans ce cas une arête est le doublet $\{i, j\}$, et l'on notera dans le reste de l'ouvrage $i \sim j$.

En informatique, la matrice d'adjacence est pratique pour tester rapidement si une arête relie deux nœuds i et j. Il suffit en effet de consulter la valeur a_{ij}. Les ressources nécessaires au stockage en mémoire de la matrice d'adjacence peuvent être cependant phénoménales. Dans l'étude des dimères tridimensionnels nous traiterons des graphes possé-

dant environ $N = 470\,000$ nœuds. La matrice d'adjacence correspondante est composée de $N^2 = 221 \cdot 10^9$ éléments. La liste d'adjacence peut être moins pratique mais plus économique que la matrice d'adjacence ; sa taille en mémoire est proportionnelle au nombre d'arêtes composant le graphe. Pour un graphe peu connecté, son utilisation est préférable.

Un « graphe bipartite » est un graphe dont l'ensemble des nœuds $\mathbf{N}$ peut être scindé en deux catégories $\mathbf{N}_1$ et $\mathbf{N}_2$ tels qu'il n'existe aucune arête reliant deux nœuds appartenant à la même catégorie. Le graphe II.1B est un graphe bipartite : les catégories $\mathbf{N}_1 = \{1, 3, 4\}$ et $\mathbf{N}_2 = \{2\}$ remplissent les conditions requises.

Les réseaux neutres composés de séquences HP sont bipartites. Deux catégories de séquences peuvent être formées. La première, $\mathbf{N}_1$, inclut toutes les séquences contenant un nombre pair de résidus P, la seconde, $\mathbf{N}_2$, toutes les séquences contenant un nombre impair de résidus P. Comme une mutation ponctuelle change nécessairement la parité du nombre de résidus P présents, une connexion ne peut exister qu'entre une séquence du groupe $\mathbf{N}_1$ et une séquence du groupe $\mathbf{N}_2$.

II.2.2 Spectre d'un graphe

La matrice d'adjacence d'un graphe, par sa nature, se prête à l'analyse habituelle des matrices introduites dans l'étude des applications linéaires. La « valeur propre » λ d'une matrice $N \times N$ (et par extension du graphe qu'elle décrit) vérifie par définition, pour un vecteur p non nul de $\mathbb{R}^N$,

$$\mathcal{A} \times p = \lambda \cdot p, \tag{20}$$

où l'opérateur « $\times$ » désigne le produit matriciel et « $\cdot$ » le produit par un scalaire[11]. Le vecteur p est un « vecteur propre » associé à la valeur propre λ. L'ensemble des valeurs propres forment le « spectre » de la matrice d'adjacence ou du graphe.

Les graphes bipartites ont la particularité suivante : si λ est une valeur propre du graphe, alors $-\lambda$ est aussi une valeur propre. Soit $p = (p_i)_{i \in \mathbf{N}}$ un vecteur propre associé à λ. On peut voir que le vecteur $\tilde{p} = (\nu_i\, p_i)_{i \in \mathbf{N}}$, où $\nu_i = 1$ si $i \in \mathbf{N}_1$ et $\nu_i = -1$ si $i \in \mathbf{N}_2$, est un vecteur propre associé à $-\lambda$. Cette particularité aura son importance pour les réseaux neutres contruits à l'aide de séquences HP.

II.2.3 Méthode de la puissance

Nous démontrerons plus bas qu'une population d'individus soumise au modèle évolutif converge vers un état stationnaire qui dépend de la topologie du réseau neutre. La plus

11. Dans la suite nous omettrons ces opérateurs pour alléger les formules.

grande valeur propre positive λ_p du réseau neutre et l'un de ses vecteurs propres caractériseront cet état stationnaire. Il sera donc nécessaire de déterminer la valeur λ_p. L'algorithme le plus simple pour calculer la plus grande valeur propre *en valeur absolue* λ_a est la « méthode de la puissance ». En ce qui concerne notre étude des réseaux neutres, nous verrons que grâce au théorème de Frobenius-Perron, la détermination de λ_a équivaut à quelques détails près à celle de λ_p.

Soit u_0 un vecteur quelconque. On applique récursivement et jusqu'à convergence la récurrence pour $t > 0$

$$v_t = \mathcal{A}\, u_{t-1}\,,$$
$$u_t = v_t / \|v_t\|_1\,,$$

où $\|\cdot\|_1$ est la « norme de Manhattan » (ou norme L^1) :

$$\|(x_i)\|_1 = \sum |x_i|.$$

Alors u_t tend vers un vecteur propre associé à λ_a. La valeur propre λ_a elle-même peut être déterminée en calculant $\|v_t\|_1 / \|u_t\|_1$. La méthode de la puissance peut être facilement adaptée aux listes d'adjacence.

Les graphes bipartites générés par l'utilisation de séquences HP possèdent deux valeurs propres maximales en valeur absolue λ_p et $-\lambda_p$. Il est possible de translater la matrice par une quantité positive δ sur la diagonale pour calculer λ_p et l'un de ses vecteurs propres. Cette méthode se nomme « méthode de la puissance translatée ».

II.2.4 Théorème de Frobenius-Perron

Une matrice « réductible » est une matrice $\mathcal{A}$ dont les lignes et les colonnes peuvent être réordonnés de telle sorte qu'elle forme une matrice triangulaire supérieure par bloc. Une matrice qui n'est pas réductible est dite « irréductible ». La matrice d'adjacence d'un graphe ne possédant qu'une seule composante connexe est nécessairement irréductible.

Le théorème de Frobenius-Perron énonce que pour une matrice N × N, $\mathcal{A} = (a_{ij})$, positive et irréductible :

1. il existe une valeur propre λ_1 positive supérieure *en valeur absolue* à toute autre valeur propre λ_i,
$$|\lambda_i| \leq \lambda_1\ ;$$
2. il existe un vecteur propre $p = (p_i)$ positif associé à λ_1, c'est-à-dire $p_i \geq 0$ pour tout i.

Les matrices d'adjacence de nos réseaux neutres sont positives (par définition) et irréductibles (parce que les réseaux neutres ne sont formés que d'une unique composante connexe). La valeur propre positive maximale est également maximale en valeur absolue. Pour cette raison, la recherche de la valeur propre positive maximale λ_p se confond, en pratique, avec celle de la valeur propre maximale en valeur absolue λ_a, à l'exception des graphes bipartites. Les graphes bipartites feront donc l'objet d'un traitement particulier.

II.2.5 Composantes connexes

Un graphe n'est pas nécessairement entièrement connecté. Un autre jeu de connexions transforme le graphe de la figure II.1B en celui de la figure II.2. Toute sous-partie d'un graphe qui est entièrement connectée est appelée « composante connexe » du graphe. Le graphe de la figure II.2 comporte deux composantes connexes : $\{1, 4\}$ et $\{2, 3\}$.

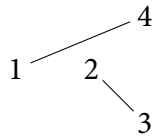


Fig. II.2 : Graphe non entièrement connecté constitué de deux composantes connexes.

II.2.6 Plus court chemin reliant deux points

Soit un graphe non orienté comportant N nœuds que nous numérotons de 1 à N. Un « chemin » du nœud i au nœud j est une suite

$$(k_1, k_2, \ldots, k_l)$$

telle que

$$k_1 = i, \tag{21a}$$

$$k_l = j, \tag{21b}$$

$$k_n \sim k_{n+1} \quad \text{pour tout } n < l. \tag{21c}$$

$l - 1$ est la « longueur » du chemin. Un tel chemin existe si et seulement si i et j appartiennent à la même composante connexe. Il existe un chemin de longueur minimale : le « plus court chemin menant de i à j ». La longueur de ce chemin est notée $\mathrm{len}\{i, j\}$. On convient habituellement que $\mathrm{len}\{i, j\} = +\infty$ lorsque i et j n'appartiennent pas à la même composante connexe.

Nous noterons $l_{\max}$ la longueur du plus court chemin maximale au sein d'une composante connexe d'un graphe, et nous nommerons « diamètre » :

$$l_{\max} = \max_{i,j} \{ \text{len}\{i, j\} \}.$$

II.3 Modèle évolutif

Les mutations affectent l'ADN. Afin de simplifier l'étape de transcription-traduction des gènes, nous nous focalisons sur le produit d'expression, la protéine. Dans notre modèle, par conséquent, les mutations affectent directement la protéine. Cela revient à ignorer les mutations synonymes au niveau de l'ADN. Dans cette section, nous décriront le modèle évolutif et les caractéristiques de son état stationnaire.

Nous considérons une population haploïde *infinie*. Nous reviendrons sur cette hypothèse dans la section « Polymorphisme, monomorphisme à l'état stationnaire » (p. 61). Intéressons-nous à un gène qui exprime une protéine de longueur L. Les types d'acide aminé sont classés en A catégories. L'ensemble des séquences possibles, l'« espace des séquences » $\mathbf{S}$, est l'ensemble de tous les L-uplets de ces A lettres. Supposons que l'ensemble des protéines fonctionnelles, d'une part, et neutres du point de vue de la sélection naturelle, d'autre part, soit le sous-ensemble $\mathbf{s}$ de $\mathbf{S}$:

$$\mathbf{s} = \{ s_1, \ldots, s_N \}.$$

Nous décidons de noter $p_i(t)$ la fréquence de la séquence s_i au temps t dans la population. La somme des fréquences vaut un :

$$\sum_i p_i(t) = 1. \tag{22}$$

II.3.1 Construction du réseau neutre

Les mutations ponctuelles transformant une séquence s_i en une séquence s_j confèrent une structure de graphe à l'ensemble $\mathbf{s}$: l'ensemble des arêtes est l'ensemble de telles paires de séquences (i, j). Le processus est supposé réversible de telle sorte que si l'arête (i, j) existe, il en va de même pour l'arête (j, i). Les arêtes ne sont pas pondérées : toute mutation est aussi probable qu'une autre. Le graphe non orienté défini par ce procédé est le « réseau neutre ». Dans la suite, nous noterons $i \sim j$ la relation « s_i et s_j sont connectés » et $\mathcal{A} = (a_{ij})$ la matrice d'adjacence du réseau neutre.

La « distance de Hamming » d_{H} entre deux séquences, $s = (s_i)$ et $s' = (s'_i)$, est le nombre de positions i où les deux séquences diffèrent. Deux séquences s_i et s_j sont connectées, $i \sim j$, si et seulement si $d_{\mathrm{H}}(s_i, s_j) = 1$. Le nombre de séquences connectées à s_i est noté n_i qui s'interprète biologiquement comme le nombre de mutations tolérées par s_i. Nous nommerons n_i la « robustesse mutationnelle » de s_i. On peut relier n_i à la matrice d'adjacence du réseau neutre :

$$n_i = \sum_j a_{ij}.$$

II.3.2 Équation d'évolution

Nous définissons μ, le taux de mutation total par protéine. Le taux de mutation par position est donc simplement μ/L. La population au temps $t + 1$ dépend de la population au temps t selon les règles suivantes.

Protocole évolutif 1

1. Chaque individu de la génération mère se reproduit fidèlement avec une probabilité 1 − μ.
2. Avec une probabilité μ, il subit une mutation ponctuelle affectant l'une des L positions de la protéine étudiée. Les mutations aux L positions possibles sont équiprobables. Le nombre de mutations accessibles est $\ell = \mathrm{L}\,(\mathrm{A} - 1)$.
3. Le mutant obtenu en 2 est conservé s'il appartient à l'ensemble s (mutation neutre).
4. Sinon, il est éliminé (mutation létale) et est remplacé par l'enfant d'un individu choisi au hasard dans la génération mère.

L'étape 4 est nécessaire pour maintenir la population constante. Elle est la source de la charge génétique dans notre modèle car chaque mutation létale doit être compensée par une naissance supplémentaire. Nous aurions pu concevoir que des mutations survinssent dans cette étape, mais le taux de mutation μ est considéré suffisamment faible pour que l'on puisse négliger cet événement dont la probabilité est de l'ordre de μ^2. Puisque la population est infinie, les fréquences $p_i(t)$ sont données par la relation de récurrence déterministe

$$p_i(t+1) = (1 - \mu + \kappa)\, p_i(t) + \frac{\mu}{\ell} \sum_{j \sim i} p_j(t), \tag{23}$$

où κ est la contribution de l'étape 4 maintenant la population constante. Cette constante peut être interprétée comme la fraction de la progéniture non viable générée par des mutations létales. Un calcul direct de la variable κ est possible mais laborieux. On peut la tirer

plus simplement de la contrainte

$$\sum_i p_i(t+1) = \sum_i p_i(t) = 1\,.$$

On obtient

$$\kappa = \mu\left(1 - \frac{1}{\ell}\sum_i \sum_{j\sim i} p_j(t)\right).$$

Nous réutiliserons à plusieurs reprise la propriété suivante. Pour toute fonction $f(i)$,

$$\begin{aligned}\sum_i \sum_{j\sim i} f(j) &= \sum_i \sum_j a_{ij} f(j)\\ &= \sum_j f(j) \sum_i a_{ij}\\ &= \sum_j f(j)\, n_j,\end{aligned} \tag{24}$$

Grâce à la propriété 24, on arrive à

$$\kappa = \mu\left(1 - \frac{\langle n\rangle}{\ell}\right). \tag{25}$$

Le terme $\langle n\rangle$ est le nombre moyen de connexions pondéré par la population. Cette valeur, qui dépend du temps, reflète la robustesse moyenne d'une séquence choisie au hasard dans la population :

$$\langle n\rangle = \sum_i p_i(t)\, n_i\,.$$

L'équation 23 prend la forme :

$$p_i(t+1) = p_i(t) + \frac{\mu}{\ell}\Big(\sum_{j\sim i} p_j(t) - \langle n\rangle\, p_i(t)\Big). \tag{23′}$$

II.3.3 Caractérisation de l'état stationnaire

L'état stationnaire, s'il existe, vérifie $p_i(t+1) = p_i(t)$ pour tout i. Nous notons p_i^∞ la valeur correspondante. Le second terme du membre droit de l'égalité 23′ doit être nul pour tout i :

$$\frac{\mu}{\ell}\Big(\sum_{j\sim i} p_j^\infty - \langle n\rangle\; p_i^\infty\Big) = 0,$$

donc,

$$\sum_{j\sim i} p_j^\infty - \langle n\rangle\; p_i^\infty = 0\,. \tag{26}$$

Si $\mathcal{A}$ désigne la matrice d'adjacence du réseau neutre, l'équation 26 prend la forme d'une équation aux valeurs propres

$$\mathcal{A}\, p^{\infty} = \langle n \rangle\ p^{\infty}, \tag{26′}$$

indépendamment de μ, L et A.

Il est important de noter que dans l'équation aux valeurs propres 26′, la valeur propre et le vecteur propre sont liés par une relation supplémentaire : $\langle n \rangle$ est une fonction de p^{∞}. Pour que l'équilibre existe, il faut, par conséquent, que la valeur propre du vecteur propre p^{∞} coïncide avec $\langle n \rangle$.

Supposons que la valeur propre λ et le vecteur p^{∞} soient solutions de l'équation 26′. Nous allons démontrer que la valeur propre λ est nécessairement égale à la robustesse mutationnelle $\langle n \rangle$. On a pour tout i :

$$\sum_{j\sim i} p_j^{\infty} = \lambda\, p_i^{\infty}.$$

En sommant sur les indices i :

$$\begin{aligned} \sum_i \sum_{j\sim i} p_j^{\infty} &= \lambda \sum_i p_i^{\infty}, \\ \sum_i n_i\, p_i^{\infty} &= \lambda, \qquad \text{(d'après la propriété 24)} \\ \langle n \rangle &= \lambda. \end{aligned}$$

De même que l'état stationnaire, l'équation d'évolution 23′ peut être s'écrire sous forme vectorielle :

$$p(t+1) = p(t) + \frac{\mu}{\ell}\,(\mathcal{A} - \langle n \rangle\, \mathcal{I})\, p(t), \tag{23″}$$

où $\mathcal{I}$ est la matrice identité.

L'indépendance de l'état stationnaire par rapport à μ est importante. En effet, le *fitness* d'un individu défini comme le nombre d'enfants fertiles qu'il engendre est

$$\begin{aligned} w_i &= 1 + \kappa - \mu\,(\ell - n_i)/\ell, \\ &= 1 + \mu\,(n_i - \langle n \rangle)/\ell, \end{aligned} \tag{27}$$

L'état stationnaire, étant indépendant de μ, est valable dans la limite des très faibles taux de mutation faisant que tous les *fitness* convergent vers la même valeur $w = 1$. On peut mettre cette propriété en regard de celle des quasi-espèces qui ne peuvent exister que lorsque la

réplication est suffisamment fidèle, c'est-à-dire quand le taux de mutation est suffisamment faible (voir par exemple référence [60]). En effet, les quasi-espèces émergent de la dispersion des taux d'autoréplication des séquences d'ARN. Dans notre modèle, des taux de mutation suffisamment faibles pour être favorables à l'apparition des quasi-espèces sont compensés par de faibles différences de *fitness* (puisque $\Delta w \propto \mu$). Au contraire, si le taux de mutation augmente, bien que les différences de *fitness* deviennent importantes, le régime mutationnel est moins favorable à l'apparition des quasi-espèces.

Dans la section suivante, nous démontrons que $p(t)$, déterminé par les formules de récurrence 23′ et 23″, tend vers un vecteur propre associé à la valeur propre positive maximale de la matrice d'adjacence décrivant le réseau neutre.

II.3.4 Convergence de la récurrence

Nous allons démontrer que le vecteur $p(t)$ converge vers un vecteur associé p^∞ à la valeur propre positive maximale de la matrice d'adjacence $\mathcal{A}$ du réseau neutre. En outre, nous verrons que ce vecteur est positif, donc physiquement acceptable. Nous procéderons en deux temps : nous traiterons en premier lieu le cas des réseaux neutres non bipartites, puis le cas particulier les réseaux neutres bipartites.

II.3.4.1 Réseau non bipartite

Nous nous limitons au cas où il existe N valeurs propres égales ou distinctes[12] et nous supposons que nos matrices ne créent pas de situation où la multiplicité de la valeur propre maximale est strictement supérieure à un. Nous ajoutons dans un premier temps l'hypothèse que la valeur propre positive maximale est également maximale en valeur absolue. Ces hypothèses sont vraies en pratique sauf dans le cas des réseaux bipartites qui sont discutés plus bas. Le spectre du réseau neutre peut donc s'écrire par ordre décroissant en valeur absolue[13] :

$$\lambda_1 > |\lambda_2| \geq \cdots \geq |\lambda_N|.$$

Notons v_1, v_2, etc., les vecteurs propres associés aux valeurs propres λ_1, λ_2, etc. L'équation 23″ peut s'écrire

$$p(t+1) = \xi(t)\, p(t) + \zeta\, \mathcal{A}\, p(t), \qquad (23''')$$

12. En somme, le total des multiplicités de chaque valeur propre est égal à la dimension N de l'espace vectoriel dans lequel agit la fonction linéaire associée à $\mathcal{A}$. Les vecteurs v_1, v_2, etc., forment une base de cet espace vectoriel.

13. Toutes les valeurs propres d'une matrice réelle symétrique sont réelles.

où $\zeta = \mu/\ell$. Nous insistons sur le fait que $\xi(t)$ dépend du temps :

$$\xi(t) = 1 - \mu \frac{\langle n \rangle}{\ell},$$

la dépendance provenant du fait que la moyenne $\langle n \rangle$, elle-même, dépend du temps car c'est une moyenne pondérée par les fréquences $p_i(t)$. On peut borner $\xi(t)$:

$$0 < 1 - \mu \leq \xi(t) \leq 1. \tag{28}$$

Les vecteurs v_i formant une base de $\mathbb{R}^N$, on peut écrire $p(t = 0)$ comme une combinaison linéraire des v_i :

$$p(0) = a_1 v_1 + \cdots + a_N v_N.$$

(Nous supposons que $a_1 \neq 0$.) On peut voir que

$$\begin{aligned}
p(1) &= \xi(0)\, p(0) + \zeta\,(\lambda_1 a_1 v_1 + \cdots + \lambda_N a_N v_N) \\
&= (\xi(0) + \zeta\lambda_1)\, a_1 v_1 + \cdots + (\xi(0) + \zeta\lambda_N)\, a_N v_N, \\
p(2) &= (\xi(0) + \zeta\lambda_1)\,(\xi(1) + \zeta\lambda_1)\, a_1 v_1 + \cdots + (\xi(0) + \zeta\lambda_N)\,(\xi(1) + \zeta\lambda_N)\, a_N v_N, \\
&\text{etc.}
\end{aligned}$$

et plus généralement :

$$p(t+1) = (\xi(0) + \zeta\lambda_1) \cdots (\xi(t) + \zeta\lambda_1)\,(a_1 v_1 + \nu_2 a_2 v_2 + \cdots + \nu_N a_N v_N),$$

où les coefficients ν_i sont donnés par la relation

$$\nu_i = \prod_{0 \leq \tau \leq t} \frac{\xi(\tau) + \zeta\lambda_i}{\xi(\tau) + \zeta\lambda_1}.$$

Il s'agit à présent de montrer que les ν_i tendent vers zéro. On peut donner une borne supérieure aux ν_i. Puisque $0 < \xi(t) \leq 1$,

$$|\nu_i| \leq \prod_{0 \leq \tau \leq t} \frac{\xi(\tau) + \zeta|\lambda_i|}{\xi(\tau) + \zeta\lambda_1} \leq \left(\frac{1 + \zeta|\lambda_i|}{1 + \zeta\lambda_1} \right)^{t+1}. \tag{29}$$

Par définition,

$$\lambda_1 > |\lambda_i|,$$

donc, l'expression majorant $|\nu_i|$ dans l'équation 29 tend vers zéro quand t tend vers l'infini. Puisque $p(t)$ tend vers un vecteur non nul (équation 22) et que a_1 est supposé non nul, on en déduit que $p(t+1)$ converge vers un vecteur non nul colinéaire à v_1.

II.3.4.2 Réseau bipartite

Le résultat précédent n'est pas valide tel quel pour toutes les matrices d'adjacence qui apparaîtront dans notre travail. Les alphabets limités à A = 2 classes d'acides aminés (les séquences de type HP), engendrent des réseaux neutres bipartites. Nous savons que pour de tels réseaux ont un spectre symétrique par rapport à zéro. Nous supposerons que dans les cas pratiques que nous allons traiter, le spectre peut s'écrire

$$\lambda_1 > \lambda_2 \geq \cdots \geq \lambda_{-2} = -\lambda_2 > \lambda_{-1} = -\lambda_1 .$$

La seule condition supplémentaire à vérifier pour que $p(t)$ converge vers v_1 est que le coefficient ν_{-1} correspondant à la valeur propre λ_{-1} tende vers zéro quand t tend vers l'infini. Nous voyons que c'est bien le cas :

$$\begin{aligned}
|\nu_{-1}| &= \prod_{0\leq\tau\leq t} \frac{|\xi(\tau) - \zeta\lambda_1|}{\xi(\tau) + \zeta\lambda_1} \\
&\leq \prod_{0\leq\tau\leq t} \frac{\max\{\xi(\tau), \zeta\lambda_1\}}{\xi(\tau) + \zeta\lambda_1} \\
&= \prod_{0\leq\tau\leq t} \max\Big\{\frac{\xi(\tau)}{\xi(\tau) + \zeta\lambda_1}, \frac{\zeta\lambda_1}{\xi(\tau) + \zeta\lambda_1}\Big\} \\
&\leq \prod_{0\leq\tau\leq t} \max\Big\{\frac{1}{1 + \zeta\lambda_1}, \frac{\zeta\lambda_1}{1 - \mu + \zeta\lambda_1}\Big\} \\
&\leq \Big(\max\Big\{\frac{1}{1 + \zeta\lambda_1}, \frac{1}{1 + (1-\mu)/(\zeta\lambda_1)}\Big\}\Big)^{t+1}.
\end{aligned}$$

Puisque $\zeta\lambda_1$, $1 - \mu$ sont strictement supérieurs à zéro, ν_{-1} tend effectivement vers zéro comme requis.

II.3.5 Autres modèles proches

D'autres processus évolutifs similaires au protocole 1 peuvent être conçus. Lorsqu'une mutation létale intervient dans le protocole que nous venons de voir, nous remplaçons l'individu moribond par l'enfant d'un individu tiré au hasard dans la génération mère. Cela peut être vu comme une mise à l'échelle *préalable* de la population. On peut à la place imaginer une mise à l'échelle *a posteriori*.

Protocole évolutif 2

1. Chaque individu de la génération mère se reproduit fidèlement avec une probabilité $1 - \mu$.

2. Avec une probabilité μ, il subit une mutation ponctuelle affectant l'une des L positions de la protéine étudiée. Les mutations aux L positions possibles sont équiprobables.
3. Le mutant obtenu en 2 est conservé s'il appartient à l'ensemble s (mutation neutre).
4. Sinon, il est éliminé (mutation létale).
5. Lorsque tous les individus se sont reproduits par répétition des étapes 1, 2, 3 et 4, la population totale a vraisemblablement diminué à cause des mutations létales (étape 4). La population est donc mise à l'échelle pour rétablir une population totale constante.

L'équation d'évolution devient

$$p_i(t+1) = \frac{1}{\kappa'}\left\{(1-\mu)\,p_i(t) + \frac{\mu}{\ell}\sum_{j\sim i} p_j(t)\right\}, \tag{30}$$

où κ' maintient la population totale constante. En utilisant la même méthode que précédemment :

$$\kappa' = 1 - \mu + \frac{\mu}{\ell}\,\langle n\rangle\,. \tag{31}$$

À l'état stationnaire,

$$\kappa'\,p_i^{\infty} = (1-\mu)\,p_i^{\infty} + \frac{\mu}{\ell}\sum_{j\sim i} p_j^{\infty},$$

En insérant l'expression de κ', comme dans l'équation 26,

$$\sum_{j\sim i} p_j^{\infty} - \langle n\rangle\; p_i^{\infty} = 0. \tag{32}$$

Ainsi, ce processus évolutif, bien que différent dans sa dynamique (équation 30), possède le même état stationnaire (équation 32).

Protocole évolutif 3

Nous proposons encore une autre voie menant aux mêmes équations : un individu se reproduit à un taux ρ, meurt à un taux δ et mute à un taux μ. On peut écrire :

$$p_i(t+1) = (1+\rho-\delta-\mu)\,p_i(t) + \frac{\mu}{\ell}\sum_{j\sim i} p_j(t). \tag{33}$$

Si l'on impose de plus que la population doit rester constante, la contrainte suivante s'applique

$$\sum_i p_i(t+1) = \sum_i p_i(t) = \sum_i (1+\rho-\delta-\mu)\,p_i(t) + \frac{\mu}{\ell}\sum_{j\sim i} p_j(t),$$

et elle sera satisfaite si et seulement si

$$\rho - \delta = \mu\,(1 - \langle n\rangle\,/\ell)\,.$$

En remplaçant dans l'équation originale, on déduit à nouveau l'équation 23′ :

$$p_i(t+1) = p_i(t) + \frac{\mu}{\ell}\Big(\sum p_j(t) - \langle n\rangle\ p_i(t)\Big). \tag{34}$$

Ce n'est pas étonnant, étant donné que $\rho - \delta$ joue exactement le rôle de κ. Néanmoins, cela nous permet de relier κ à des paramètres biologiques.

Une équation d'évolution plus générale incluant des mutations multiples a été proposée par Wilke [194].

II.3.6 Polymorphisme, monomorphisme à l'état stationnaire

La théorie de l'évolution neutre accorde un rôle important à la dérive génétique qui résulte de la finitude des populations biologiques. Néanmoins, notre modèle suppose une population infinie. Nous décrivons ici comment l'état stationnaire calculée sous l'hypothèse d'une population infinie peut s'interpréter quand la population est finie. Envisageons une situation où le nombre de séquences possibles est très faible.

Si une population est très grande et si le taux de mutation μ est assez élevé, la population est très polymorphe : toutes les séquences sont représentées dans la population. La fraction de la population qui est porteuse d'une séquence donnée est proche de la fraction calculée dans notre modèle de population infinie.

Au contraire, si la population est petite et si le taux de mutation est faible, la population est essentiellement monomorphe : il n'y a à tout moment qu'une seule séquence représentée dans la population. Cependant, ce n'est pas toujours la même. Par exemple, quand seulement deux séquences possibles, s_1 et s_2, existent, la distribution de probabilité de la fraction x d'individus porteurs de s_1 possède une forme marquée de U après un temps long [92]. En d'autres termes, la situation la plus probable est une population homogène ne comportant pratiquement que s_1 ou que s_2. L'allèle majoritaire cependant alterne. On parle de « quasi-fixation ».

La situation à trois séquences mutant selon le schéma

$$s_1 \leftrightarrow s_2 \leftrightarrow s_3$$

est très proche. La distribution des fractions d'individus porteurs de l'un ou l'autre des séquences ressemble à celle présentée dans le graphique de droite de la figure II.3. On ob-

serve que la population est l'essentiel du temps monomorphe, avec de courtes transitions qui interchangent la séquence majoritaire (graphique de gauche de la figure II.3).

L'état stationnaire représente alors la moyenne sur un temps long du temps où une séquence est quasi fixée (voir également référence [163]). Dans l'exemple des trois séquences, le vecteur propre à l'état stationnaire vaut

$$p^{\infty} = (0{,}29, 0{,}41, 0{,}29)\,.$$

La simulation de la figure II.3 montre qu'effectivement la séquence s_2 est quasi fixée plus souvent que les séquences s_1 et s_2.

La section « Effet de la taille de la population » discute plus profondément la validité de l'hypothèse de l'infinité de la population.

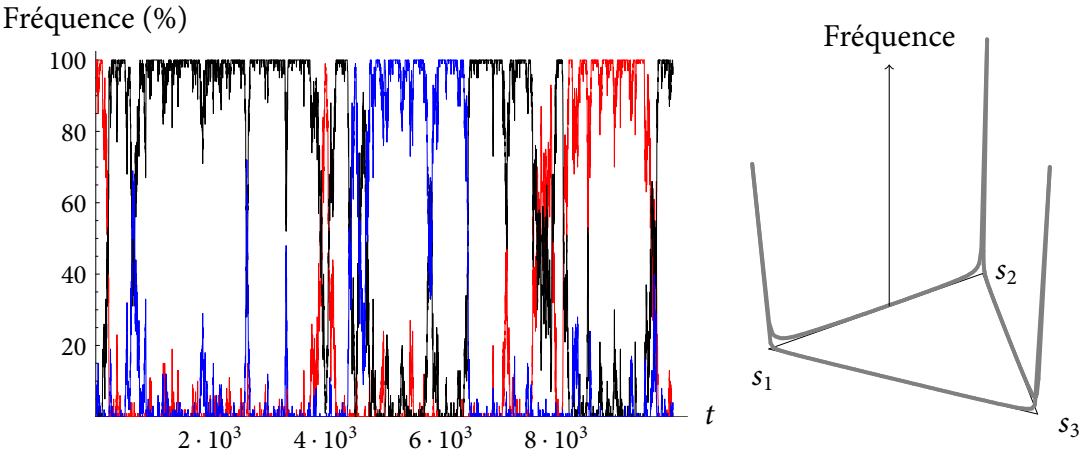


Fig. II.3 : Équilibre entre dérive et mutation dans la situation de trois allèles neutres qui mutent selon $s_1 \leftrightarrow s_2 \leftrightarrow s_3$. À gauche, dynamique pendant 10 000 générations. À droite, graphique ternaire représentant la densité des proportions observées. Les états les plus probables sont ceux où un allèle est fortement majoritaire et plus particulièrement l'allèle s_2. (Les paramètres utilisés sont : 100 individus, taux de mutation 10^{-3}.)

II.4 Modèles de protéine

Dans la section précédente, nous avons décrit comment l'évolution prend place sur un réseau neutre. La structure de réseau neutre elle-même est bâtie sur des modèles de protéines permettant de déduire quels éléments sont les nœuds du graphe, c'est-à-dire quelles sont les séquences viables et neutres. Le critère utilisé pour les protéines monomériques est structural : une séquence est déclarée viable si elle se replie et si sa conformation na-

tive est une conformation donnée. Le but de ce chapitre est donc de détailler comment est modélisé le repliement d'une séquence.

Le premier modèle de protéine est un modèle de protéine sur réseau et les acides aminés sont réduits à deux classes : hydrophobes (H) et polaires (P). Ces choix permettent la détermination exhaustive des séquences se repliant et de leur conformation native.

Le second modèle est un modèle de protéine tridimensionnelle hors-réseau. Nous devrons nous résoudre à une méthode heuristique pour évaluer la capacité d'une séquence à se replier. Par ailleurs, des alphabets plus complets seront envisagés ce qui excluera le calcul exhaustif des séquences viables. Par conséquent, nous procéderons à une exploration par une méthode de Monte Carlo et la projection dans un « sous-espace de profils » pour déterminer l'état stationnaire du modèle évolutif.

II.4.1 Protéines sur réseau

Les protéines sur réseau que nous allons décrire ont été introduites dans les années 70 grâce aux études de Gō et collaborateurs [65, 170]. Leur utilisation a bénéficié des travaux de Dill [38] qui s'attachait à trouver les mécanismes par lesquels un polymère pouvait trouver sa conformation d'énergie libre la plus basse. Depuis, le concept a été réutilisé dans de nombreux champs de la biologie, que ce soit le repliement et la dénaturation des protéines [28, 176], l'agrégation [39, 78, 86], la liaison de ligands [129] et plus récemment l'évolution des protéines.

II.4.1.1 Conformations

Les protéines sur réseau sont un modèle de protéine où les acides aminés prennent la forme de perles localisées aux intersections d'une grille bidimensionnelle ou tridimensionnelle. Les conformations sont des chemins s'inscrivant dans la grille (cf. figure II.4).

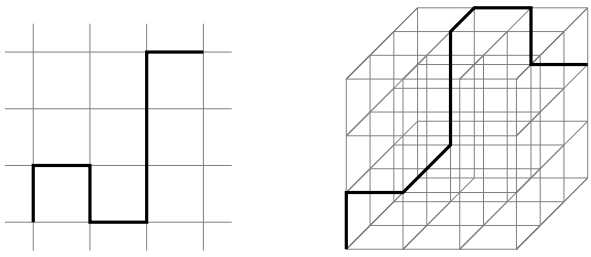


Fig. II.4 : Grilles bidimensionnelle et tridimensionnelle dans lesquelles s'inscrit une chaîne peptidique.

Les règles suivantes doivent être respectées : 1. deux acides aminés liés covalemment sont voisins dans la grille, 2. deux acides aminés distincts ne peuvent pas occuper le même

site dans la grille. On dit que deux acides aminés sont « en contact » s'il sont voisins sur la grille mais non liés covalemment. Le nombre de paires d'acides aminés en contact est le « nombre de contacts » formés par la chaîne peptidique.

La « carte de contact » d'une conformation est la liste des paires d'acides aminés en contact.

Si une conformation possède le nombre maximal de contacts possibles, elle est dite « compacte ». La figure II.5 montre deux conformations représentatives, l'une compacte et l'autre non.

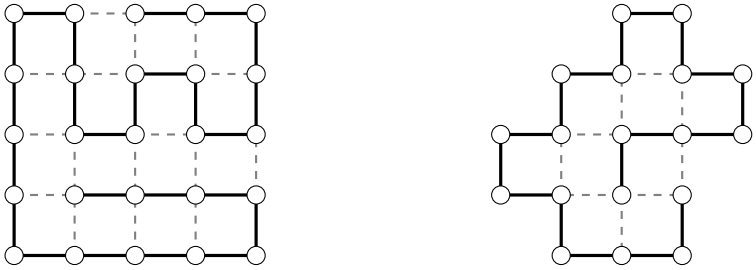


Fig. II.5 : Deux exemples de conformations sur réseau. Les contacts entre résidus sont représentés par les pointillés. À gauche, une conformation compacte de longueur L = 25 ; le nombre de contacts formés est 16. À droite, une conformation non compacte de longueur L = 18 ; le nombre de contacts formés est 9 tandis que le nombre maximal de contacts possible est 10 (d'après Bornberg-Bauer [20]).

Nous nous restreignons à présent à une grille bidimensionnelle 5×5 et aux conformations compactes de longueur L = 25. Le choix de la grille bidimensionnelle est justifié par un ratio résidus exposés sur résidus enfouis plus réalistes que dans le cas de grilles tridimensionnelles[14]. Ce modèle particulier a été utilisé à plusieurs reprises dans les travaux de Li *et al.* [112], de Taverna et Goldstein [172, 173], ceux de Wilke [195] et dans l'étude du rôle de la recombinaison de Xia and Levitt [203]. Il existe 132 contacts possibles entre les résidus de la chaîne dans ce modèle.

Du fait de l'orientation de la liaison peptidique extrémité N-terminale → extrémité C-terminale, nous distinguons une conformation de celle obtenue en inversant le sens de lecture. Ainsi, les deux conformations c et c' définies par la suite de leurs coordonnées cartésiennes

$$c = \big((x_1, y_1), \ldots, (x_L, y_L)\big),$$
$$c' = \big((x_L, y_L), \ldots, (x_1, y_1)\big),$$

14. Les conformations sur réseau tridimensionnel les plus utilisées sont restreintes à un cube $3 \times 3 \times 3$. Elles possèdent 26 résidus exposés et un unique résidu enfoui.

sont considérées distinctes. Ces conformations c et c' sont dites « inverses » l'une de l'autre, et on note inv la relation les associant : $c' = \text{inv}\, c$. Dans les figures où cela sera nécessaire, le sens de lecture sera indiqué par une flèche.

En revanche sont considérées équivalentes les conformations qui sont identiques à une rotation près où à une symétrie axiale près. Dans une terminologie mathématique, nous travaillerons sur les classes d'équivalence de la relation d'équivalence $\mathcal{R}$ opérant sur l'ensemble des conformations :

$$c \,\mathcal{R}\, c' \iff \begin{cases} c' \text{ peut être obtenue par composition de} \\ \text{symétries axiales et de rotations de } c. \end{cases} \tag{35}$$

Certaines conformations sont des points fixes de la fonction inv, c'est-à-dire qu'elles sont leur propre inverse ($c = \text{inv}\, c$). Nous qualifierons ces conformations de « symétriques ». Des exemples sont donnés dans la figure II.6.

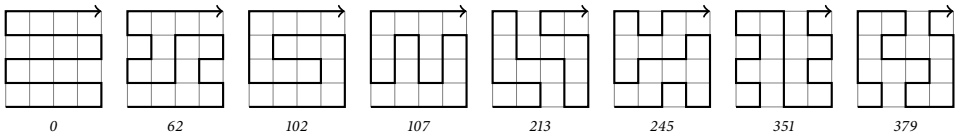


Fig. II.6 : Quelques conformations symétriques.

L'algorithme utilisé pour générer les conformations est simplement une exploration en profondeur grâce à une fonction récursive f :

sub $f(\{(x_1, y_1), \ldots, (x_i, y_i)\}) \equiv$
  if $i =$ L **then**
    Retenir $c = \{(x_1, y_1), \ldots, (x_i, y_i)\}$
  else
    for tout voisin libre (x_j, y_j) de (x_i, y_i) dans la grille **do**
      call $f(\{(x_1, y_1), \ldots, (x_i, y_i), (x_j, y_j)\})$
    end for
  end if
end sub

L'application naïve de la fonction f crée des conformations qui sont équivalentes par rotation ou symétrie. Afin de ne générer qu'une seule conformation représentative de sa classe d'équivalence induite par la relation $\mathcal{R}$ (équation 35), on applique la fonction récursive f aux segments initiaux présentés dans la figure II.7. Ces segments initiaux présentent la particularité de briser d'emblée les symétries de la grille carrée. Les conformations sont numérotées dans l'ordre dans lequel elles sont générées, de 0 à 1 080.

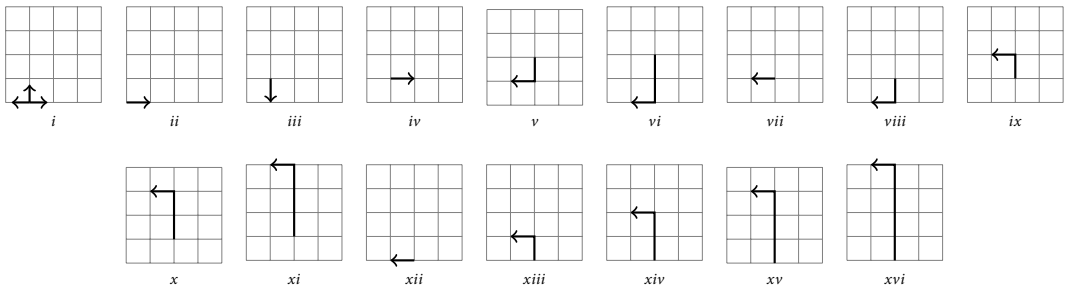


Fig. II.7 : Segment initiaux utilisés pour engendrer toutes les conformations. Ces segments brisent la symétrie de la conformation assurant qu'aucune conformation générée n'est image d'une autre par rotation ou symétrie.

La recherche en profondeur génère 1 081 conformations différentes dont 17 sont symétriques. Ce nombre est du même ordre de grandeur que le nombre de catégories de repliements des protéines globulaires pour lesquelles les prédictions vont de 1 000 à 4 000 catégories [23, 33, 69, 150, 208].

II.4.1.2 Séquences et alphabet HP

D'un point de vue général, les séquences peptidiques sont des L-uplets d'acides aminés. Le nombre de séquences possibles est fantastiquement grand même pour des valeurs relativement faibles de L : pour une petite protéine de L = 25 résidus, le nombre de séquences vaut $20^{25} \approx 3{,}4 \cdot 10^{32}$. La complexité peut être considérablement réduite si l'on tient compte du fait que certains acides aminés partagent des caractéristiques biochimiques communes. On peut les grouper en classes d'acides aminés semblables et faire l'approximation suivante : c'est la classe d'un acide aminé qui détermine ses propriétés énergétiques au sein de la protéine. Un « alphabet » est un ensemble de A classes d'acides aminés.

L'alphabet le plus simple est l'alphabet HP. Deux classes d'acides aminés existent : les acides aminés hydrophobes (H) et ceux polaires (P). Le fait que la collapse hydrophobe soit la force prédominante dans le processus de repliement explique cette ségrégation [159].

Dans ces conditions, en se limitant à des chaînes peptidiques de L = 25 résidus, l'ensemble des séquences possibles est

$$\{\mathrm{H}, \mathrm{P}\}^{25},$$

composé de $2^{25} \approx 3{,}4 \cdot 10^7$ séquences.

À l'image de ce qui a été dit à propos des conformations, une séquence possède un sens de lecture : nous différencions `HHP` et `PHH`. Deux séquences identiques au sens de lecture près sont nommées « inverses » l'une de l'autre, et nous noterons cette relation inv. Les points fixes de inv sont des séquences dites « symétriques ».

II.4.1.3 Énergie et propriétés associées

L'énergie d'une séquence s repliée dans une certaine conformation c est calculée en sommant les interactions entre les acides aminés en contact. Si la séquence $s = (s_1, \ldots, s_L)$ désigne une séquence d'acides aminés et $c = (c_1, \ldots, c_L)$ une suite de coordonnées dans la grille, l'énergie est souvent écrite sous la forme

$$E(s, c) = \sum_{i<j} \chi(c_i, c_j)\, \epsilon(s_i, s_j). \tag{36}$$

Les conventions suivantes s'appliquent : premièrement, $\chi(c_i, c_j)$ vaut un si les positions i et j sont en contact et zéro sinon. Dans une grille bidimensionnelle carrée, si $c_i = (x, y)$ et $c_j = (x', y')$, on a

$$\chi(c_i, c_j) = \begin{cases} 1 & \text{si } |x' - x| + |y' - y| = 1 \text{ et } |i - j| \neq 1, \\ 0 & \text{sinon.} \end{cases}$$

Deuxièmement, $\epsilon(s_i, s_j)$ représente l'énergie élémentaire résultant de l'interaction des acides aminés s_i et s_j. L'interaction ne dépend que des classes de s_i et s_j. La matrice d'interaction est donc résumée sous la forme

$$\mathcal{E} = \begin{array}{c} \\ \mathrm{H} \\ \mathrm{P} \end{array} \begin{array}{c} \begin{array}{cc} \mathrm{H} & \mathrm{P} \end{array} \\ \begin{pmatrix} e_{\mathrm{HH}} & e_{\mathrm{HP}} \\ e_{\mathrm{HP}} & e_{\mathrm{PP}} \end{pmatrix} \end{array}. \tag{37}$$

L'interaction $\epsilon(s_i, s_j)$ vaudra e_{HH} si s_i et s_j sont tous deux hydrophobes, e_{PP} si s_i et s_j sont tous deux polaires et e_{HP} sinon.

L'ensemble des énergies $E(s, c)$ d'une séquence s donnée lorsque c décrit l'ensemble des conformations est le « spectre énergétique » de s. Supposons que pour une séquence donnée, ce spectre s'écrive par ordre croissant

$$\{E_0, E_1, \ldots, E_n\}.$$

Le niveau E_0 est l'« état fondamental » de la séquence s. Le niveau E_1 est le « premier état excité ». On note ΔE la différence d'énergie $E_1 - E_0$. On appelle « degré de dégénérescence » du niveau d'énergie E_i, le nombre ν_i des conformations c telles que $E(s, c) = E_i$. Si $\nu_i > 1$, on dit que le niveau E_i est « dégénéré », sinon il est « non dégénéré ».

Avec ces notations, on dit que la séquence s « se replie » si son état fondamental est non dégénéré, c'est-à-dire si $\nu_0 = 1$. La « conformation native » de la séquence s est l'unique

conformation c_0 telle que $\mathrm{E}(s, c_0) = \mathrm{E}_0$. Les autres conformations jouent le rôle des états dénaturés. Pour une conformation donnée c, les séquences viables sont celles qui se replient en c.

La « température de repliement » d'une séquence s qui se replie est la température T_f à laquelle s est dans sa conformation native avec une probabilité 1/2. C'est donc la température à laquelle

$$\frac{1}{\mathrm{Z}(\mathrm{T}_f)} \exp\Big(-\frac{\mathrm{E}_0}{k\,\mathrm{T}_f}\Big) = 1/2,$$

où k est la constante de Boltzmann et $\mathrm{Z}(\mathrm{T})$ est la fonction de partition du système :

$$\mathrm{Z}(\mathrm{T}) = \sum_i \nu_i \, \exp\Big(-\frac{\mathrm{E}_i}{k\,\mathrm{T}}\Big).$$

Les propriétés énergétiques d'une séquence s repliée dans une conformation c, sont identiques à celles de la séquence inverse $\mathrm{inv}\, s$ repliée dans la conformation $\mathrm{inv}\, c$:

$$\mathrm{E}(s, c) = \mathrm{E}(\mathrm{inv}\, s, \mathrm{inv}\, c). \tag{38}$$

Il est donc possible de limiter l'exploration des séquences : d'après la table II.1, le nombre de séquences qu'il est nécessaire de traiter s'élève à à 16 781 312 au lieu de 33 546 240.

	Nombre	Traitées
Séquences symétriques	$2^{13} = 8\,192$	8 192
Séquences non symétriques	$2^{25} - 2^{13} = 33\,546\,240$	16 773 120
Total	$2^{25} = 33\,554\,432$	16 781 312

Tab. II.1 : Table présentant le nombre de séquences symétriques et non symétriques ainsi que le celui des séquences qui doivent être traitées.

Une implication de l'équation 38 est que, pour une séquence s symétrique, $\mathrm{E}(s, c) = \mathrm{E}(s, \mathrm{inv}\, c)$. L'état fondamental d'une telle séquence ne peut être non dégénéré que si la conformation minimisant l'énergie de s est symétrique : si une séquence symétrique se replie, elle doit donc se replier en une conformation symétrique.

Si une séquence non symétrique s se replie en une conformation symétrique c, alors la séquence inverse $\mathrm{inv}\, s$ se replie également en c. Le repliement établit une correspondance entre l'espace des séquences et l'espace des conformations schématisé dans la figure II.8.

II.4.1.4 Matrices d'interaction

La matrice d'interaction HP la plus employée est celle pour laquelle $e_{\mathrm{HH}} = -1$ et $e_{\mathrm{PP}} = e_{\mathrm{HP}} = 0$, avec les notations de l'équation 37. Dans notre étude, les matrices suivantes seront

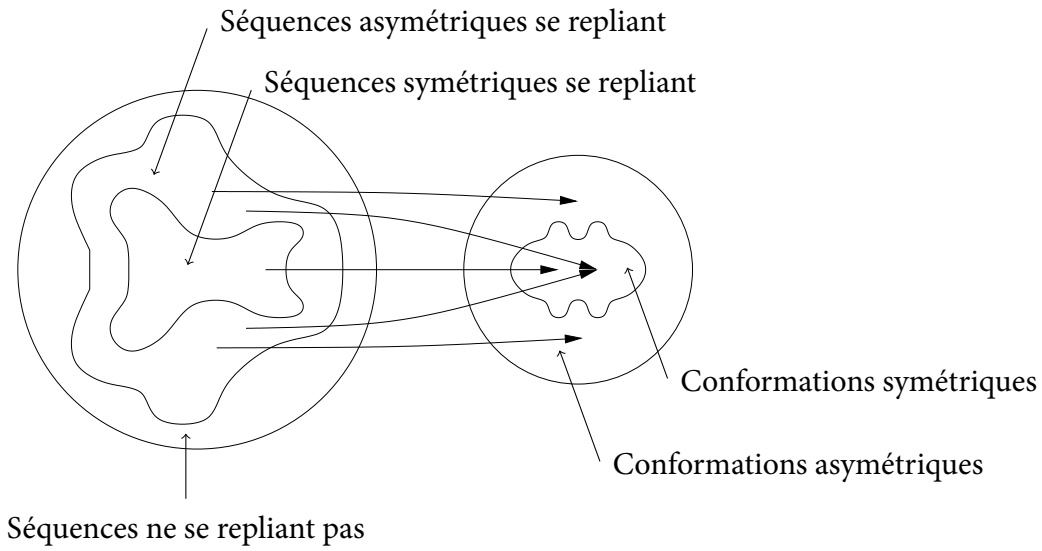


Fig. II.8 : Schéma de la correspondance entre l'espace des séquences et l'espace des conformations établie par le repliement.

utilisées :

$$\mathcal{E}_{\mathrm{GLO}} = \begin{pmatrix} 1 & 0 \\ 0 & -1 \end{pmatrix} \qquad \mathcal{E}_{\mathrm{HP}} = \begin{pmatrix} -1 & 0 \\ 0 & 0 \end{pmatrix}$$

$$\mathcal{E}_{\mathrm{HP'}} = \begin{pmatrix} -3 & -1 \\ -1 & 0 \end{pmatrix} \qquad \mathcal{E}_{\mathrm{I}} = \begin{pmatrix} 1 & -1 \\ -1 & 1 \end{pmatrix}$$

$$\mathcal{E}_{\mathrm{LHTW}} = \begin{pmatrix} -2.3 & -1 \\ -1 & 0 \end{pmatrix} \qquad \mathcal{E}_{\mathrm{TS_1}} = \begin{pmatrix} -6 & -3.5 \\ -3.5 & 1.5 \end{pmatrix}$$

$$\mathcal{E}_{\mathrm{TS_2}} = \begin{pmatrix} -6 & 3.5 \\ 3.5 & 1.5 \end{pmatrix}$$

Ces matrices sont adaptées à une constante de Boltzmann égale à un.

L'utilisation de différentes matrice d'énergie permet de distinguer les propriétés universelles de celles qui dépendent des paramètres précis du modèle. Cette approche a déjà été employée par Bornberg-Bauer *et al.* pour étudier la structure de *superfunnel* (matrices HP et AB) [21], par Buchler et Goldstein pour déterminer l'influence des alphabets sur les modifiabilités [25] ou encore par Xia et Levitt pour quantifier l'importance des conformations compactes dans l'émergence des *superfunnel* [202].

Une matrice reproduisant les caractéristiques du repliement protéique doit satisfaire les conditions suivantes (nous dirons qu'elle est « typique des protéines ») :

$$2\,e_{\mathrm{HP}} > e_{\mathrm{HH}} + e_{\mathrm{PP}}\,, \tag{39a}$$

$$e_{\mathrm{HH}} \leq e_{\mathrm{HP}}, e_{\mathrm{PP}}\,. \tag{39b}$$

La relation 39a exprime que les résidus H et P préfèrent être ségrégés. La seconde (39b) favorise l'enfouissement des résidus hydrophobes car les contacts sont plus nombreux dans le cœur de la protéine.

La matrice d'interaction GLO est typique des protéines si l'on échange les résidus H et P. Les matrices HP, HP′, LHTW et TS_2 sont toutes typiques des protéines. En revanche, TS_1 et Ising ne le sont pas. En fait, Ising possède des particularités faisant songer à des ARNt où les appariements favorisés sont ceux entre résidus complémentaires.

II.4.1.5 Séquences viables, réseaux neutres

La fonction d'une protéine étant difficile à modéliser, nous allons nous fonder sur la *structure* pour décider si une séquence est viable ou non. L'ensemble des séquences viables et la structure de graphe que les mutations ponctuelles lui confèrent constitueront un réseau neutre.

Pour chacune des 1 081 conformations, on peut calculer un ensemble de séquences viables. Soient c une conformation et $\mathrm{S}(c)$ l'ensemble des séquences se repliant en c. Le cardinal $\mathrm{D}(c) = |\mathrm{S}(c)|$ est appelé la « modifiabilité » de c. En raison du schéma de correspondance présenté dans la figure II.8, les modifiabilités de deux conformations inverses l'une de l'autre sont égales.

Puisque les mutations ponctuelles ne modifient qu'un seul résidu d'acide aminé, nous définissons une connexion entre la paire de séquences $\{s_i, s_j\}$ si

$$d_{\mathrm{H}}(s_i, s_j) = 1\,. \tag{40}$$

Les connexions confèrent une structure de graphe à l'ensemble $\mathrm{S}(c)$. Les connexions reflètent la possibilité de transformer une séquence en une autre par mutation ponctuelle.

Il est possible que certaines séquences soient isolées de certaines autres, c'est-à-dire que le graphe que nous venons de constituer soit composé de plusieurs composantes connexes. Nous appellerons « réseaux neutres » les différentes composantes connexes du graphe. À chaque conformation correspond zéro, un ou plusieurs réseau neutre. Chaque réseau neutre sera nommé dans cet ouvrage en indiquant le numéro de la conformation à laquelle il correspond suivi d'une lettre l'identifiant uniquement. Les propriétés des réseaux

neutres mentionnés dans cet ouvrage sont résumées dans « Résumé des propriétés des réseaux neutres » (p. 157).

II.4.1.6 Stockage des séquences, calcul des connexions

Nous avons décrit les principes permettant de calculer les séquences se repliant en une conformation et de construire les réseaux neutres d'une conformation donnée. Nous donnons ici des détails sur l'implémentation de ces idées.

Du point de vue technique, nous stockons les séquences sous la forme d'entiers non signés de 32 bits. Chaque position de la séquence correspond à un bit de l'entier non signé (sept positions restent inutilisées et sont simplement placées à zéro). L'entier u représenté par la séquence $s = \{s_1, \ldots s_L\}$ est

$$u = \bar{s}_1\, 2^0 + \cdots + \bar{s}_L\, 2^{L-1},$$

où $\bar{s}_i = 0$ si $s_i = \mathrm{H}$ et $\bar{s}_i = 1$ si $s_i = \mathrm{P}$. La figure II.9 schématise le stockage.

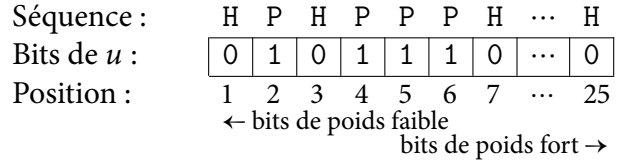


Fig. II.9 : Relation entre une séquence HP est son codage sous forme d'entier non signé. La séquence HP est convertie en séquence de 0 (H) ou de 1 (P) qui occupent les positions correspondantes des bits de l'entier u.

Les 2^{25} séquences sont donc représentées chacune par un entier différent de l'intervalle $[0, 2^{25} - 1]$. On n'utilise que quatre octets par séquence au lieu de vingt-cinq si la séquence était stockée sous forme de tableau de caractères, par exemple.

Un autre avantage que l'on peut tirer du codage en entier est la rapidité du calcul des connexions. En effet, il est possible de tester si deux séquences s_1 et s_2 sont connectées en utilisant uniquement des opérations bit-à-bit qui sont extrêmement efficaces. Supposons que les séquences s_1 et s_2 soient codées par les entiers non signés u et v, respectivement. On forme l'entier $w := u \wedge v$ où « $\wedge$ » est l'opération « ou exclusif bit-à-bit ». Les séquences s_1 et s_2 diffèrent en une et une seule position si et seulement si

$$[w > 0] \text{ et } [(w \,\&\, (w - 1)) = 0], \tag{41}$$

où « & » est l'opération « et bit-à-bit ». En pratique, la fonction C99 suivante est utilisée

```
#include <stdbool.h>
```

```
inline bool connected(unsigned int u, unsigned int v)
{
    bool result = false;
    unsigned int w = u ^ v;
    if (w && !(w & (w - 1U)))
        result = true;
    return result;
}
```

II.4.1.7 Définition de la séquence prototype

La « séquence prototype » est définie par Bornberg-Bauer comme la séquence la plus fortement connectée d'un réseau neutre [20]. Notamment parce que nous nous limitons aux conformations compactes, cette définition n'est pas appropriée, car de nombreuses séquences possèderont le nombre maximal de connexions possibles. Il serait possible de sélectionner parmi les séquences candidates, la séquence présentant la température de repliement la plus élevée. Nous choisissons cependant une autre définition, motivée par le modèle évolutif. Bornberg-Bauer et Chan observent que la séquence prototype est également la séquence la plus peuplée à l'état stationnaire [21]. Nous définirons, par conséquent, la séquence prototype comme la séquence la plus peuplée à l'état stationnaire.

II.4.1.8 Remarques concernant le modèle

Les contraintes structurales imposées sur les protéines sur réseau rendent peu réalistes certaines de leurs caractéristiques. Le potentiel HP crée beaucoup de dégénérescence et ne rend pas compte du caractère coopératif du repliement protéique [164]. Néanmoins, un bon nombre de propriétés typiques des protéines peuvent être décrites d'après ces modèles.

Des potentiels utilisant des alphabets plus complets existent. Le groupe de Goldstein recourt habituellement à une version modifiée du potentiel mj de Miyazawa and Jernigan [136] (cf., par exemple, références [171–173]). La matrice mj donne une description beaucoup plus fine de l'énergie, mais il a été montré que sa composante principale correspond aux interactions hydrophobes et polaires [113]. En outre, le détail de la matrice mj n'est pas nécessaire pour produire les structures typiques en *superfunnel* de réseau neutre qui est ce qui nous intéresse en premier lieu. *A contrario*, les alphabets complets interdisent le calcul de réseaux neutres car l'espace des séquences devient gigantesque.

Nous avons choisi de limiter l'ensemble des conformations aux conformations compactes. À cause de cela, nous surestimons le nombre de séquences se repliant. En effet, la prise en compte de conformations non compactes est susceptible de modifier le nombre de séquences se repliant, car elles affectent le spectre d'énergie de chaque séquence. Les simu-

lations menées avec une longueur de chaîne L = 18 démontrent que l'addition des conformations non compactes réduit d'un facteur 0,02 le nombre de séquences viables [30]. Par ailleurs, les conformations compactes ne sont pas les plus modifiables dans l'étude de Bornbger-Bauer [20].

Le nombre total de conformations (compactes et non compactes) augmente exponentiellement avec la longueur, comme $2{,}63^{L}$ [185]. Si L = 25, ce nombre est rédhibitoire, il est donc nécessaire de limiter l'ensemble des conformations aux conformations les plus probables. La collapse hydrophobe rend les conformations compactes plus probables [29].

Nous pensons également qu'il existe un biais combinatoire dans la conclusion de Bornberg-Bauer concernant les conformations modifiables. Elle provient du fait qu'une conformation compacte de longueur L = 18 possède *moins* de contacts entre les résidus enfouis que certaines conformations non compactes. Nous appellons « résidu enfoui » un résidu formant plus de deux contacts. Un contact entre résidus enfouis est un contact s'établissant entre deux résidus enfouis. Nous reproduisons dans la figure II.10, les deux conformations citées en exemple dans la figure 1 de l'article [20]. La conformation de gauche est compacte et sept de ses dix contacts sont formés entre des résidus enfouis. En revanche, la conformation de droite n'est pas compacte mais huit de ses neuf contacts s'établissent entre résidus enfouis.

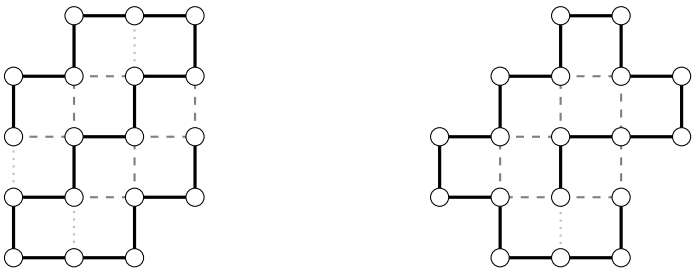


Fig. II.10 : Deux exemples de conformation sur réseau de longueur L = 18. La conformation de gauche possède dix contacts et est compacte. La conformation de droite possède neuf contacts et n'est donc pas compacte. Les contacts entre résidus enfouis sont indiqués par une ligne discontinue, les autres contacts sont en ligne pointillée et claire (d'après Bornberg-Bauer [20]).

Le nombre de contacts entre résidus enfouis est important dans la stabilité d'une conformation. Quand L = 25, ce sont bien les conformations compactes qui assurent le nombre maximal de contacts entre résidus enfouis.

II.4.2 Protéines tridimensionnelles

Les modèles de protéines sur réseau sont extrêmement utiles. Ils permettent d'explorer exhaustivement l'espace des séquences et l'espace des structures. Bien qu'ils capturent

des caractéristiques des protéines réelles, les conclusions auxquelles ils mènent demandent à être confirmées par des modèles plus réalistes. À cause des difficultés qu'ils impliquent, les modèles de protéine hors-réseau restent pour l'instant minoritaires dans les simulations évolutives [204]. Mais le besoin de les développer se fait sentir. Ainsi, la disparité de la modifiabilité des structures a émergé d'un modèle de protéine hors-réseau [130]. Cela confirme l'existence de l'inégalité des modifiabilités, indépendamment du modèle structural. Le modèle que nous présenterons aura non seulement pour but de *corroborer* les résultats obtenus avec les protéines sur réseau, mais il permettra d'*étendre* notre analyse puisque des alphabets plus complets seront considérés. Le recours à des protéines tridimensionnelles issues de la PDB, s'accompagne nécessairement d'un changement de stratégie que nous allons détailler.

Le modèle évolutif est identique à celui des protéines sur réseau. Les structures que nous allons étudier se limitent à trois squelettes peptidiques issus de la PDB. Les séquences viables sont celles qui se replient dans lesdites conformations. Elles ne sont plus trouvées par énumération exhaustive mais générées par une méthode de Monte Carlo. En effet, nous ne nous restreignons plus à deux types de résidu d'acide aminé, mais considérons les vingt acides aminés. L'espace des séquences devient donc astronomique et exclut d'emblée toute exhaustivité.

Pour la même raison, la construction des réseaux neutres fait intervenir une simplification supplémentaire. Les séquences générées sont projetées dans un « espace des profils » où chaque acide aminé est assimilé à sa classe dans un certain alphabet.

La capacité d'une séquence à se replier dans l'une des conformations natives sera évaluée en comparant l'énergie de la séquence repliée en la conformation native à celle de la séquence repliée en les conformations d'un ou de deux jeux de « fausses structures ».

II.4.2.1 Conformations et séquences natives

Trois conformations natives ont été sélectionnées parce qu'elles allient des propriétés de protéines réalistes à des longueurs de chaîne suffisamment petites pour se prêter aux calculs d'énergie et à l'exploration de l'espace des séquences. Nous appellerons « conformation native » chacune de ces conformations et « séquence native » la séquence qui lui correspond dans la PDB.

TRP-cage La protéine 1L2Y, le fragment TRP-cage, est une petit peptide de 20 résidus d'acide aminé [143]. Sa séquence native est

NLYIQ WLKDG GPSSG RPPPS
5 10 15 20

Parmi les vingt acides aminés, il y a quatre prolines. Les prolines ne seront pas soumises aux mutations, à cause de leurs propriétés particulières dans la formation des structures

secondaires et surtout de leur rôle clef dans la stabilité de la structure [143]. La structure tertiaire du trp-cage est présentée dans la figure II.11.

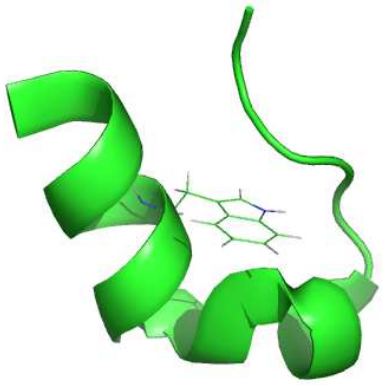


Fig. II.11 : Le fragment peptidique trp-cage (1L2Y). Le résidu en position 6 est un trp qui est important pour la stabilité de la conformation.

Domaines SH3 de Grb2 et Vav Le domaine SH3 de Grb2 (chaîne B) et celui de Vav (chaîne C) issus de l'entrée 1GCQ de la pdb sont deux protéines de 57 et 69 résidus d'acide aminé respectivement [145]. La séquence native du domaine SH3 de Grb2 est

```
                163                 168                 173                 178                 183                 188                 193                 198                 203                 208
TYVQA LFDFD PQEDG ELGFR RGDFI HVMDN SDPNW WKGAC HGQTG MFPRN
213
YVTPV NR
```

Celle du domaine SH3 de Vav est

```
                595                 600                 605                 610                 615                 620                 625                 630                 635                 640
GSHMP KMEVF QEYYG IPPPP GAFGP FLRLN PGDIV ELTKA EAEHN WWEGR
645                 650                 655
NTATN EVGWF PCNRV HPYV
```

Ces deux domaines interagissent l'un avec l'autre. La surface accessible au solvant totale perdue lors de l'interaction s'élève à 1 270 Å^2. Les conséquences de cette interaction seront étudiées dans la partie « Évolution des protéines dimériques » (p. 177). Une vue stéréoscopique des chaînes de 1GCQ est présentée dans la figure II.12.

II.4.2.2 États dénaturés

Un lot de conformations issu de la pdb sert de matrice à la préparation de « fausses conformations » qui miment l'ensemble des états dénaturés[15]. L'ensemble des fausses struc-

15. Les entrées de la pdb qui nous ont servi furent fournies par Launay : 1ds1, 1a6b, 1a7m, 1a9v, 1ab3, 1adn, 1agg, 1ah9, 1aml, 1aoy, 1aq5, 1axh, 1ayj, 1b16, 1b6f, 1b8o, 1b9u, 1bak, 1bc9, 1bci, 1bj8, 1bkr, 1bl1, 1bm4, 1bmx, 1bo9, 1bpv, 1bqv, 1br0, 1buy, 1bvh, 1bw6, 1byi, 1byq, 1bzg, 1c05, 1c1k, 1c20, 1c4e, 1c5e, 1c75, 1c7k, 1c9q, 1cdb, 1cdq, 1ce4, 1cf4, 1cfe, 1cg7, 1chc, 1ckv, 1clh, 1cmr, 1co4, 1cok, 1coo, 1cou, 1cwx, 1d1h, 1d6g, 1dbf, 1dci, 1dec, 1dfe, 1dgn, 1dj0, 1dl0, 1dlx, 1dp3, 1dp7, 1dpu, 1dqc, 1dvj, 1dxz, 1e01, 1e0l, 1e3t,

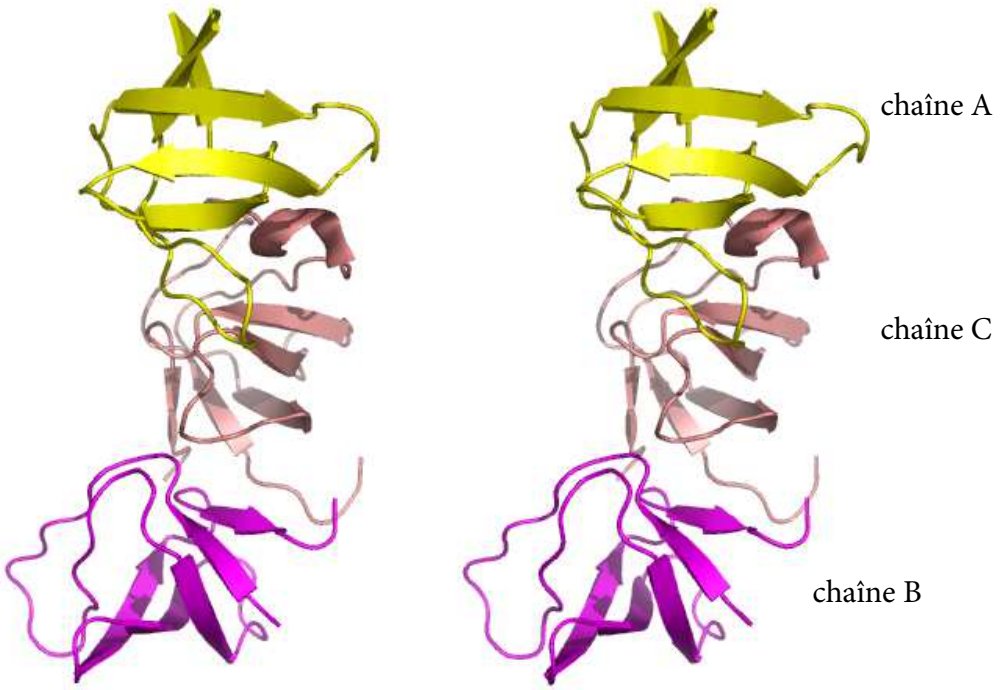


Fig. II.12 : Vue stéréoscopique de 1GCQ. Les chaînes B et C sont les domaines SH3 de Grb2 de Vav que nous allons étudier.

tures est noté $\mathbf{D}^{(1)}$. Seule l'information des squelettes peptidiques est nécessaire. Si une protéine matrice de la PDB est composée de L' résidus, le squelette des différents fragments $(1, \ldots, L)$, $(1+\Delta, \ldots, L+\Delta)$, $(1+2\Delta, \ldots, L+2\Delta)$, etc., de la protéine sont retenus pour former des fausses structures. Le paramètre Δ est un décalage qui vaut 15. Le nombre de fausses structures formées pour cette protéine matrice est $(L'-L)/\Delta$.

Pour les domaines SH3 de Grb2 et Vav, plus grands, un autre jeu de fausses structures $\mathbf{D}^{(2)}$ est produit, car trop peu de contacts pertinents sont produits par les squelettes de l'ensemble $\mathbf{D}^{(1)}$. Ce sont des structures générées à partir de la séquence biologique dans sa conformation native c_0 soumise à une dynamique moléculaire *in vacuo* réalisée par CHARMm (T = 310 K, 1 000 000 pas). Ces deux jeux de fausses structures permettent

1E4U, 1E53, 1E6U, 1E7L, 1ECI, 1EDX, 1EF4, 1EGX, 1EHS, 1EIK, 1EIT, 1EMW, 1ENW, 1EO0, 1EO1, 1EP0, 1EQO, 1ERD, 1ES9, 1EUW, 1EV0, 1EWI, 1EWS, 1EXK, 1EZG, 1F0Z, 1F53, 1F81, 1FBR, 1FCT, 1FDM, 1FHO, 1FJ2, 1FM0, 1FM0, 1FMH, 1FP0, 1FRE, 1FU9, 1FVL, 1FW9, 1FWO, 1FYC, 1FYJ, 1G25, 1G4F, 1G6E, 1G7E, 1G84, 1GAB, 1GD0, 1GGW, 1GHH, 1GNC, 1GYF, 1HA9, 1HBW, 1HDO, 1HHN, 1HS7, 1HX2, 1HYK, 1HYW, 1I1S, 1I27, 1I6W, 1ICA, 1IJA, 1IRL, 1IRS, 1ISU, 1ITF, 1IXH, 1JWE, 1KHM, 1KRS, 1KSR, 1LYP, 1MKC, 1MKN, 1MLA, 1MUT, 1NGL, 1NGR, 1NKL, 1NLS, 1PAA, 1PEH, 1PFS, 1PIH, 1PNB, 1PNB, 1PNB, 1PNB, 1PON, 1PRR, 1QA5, 1QDP, 1QFT, 1QH4, 1QHK, 1QK6, 1QK7, 1QKF, 1QLO, 1QM9, 1QOP, 1QOP, 1QP6, 1QQF, 1QTN, 1QTS, 1QU5, 1R2A, 1RES, 1RGE, 1RIE, 1RIP, 1SCY, 1SVF, 1SVF, 1SVF, 1SWU, 1TBA, 1TBA, 1TFB, 1TSG, 1XBL, 1XPA, 1YUA, 1ZTA, 1ZTO, 2BID, 2CTC, 2FMR, 2HGF, 2IFO, 2LIS, 2NLR, 3CHB, 3SIL, 4EUG, 1BXO, 1CEX, 8PRN, 1YCS, 1GCI, 1BXO, 7A3H, 1NLS, 1IXH, 1GA6, 2NLR, 1HYO, 1BK0, 1B0U, 1B8O, 1JYK, 1EKQ, 1JL0, 1KEP, 1E0C, 1AL3, 1GDO, 1AOL, 1F2D, 1CUN, 1ES6, 1BX7 et 1A6M.

de prendre en compte, d'un côté, l'accessibilité cinétique lors de la formation du *molten globule* par collapse hydrophobe, et d'un autre, les positionnements fins dans les dernières étapes du repliement.

II.4.2.3 Enfilage des séquences et calcul de l'énergie

Enfilage des séquences Contrairement à ce qui a été fait pour les protéines sur réseau, nous ne nous bornons pas à des séquences de type HP. Une séquence peut être n'importe quelle L-uplet de A = 20 lettres choisies parmi {A, C, D, E, F, G, H, I, K, L, M, N, P, Q, R, S, T, V, W, Y}, où L désigne la longueur de chaîne de la protéine d'intérêt. En outre, nous tiendrons compte de l'extension des chaînes latérales lors de la formation des contacts.

L'enfilage d'une séquence $s = (s_1, \ldots, s_L)$ sur une conformation c est réalisé en greffant sur le squelette peptidique de c le rotamère le plus fréquent de la bibliothèque de Tuffery *et al.* [182]. Ainsi, toutes les coordonnées de la protéine sont reconstituées.

Deux acides aminés sont en contact dans une chaîne s'ils sont distants de moins de 4,5 Å. La distance entre deux résidus est calculée en prenant le minimum de toutes les distances obtenues obtenues en considérant leurs atomes lourds. Par ailleurs, nous excluons les contacts entre résidus séparés par moins de deux positions dans la chaîne.

Énergie d'une séquence L'énergie de la séquence s dans la conformation c est la somme des interactions entre résidus en contact. On peut résumer comme dans l'équation 36 :

$$\mathrm{E}(s, c) = \sum_{i \leq j+3} \chi(c_i, c_j)\, \epsilon(s_i, s_j)\,. \tag{42}$$

où $\chi(c_i, c_j) = 1$ si les positions i et j de la chaîne sont en contact lorsque l'on greffe les acides aminés s_i et s_j sur le squelette peptidique, et $\chi(c_i, c_j) = 0$ sinon. La valeur $\epsilon(s_i, s_j)$ représente l'interaction entre les acides aminés s_i et s_j quand ils sont en contact ; elle ne dépend que de la classe des acides aminés.

Pour une séquence s donnée, on note $\Delta\mathrm{E}$ la différence entre l'énergie de l'état fondamental $\mathrm{E}(s, c_0)$ et le premier état excité. Pour les domaines SH3, on calcule séparément les valeurs $\Delta\mathrm{E}^{(1)}$ et $\Delta\mathrm{E}^{(2)}$ où le premier état excité est choisi parmi les conformations des jeux $\mathbf{D}^{(1)}$ et $\mathbf{D}^{(2)}$.

On définit le « Z-score » par la grandeur

$$\mathrm{Z}(s) = \frac{\mathrm{E}(s, c_0) - \langle \mathrm{E} \rangle}{\sigma},$$

où $\langle \mathrm{E} \rangle$ et σ sont la moyenne et l'écart type du spectre d'énergie de la séquence s. Pour les domaines SH3, on calcule séparément

$$\mathrm{Z}^{(1)}(s) = \frac{\mathrm{E}(s, c_0) - \langle \mathrm{E} \rangle^{(1)}}{\sigma^{(1)}},$$
$$\mathrm{Z}^{(2)}(s) = \frac{\mathrm{E}(s, c_0) - \langle \mathrm{E} \rangle^{(2)}}{\sigma^{(2)}},$$

où $\langle \mathrm{E} \rangle^{(i)}$ et $\sigma^{(i)}$ sont la moyenne et l'écart type du spectre d'énergie de la séquence s obtenu avec le jeu de fausses structures $\mathbf{D}^{(1)}$.

Le Z-score mesure l'éloignement de l'énergie de la native par rapport à la moyenne en unités d'écart type. Plus le Z-score est négatif, plus la structure native est stable [1, 22, 66].

II.4.2.4 Classes d'acides aminés, matrice d'interaction

Classes d'acides aminés Les classes d'acides aminés utilisées ont été inspirées par Murphy *et al.* [141] puis corrigées et complétées par Launay *et al.* [106]. Nous noterons de façon générique mC un alphabet à m classes. Cinq alphabets sont considérés : 20C, 6C, 4C, 3C et 2C. La figure II.13 en donne la description exacte et illustre leur organisation hiérarchique.

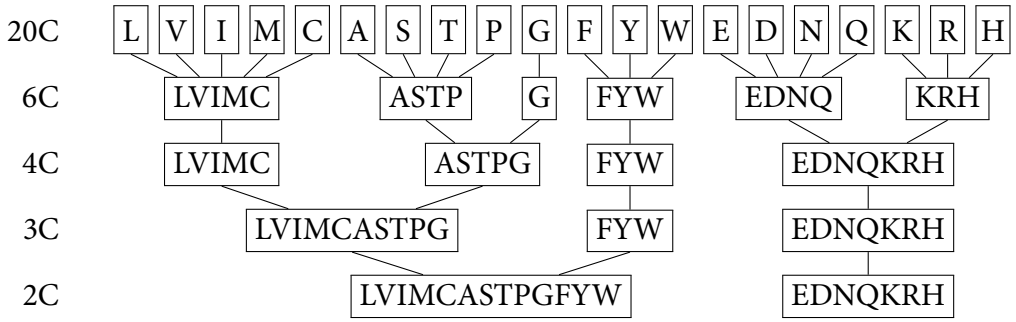


Fig. II.13 : Classification des acides aminés obtenue par Launay *et al.* en utilisant un groupage hiérarchique s'appuyant sur la similitude des lignes de la matrice blosum62 [106].

Nous appellerons le « profil » de la séquence $s = (s_1, \ldots, s_{\mathrm{L}})$ le L-uplet

$$p = \big(\mathrm{C}(s_1), \ldots, \mathrm{C}(s_{\mathrm{L}})\big)$$

où $\mathrm{C}(s_i)$ est la classe de l'acide aminé s_i. La conversion de la séquence

```
ANCKV YSWNT RKGLA FEVI
```

donne dans l'alphabet à deux classes (HP) :

```
HPHPH HHHPH PPHHH HPHH
```

Matrices d'interaction Des matrices correspondant à chaque alphabet sont disponibles. Elles furent déduites d'une procédure d'optimisation mise au point, au laboratoire, par Launay *et al.* d'après un travail de Bastolla *et al.* [5]. Nous décrivons brièvement la procédure suivie.

L'énergie d'une séquence s dans une conformation c est calculée conformément à l'équation 42. L'objectif de la méthode développée par Launay et ses collaborateurs est de trouver les valeurs optimales de $\epsilon(s_i, s_j)$ pour que les structures natives des séquences aient des énergies plus basses qu'un ensemble de fausses structures.

La similitude entre deux conformations c et c' est mesurée par le « recouvrement structural » $\mathrm{Q}(c, c')$ qui est le nombre de contacts entre résidus communs aux deux conformations c et c' divisé par le nombre de contacts maximal :

$$\mathrm{Q}(c, c') = \frac{\text{nombre de contacts communs}}{\text{nombre maximal de contacts}}. \tag{43}$$

La grandeur $\mathrm{Q}(c, c')$ est comprise entre zéro (les deux conformations n'ont aucun contact en commun) et un (tous les contacts sont communs).

Pour une séquence s et sa conformation native c_0, on définit la moyenne de Boltzmann, $\langle \mathrm{Q} \rangle_{s,c_0}$ par

$$\langle \mathrm{Q} \rangle_{s,c_0} = 1/\mathrm{Z} \sum_c \mathrm{Q}(c_0, c)\ \exp\Big(-\frac{\mathrm{E}(s, c)}{k\,\mathrm{T}}\Big),$$

où la somme comprend toutes les fausses structures et la structure native, et où Z est la fonction de partition. Une valeur $\langle \mathrm{Q} \rangle_{s,c_0}$ proche de un indique que les paramètres ϵ de la matrice d'interaction favorisent les structures proches de la structure native.

L'optimisation est conduite sur un jeu S de paires (s, c_0) de séquences et leur conformation native pour lequel on tente de maximiser la quantité

$$\langle \mathrm{Q} \rangle = \sum_{\mathrm{S}} \langle \mathrm{Q} \rangle_{s,c_0}$$

par une méthode du gradient.

Les deux matrices suivantes seront les plus largement utilisées :

$$\mathcal{E}_{2\mathrm{C}} = \begin{array}{c} \\ \mathrm{H} \\ \mathrm{P} \end{array} \begin{array}{c} \begin{array}{cc} \mathrm{H} & \mathrm{P} \end{array} \\ \begin{pmatrix} -8.5 & 9 \\ 9 & -3.5 \end{pmatrix} \end{array}$$

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

$$\mathcal{E}_{3C} = \begin{array}{c} \\ \mathrm{H} \\ \mathrm{A} \\ \mathrm{P} \end{array} \begin{array}{c} \begin{array}{ccc} \mathrm{H} & \mathrm{A} & \mathrm{P} \end{array} \\ \begin{pmatrix} -4.73 & -9.60 & 6.39 \\ -9.60 & -11.37 & 1.53 \\ 6.39 & 1.53 & -1.87 \end{pmatrix} \end{array}$$

et H = {L, V, I, M, C, A, S, T, P, G}, A = {F, Y, W}, and P = {E, D, N, Q, K, R, H}. Le tableau II.2 présente les valeurs relatives à l'alphabet complet 20C.

Il est important de noter que nos simulations prennent place dans des espaces à vingt acides aminés quelle que soit la matrice utilisée. Deux séquences partageant le même profil, peuvent posséder des spectres d'énergie différents. En effet, même si deux acides aminés appartiennent à la même classe, une différence dans les contacts formés peut résulter de la différence de la géométrie dans l'espace de leur chaîne latérale.

II.4.2.5 Trajectoires

Les paragraphes précédents ont développé les concepts nécessaires à la formation d'une séquence et à l'évaluation de son énergie. Nous allons décrire, à présent, la méthode de Monte Carlo utilisée pour générer des séquences. Les séquences viables devront se replier dans la conformation native, ce que nous évaluerons à l'aide des grandeurs ΔE, $Z(s)$ définies plus haut.

Nous notons c_0 l'une des trois conformations natives proposées plus haut. Nous disposons d'une « séquence de référence » s_0 se repliant en la conformation c_0. La séquence de référence possède une double fonction. Elle sert de « graine » à la génération des séquences et elle fixe les paramètres d'acceptation des séquences générées. Une procédure de Monte Carlo génère une suite de séquences $(s_1, \ldots, s_N)$, en exécutant les règles suivantes :

for $t := 0$ à N **do**
  Créer une séquence mutante s_t^* par mutation ponctuelle de la séquence s_t
  if s_t^* se replie **then**
    $s_{t+1} := s_t^*$
  else
    $s_{t+1} := s_t$
  end if
end for

La condition « s_t^* se replie » résume plusieurs contraintes. Notons $\Delta E_0^{(1)}$ et $\Delta E_0^{(2)}$ les différences d'énergie de la séquence s_0 entre l'état fondamental et les premiers états excités

	ALA	ARG	ASN	ASP	CYS	GLN	GLU	HIS	ILE	LEU	LYS	MET	PHE	PRO	SER	THR	TRP	TYR	VAL	GLY
ALA	−0.15	0.48	2.44	4.35	0.06	2.28	1.67	2.57	−0.76	−2.42	3.59	−2.06	−4.29	13.40	0.78	3.74	6.80	−3.97	−5.60	3.38
ARG	0.48	8.78	−1.91	−4.08	3.80	3.79	−10.34	1.74	3.78	4.15	8.13	−1.45	1.03	5.31	−0.52	5.73	3.10	−5.35	6.18	0.35
ASN	2.44	−1.91	0.52	−1.22	0.31	3.93	−1.84	2.80	7.83	−2.96	−4.48	1.62	0.53	1.77	−1.07	−0.80	0.17	7.77	0.91	−0.77
ASP	4.35	−4.08	−1.22	5.06	−1.58	0.34	12.51	−0.97	6.97	13.54	−12.12	2.07	6.93	7.03	0.52	−2.46	3.19	0.61	7.28	7.88
CYS	0.06	3.80	0.31	−1.58	−28.90	−2.34	−5.07	−2.18	−4.07	−6.30	−0.71	−4.40	−7.46	2.23	5.04	0.63	0.06	−3.51	−3.13	−8.00
GLN	2.28	3.79	3.93	0.34	−2.34	0.87	−1.72	3.63	7.24	4.57	−3.30	−3.23	1.93	2.39	−0.40	−0.47	−2.15	−3.24	−1.48	5.67
GLU	1.67	−10.34	−1.84	12.51	−5.07	−1.72	7.05	−0.42	5.33	2.75	−12.12	−0.10	3.97	6.07	4.72	4.28	0.52	3.44	12.91	8.54
HIS	2.57	1.74	2.80	−0.97	−2.18	3.63	−0.42	−0.29	−3.84	0.33	4.31	−0.19	−0.87	0.08	1.84	0.73	−2.68	−3.12	2.27	2.88
ILE	−0.76	3.78	7.83	6.97	−4.07	7.24	5.33	−3.84	−11.69	−18.95	9.19	−10.33	−9.14	4.62	−1.16	−3.15	−6.00	−11.38	−11.35	1.07
LEU	−2.42	4.15	−2.96	13.54	−6.30	4.57	2.75	0.33	−18.95	−13.70	5.28	−6.49	−8.52	8.56	4.24	2.43	−6.86	−8.31	−11.47	3.74
LYS	3.59	8.13	−4.48	−12.12	−0.71	−3.30	−12.12	4.31	9.19	5.28	10.07	5.43	5.11	0.55	4.04	5.28	5.29	−7.86	4.10	1.40
MET	−2.06	−1.45	1.62	2.07	−4.40	−3.23	−0.10	−0.19	−10.33	−6.49	5.43	−0.24	−9.80	0.69	1.25	0.79	−2.57	−5.13	−7.52	−4.14
PHE	−4.29	1.03	0.53	6.93	−7.46	1.93	3.97	−0.87	−9.14	−8.52	5.11	−9.80	−12.20	4.42	1.81	−2.01	−3.17	−7.06	−8.21	2.26
PRO	13.40	5.31	1.77	7.03	2.23	2.39	6.07	0.08	4.62	8.56	0.55	0.69	4.42	0.30	13.95	8.62	−3.94	−6.70	4.81	4.10
SER	0.78	−0.52	−1.07	0.52	5.04	−0.40	4.72	1.84	−1.16	4.24	4.04	1.25	1.81	13.95	4.23	4.01	1.17	−5.47	0.20	3.64
THR	3.74	5.73	−0.80	−2.46	0.63	−0.47	4.28	0.73	−3.15	2.43	5.28	0.79	−2.01	8.62	4.01	0.12	−2.63	3.11	−0.09	−7.25
TRP	6.80	3.10	0.17	3.19	0.06	−2.15	0.52	−2.68	−6.00	−6.86	5.29	−2.57	−3.17	−3.94	1.17	−2.63	−4.92	−9.33	−7.39	2.72
TYR	−3.97	−5.35	7.77	0.61	−3.51	−3.24	3.44	−3.12	−11.38	−8.31	−7.86	−5.13	−7.06	−6.70	−5.47	3.11	−9.33	−7.70	−2.13	−11.56
VAL	−5.60	6.18	0.91	7.28	−3.13	−1.48	12.91	2.27	−11.35	−11.47	4.10	−7.52	−8.21	4.81	0.20	−0.09	−7.39	−2.13	−19.02	−0.51
GLY	3.38	0.35	−0.77	7.88	−8.00	5.67	8.54	2.88	1.07	3.74	1.40	−4.14	2.26	4.10	3.64	−7.25	2.72	−11.56	−0.51	−1.05

Tab. II.2 : Matrice d'interaction pour l'alphabet complet (20C).

des jeux $\mathbf{D}^{(1)}$ et $\mathbf{D}^{(2)}$ respectivement. Notons $Z_0^{(1)}$ et $Z_0^{(2)}$ les Z-scores du spectre d'énergie de s_0 calculés sur les jeux $\mathbf{D}^{(1)}$ et $\mathbf{D}^{(2)}$ respectivement. Les contraintes pour que la séquence s_t^* se replie sont les suivantes.

1. L'énergie $E(s_t^*, c)$ doit être minimale quand s_t^* est dans la conformation native $c = c_0$. Pour toute conformation de $\mathbf{D}^{(1)}$ et $\mathbf{D}^{(2)}$, on doit avoir

$$E(s_t^*, c_0) < E(s_t^*, c).$$

2. La différence d'énergie entre l'état fondamental et le premier état excité du jeu $\mathbf{D}^{(1)}$ doit être supérieure à $\Delta E_0^{(1)}$.

$$\forall c \in \mathbf{D}^{(1)}, E(s_t^*, c_0) + \Delta E_0^{(1)} \leq E(s_t^*, c).$$

3. La différence d'énergie entre l'état fondamental et le premier état excité du jeu $\mathbf{D}^{(2)}$ doit être supérieure à $\Delta E_0^{(2)}$.

$$\forall c \in \mathbf{D}^{(2)}, E(s^*, c_0) + \Delta E_0^{(2)} \leq E(s^*, c).$$

4. Le Z-score de la séquence s_t^* calculé sur le jeu $\mathbf{D}^{(1)}$ doit être inférieur à $Z_0^{(1)}$.

$$Z^{(1)}(s^*) \leq Z_0^{(1)}.$$

5. Le Z-score de la séquence s_t^* calculé sur le jeu $\mathbf{D}^{(2)}$ doit être inférieur à $Z_0^{(2)}$.

$$Z^{(2)}(s^*) \leq Z_0^{(2)}.$$

Pour le trp-cage qui ne dispose que d'un seul jeu de fausses structures, le point 1 ne prend en compte que $\mathbf{D}^{(1)}$. De même, les points 3 et 5 ne s'appliquent pas pour le trp-cage.

II.4.2.6 Préoptimisation

La séquence de référence fixe la limite des séquences qui peuvent être acceptées par la procédure de Monte Carlo. Les séquences menant à une stabilité moindre que celle conférée par s_0 sont refusées, celles menant à une plus grande stabilité sont acceptées.

L'utilisation de la séquence native (c'est-à-dire issue de la pdb) comme séquence de référence n'est pas satisfaisante, car elle mène à l'acceptation de trop de séquences (environ une séquence sur deux est acceptée). L'objet de la procédure que nous allons décrire, la « préoptimisation », est de trouver une séquence de référence limitant suffisamment l'exploration du réseau neutre.

La procédure de « préoptimisation » est analogue à une méthode de repliement inverse. Nous pouvons résumer cette étape ainsi : puisque le modèle énergétique n'est pas adapté à la séquence biologique, nous adaptons la séquence de référence au modèle énergétique. Le principe de la préoptimisation est schématisé dans la figure II.14.

Nous effectuons pour cela une courte trajectoire (quelques milliers de pas) et nous recherchons parmi les séquences acceptées celle qui possède une stabilité accrue lorsqu'elle est repliée en la conformation c_0. Plusieurs cycles peuvent être nécessaires avant que la stabilité soit suffisante.

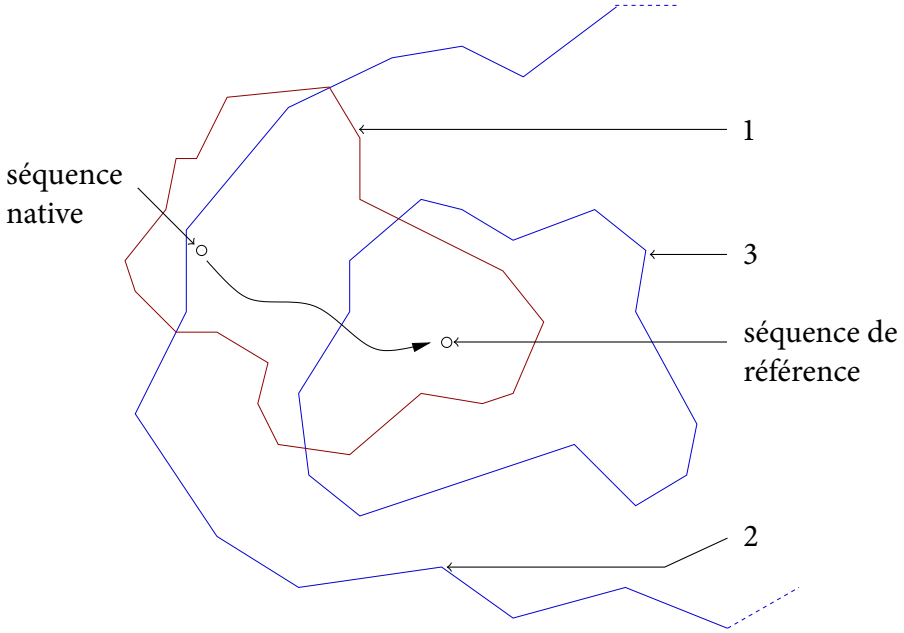
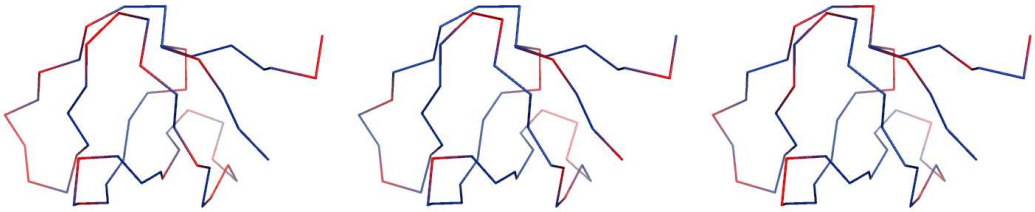


Fig. II.14 : Principe de la préoptimisation. En rouge (1), le véritable réseau neutre qui serait exploré avec un modèle d'énergie fin. En bleu (2), le réseau neutre exploré avec notre modèle d'énergie en utilisant la séquence biologique comme référence est beaucoup trop vaste. Il ne possède certainement pas des caractéristiques réalistes du réseau neutre réel (1). La préoptimisation trouve une séquence de référence qui permet d'échantillonner un réseau neutre (3) qui possède des caractéristiques du réseau neutre réel.

L'effet de la préoptimisation est un caractère plus hydrophobe du cœur de la protéine et plus hydrophile de la surface de la protéine. La figure II.15 représente les acides aminés hydrophobes et hydrophiles de Grb2 après différentes étapes de préoptimisation.

II.4.2.7 Profils, réseaux neutres

À cause de l'immensité de l'espace des séquences, il n'est pas possible de reconstruire de réseau neutre à partir des séquences acceptées au cours d'une trajectoire. Les réseaux neutres sont donc construits à partir de l'ensemble des profils obtenus à partir des séquences viables.



Cycle 2 Cycle 4 Cycle 6

Fig. II.15 : Effet de la préoptimisation sur la ségrégation des résidus hydrophobes et hydrophiles au cours de différents cycles de préoptimisation de Grb2. Les résidus hydrophobes sont en bleu, les hydrophiles sont en rouge.

Soient s_1 et s_2 deux séquences viables et p_1 et p_2 les profils associés. Les deux profils p_1 et p_2 sont connectés s'ils diffèrent l'un de l'autre en une seule position. En général, les séquences s_1 et s_2 ne sont pas connectées, même si leurs profils le sont.

II.4.2.8 Schéma global

Les procédures décrites dans ce chapitre étant relativement complexes, nous nous proposons d'en résumer ici le schéma global. Nous notons c_0 l'une des conformations natives parmi le trp-cage, le domaine SH3 de Grb2 et le domaine SH3 de Vav. La séquence s_{nat} est la séquence native de c_0 et s_0 désigne la séquence de référence.

Préoptimisation Une ou plusieurs courtes trajectoires sont réalisées à partir de la séquence native s_{nat} jusqu'à rencontrer une séquence s_0 se repliant de façon stable en c_0.

Trajectoire de production Une trajectoire de production est initiée en se servant de la séquence s_0 comme séquence de référence. Un ensemble de séquences viables $\mathbf{s}$ est généré.

Construction du réseau neutre Nous construisons le réseau neutre à partir de l'ensemble $\mathbf{p}$ des profils des séquences viables $\mathbf{s}$. Le réseau neutre sert de support à l'évolution telle qu'elle est décrite dans « Modèle évolutif ».

La figure II.16 schématise ces trois étapes.

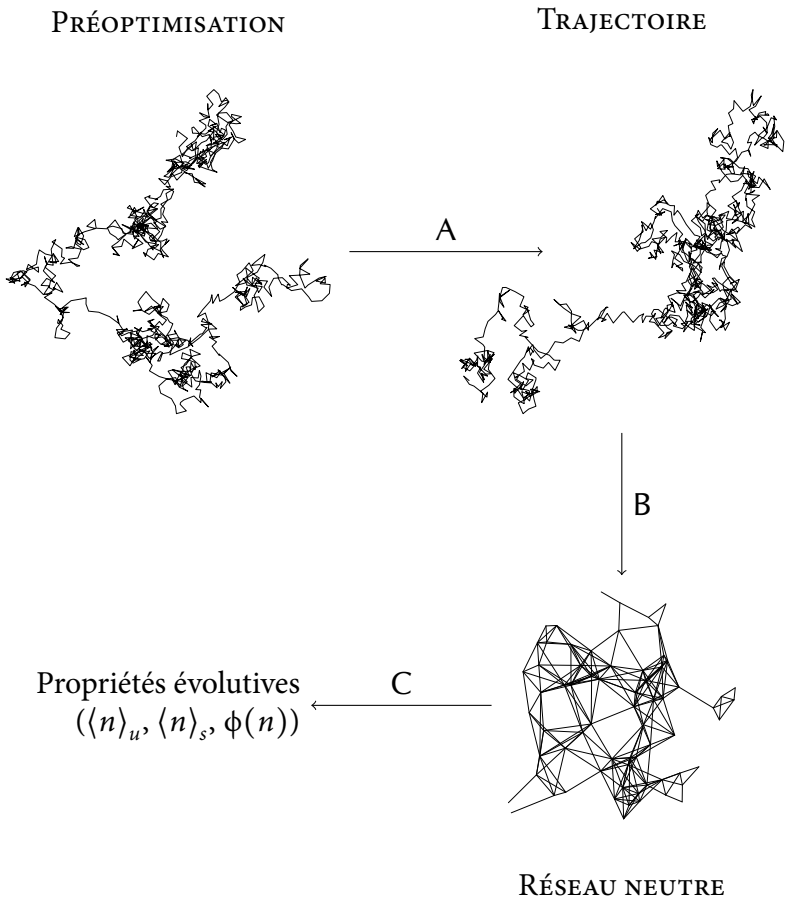


Fig. II.16 : Schéma des étapes menant de la préoptimisation à la reconstruction des réseaux neutres. A — Une séquence de référence est extraite d'une trajectoire de préoptimisation. B — La séquence de référence trouvée initie une trajectoire générant des séquences aléatoirement et les acceptant ou les rejetant selon les conditions exposées dans « Trajectoires » (p. 80). C — Les séquences acceptées sont converties en profils et la structure de réseau neutre est construite en pontant les profils ne différant qu'en une seule position.

Chapitre III

Évolution des protéines monomériques

Era de las que rompen los puentes con solo cruzarlos…

Julio Cortázar, *Rayuela*

III.1 Introduction

La modélisation de la fonctionnalité est l'un des enjeux des simulations récentes et futures. Dans l'introduction, nous avons annoncé que l'inclusion des interactions protéine-protéine rendait possible cette modélisation. Nous nous intéresserons, dans cette thèse, aux hétérodimères. Dans un dimère, chaque monomère doit généralement se replier de façon stable. Dans notre modèle, les séquences capables de dimériser constitueront donc un *sous-ensemble* des séquences capables de se replier. Par conséquent, notre thèse débute par l'étude des protéines monomériques, c'est-à-dire des séquences dont la viabilité n'est évaluée que par la stabilité du repliement dans une conformation cible. L'évolution est neutre, car nous ne sélectionnons pas positivement une séquence pour sa stabilité : une séquence modérément stable est viable tout autant qu'une séquence très stable.

Deux modèles seront présentés et plus tard étendus pour tenir compte des interactions protéine-protéine.

— La première partie détaille les résultats d'un modèle de protéine sur réseau fortement inspiré de travaux antérieurs. Le sujet a été déjà abondamment couvert. Une partie de ces résultats peut donc être retrouvée dans l'un ou l'autre des ar-

ticles qui ont été ou seront cités. Nous produirons cependant un grand nombre de résultats originaux.

— La seconde partie développe les résultats d'un modèle d'évolution de protéine tridimensionnelle hors-réseau. Ce type d'approche a été nettement moins utilisé et la totalité des résultats est originale.

Nous conclurons ce chapitre en étudiant de manière plus approfondie la validité de l'hypothèse de la population infinie.

III.2 Résultats

III.2.1 Protéines sur réseaux

Les séquences considérées dans cette partie sont composées de résidus hydrophobes (H) ou polaires (P). Les séquences viables sont celles qui se replient en une conformation cible qui est leur état d'énergie fondamental non dégénéré. Les séquences viables, connectées par des mutations ponctuelles, forment un réseau neutre. La majorité des résultats est consacrée à trois matrices typiques des protéines : LHTW, HP et HP′. Certaines analyses seront étendues à d'autres matrices en annexe.

Nous étudierons, dans un premier temps, les statistiques sur le repliement et sur les réseaux neutres. Nous relèverons quelles sont les propriétés thermodynamiques des séquences appartenant à un réseau neutre et, en particulier, comment elles s'organisent en *superfunnel*. Ensuite, nous examinerons les propriétés de l'évolution neutre au sein d'un réseau neutre et de son impact sur la robustesse et la stabilité des séquences présentes dans une population infinie. Enfin, une annexe présentera quelques résultats plus anecdotiques, en particulier ceux obtenus pour les matrices non typiques des protéines.

III.2.1.1 Statistiques de repliement, réseaux neutres

Statistiques de repliement Nous avons calculé les énergies des 16 781 312 séquences (cf. table II.1) enfilées sur les 1 081 conformations et établi quelles séquences se repliaient et, le cas échéant, quelle était leur conformation native. Nous avons construit les réseaux neutres de chaque conformation pour chaque matrice d'énergie. Le tableau III.1 présente le nombre de séquences qui se replient, le nombre de réseaux neutres, la taille et l'identifiant du plus grand réseau neutre.

Le nombre total de séquences s'élève à $2^{25} \approx 33 \cdot 10^6$. La fraction de séquences se repliant varie de 18 %, pour la matrice HP, à 36 % pour la matrice LHTW. Les réseaux neutres sont plus nombreux quand le nombre de séquences se repliant est faible. Dans les

trois cas étudiés, le plus grand réseau neutre formé correspond à la conformation 786 (reproduite dans la figure III.1). Cette conformation possède, d'après Li *et al.*, des régularités réminiscentes des structures secondaires des protéines [112].

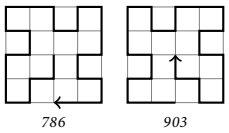


Fig. III.1 : Les conformations natives des plus grands réseaux neutres en utilisant la matrice HP, HP′ et LHTW : 786 et 903, inverses l'une de l'autre.

Entre environ 3 000 et 7 000 réseaux neutres ont pu être construits. Chacune des 1 081 conformations génère plusieurs réseaux neutres. La figure III.2 présente l'histogramme des tailles des plus grands réseaux neutres de chaque conformation, l'histogramme des tailles des seconds plus grands réseaux neutres, etc.

On voit que les réseaux neutres de chaque conformation ne sont donc pas égaux. Généralement, le plus grand réseau neutre d'une conformation c regroupe l'essentiel des séquences se repliant en c, les autres réseaux neutres sont de taille beaucoup plus réduite. Dans notre modèle, tout du moins, l'étude des séquences se repliant en une conformation revient pratiquement à étudier le plus grand réseau neutre existant pour cette conformation (voir aussi références [161, 181]).

Modifiabilité La modifiabilité d'une conformation (*designability*) caractérise l'aptitude de cette conformation a s'accommoder de nombreuses séquences. Formellement, la modifiabilité d'une conformation c est le nombre D des séquences qui se replient en c.

L'intérêt pour la modifiabilité est apparu avec l'article de Finkelstein *et al.*, *Why are the same protein folds used to perform different functions?*, dans lequel les auteurs ont observé que certains repliements sont plus fréquents que d'autres [55] et avancé des arguments théoriques permettant d'expliquer ce constat. Si la conformation native d'une

Matrice d'énergie	*Nombre de séquences se repliant*	*Nombre de réseaux neutres*	*Plus grand réseau neutre*
HP	6 181 800	7 051	54 801 (786)
HP′	10 382 779	3 990	31 935 (786)
LHTW	12 386 286	2 977	67 614 (786)

Tab. III.1 : Nombre de séquences se repliant et de réseaux neutres reconstruits à l'aide des matrices d'énergie LHTW, HP et HP′. La dernière colonne indique la taille du plus grand réseau neutre et la conformation qui lui correspond.

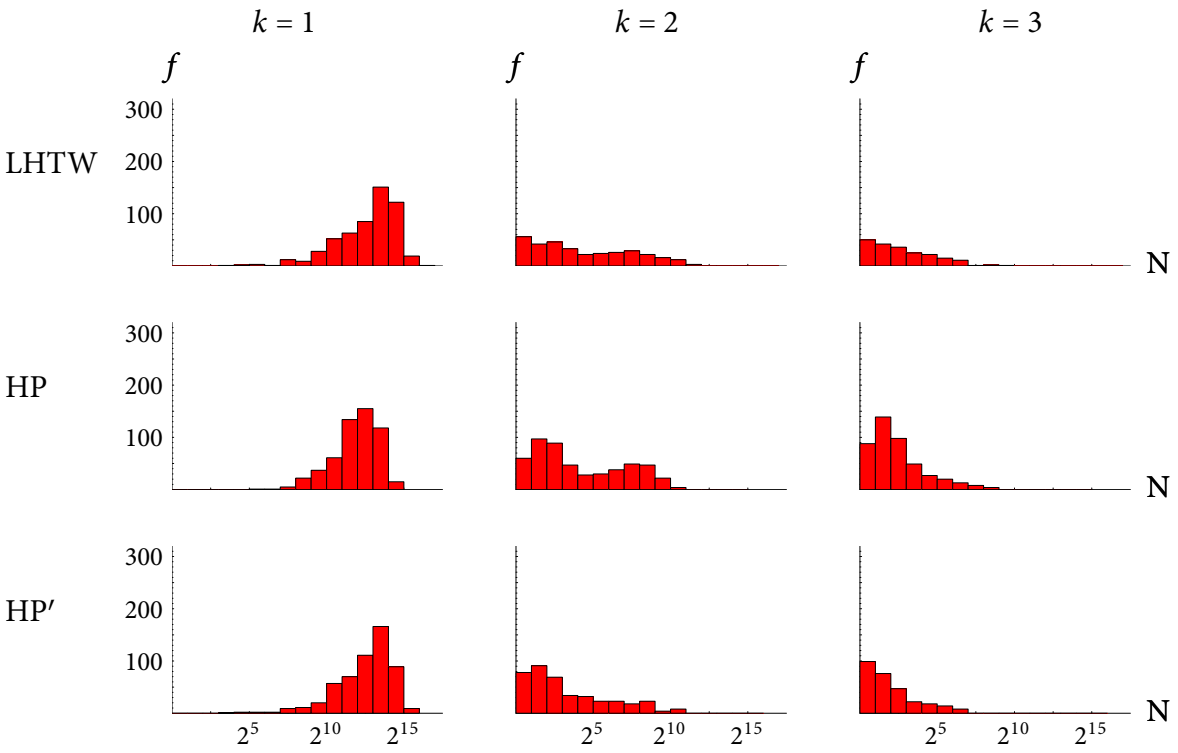


FIG. III.2 : Histogramme des tailles N des plus grands réseaux neutres de chaque conformation ($k = 1$), histogramme des tailles des seconds plus grands réseaux neutres de chaque conformation ($k = 2$), etc. (Note : $2^5 = 32$, $2^{10} = 1\,024$, and $2^{15} = 32\,768$.)

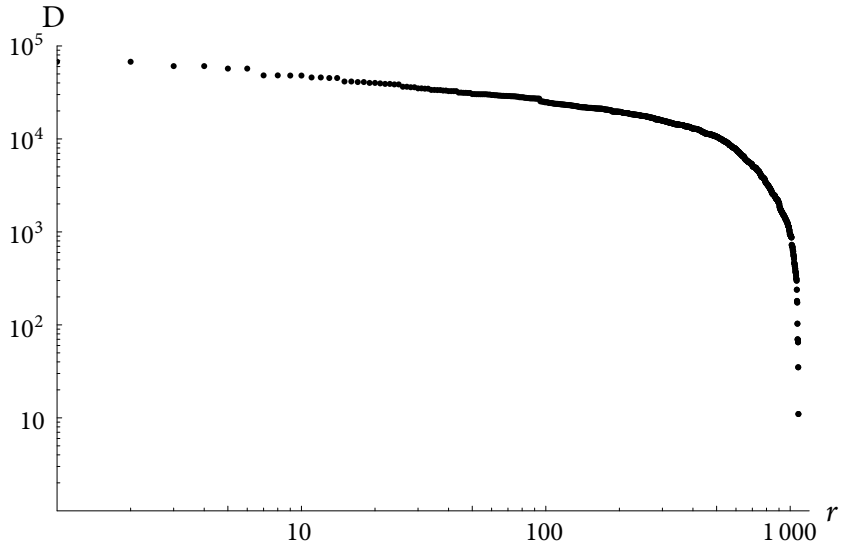


FIG. III.3 : Graphe de Zipf des modifiabilités avec la matrice LHTW. Les modifiabilités D sont classées par ordre décroissant et sont représentées en fonction de leur rang r. La linéarité vaut jusqu'au rang $r = 200$.

séquence était choisie au hasard, la modifiabilité des conformations obéirait à une loi de Poisson. C'est l'hypothèse initialement formulée par Wang [189].

Dans toutes les simulations qui s'y sont intéressées, l'hypothèse de Wang a été infirmée (voir la revue [204]). Une loi de Zipf a été communément invoquée pour décrire la distribution des modifiabilités. Cette loi stipule que si les modifiabilités sont classées par ordre décroissant

$$D_1 \geq D_2 \geq \cdots ,$$

alors D_r est proportionnelle à $1/r^{\alpha}$, où α est un exposant caractérisant la distribution. La relation liant r et D_r est linéaire dans une représentation log-log. On trouvera dans la référence [59], une illustration du même phénomène pour les ARNt et dans la référence [130] une intéressante extension à des modèles tridimensionnels de protéine. Une loi de Zipf classique avec $\alpha \approx 0{,}28$ s'applique dans notre cas jusqu'aux alentours du rang $r = 200$ en utilisant la matrice d'énergie LHTW (cf. figure III.3). Pour des valeurs supérieures à 200, la modifiabilité decroît brusquement.

Les mêmes conformations apparaissent le plus souvent comme les plus modifiables avec les trois matrices LHTW, HP et HP′ (cf. table III.2). Les conformations 781 (194), 787 (241), 903 (786) et 906 (902) apparaissent deux fois dans la table (entre parenthèses est indiquée la conformation inverse). Les conformations 753 (509), 774 (580), 901, 905 (60) et 976 (908) y apparaissent trois fois. Dans le cas de la matrice LHTW, trente structures sont les conformations natives de 10 % des séquences qui se replient.

HP		HP′		LHTW	
903-786	(54 802)	903-786	(19 889)	903-786	(67 615)
787-241	(48 311)	787-241	(19 428)	753-509	(60 708)
753-509	(44 614)	906-902	(18 745)	787-241	(57 097)
906-902	(38 125)	901	(18 224)	976-908	(48 238)
781-194	(34 582)	905-60	(31 953)	906-902	(48 047)
774-580	(34 488)	756-100	(30 469)	774-580	(45 830)
976-908	(34 119)	781-194	(21 942)	635-486	(45 195)
905-60	(33 062)	769-253	(21 067)	754-390	(41 469)
901	(32 840)	245	(20 826)	781-194	(40 904)

Tab. III.2 : Liste des conformations les plus modifiables pour les matrices d'énergie HP, HP′ et LHTW. Les conformations inverses vont par paire, séparées par un tiret, sauf quand la conformation est symétrique. Les modifiabilités sont indiquées entre parenthèses.

La même conformation (et son inverse) est systématiquement la moins modifiable pour les trois matrices d'énergie étudiées. Elle est représentée dans la figure III.4.

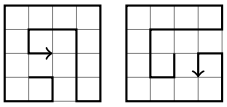


Fig. III.4 : Les conformations les moins modifiables en utilisant les matrices d'énergie LHTW, HP et HP′, 700 et 800, inverses l'une de l'autre. La modifiabilité de ces conformations vaut 11, 107 et 13 avec les matrices LHTW, HP et HP′ respectivement.

Robustesse mutationnelle La robustesse mutationnelle d'une séquence se repliant en une conformation c est le nombre de mutations qui la transforme en une séquence qui se replie, elle aussi, en c. La distribution de la robustesse mutationnelle calculée pour l'ensemble des séquences qui se replient est présentée dans la figure III.5 pour les trois matrices d'énergie LHTW, HP et HP′.

Les trois distributions de la figure III.5 possèdent toutes les trois une forme en cloche asymétrique, en particulier pour HP et HP′. On peut s'intéresser à la distribution de la robustesse mutationnelle au sein d'un réseau neutre. Nous en présentons quatre représentatives dans la figure III.6 : celles des réseaux neutres 664 (a) (LHTW), 755 (a) (LHTW), 786 (a) (LHTW) et 786 (a) (HP). Ces quatre distributions reproduisent la forme observée dans la figure III.5

Quoique leur forme soit identique, les distributions de la figure III.6 ne sont pas exactement superposables : les robustesses mutationnelles moyennes valent 9,94, 10,27, 11,20 pour les réseaux neutres 664 (a) (LHTW), 755 (a) (LHTW) et 786 (a) (LHTW), respectivement. Les tailles respectives de ces réseaux neutres sont 25 367, 35 970 et 67 614. Ces résultats suggèrent que la robustesse mutationnelle moyenne augmente avec la taille du réseau neutre. La figure III.7 montre qu'effectivement la robustesse mutationnelle $\langle n \rangle$ augmente linéairement avec le logarithme de la taille N du réseau neutre. Une relation similaire lie l'écart type de la robustesse mutationnelle à la taille.

III.2.1.2 Propriétés thermodynamiques

Stabilité La température de repliement T_f et la différence d'énergie ΔE définies dans « Modèle et méthodes » mesurent la stabilité d'une séquence repliée dans sa conformation native. On peut se demander si la taille du réseau neutre détermine également les valeurs moyennes $\langle T_f \rangle$ et $\langle \Delta E \rangle$. La figure III.8 indique un comportement plus complexe.

Globalement, la stabilité augmente avec la taille du réseau neutre. Mais la croissance de $\langle T_f \rangle$ n'est pas liée linéairement au logarithme de la taille du réseau neutre et change significativement d'une matrice à l'autre. La différence d'énergie moyenne $\langle \Delta E \rangle$ est plus surprenante encore, puisque $\langle \Delta E \rangle$ forme un plateau pour des tailles de réseau neutre allant jusqu'à 10^3-10^4 environ. À partir d'une taille minimale, $\langle \Delta E \rangle$ peut se « détacher ». La transition est particulièrement spectaculaire dans le cas de la matrice d'énergie LHTW et

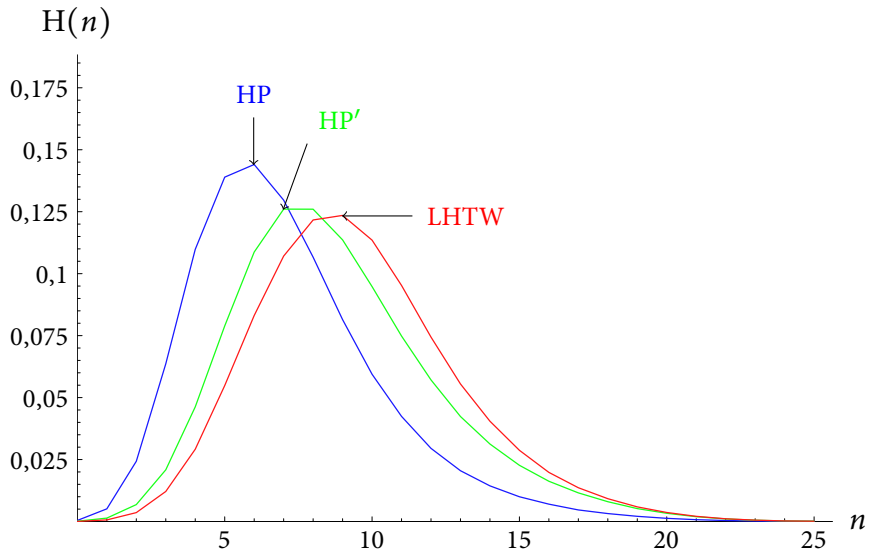


FIG. III.5 : Distribution de la robustesse mutationnelle de l'ensemble des séquences se repliant pour les matrices d'énergie présentées dans « Modèle et méthodes ».

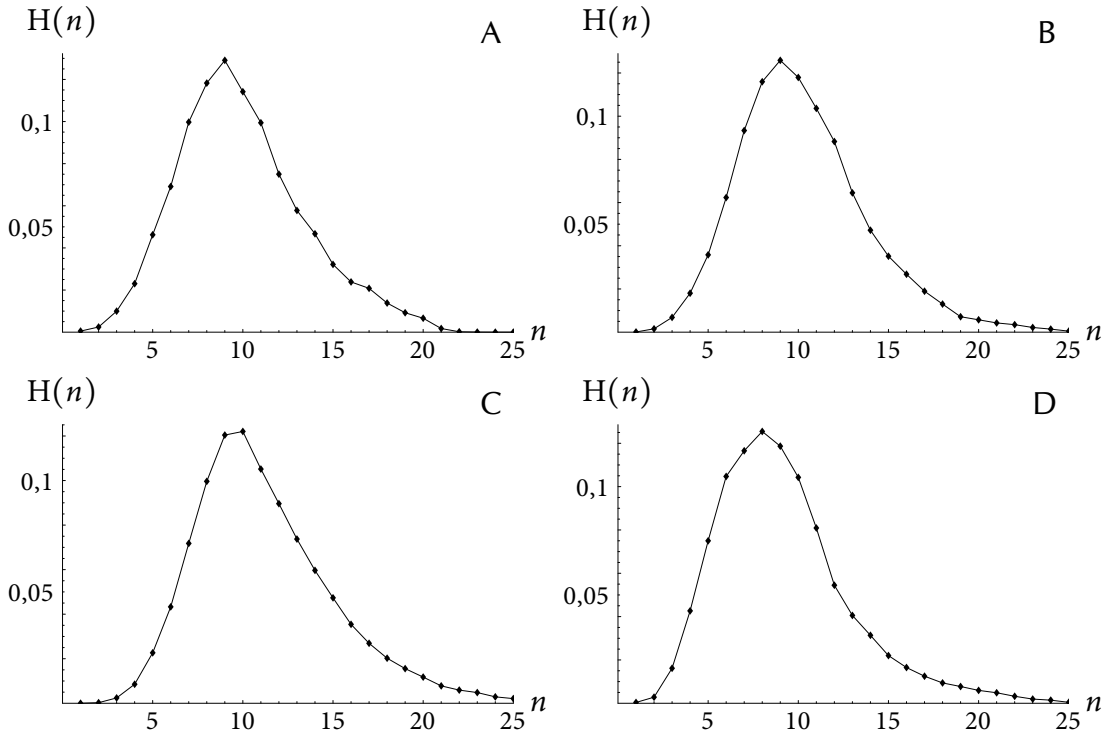


FIG. III.6 : La distribution de la robustesse mutationnelle pour diverses matrices d'énergie et divers réseaux neutres : 664 (a) (LHTW), 755 (a) (LHTW), 786 (a) (LHTW) et 786 (a) (HP). La matrice d'énergie LHTW est représentée plusieurs fois pour illustrer l'influence de la taille du réseau.

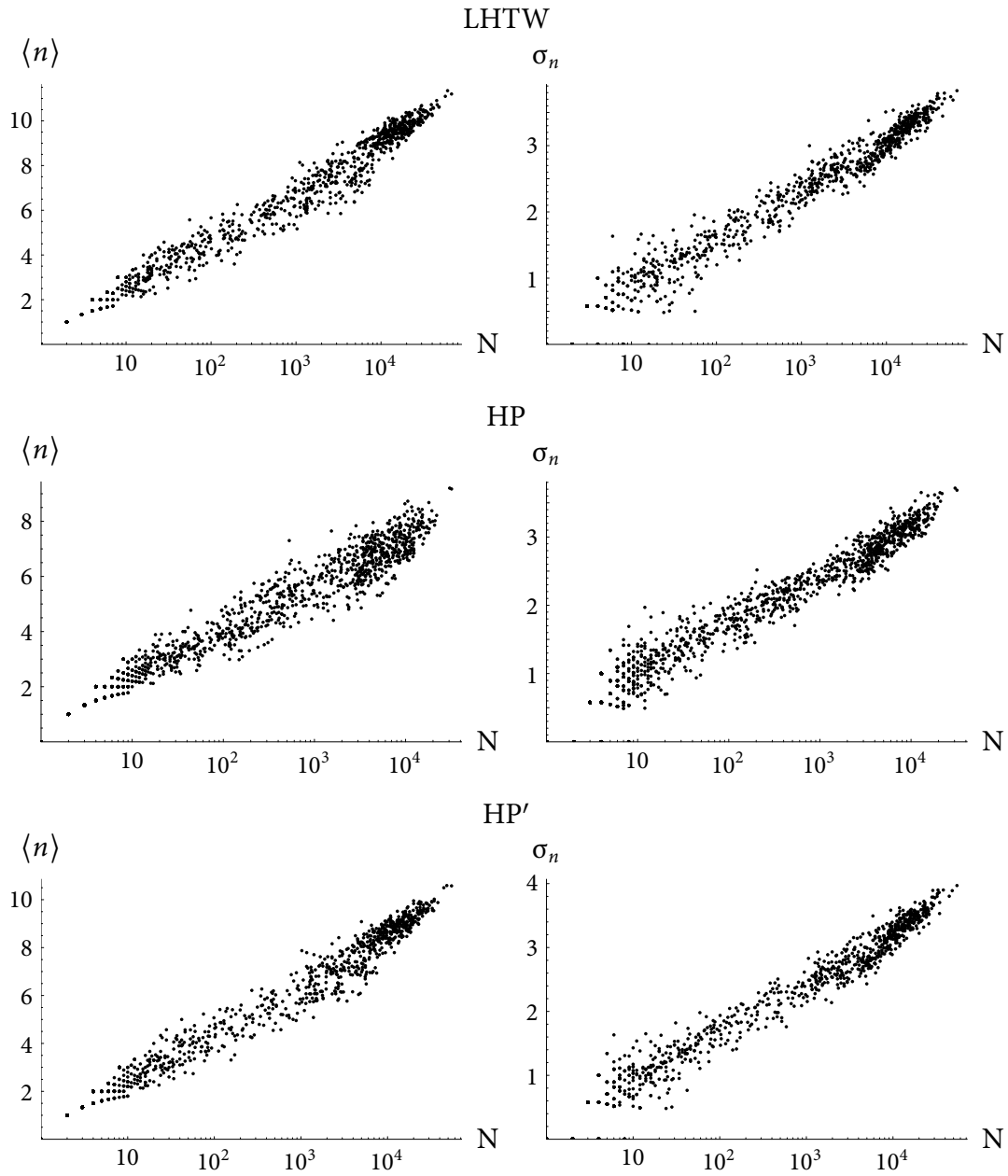


FIG. III.7 : Robustesse mutationnelle moyenne $\langle n \rangle$ et écart type σ_n en fonction de la taille N des réseaux neutres pour les trois matrices LHTW, HP et HP′.

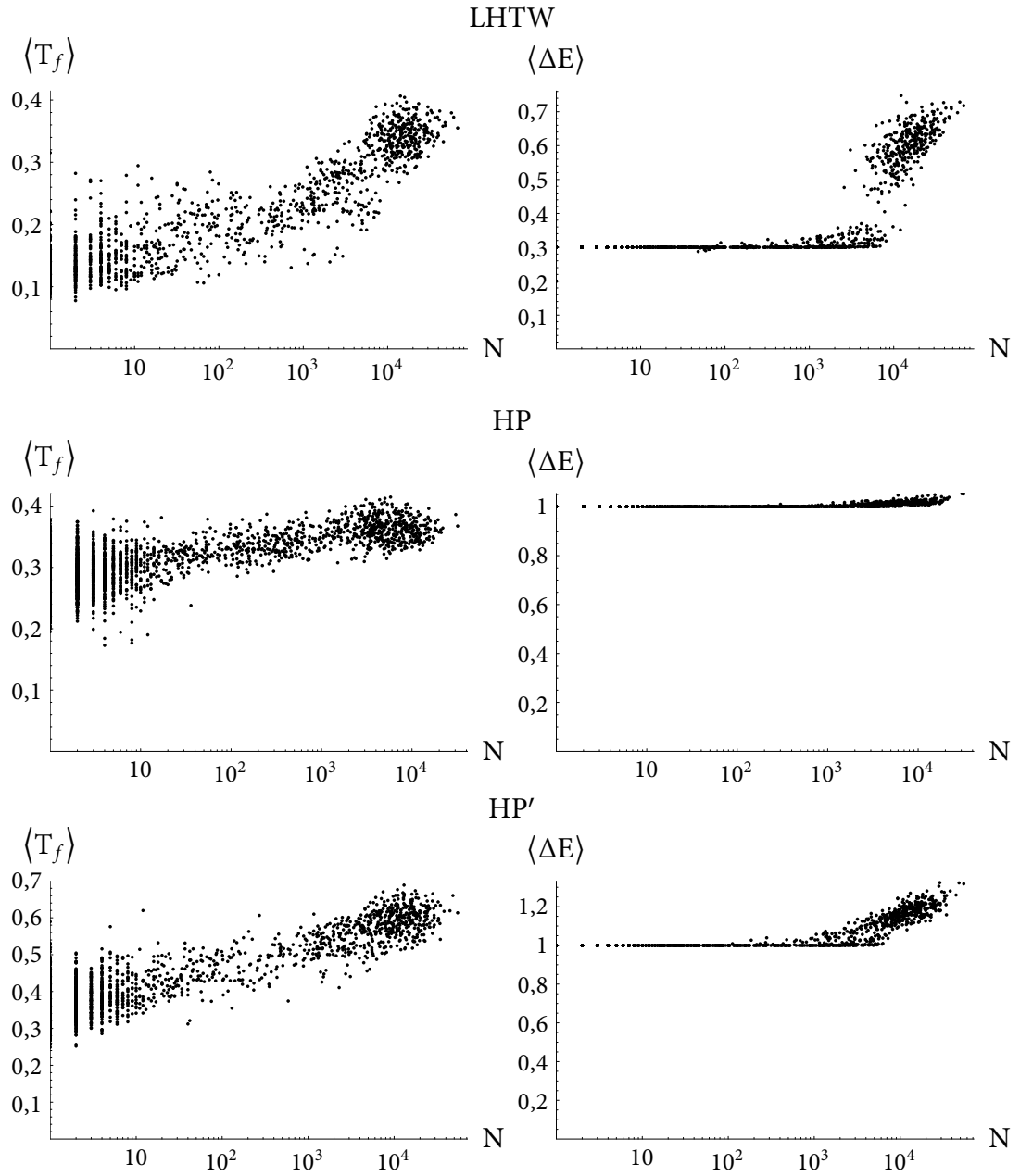


FIG. III.8 : Température de repliement moyenne $\langle T_f \rangle$ et différence d'énergie moyenne $\langle \Delta E \rangle$ en fonction de la taille N des réseaux neutres.

a été commentée par Li *et al.* [112]. La matrice HP′ est à peu près similaire mis à part que la transition est moins franche. Le même phénomène se produit pour HP mais le saut est faible devant la valeur du plateau.

Organisation en *superfunnel* D'après Bornberg-Bauer, un réseau neutre s'organise autour d'une séquence prototype qui ressemble à la séquence consensus [20]. Par séquence consensus, nous entendons la « séquence moyenne » du réseau neutre. La séquence prototype et la séquence consensus du réseau neutre 786 (a) obtenu avec la matrice d'énergie LHTW sont représentées dans la figure III.9. La séquence prototype est l'image en noir et blanc de la séquence consensus.

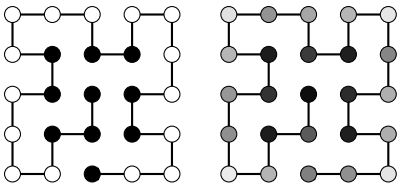


Fig. III.9 : Séquence prototype (à gauche) et séquence consensus (à droite) du réseau neutre 786 (a) obtenu avec la matrice LHTW. Les résidus noirs de la séquence prototype sont les résidus hydrophobes et les résidus blancs sont les résidus polaires. Les niveaux de gris dans la séquence consensus indiquent la fraction de résidus hydrophobes à une position.

Une grande diversité de séquences existent dans les grands réseaux neutres. La majorité des séquences d'un réseau neutre diffèrent de la séquence prototype en un nombre appréciable de positions. La distance de Hamming les séparant est de l'ordre de cinq à huit. Ce constat a été utilisé par Xia et Levitt pour introduire leurs travaux sur le rôle de la recombinaison [203] et, à une autre reprise, par Taverna et Golstein pour rendre compte de la stabilité marginale des protéines [172]. En illustration, trois séquences ont été choisies au hasard dans le réseau neutre 786 (a) (LHTW) et représentées dans la figure III.10.

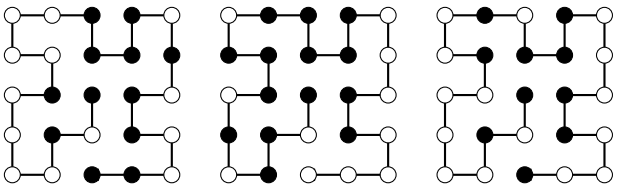


Fig. III.10 : Trois séquences choisies au hasard dans le réseau neutre 786 (a) obtenu avec la matrice LHTW.

L'organisation en *superfunnel* décrite par Bornberg-Bauer est caractérisée par une grande stabilité des séquences dans le voisinage de la séquence prototype. À mesure que l'on s'éloigne de la séquence prototype, la stabilité des séquences diminue [20, 202].

L'image la plus typique du *superfunnel* est obtenue pour les grands réseaux neutres. Pour la mettre en évidence, nous représentons la robustesse mutationnelle n, la tempéra-

ture de repliement T_f et la différence d'énergie ΔE à une distance l de la séquence prototype. La distance l est mesurée par la longueur du plus court chemin dans le réseau neutre menant d'une séquence à la séquence prototype (cf. « Modèle et méthodes », section II.2.6, p. 52). Nous représentons également le nombre de séquences se trouvant à distance l de la séquence prototype pour illustrer que la majorité des séquences en sont relativement éloignées. Les graphiques correspondant au réseau neutre 786 (a) sont présentés dans la figure III.11.

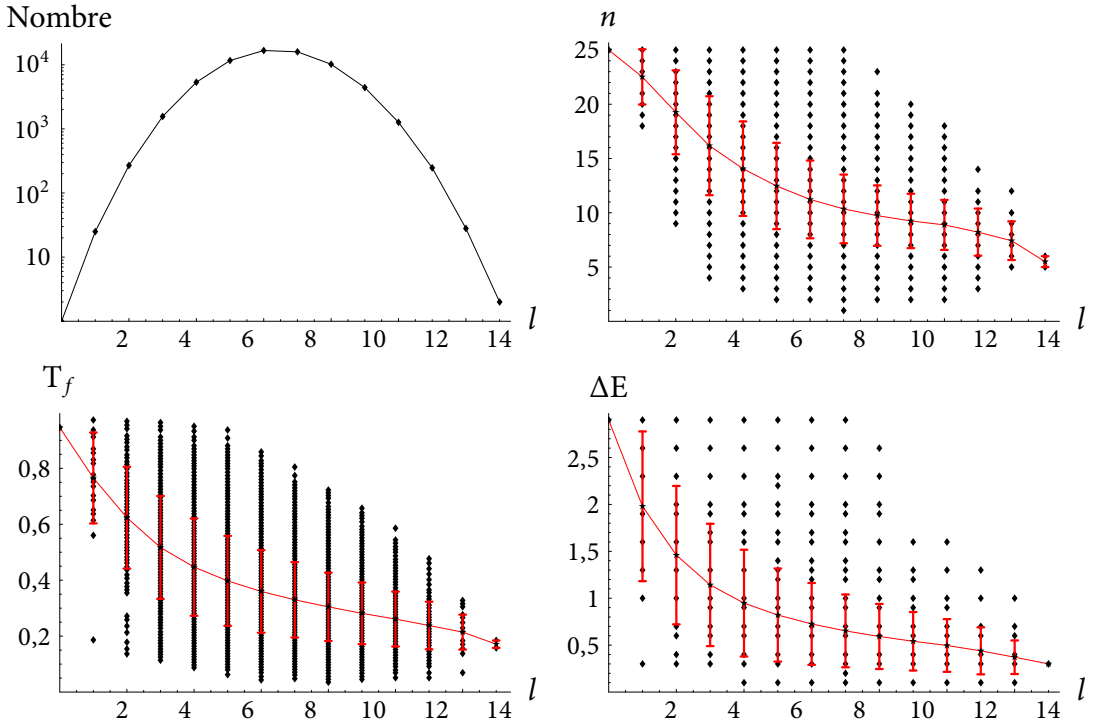


Fig. III.11 : Propriétés du réseau neutre 786 (a) obtenu avec la matrice LHTW(N = 67 614), en fonction de l'éloignement l à la séquence prototype. Nombre de séquences (en haut, à gauche), robustesse mutationnelle n (en haut, à droite), température de repliement T_f (en bas, à gauche) et différence d'énergie ΔE (en bas, à droite) en fonction de l'éloignement l à la séquence prototype. Les données sont représentées par les points noirs, la moyenne par la ligne brisée rouge et les écarts types par les barres verticales. Ces mêmes conventions s'appliquent pour les figures analogues qui suivent.

Le terme *superfunnel* est inspiré du *funnel* du repliement des protéines. Une représentation plus classique du *superfunnel* peut être obtenue en représentant la stabilité moyenne (T_f) dans un système de deux coordonnées. Les deux coordonnées que nous avons choisies sont : le nombre de résidus polaires présents dans une séquence et l'éloignement l à la séquence prototype (cf. figure III.12).

Nous avons soumis d'autres réseaux neutres à une analyse analogue à celle de la figure III.11. Nous en présentons deux représentatifs.

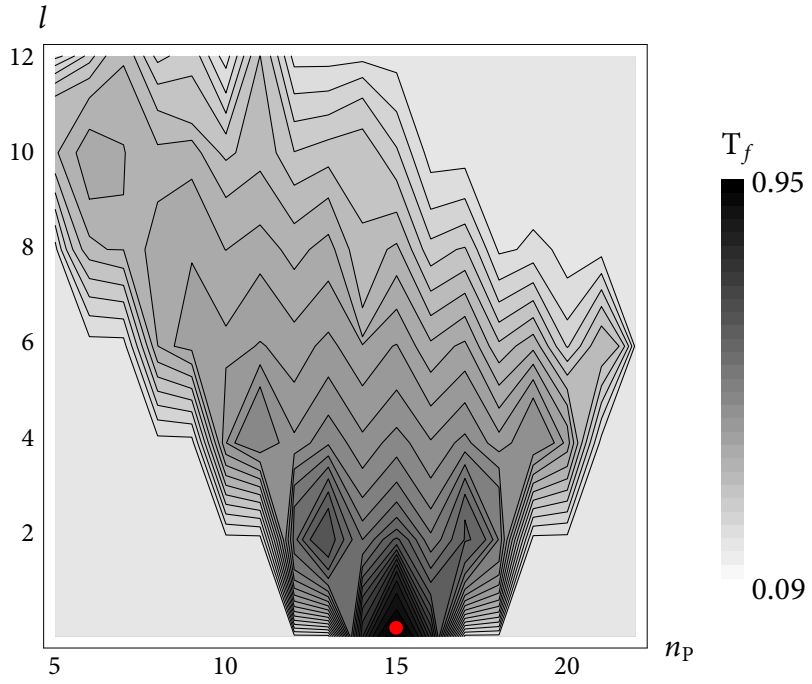


Fig. III.12 : Structure de *superfunnel* dans l'espace des séquence. La température de repliement T_f est représentée selon deux coordonnées : le nombre n_P de résidus polaires présents dans les séquence et l'éloignement l à la séquence prototype. La séquence prototype elle-même est représentée par un disque rouge.

710 (a) Ce réseau est composé d'un nombre relativement réduit de séquences (N = 3 520), mais sa différence d'énergie moyenne $\langle \Delta E \rangle$ est élevée ($\langle \Delta E \rangle$ = 5,26).

198 (a) Ce réseau est composé d'un nombre relativement élevé de séquences (N = 6 886), mais sa différence d'énergie moyenne $\langle \Delta E \rangle$ est faible ($\langle \Delta E \rangle$ = 3,04).

Le premier réseau neutre, en dépit d'une taille réduite, se détache du plateau observé dans la figure III.8. Quant au second, malgré une taille plus importante, il est encore sur ledit plateau. Les résultats sont présentés dans la figure III.13.

La plus faible stabilité moyenne $\langle \Delta E \rangle$ du réseau 198 (a) ne semble pas corrélée à une plus faible robustesse mutationnelle ou une température de repliement diminuée. En revanche, la répartition des séquences autour de la séquence prototype présente est bimodale. Cette caractéristique est observée pour tous les réseaux neutres qui, comme 198 (a), possèdent une faible stabilité moyenne $\langle \Delta E \rangle$.

Les réseaux neutres réduits possédant une grande stabilité, sont caractérisés par une répartition des séquences autour de la séquence prototype en forme de cloche et un *superfunnel* plus lisse.

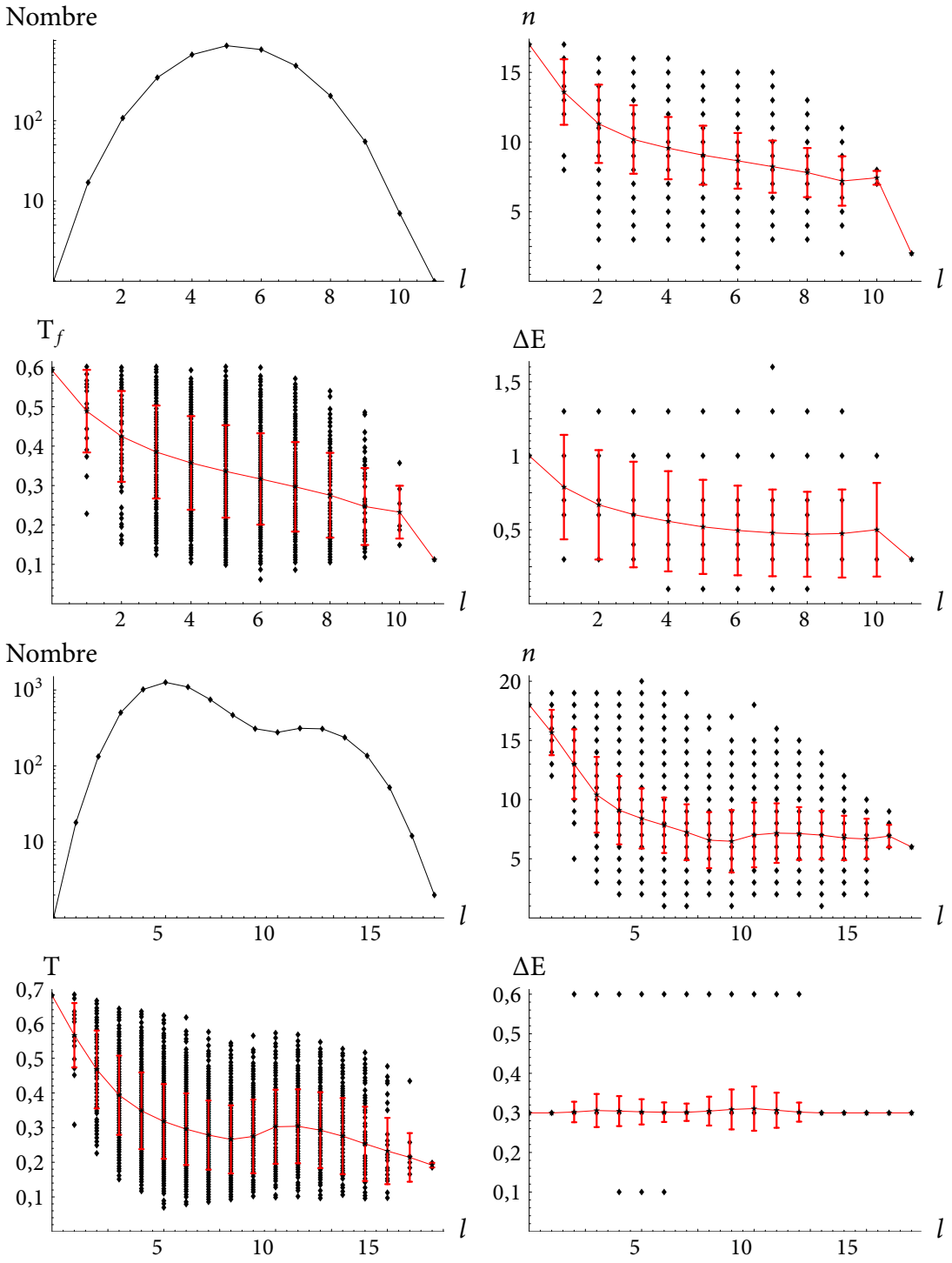


FIG. III.13 : Propriétés du réseau neutre 710 (a) obtenu avec la matrice LHTW(N = 3 520, $\langle \Delta E \rangle$ = 5,26), en fonction de l'éloignement l à la séquence prototype (les quatre graphiques du haut). Propriétés du réseau neutre 198 (a) obtenu avec la matrice LHTW(N = 6 885, $\langle \Delta E \rangle$ = 3,04), en fonction de l'éloignement l à la séquence prototype (les quatre graphiques du haut).

III.2.1.3 Propriétés topologiques

Nous considérons à présent deux quantités purement topologiques. Pour chaque réseau neutre nous calculons, premièrement, la distance de Hamming maximale d_{max} séparant deux séquences du réseau neutre. Deuxièmement nous calculons le diamètre du réseau : la longueur l_{max} du plus long plus court chemin reliant deux séquences par des arêtes du réseau neutre (cf. « Modèle et méthodes », section II.2.6, p. 52). Pour un réseau neutre donné, on a $d_{max} \leq l_{max}$. Ces quantités ont été calculées pour tous les réseaux neutres construits avec la matrice d'énergie LHTW et sont représentées en fonction de la taille des réseaux neutres dans la figure III.14.

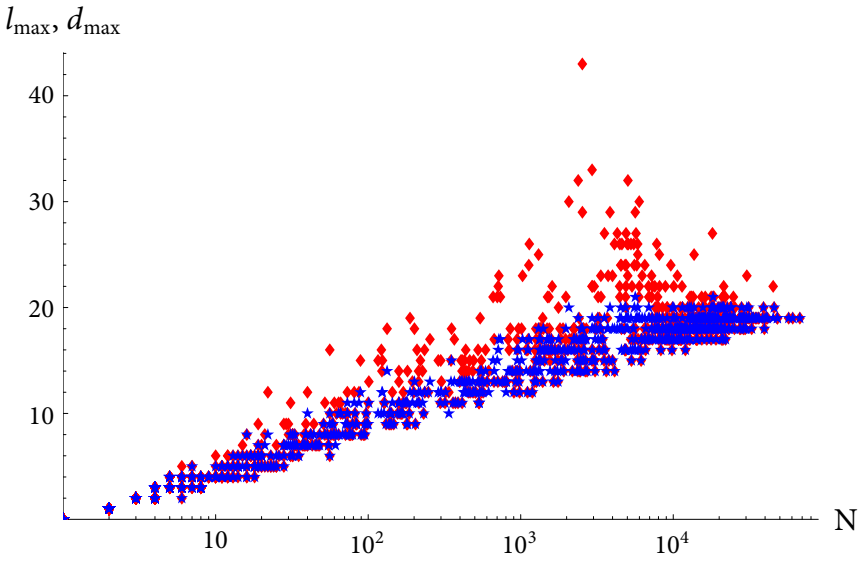


Fig. III.14 : Diamètre l_{max} (en rouge) et distance de Hamming maximale d_{max} (en bleu) en fonction de la taille N des réseaux neutres de la matrice LHTW. Le réseau neutre anomal possédant un diamètre l_{max} fait l'objet d'une étude plus approfondie (cf. p. 100 et figure III.15, p. 101).

Le diamètre et la distance de Hamming maximale augmentent linéairement avec le logarithme de la taille du réseau neutre. Ce fait suggère que les réseaux neutres sont du type « petit monde » : on peut raccorder n'importe quelle séquence à n'importe quelle autre en peu de mutations.

Un phénomène de transition comme celui déjà remarqué pour $\langle \Delta E \rangle$ (cf. figure III.8) apparaît. La distance de Hamming maximale et le diamètre coïncident la plupart du temps pour les réseaux neutres de taille modérée ($N \leq 2\,000$). Ils divergent pour des tailles comprises entre 2 000 et 10 000 environ. Au-dessus de 10 000, ils coïncident à nouveau.

La divergence la plus marquée est obtenue pour le réseau neutre 155 (a). La distance de Hamming maximale de ce réseau de 2 534 séquences vaut 18, le diamètre correspondant vaut 43. Nous avons procédé aux mêmes analyses que dans les figures III.11 et III.13 pour ce réseau neutre. Les résultats apparaissent dans la figure III.15.

La répartition des séquences autour de la séquence prototype suggèrent que deux groupes de séquences bien distincts existent. Nous avons calculé les séquences moyennes de ces deux groupes : le groupe des séquences dont l'éloignement à la séquence prototype est inférieur à 16 et celui des séquences dont l'éloignement est supérieur à 16. Les deux séquences moyennes sont présentées dans la figure III.16.

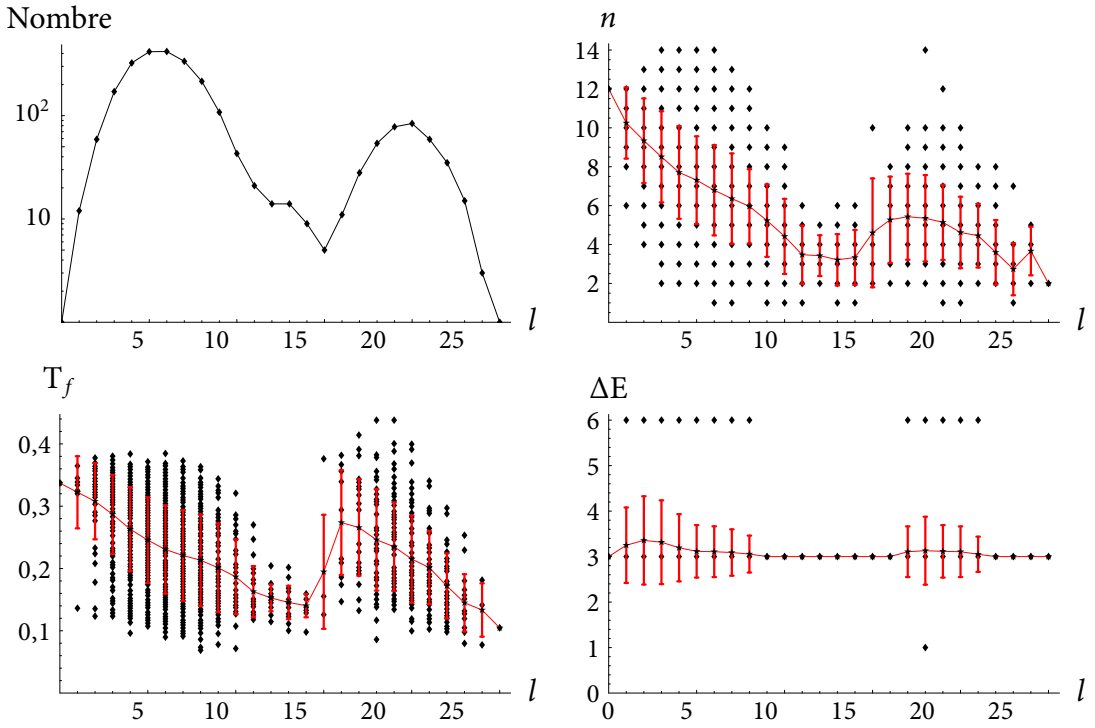


Fig. III.15 : Propriétés du réseau neutre 155 (a) obtenu avec la matrice LHTW(N = 2 534, $l_{\max} = 43$, $d_{\max} = 18$), en fonction de l'éloignement l à la séquence prototype.

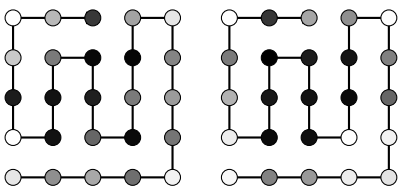


Fig. III.16 : Séquences moyennes du réseau neutre 155 (a). À gauche, moyenne des séquences dont l'éloignement à la séquence prototype est strictement inférieure à 16 ($l < 16$). À droite, celle des séquences dont l'éloignement est supérieur à 16 ($l \geq 16$).

III.2.1.4 Propriétés évolutives

Les propriétés évolutives d'un réseau neutres peuvent être appréhendées par le calcul de deux moyennes distinctes. Pour une certaine propriété x dépendant des séquences composant le réseau neutre (par exemple, la robustesse mutationnelle n ou la température de repliement T_f), on peut évaluer la moyenne $\langle x \rangle_u$

$$\langle x \rangle_u = 1/N \sum_i x_i . \tag{44}$$

Cette moyenne est l'espérance de x lorsque l'on tire au hasard une séquence dans le réseau neutre. Si au contraire on tire au sort deux séquences dans une population évoluant selon le protocole défini dans « Modèle et méthodes », alors l'espérance vaut

$$\langle x \rangle_s = \sum_i p_i x_i , \tag{45}$$

où p_i est le vecteur des fréquences à l'état stationnaire.

L'effet de l'évolution de la population peut être mesuré simplement par le rapport

$$\phi(x) = \frac{\langle x \rangle_s}{\langle x \rangle_u} . \tag{46}$$

Nous nommerons $\phi(n)$ « facteur d'enrichissement » ou « facteur d'amélioration ».

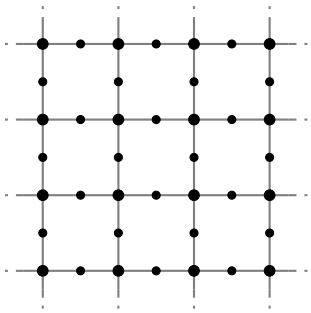


Fig. III.17

Les modifications résultant de l'évolution neutre ne sont pas caractéristiques de la topologie des réseaux neutres, quelle qu'elle soit. Dans une grille infinie comme celle de la figure III.17 on peut calculer — grâce aux symétries du réseau — la fréquence d'un nœud à une intersection relative à celle d'un nœud qui n'est pas à une intersection : elles sont dans un rapport $\sqrt{2} : 1$. Nous avons donc :

$$\begin{aligned} \langle n \rangle_u &= 8/3 \approx 2{,}667 , \\ \langle n \rangle_s &= \sqrt{8} \approx 2{,}828 , \\ \phi(n) &= 3/\sqrt{8} \approx 1{,}061 . \end{aligned}$$

En d'autres termes, à l'état stationnaire, la robustesse aux mutations de ce réseau neutre hypothétique est accrue de 6,1 %.

Dans une situation extrêmement simplifiée où une séquence prototype est entourée d'une seule couche de séquences peu connectées, Bornberg-Bauer et Chan ont estimé

que la fréquence p_i à l'état stationnaire d'une séquence était proportionnelle à la racine carrée du nombre de connexions n_i de cette séquence :

$$p_i \propto \sqrt{n_i}.$$

Avant de revenir sur nos résultats, nous faisons remarquer que leur conclusion essentielle est correcte : la fréquence à l'état stationnaire d'une séquence augmente avec sa robustesse mutationnelle. Or les séquences très robustes (fortement connectées) du graphe se concentrent autour de la séquence prototype, d'après la structure de *superfunnel*. Un réseau neutre de 73 séquences est représenté dans la figure III.18 et illustre cette propriété : les séquences de la périphérie ont une fréquence à l'état stationnaire quasi nulle, l'essentiel de la population se concentrant au centre du graphe.

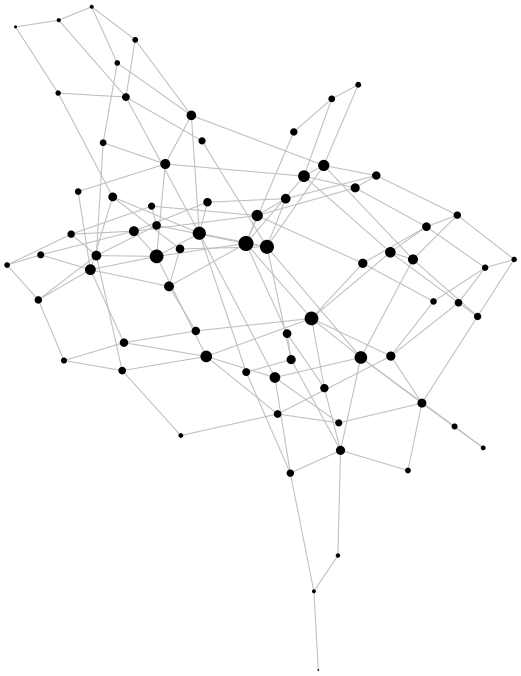


Fig. III.18 : Un petit réseau neutre de 73 séquences obtenu avec la matrice d'énergie LHTW. Le nombre de connexions dans le graphe s'identifie avec la robustesse mutationnelle. Chaque séquence est représentée par un disque dont la surface est proportionnelle à sa fréquence à l'état stationnaire. Les séquences peu connectées, et en particulier, celles de la périphérie du réseau, sont peu probables à l'état stationnaire, tandis que celles du centre, plus connectées, ont une probabilité augmentée.

La loi $p_i \propto \sqrt{n_i}$ ne s'applique pas aux réseaux sur lesquels nous travaillons, du fait de la trop grande simplicité envisagée par Bornberg-Bauer et Chan [21]. La fréquence à l'état stationnaire augmente plus rapidement que linéairement avec la robustesse mutationnelle (cf. figure III.19A). La même observation se confirme lorsque l'on considère la température de repliement (cf. figure III.19B).

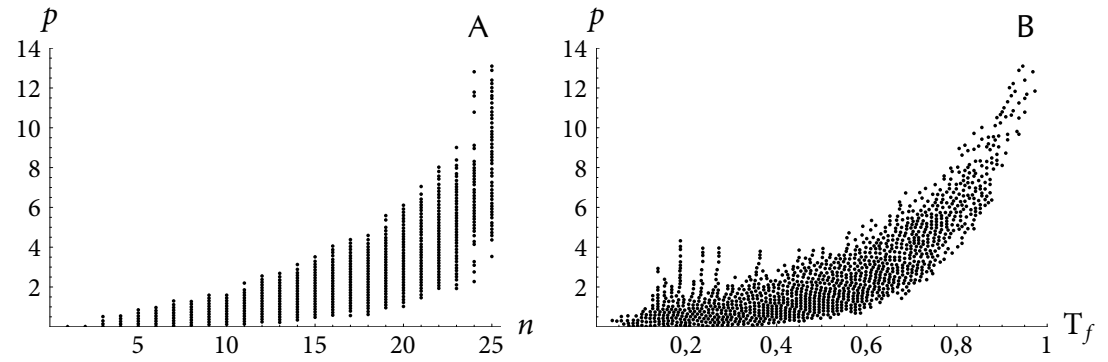


Fig. III.19 : Corrélation entre la fréquence à l'état stationnaire (unité arbitraire) et la robustesse mutationnelle (A) ou la température de repliement (B) pour le réseau neutre 786 (a) obtenu avec la matrice d'énergie LHTW.

La corrélation entre la fréquence à l'état stationnaire p_i et la robustesse mutationnelle n_i, biaise la distribution de la robustesse mutationnelle (cf. graphique gauche de la figure III.20). La corrélation entre p_i et la température de repliement $T_{f\,i}$ biaise de même la distribution de la température de repliement (cf. graphique droite de la figure III.20).

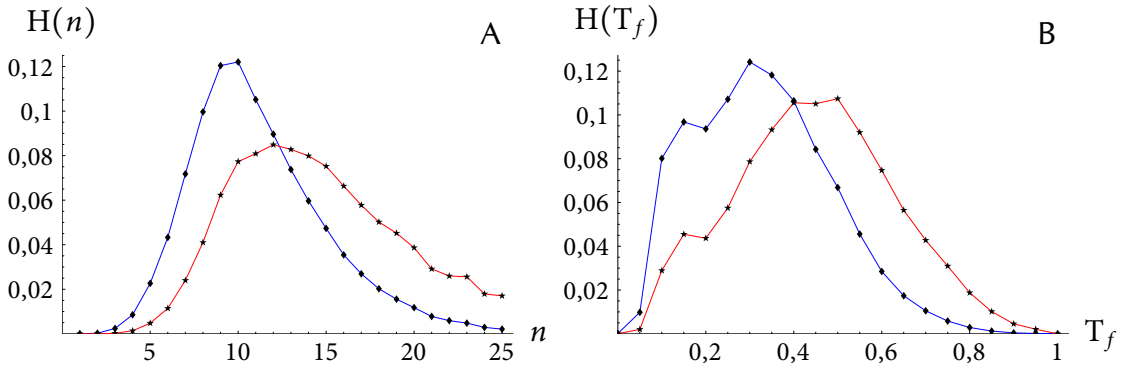


Fig. III.20 : Distribution de la robustesse mutationnelle et de la température de repliement dans le réseau neutre 786 (a) construit avec la matrice d'énergie LHTW. En bleu, distribution résultant d'une distribution uniforme, en rouge à l'état stationnaire.

Si l'on part d'une répartition uniforme de la population, c'est un flux de population vers le centre du réseau neutre, vers la séquence prototype, qui s'amorce. La figure III.21 illustre ce fait dans le cas du réseau neutre 786 (a) construit avec la matrice d'énergie LHTW. On observe que la population se concentre vers la séquence prototype. Les séquences éloignées de la séquence prototype sont elles dépeuplées sous l'effet de l'évolution neutre.

Ce flux de population tend à faire ressembler la séquence moyenne du réseau neutre encore plus à la séquence prototype (cf. figure III.22). Les résidus placés en surface deviennent plus hydrophiles en moyenne, les résidus enfouis deviennent plus hydrophobes.

Le flux de population observé dans la figure III.21 n'est pas spectaculaire. La séquence moyenne ne se rapproche pas énormément de la séquence prototype, on constate en effet

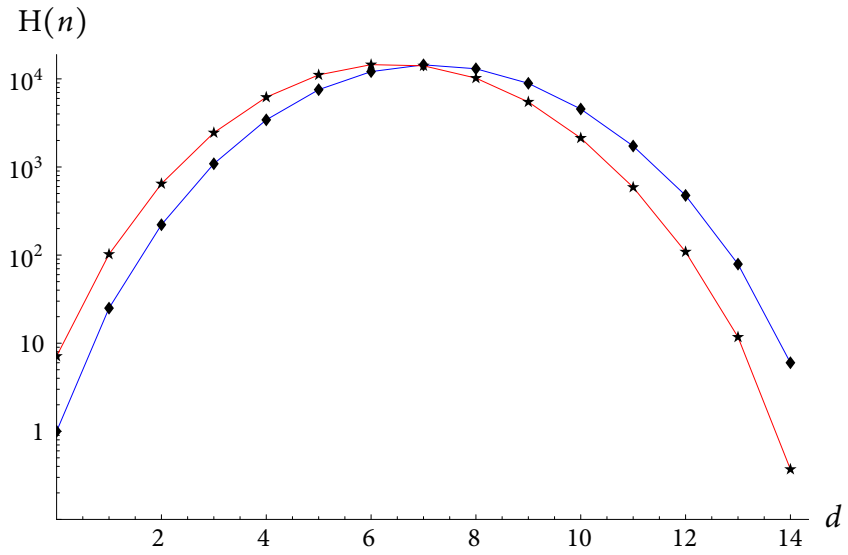


FIG. III.21 : Distribution de la population autour de la séquence prototype sous une distribution uniforme (en bleu) et à l'état stationnaire (en rouge) pour le réseau neutre 786 (a) construit avec la matrice d'énergie LHTW.

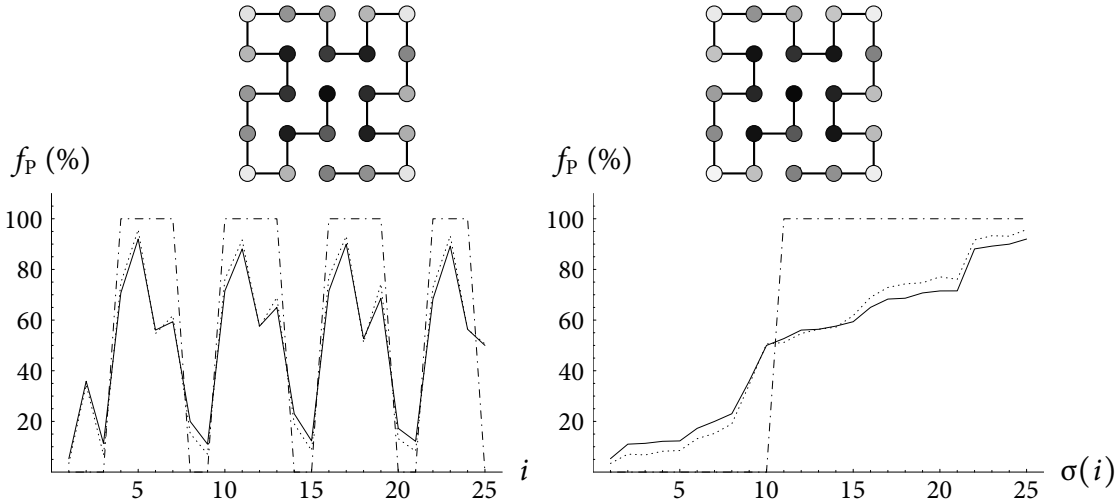


FIG. III.22 : Profils hydrophobe-hydrophiles sous une distribution uniforme (en haut, à gauche) et à l'état stationnaire (en haut, à droite) pour le réseau neutre 786 (a). En bas, les profils sont représentés par la proportion de résidus polaires apparaissant à une position i. La ligne continue (—) correspond au profil moyen calculé en supposant une distribution uniforme de la population sur le réseau neutre, la ligne pointillée (···) correspond à l'état stationnaire et la ligne discontinue (– · –) correspond à la séquence prototype. $\sigma(i)$ est une renumérotation des positions par ordre d'hydrophilie croissante.

que les séquences moyennes sous une distribution uniforme et à l'état stationnaire sont quasiment confondues. La recombinaison permet à ce phénomène de prendre plus d'importance [203].

Les quantités $\phi(n)$ et $\phi(T_f)$ quantifient l'impact de l'évolution neutre sur les propriétés moyennes d'une séquence tirée au hasard dans la population en comparaison d'un tirage aléatoire d'une séquence du graphe. Cet impact est déterminé en grande partie par la taille N du réseau neutre comme l'indique la figure III.23. La population évolue vers les génotypes plus robustes et plus stables de manière plus sensible dans les réseaux neutres plus grands et plus modifiables.

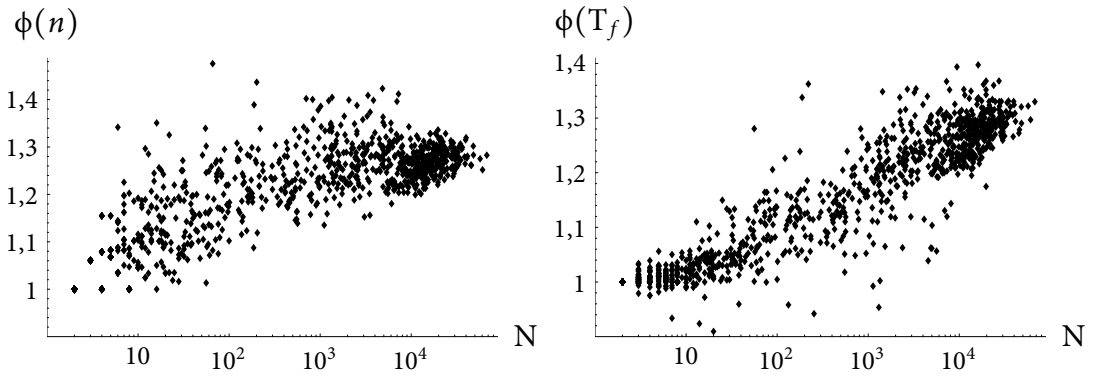


Fig. III.23 : Amélioration de la robustesse mutationnelle moyenne (à gauche) et de la température de repliement moyenne (à droite) à l'état stationnaire pour la matrice d'énergie LHTW.

Les annexes présentant les données pour les autres matrices d'énergie et des résultats complémentaires sont reportées après la discussion (p. 145 et suivantes).

III.2.2 Protéines tridimensionnelles

Dans le modèle de protéine tridimensionnelle hors-réseau, la viabilité est toujours évaluée grâce à la stabilité de la conformation native (l'une des trois conformations trp-cage, Grb2 ou Vav). La stabilité est mesurée par la différence d'énergie séparant l'état fondamental du premier état excité mais aussi par le Z-score de son spectre d'énergie. Les états dénaturés sont mimés par un ou deux jeux de fausses structures.

Nous avons défini un lot d'alphabets qui catégorisent les acides aminés en m classes (où m peut prendre les valeurs 2, 3, 4, 6 et 20, le cas 20 correspondant à un alphabet complet). Une séquence est convertie en son profil en assimilant chaque acide aminé à sa classe. Les réseaux neutres sont construits à partir des profils des séquences acceptées lors des trajectoires.

III.2.2.1 Trajectoires réalisées

Le tableau III.3 résume les trajectoires menées à terme dans notre étude.

Trajectoires libres Deux trajectoires ont été réalisées pour Vav. Nous les notons C1 et C2. Les séquences de référence ont été obtenues grâce à des cycles de préoptimisation avec une matrice à deux classes. Le rôle de la préoptimisation est de fournir une séquence de référence suffisamment stable. Cent millions de pas ont été exécutés par la méthode de Monte Carlo décrite dans « Modèle et méthodes ». La matrice à deux classes a été utilisée pour la détermination des énergies.

Nous avons généré des séquences pour TRP-cage selon trois voies : une trajectoire a utilisé une matrice à vingt classes, une autre celle à trois classes et la dernière une matrice à deux classes. La séquence de référence de ces trois simulations est la même et a été obtenue par une phase de préoptimisation à vingt classes. Nous nommerons ces trois trajectoires, C20, C3 et C2. La longueur de ces trois trajectoires est $50 \cdot 10^6$ pas.

Trajectoires d'exploration dimériques Deux trajectoires vont mettre en scène Grb2 et Vav pour étudier l'interaction protéine-protéine. À cet effet, dix résidus sont figés à l'interface *lors de la préoptimisation* pour préserver un motif favorable à l'interaction. En revanche, cette contrainte est supprimée durant les trajectoires elles-mêmes. Les positions fixées sont

— Dans Grb2, VAL161, GLN162, PHE165, ARG179, TRP193, ARG207, ASN208, TYR209, VAL210 et THR211 ;

— Dans Vav, MET594, PRO595, GLU626, LYS629, ALA632, GLY639, HIS656, PRO657, TYR658 et VAL659.

Elles sont représentées dans la figure III.24. La préoptimisation est réalisée avec la matrice à deux classes.

Après cette préoptimisation, deux trajectoires, que nous nommerons B10 et C10, ont été réalisées. Cent millions de pas ont été exécutés par la méthode de Monte Carlo décrite dans « Modèle et méthodes ». La matrice à deux classes à été utilisée pour la détermination des énergies.

III.2.2.2 Exemple de préoptimisation

La séquence de préoptimisation produit à chaque cycle une séquence de référence plus stable. Nous l'illustrons avec les cycles de préoptimisation de la trajectoire B10 (cf. tableau III.4). On observe une diminution des valeurs de référence $\Delta E_0^{(i)}$ et $Z_0^{(i)}$.

Nom	*Séquence de référence*	*Cycles*	*Classes*	*Fausses structures*	*Positions gelées*	*Séquences acceptées*	*(%)*	*Nombre de profils*
B10	TYVQW LFQYF PKAQC YPIHI RQGFP VWVAC KRKHW GIVLL QDPWC MISRN YVTNM LQ	5	2	2	10	10143897	10,1	20840
C10	DIDMP YCFPH HWGCA KDWMH AHRSY CSLLC HPSLA ELGKW QAKGG YYWGR YSLFA RDLYM QIERY HPYV	5	2	2	10	10472067	10,5	29667
C1	WREKV LLHLS AFCCD GRKKL SHIDE DCCVC VEVII VYPNE DHPCE ECKEL LVMQA EICKC NCPVI SWIT	7	2	2	0	1627612	1,6	142113
C2	WSTKV LLHLS AFCCD VPKKL GQIDE DCCVC VDIII VYWEE ETNCE ECKEL LVMHA EICKC NCPVI SWIS	8	2	2	0	1242953	1,2	36574
2C	PNLQT YFTLW IPSYR YPPPD	1	2	1	4	22738262	46	3042
3C	PNLQT YFTLW IPSYR YPPPD	1	3	1	4	24821577	49	528472
20C	PNLQT YFTLW IPSYR YPPPD	1	20	1	4	11455666	22	

Tab. III.3 : Les différentes simulations réalisées dans notre étude. « Nom » est la dénomination des différentes trajectoires. La colonne « Cycles » indique le nombre de cycles de préoptimisation qui ont été nécessaires à l'obtention de la séquence de référence. « Classes » est le nombre de classes utilisée lors de la trajectoire. « Fausses structures » contient le nombre de fausses structures. « Positions gelées » donne le nombre de positions qui ne sont pas soumises à des mutations. La colonne « Séquences acceptées » et la colonne qui suit donnent le nombre de séquences acceptées au cours de la simulation et le pourcentage correspondant.

Cycle	*Séquence de référence*	*Pourcentage de séquences acceptées*	Z_0^1	Z_0^2	ΔE_0^1	ΔE_0^2
1	TYVQA LFDFD PQEDG ELGFR RGDFI HVMDN SDPNW WKGAC HGQTG MFPRN YVTPV NR	67,575	−2,65	−1,56	12	1
2	KWVQM EFLSW KVKCD MIGMV RDGML INAKI DGDCW ESQSF DRWAK KLTRN YVTCS QC	48,897	−5,67	−7,30	142	161
3	NYVQW IFVCW EMVQC MSIIS RQWWC CESEA TRRKW CMYPL DCLVV TVPRN YVTLM RT	32,224	−6,92	−11,56	262,5	325
4	YYVQM VFECW ANCKV YSWNT RKGLA FEVIE NQYWW EYCCL NNTWC TMCRN YVTSI LC	19,462	−8,23	−17,20	387	377,5
5	FLVQW NFKYS PRCRL YSIHV RCIFC WWCHV RYNGW PIACI ERFFG WMPRN YVTDI PD	13,843	−9,80	−17,80	503	415
6	TYVQW LFQYW ADVRF YPVHC REGFP TWCVT RRKKW AIMML QHPWC MMSRN YVTNM LR	7,596	−10,75	−18,22	632	431
7	TYVQW LFQYF PKAQC YPIHI RQGFP VWVAC KRKHW GIVLL QDPWC MISRN YVTNM LQ	4,891	−10,88	−20,70	649	463,5

Tab. III.4 : Exemple d'une phase de préoptimisation. Celle présentée est celle menée sur Grb2 en prélude de la trajectoire B10. Dix résidus sont gelés comme cela a été expliqué plus haut. Nous indiquons à chaque cycle la séquence de référence et les valeurs de référence $\Delta E_0^{(i)}$ et $Z_0^{(i)}$ fixant la limite minimale de stabilité.

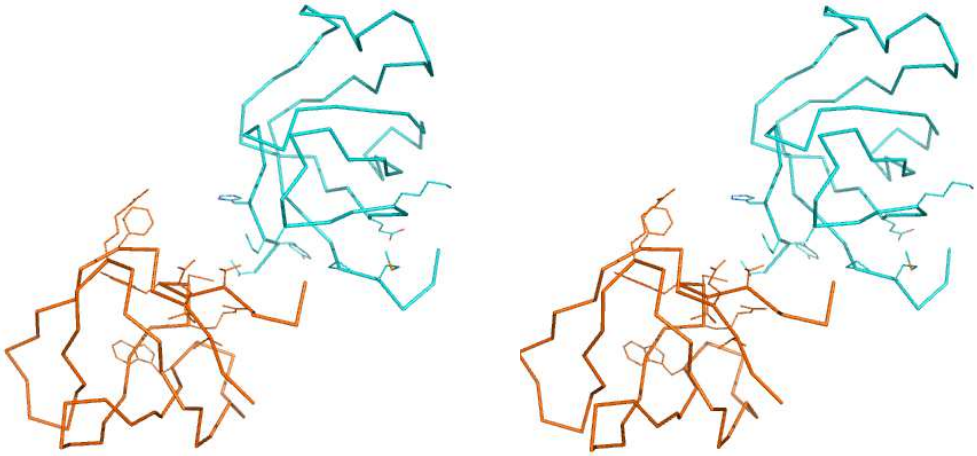


Fig. III.24 : Vue stéréoscopique des squelettes peptidiques de Grb2 et Vav. Les chaînes latérales des positions fixées à l'interface sont représentées. Pour améliorer la lisibilité de la figure, nous avons éloigné les deux monomères d'une distance de 10 Å.

III.2.2.3 Propriétés topologiques

Robustesse aux mutations Nous avons reconstruit les réseaux neutres à deux classes à partir des trajectoires B10 et C10. La distribution de la robustesse mutationnelle de ces réseaux neutres est présentée en bleu dans la figure III.25 (la courbe rouge est la distribution à l'état stationnaire sur lequel nous reviendrons plus tard). Nous retrouvons la forme en cloche allongée typique observée pour les protéines sur réseau : peu de séquences sont très peu robustes, peu de séquences sont très robustes. Néanmoins, la présence de quelques séquences très robustes confère une topologie semblable à celle des graphes sans échelle. Nous y revenons plus loin.

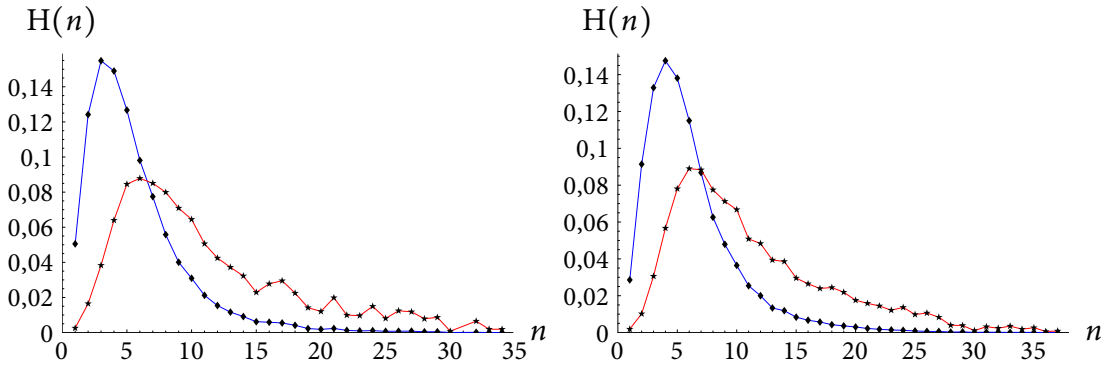


Fig. III.25 : Distribution de la robustesse mutationnelle sous une distribution uniforme (en bleu) et à l'état stationnaire (en rouge) pour les graphes à deux classes reconstruits à partir des trajectoires B10 (à gauche) et C10 (à droite).

La forme allongée de la distribution est retrouvée pour les autres réseaux neutres reconstruits à partir des trajectoires C1 et C2. Une forme similaire est trouvée pour le TRP-cage, mais moins allongée à cause de la faible longueur de chaîne.

Nous avons procédé à la relecture des trajectoires 2C, 3C et 20C. Toutes les séquences acceptées ont été examinées et soumises à des mutations aux 16 positions possibles. Cette approche permet de connaître la distribution de la robustesse mutationnelle du réseau neutre m classes. Les résultats pour m = 20, 3 et 2 sont présentés dans la figure III.26.

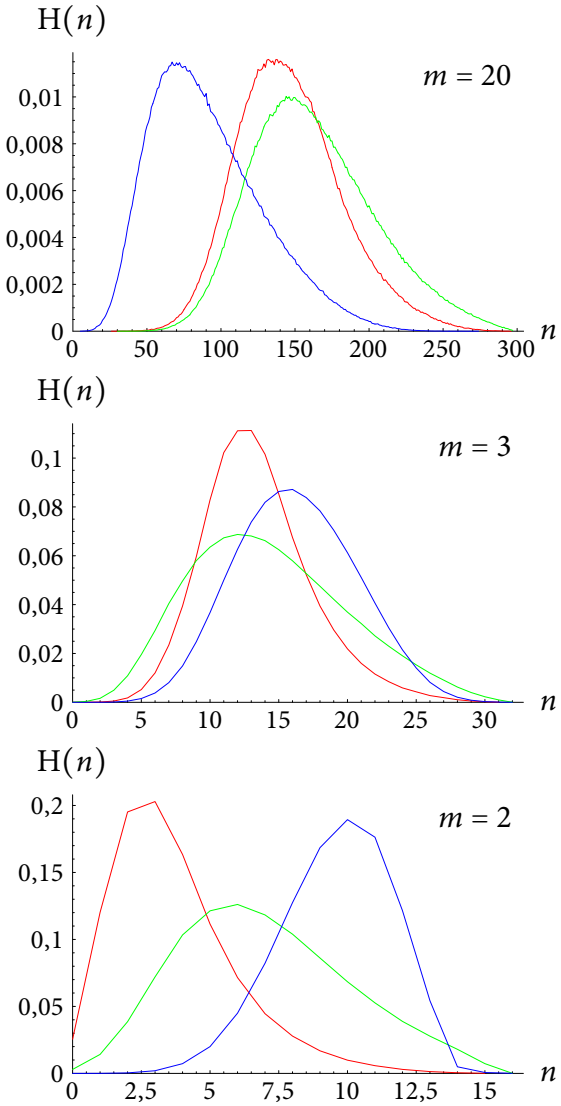


Fig. III.26 : Distribution de la robustesse mutationnelle calculée par relecture des trajectoires 2C, 3C et 20C du TRP-cage. On évalue pour chaque séquence acceptée le nombre de mutations tolérées. On accepte une mutation par classe d'acides aminés (avec m = 2, 3 et 20 classes). La trajectoire 20C est représentée en bleu, la trajectoire 3C en vert, et la trajectoire 2C en rouge.

Ces résultats indiquent que les formes en cloche, résultant de la projection en profil avant la construction des réseaux neutres, n'est pas un artefact de notre méthode de

reconstruction. Les réseaux neutres réels possèdent une distribution de la robustesse mutationnelle similaire.

Réseau sans échelle Les « réseaux sans échelle » (*scale-free networks*) sont un type de réseau décrit intialement par Barabási et Albert dans l'étude des connexions entre sites internet [3]. Ces réseaux sont caractérisés par une distribution très inégale des connexions et l'existence de quelques nœuds intensément connectés. Ce type de réseau apparaît dans les réseaux d'interaction entre protéines notamment (voir par exemple, référence [110].

Le nombre de nœuds possédant k connexions dans un réseau sans échelle est une loi de puissance : il est proportionnel à $k^{-\gamma}$. La représentation de la distribution de la robustesse mutationnelle n en échelle logarithmique est montrée en figure III.27.

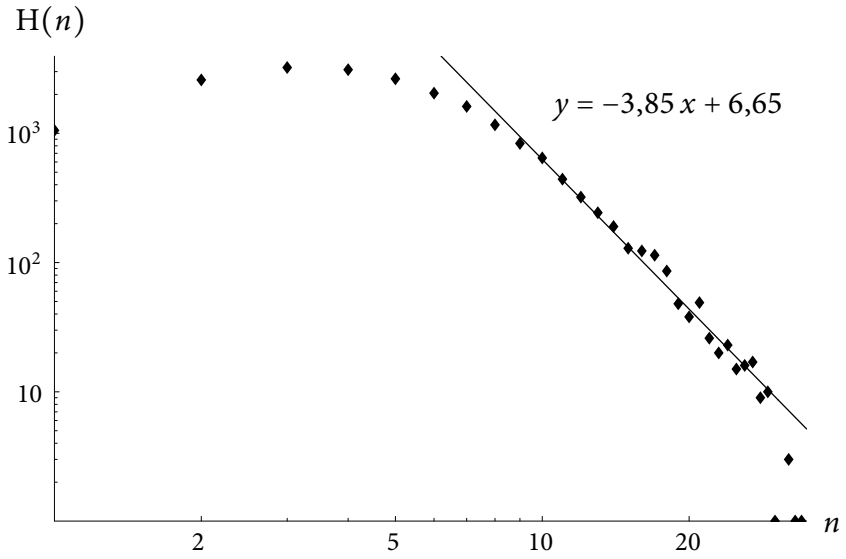


Fig. III.27 : Représentation log-log de la distribution de la robustesse mutationnelle n dans le graphe deux classes reconstruit à partir de la trajectoire B10 (voir également figure III.25). La queue de la distribution peut être approximée par une loi de puissance.

La queue de la distribution peut être approximée par une relation affine. Nos réseaux neutres possèdent donc des propriétés de réseau sans échelle : la présence de nœuds très connectés (les *hub*) correspondent au centre du *superfunnel*.

Superfunnel dans l'espace des séquences La structure de *superfunnel* a été démontrée auparavant à l'aide de simulations menées sur des protéines sur réseau et avec un alphabet HP. Nous souhaitons confirmer l'existence d'une telle structure pour des protéines tridimensionnelles et des alphabets plus complets. Il n'est pas nécessaire de reconstruire les réseaux neutres pour mettre en évidence la structure de *superfunnel*, car il suffit de mesurer la stabilité en fonction de l'éloignement à la séquence prototype. Nous utilisons pour cela la trajectoire 20C du trp-cage.

Il nous faut donc déterminer la séquence prototype. Les résultats obtenus grâce aux protéines sur réseau indiquent que la séquence prototype ressemble à la séquence consensus ou séquence moyenne. Les séquences consensus obtenues en moyennant les séquences générées lors des trajectoires 2C, 3C et 20C sont proches (cf. colonne gauche de la figure III.28). Il en va de même pour les profils calculés sous une distribution uniforme et à l'état stationnaire pour les réseaux neutres 2C et 3C établis à partir des trajectoires 2C et 3C respectivement (cf. colonne droite de la figure III.28).

D'après ces résultats et la séquence moyenne de la trajectoire 20C, on peut raisonablement faire l'hypothèse que la séquence prototype $\hat{s}$ est :

NYMNN WTDGY FPRYK YPPPT

en reprenant à chaque position le résidu le plus fréquent.

Pour toutes les séquences s acceptées au cours des 500 000 000 pas de trajectoire[16], on calcule 1. la distance de Hamming $d_{\mathrm{H}}(\hat{p}, p)$ entre les profils $\hat{p}$ et p de $\hat{s}$ et s respectivement (en utilisant un alphabet mC), 2. le Z-score de la séquence s (en utilisant l'alphabet 20C de la trajectoire). Pour différentes valeurs de m, on représente le Z-score en fonction de la distance de Hamming entre les profils dans la figure III.29. Le Z-score est d'autant plus négatif que la conformation native d'une séquence est stable. En conséquence, nous observons un *superfunnel* qui est dirigé vers le bas.

On peut également observer comment évolue la robustesse mutationnelle en fonction de la distance au profil prototype. Les résultats sont présentés dans la figure III.30. Exactement de la même façon que pour les protéines sur réseau, la robustesse elle aussi obéit à l'organisation en *superfunnel*.

Un autre choix possible de « séquence prototype » est de se conformer aux idées initiales de Bornberg-Bauer [20] et de choisir la séquence $\tilde{s}$ la plus robuste de la trajectoire (obtenue lors de la relecture de la trajectoire à vingt classes) :

DWMNN WREGY IPQYK YPPPT

Cette séquence tolère 274 mutations soit 90 % de toutes les mutations possibles.

Les deux séquences $\hat{s}$ et $\tilde{s}$ diffèrent en six positions mais possèdent des profils en trois classes proches : les profils diffèrent en deux positions seulement (cf. figure III.31D et E) et sont proches des profils les plus peuplés à l'état stationnaire (cf. figure III.31A, B et C).

Quoi qu'il en soit l'évolution de la robustesse mutationnelle à distance d du profil prototype ne dépend pas énormément du choix de la séquence prototype, $\hat{s}$ ou $\tilde{s}$. Les résultats sont comparés dans la figure III.32 en mesurant la fraction moyenne f_0 de mutations neutre à distance d des profils prototypes.

16. À cause du coût de calcul requis pour traiter toutes les séquences, en réalité, seule une séquence sur quinze est examinée.

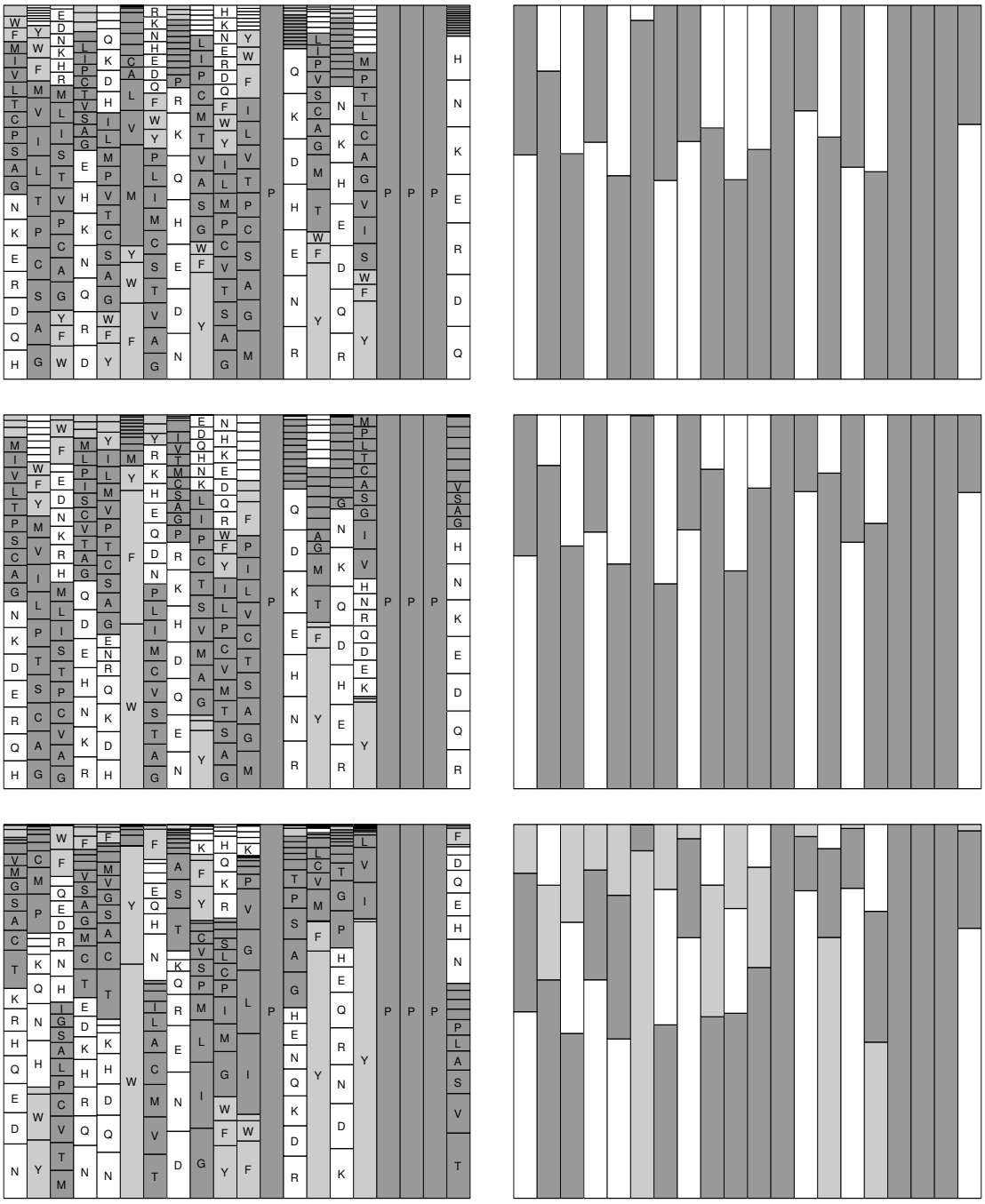


Fig. III.28 : En haut, à gauche, séquence moyenne de la trajectoire 2C. Au centre, à gauche, séquence moyenne de la trajectoire 3C. En bas, à gauche, séquence moyenne de la trajectoire 20C. En haut, à droite, profil à deux classes moyen sous une distribution uniforme provenant du réseau neutre à deux classes reconstruit à partir de la trajectoire 2C. Au centre, à droite, profil à deux classes moyen à l'état stationnaire provenant du réseau neutre à deux classes reconstruit à partir de la trajectoire 2C. Au centre, à droite, profil à deux classes moyen à l'état stationnaire provenant du réseau neutre à trois classes reconstruit à partir de la trajectoire 3C.

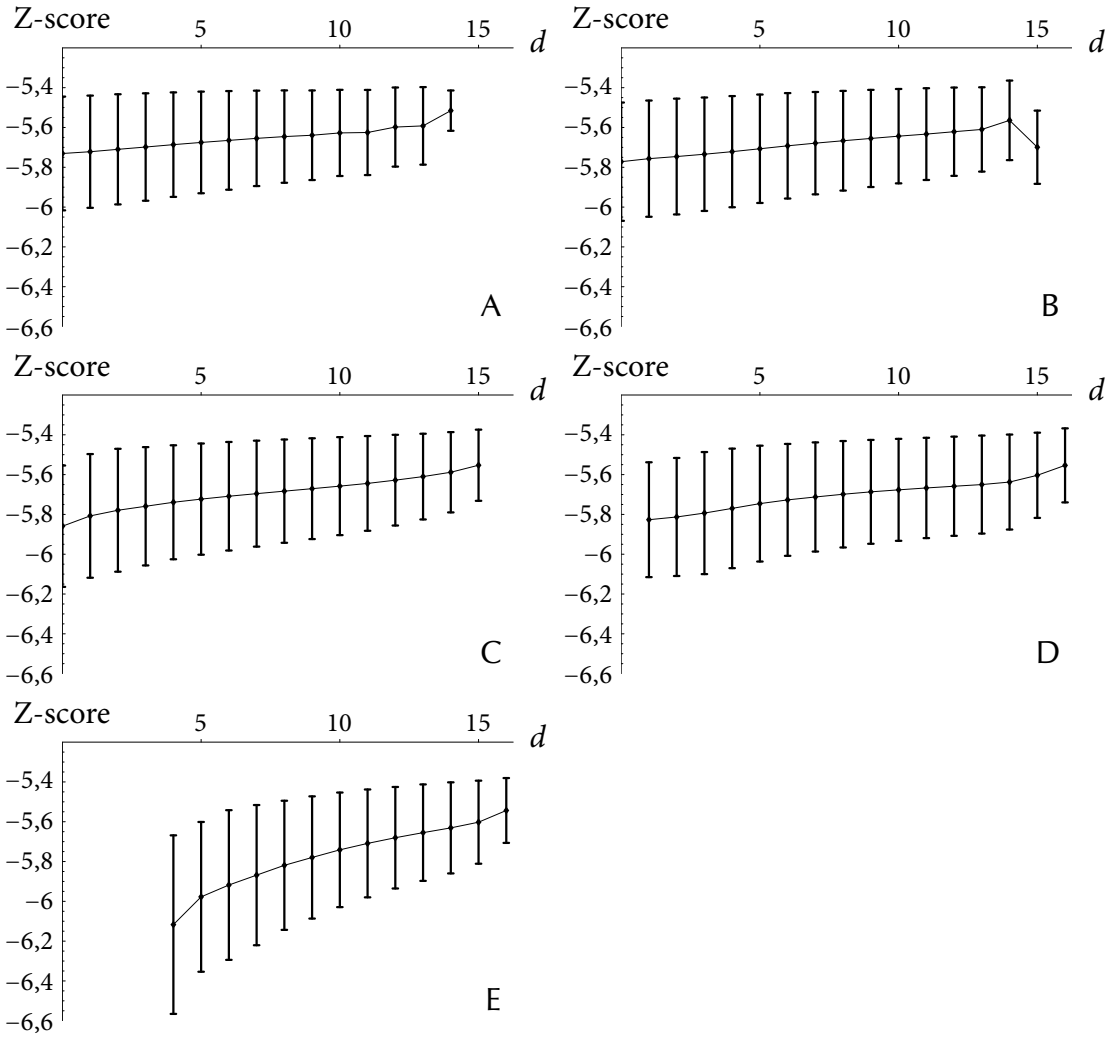


FIG. III.29 : Organisation en *superfunnel* dans l'espace des séquences pour la trajectoire 20C. Nous avons représenté le Z-score moyen et l'écart type en fonction de la distance au profil prototype $\hat{p}$. La même échelle a été utilisée pour tous les graphiques pour faciliter la comparaison. A — La distance de Hamming est calculée entre les profils à deux classes. B — La distance de Hamming est calculée entre les profils à trois classes. C — La distance de Hamming est calculée entre les profils à quatre classes. D — La distance de Hamming est calculée entre les profils à six classes. E — La distance de Hamming est calculée entre les profils à vingt classes.

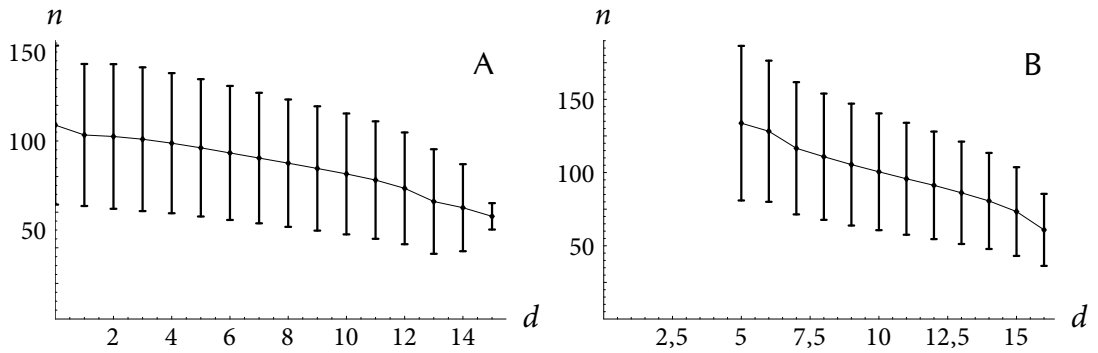


FIG. III.30 : Robustesse mutationnelle moyenne n et son écart type en fonction de la distance d au profil prototype $\hat{p}$. A — La distance de Hamming est calculée entre les profils à trois classes. B — La distance de Hamming est calculée entre les profils à vingt classes.

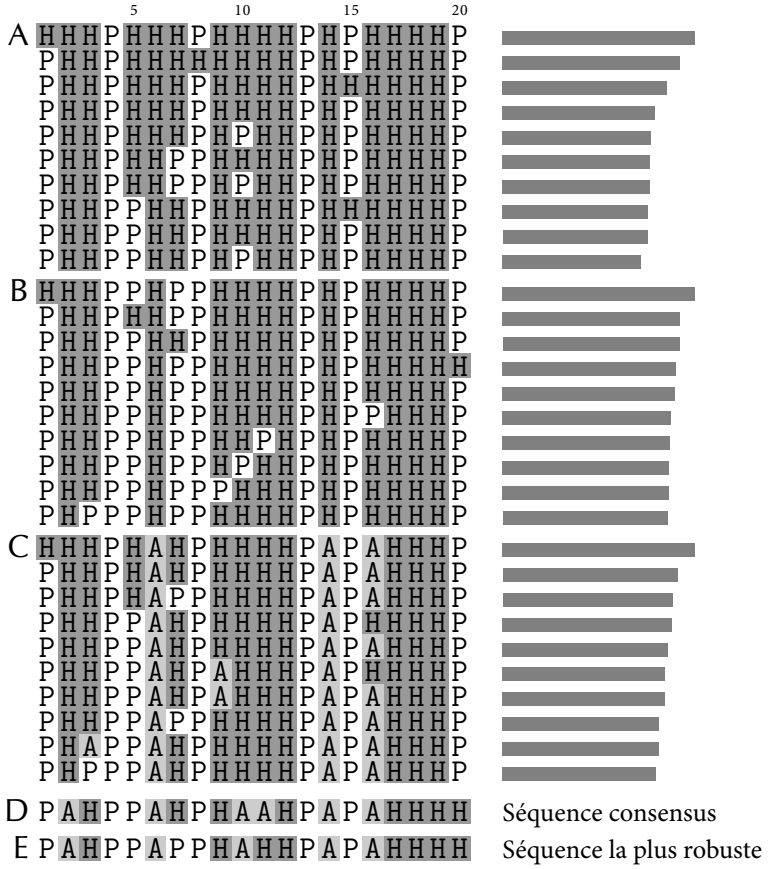


FIG. III.31 : A — Profils deux classes les plus peuplés à l'état stationnaire calculés sur le graphe deux classes construit à partir de la trajectoire 2C. B — Profils deux classes les plus peuplés à l'état stationnaire calculés sur le graphe deux classes construit à partir de la trajectoire 3C. C — Profils deux classes les plus peuplés à l'état stationnaire calculés sur le graphe trois classes construit à partir de la trajectoire 3C. D — Profil trois classes de la séquence consensus $\hat{s}$. E — Profil trois classes de la séquence consensus $\tilde{s}$. Les barres grises représente la fréquence à l'état stationnaire.

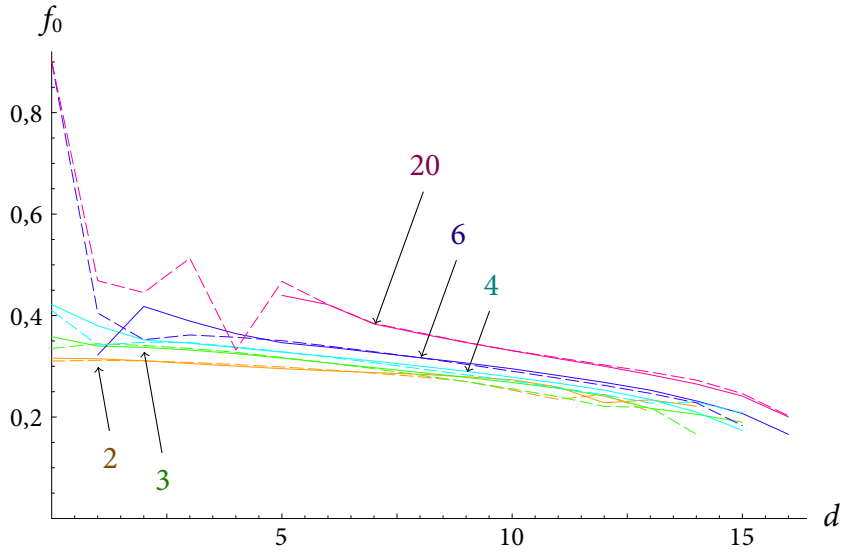


Fig. III.32 : Fraction f_0 de mutations neutres autour des séquences prototypes $\hat{s}$, bâtie à partir de la séquence consensus (ligne continue), et $\tilde{s}$, séquence la plus connectée de la trajectoire (pointillé). Les distances sont calculées sur les profils m classes des séquences acceptées durant la trajectoire 20C avec m = 20, 6, 4, 3 et 2.

La robustesse moyenne à distance l de l'une ou l'autre des séquences prototype $\hat{s}$ et $\tilde{s}$ se confond. Ces deux choix paraissent aussi pertinents l'un que l'autre ce qui confirme la conclusion de Bornberg-Bauer : la séquence prototype définie comme la plus robuste et la séquence consensus se confondent.

III.2.2.4 Propriétés évolutives

Évolution de la robustesse mutationnelle Comme pour les protéines sur réseau, deux distributions de la population seront utilisées pour estimer l'effet de l'évolution neutre : une distribution uniforme de la population et la distribution à l'état stationnaire. Elles ont été calculées pour tous les graphes mentionnés dans le tableau III.3. Les robustesses aux mutations moyennes sont données dans le tableau III.5. Des valeurs $\phi(n)$, jusqu'alors jamais observées, de 1,8-2,0 sont trouvées pour les protéines Grb2 et Vav.

Nom	$\langle n \rangle_u$	$\langle n \rangle_s$	$\phi(n)$
B10	5,6	11,0	1,96
C10	6,1	11,2	1,83
C1	8,7	14,7	1,70
C2	8,9	13,3	1,49
2C	7,5	9,5	1,27
3C	14,6	20,0	1,37

Tab. III.5 : Robustesse mutationnelle moyenne et augmentation de la robustesse moyenne pour toutes les trajectoires pour lesquelles une reconstruction du réseau neutre est possible.

Les distributions de la robustesse mutationnelle sont très similaires ; nous les illustrons à l'aide des résultats obtenus sur les graphes à deux classes issus des trajectoires B10 et C10 (cf. figure III.25, p. 110).

De la même manière que pour les réseaux neutres des protéines sur réseau, il est utile de mesurer l'impact de l'évolution par le facteur d'enrichissement $\phi(n) = \langle n \rangle_u / \langle n \rangle_s$. La figure III.33 et la table III.5 montrent six valeurs. Cinq de ces valeurs se rapportent à des réseaux deux classes (ceux reconstruits à partir des trajectoires B10, C10, C1, C2 et 2C) et l'une se rapporte à un réseau trois classes (trajectoire 3C). La figure III.33 les compare aux résultats des protéines sur réseau. Les deux réseaux neutres 2C et 3C de la protéine TRP-cage sont exactement alignés avec les protéines sur réseau. En ce sens, le TRP-cage est une protéine hors-réseau qui se comporte comme une protéine sur réseau. Les réseaux neutres C1 et C2 ne sont pas loin également du nuage de points obtenu à partir des protéines sur réseau. Les réseaux neutres des domaines SH3 de Grb2 et Vav suggèrent que la longueur de la chaîne protéique pourrait accroître l'effet de la dynamique de la population.

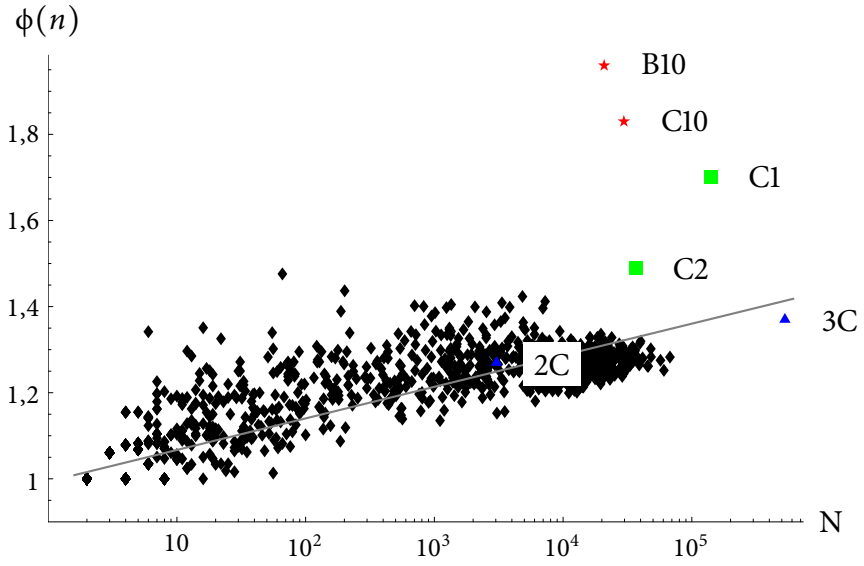


Fig. III.33 : Facteurs d'amélioration $\phi(n)$ en fonction de la taille N des réseaux neutres. En noir, les valeurs correspondant aux protéines sur réseau.

Exclusion des presqu'îles neutres À mesure que de nouvelles séquences sont acceptées au cours d'une trajectoire, la vitesse à laquelle sont découverts de nouveaux *profils* diminue. En particulier, dans les trajectoires réalisées avec des matrices à deux classes, cette observation est la manifestation du fait que l'énergie est principalement déterminée par le profil deux classes d'une séquence. Une fois que l'essentiel des profils compatibles avec une structure sont découverts, la trajectoire accepte de plus en plus des séquences dont le profil est déjà connu. À certaines occasions, cependant, le nombre de profils découverts

en fonction du temps bénéficie d'un regain de vitesse, comme montré dans la figure III.34 pour la trajectoire C2.

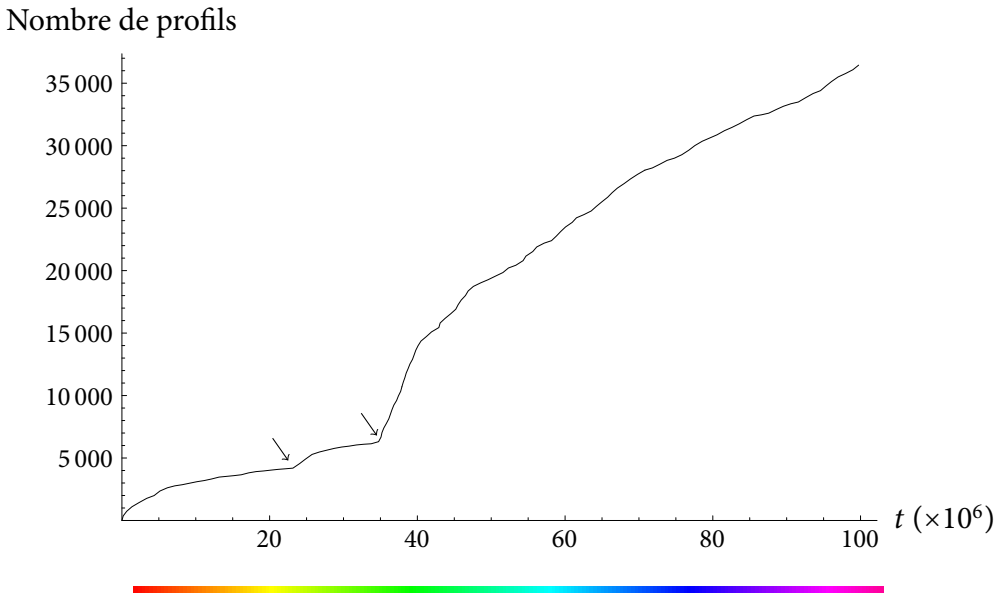


Fig. III.34 : Nombre de profils découverts au cours de la trajectoire C2 en fonction du temps (en nombre de pas de trajectoire). Les flèches indiquent les moments où la vitesse d'exploration des profils augmente soudainement.

Ces gains de vitesse surviennent lorsqu'une combinaison heureuse de mutations permet d'explorer une nouvelle région de l'espace des séquences. Principalement, trois régions sont explorées au cours de la trajectoire 2C pendant les périodes de temps $[0, 23 \cdot 10^6]$, $[23 \cdot 10^6, 35 \cdot 10^6]$ et $[35 \cdot 10^6, 100 \cdot 10^6]$. Nous nommons ces régions R_1, R_2 et R_3 respectivement.

L'analyse en composantes principales permet de décomposer des données selon des axes qui maximisent la quantité d'information. Ces axes sont les « composantes principales ». Cette technique peut donc être utile pour représenter en deux ou trois dimensions des objets complexes. En deux classes, tous les profils de la trajectoire C2 sont des coins de l'hypercube $\{0, 1\}^L$ (en codant H = 0 et P = 1). Nous avons recherché les deux premières composantes principales de l'ensemble des profils. Ces deux composantes définissent un plan. Ce plan est tel que la projection des données sur lui perde le moins d'information possible. Une analyse en composantes principales des profils générés lors de la trajectoire confirme en effet que plusieurs régions peuvent être distinguées : on distingue précisément R_1 d'une part et R_2 et R_3 d'autre part (cf. figure III.35, en haut à gauche). Ces trois régions forment des « presqu'îles neutres ».

Le calcul de l'état stationnaire sur le réseau neutre des profils deux classes de la trajectoire C2 met en évidence un phénomène intéressant : l'essentiel de la population se répartit dans les régions R_2 et R_3 laissant R_1 appauvrie (cf. figure III.35, en haut à droite).

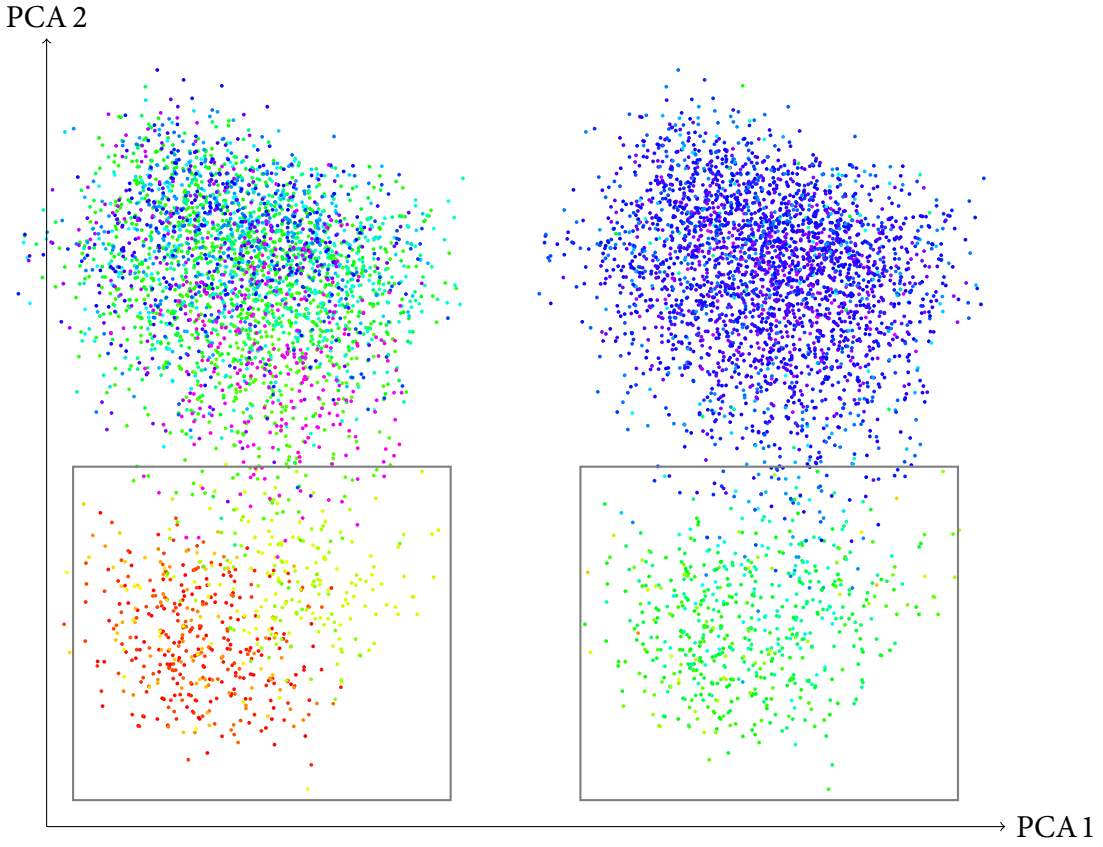


FIG. III.35 : Analyse en composantes principales des profils deux classes générés lors de la trajectoire C2. La région R_1 est encadrée. À gauche, la couleur représente le temps conformément à l'échelle de la figure III.34. À droite, la couleur représente la fréquence à l'état stationnaire conformément à l'échelle de la figure III.36.

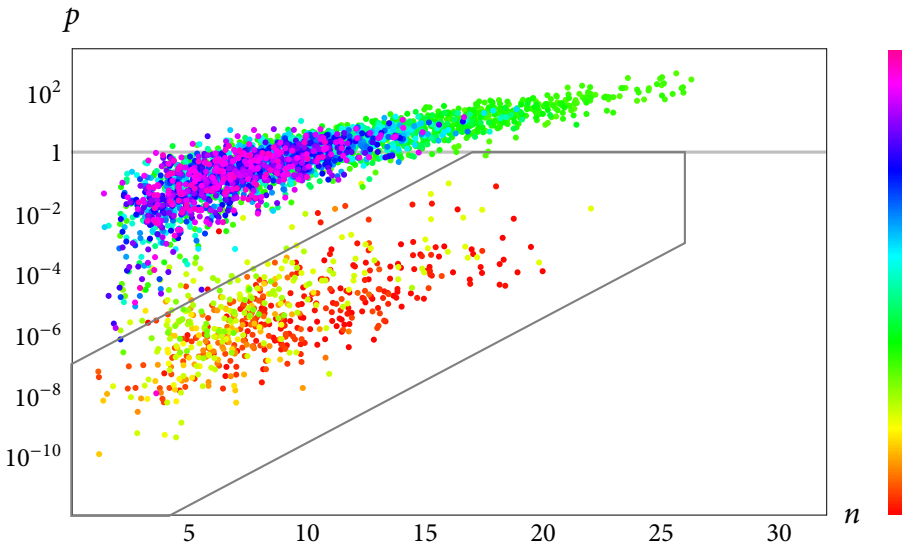


FIG. III.36 : Fréquences à l'état stationnaire (unité arbitraire) en fonction de la robustesse mutationnelle. Les couleurs du graphique représentent le temps selon l'échelle fixée dans la figure III.35. L'échelle de couleur à droite est utile à la compréhension du graphique droit de la figure III.35.

L'appauvrissement relatif de R_1 n'est cependant pas dû uniquement à des séquences plus faiblement connectées que dans R_2 ou R_3. Cet effet est plus prononcé : la représentation de la fréquence à l'état stationnaire en fonction de la robustesse mutationnelle, prouve bien que des séquences possédant la même robustesse mutationnelle dans R_1 et R_3 (n = 15–20) sont très différemment peuplées (cf. figure III.36). On en déduit que les caractéristiques évolutives de ce réseau neutre dépendent principalement de la plus vaste presqu'île neutre qui le compose.

III.2.3 Effet de la taille de la population

III.2.3.1 Du régime uniforme à l'état stationnaire

Notre modèle fait l'hypothèse d'une population infinie. La section « Polymorphisme, monomorphisme à l'état stationnaire » (p. 61), expose comment cette hypothèse permet de rendre compte de la dynamique d'une population finie mais ne discute pas la validité de cette hypothèse. La taille de nos réseaux neutres, que ce soit ceux du modèle de protéine sur réseau ou ceux reconstruits à partir de profils, sous-estiment la taille des réseaux neutres réels. Nous portons notre attention sur les trajectoires.

Il est bien sûr impossible que la composition de la population corresponde aux fréquences de l'état stationnaire dès que la population est finie. Par exemple, $22{\cdot}10^6$ séquences ont été acceptées au cours de la trajectoire 20C. Le réseau neutre qui contient cette trajectoire est certainement beaucoup plus grand encore. Si la taille de la population est inférieure à $22 \cdot 10^6$, il est évident que toutes les séquences ne peuvent être représentées. Dans le reste de cette section, nous utiliserons M pour la taille de la population et N pour la taille du réseau neutre d'une protéine.

Une question particulièrement importante se pose : sur un temps long T, peut-on s'attendre à ce que le nombre d'individus ayant porté une séquence s_i soit bien approximé par $p_i\,\mathrm{M}\,\mathrm{T}$ où p_i est la fréquence à l'état stationnaire théorique ? Autrement dit, a-t-on

$$\lim_{\mathrm{T}\to+\infty} \frac{1}{\mathrm{T}} \sum_{1\leq t\leq \mathrm{T}} \frac{\mathrm{M}_i(t)}{\mathrm{M}} = p_i\,, \tag{47}$$

où $\mathrm{M}_i(t)$ est le nombre d'individus porteurs de la séquence s_i au temps t ? Les résultats les plus importants à ce sujet ont été apportés par van Nimwegen et ses collègues [184, figure 3].

van Nimwegen *et al.* commencent par envisager une population finie (de taille M) dans laquelle le taux de mutation par séquence μ est si faible que, lorsqu'un mutant neutre apparaît, aucune autre mutation ne puisse survenir avant que le mutant ne se fixe ou ne soit éliminé. Le temps de fixation de mutants neutres étant de l'ordre de M générations,

cette situation se produit lorsque

$$M\mu \ll 1. \tag{48}$$

Lorsque la condition 48 est remplie, la population est composée d'un unique génotype en permanence sauf pendant de courts intervalles correspondant à la fixation d'un mutant.

En d'autres termes, la population se concentre quasiment à tout moment sur un nœud du réseau neutre, ne pouvant se déplacer vers des nœuds voisins que lorsqu'elle émet un mutant neutre qui par dérive génétique parvient à se fixer. La dynamique de la population sur le réseau neutre, si l'on fait abstraction des phases de fixation, est un « chemin aléatoire » sur le réseau neutre. La probabilité que la population se trouve sur un nœud particulier du réseau neutre après un temps long *dans cette situation particulière* est connue[17] : elle vaut exactement 1/N (une distribution uniforme). La distribution uniforme que nous avons utilisée jusqu'à présent pour mesurer l'impact de l'évolution neutre d'une population infinie, est en fait le régime sous lequel évolue une population finie vérifiant 48.

Lorsque la population devient de plus en plus grande, l'expression 47 devient vraie. La relation 48 suggère que cela dépend du produit M μ plutôt que de la taille M seule. La figure 3 de la référence [184] confirme que la transition du régime uniforme à l'état stationnaire est uniquement une fonction de M μ (nous la reproduisons dans la figure III.37). Quand M μ est supérieur à 500, l'égalité 47 est vérifiée.

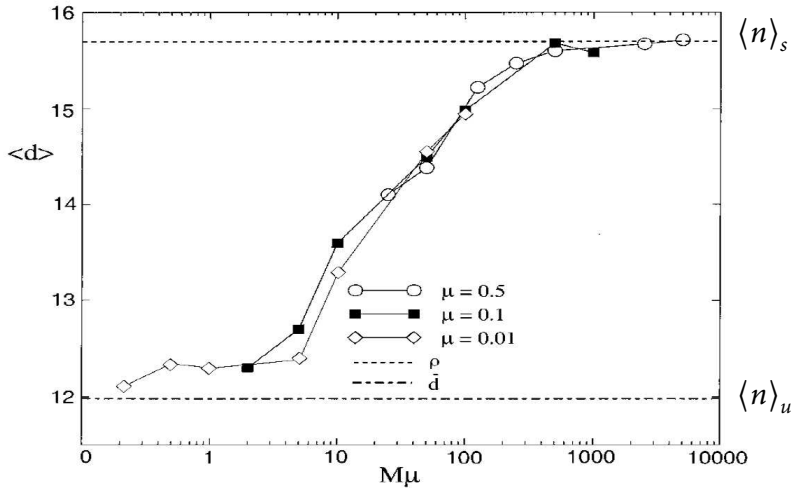


Fig. III.37 : Évolution de la robustesse mutationnelle moyenne $\langle n \rangle$ en fonction de la quantité M μ (d'après la référence [184]).

17. Techniquement, la dynamique de la population obéit aux règles d'une chaîne de Markov dont il est possible de connaître le vecteur probabilité quand le temps t tend vers l'infini.

On note, dans ce qui suit, $p_i^{(\mathrm{M})}$ le membre de gauche de l'équation 47

$$p_i^{(\mathrm{M})} = \lim_{\mathrm{T}\to+\infty} \frac{1}{\mathrm{T}} \sum_{1\leq t\leq \mathrm{T}} \frac{\mathrm{M}_i(t)}{\mathrm{M}}.$$

Des simulations ont été réalisées sur un réseau de taille 99, par Éric Brunet[18]. Les valeurs $p_i^{(\mathrm{M})}$ et de p_i ont été comparées pour différentes valeurs de M et avec un taux de mutation $\mu = 1$. Les résultats sont présentés dans la figure III.38. On peut approximer $p_i^{(\mathrm{M})}$ par les relations affines suivantes :

$$\begin{aligned}
p_i^{(1)} &= 1/\mathrm{N},\\
p_i^{(2)} &\approx 1/\mathrm{N} + 0{,}101\,(p_i - 1/\mathrm{N}),\\
p_i^{(3)} &\approx 1/\mathrm{N} + 0{,}183\,(p_i - 1/\mathrm{N}),\\
p_i^{(10)} &\approx 1/\mathrm{N} + 0{,}516\,(p_i - 1/\mathrm{N}),\\
p_i^{(20)} &\approx 1/\mathrm{N} + 0{,}706\,(p_i - 1/\mathrm{N}),\\
p_i^{(50)} &\approx 1/\mathrm{N} + 0{,}872\,(p_i - 1/\mathrm{N}).
\end{aligned}$$

Pour des valeurs supérieures de M on a à peu près :

$$p_i^{(\mathrm{M})} - 1/\mathrm{N} \approx (1 - 7/\mathrm{M})\,(p_i - 1/\mathrm{N}). \tag{49}$$

Si l'approximation donnée par l'équation 49 est valide, en prenant la valeur $\mathrm{M}\,\mu = 500$ de van Nimwegen *et al.*, l'hypothèse de population infinie crée une erreur de l'ordre de 1 %.

III.2.3.2 Influence de la taille du réseau neutre

La valeur limite avancée par van Nimwegen *et al.* est-elle dépendante du réseau neutre sur lequel ils ont travaillé ? En particulier, cette valeur dépend-elle de la taille N du réseau ? L'importance de cette question peut être réalisée encore une fois en imaginant les tailles des réseaux neutres réels.

Diverses simulations ont été menées avec M = 1, 2, 10 et 50 individus et un taux de mutation élevé $\mu = 1$. Neuf réseaux neutres, construits avec le modèle des protéines sur réseau et la matrice LHTW, de tailles diverses ont été utilisés pour chacune des population : N = 99, 199, 504, 1 013, 2 037, 5 028, 10 079, 20 253 et 48 248. Les deux membres de l'égalité 47, $p_i^{(\mathrm{M})}$ et p_i sont comparés. Les résultats sont présentés dans la figure III.39

18. Je lui suis très reconnaissant de m'avoir accordé beaucoup de son temps pour s'intéresser à mes problèmes, répondre à mes questions, et mettre le doigt sur mes défaillances.

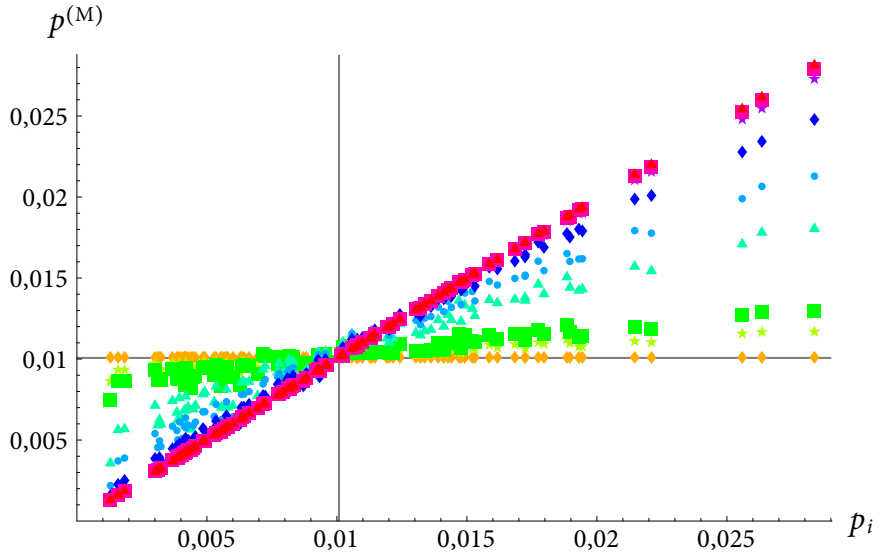


Fig. III.38 : Calculs réalisés sur un graphe de N = 99 nœuds avec les valeurs M = 1, 2, 3, 10, 20, 50, 200 et 1000 et un taux de mutation μ = 1. On peut approcher les données par des relations affines se croisant toutes au point $(1/N, 1/N)$ (communication d'Éric Brunet).

D'après ces résultats, il est tout à fait remarquable que la taille du réseau neutre n'intervienne pas la convergence des $p_i^{(M)}$ vers l'état stationnaire à population infinie.

III.2.3.3 Effet du modèle évolutif : modèle de Wright-Fisher

Dans notre modèle (protocole 1), le mode reproductif est exact, chaque individu tente de se reproduire une fois. Les fluctuations dans la taille de la progéniture ne proviennent que des mutations létales et du remplacement de l'enfant moribond par l'enfant d'un individu tiré au hasard dans la population (étape 4 du modèle évolutif, p. 54). Le modèle de Wright-Fisher modélise des fluctuations dans la taille de la progéniture indépendamment de la réussite ou de l'échec des mutations (cf. aussi l'introduction, p. 28). Au lieu de faire se reproduire les N individus composant la population, N individus sont tirés au sort avec remise et ce sont ces individus qui tentent de se reproduire. La taille de la progéniture suit approximativement une loi de Poisson dont le paramètre vaut un. Ce modèle stochastique rend compte de la dérive génétique, c'est-à-dire de l'extinction de certaines lignées, et la fixation d'un allèle.

Nous proposons d'intégrer ces fluctuations qui sont importantes dans la compréhension de la composition allélique d'une population dans notre modèle. Nous proposons donc le protocole suivant.

Protocole évolutif 4

On exécute M fois les étapes suivantes.

1. Un individu de la génération mère *choisi au hasard* se reproduit fidèlement avec une probabilité 1 − μ.

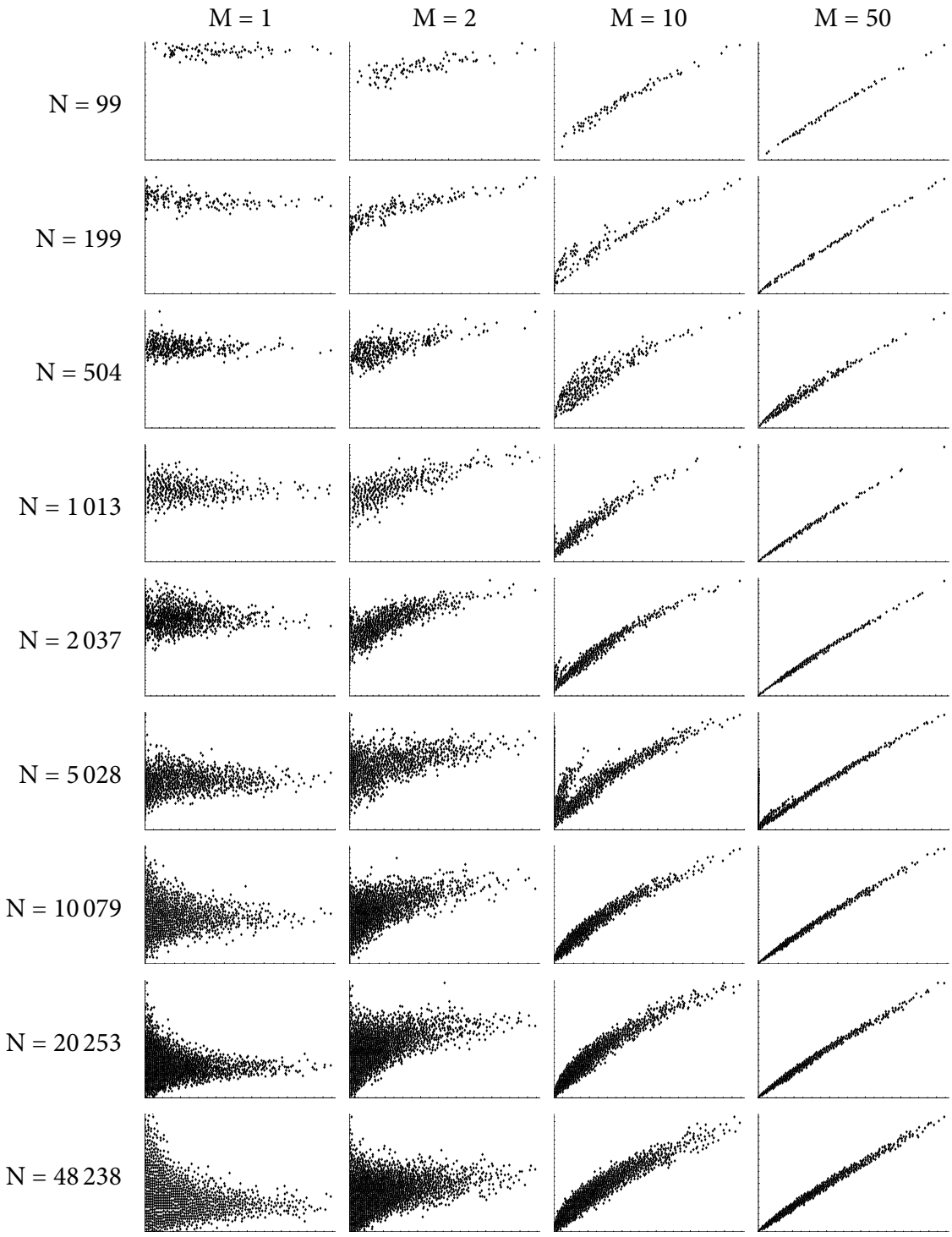


Fig. III.39 : Comparaison de la fréquence des séquences d'une population finie (en abscisse) à celle d'une population infinie (en ordonnée) pour différentes tailles N de réseau neutre et différentes tailles M de population finie. Les unités sont arbitraires.

2. Avec une probabilité μ, il subit une mutation ponctuelle affectant l'une des L positions de la protéine étudiée. Les mutations aux L positions possibles sont équiprobables. Le nombre de mutations accessibles est $\ell = L(A-1)$.
3. Le mutant obtenu en 2 est conservé s'il appartient à l'ensemble s (mutation neutre).
4. Sinon, il est éliminé (mutation létale) et est remplacé par l'enfant d'un individu choisi au hasard dans la génération mère.

Une variante possible serait le protocole suivant.

Protocole évolutif 5

On exécute jusqu'à obtenir M enfants viables les étapes suivantes.

1. Un individu de la génération mère *choisi au hasard* se reproduit fidèlement avec une probabilité $1-\mu$.
2. Avec une probabilité μ, il subit une mutation ponctuelle affectant l'une des L positions de la protéine étudiée. Les mutations aux L positions possibles sont équiprobables. Le nombre de mutations accessibles est $\ell = L(A-1)$.
3. Le mutant obtenu en 2 est conservé s'il appartient à l'ensemble s (mutation neutre).
4. Sinon, il est éliminé (mutation létale).

Les graphiques de la figure III.6 représentent $\langle n \rangle$ en fonction du temps avec les paramètres donnés dans la table III.6, avec et sans implémentation de la reproduction selon le modèle de Wright-Fisher.

La dérive génétique du modèle de Wright-Fisher est une force agissant contre la diversité. Ainsi le nombre de séquences N_s est-il réduit. Les courbes de N_s et $\langle n \rangle$ en fonction du temps sont par conséquent plus bruitées. La population se reproduisant sans fluctuation dans la taille de la progéniture (protocole 1) se comporte comme une population plus nombreuse évoluant selon le protocole 4.

La robustesse mutationnelle $\langle n \rangle_T$ moyennée sur un temps T long, indique l'état de transition entre le régime uniforme ($M\mu \gg 1$, $\langle n \rangle_u = 5{,}71$) et l'état stationnaire théorique ($M\mu \ll 1$, $\langle n \rangle_s = 7{,}92$). On résume les différents résultat dans la table III.6. On voit que le protocole 4 nécessite des populations plus nombreuses pour atteindre le régime des populations infinies.

III.3 Discussion

Dans les paragraphes qui suivent, nous ferons référence à des figures et des tables des annexes (p. 145 et suivantes). Pour faciliter la lecture, elles seront indiquées par une astérisque.

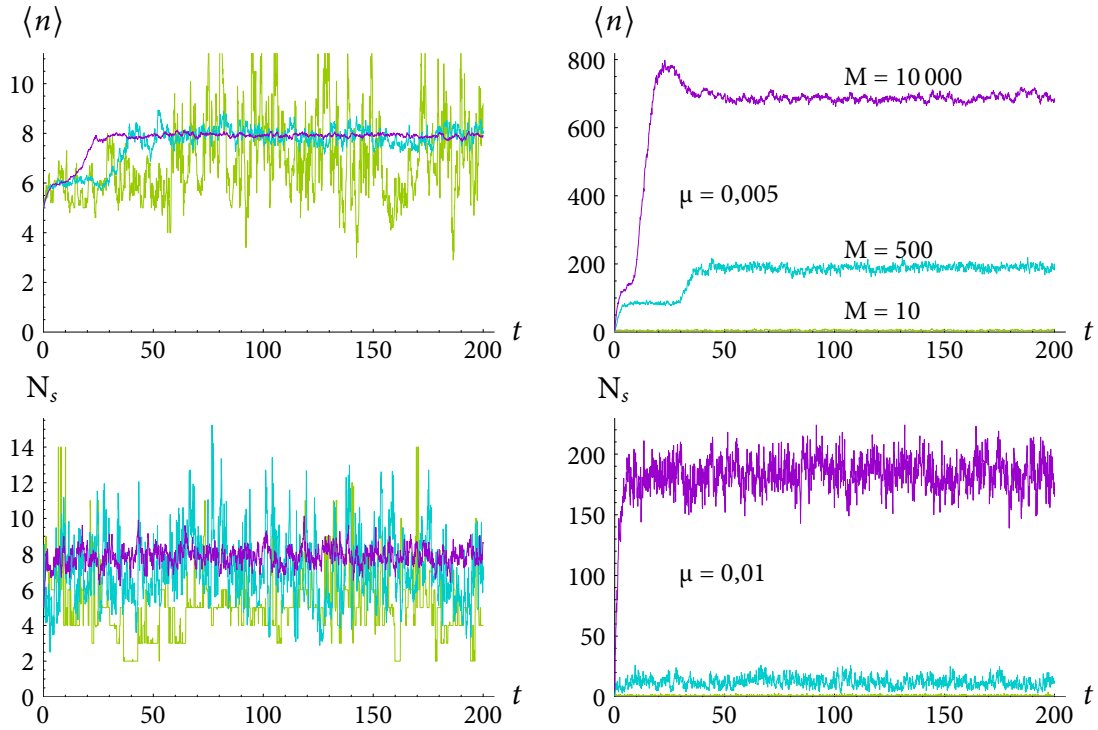


FIG. III.40 : Évolution de $\langle n \rangle$ et nombre de séquences dans la population en fonction du temps selon les paramètres de la table III.6. En haut, chaque séquence se reproduit au moins une fois selon le protocole défini dans « Modèle et méthodes ». En bas, les séquences se reproduisant sont choisies aléatoirement par un tirage avec remise (modèle de Wright-Fisher). N_s est le nombre de séquences différentes présentes à tout moment dans la population. L'unité de temps représente 1 000 générations.

	Taux de mutation μ	*Taille de la population* M	M μ	*Diversité* $\langle N_s \rangle_T$	*Robustesse* $\langle n \rangle_T$
	0,005	10	0,05	5	7,13
non WF	0,005	500	2,5	190	7,92
	0,005	10 000	50	700	7,92
	0,01	10	0,1	1	5,22
WF	0,01	500	5	10	7,25
	0,01	10 000	100	180	7,90

TAB. III.6 : Valeurs de $\langle n \rangle$ pour les différentes simulations sans et avec implémentation du modèle de Wright-Fisher. Les moyennes sont calculées entre les temps $50 \cdot 10^3$ et $200 \cdot 10^3$ générations où la diversité N_s des séquences est stable et sont notées $\langle \cdot \rangle_T$. La diversité $\langle N_s \rangle_T$ est le nombre moyen de séquences présentes dans la population dans ledit intervalle. La robustesse moyenne $\langle n \rangle_T$ est la moyenne de la robustesse moyenne instantanée présentée dans les graphiques de la figure III.40.

III.3.1 Structure des réseaux neutres

III.3.1.1 Organisation de *superfunnel*

Pour décrire la structure des réseaux neutres, le terme de *superfunnel* a été introduit [21, 202]. Ce terme recouvre plusieurs phénomènes :

1. une organisation autour d'une séquence prototype qui ressemble à la séquence consensus obtenue en « alignant » l'ensemble des séquences ;
2. une stabilité thermodynamique qui croît à mesure que l'on s'approche de la séquence prototype.

Le premier point est illustré par la figure III.9 pour les protéines sur réseau et les figures III.28 et III.31 pour les protéines tridimensionnelles. Le second point est mis en évidence par les figures III.11, III.12, III.13, III.57* dans le cas des protéines sur réseau et la figure III.29 pour le trp-cage. On notera que le *superfunnel* mesuré par le Z-score (voir la figure III.29) n'est pas très marqué, sans doute parce que le Z-score est une mesure peu sensible aux énergies des premiers états excités qui affectent énormément la stabilité de la protéine. On sait qu'en général, des valeurs de ΔE élevées sont associées à une meilleure accessibilité cinétique.

Les séquences les plus proches de la séquence prototype sont également plus robustes aux mutations. Les mêmes figures que précédemment peuvent être citées pour en rendre compte. Les figures III.30 et III.32 prouvent que ce n'est pas un artefact de l'utilisation d'alphabets réduits ou de modèles de protéine simplifiés. Nous pouvons expliquer ce constat par le fait que les séquences proches de la séquence prototype, étant plus stables, peuvent s'accomoder de plus de mutations déstabilisatrices que celles situées dans la périphérie du réseau neutre.

Un point intéressant est à noter dans les résultats utilisant des alphabets possédant plus de deux classes. Les travaux antérieurs de Bornberg-Bauer se fondaient sur un alphabet HP. Babajide *et al.* ont trouvé que les séquences des protéines étaient flexibles dans un alphabet à vingt acides aminés [1] mais les profils HP restent conservés : « [sequences belonging to a neutral network are] very flexible at the level of individual amino acids but require a significant level of conservation of amino acid classes ». Puisqu'il existe une forte conservation dans l'alphabet HP, la diminution de la robustesse à mesure qu'augmente la distance à la séquence prototype pourrait donc être un artefact de l'utilisation d'un alphabet HP. La figure III.32 montre que cette hypothèse est vérifiée également dans des alphabets plus complets.

La moyenne sur un réseau neutre de la robustesse mutationnelle augmente linéairement avec le logarithme de la taille N du réseau neutre (cf. figures III.7 et III.46*). Cette

propriété paraît universelle, car elle est commune à toutes les matrices d'énergie, y compris celles qui n'ont pas un caractère peptidique (cf. relations 39a et 39b, p. 70) comme Ising ou TS_2 (cf. figure III.46*). Par ailleurs, les distances de Hamming maximales et les diamètres augmentent également linéairement avec le logarithme de la taille du réseau neutre (cf. figure III.14). Ces grandeurs caractérisent l'*étendue* et donc la diversité des séquences au sein d'un réseau neutre. Cela suggère qu'un réseau neutre ne croît pas seulement par accrétion en périphérie mais également par densification au centre

III.3.1.2 Mutations compensatrices dans les petits réseaux

La taille du réseau neutre influence d'autres propriétés. Peut-être la relation la plus spectaculaire est celle relative à $\langle \Delta E \rangle$, la différence d'énergie moyenne entre l'état fondamental et le premier état excité (cf. figures III.8 et III.49*). Pour la plupart des matrices d'énergie, il y a un phénomène de seuil : en dessous d'une certaine taille N_{min}, $\langle \Delta E \rangle$ est constant et minimal, lorsque la taille du réseau devient supérieure à ce seuil, la stabilité moyenne augmente soudainement[19]. Le saut est important pour LHTW et HP′. La valeur du seuil dépend de la matrice d'énergie et se situe aux alentours de $5 \cdot 10^3$ pour les matrices LHTW et HP′.

Li *et al.* ont observé que le seuil N_{min} était également la taille minimale au-dessus de laquelle il devenait possible d'estimer approximativement la taille N d'un réseau neutre à partir des entropies de Shannon à chaque position [112]. Si l'on note σ_i l'entropie de Shannon à la position i, on a

$$\log N \approx \sum \sigma_i .$$

Les entropies deviennent additives. Cette observation signifie qu'à partir de cette taille, les positions d'une protéine mutent indépendamment les unes des autres. Il est intéressant de noter que le seuil N_{min} correspond également à une transition dans les diamètres et distances de Hamming maximales (cf. figure III.14) : en dessous de N_{min}, la tendance est à la divergence entre l_{max} et d_{max} ; pour des tailles supérieures à N_{min}, l_{max} et d_{max} coïncident à nouveau.

Deux « mutations compensatrices » sont deux mutations létales, prises isolément, qui deviennent viables lorsqu'elles surviennent de concert. Nous négligeons les mutations doubles dans notre modèle. Nous étendons donc ce concept pour l'adapter aux réseaux neutres. Pour nous, des mutations compensatrices seront un couple de mutations tel que la seconde soit létale si la première n'a pas été réalisée en premier lieu.

19. La stabilité d'une conformation peut être mesurée par la température de repliement ou la différence d'énergie libre ΔG entre les états dénaturés et la conformation native. On sait que la stabilité est généralement corrélée à la différence d'énergie ΔE entre l'état fondamental et le premier niveau d'énergie excité.

L'existence d'un phénomène de seuil pour $\langle \Delta E \rangle$ et dans la coïncidence de l_{max} et d_{max} suggère, que dans les réseaux neutres de tailles $N < N_{min}$, les mutations compensatrices prédominent. Car dans ce cas, l'additivité des entropies de Shannon n'est plus vraie et les mutations ne sont plus indépendantes. Dans le même temps, si le recours à des mutations compensatrices est nécessaire pour assurer la connectivité du réseau neutre, cela signifie que la stabilité des séquences est marginale et incapable de tolérer une seule mutation déstabilisatrice sans passer par une première mutation qui la rende moins délétère. Enfin, si les mutations compensatrices prédominent, le diamètre augmente nécessairement plus vite que la distance de Hamming maximale (cf. le schéma de la figure III.41). Ce type de mutation est, à quelques détails près, ce que Kimura appela « mutation neutre compensatrice » [96].

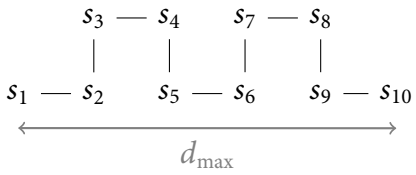


Fig. III.41 : Quand dans un réseau neutre, les mutations compensatrices dominent, le diamètre $(s_1, \ldots, s_{10})$ est plus grand que la distance de Hamming maximale (l_{max}).

III.3.1.3 Distribution de la robustesse mutationnelle

Nous avons observé que la distribution du nombre de connexions possède une forme similaire quelle que soit la matrice d'énergie utilisée (figures III.6, III.5, III.51* et III.50*). Les distributions issues du modèle de protéine tridimensionnelle sont également extrêmement semblables, que ce soit dans les graphes reconstruits (cf. figure III.25) ou d'après la relecture des trajectoires (cf. figure III.26).

Ces distributions sont caractérisées par une forme allongée indiquant la présence de quelques rares séquences extrêmement robustes aux mutations. Ces séquences robustes confèrent une structure proche des réseaux sans échelle (cf. figure III.27). Des exceptions existent lorsque la longueur de chaîne limite l'extension de la distribution et provoque un phénomène de saturation (cf. la courbe GLO de la figure III.50*, certaines courbes de la figure III.26).

Si l'alphabet est composé de A lettres et la longueur de chaîne est L, la fraction de mutations neutres, f_0, peut être écrite comme une fonction de la robustesse mutationnelle $\langle n \rangle$

$$f_0 = \frac{\langle n \rangle}{L(A-1)}. \tag{50}$$

L'utilisation de $\langle n \rangle_u$ dans le membre droit de l'équation 50 donne des fractions valant 9,8 %, 8,8 %, 12,6 % et 12,9 % respectivement pour les réseaux neutres reconstruits à partir des trajectoires B10, C10, C1 et C2. L'utilisation de $\langle n \rangle_s$ mène à 19,3 %, 16,2 %, 21,3 % et 19,3 %. Ces fractions sont cohérentes avec les valeurs typiquement avancées. Kimura estima, d'après son taux d'évolution, à 14,3 % la fraction de mutations neutres dans l'hémoglobine [94, p. 101, 102]. King et Jukes proposèrent, en se fondant sur les travaux de Mukai [139] et sur ceux de Whitfield [73], des fractions typiques allant de 5 % à 10 % [100]. La mutation systématique en glycine et alanine est déstabilisatrice dans 82 % des cas [70, 123, 166]. Il semble donc que la forme des distributions observées pour les domaines SH3 de Grb2 et Vav sont proches de la réalité biologique. Bastolla et ses collaborateurs ont calculé une distribution de la fraction de mutations neutres qui diffère sensiblement de la nôtre [7]. Elle nous paraît irréaliste par le nombre important de mutations neutres tolérées (en moyenne plus de 60 %). La forme que nous observons par ailleurs ne semble pas être un artefact du modèle d'énergie par interaction entre les résidus en contact puisque Bastolla *et al.* ont mis en œuvre un modèle très proche.

La forme caractéristique des distributions de la robustesse mutationnelle est assez bien rendue par le modèle d'énergie aléatoire que nous exposons en annexe (cf. figure III.52*). Elle semble pouvoir s'expliquer sans faire intervenir d'hypothèses *ad hoc* sur les conformations. (Voir également les résultats obtenus avec des alphabets plus complets, figure III.62*, p. 159.)

Du modèle d'énergie aléatoire peuvent être tirées deux conclusions. Premièrement, peu de séquences ne tolèrent aucune mutation (ou peu de mutations). L'état fondamental dispose d'un avantage énergétique, quantifié par ΔE, sur les premiers niveaux excités. Pour qu'une mutation soit rejetée, il faut que les premiers états excités arrivent à franchir le fossé ΔE. En essayant de muter les vingt-cinq positions de la chaîne, on dispose de vingt-cinq tentatives pour réaliser ce gain d'énergie. Les chances d'y parvenir vingt-cinq fois sont faibles.

Néanmoins, les mutations sont généralement plus déstabilisatrices pour le niveau d'énergie fondamental que pour les premiers niveaux excités. Il est, du coup, improbable qu'aucune des vingt-cinq mutations ne parviennent à déstabiliser suffisamment l'état fondamental pour qu'il passe au-dessus de l'un des premiers niveaux excités. On montre que si les mutations ne sont pas plus déstabilisatrices pour le niveau d'énergie fondamental[20], alors de nombreuses séquences tolèrent le nombre maximal de mutations neutres.

20. Nous avons simulé ce scénario en supposant que la déstabilisation d'un niveau d'énergie était une gaussienne centrée sur le niveau d'énergie.

III.3.2 Modifiabilité

III.3.2.1 *Designability principle*

La modifiabilité d'une conformation est le nombre de séquences dont elle est la structure native. La distribution des modifiabilités parmi les structures est souvent décrite par une loi de Zipf. Cette loi exprime une forte inégalité dans la répartition des séquences : quelques rares structures collectent un nombre considérable de séquences, tandis que la majorité des structures est très peu modifiable.

Pour les potentiels inspirés des interactions hydrophobes (cf. équations 39a et 39b, p. 70), on peut voir que l'essentiel des séquences se repliant en une conformation est regroupé dans un grand réseau neutre (cf. les matrices LHTW, HP, HP′ dans la figure III.2 et les matrices TS_1 et GLO dans la figure III.45*). La modifiabilité d'une structure se confond donc quasiment avec la taille du plus grand de ses réseaux neutres. Ce résultat peut être sujet à caution à cause de la considération de structures compactes seulement.

Les conformations très modifiables sont caractérisées par une plus grande stabilité en moyenne (cf. figures III.8, III.48* et III.49*, et références [102, 124, 207]) et une plus forte robustesse aux mutations (cf. figure III.7). L'importance théorique, notamment dans le cadre du *protein design* [80, 87], de ce phénomène a pris le nom de « designability principle » [47, 111, 112, 114, 130] et a motivé le développement de méthodes permettant d'estimer la modifiabilité d'une structure. Meyerguz *et al.* établirent que, sous certaines conditions, l'estimation de la modifiabilité (ce qu'il appelle *evolutionary capacity*) prend la forme d'un problème de Knapsack[21] [126]. D'autres méthodes recourent à des analyses en composantes principales [117, 206]. Sans doute l'un des résultats les plus importants est celui qu'ont apporté England et Shakhnovich qui montrèrent qu'une grande modifiabilité était caractérisée par de plus grandes traces des puissances de la matrice de contact [49]. Pour une application à la modifiabilité des protéines des organismes thermophiles, voir référence [48].

III.3.2.2 Modifiabilité et taux d'évolution

Il existe un lien au moins théorique entre la modifiabilité et le taux de substitution. En effet, la vitesse d'évolution d'une protéine dépend de la fraction f_0 de mutations neutres. La robustesse mutationnelle $\langle n \rangle$ croissant avec la modifiabilité, il doit en aller de même pour la fraction f_0 de mutations neutres, d'après l'égalité 50. Cette hypothèse apparaît également dans les travaux de Bloom *et al.* [16].

21. Ce problème consiste à trouver un sous-ensemble d'une liste de couples valeur-coût (v_i, c_i) tel que la somme des valeurs soit maximale et la somme des coûts soit bornée par une certaine constante C.

On s'attend à ce que l'effet soit encore renforcé par la tendance de l'évolution neutre à peupler les séquences très connectées. Si l'hypothèse de la population infinie s'applique, c'est $\langle n \rangle_s$ qui doit être utilisé dans l'équation 50. Or, $\langle n \rangle_s$ s'écrit comme le produit de $\langle n \rangle_u$ et $\phi(n)$, qui sont deux fonctions croissantes de N (cf. figures III.7 et III.33).

III.3.2.3 Modifiabilité et structures secondaires

Les mêmes conformations apparaissent dans la liste des conformations les plus modifiables pour plusieurs matrices d'énergie (cf. tableaux III.1, III.2, III.7* et III.8* et figure III.2). Il a été remarqué que ces conformations (en particulier la conformation 786) possèdent des régularités réminiscentes de structures secondaires [112]. Nous pensons que ce résultat est, au moins partiellement, artéfactuel (cette opinion est partagée par Blackburne et Hirst [13]). Effectivement, en se limitant aux conformations compactes, on introduit une contrainte structurale importante. Les structures en hélice comme celles de la conformation 786, à cause de cette contrainte, sont peu propices à de petits réarrangements structuraux. Nous pouvons quantifier cela par le recouvrement structural Q que nous avons introduit à la page 79. Un recouvrement entre deux structures proches de un signifie qu'il est possible de passer de l'une à l'autre par de petits réarrangements structuraux. La structure 786 bénéficie d'un faible recouvrement structural. En effet, les structures les plus proches de 786 (cf. figure III.42) possèdent moins de dix contacts en commun avec elle. En comparaison, la conformation 700 est la moins modifiable pour les trois matrices d'énergie HP, HP′ et LHTW (cf. figure III.4). La conformation la plus proche possède treize contacts en commun (cf. figure III.43).

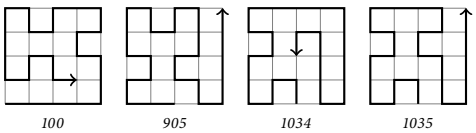


Fig. III.42 : Les conformations les plus proches de la conformation 786. Le recouvrement Q vaut 10, 9, 8 et 8, respectivement.

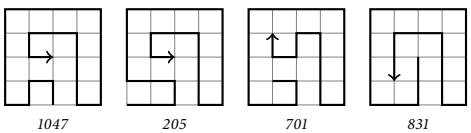


Fig. III.43 : Les conformations les plus proches de la conformation 700. Le recouvrement Q vaut 13, 12, 12 et 12, respectivement.

Or la modifiabilité est en partie déterminée par le recouvrement structural Q (cf. figure III.54*). Intuitivement, deux structures proches structuralement (Q ~ 1) sont en compétition pour « obtenir des séquences ». Cette compétition engendre de la dégénérescence

et diminue donc le nombre de séquences capables de se replier dans l'une ou l'autre conformation.

Dans le même temps, les hélices de la structure 786 sont en surface, c'est-à-dire en interaction avec le solvant ; elles garantissent un grand nombre de contacts entre résidus enfouis ce qui stabilise la structure. Ce n'est donc pas la structure secondaire elle-même qui est intéressante, mais son aptitude à augmenter les contacts hydrophobes. Cet argument permet de prédire qu'une hélice n'est favorable que lorsqu'elle est contact avec le solvant. Il existe en effet des structures possédant des hélices enfouies qui n'ont pas de modifiabilité particulièrement élevée (cf. figure III.44). De plus, les hélices ne sont avantageuses que si les interactions H-H sont les plus stabilisatrices. On observe effectivement que la conformation 786 disparaît des structures très modifiables avec la matrice Ising dont les interactions H-H et P-P sont répulsives (cf. tableau III.8*).

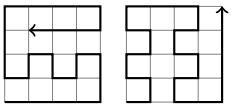


Fig. III.44 : Deux conformations possédant des « structures secondaires » mais formant moins de contacts entre résidus enfouis et susceptibles d'accomoder de plus faibles réarrangements structuraux. Les modifiabilités de ces conformations valent, avec la matrice LHTW, 8 952 et 11 919, respectivement (la moyenne est 11 000 environ).

En résumé, l'apparition des structures secondaires observée par Li *et al.* est due à deux effets qui jouent conjointement.

1. Une structure qui compte tenu de la contrainte structurale des conformations compactes est très dissemblable de toute autre (faible recouvrement structural avec les autres structures).
2. Une structure qui augmente le nombre de contacts entre résidus enfouis.

L'artefact réside à notre avis en ce qu'aucun de ces points n'est clairement établi dans le cas des protéines biologiques. Premièrement, le recours à des structures compactes n'est pas fondé sur une contrainte *a priori*, mais bien, *a posteriori*, sur l'observation qu'elles sont favorisées par rapport aux structures non compactes. Deuxièmement, les structures secondaires des conformations biologiques charpentent les structures en les traversant, la surface d'une protéine étant la plupart du temps composée de boucles. Nous pensons par conséquent que l'existence des structures secondaires est bien mieux expliquée par la théorie classique qui postule que les structures secondaires permettent de masquer les donneurs et les accepteurs de liaison hydrogène.

III.3.3 Évolution neutre adaptative

III.3.3.1 Évolution adaptative vers la séquence prototype

L'observation la plus surprenante parmi nos résultats est que, du modèle d'évolution neutre, découle un *caractère adaptatif* de la population. Un flux de population s'amorce vers les séquences stables thermodynamiquement et robustes aux mutations du réseau neutre (cf. figure III.21). Une séquence tirée au hasard dans une population à l'état stationnaire ressemble plus à la séquence prototype qu'une séquence tirée au hasard dans le réseau neutre (cf. figure III.22). À cause de la structure en *superfunnel*, les caractéristiques comme la robustesse aux mutations ou la température de repliement sont donc biaisées (cf. figure III.20) : en moyenne les séquences de la population sont plus robustes aux mutations et se replient plus efficacement.

L'ampleur du biais que nous venons de souligner dépend principalement de la taille du réseau neutre (cf. figures III.23 et III.33). Nous sommes donc tenté de penser que les réseaux neutres biologiques réels pourraient donner lieu à des phénomènes plus notables encore. Nos études de protéines tridimensionnelles appuient cette prédiction, mais il faut reconnaître que les résultats sont trop peu nombreux pour être catégorique, et des études supplémentaires sont requises pour confirmer cette théorie. Une approche possible serait de procéder à des dynamiques évolutives en alphabet complet comme l'ont fait Taverna et Goldstein [173].

III.3.3.2 Presqu'îles neutres et scénarios évolutifs

Un phénomène à plus large échelle que celui décrit dans le paragraphe précédent se déploie. Il est très possible que les réseaux neutres s'organisent en presqu'îles séparées les unes des autres par des régions pauvres en connexions. Ce schéma est conforme à l'organisation hiérarchique étudiée par Dokholyan et Shakhnovich [40]. La figure III.15 en donne une illustration dans le cas des protéines sur réseau et la figure III.35 dans le cas de Vav. L'état stationnaire peuple majoritairement la presqu'île neutre possédant la plus forte robustesse mutationnelle moyenne. Ce cas s'apparente au *survival of the flattest* de Wilke *et al.* [193]. Wilke a aussi décrit une évolution adaptative nommée *neutral staircase* [194] : l'évolution neutre procède par diffusion dans une presqu'île jusqu'à ce qu'elle trouve un accès à une presqu'île offrant une robustesse mutationnelle supérieure. La robustesse mutationnelle augmente donc par paliers successifs. Après un temps long, une presqu'île de robustesse mutationnelle maximale est trouvée : c'est notre état stationnaire.

On peut justifier mathématiquement le phénomène de dépeuplement des presqu'îles neutres. La matrice d'adjacence d'un réseau neutre composé de deux presqu'îles R_1 et R_2

peut s'écrire (au besoin en renumérotant ses nœuds) :

$$\mathcal{A} = \begin{pmatrix} \boxed{\mathcal{A}_1} & \sim 0 \\ \sim 0 & \boxed{\mathcal{A}_2} \end{pmatrix},$$

où $\mathcal{A}_1$ et $\mathcal{A}_2$ sont les matrice d'adjacence décrivant les presqu'îles neutres et où ~ 0 signifie que peu de connexions relient R_1 et R_2. Nous avons supposé aussi que R_2 est plus petite que R_1 : les tailles des presqu'îles, N_i, vérifient $N_2 < N_1$.

Supposons dans un premier temps que l'on calcule l'état stationnaire de chaque région séparément, comme si elles étaient complètement séparées. Appelons $p^{(1)}$ et $p^{(2)}$ les vecteurs fréquences de chaque presqu'île, de tailles respectives N_1 et N_2. Nommons $\langle n \rangle_s^{(1)}$ et $\langle n \rangle_s^{(2)}$ les robustesses mutationnelles associées. On a donc :

$$\mathcal{A}_1 \, p^{(1)} = \langle n \rangle_s^{(1)} \; p^{(1)},$$
$$\mathcal{A}_2 \, p^{(2)} = \langle n \rangle_s^{(2)} \; p^{(2)}.$$

Étant donné les relations existant entre les tailles N des réseaux neutres et les quantités $\langle n \rangle_u$ et $\phi(n)$ (cf. figures III.7 et III.33), il est raisonnable de supposer que $\langle n \rangle_s^{(2)} < \langle n \rangle_s^{(1)}$.

On prolonge, à présent, les vecteurs $p^{(1)}$ et $p^{(2)}$ de façon à les rendre de taille $N = N_1 + N_2$.

$$q^{(1)} = (\quad p^{(1)} \quad , \underbrace{0, \ldots, 0}_{N_2 \text{ fois}}),$$
$$q^{(2)} = (\underbrace{0, \ldots, 0}_{N_1 \text{ fois}}, \quad p^{(2)} \quad).$$

Ce sont deux vecteurs fréquences candidats de la matrice $\mathcal{A}$; les robustesses mutationnelles associées sont simplement $\langle n \rangle_s^{(1)}$ et $\langle n \rangle_s^{(2)}$. L'état stationnaire correspond à la valeur propre maximale[22]. Comme $\langle n \rangle_s^{(2)} < \langle n \rangle_s^{(1)}$, on en déduit que les fréquences sont décrites, à l'état stationnaire, par la valeur propre $\langle n \rangle_s^{(1)}$ et le vecteur propre $q^{(1)}$.

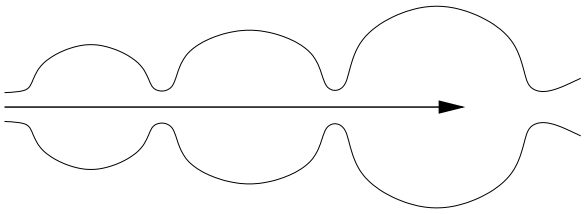
Compte tenu des remarques faites précédemment, on peut envisager différents scénarios d'évolution. Dans les schémas qui suivent, les presqu'îles sont représentées par des

22. On peut sans difficulté démontrer *ab absurdo* qu'il ne peut pas y avoir de valeur propre supérieure à $\langle n \rangle_s^{(1)}$.

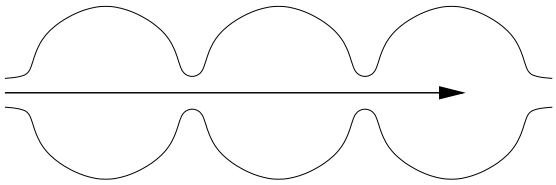
bulles dont la taille représente la robustesse mutationnelle.

1. Si l'évolution est neutre et la population assez grande,

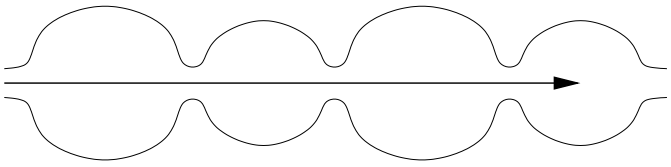
   (a) la population peut visiter des presqu'îles neutres par ordre de robustesse mutationnelle $\langle n \rangle$ croissante (*neutral staircase*) ;



   (b) la population peut visiter des presqu'îles neutres possédant des robustesses mutationnelles équivalentes.



2. L'évolution neutre d'une population infinie ne permet pas de passer d'un ensemble de séquences à un autre possédant une robustesse mutationnelle inférieure. Néanmoins, deux situations permettent que cela se produise. Premièrement, l'évolution peut être darwinienne, c'est-à-dire qu'elle peut faire intervenir une sélection *positive*. Il faut alors que l'ensemble de séquences de robustesse inférieure possède un avantage sélectif significatif. Le deuxième cas à envisager est une l'évolution neutre d'une population peu nombreuse. D'après notre étude sur la taille de la population, l'effet de dynamique de la population ne peut pas avoir lieu. Dans ce cas, la distribution de la population (après un temps long) obéit à une distribution uniforme sur le réseau neutre.



III.3.3.3 L'hétérozygotie à l'origine de l'adaptation

Le comportement adaptatif de l'évolution neutre est surprenante au premier abord. Elle découle de deux causes.

1. Le *fitness* w_i d'une séquence est donné par l'équation 27. Le *fitness* de toutes les séquences tend vers un quand le taux de mutation, μ, tend vers zéro. Dès que le taux de mutation est différent de zéro, une différence de *fitness* non nulle apparaît entre les séquences possédant un nombre de connexions différent.

2. Le nombre de connexions détermine le *fitness* mais n'est pas suffisant pour rendre compte des caractéristiques de l'état stationnaire. Dans le cas contraire la relation existant entre le nombre de connexions d'une séquence n_i et sa fréquence à l'état stationnaire ne serait pas bruitée (cf. figure III.19A). En réalité, une séquence peut posséder de nombreux voisins ; si ses voisins sont, eux, peu connectés, elle sera peu représentée à l'état stationnaire. De la même façon, si elle possède peu de voisins et si ses voisins sont très peuplés, la séquence sera appréciablement peuplée. La fréquence du voisinage d'une séquence a donc une influence sur celle de la séquence elle-même. Dans une certaine mesure le voisinage du voisinage, etc., a une influence aussi.

Un taux de mutation élevé agit comme un révélateur : il « revèle » par le nombre de mutants viables produits quels sont les génotypes fortement connectés ou plus généralement appartenant à une région densément connectée du réseau neutre. En outre, le taux de mutation augmente le polymorphisme génétique d'une population (cf. équation 16, p. 30), ce qui profite à l'adaptation.

1. La présence concomitante de nombreux génotypes augmente les chances que, au sein de la population, émerge un mutant très robuste.
2. Le caractère adaptatif ne peut se manifester que s'il existe au moins deux séquences différentes dans la population. L'évolution neutre favorise alors celle qui génère le plus souvent des mutants viables. Par la présence de nombreux génotypes, l'évolution est capable de « sentir » la topologie du réseau neutre. Pour la même raison, les presqu'îles neutres de robustesse mutationnelle inférieure sont éliminées par compétition entre les différentes presqu'îles (par compétition inter-groupe).

La condition exprimée par van Nimwegen *et al.* est évocatrice de la formule 16 de l'introduction liant l'hétérozygotie au taux de mutation neutre et à la taille de la population [184]. La condition de van Nimwegen *et al.*

$$M\mu \gg 1. \tag{51}$$

possède donc un lien au moins formel entre l'hétérozygotie et la transition d'un régime uniforme vers l'état stationnaire.

On peut arriver à l'équation 51 par une autre voie. L'équation 27 (p. 56) nous dit que le *fitness* de la séquence i est donné par

$$w_i = 1 + \mu\,(n_i - \langle n \rangle)/\ell,$$

avec les notations habituelles. Le *fitness* moyen vaut un (puisque la population totale reste constante). L'avantage sélectif de la séquence i vaut $\mu\,(n_i - \langle n\rangle)/\ell$. La différence de *fitness* typique entre deux séquences choisies au hasard dans le réseau neutre est donc

$$\Delta w = \mu\,\sigma_n/\ell, \tag{52}$$

où σ_n est l'écart type de la robustesse mutationnelle

La différence de *fitness* donnée par l'équation 52 ne peut infléchir l'évolution d'une population que si son importance surpasse celle de la dérive génétique. Or le jeu de la sélection et de la dérive génétique est exactement ce que s'est attaché à étudier Kimura. D'après l'équation 18 de l'introduction (p. 32), la sélection joue un rôle plus important que la dérive génétique dès lors que

$$|\mathrm{N}\,s| \gg 1\,,$$

avec les notations de l'introduction. Avec les notations de ce chapitre cela se transforme en

$$\mathrm{M}\,\Delta w = \mathrm{M}\,\mu\,\sigma_n/\ell \gg 1\,. \tag{53}$$

L'équation 53 suggère que l'influence de la population est liée à la topologie du réseau. Et donc le seuil de 500 avancé par van Nimwegen *et al.* n'aurait donc rien d'universel [184]. Dans notre cas, une valeur de 2,5 suffit quand l'évolution obéit au protocole 1. Si le mode de reproduction est inspiré du modèle de Wright-Fisher (protocole 4), une valeur de 100 est nécessaire (cf. tableau III.6, p. 127). Or le protocole 4 diminue le nombre de séquences présentes dans la population.

Il est intéressant de noter que le nombre moyen de séquences différentes présentes dans la population est similaire dans deux situations :

1. en l'absence du mode de reproduction de Wright-Fisher, $\mathrm{M} = 500$ et $\mu = 0{,}005$, on a $\langle \mathrm{N}_s\rangle_\mathrm{T} = 190$;
2. avec le mode de reproduction de Wright-Fisher, $\mathrm{M} = 10\,000$ et $\mu = 0{,}01$, on a $\langle \mathrm{N}_s\rangle_\mathrm{T} = 180$.

Dans ces deux situations différentes, la robustesse mutationnelle moyenne approche effectivement la valeur propre maximale du graphe : 7,92. Ce résultat suggère fortement un lien entre la diversité des séquences présentes dans une population et la possibilité que l'approximation de la population infinie soit valide.

Si notre théorie est correcte, le nombre de variants d'une protéine dans une population serait un indicateur permettant de savoir si l'évolution neutre peut développer son caractère adaptatif.

III.3.4 Qualité du modèle structural

Notre modèle structural tridimensionnel se distingue de ceux inventés par divers auteurs [40, 186] qui assignent des contacts entre résidus proches en se fondant sur les distances entre C_α ou C_β. Notre modèle est de ce point de vue identique à celui de Bastolla *et al.* :

> We found the best results using the following definition: two residues are defined to be in contact if any pair of their heavy atoms is closer than the threshold distance $R_c \leq 4.5$ Å. Contacts among residues separated by less than three positions along the sequence are not considered [5].

En outre, la greffe de la chaîne latérale est cohérente avec la méthode d'apprentissage qui permit de construire les matrices d'interaction [106]. Nous pensons donc que cette différence entre ces deux définitions ne modifie pas profondément la matrice de contact[23] et n'est donc pas décisive dans, par exemple, le degré de conservation d'une position[24]. On verra en annexe les profils obtenus avec un critère de distance entre les C_β et un alphabet à vingt classes, figure III.64*, p. 161. Bien plus encore que la greffe des chaînes latérales, l'étape la plus novatrice de notre approche est le recours à la préoptimisation, de principe similaire aux méthodes de Babajide *et al.* [1], qui assure d'adéquation entre les séquences et la structure native. Ce faisant, nous limitons les problèmes liés à notre description physique simplifiée que nous discutons ci-après.

Les modèles d'évolution reposent sur des réseaux neutres qui cristallisent la correspondance génotype-phénotype-*fitness*. Cette correspondance est elle-même fondée sur des « modèles structuraux minimalistes » : d'une part, le critère utilisé se fie uniquement à la stabilité de la conformation native, d'autre part, cette stabilité est estimée à partir de modèles de protéine extrêmement simplifiés. Ces simplifications réduisent le caractère corroborable de nos modèles par rapport aux théories plus satisfaisantes que nous avons à notre disposition concernant le repliement protéique (cf. Popper, *La Connaissance objective* [154, p. 62]). Notre partie « Modifiabilité et structures secondaires » (cf. p. 133) démontre notre souci de ne pas vouloir extraire plus de ces modèles que ce qu'ils ont à offrir.

Diverses études ont prouvé que les hypothèses sur l'énergie que nous utilisons ne sont pas suffisantes. Ainsi, Thomas et Dill, en recourant à des modèles de protéine sur réseau, se sont interrogés sur la pertinence des potentiels statistiques (*extracted energies*) comparés aux potentiels vrais (*true energies*) [177]. Ils concluent qu'il est généralement difficile d'ob-

23. Cependant, l'utilisation des carbones β nécessite l'introduction d'une distance limite R_c plus grande, 8,5 Å (référence [186]) ou 7,5 Å (référence [40]), par exemple.

24. Bien que le détail précis des séquences soit vraisemblablement modifié.

tenir une information pertinente à partir de matrices d'interactions apprises à partir de bases de données et, résultat important touchant à nos hypothèses, que ces potentiels statistiques sont peu performants à identifier les conformations natives de séquences choisies au hasard. Vendruscolo et Domany ont montré que le calcul d'une énergie par somme d'interactions entre paires d'acides aminés en contact ne peuvent pas satisfaire la condition que l'on attendrait d'une « bonne fonction d'énergie » : que l'énergie d'une séquence s dans la conformation native c_0 soit inférieure à celle de cette même séquence dans toute autre conformation c [186] :

$$\mathrm{E}(s, c_0) < \mathrm{E}(s, c)\,. \tag{54}$$

Leur premier résultat est particulièrement définitif : il est aisé de mettre en échec une matrice d'énergie de type HP. Ce manque de robustesse des modèles ne considérant qu'un alphabet réduit est prégnant dans l'étude de Buchler et Goldstein : ils démontrèrent que les modifiabilités calculées à l'aide de matrices à deux classes sont très différentes de celles calculées à l'aide d'alphabets complets [25]. Leurs résultats suggèrent néanmoins que des alphabets de taille intermédiaire ($m = 4$ classes) reconstituent une partie importante de l'information présente dans les interactions à vingt classes. Par ailleurs, il semble que leur résultat négatif tienne à une erreur dans le choix de la matrice de Miyazawa et Jernigan [135, 136] : les modifiabilités calculées avec un alphabet à vingt classes et celles calculées avec un alphabet à deux classes sont en excellent accord lorsque la matrice à vingt classes utilisée correspond à la moitié supérieure du tableau V de la référence [135] (cf. la référence [111]). Or cette même matrice est celle qui permit de dériver la matrice LHTW. Buchler et Goldstein avaient, eux, recouru à la matrice VI de la référence [135]. On se reportera aux figures III.61* et III.62* (p. 158, 159) des annexes pour des résultats plus détaillés qui montrent bien que l'abstraction HP n'est pas aussi peu pertinente que ne voulurent le laisser penser Buchler et Goldstein. Néanmoins, la question du nombre minimal de classes nécessaires pour obtenir des propriétés typiques des protéines a fait l'objet d'études expérimentales et théoriques qui indiquent que cinq est le nombre d'acides aminés minimal requis [31, 51, 156, 188]. En particulier, Riddle *et al.* démontrent que cinq types suffisent à concevoir un repliement de type domaine SH3 mais que trois est insuffisant [156]. La réduction à dix types d'acides aminés a été proposée dans la reconnaissance de repliement par homologie [115, 141]. La méthode d'identification des repliements natifs de séquences naturelles mise au point par Launay *et al.* indique que six classes suffisent à atteindre un succès dans l'identification des repliements natifs comparable à celui obtenu avec vingt classes [106].

L'ensemble de ces résultats amène à considérer avec suspicion le calcul d'une énergie par somme d'interactions entre résidus en contact et la réduction des alphabets. Ce-

pendant, le succès de la réduction opérée par Wang et Wang à partir de la matrice d'interaction de Miyazawa et Jernigan plaide en faveur de la pertinence de ces matrices et souligne l'importance d'une classification efficace [188]. Il est à noter que les résultats de Buchler et Goldstein, ceux de Vendruscolo et Domany, ni, enfin, ceux de Thomas et Dill [25, 177, 186] ne prennent en considération le fait que les séquences aussi bien que les structures sont modelées par l'évolution et pourraient donc mieux se prêter aux méthodes d'analyse utilisant des modèles d'énergie approchés que ce que ne prédisent ces études exhaustives [5, 101, 157]. Par ailleurs, dans leur article, Vendruscolo et Domany cherchent à mettre en défaut la relation 54 en faisant évoluer d'une part l'ensemble des conformations $\{c\}$ et les paramètres d'interaction de la fonction $\mathrm{E}(\cdot,\cdot)$ dans des directions « contraires ». Les conformations $\{c\}$ évoluent dans le sens le plus favorable à invalider la relation 54, tandis que les paramètres de $\mathrm{E}(\cdot,\cdot)$ évoluent dans un sens favorable au maintien de la vérité de la relation 54. Dans notre étude, la préoptimisation modifie, elle, la séquence *s*. Nous échappons donc aux conclusions de Vendruscolo et Domany dont l'objet était de répondre à la question (et ils y répondirent par l'affirmative) : *étant donné une séquence s fixée*, est-il possible de trouver une conformation *c* inversant la relation 54 ? Pour invalider définitivement notre démarche, le prédicat qu'il faudrait s'attaquer à démonter serait le suivant : pour une conformation c_0 fixée, est-il possible qu'il n'existe aucune séquence *s* telle qu'il n'existe aucune fonction d'énergie par interaction de paires $\mathrm{E}(\cdot,\cdot)$, vérifiant la relation 54 pour toute autre conformation *c* que c_0 ?

Enfin, même en se bornant à un alphabet HP, le détail des séquences en lui-même n'est pas ce qui intéresse : l'évolution telle que nous la modélisons ne dépend pas des séquences proprement dites mais uniquement de la topologie des réseaux neutres. En d'autres termes, plus encore que la pertinence des séquences générées par nos méthodes, ce sont les *relations* qui lient les séquences viables, les mutations neutres ou encore les arêtes du réseau neutre, qui sont au centre de nos préoccupations. C'est le message que porte la figure II.14, p. 83. Notre préoccupation devient donc : les mutations qui lient les séquences que nous générons obéissent-elles à des patrons vraisemblables ? Nous avons présenté quelques éléments de réponse dans la section « Distribution de la robustesse mutationnelle » (p. 130) en comparant les fractions de mutations neutres expérimentales à celles obtenues dans le cadre de notre modèle. Même les protéines sur réseau présentent des caractéristiques des protéines dans leurs motifs de mutations : certains résidus, principalement ceux enfouis, sont plus conservés que ceux en surface ; nous avons également défendu que dans les réseaux neutres de tailles petites et modestes, les mutations compensatrices pourraient constituer l'essentiel des mutations neutres. La conclusion du chapitre « Évolution des protéines dimériques » (cf. p. 207) apportera des compléments concernant la conservation des résidus pour les modèles de protéine tridimensionnelle.

Certaines de nos études ont pu être conduites pour des alphabets allant de $m = 2$ à $m = 20$ classes (cf. figures III.26, III.28, III.29, III.30 et III.32). Dès lors qu'un réseau neutre a dû être reconstruit, nous nous sommes limité à $m = 2$ ou $m = 3$ classes. La taille des réseaux neutres variant comme une exponentielle de $\mathrm{L}(m-1)$, il semble donc possible d'étudier, à la limite, des réseaux neutres construits avec quatre classes pour le trp-cage, puisqu'en vertu de cette loi, les réseaux neutres auraient une taille de l'ordre de $100 \cdot 10^6$ séquences ce que nous pouvons techniquement traiter[25]. Un alphabet à trois classes pourrait à la limite être traité pour les protéines sur réseau. Supposant que le même critère de repliement que celui employé pour les alphabets HP soit appliqué dans le cas d'un alphabet à trois classes, nous choisissons la matrice d'énergie de la page 80. Avec ce choix, une évaluation par Monte Carlo montre que 10^6 séquences sur $1{,}36 \cdot 10^6$ se replient, soit une fraction de 73 %[26]. Comme ce qui a été trouvé pour la plupart des matrices HP, les conformations 700 et 800 sont les moins modifiables avec 140 et 156 séquences identifiées au cours de la simulation de Monte Carlo. La modifiabilité de ces conformations s'élèvent donc probablement à

$$3^{25}/1\,360\,718 \times 148 \approx 90 \cdot 10^6 .$$

Quant aux plus gros réseaux neutres, ils auraient une taille approximative de $1{,}8 \cdot 10^9$ séquences. En d'autres termes, avec des ressources de calcul accessibles actuellement, il est possible de produire des réseaux neutres dans des alphabets à trois classes, du moins pour les plus petits réseaux neutres (mais ce ne sont pas forcément les réseaux les plus intéressants). Pour cela, il suffit de partir d'une séquence s_0 qui se replie en la conformation d'intérêt et d'explorer son voisinage par mutagenèse systématique au moyen d'une pile :

$r := \{s_0\}$ # *Le réseau neutre*
$stack := \{s_0\}$
while $s := \mathrm{pop}(stack)$ **do**
  for toute séquence s^* mutée à partir de s **do** # $(3-1) \times 25 = 50$ *mutations*
    if $s^* \notin r$ et s^* se replie **then**
      $\mathrm{push}(stack, s^*)$ # *On étudiera le voisinage de* s^* *plus tard*
      $\mathrm{push}(r, s^*)$ # *Marquer* s^* *comme déjà rencontrée*
    end if

25. Il convient de remarquer que s'il a une taille théorique abordable, le réseau neutre à quatre classes requerrait un temps de calcul sans doute très grand si ce n'est rédhibitoire. Si $500 \cdot 10^6$ pas de trajectoire sont nécessaires à la découverte de $500 \cdot 10^3$ profils et si l'on suppose une loi linéaire, on peut estimer à $500 \cdot 10^9$ le nombre de pas de trajectoire nécessaires à l'exploration d'un réseau neutre à quatre classes.

26. Ce qui rend le critère de non-dégénérescence de l'état fondamental en réalité inapproprié puisque, expérimentalement, la majorité des séquences ne se replient pas. Mais ce calcul fournit des ordres de grandeur.

end for
end while

La prise en compte de cinq classes semble, quant à elle, pour l'instant hors d'atteinte. De façon similaire, la reconstruction de réseaux neutres de séquences se repliant en l'un des deux domaines SH3 étudiés ici n'est possible qu'avec des alphabets à deux classes. Des modèles d'énergie plus sophistiqués intégrant un positionnement précis des chaînes latérales et un solvant implicite est en cours de développement (cf. « Modèle tous atomes », p. 160).

Plusieurs classifications peuvent être obtenues selon la méthode utilisée pour les générer. Nous pensons que le choix n'est pas capital tant qu'une méthode de classification raisonnable est mise en œuvre et qu'une matrice d'interaction spécifique est produite. Cette remarque est également celle de Fan et Wang qui concluent

> It should be noted that our results are not sensitive to the precise grouping of amino acids, provided that the grouping method is reasonable. In fact, the difference is subtle for grouping of amino acids using different criteria, and the subtle difference has a minor effect on our results [51].

Enfin, la méthode de construction de fausses structures employée par Vendruscolo et Domany pourrait être adaptée à nos modèles [186].

III.4 Annexes

III.4.1 Statistiques

Les statistiques des tables III.1 et III.2 sont reproduites pour les autres matrices d'énergie dans les tables III.7 et III.8.

Matrice d'énergie	*Nombre de séquences se repliant*	*Nombre de réseaux neutres*	*Plus grand réseau neutre*
GLO	5 170 505	144	38 391 (509)
Ising	8 331 268	44 811	25 476 (61)
TS_1	7 008 342	3 475	42 143 (908)
TS_2	14 237 267	3 665	50 819 (241)

Tab. III.7 : Nombre de séquences se repliant et de réseaux neutres reconstruits à l'aide des diverses matrices d'énergie. La dernière colonne indique la taille du plus grand réseau neutre et la conformation qui lui correspond.

GLO		Ising		TS_1		TS_2	
753-509	(38 391)	169-125	(25 558)	976-908	(42 143)	787-241	(50 888)
903-786	(37 343)	61-103	(25 532)	903-786	(33 192)	903-786	(50 165)
976-908	(31 171)	102	(24 964)	951-658	(32 787)	901	(43 753)
787-241	(28 548)	964-52	(24 932)	774-580	(32 036)	905-60	(41 716)
906-902	(27 597)	246-106	(24 600)	753-509	(30 959)	897-66	(40 197)
774-580	(26 730)	62	(24 550)	906-902	(29 275)	756-100	(40 173)
635-486	(26 299)	275-105	(24 360)	596-101	(27 013)	245	(38 358)
634-367	(23 915)	116-1014	(24 260)	635-486	(24 562)	906-902	(38 251)
754-390	(23 582)	99-766	(24 242)	898-770	(24 068)	898-770	(37 966)

Tab. III.8 : Liste des conformations les plus modifiables pour les matrices d'énergie GLO, Ising, TS_1, TS_2. Les conformations inverses vont par paire, séparées par un tiret, sauf quand la conformation est symétrique. Les modifiabilités sont indiquées entre parenthèses.

La figure III.2 est reproduite pour les autres matrices d'énergie dans la figure III.45.

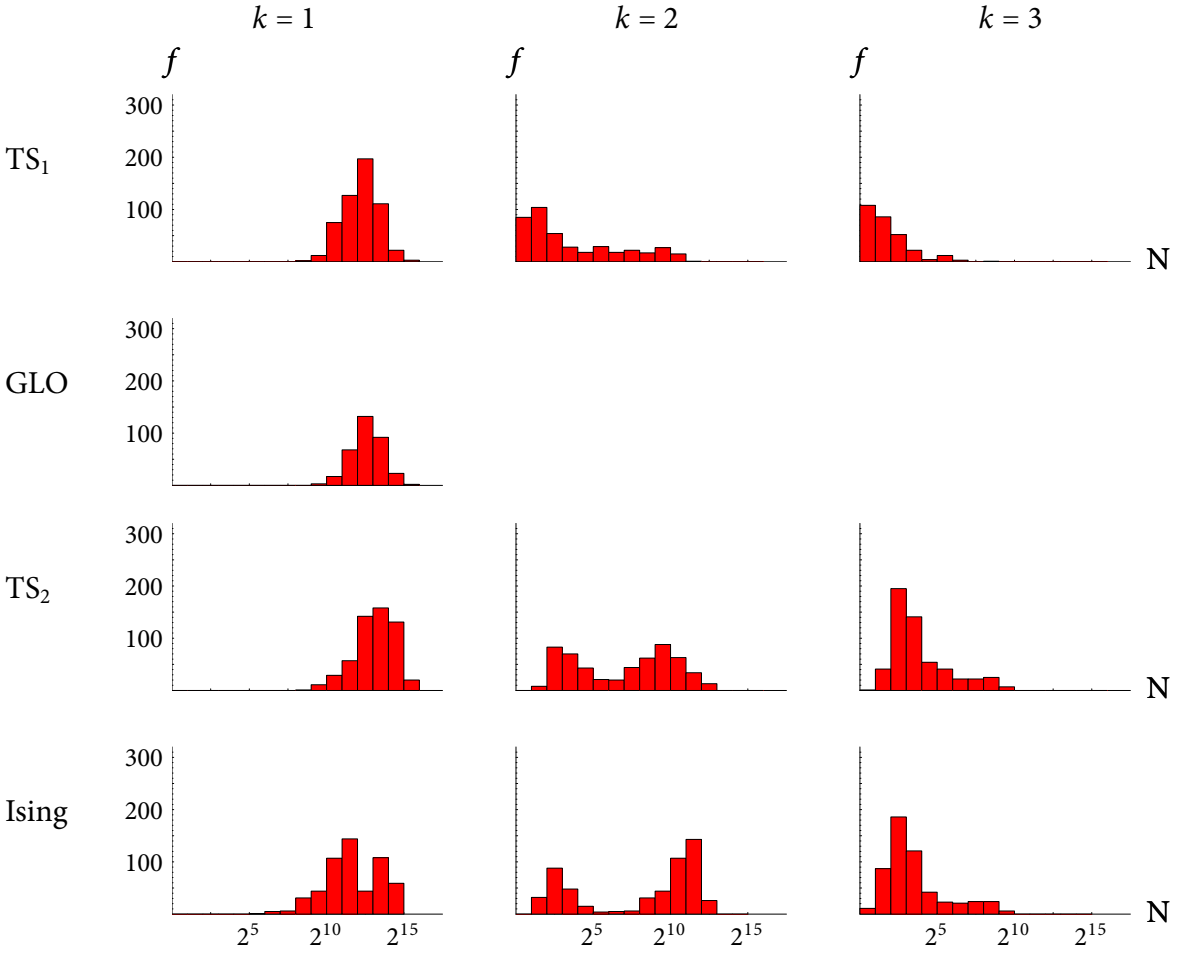


Fig. III.45 : Histogramme des tailles N des plus gros réseaux neutres de chaque conformation ($k = 1$), histogramme des tailles des seconds plus gros réseaux neutres de chaque conformation ($k = 2$), etc. (Note : $2^5 = 32$, $2^{10} = 1\,024$, and $2^{15} = 32\,768$.)

On remarquera que TS_1 donne les mêmes résultats que les trois matrices étudiées précédemment au contraire de TS_2 et Ising. Avec la matrice GLO, les conformations ne possèdent qu'un seul réseau neutre au maximum. En fait, d'après la table III.7, les séquences ne peuvent se replier que dans 144 des 1 081 conformations.

III.4.2 Propriétés moyennes des réseaux neutres

Nous complétons les figures III.7 et III.8 pour les autres matrices d'énergie. Les conclusions tirées pour les matrices d'énergie typiques des protéines restent essentiellement les mêmes.

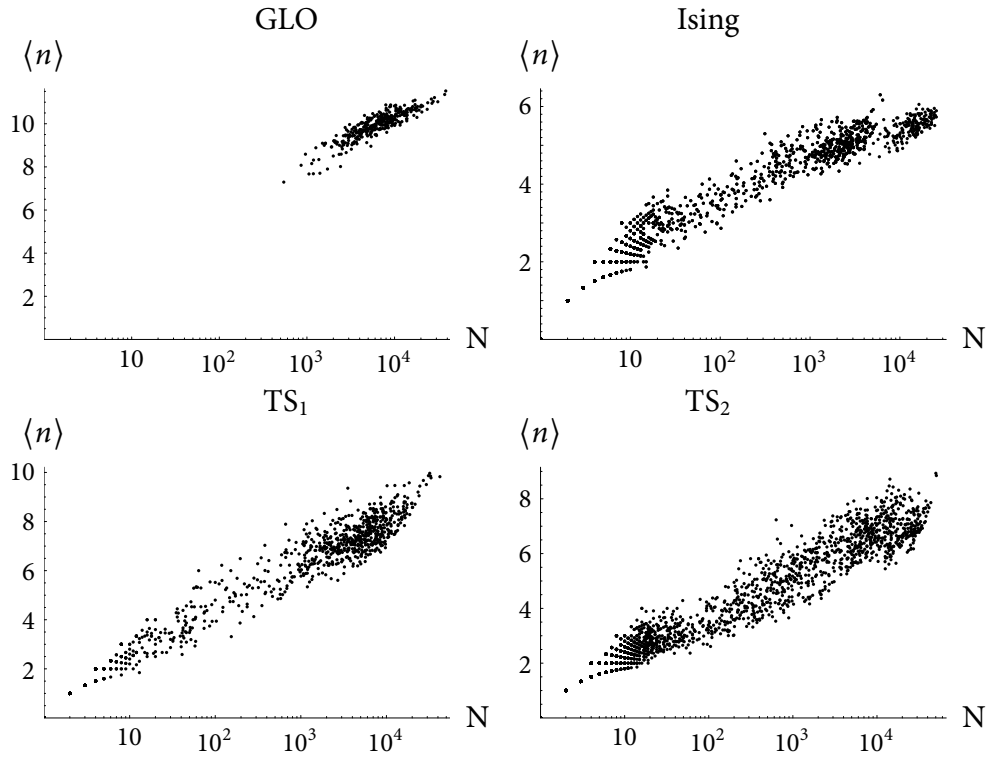


FIG. III.46 : Robustesse moyenne $\langle n \rangle$ en fonction de la taille N des réseaux neutres calculés avec les autres matrices d'énergie.

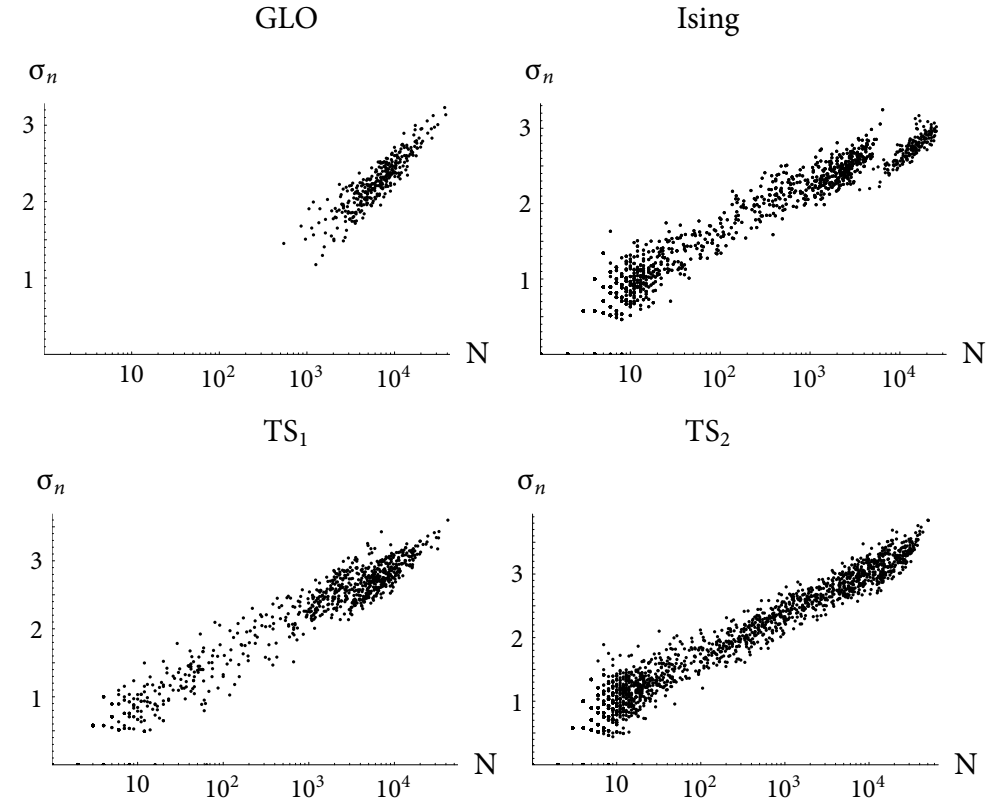


FIG. III.47 : Écart type σ_n en fonction de la taille N des réseaux neutres calculés avec les autres matrices d'énergie.

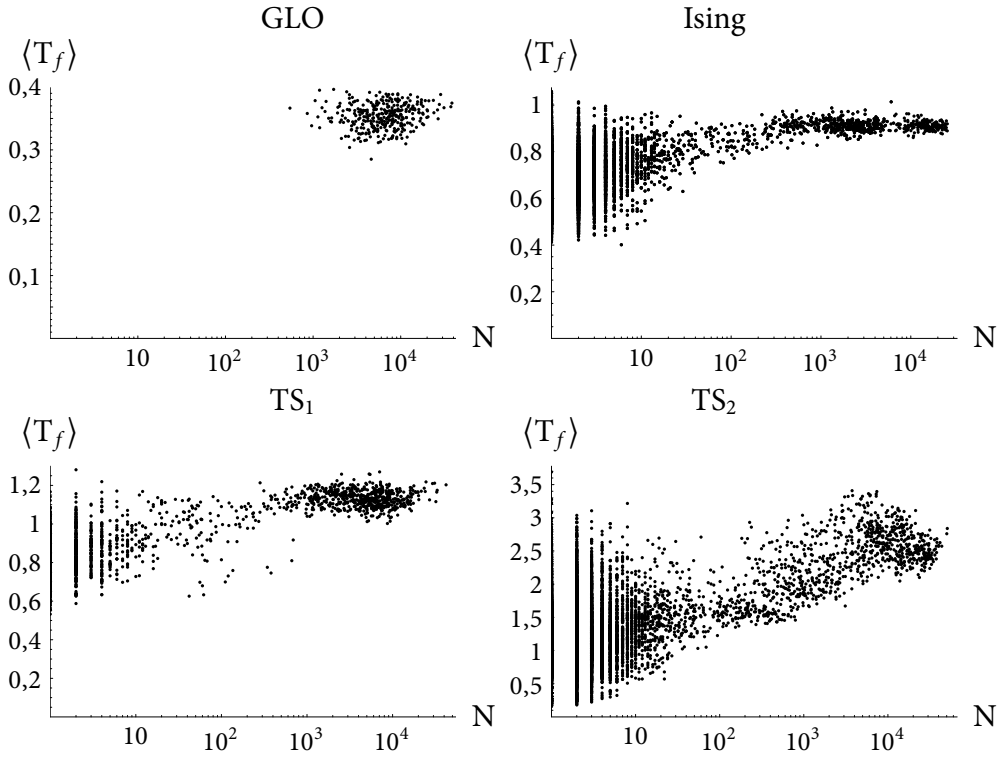


Fig. III.48 : Température de repliement moyenne $\langle T_f \rangle$ en fonction de la taille N des réseaux neutres calculés avec les autres matrices d'énergie.

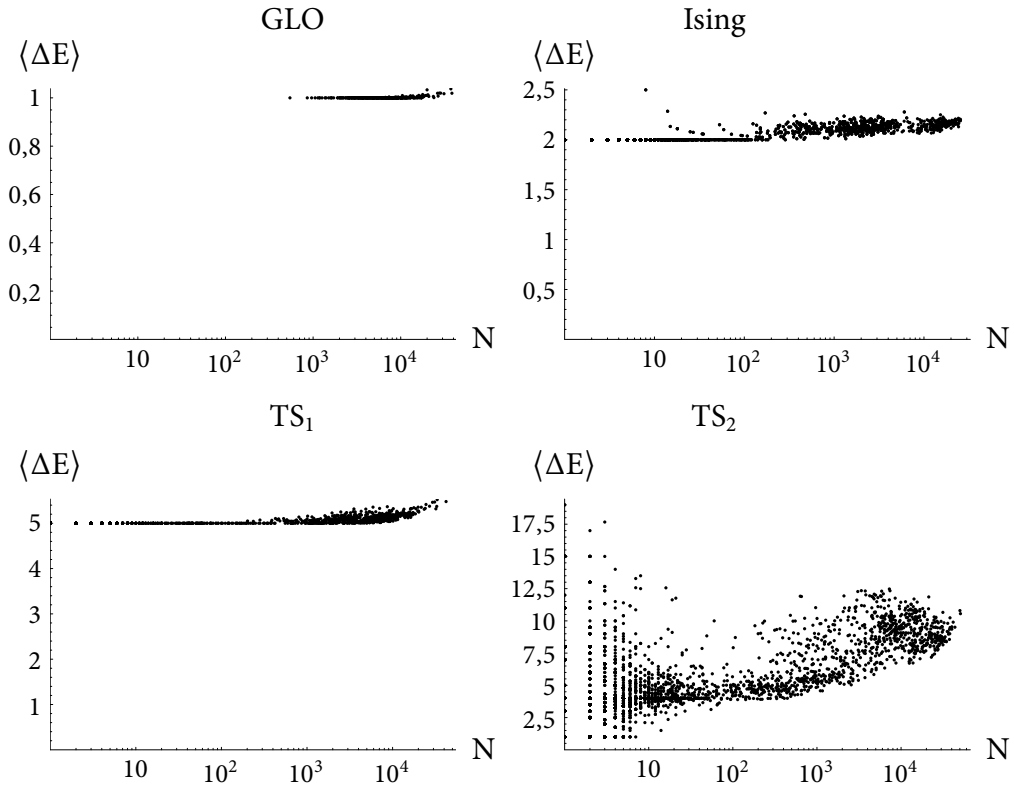


Fig. III.49 : Différence d'énergie moyenne $\langle \Delta E \rangle$ en fonction de la taille N des réseaux neutres calculés avec les autres matrices d'énergie.

III.4.3 Distribution de la robustesse mutationnelle

III.4.3.1 Distribution universelle

La forme caractéristique de la distribution de la robustesse mutationnelle est retrouvée pour les matrices d'énergie non typiques des protéines.

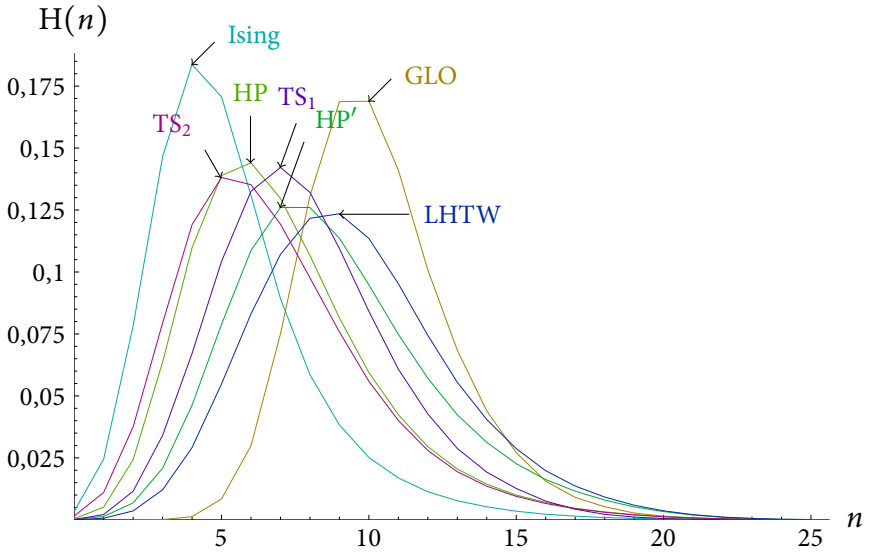


Fig. III.50 : Distribution de la robustesse mutationnelle de l'ensemble des séquences se repliant pour les matrices d'énergie présentées dans « Modèle et méthodes ».

Nous complétons les distributions de la figure III.6 par les autres matrices d'énergie (cf. figure III.51).

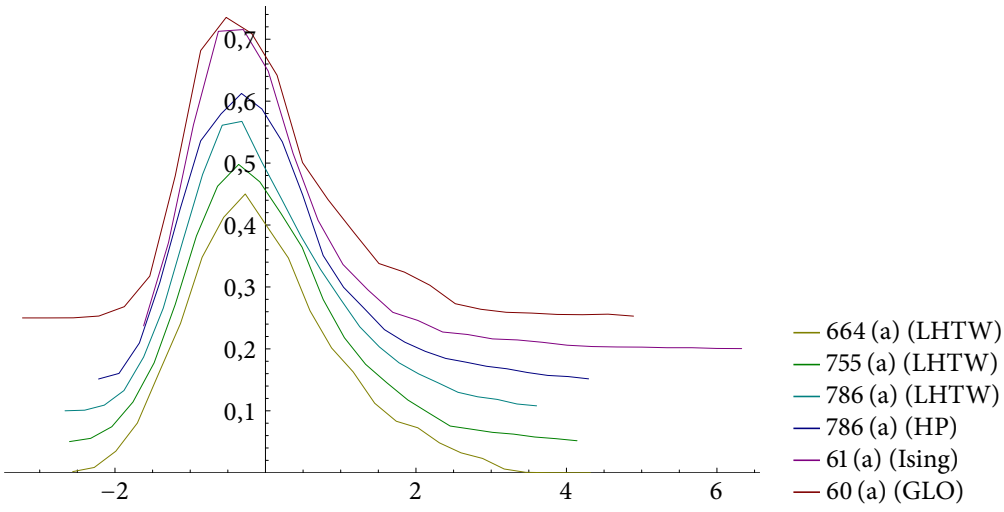


Fig. III.51 : La distribution de la robustesse mutationnelle pour diverses matrices d'énergie et divers réseaux neutres. La matrice d'énergie LHTW est représentée plusieurs fois pour illustrer l'influence de la taille du réseau. Les distributions sont normalisées et superposées.

III.4.3.2 Origine de la forme des distributions

La forme des distributions de la robustesse mutationnelle que nous venons de voir semble universelle. Dans le modèle des protéines sur réseau, elle est retrouvée avec toutes les matrices d'énergie. Elle émerge également du modèle tridimensionnel. Deux des traits caractérisques de ces distributions

1. le faible nombre de séquences possédant peu de voisins et
2. le faible nombre de séquences possédant de nombreux voisins

peuvent essentiellement s'interpréter par un modèle d'énergie aléatoire. Soit s une séquence se repliant en une conformation c_0. On numérote les conformations différentes de c_0 par ordre d'énergie croissante :

$$\mathrm{E}(s, c_0) < \mathrm{E}(s, c_1) \leq \mathrm{E}(s, c_2) \leq \cdots .$$

Lorsque l'on mute la séquence s en s^*, la séquence mutante forme un lien avec s si et seulement si $\mathrm{E}(s^*, c_0) < \mathrm{E}(s, c_i)$. Au contraire, s^* ne forme pas de connexion si l'une des structures c_i ($i > 0$) procure une énergie au plus égale à $\mathrm{E}(s^*, c_0)$. Il existe évidemment une corrélation entre les énergies $\mathrm{E}(s, c_i)$ et $\mathrm{E}(s^*, c_i)$. On peut donc se focaliser sur les premiers niveaux d'énergie excités : les conformations c_i avec i élevé n'ont aucune chance d'avoir un quelconque rôle dans la détermination de la connexion entre s et s^*. Dans un modèle d'énergie aléatoire, le nombre de connexions contractées par s est liée au nombre de fois que la différence d'énergie $\Delta\mathrm{E} = \mathrm{E}(s, c_1) - \mathrm{E}(s, c_0)$ peut tolérer une perturbation aléatoire des niveaux d'énergie[27]. L'implémentation de ce modèle avec la matrice d'énergie LHTW donne la distribution présentée dans la figure III.52 et une longueur de chaîne L = 25.

Les distributions de la figure III.52 possèdent la même allure. Cependant, la distribution résultant du modèle d'énergie aléatoire a tendance à surestimer la fréquence des séquences possédant plus de douze connexions. Une partie au moins de l'explication provient de l'utilisation aléatoire des contacts dans ce modèle alors que les 132 contacts ne sont pas utilisés uniformément dans les conformations de protéine sur réseau (cf. figure III.53).

Pour une comparaison avec des modèles utilisant des alphabets plus complets, cf. figure III.62, p. 159.

27. Dans la formation d'une conformation, 132 contacts entre résidus sont possibles : le résidu à la position i ne peut interagir avec un résidu à la position j que si $|i - j|$ est impair et supérieur à trois. Seize contacts aléatoires sont attribués pour chaque « conformation », que nous nommerons « conformation aléatoire ». Nous créons 1 000 conformations aléatoires. Cent trente deux interactions sont choisies aléatoirement (LHTW) et constituent la « matrice d'interaction aléatoire ». Nous obtenons un spectre d'énergie. La conformation native est celle donnant l'énergie la plus faible. Pour chaque position i, nous étudions les perturbations de ce spectre. Nous modifions aléatoirement tous les contacts faisant intervenir un contact entre le résidu numéro i et tout autre résidu. La robustesse mutationnelle est le nombre de positions i dont la perturbation ne modifie pas la conformation native.

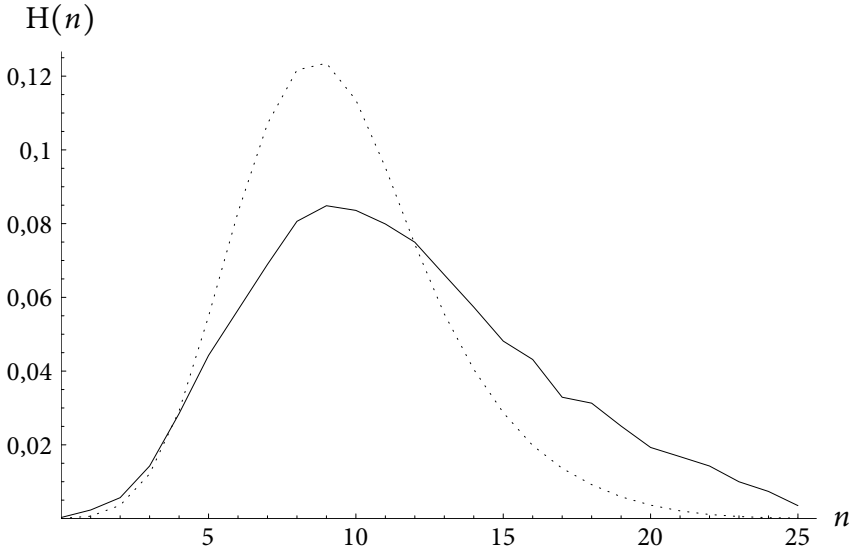


FIG. III.52 : Distribution de la robustesse mutationnelle dans le modèle d'énergie aléatoire (ligne continue) et expérimentale (LHTW).

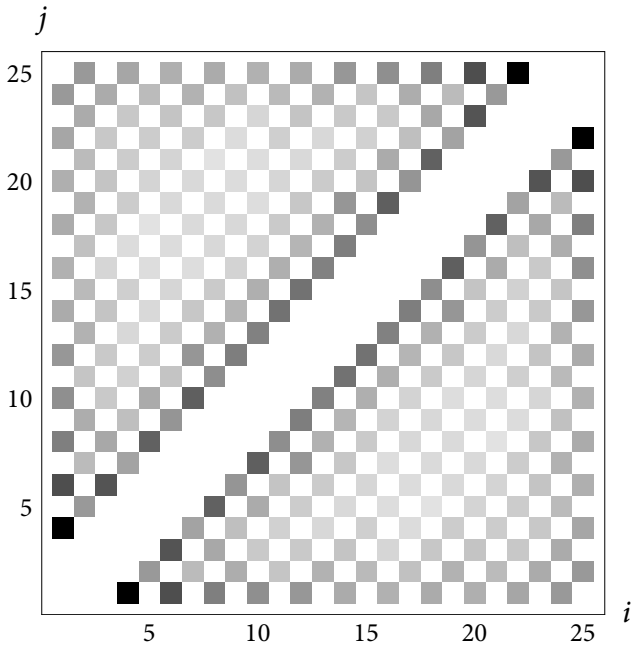


FIG. III.53 : Utilisation non uniforme des contacts.

III.4.4 Modifiabilité

Le recouvrement structural Q mesure la similitude entre deux conformations. Si une conformation est dissemblable de toute autre, les états dénaturés seront moins « compétitifs » et donneront moins de dégénérescence. Ce principe est appelé « designing out ». Le recouvrement structural Q détermine essentiellement la modifiabilité d'une conformation (cf. figure III.54).

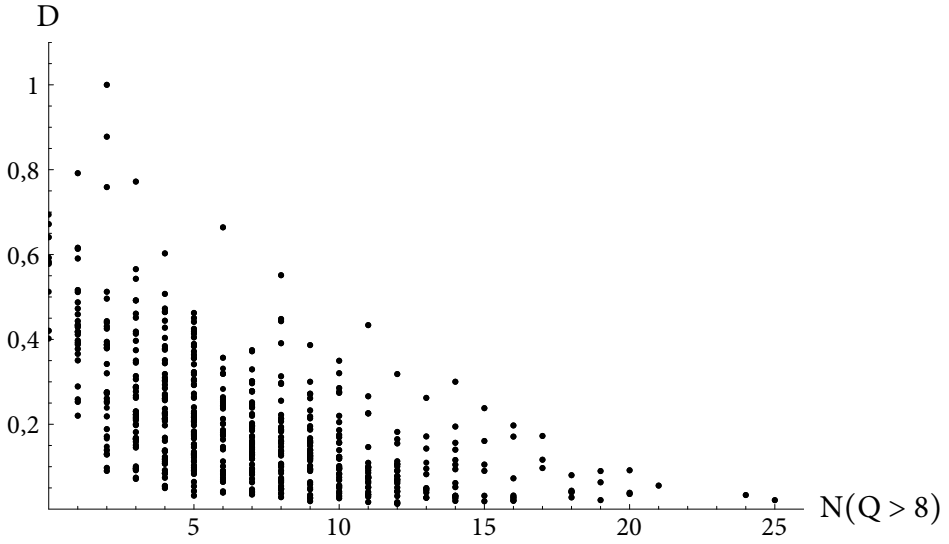


Fig. III.54 : Le recouvrement structural Q détermine en partie la modifiabilité d'une structure. Pour chaque structure *c*, on représente en abscisse le nombre de structures ayant un recouvrement d'au moins 9 avec *c* ; en ordonnée, la modifiabilité obtenue en sommant les modifiabilités observées pour les différentes matrices d'énergie (unités arbitraires). Le coefficient de corrélation est −0,568. Les coefficients de corrélation découlant de l'utilisation séparée des matrices d'énergie sont moins élevés en valeur absolue.

III.4.5 Réseaux neutres de la matrice Ising

La conformation la plus modifiable obtenue avec la matrice Ising est présentée dans la figure III.55

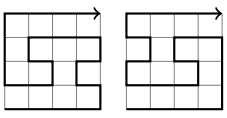


Fig. III.55 : Les conformations natives des plus grands réseaux neutres en utilisant la matrice Ising : 61 et 103, inverses l'une de l'autre.

Les réseaux neutres formés avec la matrice d'énergie Ising peuvent former un double *superfunnel*. Avec cette matrice, deux séquences « complémentaires » (par exemple, HPP et PHH) repliées dans une même conformation possèdent la même énergie. Par symétrie deux *superfunnel* peuvent émerger et si la connectivité est suffisante, ces deux *superfunnel* peuvent fusionner en un seul réseau neutre (cf. figure III.56).

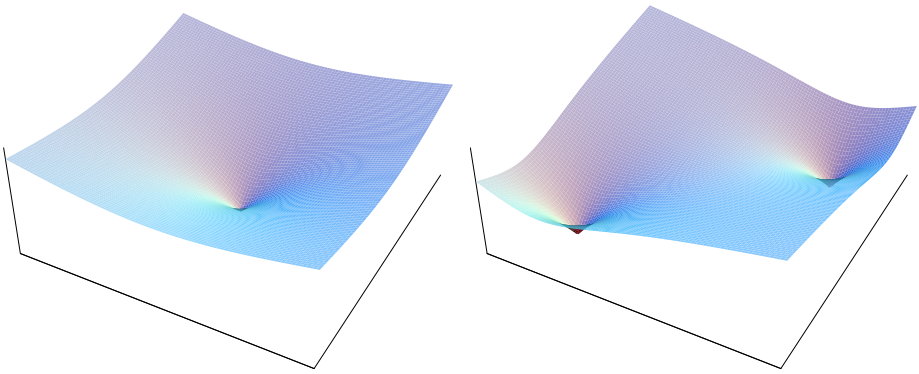


Fig. III.56 : Représentation schématique des *superfunnel* dans l'espace des séquences pour des potentiels inspirés des protéines (LHTW, HP, etc.) et la matrice d'énergie Ising. Deux *superfunnel* symétriques peuvent émerger avec la matrice Ising et, si la connectivité est suffisante, être connectés par des chemins neutres traversant l'espace des séquences.

L'analyse, à présent classique, des propriétés à distance l d'une des deux séquences prototypes du réseau neutre 61 (a) obtenu avec la matrice Ising est réalisée et les résultats sont montrés dans la figure III.57.

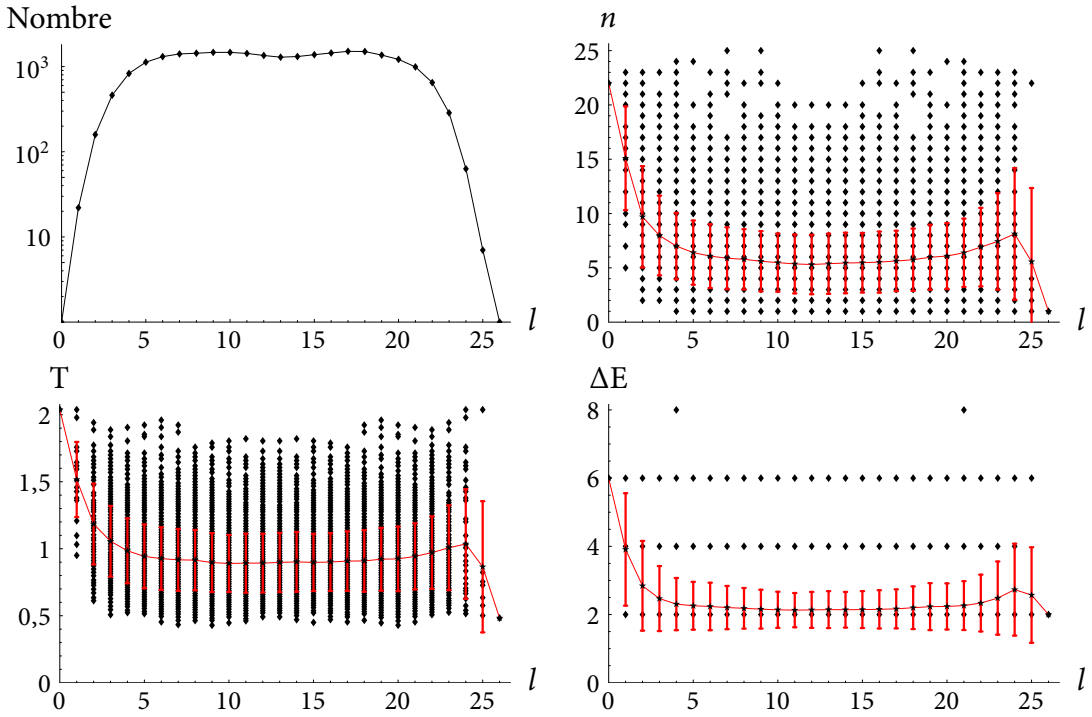


FIG. III.57 : Propriétés du réseau neutre 61 (a) obtenu avec la matrice Ising (N = 25 476), en fonction de l'éloignement l à la séquence prototype.

En dépit de son caractère non typique des protéines et de la forme singulière de son *superfunnel*, l'amélioration de la robustesse mutationnelle et de la température de repliement, $\phi(n)$ et $\phi(T_f)$, dépendent de la taille du réseau neutre selon la même loi que celle observée pour LHTW (cf. figure III.58).

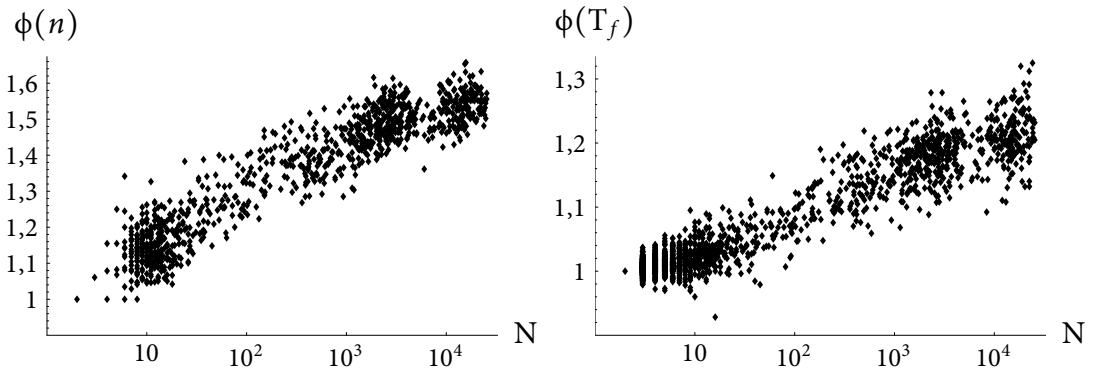


Fig. III.58 : Amélioration de la robustesse mutationnelle moyenne (à gauche) et de la température de repliement moyenne (à droite) à l'état stationnaire pour la matrice d'énergie Ising (en bas).

La distribution de population se scinde en deux pour se diriger vers le *superfunnel* le plus proche (cf. figure III.59).

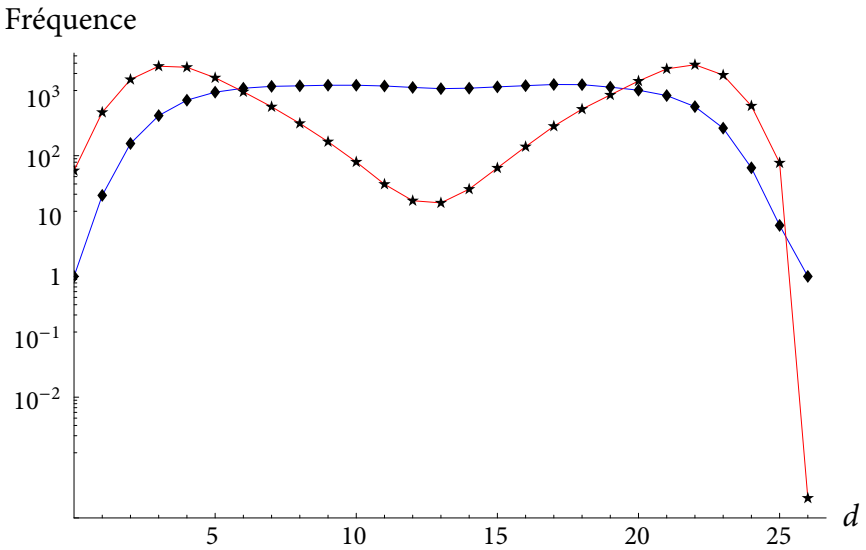


Fig. III.59 : Distribution de la population autour de la séquence prototype sous une distribution uniforme (en bleu) et à l'état stationnaire (en rouge) pour le réseau neutre réseau neutre 61 (a) construit avec la matrice d'énergie Ising.

III.4.6 Résumé des propriétés des réseaux neutres

Nous évoquons de nombreuses conformations et de nombreux réseaux neutres dans cet ouvrage. Cette section donne les propriétés de la plupart des réseaux neutres et les conformations associées (cf. table III.9 et figure III.60).

Code	D	$\langle n \rangle_u$	$\langle n \rangle_s$	$\phi(n)$	$\langle T_f \rangle_u$	$\langle T_f \rangle_s$	$\phi(T_f)$	ΔE
5 (a)	10 079	8,61	11,21	1,30	0,31	0,40	1,31	5,09
33 (a)	3 032	7,22	9,17	1,27	0,22	0,26	1,18	3,03
37 (a)	4 295	7,27	9,32	1,28	0,27	0,35	1,28	3,01
60 (a)	40 036	10,24	13,51	1,32	0,34	0,45	1,32	6,69
77 (b)	3 074	7,67	9,73	1,27	0,28	0,36	1,28	3,04
134 (a)	3 849	6,37	8,91	1,40	0,20	0,25	1,24	3,11
136 (a)	9 809	8,96	11,79	1,32	0,30	0,38	1,29	3,61
155 (a)	2 534	6,57	8,76	1,33	0,23	0,28	1,21	3,12
179 (a)	3 995	7,14	9,49	1,33	0,28	0,37	1,32	3,02
198 (a)	6 884	8,07	10,70	1,33	0,31	0,40	1,30	3,04
238 (e)	3 998	8,12	10,39	1,28	0,29	0,36	1,25	3,52
249 (a)	5 071	7,30	9,82	1,35	0,23	0,28	1,20	3,20
400 (a)	3 133	7,94	10,05	1,27	0,30	0,37	1,23	3,01
452 (a)	4 115	7,49	9,62	1,28	0,24	0,26	1,11	3,13
528 (a)	6 254	7,88	10,25	1,30	0,28	0,37	1,32	3,02
556 (a)	3 207	6,78	8,75	1,29	0,24	0,32	1,35	3,17
627 (b)	5 028	7,67	9,64	1,26	0,20	0,22	1,10	3,13
664 (a)	25 367	9,94	12,77	1,28	0,32	0,41	1,29	5,95
686 (g)	3 072	7,93	10,19	1,28	0,30	0,37	1,24	3,59
687 (a)	4 127	8,58	10,71	1,25	0,30	0,37	1,23	3,70
710 (a)	3 520	9,04	10,46	1,16	0,33	0,40	1,20	5,26
755 (a)	35 970	10,27	13,30	1,30	0,33	0,44	1,31	6,31
786 (a)	67 614	11,20	14,37	1,28	0,36	0,47	1,33	7,18
854 (a)	9 798	9,19	11,48	1,25	0,33	0,40	1,22	5,16
907 (a)	4 835	6,38	9,09	1,42	0,23	0,30	1,30	3,23
1057 (a)	3 061	8,94	10,31	1,15	0,33	0,40	1,20	5,87
61 (a)*	25 476	5,89	9,26	1,57	0,92	1,13	1,23	3,02

Tab. III.9 : Résumé des propriétés de quelques réseaux neutres mentionnés dans cet ouvrage. À l'exception de celui 61 (a) marqué d'une astérisque, construit à l'aide de la matrice Ising, les réseaux neutres présentés dans la table ont été obtenus avec la matrice LHTW.

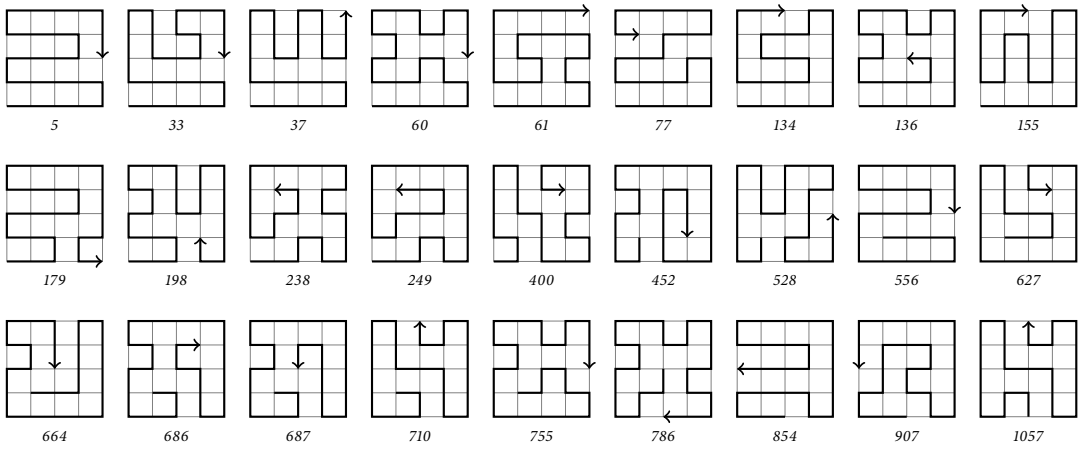


Fig. III.60 : Quelques conformations mentionnées dans cet ouvrage.

III.4.7 Alphabets plus complets

Nous avons évalué par un échantillonage aléatoire la modifiabilité des 1081 conformations sur réseau décrites dans cet ouvrage. Nous avons envisagé quatre critères de repliement : la séquence s se replie en la conformation c_0 si et seulement si

1. la conformation c_0 est celle d'énergie minimale ($\mathrm{E}(s, c_0) < \mathrm{E}(s, c)$ pour toute autre conformation c) ;
2. en plus du critère 1, on impose que le Z-score soit inférieur à un seuil Z_0 ;
3. en plus du critère 2, on impose que la barrière énergétique $\Delta\mathrm{E}$, entre l'état fondamental et le premier état excité, soit supérieure à un seuil $\Delta\mathrm{E}_0$;
4. en plus du critère 1, on impose que l'énergie libre de repliement $\Delta\mathrm{G}_f$ soit inférieure à un seuil $\Delta\mathrm{G}_0$. Cette énergie est calculée conformément aux travaux de Taverna et Goldstein [173], entre autres :

$$\Delta\mathrm{G}_f = \mathrm{E}(s, c_0) + k\,\mathrm{T}\log\left\{\mathcal{Z} - \exp\left(-\mathrm{E}(s, c_0)/(k\,\mathrm{T})\right)\right\},$$

où $\mathcal{Z}$ est la fonction de partition du système et $k\,\mathrm{T} = 0{,}6$ [173].

Nous envisageons en outre des alphabets à trois et vingt classes. La matrice d'énergie pour l'alphabet à trois classes est celle décrite à la page 80 tandis que celle pour l'alphabet complet est la matrice supérieure de Miyazawa et Jernigan du tableau V de la référence [135]. Nous présentons des résultats parcellaires dans la figure III.61.

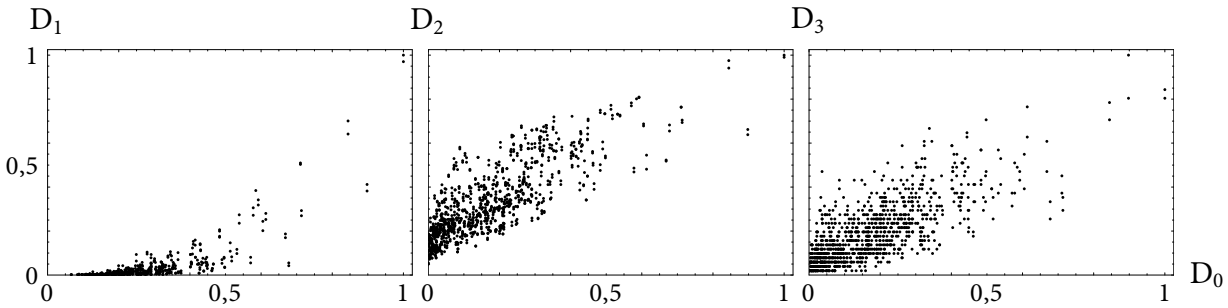


Fig. III.61 : Comparaison des modifiabilités obtenues avec différents alphabets et différents critères de repliement. Les modifiabilités sont normalisées de sorte que la structure la plus modifiable possède une modifiabilité unité. En abscisse, les modifiabilités D_0 obtenues avec la matrice LHTW et un alphabet HP. Elles sont comparées à celles, D_1, D_2 et D_3, calculées dans trois situations différentes. En haut : utilisation du critère 4 avec $\Delta\mathrm{G}_0 = -3.5$ et l'alphabet complet. Au centre : utilisation du critère 1 et l'alphabet à trois classes. En bas : utilisation du critère 3 avec $\mathrm{Z}_0 = -3{,}0$ et $\Delta\mathrm{E}_0 = 0{,}75$ et l'alphabet complet.

La figure III.62 représente quelques distributions de la robustesse mutationnelle représentatives.

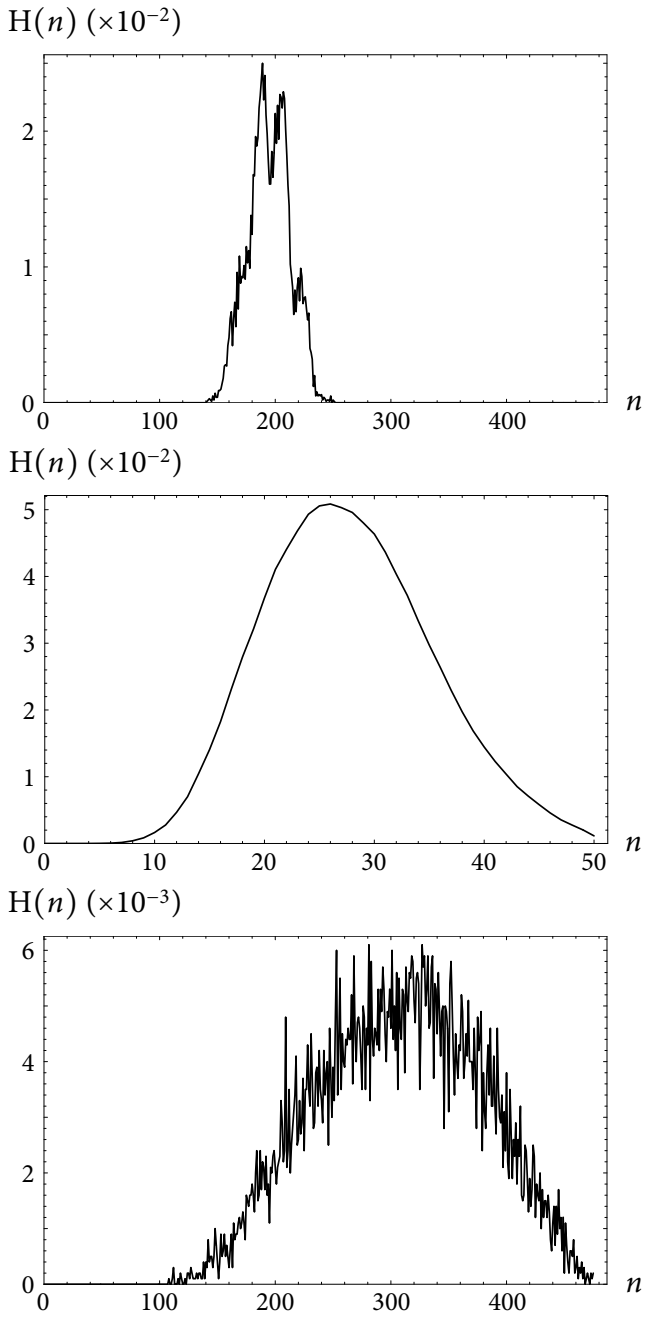


Fig. III.62 : Distribution de la robustesse mutationnelle avec différents alphabets et différents critères de repliement. À gauche : utilisation du critère 4 avec $\Delta G_0 = -1.5$ et l'alphabet complet. Au centre : utilisation du critère 1 et l'alphabet à trois classes. À droite : utilisation du critère 3 avec $Z_0 = -3{,}0$ et $\Delta E_0 = 0{,}75$ et l'alphabet complet.

III.4.8 Modèle tous atomes

Le degré de mutabilité hydrophobe d'une position i peut s'écrire comme une entropie de Shannon :

$$m = -h \log h - (1-h) \log(1-h), \tag{55}$$

où h est la fraction de résidus hydrophobes observés à cette position. Plus m augmente, plus une position peut être aisément mutée.

Soixante mille séquences pour Grb2 ont été générées avec une méthode dérivée de celle de Wernish *et al.* (cf. référence [191]) qui inclut un modèle de solvant implicite. Nous avons converti les séquences en profils. La mutabilité hydrophobe m a été calculée pour ces profils et pour ceux issus de la trajectoire B10, un lissage sur une fenêtre de cinq résidus a été opéré. Les résultats sont présentés dans la figure III.63.

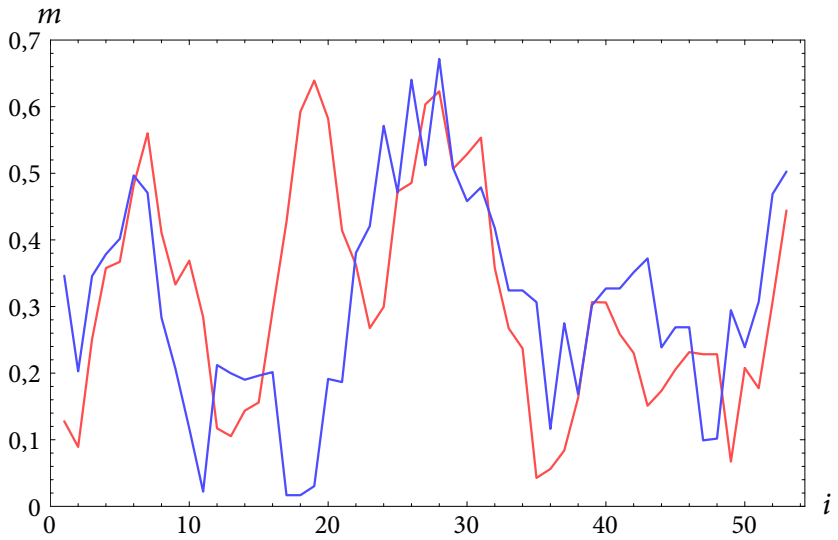


Fig. III.63 : Mutabilité m en fonction de la position i dans la chaîne de Grb2. En bleu, le modèle tous atomes adaptée de Wernish *et al.* ; en rouge, les données de la trajectoire B10.

III.4.9 Distance fondée sur les carbones Cβ

Nous avons modifié le critère de contact : deux résidus sont en contact si la distance qui sépare leurs carbones β est inférieure à 8,5 Å, nous présentons ici les résultats obtenus avec un alphabet à vingt classes. La matrice d'énergie est celle de cet ouvrage. Ces résultats sont donnés à titre indicatif, pour qu'une comparaison puisse être faite, il faudrait reconstruire les réseaux neutres avec un alphabet à trois classes et utiliser des matrices optimisées avec ce nouveau critère de contact.

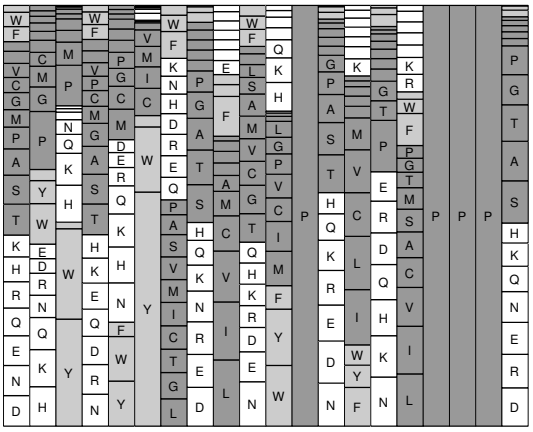


Fig. III.64 : Profil des séquences générées lors d'une trajectoire de $300 \cdot 10^6$ pas avec un critère de contact fondé sur la distance entre carbone β.

III.4.10 Randomisation du graphe

Dans notre discussion, nous nous sommes restreints à comparer deux quantités : la moyenne sous une distribution uniforme de la grandeur x, $\langle x \rangle_u$, et la moyenne pondérée par les fréquences à l'état stationnaire, $\langle x \rangle_s$. Nicolas Lartillot a proposé d'observer la variation de x lorsque l'on randomise les connexions. On part d'un réseau neutre $\mathcal{G} = (\mathbf{N}, \mathbf{E})$ avec le formalisme de « Modèle et méthodes » (p. 47), et l'on choisit aléatoirement T fois, avec T supposé grand, deux arêtes $\{i, j\}$ et $\{k, l\}$ et on les remplace par $\{i, k\}$ et $\{j, l\}$. Ainsi, on déstructure le réseau neutre sans modifier le nombre de connexions de chaque séquence. Ce faisant, on conserve la corrélation qu'il existe entre température de repliement et robustesse mutationnelle et l'on est en mesure de quantifier l'impact de la dynamique de population sur un graphe structuré comparée à celle d'un graphe possédant la même distribution $\mathrm{H}(n)$ mais désorganisé. Les moyennes pondérées par les fréquences à l'état stationnaire de ces graphes déstructurés sont calculées et présentées dans les figures III.65 et III.66.

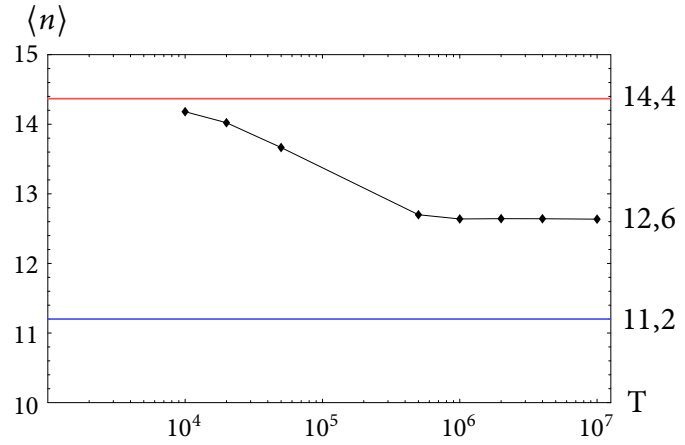


Fig. III.65 : Évolution de $\langle n \rangle$ après T échanges. En bleu est figurée la valeur $\langle n \rangle_u$ du réseau neutre et en rouge, la valeur $\langle n \rangle_s$ du réseau neutre avant que ne soit opéré les échanges.

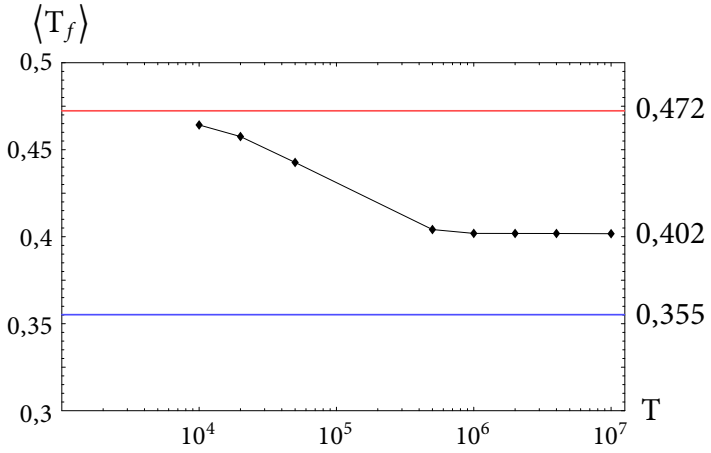


Fig. III.66 : Évolution de $\langle \mathrm{T}_f \rangle$ après T échanges. En bleu est figurée la valeur $\langle \mathrm{T}_f \rangle_u$ du réseau neutre et en rouge, la valeur $\langle \mathrm{T}_f \rangle_s$ du réseau neutre avant que ne soit opéré les échanges.

Chapitre IV

Modèle et méthodes
(protéines dimériques)

Pleurant, comme Diane au bord de ses fontaines,
Ton amour taciturne et toujours menacé.

Alfred de Vigny, *Les Destinées*

Le chapitre « Modèle et méthodes » a introduit les concepts utilisés pour modéliser l'évolution des protéines monomériques. Nous donnons les compléments nécessaires à la prise en considération d'une interaction protéine-protéine.

IV.1 Adaptation du modèle évolutif

IV.1.1 Construction des hyper-réseaux neutres

Nous nous intéressons à des hétérodimères protéine-protéine. La formation des dimères peut procéder par deux mécanismes [8,120,165,205]. Le premier, dit à « trois états » impose que les deux monomères se replient séparément pour, ensuite, interagir. Ce premier mécanisme peut être schématisé par

$$D_1 + D_2 \rightarrow N_1 + N_2 \rightarrow N_1 \cdots N_2$$

où D représente l'état dénaturé d'un monomère et N sa conformation native. L'aspartate aminotransférase obéit à ce schéma [81]. Ce mode de dimérisation est le plus commun. Le second mécanisme, beaucoup plus rare, dit à « deux états » : le repliement des monomères et la formation du dimère sont concomitants. On résume alors :

$$D_1 + D_2 \rightarrow N_1 \cdots N_2 .$$

La formation du répresseur Arc P22 obéit à ce schéma [127].

Une autre terminologie est également utilisée pour différencier ces mécanismes : dimère « transitoire » ou « obligatoire ».

Le mécanisme que nous envisageons est celui à trois états. C'est de loin le mécanisme le plus courant. En outre, il permet de rechercher les séquences favorables à la dimérisation dans les réseaux neutres fondés sur la stabilité structurale. Cette approche a, de plus, l'avantage de pouvoir être traitée avec les ressources de calcul qui sont à notre disposition.

Le concept de réseau neutre a besoin d'être élargi, car il ne suffit plus de compter les séquences compatibles avec une structure ou plus généralement une fonction. Il nous faut énumérer les *paires de séquences* capables de dimériser. Nous nommerons explicitement « génotype » une paire de séquences. On dira qu'un génotype est « viable » si les protéines qui le constituent se replient et dimérisent et « non-viable » dans le cas contraire.

Nous convenons d'appeler A et B les deux protéines. Notons par ailleurs $\mathbf{A}$ l'ensemble des séquences se repliant en A, et $\mathbf{B}$ celui des séquences se repliant en B. L'ensemble des génotypes $\mathbf{s}$ menant à la formation d'un dimère AB stable est un sous-ensemble de l'ensemble produit $\mathbf{S} = \mathbf{A} \times \mathbf{B}$. Sa détermination est l'objet des modèles d'interaction présentés plus tard dans ce chapitre. Pour l'instant, nous nous préoccupons d'adapter le modèle évolutif à ce nouvel espace.

Les ensembles $\mathbf{A}$ et $\mathbf{B}$ sont des réseaux neutres et possèdent par conséquent une structure de graphe. Nous continuerons à utiliser la notation $i \sim j$ pour signifier que les séquences i et j appartenant à un même réseau neutre sont connectées. Étant donné que les mutations doubles sont rares, nous décidons de connecter les génotypes qui diffèrent par une mutation ponctuelle dans une et seulement une des deux séquences. Les deux génotypes (i, j) et (k, l) sont connectés, et nous noterons $(i, j) \sim (k, l)$, si et seulement si

$$\big((i \sim k) \text{ et } (j = l)\big) \text{ ou } \big((i = k) \text{ et } (j \sim l)\big). \tag{56}$$

L'équation 56 confère une structure de graphe aux ensembles $\mathbf{S}$ et $\mathbf{s}$. Comme les nœuds du graphe dimérique appartiennent au produit des réseaux neutres monomériques, nous nommerons le graphe construit « hyper-réseau neutre » ou « réseau neutre dimérique ». S'il

est évident que le graphe **S** est entièrement connecté par la relation 56, il sera nécessaire de s'assurer que le graphe s n'est pas (trop) morcelé.

La robustesse mutationnelle, $n_{(i,j)}$, d'un génotype (i,j) est le nombre total de mutations tolérées par i et par j. La figure IV.1 propose un schéma explicatif de la relation 56. Les figures IV.1C et D schématisent la relation qui existe entre les hyper-réseaux **S** et s.

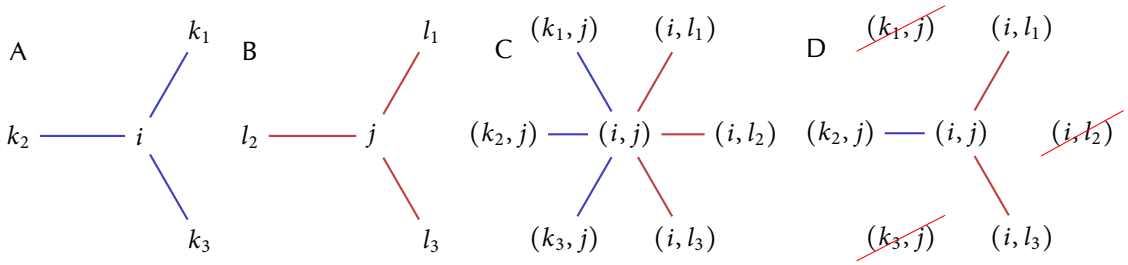


Fig. IV.1 : Topologie de l'hyper-réseau **S** en fonction de celle des réseaux neutres **A** et **B**. A — Une séquence i de **A** et ses connexions. B — Une séquence j de **B** et ses connexions. C — Dans l'hyper-réseau **S** le génotype (i,j) peut muter la séquence i ou la séquence j, mais pas les deux à la fois. Le nombre de connexions est donc la somme du nombre de liens de i dans **A** et du nombre de liens de j dans **B**. D — Après que les génotypes non viables ont été éliminés (ceux qui sont barrés dans le schéma B), on obtient l'hyper-réseau neutre s. La robustesse du génotype (i,j) vaut trois dans l'exemple présenté.

IV.1.2 Équation d'évolution et état stationnaire

L'équation d'évolution est adaptée d'après l'équation 23′ (p. 55). En désignant par $p_{(i,j)}(t)$ la fréquence du génotype (i,j), par μ le taux de mutation par séquence et par L la longueur de chaîne de chaque monomère,

$$p_{(i,j)}(t+1) = p_{(i,j)}(t) + \frac{\mu}{L}\Big(\sum_{(k,l)\sim(i,j)} p_{(k,l)}(t) - \langle n\rangle\, p_{(i,j)}(t)\Big). \tag{57}$$

Dans l'équation 57, $\langle n\rangle$ est le nombre moyen de connexions contractées par les génotypes viables.

$$\langle n\rangle = \sum_{(i,j)} n_{(i,j)}\, p_{(i,j)}\,.$$

Si l'hyper-réseau s est connecté, l'état stationnaire est comme auparavant un vecteur propre associé à la plus grande valeur propre de la matrice d'adjacence décrivant l'hyper-réseau neutre. Si l'hyper-réseau s n'est pas complètement connecté, deux approches sont possibles : soit une composante connexe principale regroupe l'essentiel des génotypes qui sont viables et on peut négliger quantitativement l'effet des plus petites, soit on se restreint à la plus grande composante connexe.

IV.1.3 État stationnaire sans contrainte

Nous nous proposons ici d'étudier l'état stationnaire d'une population infinie évoluant sur l'hyper-réseau neutre **S** dont tous les nœuds sont considérés viables. Cela revient à se demander comment se comporte l'hyper-réseau neutre lorsqu'aucune contrainte de dimérisation n'est ajoutée : on a alors $\mathbf{s} = \mathbf{S}$. Comment l'état stationnaire de cet hyper-réseau neutre se compare aux états stationnaires calculés à l'aide des réseaux neutres **A** et **B** ?

Dans ces paragraphes, nous notons $\mathrm{N_A}$ et $\mathrm{N_B}$ le nombre de séquences appartenant aux réseaux neutres **A** et **B** respectivement. On note $\mathcal{A}$ la matrice d'adjacence du réseau neutre **A**, $\mathcal{B}$ celle du réseau neutre **B**. Soient, α et β des valeurs propres quelconques de $\mathcal{A} = (a_{ik})$ et $\mathcal{B} = (a_{jl})$ respectivement ; soient u et v les vecteurs de $\mathbb{R}^{\mathrm{N_A}}$ et $\mathbb{R}^{\mathrm{N_B}}$ associés aux valeurs propres α et β. Par définition,

$$\mathcal{A}\,u = \alpha\,u, \qquad\qquad \mathcal{B}\,v = \beta\,v, \tag{58a}$$

ou encore pour tout i et tout j,

$$\sum_k a_{ik}\,u_k = \alpha\,u_i, \qquad\qquad \sum_l b_{jl}\,v_l = \beta\,v_j, \tag{58b}$$

Nous formons le vecteur p de $\mathbb{R}^{\mathrm{N_A}\times\mathrm{N_B}}$:

$$p_{(i,j)} = u_i\,v_j. \tag{59}$$

La matrice d'adjacence décrivant $\mathbf{s}$ est une matrice $(\mathrm{N_A}\times\mathrm{N_B})\times(\mathrm{N_A}\times\mathrm{N_B})$, $\mathcal{C} = (c_{(i,j),(k,l)})$ telle que, pour tout i, tout j, tout k et tout l (d'après l'équation 56) :

$$c_{(i,j),(k,l)} = \begin{cases} a_{ik} & \text{si } j = l, \\ b_{jl} & \text{si } i = k, \\ 0 & \text{sinon.} \end{cases}$$

Nous calculons le produit $q = \mathcal{C}\,p$. L'élément (i,j) de q vaut :

$$\begin{aligned} \sum_{(k,l)} c_{(i,j),(k,l)}\,p_{(k,l)} &= \Big(\sum_{(k,l=j)} + \sum_{(k=i,l)}\Big) c_{(i,j),(k,l)}\,u_k\,v_l\,, \\ &= \sum_k a_{ik}\,u_k\,v_j + \sum_l b_{jl}\,u_i\,v_l\,. \\ &= v_j\,\alpha\,u_i + u_i\,\beta\,v_j = (\alpha+\beta)\,u_i\,v_j = (\alpha+\beta)\,p_{(i,j)}\,. \end{aligned} \tag{60}$$

On en déduit que $\alpha + \beta$ est une valeur propre de $\mathcal{C}$ dont p est un vecteur propre. Le spectre de $\mathcal{C}$ est donc l'ensemble des valeurs $\alpha + \beta$ où α parcourt le spectre de $\mathcal{A}$ et β celui de $\mathcal{B}$.

La plus grande valeur propre positive de $\mathcal{C}$ est clairement celle obtenue quand α est la plus grande valeur propre positive de $\mathcal{A}$ et β la plus grande valeur propre positive de $\mathcal{B}$. Notons u_i et v_j les fréquences des séquences i et j dans les états stationnaires monomériques. D'après ce qui précède, le vecteur p : est un vecteur propre associé à la plus grande valeur propre positive de $\mathcal{C}$. En outre, on a

$$\sum_{(i,j)} p_{(i,j)} = \sum_i \sum_j u_i v_j = \sum_i u_i \sum_j v_j = \sum_i u_i \times 1 = 1,$$

ce qui prouve que le vecteur p est bien normalisé pour que $p_{(i,j)}$ s'interprète comme la fréquence du génotype (i,j).

En l'absence de toute contrainte, les événements « être porteur de la séquence i pour A » et « être porteur de la séquence j pour B » sont indépendants (cf. équation 59). Aucun phénomène d'épistase n'a donc lien et le modèle à deux protéines est une extension du modèle monomérique.

On peut calculer également le facteur d'amélioration $\phi(n)$. En effet, en l'absence de contrainte, le nombre de connexions du génotype (i,j) dans s est (cf. figure IV.1)

$$n_{(i,j)} = n_i + n_j,$$

où n_i est le nombre de connexions de i dans $\mathbf{A}$ et n_j est le nombre de connexions de j dans $\mathbf{B}$. On a donc :

$$\langle n \rangle_u = \langle n \rangle_u^{\mathrm{A}} + \langle n \rangle_u^{\mathrm{B}}, \tag{61a}$$

$$\langle n \rangle_s = \langle n \rangle_s^{\mathrm{A}} + \langle n \rangle_s^{\mathrm{B}}, \tag{61b}$$

où $\langle n \rangle^{\mathrm{A}}$ et $\langle n \rangle^{\mathrm{B}}$ sont les moyennes monomériques. D'où,

$$\begin{aligned} \phi(n) &= \frac{\langle n \rangle_s}{\langle n \rangle_u} = \frac{\langle n \rangle_s^{\mathrm{A}} + \langle n \rangle_s^{\mathrm{B}}}{\langle n \rangle_u^{\mathrm{A}} + \langle n \rangle_u^{\mathrm{B}}} \\ &= \frac{\phi(n)^{\mathrm{A}} \langle n \rangle_u^{\mathrm{A}} + \phi(n)^{\mathrm{B}} \langle n \rangle_u^{\mathrm{B}}}{\langle n \rangle_u^{\mathrm{A}} + \langle n \rangle_u^{\mathrm{B}}}. \end{aligned} \tag{62}$$

où $\phi(n)^{\mathrm{A}}$ et $\phi(n)^{\mathrm{B}}$ sont les facteurs d'amélioration monomérique. Le facteur d'amélioration $\phi(n)$ est donc une moyenne arithmétique des facteurs $\phi(n)^{\mathrm{A}}$ et $\phi(n)^{\mathrm{B}}$, pondérés par $\langle n \rangle_u^{\mathrm{A}}$ et $\langle n \rangle_u^{\mathrm{B}}$.

IV.2 Modélisation de la dimérisation

Cette section détaille la méthode utilisée pour connaître quels sont les génotypes viables, c'est-à-dire les paires de séquences dimérisant de façon stable.

IV.2.1 Protéines sur réseau

IV.2.1.1 Modèle d'interaction

Les séquences compatibles avec une conformation sur réseau A (respectivement, B) est un des réseaux neutres, **A** (respectivement, **B**), de cette conformation. Les deux conformations peuvent interagir par n'importe laquelle de leurs faces comme indiqué dans la figure IV.2. Nous considérons également que les monomères peuvent former des homodimères.

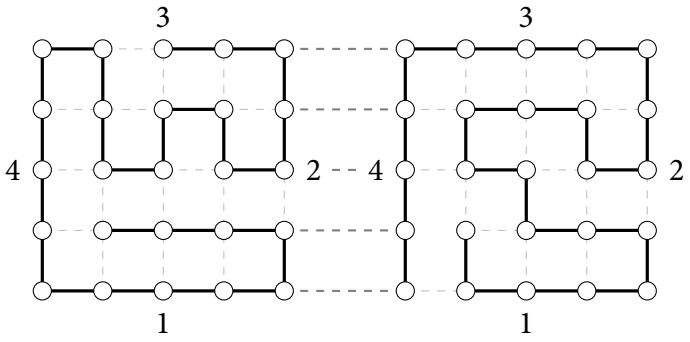


Fig. IV.2 : Deux monomères en interaction. Chaque face de chaque protéine est numérotée de 1 jusqu'à 4. Cinq résidus par protéine sont engagés dans des interactions (signalées en gris foncé) avec l'autre partenaire.

Si les faces des conformations sont numérotées comme dans la figure IV.2, les interactions possibles des hétérodimères sont les seize combinaisons : 1-1, 1-2, 1-3, 1-4, 2-1, 2-2, 2-3, 2-4, 3-1, 3-2, 3-3, 3-4, 4-1, 4-2, 4-3 et 4-4. Les interactions possibles des homodimères sont les dix combinaisons : 1-1, 1-2, 1-3, 1-4, 2-2, 2-3, 2-4, 3-3, 3-4 et 4-4. Nous avons donc dix homodimères AA_{I}, dix homodimères BB_{J} et seize hétérodimères AB_{K}. Si nous ajoutons à cette liste les formes libres A et B, nous avons un mélange de 38 espèces chimiques plus tous les états dépliés. Les états dépliés seront négligés.

Chaque interface comporte cinq contacts. L'énergie d'interaction est la somme des interactions entre les résidus en contact en utilisant la matrice LHTW :

$$\mathcal{E} = \begin{pmatrix} -2.3 & -1 \\ -1 & 0 \end{pmatrix}$$

L'une des interfaces des hétérodimères, parmi les seize possibles, est choisie comme l'interface fonctionnelle. L'hétérodimère correspond sera noté par convention AB_1. L'interface fonctionnelle est choisie de façon à minimiser la moyenne sur l'ensemble des génotypes de $\mathbf{S}$ de l'énergie d'interaction. Il est en effet plus probable qu'une interaction se développe entre deux faces présentant une prédisposition à s'associer.

IV.2.1.2 Équilibre chimique

Établissement des équations L'équilibre entre les formes dissociées et la forme associée de deux protéines obéit à l'équilibre :

$$X + Y \rightleftharpoons XY \quad \mathcal{K} = \frac{[XY]}{[X]\,[Y]} = \exp\Big(-\frac{\mathrm{E} + \gamma}{k\,\mathrm{T}}\Big), \tag{63}$$

où $\mathcal{K}$ est la constante d'association, E l'énergie d'interaction, et $k = 1$ la constante de Boltzmann. La contribution γ est une « pénalité entropique » qui rend compte du coût entropique nécessaire afin de rapprocher les deux partenaires. Le paramètre γ est supposé constant : les entropies de rotation, translation et vibration sont identiques pour toutes les structures (de taille égale).

Étant donné un génotype (i, j), il est possible de connaître toutes les constantes d'association $\mathcal{K}$ pour les 36 équilibres impliquant les hétérodimères et les homodimères. Les équations se résument à

$$[AA_\mathrm{I}] = a_\mathrm{I}\,[A]^2 \qquad 1 \le \mathrm{I} \le 10, \tag{64a}$$

$$[BB_\mathrm{J}] = b_\mathrm{J}\,[B]^2 \qquad 1 \le \mathrm{J} \le 10, \tag{64b}$$

$$[AB_\mathrm{K}] = c_\mathrm{K}\,[A]\,[B] \qquad 1 \le \mathrm{K} \le 16, \tag{64c}$$

où les a_I, b_J et c_K sont les constantes d'associations $\mathcal{K}$ des dimères AA_I, BB_J et AB_K. (Remarque : c_1 est la constante d'association du dimère fonctionnel.) Nous négligeons les états dépliés et nous supposons que la quantité totale de protéine A et B est constante :

$$[A]_\mathrm{tot} = \sum_{1\le \mathrm{I}\le 10} 2\,[AA_\mathrm{I}] + \sum_{1\le \mathrm{K}\le 16} [AB_\mathrm{K}] + [A]\,, \tag{65a}$$

$$[B]_\mathrm{tot} = \sum_{1\le \mathrm{J}\le 10} 2\,[BB_\mathrm{J}] + \sum_{1\le \mathrm{K}\le 16} [AB_\mathrm{K}] + [B]\,, \tag{65b}$$

Résolution numérique Soient les notations suivantes : $a = \sum a_\mathrm{I}$, $b = \sum b_\mathrm{J}$ et $c = \sum c_\mathrm{K}$. Notons x la concentration $[A]$, y la concentration $[B]$, v et w les concentrations totales

$[A]_{\text{tot}}$ et $[B]_{\text{tot}}$. À partir des équations 64a, 64b, 64c, 65a et 65b, on obtient

$$v = 2\,a\,x^2 + c\,x\,y + x, \tag{66a}$$
$$w = 2\,b\,y^2 + c\,x\,y + y. \tag{66b}$$

L'élimination de la variable y mène à

$$y = \frac{v - x - 2\,a\,x^2}{cx}, \tag{67a}$$
$$\mathrm{P}(x) = \alpha_4\,x^4 + \alpha_3\,x^3 + \alpha_2\,x^2 + \alpha_1\,x + \alpha_0 = 0, \tag{67b}$$

avec les coefficients

$$\begin{aligned}
\alpha_4 &= 2\,a\,(4\,a\,b - c^2),\\
\alpha_3 &= 8\,a\,b - 2\,a\,c - c^2,\\
\alpha_2 &= 2\,b - c - 8\,a\,b\,v + c^2\,v - c^2\,w,\\
\alpha_1 &= v\,(c - 4\,b),\\
\alpha_0 &= 2\,b\,v^2.
\end{aligned}$$

Une fois l'équation $\mathrm{P}(x) = 0$ résolue, on peut déduire y, la concentration en B libre grâce à l'équation 67a. La concentration en dimère fonctionnel s'obtient simplement comme le produit $c_1\,x\,y$.

Le polynôme $\mathrm{P}(x)$ est du quatrième degré, il est donc résoluble analytiquement. Cette approche est lente, cependant. Nous avons donc conçu une méthode de résolution numérique que nous décrivons à présent.

Nous rappelons, avant d'en venir à la méthode proprement dite, la relation

$$1 + x \geq \sqrt{1 + 2x} \tag{69}$$

pour toute valeur x telle que $\sqrt{1 + 2x}$ soit définie.

Soient x_1 et x_2 les racines de l'équation $y(x) = 0$ où $y(x)$ est défini par l'équation 67a :

$$x_1 = \frac{-1 - \sqrt{1 + 8\,a\,v}}{4\,a}, \tag{70a}$$
$$x_2 = \frac{-1 + \sqrt{1 + 8\,a\,v}}{4\,a}. \tag{70b}$$

Évidemment, x_1 et x_2 sont respectivement négative et positive. La valeur de P en x_2 est

$$\mathrm{P}(x_2) = c^2\, w\, \frac{-1 - 4\,a\,v + \sqrt{1 + 8\,a\,v}}{8\,a^2}. \tag{71}$$

En vertu de la relation 69, on peut voir que $\mathrm{P}(x_2)$ est négatif. D'autre part, $\mathrm{P}(0) = \alpha_0$ est positif. Il est par conséquent certain qu'il existe au moins une solution à l'équation $\mathrm{P}(x) = 0$ dans l'intervale $]0, x_2[$.

Il est important de noter que la relation 69 garantit également que x_2 fournit une borne physiquement acceptable. En effet, on tire successivement

$$\begin{aligned} x_2 &= \frac{-1 + \sqrt{1 + 8\,a\,v}}{4\,a} \\ &\leq \frac{-1 + (1 + 4\,a\,v)}{4\,a} = \frac{4\,a\,v}{4\,a} = v. \end{aligned}$$

Prouvons à présent qu'une unique solution existe dans l'intervalle $]0, x_2[$. Les racines de $\mathrm{P}(x)$ sont les mêmes que celles de $\mathrm{Q}(x) = \mathrm{P}(x)/(c^2\,x^2)$. $\mathrm{Q}(x)$ est par ailleur du même signe que $\mathrm{P}(x)$. Étudions la dérivée de $\mathrm{Q}(x)$:

$$\begin{aligned} \frac{\mathrm{d}}{\mathrm{d}x}\left[\frac{\mathrm{P}(x)}{c^2\,x^2}\right] = -1 - 4\,a\,x - \frac{1 + 4\,a\,x}{c\,x} \\ - (v - x - 2\,a\,x^2)\left(\frac{1}{c\,x^2} + \frac{4\,b\,(1 + 4\,a\,x)}{c^2\,x^2} + \frac{4\,b}{c^2\,x^3}\right). \end{aligned} \tag{72}$$

Dans l'intervalle, $]0, x_2[$, la quantité $v - x - 2\,a\,x^2$ est positive : car c'est exactement le polynôme dont x_2 est une racine (comparer à l'équation 67a). La dérivée de $\mathrm{Q}(x)$ est du coup manifestement négative.

Comme $\lim_{x\to 0^+} \mathrm{Q}(x) = +\infty$ et $\mathrm{Q}(x_2) = \mathrm{P}(x_2)/(c^2\,x_2^2) < 0$ et comme la dérivée de $\mathrm{Q}(x)$ est négative sur l'intervalle $]0, x_2[$, il existe une unique solution à l'équation $\mathrm{Q}(x) = 0$ dans l'intervalle $]0, x_2[$. La même conclusion s'étend à $\mathrm{P}(x) = 0$

La méthode de résolution numérique s'appuie sur ces conclusions : 1. on calcule la racine x_2, 2. on estime par dichotomie l'intervalle contenant la solution à l'équation $\mathrm{P}(x) = 0$ (précision 10^{-6}), 3. enfin on interpole linéairement la solution à partir des bornes de la l'intervalle obtenu par dichotomie. Les calculs ont été intensément validés à l'aide d'un logiciel de calcul formel (Mathematica).

IV.2.1.3 Température de repliement

La température de repliement d'un génotype (i,j) est le minimum des températures de repliement $T_{f\,i}$ et $T_{f\,j}$ des deux séquences i et j.

$$T_{f\,(i,j)} = \min\{T_{f\,i}, T_{f\,j}\}. \tag{73}$$

IV.2.1.4 Critère sélectif

Le critère sélectif utilisé pour qu'un génotype (i,j) soit considéré viable est que la concentration en dimère fonctionnel soit supérieure à un seuil δ. Le cas $\delta = 0$ correspond à la situation sans contrainte étudiée précédemment (cf. p. 166). Nous ferons varier le paramètre δ pour étudier l'incidence d'une contrainte croissante.

IV.2.2 Protéines tridimensionnelles

Parmi les trois que nous avons étudiées dans le chapitre « Évolution des protéines monomériques », deux protéines, les domaines SH3 et Grb2 et Vav, sont capables d'interagir.

IV.2.2.1 Principe de l'étude

L'expérience gagnée grâce aux simulations d'évolution de modèles tridimensionnels de protéine nous a montré combien il était difficile d'explorer l'espace des séquences même en utilisant un alphabet binaire de type HP. Le problème critique réside en l'obtention d'un nombre suffisant de profils HP tout en s'assurant qu'après un temps assez long, quasiment plus aucun nouveau profil n'est découvert. La difficulté augmente rapidement avec la taille de la protéine. Par conséquent, simuler par Monte Carlo l'évolution du dimère Grb2-Vav devient impensable puisque cela revient à entreprendre l'évolution d'une protéine de taille $L = 57 + 69 = 126$. La stratégie envisagée nous a donc conduit à simuler séparément les protéines monomériques comme décrit dans la section « Protéines tridimensionnelles », p. 106 et de tester leur capacité à dimériser *a posteriori* conformément à ce qui a été annoncé en introduction de ce chapitre.

Malgré tout, il est hors de question de traiter toutes les séquences acceptées lors des trajectoires B10 et C10. En effet, au cours de ces trajectoires quelque $10 \cdot 10^6$ séquences furent acceptées pour chaque monomère, le nombre de combinaisons atteignant donc 10^{14}. La simplification suivante est donc mise en œuvre : pour chaque simulation, une seule séquence est choisie pour représenter un profil HP, la première dans l'ordre dans lequel les profils sont découverts. En pratique, bien que nous utilisions une matrice binaire, deux séquences différentes partageant le même profil n'ont pas, en toute rigueur, la même énergie car les différences d'extension tridimensionnelle des chaînes latérales induisent des différences dans les contacts établis par l'une et l'autre chaîne. Cette remarque s'applique aussi

bien au calcul de l'énergie d'une séquence repliée dans une conformation (comme nous l'avons fait plus tôt), que dans celui de l'énergie d'interaction entre deux chaînes (comme nous le ferons). Nous négligerons cet effet, car, malgré tout, la composante hydrophobe-hydrophile détermine principalement l'interaction entre protéines.

La sélection du sous-ensemble des paires de séquences dimérisant se fait selon un critère similaire à celui de l'acceptation de mutations dans les simulations d'évolution monomériques. Chaque paire de séquences est utilisée pour calculer une énergie empirique d'interaction pour le dimère natif et un jeu de faux dimères représentant les interactions non spécifiques. Les faux dimères sont composés des structures natives de Grb2 et Vav interagissant via une interface non native. Leur construction se fait par une technique d'amarrage décrite plus bas. Si l'interaction du dimère natif est suffisamment basse en comparaison de celles des faux dimères, la paire de séquences est acceptée.

En pratique, pour toute paire de séquences, on estime toutes les énergies des faux dimères et du dimère natif. Si le dimère natif est classé dixième ou mieux, la paire de séquence est conservée. On calcule également le Z-score du spectre des énergies d'interaction.

Afin de rester cohérent avec les simulations menées jusqu'alors, nous avons utilisé la même matrice HP que précédemment :

$$\mathcal{E}_{2C} = \begin{array}{c} \\ \mathrm{H} \\ \mathrm{P} \end{array} \begin{array}{c} \begin{array}{cc} \mathrm{H} & \mathrm{P} \end{array} \\ \begin{pmatrix} -8.5 & 9 \\ 9 & -3.5 \end{pmatrix} \end{array}$$

bien que des matrices différences puissent être obtenues par une optimisation spécifique aux dimères [106]. Notre matrice d'interaction est capable de détecter les régions favorables à une interaction comme le démontre la figure IV.3. Pour chaque faux dimère, nous avons projeté le centre de masse de Grb2 sur une sphère centrée sur Vav. L'énergie de chaque faux dimère est calculée avec les séquences natives (celle de la PDB) et est représentée par un code de couleur : bleu pour les énergies basses, gris pour les énergies intermédiaires et rouge pour les énergies élevées. Il est remarquable que les régions d'interaction favorables soient les deux interfaces présentes dans la structure cristalline de la PDB.

Les deux interfaces identifiées dans le cristal de la structure 1GCQ sont deux interfaces d'interaction putatives entre Grb2 et Vav. Certains arguments suggèrent que l'interface biologique serait celle impliquant la chaîne A de Grb2 [145]. Nous nous sommes intéressé à la chaîne B qui est néanmoins identifiée par notre matrice d'énergie comme une interface propice à l'interaction.

L'hypothèse d'un dimère transitoire est justifiée par plusieurs observations biologiques. Premièrement, l'interaction Grb2-Vav est transitoire, la constante d'affinité étant $5 \cdot 10^{-6}$ M

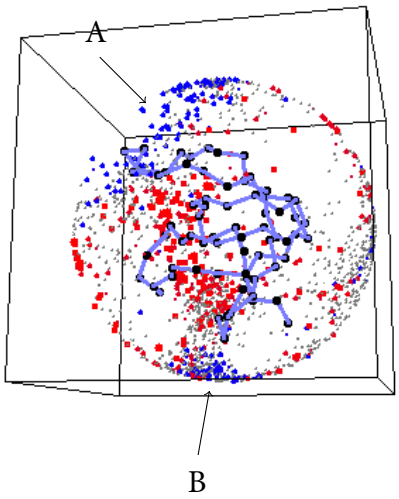


Fig. IV.3 : Projection des centres de masse de Grb2 sur une sphère entourant Vav pour tous les decoys. Les énergies sont représentées par un code de couleur : les énergies les plus favorables sont en bleu, les plus défavorables en rouge, les énergies intermédiaires sont grisées. Les deux surfaces d'interaction identifiées lors de la cristallisation sont indiquées par des flèches.

environ. Deuxièmement, la surface d'interaction est relativement faible : la surface accessible au solvant totale perdue lors de l'interaction est de 1 270 Å^2, ce qui est relativement limité [88]. Enfin, troisièmement, la structure des domaines liés est identique à celle sous forme libre, l'interaction se faisant par recrutement rigide des partenaires [145].

IV.2.2.2 Séquences de référence

Les trajectoires à partir desquelles sont extraites les séquences dont on testera l'aptitude à dimériser sont les trajectoires B10 et C10 décrites dans « Trajectoires d'exploration dimériques » (p. 107). Les séquences de référence de ces trajectoires présentent la particularité de ne pas avoir été mutées en dix résidus chacune. En effet, la préoptimisation produit des séquences très hydrophobes dans le cœur de la structure et très polaire à la surface. Le gel de ces dix positions permet de conserver les résidus présents à l'interface fonctionnelle, hydrophobes et polaires, dans une disposition favorable à l'interaction.

IV.2.2.3 Formation des fausses structures

Les fausses structures dimériques sont formées à partir des structures natives par une procédure d'amarrage selon une méthode originale conçue au laboratoire. Nous en résumons ici les grands principes. Nous renvoyons le lecteur à la référence [106] pour de plus amples informations.

1. Les deux conformations natives sont dans un premier temps écartées dans une direction aléatoire et une rotation aléatoire est appliquée à l'un des deux partenaires.

2. Une minimisation dans le vide en corps rigide est réalisée avec charmm19 [24].
3. Les partenaires sont à nouveau rapprochés en incluant un potentiel harmonique attirant les deux centres de masse des deux protéines jusqu'à atteindre la distance souhaitée correspondant à la distance observée pour le dimère natif. La constante de rappel vaut 60 kcal $\cdot$ mol^{-1} $\cdot$ Å^{-1}.
4. Une nouvelle étape de minimisation en corps rigide de cent pas est entreprise en incluant le potentiel harmonique et le nombre de contacts intermoléculaires est estimé selon le critère d'une distance inférieure à 4,5 Å. Si ce nombre représente moins de 80 % des contacts observés dans le complexe natif, la structure est écartée.
5. Le potentiel harmonique est supprimé et une étape de minimisation est reprise en autorisant des changements conformationnels de l'un des deux partenaires. À cette étape, les charges des deux monomères sont modifiées de $\pm e/4$ pour favoriser la formation de contacts. Cette différence de charge n'est pas prise en compte pour les interactions intramoléculaires.
6. À cinquante pas de minimisation de Powell, succède une phase de comparaison structurale au complexe natif. Le dimère généré est comparé à celui natif en superposant le monomère de chaque dimère n'ayant pas subi de déformation du squelette peptidique — appelons-le A et B, son partenaire. Le rms des chaînes latérales entre les deux structures de A est comparé à 4,5 Å, s'il lui est supérieur, la structure est écartée.
7. Si le nombre de contacts entre les partenaires est inférieur à 45 % de ce qu'il est dans le complexe natif, la structure est écartée.
8. Si l'énergie de van der Waals est plus élevée que 4 000 kcal $\cdot$ mol^{-1}, la structure est écartée.
9. Si le rms entre les deux versions du monomère B est inférieur à 3 Å, c'est-à-dire, si la fausse structure est trop proche du dimère natif, la structure est également écartée.

Pour nos simulations, le détail des chaînes latérales n'intervient pas, puisqu'à nouveau ce sont les rotamères définis par Tuffery *et al.* [182] qui sont greffés sur les squelettes peptidiques pour les évaluations énergétiques. Pour 1GCQ B et C, le nombre des fausses structures est 1 695 desquels nous rejetons 912 qui présentent une interaction plus favorable que le dimère natif lui-même. Ce taux élevé de faux positifs est dû au fait que l'interface des séquences pseudo-natives n'ont pas été préoptimisées. Par ailleurs, il faut noter qu'en comparaison des simulations des dimères sur réseau, le nombre de compétiteurs est extrêmement élevé : le nombre de faux dimères s'élevait à 35 seulement.

Chapitre V

Évolution des protéines dimériques

Such questions need to be answered despite the intricacies of their pursuit. Whatever the answers may be, the process of adopting a preexisting, additional factor into a regulatory complex must have been a frequent occurrence in evolution.

Zuckerkandl, *Neutral and nonneutral mutations*

V.1 Introduction

Zuckerkandl proposa que les taux d'évolution des protéines sont déterminés par leur « densité fonctionnelle » (*functional density*) qui est la fraction des sites qui contribuent de façon importante à la fonctionnalité d'une protéine [209]. Plus tard une autre expression fut introduite pour désigner la même caractéristique : « densité de *fitness* » (*fitness density*) [42]. Quelle que soit la terminologie, l'hypothèse sous-jacente est celle que nous avons énoncée dans l'introduction : la contrainte fonctionnelle détermine le taux de substitution. On s'attend, par exemple, que la contrainte imposée sur les histones soit très grande, compte tenu de la forte conservation de ces protéines.

Intuitivement, les associations multimériques devraient augmenter la densité fonctionnelle et donc diminuer les taux de substitution dans les protéines dont la fonction requiert une interaction. À ce sujet, aucune preuve expérimentale n'a permis de conclure clairement. D'un côté, le polymorphisme des protéines multimériques semble diminué indiquant une sélection négative plus stringente [190] (cf. l'équation 16, p. 30 de l'intro-

duction), et une première étude a révélé une corrélation négative entre le nombre de partenaires dans le réseau d'interaction de la levure et la vitesse d'évolution [61]. D'un autre côté, cette corrélation n'a pas pu être confirmée par d'autres groupes [15, 74]. D'autres résultats sont surprenants : les protéines capables d'interagir avec elles-mêmes sont généralement plus stables [43, 79]. Faut-il en conclure qu'il existe une pression sélective faisant évoluer les dimères vers des structures plus modifiables ? Cet argument n'est pas suffisant. En effet, on peut également remarquer que l'hémoglobine et la myoglobine, qui ne diffèrent à première vue que par leur capacité à se complexer, ont un taux d'évolution très semblable (1,0 substitutions par milliard d'année et par position pour la première, 0,9 pour la seconde [122]). Quoi qu'il en soit, si l'association protéine-protéine est contraignante, elle l'est faiblement tant semble difficile sa mise en évidence.

L'importance des interactions protéine-protéine est à présent pleinement réalisée. De nombreux travaux tentent de démêler les principes de la reconnaissance spécifique entre protéines (voir la revue [198]). Les méthodes de prédiction d'interaction s'améliorent [125], et les mécanismes de liaison à deux ou trois états sont en voie d'élucidation par des méthodes de simulation moléculaire [107–109]. Mais les interactions protéine-protéine n'ont quasiment pas été abordées depuis un angle évolutif.

Quelques exceptions existent. Des études d'évolution *in vitro* ont été menées pour augmenter l'affinité d'un anticorps par la technique de *phage display* [175]. Mais il s'agit d'évolution dirigée et en aucun cas neutre. L'interaction protéine-protéine est au cœur des études portant sur l'agrégation dont l'importance est démontrée par les nombreuses maladies qui y sont associées [39, 78, 86]. Le groupe de Goldstein modélise très souvent la fonctionnalité d'une protéine par son aptitude à lier un ligand peptidique (qui toutefois n'évolue pas) [196, 197]. Enfin, Tiana et ses collaborateurs ont étudié les motifs de conservation des résidus dans le cas de protéines interagissant selon un mécanisme à deux ou trois états [178, 179].

Nous modélisons les interactions protéine-protéine par la formation d'un dimère AB à trois états. Cette hypothèse permet de rechercher les génotypes viables dans l'ensemble des paires de séquences (i, j) où i et j sont deux séquences se repliant dans les conformations A et B. Les séquences i et j font donc partie des réseaux neutres monomériques qui ont été présentés dans le chapitre « Modèle et méthodes » (p. 47) et étudiés dans le chapitre « Évolution des protéines monomériques » (p. 47). La viabilité du génotype (i, j) repose sur la comparaison de l'énergie d'interaction du dimère fonctionnel à celle de faux dimères.

Le modèle de protéine sur réseau résout la concentration du dimère fonctionnel et cette concentration est utilisée comme critère sélectif : ne peuvent survivre que les génotypes dont la concentration est supérieure à une valeur minimale. Le modèle de protéine

tridimensionnelle hors-réseau se fonde sur le rang du dimère fonctionnel pour accepter ou rejeter un génotype.

V.2 Résultats

Pour le modèle de dimère sur réseau, le critère sélectif est la concentration en dimère fonctionnel. Une concentration supérieure à un seuil est requise pour assurer la viabilité d'un individu. Le critère sélectif utilisé pour le modèle de dimère tridimensionnel et hors-réseau est le rang k du dimère fonctionnel comparé aux faux dimères par ordre d'énergie croissante.

V.2.1 Protéines sur réseau

Nous étude portera sur les statistiques de viabilité puis répondra à la question de la localisation des génotypes viables. Nous mettrons en évidence les traits distinctifs de l'évolution de la robustesse mutationnelle, la température de repliement et de la concentration en dimère fonctionnel. Nous observerons l'existence d'une interaction épistatique entre les gènes exprimant les protéines du complexe. Enfin, nos résultats suggéreront un l'organisation en *superfunnel* pour la concentration en dimère fonctionnel, ce que nous nommerons « *superfunnel* fonctionnel ».

On notera $\mathbf{s}^{(\delta)}$ l'hyper-réseau obtenu pour la contrainte fonctionnelle δ. On a $\mathbf{s}^{(0)} = \mathbf{S}$, l'ensemble de tous les génotypes (i, j) tels que i et j se replient chacun dans leur conformation native. Par ailleurs, on a évidemment la relation

$$\mathbf{s}^{(\delta_1)} \subset \mathbf{s}^{(\delta_2)},$$

si $\delta_1 \geq \delta_2$.

Les valeurs suivantes ont été utilisées pour les paramètres de dimérisation des protéines sur réseau : $\gamma = 5$, $[A]_{\text{tot}} = [B]_{\text{tot}} = 10$. La valeur de γ a été choisie pour qu'environ deux contacts H-H soient nécessaires pour compenser la pénalité entropique du rapprochement des partenaires. D'autres valeurs ont été testées, quoique moins intensivement, et mènent aux mêmes conclusions : $[A]_{\text{tot}} = [B]_{\text{tot}} \in \{1, 5, 10\}$, et $\gamma \in \{1, 2, 5, 10\}$. Un choix différent de γ peut être compensé par changement d'échelle de la contrainte δ : en effet, γ influence la concentration totale des différents dimères AA_{I}, BB_{J} et AB_{K} mais pas la concentration relative du dimère fonctionnel par rapport aux autres dimères.

Nous nous limiterons à des réseaux neutres monomériques de taille approximative $N = 10\,000$ pour pouvoir calculer l'état stationnaire de la population. Le filtrage fera

intervenir des réseaux de toute taille, car il ne nécessite pas de reconstruire les hyper-réseaux neutres. La reconstitution des composantes connexes a été entreprise sur des réseaux neutres monomériques de taille maximale N = 5 000.

V.2.1.1 Concentration en dimère fonctionnel

La distribution de la concentration en dimère fonctionnel a une allure de distribution exponentielle s'étalant de 0 jusqu'à environ 3 ; elle est présentée dans la figure V.1 pour l'association 854 (a)×136 (a). La concentration moyenne en dimère fonctionnel vaut 0,32 et l'écart type 0,13. L'histogramme atteste que la majorité des génotypes ont une faible capacité à dimériser.

La concentration maximale en dimère fonctionnel est environ le tiers des concentrations monomériques totales. Cela se produit lorsqu'une paire de séquences possède une interface quasi entièrement hydrophobe. Dans ce cas, la même interface peut servir à l'association de deux homodimères *AA* et *BB* possédant approximativement la même énergie d'interaction. Le dimère fonctionnel et ces deux homodimères se partagent alors à peu près équitablement la concentration totale.

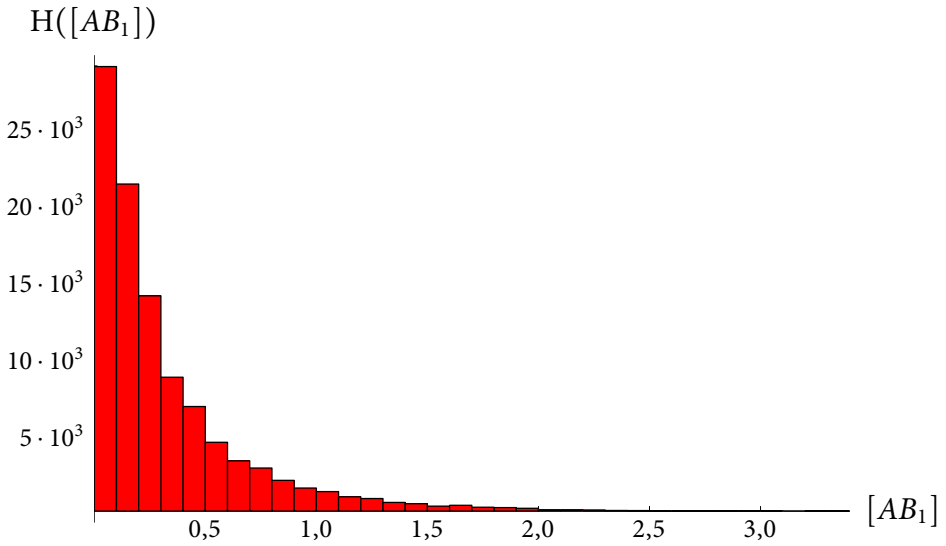


Fig. V.1 : Distribution de la concentration en dimère fonctionnel calculée avec les paramètres indiqués dans le texte avec les réseaux neutres monomériques 854 (a) et 136 (a).

Les profils moyens obtenus en appliquant des contraintes fonctionnelles croissantes sont présentés dans la figure V.2 pour le dimère 509 (a)×786 (a). Les profils obtenus sans contrainte ($\delta = 0$) sont simplement les profils moyens monomériques.

Le nombre de génotypes viables en fonction de la contrainte fonctionnelle δ est indiqué dans le tableau V.1 pour le dimère 854 (a)×136 (a).

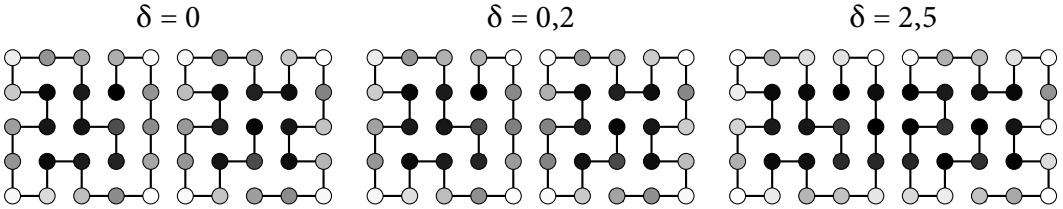


Fig. V.2 : Profils moyens obtenus pour le dimère 509 (a)×786 (a) pour les contraintes fonctionnelles δ = 0, 0,2 et 2,5.

Contrainte fonctionnelle δ	*Nombre de survivants*	*Pourcentage (%)*
0	96 108 582	100
0,1	68 290 887	71
0,2	47 695 260	50
0,3	34 273 488	36
0,4	25 785 086	27
0,6	14 825 259	15
0,8	8 983 407	9,3
1	5 522 233	5,8
1,5	1 626 428	1,7
2	353 538	0,37
2,5	47 192	0,05

Tab. V.1 : Nombre de survivants en fonction de la contrainte fonctionnelle δ pour le dimère 854 (a)×136 (a).

V.2.1.2 Composantes connexes de l'hyper-réseau neutre

Nous avons souligné dans la description du modèle l'importance de s'assurer que l'hyper-réseau neutre n'était pas trop morcelé. Nous avons donc formé les composantes connexes pour sept paires de réseaux neutres de tailles approximatives 5 000 × 5 000 : 134 (a)×37 (a), 179 (a)×687 (a), 238 (e)×452 (a), 33 (a)×1057 (a), 400 (a)×556 (a), 686 (g)× 77 (b) et 627 (b)×249 (a). La même figure ressort de toutes les simulations : nous présentons le résultat correspondant à 627 (a)×249 (a) dans la figure V.3. Une grande composante connexe émerge et regroupe plus de 95 % des génotypes viables pour la plupart des valeurs sélectives δ.

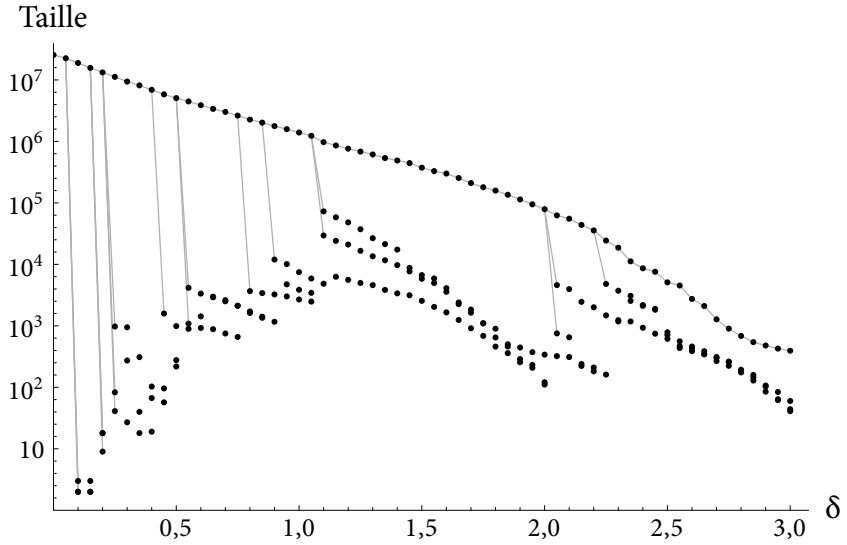


Fig. V.3 : Taille des quatre plus grandes composantes connexes issues de la sélection des génotypes viables en fonction de la contrainte fonctionnelle δ.

V.2.1.3 Localisation des génotypes viables

Soit l'hyper-réseau neutre $\mathbf{s}^{(\delta)} \subset \mathbf{S} = \mathbf{A} \times \mathbf{B}$. Une séquence « viable » de $\mathbf{A}$ est une séquence $i \in \mathbf{A}$ telle qu'il existe au moins une séquence $j \in \mathbf{B}$ avec laquelle elle forme un dimère stable :

$$i \in \mathbf{A} \text{ viable} \iff \exists j \in \mathbf{B}, (i,j) \in \mathbf{s}^{(\delta)}.$$

Et réciproquement pour les séquences viables de $\mathbf{B}$. Les séquences viables sont simplement les séquences de l'un des deux monomères qu'il est possible d'observer. La figure V.4 représente le réseau neutre 854 (a) disposé circulairement autour de la séquence prototype monomérique au centre et signale par des points rouges les séquences viables du réseau neutre.

La figure V.4 semble indiquer que les séquences viables n'ont pas de localisation spécifique dans le réseau neutre. Si l'on s'attendait à ce que les séquences viables soient moins

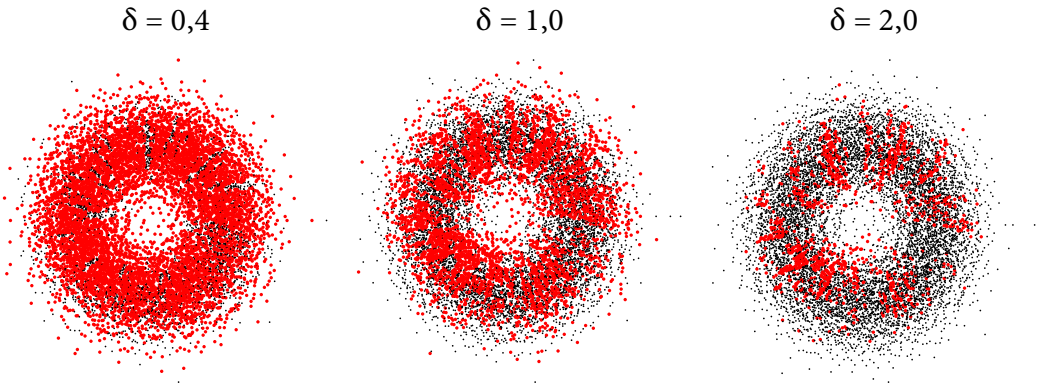


Fig. V.4 : Le réseau neutre 854 (a) est disposé circulairement autour de la séquence prototype monomérique. Les points rouges signalent les séquences viables du réseau neutre, pour différentes valeurs de δ (dimère 854 (a)×136 (a)). La distance au centre représente l'éloignement (cf. p. 52) à la séquence prototype. Les coordonnées ont été modifiées aléatoirement pour permettre de distinguer la densité des points.

stables, elles devraient être plus nombreuses en périphérie du réseau neutre qu'au centre, conformément à la structure en *superfunnel*. Les graphiques de la figure V.5 confirment que ce n'est pas le cas. Nous avons représenté pour chaque génotype,

1. le nombre de connexions formées dans le graphe complet **S** en fonction de la concentration en dimère fonctionnel et
2. la température de repliement T_f en fonction de la concentration en dimère fonctionnel.

Les autres dimères amènent à la même conclusion : il n'existe essentiellement pas de corrélation entre ces quantités.

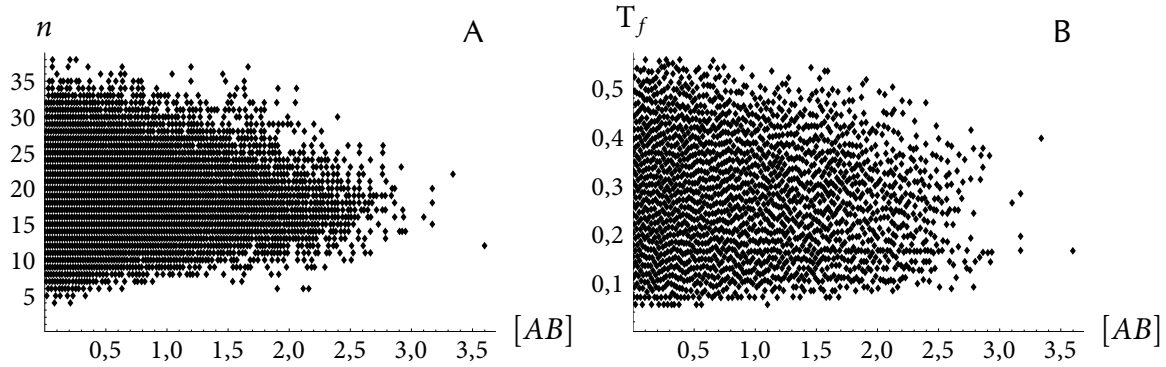


Fig. V.5 : À gauche, représentation du nombre de connexion dans l'hyper-réseau neutre complet **S** en fonction de la concentration en dimère natif. À droite, représentation de la température de repliement en fonction de la concentration en dimère natif.

L'absence de corrélation entre la température de repliement et la concentration en dimère fonctionnel est surprenante de prime abord. En effet, les séquences capables de

former un dimère fonctionnel stable doivent exposer des résidus hydrophobes à la surface, sur l'interface fonctionnelle des protéines. Or, l'exposition de résidus hydrophobes à la surface déstabilise une protéine. Par exemple, la séquence prototype du réseau neutre monomérique 786 (a), ne possède qu'un seul résidu exposé à la surface (cf. figure III.9, p. 96).

Nous avons donc voulu savoir si les génotypes viables étaient composés de protéines exposant plus de résidus hydrophobes à mesure que la contrainte δ augmentait. Nous présentons le résultat pour la protéine 854 (a) dans le dimère 854 (a)×136 (a) dans la figure V.6.

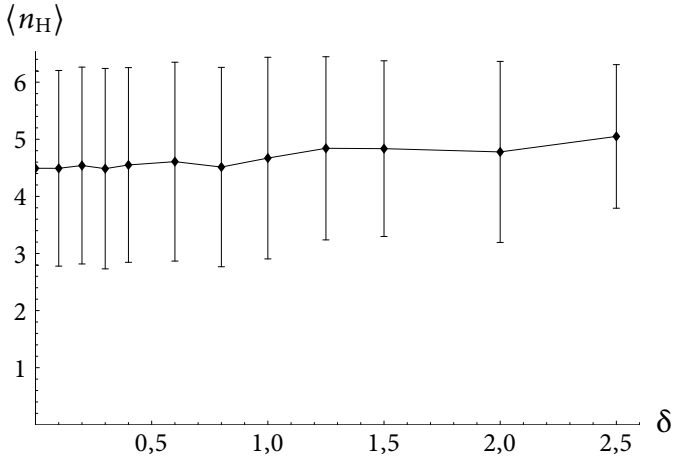


Fig. V.6 : Nombre moyen de résidus exposés à la surface de la protéine 854 (a), $\langle n_{\mathrm{H}} \rangle$, quand elle interagit avec 136 (a), en fonction de la contrainte δ.

La figure V.6 suggère qu'une plus grande contrainte fonctionnelle ne requiert pas l'exposition d'un surcroît de résidus hydrophobes, ou tout du moins que ce surcroît n'est pas très sensible. Cela est rendu possible par la plasticité des réseaux neutres. L'introduction de la contrainte fonctionnelle ne fait que sélectionner les séquences qui possèdent les résidus hydrophobes à l'interface fonctionnelle, c'est-à-dire bien positionnés.

V.2.1.4 Filtrage

Nous identifions une protéine (une fonction enzymatique, par exemple) à un réseau neutre. Étant donné une contrainte δ et une protéine « pivot », il est possible de déterminer toutes les séquences viables de tous les réseaux neutres. Nous appellerons cette procédure « filtrage » car les réseaux neutres sont épurés des séquences qui ne sont pas capables de dimériser avec la protéine pivot. Nous noterons, dans ce paragraphe, **A** le réseau neutre correspondant à la protéine pivot. Chaque réseau neutre **B** est examiné successivement. Une séquence j de **B** est viable s'il existe au moins une séquence i de **A** tel que le génotype (i, j) soit viable (la concentration en dimère fonctionnel est supérieure à δ). L'ensemble des séquences j viables de **B** forme le « réseau neutre **B** filtré ».

Nous choisissons le réseau neutre pivot 5 (a) (N = 10 079) et la contrainte $\delta = 2{,}0$. Nous évaluons la distance de Hamming maximale, $d'_{\max}$ entre deux séquences de chaque réseau neutre filtré. En outre, nous calculons la distance de Hamming minimale, $d'_{\min}$ séparant chaque paire de réseaux neutres filtrés. Ces quantités sont comparées aux valeurs $d_{\max}$ et $d_{\min}$, obtenues lorsque les réseaux neutres ne sont pas filtrés. Nous avons schématisé cette approche dans la figure V.7. Les résultats sont présentés dans la figure V.8.

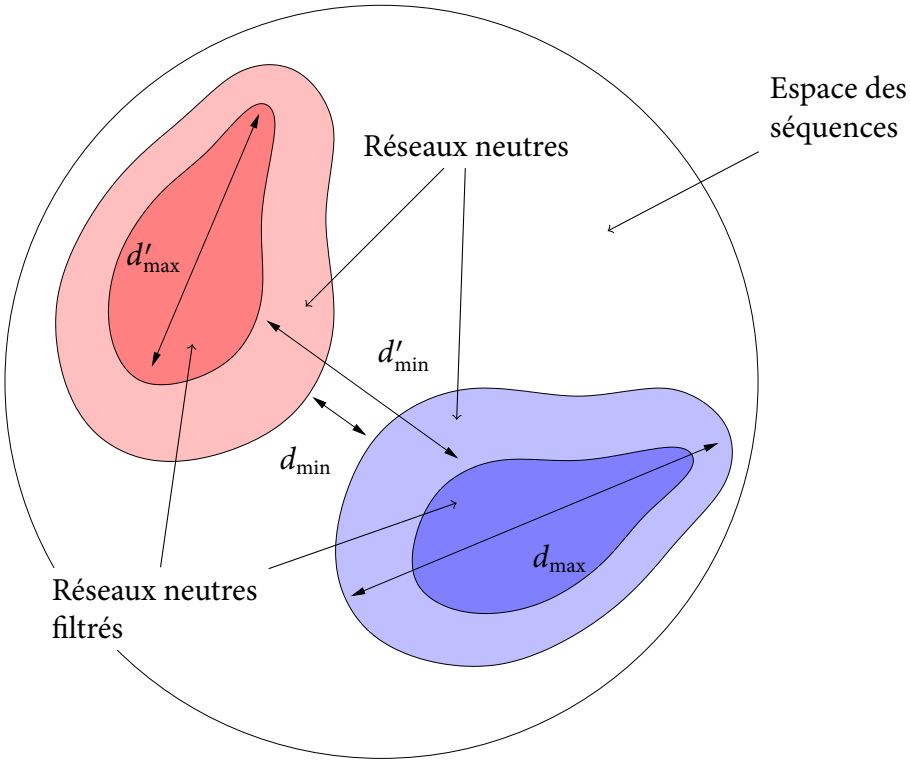


Fig. V.7 : Le filtrage correspond à l'élimination de toutes les séquences non viables pour former un dimère fonctionnel avec la protéine pivot. Les distances de Hamming maximales intra-réseau neutre sont comparées avant ($d_{\max}$) et après filtrage ($d'_{\max}$). Les distances de Hamming minimales inter-réseaux neutres sont comparées avant ($d_{\min}$) et après filtrage ($d'_{\min}$).

Le filtrage réduit la distance de Hamming maximale intra-réseau neutre et augmente la distance de Hamming minimale inter-réseaux neutres. Cependant en dépit d'un nombre restreint de génotypes survivants (de l'ordre de 0,5 % de la taille des hyper-réseaux neutres non filtrés), l'effet du filtrage sur la diversité des séquences est relativement limité. Cette observation est conforme à la répartition délocalisée des séquences viables comme indiqué dans la figure V.4.

L'uniformité des résultats suggère qu'il n'existe pas de structures privilégiées par la dimérisation : toute structure est une bonne candidate autant qu'une autre pour dimériser avec la protéine pivot.

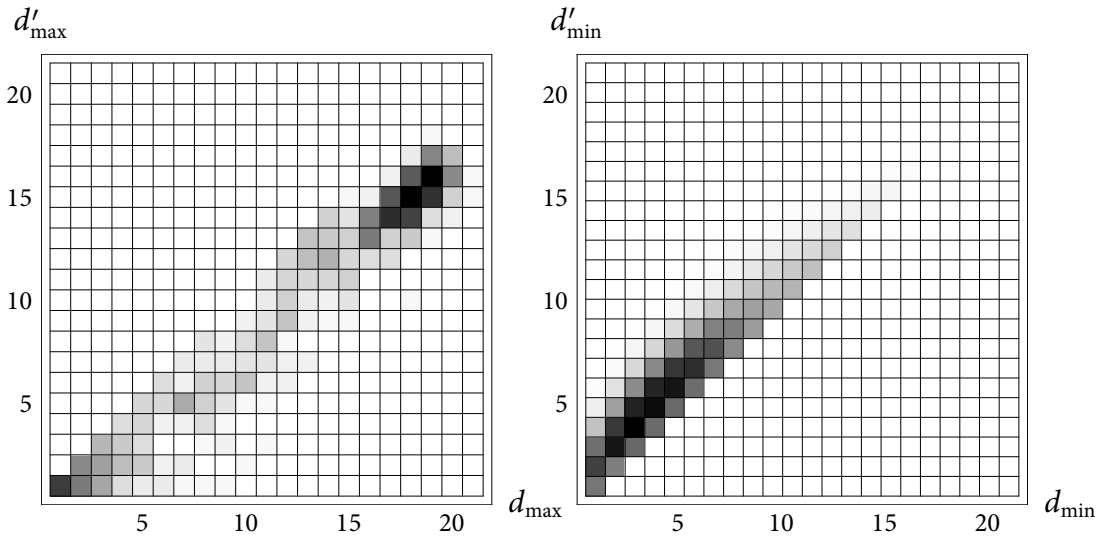


Fig. V.8 : Filtrage des réseaux neutres par le réseau neutre pivot 5 (a) avec la contrainte fonctionnelle $\delta = 0{,}2$. À gauche, graphique de densité représentant la distance de Hamming intra-réseau neutre après filtrage ($d'_{\max}$) en fonction de celle avant filtrage ($d_{\max}$). À droite, graphique de densité représentant la distance de Hamming inter-réseaux neutres après filtrage ($d'_{\min}$) en fonction de celle après filtrage ($d_{\min}$). Les diagonales sur ces graphiques correspondent à l'absence de modification des quantités $d_{\max}$ et $d_{\min}$ par le filtrage.

V.2.1.5 Propriétés à l'état stationnaire

Nous proposons à présent de comparer les propriétés moyennes à l'état stationnaire à celles obtenues lorsque la population est répartie uniformément sur l'hyper-réseau neutre. En plus des propriétés qui ont été étudiées pour les monomères, la robustesse mutationnelle n et la température de repliement T_f, nous nous intéressons à une troisième propriété : la concentration en dimère fonctionnel. La figure V.9 représente l'évolution des moyennes de ces trois propriétés en fonction de la contrainte fonctionnelle δ (à gauche). Les distributions obtenues pour une contrainte $\delta = 0{,}4$ sont placées à droite de la figure V.9

Les conclusions importantes que l'on peut tirer de ce résultat sont multiples.

1. (a) La robustesse mutationnelle diminue à mesure qu'augmente la contrainte fonctionnelle δ. En effet, plus la contrainte δ est grande, moins il y a de génotypes viables. L'hyper-réseau neutre perd donc des connexions.

   (b) À l'état stationnaire, la robustesse mutationnelle est renforcée, comme c'était le cas pour l'évolution des protéines monomériques.

2. (a) La température de repliement reste constante quelle que soit la contrainte fonctionnelle δ. Cela est dû à l'absence de corrélation entre la température de repliement T_f et la concentration en dimère fonctionnel, comme nous l'avons observé dans la figure V.5. Les séquences s'associant en un dimère stable ne sont pas plus instables. En outre, contrairement à la robustesse mutationnelle, la sup-

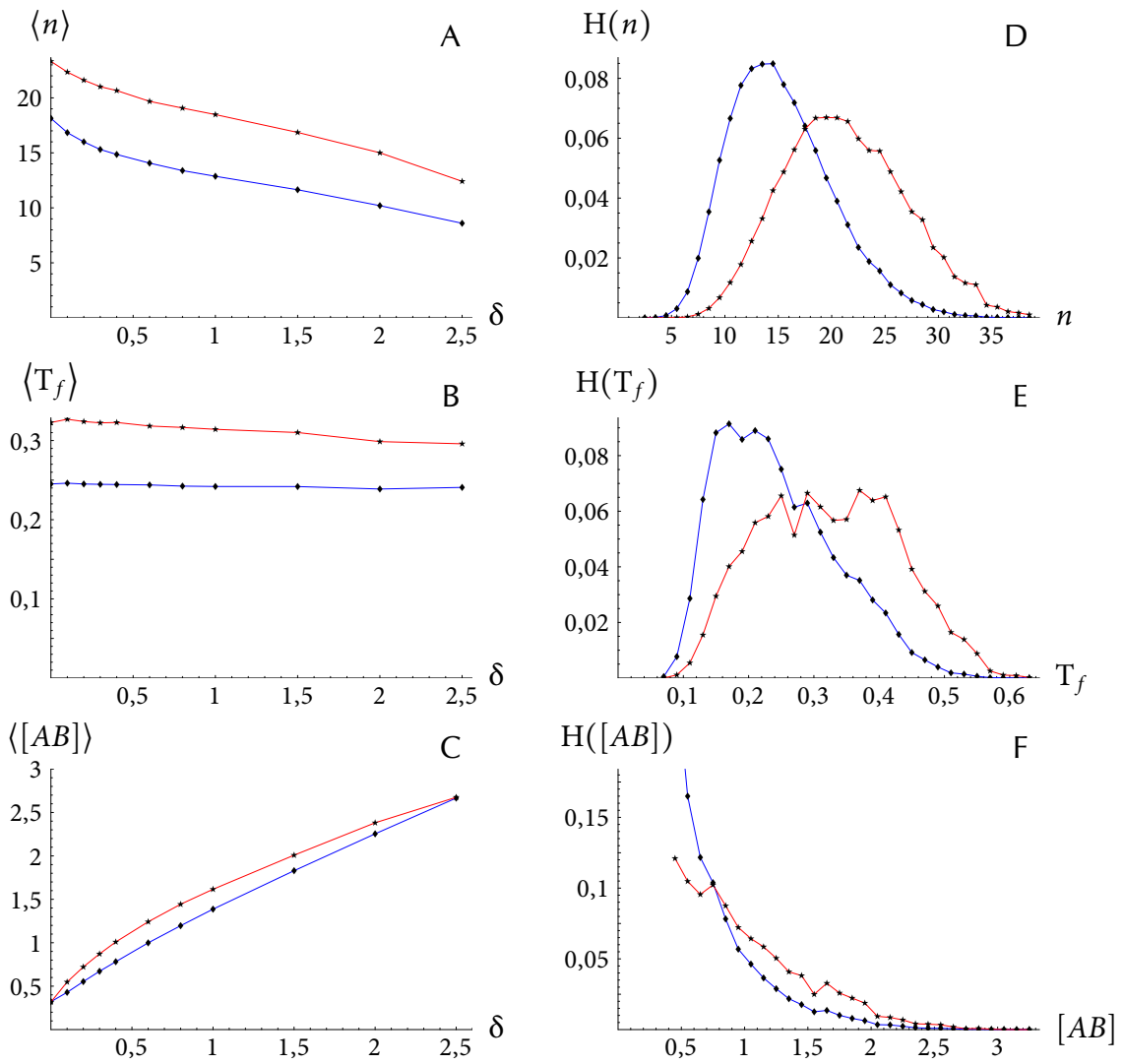


FIG. V.9 : Étude de la robustesse mutationnelle, de la température de repliement et de la concentration en dimère fonctionnel du dimère 854 (a)×136 (a). A — Robustesse mutationnelle moyenne sous une distribution uniforme (en bleu) et à l'état stationnaire (en rouge) en fonction de la contrainte fonctionnelle δ. B — Distribution de la robustesse mutationnelle sous une distribution uniforme (en bleu) et à l'état stationnaire (en rouge) pour une contrainte fonctionnelle δ = 0,4. C — Température de repliement moyenne sous une distribution uniforme (en bleu) et à l'état stationnaire (en rouge) en fonction de la contrainte fonctionnelle δ. D — Distribution de la température de repliement sous une distribution uniforme (en bleu) et à l'état stationnaire (en rouge) pour une contrainte fonctionnelle δ = 0,4. E — Concentration en dimère fonctionnel moyenne sous une distribution uniforme (en bleu) et à l'état stationnaire (en rouge) en fonction de la contrainte fonctionnelle δ. F — Distribution de la concentration en dimère fonctionnel sous une distribution uniforme (en bleu) et à l'état stationnaire (en rouge) pour une contrainte fonctionnelle δ = 0,4.

pression des génotypes non viables n'affecte pas *mécaniquement* la température de repliement : si l'on retire le voisin d'un génotype (i,j), sa température de repliement n'est pas modifiée.

(b) Là aussi, l'évolution neutre mène à une plus grande stabilité en moyenne.

3. (a) La concentration en dimère fonctionnel augmente avec la contrainte car à mesure que la contrainte fonctionnelle δ augmente, les séquences ne s'associant pas avec suffisamment d'affinité sont éliminées par sélection négative.

(b) La relation entre l'état stationnaire et la distribution uniforme est complexe : à l'état stationnaire, la concentration en dimère fonctionnel n'est pas augmentée ni en l'absence de contrainte fonctionnelle ($\delta = 0$) ni quand la contrainte est trop élevée. Pour des valeurs intermédiaires de δ, la concentration en dimère fonctionnel est augmentée par rapport à la moyenne obtenue sous une distribution uniforme.

La capacité à dimériser peut être également mesurée par le Z-score du dimère natif. Le Z-score est calculé en comparant l'énergie d'interaction du dimère natif, $\mathrm{E}(AB_1)$, à l'énergie d'interaction moyenne, $\langle \mathrm{E} \rangle$,

$$\mathrm{Z}(s) = \frac{\mathrm{E}(AB_1) - \langle \mathrm{E} \rangle}{\sigma},$$

où σ est l'écart type des énergies d'interaction.

La figure V.10 présente l'évolution du Z-score moyen et du facteur de diminution $\phi(\mathrm{Z})$ en fonction de δ. La courbe du Z-score en fonction de δ décroît en s'aplatissant. À l'état stationnaire, le Z-score est plus négatif que sous une distribution uniforme. Sauf pour de très faibles contraintes ($\delta \leq 0{,}4$) et des contraintes très fortes ($\delta \geq 2{,}7$), $\phi(\mathrm{Z})$ n'évolue quasiment pas et prend des valeurs typiques de l'ordre de 1,04-1,05. Ces valeurs sont en accord avec celles que nous trouverons pour les dimères tridimensionnels hors-réseau.

L'état stationnaire enrichit les génotypes possédant de nombreuses connexions. Par conséquent, en l'absence de contrainte fonctionnelle, l'absence d'augmentation en moyenne de la concentration en dimère fonctionnel est due à l'absence de corrélation démontrée par la figure V.5.

La table V.2 donne les valeurs de la robustesse mutationnelle du dimère 854 (a) × 136 (a) en fonction de la contrainte structurale. Pour des réseaux de taille comparable, la robustesse mutationnelle moyenne $\langle n \rangle$ se répartit équitablement entre les deux séquences composant les génotypes, sous une distribution uniforme de la population ou à l'état stationnaire.

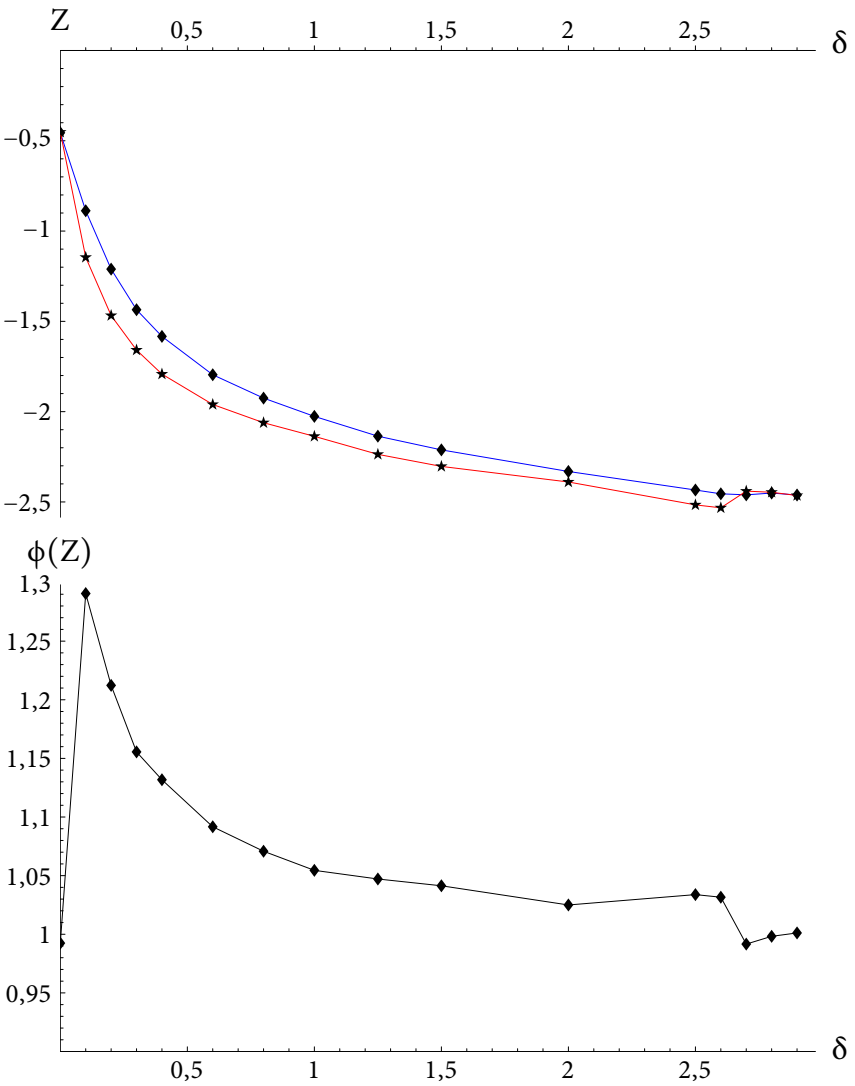


FIG. V.10 : En haut : Évolution du Z-score en fonction de la contrainte fonctionnelle δ pour les modèles de protéine sur réseau (854 (a)×136 (a)) sous une distribution uniforme (en bleu —) et à l'état stationnaire (en rouge —). En bas : Évolution de $\phi(Z)$ en fonction de δ.

V.2.1.6 Effet de la dynamique de population

Épistase Comme pour les protéines monomériques, nous mesurons l'effet de la dynamique de la population par le rapport entre les propriétés à l'état stationnaire et celles sous une distribution uniforme de la population. La température de repliement n'étant que peu affectée par la contrainte fonctionnelle δ (cf. figures V.5 et V.9), nous nous concentrons sur la robustesse mutationnelle et la concentration en dimère fonctionnel. On définit, comme pour les protéines monomériques, les enrichissements $\phi(n)$ et $\phi(\mathrm{T}_f)$:

$$\phi(n) = \langle n \rangle_s / \langle n \rangle_u ,$$
$$\phi(\mathrm{T}_f) = \langle \mathrm{T}_f \rangle_s / \langle \mathrm{T}_f \rangle_u .$$

Nous représentons dans la figure V.11, les enrichissements en fonction de la taille de l'hyperréseau neutre pour quinze dimères[28].

L'enrichissement de la robustesse mutationnelle sous l'effet de l'évolution neutre de la population est tout à fait inattendu. En effet, il est contraire à la tendance observée pour les protéines monomériques, pour lesquelles l'augmentation de la robustesse mutationnelle, $\phi(n)$, augmente avec la taille du réseau neutre (cf. figure III.33, p. 118). Dans le cas des dimères, sauf pour des valeurs élevées de δ ($\delta \geq 2{,}9$), on observe une augmentation de

δ	$\langle n \rangle_u^{\mathrm{A}}$	$\langle n \rangle_u^{\mathrm{B}}$	$\langle n \rangle_u$	$\langle n \rangle_s^{\mathrm{A}}$	$\langle n \rangle_s^{\mathrm{B}}$	$\langle n \rangle_s$
0	9,18	8,96	18,14	11,48	11,84	23,32
0,1	8,47	8,36	16,84	10,95	11,39	22,34
0,2	8,03	7,96	16,00	10,61	11,00	21,61
0,3	7,67	7,63	15,31	10,30	10,71	21,01
0,4	7,43	7,42	14,85	10,10	10,57	20,66
0,6	7,06	7,01	14,07	9,69	9,99	19,68
0,8	6,71	6,68	13,39	9,38	9,69	19,07
1	6,45	6,42	12,87	9,08	9,40	18,48
1,5	5,86	5,78	11,64	8,32	8,53	16,86
2	5,18	5,02	10,20	7,48	7,52	15,01
2,5	4,43	4,16	8,60	6,21	6,20	12,41

Tab. V.2 : Robustesse mutationnelle en fonction de la contrainte fonctionnelle δ pour le dimère 854 (a)×136 (a), sous une distribution uniforme et à l'état stationnaire. La robustesse mutationnelle peut se décomposer en deux composantes pour chaque protéine. Ces composantes sont indiquées par $\langle n \rangle^{\mathrm{A}}$ et $\langle n \rangle^{\mathrm{B}}$.

28. Les dimères utilisés sont : 143 (a)×456 (a), 1 (a)×351 (a), 28 (a)×91 (a), 30 (a)×383 (a), 337 (a)×5 (a), 350 (a)×768 (a), 388 (a)×446 (a), 419 (a)×384 (a), 422 (a)×49 (a), 506 (a)×338 (a), 509 (a)×786 (a), 605 (a)×54 (a), 612 (a)×97 (a), 627 (b)×249 (a) et 854 (a)×136 (a).

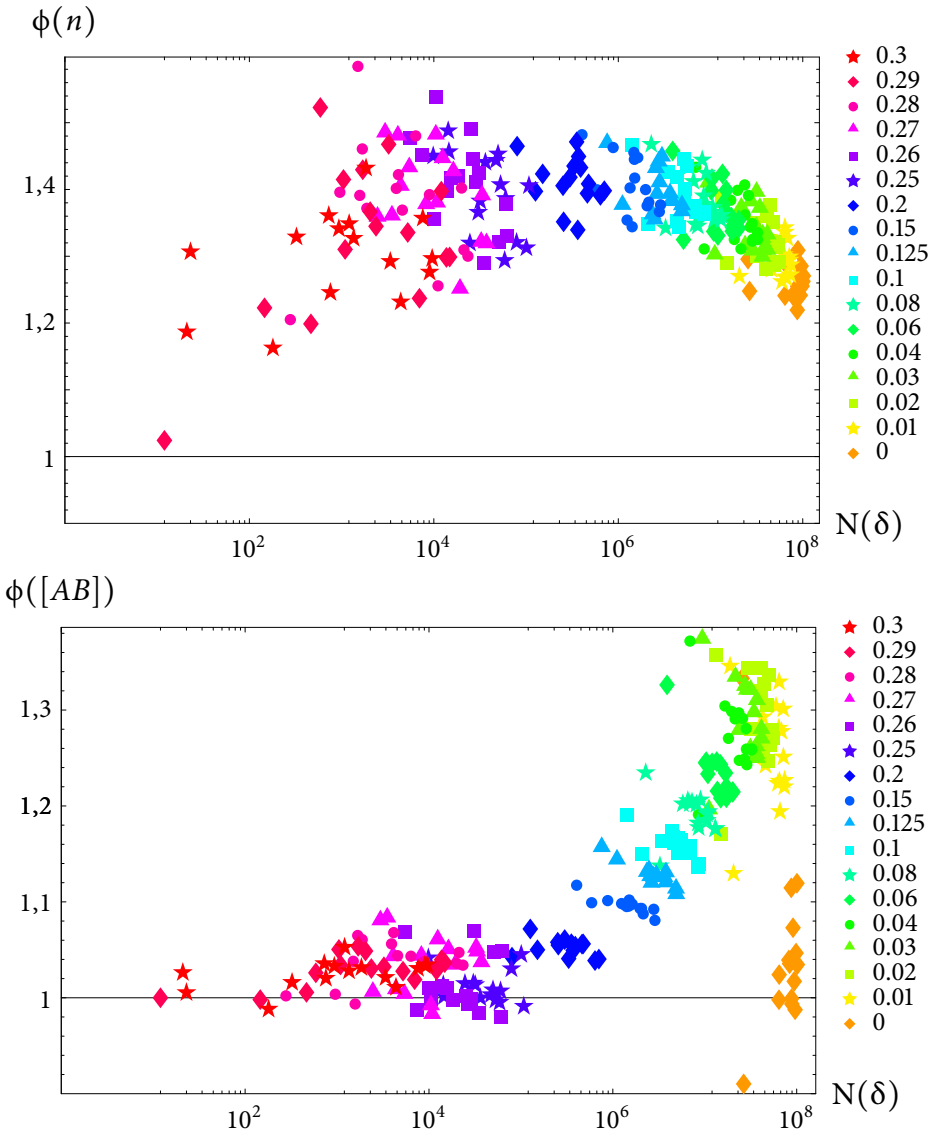


FIG. V.11 : Enrichissement de la robustesse mutationnelle, $\phi(n)$, et de la concentration en dimère fonctionnel, $\phi([AB])$, en fonction de la taille de l'hyper-réseau neutre, $N(\delta)$, pour différentes valeurs de δ représentées par le code de couleur de droite.

$\phi(n)$ à mesure que décroît la taille $N(\delta)$ de l'hyper-réseau neutre.

La valeur $\phi(n)$ en l'absence de contrainte ($\delta = 0$) est typique de celle rencontrée pour les protéines monomériques, en vertu de la relation 62 (p. 167) liant le facteur d'amélioration $\phi(n)$ du dimère à celle des protéines monomériques le constituant. On rencontre des valeurs d'enrichissement sensiblement plus élevées que pour les protéines monomériques pour des tailles équivalentes.

L'augmentation de la concentration en dimère fonctionnel sous l'effet de l'évolution neutre de la population est à l'image du schéma en arc observé précédemment dans la figure V.9C. Des valeurs proches de un sont rencontrées

1. en l'absence de contrainte ($\delta = 0$),
2. pour des valeurs élevées de δ ($\delta \geq 2{,}0$).

La relation 62 (p. 167) liant l'augmentation de la robustesse mutationnelle $\phi(n)$ aux augmentations monomériques $\phi(n)^{A}$ et $\phi(n)^{B}$ est, en réalité, plus générale que ce que nous annoncions. Elle est valable dès lors que l'ensemble des génotypes viables $\mathbf{s}^{(\delta)}$ est le produit de sous-ensembles **a** et **b** des réseaux neutres monomériques **A** et **B**, respectivement.

L'épistase est la dépendance entre les gènes : l'effet d'une mutation du génotype (i, j) dans la séquence i dépend de la séquence j. La situation que nous venons de décrire, où $\mathbf{s}^{(\delta)} = \mathbf{a} \times \mathbf{b}$, correspond à un cas *sans interaction épistatique*. En effet, une mutation du génotype (i, j) dans la séquence i est viable uniquement si la séquence mutée i^{*} est viable (appartient à **a**).

Dès lors qu'il n'y a pas d'épistase, on peut calculer une augmentation à la robustesse mutationnelle sur les sous-ensembles **a** et **b** des réseaux neutres monomériques **A** et **B**. L'augmentation de la robustesse du dimère est donnée par la formule 62. Par conséquent, en l'absence d'épistase, l'augmentation de la robustesse mutationnelle sera la moyenne arithmétique des augmentations $\phi(n)^{A}$ et $\phi(n)^{B}$ de chaque réseau neutre **a** et **b** et sera comprise entre elles :

$$\min\{\phi(n)^{A}, \phi(n)^{B}\} \leq \phi(n) \leq \max\{\phi(n)^{A}, \phi(n)^{B}\}. \tag{74}$$

La croissance de $\phi(n)$, alors que la taille de l'hyper-réseau neutre $N(\delta)$ décroît, ne peut donc s'expliquer que par la dépendance entre les gènes, c'est-à-dire l'épistase.

Profils moyens Dans l'étude des protéines monomériques, nous avons montré que le profil moyen, sous l'effet de l'évolution neutre de la population, se rapprochait de la séquence prototype. L'effet est plus sensible encore pour les dimères. La figure V.12 présente

les profils de 786 (a) formant un dimère avec 509 (a). Les positions à l'interface sont les positions 6, 7 et 10. L'effet de l'évolution neutre sur ces positions est particulièrement forte.

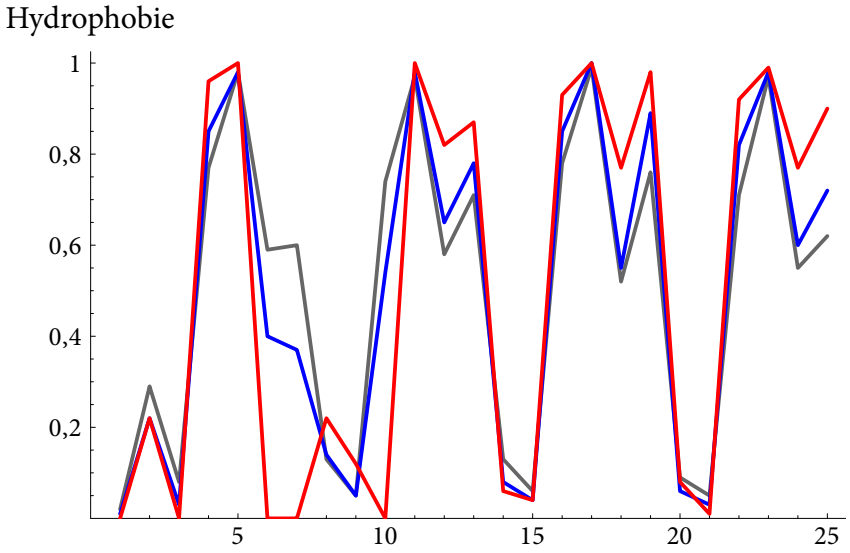


Fig. V.12 : Profils moyen de 786 (a) dans son interaction avec 509 (a). En gris, est rappelé le profil moyen monomérique. Les profils pour une contrainte fonctionnelle $\delta = 2{,}5$ sont en bleu (distribution uniforme) et en rouge (à l'état stationnaire).

V.2.1.7 *Superfunnel* fonctionnel

L'augmentation de la concentration moyenne en dimère fonctionnel, tout du moins pour certaines valeurs de δ, suggère que les séquences possédant une capacité à dimériser s'organisent en « *superfunnel* fonctionnel ». Par cette expression, nous entendons que la concentration en dimère fonctionnel d'un génotype doit être essentiellement déterminée par sa distance à un « génotype prototype » : le génotype le plus peuplé à l'état stationnaire. La figure V.13 montre qu'effectivement une organisation de type *superfunnel* existe sauf lorsque $\delta = 0$. Pour des fortes contraintes fonctionnelles, la forme est très aplatie.

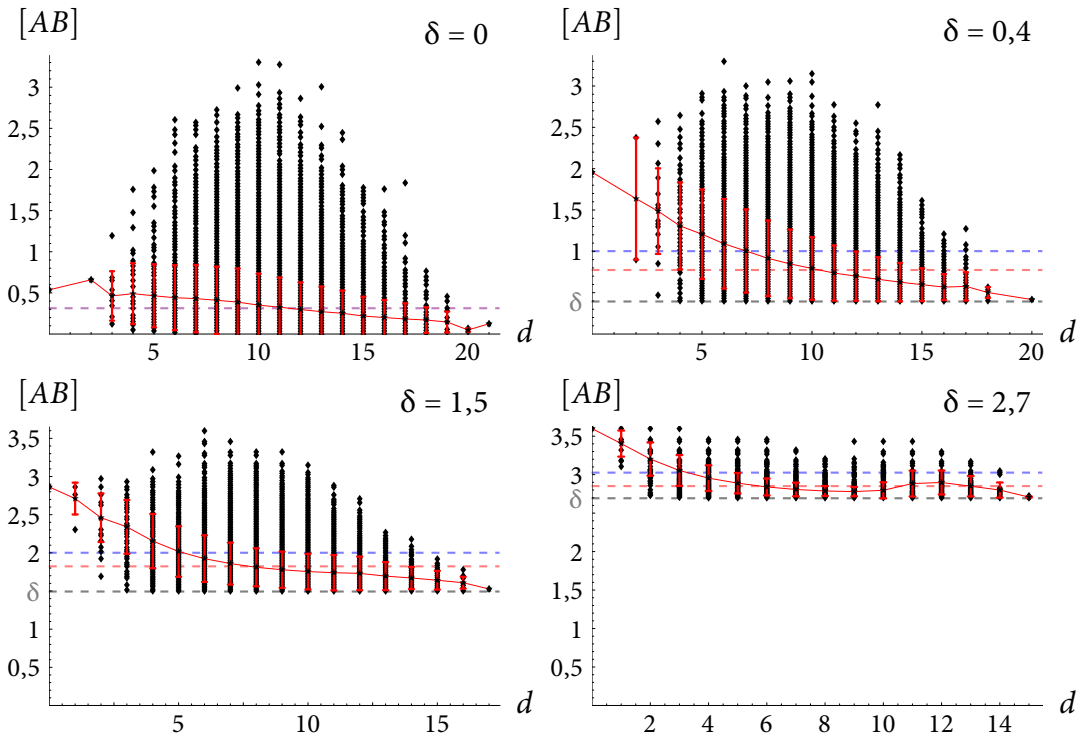


FIG. V.13 : Concentration en dimère fonctionnel $[AB]$ en fonction de la distance de Hamming d au génotype prototype (génotype le plus peuplé) pour quatre valeurs de δ : 0, 0,4, 1,5 et 2,7. Les données sont représentées par les points noirs ; les moyennes et les écarts types sont en rouge. Les lignes pointillées bleue et rouge figurent les moyennes sous une distribution uniforme et à l'état stationnaire respectivement.

V.2.2 Protéines tridimensionnelles

Lors de l'étude monomérique, nous nous sommes restreints à l'espace des profils pour construire les réseaux neutres des protéines tridimensionnelles hors-réseau. Les profils ne sont pas utilisables tel quel dans le calcul des énergies d'interaction, car ils ne disposent plus de l'information structurale nécessaire à l'établissement des contacts entre les acides aminés. Il nous faut donc revenir aux séquences générées lors de la trajectoire. Une séquence par trajectoire est donc retenue pour *représenter* chaque profil HP.

Le réseau de Grb2 contient 31 469 profils[29] et celui de Vav en contient 29 667. Ont été examinées $31\,469 \times 29\,667 = 933 \cdot 10^6$ paires de séquences. Nous avons retenu celles qui permettaient au dimère natif d'être classé dans les dix premiers dimères par ordre d'énergie croissante. En effet, si nous faisons l'hypothèse que l'activité biologique nécessite la formation du dimère dix pour cent du temps, une condition nécessaire à remplir est qu'il existe moins de dix fausses structures ayant une énergie plus basse que le dimère fonctionnel.

Une fois les génotypes viables identifiés, ils sont reconvertis en profils. L'hyper-réseau neutre est alors construit suivant les mêmes règles que pour les dimères de protéines sur réseau : une connexion est établie entre deux génotypes (i, j) et (k, l) si une et une seule mutation intervient dans les deux profils.

V.2.2.1 Statistiques de viabilité

Nous avons dénombré 470 334 paires de séquences remplissant la condition de viabilité, soit une fraction $5 \cdot 10^{-4}$ des génotypes compatibles avec les repliements de Grb2 et Vav. Si l'on se restreint à des rangs k inférieurs à 10, le nombre de génotypes viables est encore plus faible (cf. tableau V.3). En réduisant le rang maximal autorisé pour le dimère fonctionnel, k, par ordre d'énergie d'interaction croissante, on augmente la contrainte. Une diminution de k correspond à une augmentation de la contrainte δ des dimères de protéines sur réseau.

V.2.2.2 Propriétés à l'état stationnaire

Contrairement aux hyper-réseaux neutres des dimères de protéines sur réseau, les plus grandes composantes connexes ne rassemblent pas la majorité des génotypes. Nous

29. Le tableau III.3, p. 108 indique que le réseau comporte 20 840 séquences. Pour la simulation monomérique, seule une partie de la trajectoire a été utilisée pour générer le graphe des profils, au moment où le nombre de profils semblait converger. Les 10 000 autres profils résultent de la découverte d'un vaste bassin de séquences qui n'a pas eu le temps de converger dans le temps de la simulation. Ils ont par conséquent été exclus. En revanche, dans le cadre de la dimérisation, il nous a semblé important de tirer profit de la variabilité de séquences qu'offrait ce bassin. Les caractéristiques ne sont pas énormément modifiées : de $\langle n \rangle_u = 5{,}59$, $\langle n \rangle_u = 10{,}97$ et $\phi(n) = 1{,}96$, on passe à $\langle n \rangle_u = 5{,}79$, $\langle n \rangle_u = 11{,}49$ et $\phi(n) = 1{,}98$.

introduisons donc une nouvelle notation : la moyenne $\langle x \rangle_r$ calculée en distribuant uniformément la population sur la composante connexe la plus importante. En effet, l'état stationnaire ne peuple que celle-ci et la comparaison entre une distribution parfaitement uniforme et l'état stationnaire n'a que peu de sens et biaise les facteurs d'amélioration. Nous définissons le facteur d'augmentation associé, $\phi_r(x)$,

$$\phi_r(x) = \frac{\langle x \rangle_s}{\langle x \rangle_r}. \tag{75}$$

Nous avons souhaité examiner ces résultats, plus précisément, en fonction du rang k maximal autorisé. Nous avons constitué plusieurs ensembles de séquences : $\mathbf{s}_1$, l'ensemble des paires de séquences classant premier le dimère natif, $\mathbf{s}_2$, l'ensemble des paires de séquence classant second ou mieux le dimère natif, etc. Chaque ensemble $\mathbf{s}_k$ a été étudié séparément : construction de l'hyper-réseau neutre, calcul de l'état stationnaire et comparaison des valeurs moyennes.

Nous avons comparé trois grandeurs.

1. la robustesse mutationnelle n,
2. le Z-score (calculé à partir du spectre des énergies d'interaction),
3. le rang moyen.

Le détail des résultats, la reconstruction des graphes et le calcul des états stationnaires pour les génotypes viables selon leur rang, sont présentés dans le tableau V.4.

Robustesse mutationnelle La distribution de la robustesse mutationnelle moyenne $\langle n \rangle$, sous une distribution uniforme et à l'état stationnaire, est présentée dans la figure V.14 pour les 470 334 séquences sélectionnées ($k = 10$). Ces distributions possèdent la forme

Rang k	*Fréquence*	*Fréquence cumulée*
1	537	537
2	30 264	30 801
3	36 490	67 291
4	42 663	109 954
5	47 949	157 903
6	53 230	211 133
7	57 889	269 022
8	62 830	331 852
9	66 924	398 776
10	71 558	470 334

Tab. V.3 : Fréquence d'apparition de paires de séquences classant k-ième par ordre d'énergie croissante.

classique déjà rencontrée pour les monomères et les dimères de protéines sur réseau. Le biais de population vers les génotypes fortement connectés, dans la distribution à l'état stationnaire, est très prononcé. Les augmentations de la robustesse $\phi(n)$ et $\phi_r(n)$ valent 3,00 et 2,23, respectivement.

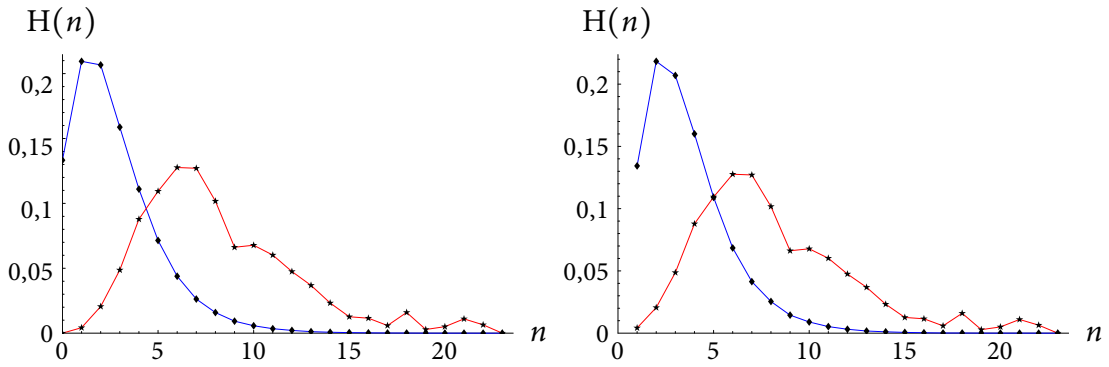


Fig. V.14 : Robustesse mutationnelle n du dimère Grb2 et Vav. Figure de gauche — Comparaison entre la distribution uniforme des génotypes (en bleu —) et l'état stationnaire (en rouge —). Figure de droite — Comparaison entre la distribution uniforme des génotypes appartenant à la plus grande composante connexe (en bleu —) et l'état stationnaire (en rouge —).

Aptitude à dimériser L'aptitude à dimériser augmente sous l'effet de l'évolution neutre : le Z-score est plus grand, en valeur absolue, à l'état stationnaire que sous une distribution uniforme. Les facteurs d'augmentation du Z-score sont du même ordre de grandeur que ceux observés pour les protéines sur réseau (cf. figure V.10). De manière similaire, le rang moyen est diminué à l'état stationnaire par rapport à une distribution uniforme.

Effet de l'épistase Étant donné la taille de la composante connexe majeure, il est probable que les résultats obtenus pour les rangs $k = 1$ et 2 ne soient pas exploitables. Pour $k > 2$, nous remarquons que l'effet de la dynamique de la population déjà observé pour les dimères sur réseau est restauré. En effet, l'amélioration $\phi(n)$ pour l'ensemble de l'hyper-réseau neutre vaut d'après la formule 62 (p. 167) :

$$\phi(n) = \frac{5{,}6 \times 1{,}96 + 6{,}1 \times 1{,}83}{5{,}6 + 6{,}1} = 1{,}89.$$

Les valeurs obtenues lorsque la contrainte dimérique est appliquée varient entre 1,93 et 2,23 et augmentent à mesure que la contrainte diminue (quand k augmente). C'est-à-dire que l'épistase se manifeste selon la même loi que celle observée pour les dimères sur réseau pour des valeurs élevées de δ.

Rang	*Nombre de génotypes viables*[*]	*Plus grande composante connexe*[†]	*Robustesse mutationnelle*[‡]	*Z-score*[‡]	*Rang moyen*[‡]
1	537 421 177	4 (0,74 %)	0,14 2 2 14,13 1	−2,89 −3,03 −3,02 1,05 1,00	1 1 1 1 1
2	30 801 5 391 702	1 809 (5,87 %)	1,09 2,45 5,09 4,65 2,08	−2,89 −2,97 −3,10 1,07 1,04	1,98 1,97 2,00 1,01 1,01
3	67 291 8 550 961	16 671 (24,77 %)	1,57 2,73 5,49 3,50 2,01	−2,82 −2,84 −3,05 1,08 1,07	2,53 2,47 2,18 0,86 0,89
4	109 954 11 396 1 207	37 484 (34,09 %)	1,88 2,99 5,77 3,07 1,93	−2,76 −2,77 −3,26 1,18 1,18	3,10 2,97 2,82 0,91 0,95

Tab. V.4 : Calcul des réseaux neutres dimériques et des états stationnaires correspondants. Nous présentons les résultats obtenus en nous restreignant aux séquences classant le dimère natif dans les k premiers par ordre d'énergie croissante, avec k allant de 1 à 10. Nombre de génotypes viables (*) : trois nombres sont indiqués, de haut en bas, 1. le nombre de génotypes (i, j) viables, 2. le nombre de séquences viables pour Grb2, 3. le nombre de séquences viables pour Vav. Plus grande composante connexe (†) : nombre de génotypes dans la plus grande composante connexe et le pourcentage de l'ensemble des génotypes viables qu'elle représente. Robustesse mutationnelle, Z-score et rang moyen (‡) : sont donnés cinq nombres, de haut en bas, 1. $\langle x \rangle_u$ (la moyenne de x supposant une distribution uniforme des génotypes), 2. $\langle x \rangle_r$ (la moyenne de x supposant une distribution uniforme des génotypes appartenant à la plus grande composante connexe), 3. $\langle x \rangle_s$ (la moyenne de x à l'état stationnaire), et les enrichissements 4. $\phi(x) = \langle x \rangle_s / \langle x \rangle_u$ et enfin 5. $\phi_r(x) = \langle x \rangle_s / \langle x \rangle_r$.

Rang	*Nombre de génotypes viables**	*Plus grande composante connexe†*	*Robustesse mutationnelle‡*	*Z-score‡*	*Rang moyen‡*
5	157 903 13 900 1 397	60 794 (38,50 %)	2,10 3,18 6,43 3,07 2,02	−2,71 −2,72 −3,19 1,18 1,17	3,68 3,48 3,21 0,87 0,92
6	211 133 15 988 1 595	84 347 (39,95 %)	2,27 3,34 6,82 3,01 2,04	−2,66 −2,67 −2,82 1,06 1,06	4,26 4,00 3,63 0,85 0,91
7	269 022 17 704 1 805	121 498 (45,16 %)	2,41 3,41 7,21 3,00 2,11	−2,62 −2,63 −2,79 1,07 1,06	4,85 4,54 3,97 0,82 0,87
8	331 852 19 313 1 996	155 946 (46,99 %)	2,52 3,51 7,48 2,96 2,13	−2,58 −2,59 −2,76 1,07 1,07	5,45 5,09 4,30 0,79 0,84
9	398 776 20 693 2 187	194 442 (48,76 %)	2,62 3,59 7,85 2,99 2,18	−2,54 −2,56 −2,64 1,04 1,03	6,04 5,64 5,07 0,84 0,90
10	470 334 21 877 2 386	239 960 (51,02 %)	2,72 3,66 8,16 3,00 2,23	−2,51 −2,53 −2,61 1,04 1,03	6,65 6,22 5,47 0,82 0,88

Tab. V.4 : (Suite.)

V.2.2.3 Localisation des génotypes viables

L'un des résultats les plus frappants de l'étude des protéines sur réseau est l'indépendance entre la robustesse mutationnelle et la fonctionnalité (cf. figure V.5). Ce résultat indique notamment que la fonctionnalité est délocalisée géographiquement au sein de l'hyper-réseau neutre et des réseaux neutres monomériques.

Nous examinons sa validité dans le cadre du modèle de dimère de protéines tridimensionnelles hors-réseau. Pour chaque génotype viable, nous calculons le nombre de voisins n viables et non viables, c'est-à-dire le nombre de connexions dans l'hyper-réseau neutre complet. En d'autres termes, c'est le nombre de connexions tel qu'il est représenté dans la figure IV.1C.

La distribution de n contient la même information que le graphique de la figure V.5A. Une représentation telle que celle de la figure V.5A n'est cependant pas possible, car nous n'avons pas les moyens de calculer la concentration du dimère fonctionnel.

Si la conclusion tirée du modèle sur réseau est encore valable les distributions devraient se confondre avec la distribution de la robustesse mutationnelle de l'ensemble des $9 \cdot 10^8$ génotypes. Les distributions (sauf celle correspondant à $k = 1$) sont présentées dans la figure V.15. On ne constate pas d'écart important entre les distributions du nombre de connexions dans l'hyper-réseau neutre complet pour les génotypes classant k-ième le dimère fonctionnel.

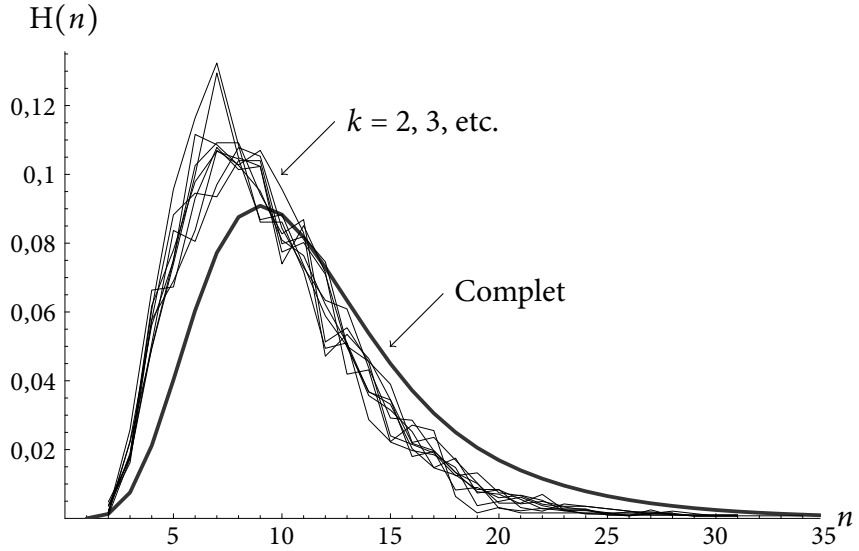


Fig. V.15 : Distribution des $n_r^A + n_s^B$ pour les génotypes (r, s) classant k-ième le dimère natif, pour $k = 2, 3$, etc. « Complet » correspond à la distribution attendue pour l'ensemble des génotypes ($k = \infty$).

L'écart entre la distribution complète et les distributions obtenues pour les génotypes viables suggère que les génotypes viables ne sont pas dispersés uniformément dans l'hyper-

réseau neutre des génotypes compatibles avec les repliements de Grb2 et Vav. Cependant, certaines génotypes viables possèdent malgré tout un nombre de connexions important et l'écart entre les distributions pour $k = 2 \ldots 10$ et la distribution complète reste relativement faible. La table V.5 complète les données de la figure V.15 en précisant les moyennes et écarts types correspondant aux distributions précitées. Elle confirme l'existence d'un écart entre les différentes distributions.

Rang k	$\langle n \rangle$	σ_n
1	9,153 85	2,178 43
2	8,889 95	3,738 97
3	9,373 61	3,980 81
4	9,477 94	4,098 26
5	9,501 38	4,198 67
6	9,655 09	4,248 99
7	9,519 1	4,364 76
8	9,534 22	4,148 83
9	9,729 77	4,286 39
10	9,755 87	4,306 21
	11,919 9	5,604 55

Tab. V.5 : Moyennes et écarts types des tolérances aux mutations au sein de l'hypergraphe total pour les ensembles de séquences classant k-ième le dimère natif. La dernière ligne correspond à la distribution attendue pour l'hypergraphe complet.

L'écart observé tient en grande partie au faible échantillonnage du réseau neutre monomériques de Vav. En effet, seules 2 386 séquences du réseau neutre de Vav sont viables (8 %), tandis que 21 877 séquences du réseau neutre de Grb2 le sont (70 %). Les séquences viables de Vav sont plus localisées que les séquences viables de Grb2. Elles se situent dans une région du réseau monomérique possédant une robustesse mutationnelle réduite.

Le choix des faux dimères peut également être responsable de l'écart entre les distributions. Les faux dimères qui possédaient une énergie plus faible que le dimère fonctionnel, avec les séquences de référence, ont été exclus lors de la recherche des génotypes viables. On peut donc penser que l'exclusion de ces faux dimères favorise l'acceptation des génotypes proches du génotype composé des séquences de référence. Or, lors des trajectoires générant les séquences des réseaux neutres monomériques, ces mêmes séquences de référence ont servi à fixer la limite basse de stabilité. Durant les trajectoires, toute séquence plus stable dans la conformation native est viable. Il ne peut donc pas exister de séquence viable plus instable que la séquence de référence. En d'autres termes, la séquence de référence est

la séquence viable la plus instable possible[30]. Par conséquent, si l'exclusion spécifique des faux dimères avantage les séquences proches de la séquence prototype, alors les séquences acceptées auront sans doute une stabilité réduite. Une autre démonstration est fournie par la figure V.16.

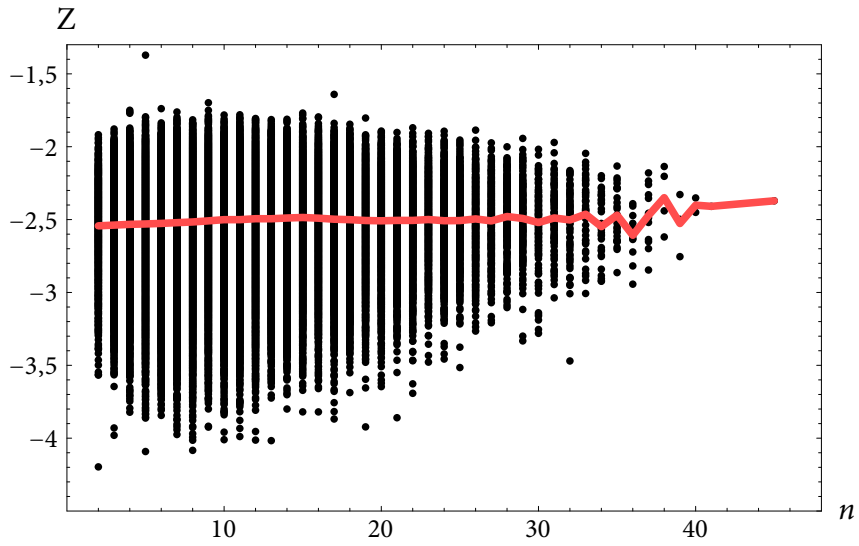


Fig. V.16 : Le Z-score dimérique — c'est-à-dire des énergies d'association — en fonction du nombre de connexions dans l'hyper-réseau neutre complet n. La moyenne est représentée par la ligne rouge en fonction de n.

V.2.2.4 *Superfunnel* fonctionnel

La structure de *superfunnel* fonctionnel est plus difficile à évaluer par le Z-score. Elle n'est clairement visible que pour les ensembles s_4 (voir figure V.17) et s_5. Pour ces ensembles de séquences viables, on observe, autour du génotype prototype, un bassin de génotypes dont le dimère fonctionnel est plus stable.

V.2.2.5 Conservation des résidus

Nous avons classé les résidus en trois catégories à partir de la structure cristallographique.

1. Les *résidus enfouis* sont ceux qui présentent une surface accessible au solvant (ASA) inférieure à 10 % de leur ASA de référence calculée à partir de tripeptides GXG et d'une sphère d'eau de rayon 1,6 Å (161, 163, 167, 175, 177, 183, 185, 196, 203, 205 pour Grb2, 599, 603, 605, 609, 617, 619, 623, 625, 627, 636, 639, 641, 648, 650, 651, 654 pour Vav).

30. Pour cette raison, d'ailleurs, on s'attend à ce que les séquences prototypes soient en périphérie de leur réseau neutre monomérique et non au centre.

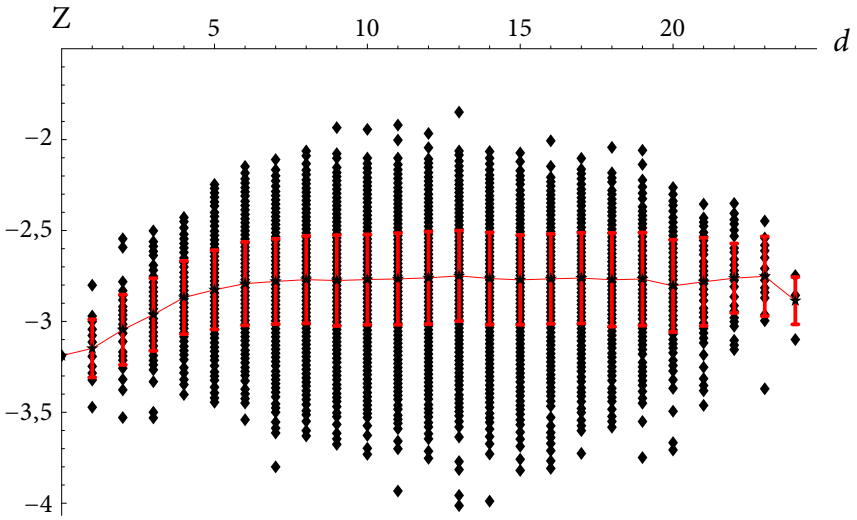


Fig. V.17 : Structure de *superfunnel* fonctionnel autour du génotype, c'est-à-dire de la paire de séquences la plus peuplée à l'état stationnaire calculé sur l'hyper-réseau neutre s_4. La capacité à dimériser est mesurée par le Z-score de l'association native par rapport aux fausses structures. Les moyennes et les écarts types sont représentés en rouge.

2. Les résidus à l'interface sont ceux qui perdent plus de 1 % de leur ASA de référence lors de l'interaction des deux protéines (194, 210, 162, 164, 165, 170, 179, 180, 190, 206, 207, 208, 209, 211, 212, 215 pour Grb2, 597, 655, 592, 593, 594, 595, 596, 628, 630, 631, 632, 633, 637, 652, 656, 657, 658, 659 pour Vav).
3. Les résidus exposés sont les résidus n'appartenant à aucune des deux catégories précédentes (159, 160, 166, 168, 169, 171, 172, 173, 174, 176, 178, 181, 182, 184, 186, 187, 188, 189, 191, 192, 193, 195, 197, 198, 199, 200, 201, 202, 204, 213, 214 pour Grb2, 591, 598, 600, 601, 602, 604, 606, 607, 608, 610, 611, 612, 613, 614, 615, 616, 618, 620, 621, 622, 624, 626, 629, 634, 635, 638, 640, 642, 643, 644, 645, 646, 647, 649, 653 pour Vav).

Voir également la table V.6.

Chaîne	*Résidus à l'interface*		*Résidus non impliqués*	
	superficiels	*enfouis*	*superficiels*	*enfouis*
B	14	2	31	10
C	16	2	35	16

Tab. V.6 : Nombre de résidus exposés ou enfouis, impliqués ou non dans l'interface Grb2-Vav.

Nous avons analysé la conservation des différentes catégories en utilisant pour mesure l'entropie de Shannon. Pour chaque catégorie G (« enfoui », « interface », « autre »), nous

définissons l'entropie de Shannon moyenne *par résidu* :

$$s(\mathrm{G}) = -\frac{1}{|\mathrm{G}|}\sum_{i\in\mathrm{G}} p_i \log_2 p_i + (1-p_i)\log_2(1-p_i),$$

où p_i désigne la proportion de résidus polaires observés à ladite position et $|\mathrm{G}|$ le nombre de résidus de la catégorie. (La fonction $f(x) = x\log_2 x$ est prolongée par continuité en 0.) La grandeur $s(\mathrm{G})$ peut varier de 0 — dans le cas de résidus parfaitement conservés — jusqu'à 1 — dans le cas d'une absence de préférence pour l'un des résidus. Les résultats numériques sont synthétisés dans la table V.7 et les profils sont présentés dans la figure V.18.

Conformément aux observations expérimentales, les positions dans le cœur des protéines sont sensiblement plus conservées, en particulier à l'état stationnaire. L'état stationnaire a tendance à augmenter la conservation en profil HP. En revanche, les résidus à l'interface ne semblent pas plus conservés que le reste de la surface. Dans le cas de la chaîne Vav, les résidus de l'interface sont même moins conservés en moyenne que les autres résidus de surface (l'écart type est très grand cependant). Ce résultat est en accord avec l'étude de Caffrey *et al.* [26] qui démontre que la conservation des résidus d'interface est loin d'être très prononcée (cf. la figure 1 de l'article cité). Les profils de la figure V.18 montrent également qu'il n'existe pas de « signature » de l'interface : le schéma d'hydrophobie des résidus de l'interface est similaire à celui des autre résidus de surface. Les profils indiquent que l'interface n'est pas plus hydrophobe, en particulier dans le cas de Vav, que le reste de la surface. Les profils moyens sont très similaires avec et sans la contrainte dimérique. En revanche, les résidus enfouis sont très conservés et très majoritairement hydrophobes, même si quelques résidus, conservés eux aussi, sont hydrophiles [90]. Nous sous-estimons le nombre de résidus hydrophiles enfouis car la procédure de préoptimisation les a, en grande partie, éliminés.

Chaîne		*Monomère*		*Dimère*	
		Distribution uniforme	*État stationnaire*	*Distribution uniforme*	*État stationnaire*
B	*Interface*	0,45 (0,37)	0,28 (0,27)	0,49 (0,43)	0,21 (0,30)
	Résidus exposés	0,54 (0,42)	0,38 (0,39)	0,50 (0,40)	0,21 (0,30)
	Résidus enfouis	0,10 (0,17)	0,04 (0,09)	0,14 (0,28)	0,01 (0,03)
C	*Interface*	0,48 (0,40)	0,36 (0,34)	0,40 (0,40)	0,16 (0,26)
	Résidus exposés	0,33 (0,40)	0,26 (0,36)	0,27 (0,35)	0,10 (0,28)
	Résidus enfouis	0,16 (0,24)	0,07 (0,13)	0,09 (0,19)	0,02 (0,06)

Tab. V.7 : Conservation des résidus mesurée par l'entropie de Shannon par résidu dans les chaînes Grb2 et Vav selon les trois catégories : résidus à l'interface, résidus exposés mais non impliqués dans l'interface, résidus enfouis. Les résultats monomériques et dimériques sont présentés, en supposant soit une utilisation uniforme des génotypes viables, soit une utilisation biaisée par l'état stationnaire. Les nombres entre parenthèses indiquent les écarts types.

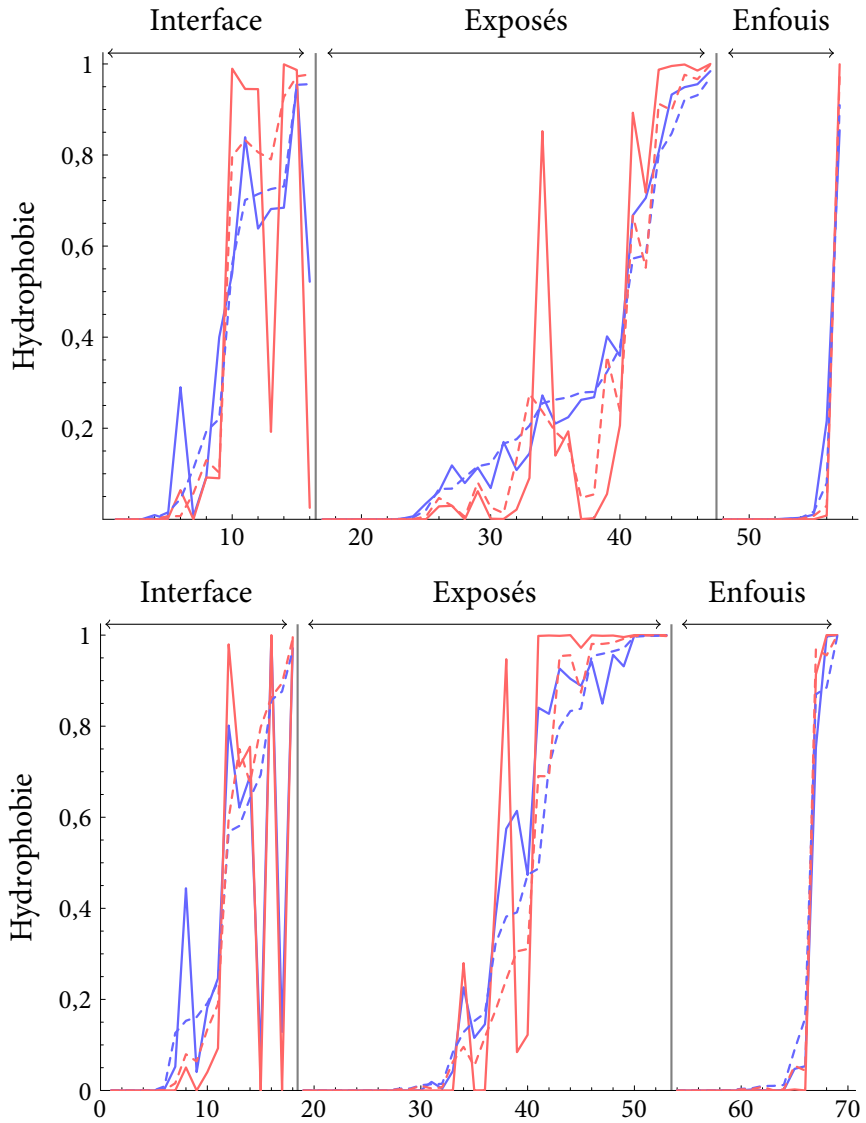


FIG. V.18 : Profils HP moyens pour les chaînes Grb2 (en haut) et Vav (en bas). Les profils monomériques sont inclus pour comparaison en pointillés (bleu : distribution uniforme ⋯, rouge : état stationnaire ⋯). Les profils continus sont ceux des dimères sélectionnés au rang $k = 10$ (bleu : distribution uniforme —, rouge : état stationnaire —).

V.3 Discussion

Nous avons présenté dans ce chapitre deux modèles d'évolution de dimères. Ils diffèrent en de nombreux points.

1. L'un est un modèle de dimères de protéines sur réseau bidimensionnel, l'autre de protéines tridimensionnelles hors-réseau. Le premier simule l'évolution d'une petite interface formées par cinq résidus sur chaque monomère, le second celle d'une interface engageant plusieurs dizaines d'acides aminés au total (seize et dix-huit pour Grb2 et Vav respectivement, selon le critère fondé sur la perte d'ASA, cf. p. 203).
2. La sélection est fondée dans un cas sur la concentration en dimère fonctionnel, l'autre sur le rang du dimère fonctionnel par ordre d'énergie d'interaction croissante.
3. Les matrices d'énergie et les méthodes pour générer les séquences appartenant aux réseaux neutres monomériques sont très différentes.

En dépit des différences qui séparent les modèles, les résultats obtenus partagent de nombreuses caractéristiques équivalentes, et les principaux résultats des deux modèles sont en bon accord.

V.3.1 Dimérisation comme modèle de fonctionnalité

V.3.1.1 Une fonctionnalité « non structurale »

La fonctionnalité d'une protéine ou d'un complexe est difficile à modéliser. Dans les fonctions enzymatiques, certains résidus sont absolument nécessaires pour leur fonctions chimiques ; dans les fonctions de signalisation, des changements conformationnels surviennent. Ces éléments ne sont pris en compte dans aucun modèle de fonctionnalité. Néanmoins, l'ubiquité des interactions protéine-protéine fait de la dimérisation un candidat naturel pour la fonctionnalité.

Les interactions protéine-protéine présentent l'avantage de libérer le modélisateur de considérations purement structurales et possèdent des points en commun avec les fonctions enzymatiques. Les fonctions enzymatiques reposent sur le positionnement précis de certains résidus [72, 128]. Avec la modélisation des interactions protéine-protéine, nous reproduisons deux aspects du fonctionnement d'une enzyme : une charpente, et le positionnement de résidus (à l'interface). Il n'est donc pas étonnant que la fixation de ligand soit centrale dans la modélisation de la fonctionnalité qui s'est développée dans les dernières années (voir, par exemple, référence [196]). Blackburne et Hirst n'utilisent pas de

ligand mais font l'hypothèse que la fonctionnalité coïncide avec la présence d'une poche censée accueillir un ligand hydrophobe [12–14]. Dans notre thèse, le modèle structural définissant la charpente correspond à la construction des réseaux neutres monomériques. Le placement des résidus importants pour la fonction (la dimérisation) n'arrive qu'en second lieu. L'absence de corrélation entre la stabilité d'un génotype et son aptitude à dimériser, conforte l'idée que la fonctionnalité est « orthogonale » au modèle structural (cf. figure V.5).

Une des limites de notre modèle dans sa capacité à décrire correctement la fonctionnalité est la simplicité des alphabets utilisés. Pour les protéines sur réseau comme pour celles hors-réseau, un potentiel de type HP a été utilisé. Cette simplification de l'alphabet explique qu'on ne puisse atteindre de très hautes concentrations en dimère fonctionnel avec le modèle sur réseau. De fait, les plus hautes concentrations sont atteintes lorsque à l'interface quasiment tous les résidus sont hydrophobes. Il est, par conséquent, inévitable que des homodimères possédant une bonne affinité rivalisent avec le dimère fonctionnel. Un alphabet plus complet pourrait donner lieu à des concentrations plus élevées en disposant, par exemple, des résidus chargés complémentaires sur les deux interfaces des monomères.

V.3.1.2 Importance de la sélection négative

Les génotypes capables de former un complexe stable sont rares. Comme le montre notamment la distribution de la concentration en dimère fonctionnel (figure V.1), l'essentiel des paires de séquences ne possèdent qu'une faible aptitude à dimériser. Les génotypes pour lesquels la concentration en dimère fonctionnel dépasse 2,0 représentent seulement 0,5 % de tous les génotypes dont les structures des monomères sont stables (cf. tableau V.1). La fraction de génotypes classant premier le dimère fonctionnel (protéines hors-réseau) est de $5 \cdot 10^{-7}$. Si les énergies d'interaction étaient aléatoires, cette fraction devrait être l'inverse du nombre de faux dimères, soit environ 10^{-3}.

Dans notre modèle, les monomères doivent se replier avant d'interagir. Cette condition impose une ségrégation des résidus hydrophobes dans le cœur de la protéine et polaires en surface. Les résidus polaires en surface rendent difficile l'association protéine-protéine, particulièrement dans le modèle de protéine sur réseau car, pour ce dernier, l'interaction P-P est la seule à ne pas être favorable.

L'absence de complémentarité structurale est un autre aspect qui est négligé dans notre étude des protéines sur réseau : les interfaces sont simplifiées au maximum. En conséquence, il n'existe pas dans ce modèle de structure préférentielle plus apte à former des interactions qu'une autre. En effet, le filtrage diminue par un facteur 0,5 % la taille des hyper-réseaux neutres, indépendamment de la structure des partenaires. Cependant, le résultat du filtrage suggère qu'il existe plus de séquences aptes à dimériser si la structure

est très modifiable. Une structure modifiable s'accomode de plus de mutations et en particulier de plus de résidus hydrophobes à l'interface. Les résultats du filtrage sont cohérents avec ceux de Williams *et al.*. Ces auteurs ont comparé la représentation des structures en utilisant un critère structural puis un critère fonctionnel (liaison à un ligand). La représentation n'est pas altérée de façon critique par le critère fonctionnel [196].

V.3.2 Propriétés des génotypes viables

V.3.2.1 Diversité et connectivité des génotypes viables

Nous venons de voir que la contrainte fonctionnelle est draconienne : rares sont les génotypes qui sont viables. Où sont donc situés les séquences et les génotypes viables dans les réseaux neutres monomériques et dimériques ? L'aptitude à dimériser apparaît comme une propriété orthogonale aux considérations structurales habituelles. La localisation des séquences viables n'est pas cantonnée à une partie restreinte du réseau neutre. Les figures V.4, V.5 et V.8 pour les protéines sur réseau indiquent une vaste répartition des séquences viables dans les réseaux neutres monomériques. La figure V.15 donne la même indication pour les protéines hors-réseau. Par ailleurs, il est remarquable que, dans le réseau neutre reconstruit à partir des génotypes classant dixième ou mieux le dimère natif, 70 % des séquences du réseau neutre monomérique de Grb2 soient représentés bien qu'une infime fraction des paires de séquences soit viable (0,05 %).

Dans le même temps, les génotypes viables ne sont pas dispersés au point de ne plus former de composante connexe. L'évolution des dimères peut donc procéder comme Maynard Smith l'avait envisagé : la densité de connexions reste suffisante pour connecter la majeure partie des génotypes viables (figure V.3). Les hyper-réseaux neutres des protéines hors-réseau sont eux aussi connectés : une composante connexe principale regroupe de l'ordre 50 % des génotypes viables dès que l'hyper-réseau neutre est suffisamment grand (cf. table V.4). En dépit de la sélection négative importante, les réseaux neutres dimériques apparaissent donc comme à la fois très diversifiés et bien connectés (réseau « petit monde »).

V.3.2.2 Stabilité des séquences fonctionnelles

De l'absence de localisation spécifique dans le réseau neutre (cf. figure V.4) découle une indépendance entre le critère sélectif et la stabilité d'une séquence (cf. figure V.5). Nous rejoignons les conclusions de Bloom *et al.* : « there are sequences that exhibit both good stability and strong binding » [18, p. 2762]. La stabilité dans le modèle tridimensionnel hors-réseau n'a pas été directement mesurée. Les distributions du nombre de connexions viables et non-viables des génotypes viables fournit cependant une approximation (cf. figures V.15 et V.16).

Dans le modèle des protéines sur réseau, la stabilité n'est pas diminuée par l'interaction entre les protéines, car l'interaction ne requiert pas l'exposition d'un surcroît de résidus hydrophobes (cf. figure V.6). De manière imagée, on peut dire que l'interaction protéine-protéine est rendue possible par un réarrangement des résidus hydrophobes déjà présents à la surface (cf. figure V.2). Cette différence de composition quasiment nulle entre des protéines monomériques et des protéines monomériques est étudiée en annexe à la lumière des données structurales de la PDB (cf. figure V.19*, p. 217).

Cette manière de voir rapproche énormément nos résultats des méthodes utilisées par Tiana et Broglia dans leur étude des motifs de conservation dans les dimères à deux et trois états [179]. Effectivement, afin de préserver une composition réaliste en acides aminés, ils génèrent des séquences en permutant aléatoirement des acides aminés dans la séquence. Les séquences capables de dimériser selon un mécanisme à trois états (séquences 3 et 4 dans le tableau I de la référence [179]) possèdent des énergies de repliement E_1 similaires à celle d'une protéine purement monomérique (séquence 6).

V.3.2.3 Organisation en *superfunnel* fonctionnel

L'organisation en *superfunnel* fonctionnel est présente dans le travail de Blackburne et Hirst [14]. Les figures V.13 et V.17, pour les deux modèles que nous avons présentés, prouvent que l'aptitude à dimériser s'organise également selon un *superfunnel*, autour d'un génotype prototype. Comme ce génotype est le plus peuplé à l'état stationnaire, il possède vraisemblablement un grand nombre de connexions le liant à d'autres génotypes viables. Un génotype possédant beaucoup de voisins viables est probablement un génotype ayant une bonne propension à former un dimère stable. L'aptitude moyenne à dimériser diminue à mesure que l'on s'éloigne du génotype prototype.

Le *superfunnel* présenté dans la figure V.13 n'est pas très prononcé. Des alphabets plus complets rendraient probablement possible l'apparition de *superfunnel* plus marqués.

V.3.3 Évolution neutre de la fonctionnalité

L'évolution neutre aboutit à une augmentation en moyenne de la fonctionnalité (facteur d'enrichissement $\phi([AB])$ supérieur à un). Les génotypes dont les séquences dimérisent efficacement tolèrent généralement plus de mutations que ceux dont le dimère possède une stabilité réduite. Les génotypes hautement fonctionnels possèdent donc plus de connexions dans l'hyper-réseau neutre. Par un phénomène de rétroaction des populations identique à celui des protéines monomériques, ces connexions favorisent la surreprésentation de ces génotypes. L'aptitude moyenne à dimériser s'en trouve augmentée (cf. figure V.11 pour les protéines sur réseau et table V.4 pour les protéines hors-réseau).

Ce phénomène n'apparaît pas en l'absence de contrainte ($\delta = 0$) parce que la concentration en dimère fonctionnel est non corrélée au nombre de connexions dans l'hyper-réseau neutre (cf. figure V.5). De même, pour des contraintes fonctionnelles trop élevées (δ trop grand), l'augmentation $\phi([AB])$ devient plus faible (cf. figure V.9). La figure V.13 permet d'avancer une explication à cet « essoufflement » pour des valeurs élevées de δ. Pour $\delta = 2{,}7$, une structure de *superfunnel* apparaît, cependant la plage de concentrations dans laquelle il prend place est bornée

1. d'une part, par la contrainte fonctionnelle ($\delta = 2{,}7$),
2. d'autre part, par la concentration maximale que permette d'atteindre l'alphabet HP ($[AB]_{\max} \approx 3{,}33$, cf. figure V.1).

La plage étant limitée (de 2,7 à 3,33), l'évolution neutre ne peut pas augmenter notoirement la concentration en dimère fonctionnel moyenne.

Un deuxième phénomène prend place. Il est identique à celui décrit par Taverna et Golstein [172]. Les génotypes du centre du *superfunnel* sont en réalité peu nombreux, et sont défavorisés « entropiquement ».

Si des alphabets plus complets permettent, comme nous l'avons suggéré plus haut, de créer des *superfunnel* plus marqués, l'augmentation de l'aptitude moyenne à dimériser pourrait persister pour des valeurs plus hautes de δ. Une approche comme celle employée déjà par Taverna et Goldstein permettrait de valider cette hypothèse [173]. Ces auteurs ont simulé l'évolution d'une population[31] et l'ont comparée à celle d'un individu unique[32]. Ils ont mis au point cette méthode pour étudier la déstabilisation énergétique des mutations. Mais cette stratégie est parfaitement envisageable avec notre modèle de dimérisation.

V.3.4 Émergence d'une fonction

Le fait que les génotypes fonctionnels soient rares mais connectés implique que ce qui est limitant dans l'émergence d'une fonction est de trouver un génotype fonctionnel. Une fois ce génotype trouvé, la fonction comme la stabilité peuvent être optimisées en parcourant l'hyper-réseau neutre. Une conclusion très similaire ressort également de l'étude de Bloom *et al.* [18] : « once highly functional sequences are found, they can be optimised for stability ».

31. La population est suffisamment nombreuse et le taux de mutation est assez élevé pour que l'effet de dynamique de la population se manifeste.

32. L'effet de dynamique de la population ne peut pas exister avec un seul individu. Les propriétés de l'évolution de cet individu sont celles de la distribution uniforme.

L'optimisation de la fonction est rendue possible par les connexions entre les génotypes fonctionnels mais aussi par l'organisation en *superfunnel* fonctionnel. L'efficacité de l'évolution et de l'évolution dirigée dépend de cette organisation. Sans la connectivité entre les génotypes viables ou sans l'organisation en *superfunnel* fonctionnel, l'évolution ne peut tout simplement pas se produire. C'est le paradoxe de Salisbury : étant donné l'immensité de l'espace des séquences, comment est-il possible que l'évolution puisse atteindre des séquences fonctionnelles si elle ne peut pas élaborer de *stratégie* [160] ?

La recherche du génotype fonctionnel initial semble donc être l'étape limitante dans la création d'une nouvelle fonction. Elle repose sur une exploration du réseau neutre monomérique et souligne donc l'importance de l'évolution neutre dans la prospection de nouvelles fonctions.

La structure des réseaux neutres des protéines et de ceux des ARNt est singulièrement différente. Ces différences affectent la manière dont l'évolution peut explorer de nouveaux phénotypes, c'est-à-dire de nouvelles structures, de nouvelles fonctions.

1. Les réseaux neutres d'ARNt sont épars mais traversent tout l'espace des séquences. Ils obéissent au même schéma que celui dressé pour la matrice d'Ising qui crée une complémentarité H-P sur le modèle de la complémentarité des bases de l'ARN. De nombreux réseaux neutres, correspondant à des structures différentes, sont en contact les uns avec les autres. Ces contacts entre réseaux rendent possible la « découverte phénotypique » par diffusion dans un réseau neutre [84]. Ce sont grâce à eux qu'il est possible de simuler une évolution par « équilibre ponctué » mis en évidence par Fontana et Schuster [59] : l'évolution prend place dans un réseau neutre jusqu'à trouver un passage vers un autre réseau neutre, plus avantageux.
2. D'après une étude réalisée sur des protéines sur réseau et un alphabet HP, les réseaux neutres de protéine sont plus localisés [20]. Babajide *et al.* ont trouvé que les séquences des protéines étaient flexibles dans un alphabet à vingt acides aminés [1] mais pas dans un alphabet HP. Les profils HP restent conservés : « [the sequences of a neutral network are] very flexible at the level of individual amino acids but require a significant level of conservation of amino acid classes ».

Les réseaux neutres de protéine n'autorisent que peu d'exploration phénotypique : ils sont bien isolés les uns des autres[33]. L'exploration phénotypique est nettement moins efficace, même si la recombinaison permet de l'accroître [35]. Dans ces conditions, la réutilisation d'une même charpente pour réaliser différentes réactions enzymatiques paraît nettement moins surprenante [4]. Par ailleurs, la séparation des réseaux neutres dans l'espace

33. Cette propriété est absente de notre modèle de protéine sur réseau, car nous nous sommes limité aux conformations compactes.

des séquences fournit une base théorique au défi de Paracelsus, c'est-à-dire la difficulté de produire une nouvelle structure protéique stable sans modifier très substantiellement sa séquence [158]. Et bien que des séquences sans similitude apparente puissent être connectées par des intermédiaires, les profils HP de ces séquences sont très voisins [103,144,174].

La capacité d'une protéine à évoluer dépend de sa stabilité et de sa robustesse mutationnelle : une séquence plus stable et plus robuste peut acquérir une nouvelle fonction plus efficacement [17]. Ce constat fait à nouveau le lien avec les structures modifiables qui sont plus stables en moyenne (cf. figure III.8, p. 95). Les structures modifiables sont en outre plus robustes aux mutations. La fraction de mutations neutres augmente et la vitesse d'évolution également. Si la recherche d'un premier génotype fonctionnel, est une étape limitante, la vitesse de recherche est liée à la capacité à évoluer (*evolvability*). Les structures modifiables pourraient permettre d'explorer plus rapidement de nouvelles fonctions.

L'évolution neutre augmente en moyenne la capacité à dimériser, la stabilité (T_f) et la robustesse mutationnelle (cf. figure V.9 et table V.4). L'évolution neutre stabilise donc les monomères dans leur conformation native et l'interaction entre les monomères. Cette double stabilisation pourrait être importante dans le recrutement d'une nouvelle interaction.

V.3.5 Conservation des résidus

L'interface des dimères tridimensionnels hors-réseau n'est pas plus conservée que le reste de la surface des monomères. L'absence de conservation de l'interface est en accord avec Caffrey *et al.* [26]. La recherche systématique des interfaces par un biais dans la conservation est, d'après ces résultats, vouée à l'échec. Les succès obtenus dans ce sens ont toujours employé des facteurs additionnels [45, 183]. Des méthodes, censément plus fiables, reposant sur la conservation du cœur et de l'anneau des interfaces ont été proposées [71].

Des études récentes ont montré que seuls les dimères permanents possédaient une interface nettement conservée [131]. Les travaux Caffrey *et al.* mènent à la même conclusion : les dimères transitoires, à trois états, ne présentent pas de conservation évidente. L'existence de quelques « points chauds » cumulant la majorité de l'énergie d'interaction [19, 118] ou de résidus d'ancrage [155] pourrait expliquer que la majorité des résidus à l'interface *ne* sont *pas* conservés.

Néanmoins, l'absence totale de différence dans la conservation de l'interface par rapport au reste de la surface peut être aussi due à la simplification de la matrice d'énergie. Ainsi, la présence de charges répulsives peut réduire la variabilité à l'interface. Le modèle HP, ne distinguant pas les charges parmi les résidus polaires, ne peut pas rendre compte

de cet effet. Notre modèle néglige également les réarrangements des chaînes latérales. Au niveau de l'interface, la nécessité d'un bon agencement des chaînes latérales pourrait être un facteur important de la conservation des résidus.

La contrainte fonctionnelle agit de manière diffuse et non seulement au niveau de l'interface, la conservation est augmentée sensiblement sous l'effet de la pression sélective pour les résidus à l'interface, mais aussi ceux surfaciques ou enfouis.

Tandis que le cœur des monomères est essentiellement hydrophobe et très conservé, l'interface n'est pas particulièrement hydrophobe. Cette observation surprenante découle en partie de la nécessité pour les monomères de se replier. En outre, la présence de résidus polaires aux interfaces n'est pas rare :

> It has often been assumed that proteins will associate through hydrophobic patches on their surfaces. However, polar interactions between subunits are also common and, in terms of the driving force for complexation, it is important to explore the relative contributions of these effects, including reference to the subunits' ability to exist independently [88].

V.3.6 Taux d'évolution

Nous avons abordé en introduction l'énigme des vitesses d'évolution des protéines multimériques. Les données expérimentales suggèrent que la contrainte impliquée par la complexation, si elle existe, est, pour le moins, difficile à mettre en évidence.

Tout d'abord, il est nécessaire de faire le lien entre la vitesse d'évolution d'une protéine et la robustesse mutationnelle. Dans l'introduction, nous avons vu que, d'après Kimura, la vitesse d'évolution « macroscopique » est reliée au taux de mutation neutre (équation 19′, p. 32) :

$$k = \mu_T \, f_0 \,,$$

où k est le taux de substitution, μ_T le taux de mutation total et f_0 la fraction des mutations qui sont neutres. Dans notre modèle, la fraction de mutations neutres s'écrit

$$f_0 = \frac{\langle n \rangle}{(L_A + L_B)\,(A - 1)} \,,$$

où L_A et L_B sont les longueurs de chaîne des monomères et A la taille de l'alphabet. On a donc une relation très simple : le taux de substitution est proportionnel à la robustesse mutationnelle moyenne : $k \propto \langle n \rangle$.

Les données brutes de nos modèles soutiennent une diminution du taux de substitution lorsque la contrainte de dimérisation est ajoutée. En effet, l'addition d'une contrainte dimérique diminue la robustesse mutationnelle (cf. tableaux V.2 et V.4) et donc le taux de

substitution attendu. Certains résultats de notre modèle suggèrent néanmoins des pistes qui permettraient d'expliquer pourquoi les taux d'évolution expérimentaux des protéines multimériques ne semblent pas affectés par la contrainte de complexation.

1. Dans les profils des protéines sur réseau (cf. figure V.12) et les données de conservation des résidus (cf. table V.7), tous les résidus apparaissent plus conservés lorsqu'une contrainte dimérique est ajoutée. Cela signifie que la contrainte fonctionnelle est diffusément répartie sur la structure. Autrement dit, la contrainte fonctionnelle n'incombe pas uniquement à l'interface.
2. L'accroissement de $\phi(n)$ par un effet d'épistase entre les gènes peut expliquer un regain de vitesse d'évolution. Si l'on consulte le tableau V.2, nous pouvons voir qu'en l'absence de contrainte, les monomères *A* et *B* évoluent à une vitesse proportionnelle à 9,18 et 8,96 si le régime est celui d'une distribution uniforme ($M\mu \ll 1$). Une contrainte fonctionnelle $\delta = 0{,}6$ réduit de 23 % et 29 % les vitesses d'évolution alors que 85 % des génotypes ont été éliminés par sélection négative (cf. table V.1). À l'état stationnaire ($M\mu \gg 1$), la même contrainte ne réduit les vitesses d'évolution que de 16 %. L'effet de dynamique de la population limite la diminution de vitesse d'évolution.
3. La robustesse mutationnelle obtenue sans contrainte ($\delta = 0$) vaut 23,32. Lorsque la contrainte fonctionnelle est élevée, prenons pour valeur $\delta = 2{,}5$, seule une petite fraction des génotypes sont viables (0,05 %). La robustesse mutationnelle vaut alors 12,41. Une forte contrainte fonctionnelle réduit donc la robustesse mutationnelle de seulement 46 %. La même conclusion d'applique au taux d'évolution puisque le taux d'évolution est proportionnel à la robustesse mutationnelle.

   On peut donc imaginer, qu'une protéine puisse évoluer rapidement même si elle est soumise à une sélection négative extrêmement stringente. Cela est possible si les séquences viables forment un réseau suffisamment connecté. Nous pensons que cet effet pourrait être mis en évidence plus clairement avec un alphabet plus complet, car dans ce cas l'organisation en *superfunnel* serait plus marquée et les séquences viables plus densément connectées en son centre.

Les deux modèles que nous avons présentés sont simples mais en bon accord qualitatif. L'étude de l'évolution des dimères par simulation nous a forcé à procéder à des approximations : utilisation de protéines sur réseau, d'un alphabet simplifié, etc. Néanmoins, ils offrent des perspectives nouvelles sur des sujets aussi divers que l'apparition de nouvelles fonctions, la conservation des résidus ou les taux d'évolution. Nous pensons que ces résultats, encore relativement préliminaires, sont encourageants et doivent faire l'objet d'approfondissements futurs.

V.4 Annexes

V.4.1 Composition hydrophobe expérimentales

Les protéines suivantes sont étudiées[34] :

Protéines monomériques 1A41, 1A4I, 1AH5, 1AK0, 1AKO, 1AL3, 1ALN, 1AMF, 1AOD, 1AOL, 1AUQ, 1AUX, 1AXC, 1B0U, 1B35, 1B5E, 1B9N, 1BJY, 1BOU, 1BRT, 1BTM, 1C3W, 1C8Z, 1CBY, 1CL7, 1CMX, 1CN4, 1CNV, 1CUN, 1D2H, 1DEL, 1E0C, 1E2T, 1E3K, 1E42, 1E9G, 1EF1, 1EG4, 1EH4, 1EKQ, 1EKU, 1EL6, 1EM2, 1EOI, 1EON, 1EQ2, 1ES5, 1ES6, 1EUV, 1EYP, 1EZM, 1F5V, 1FG5, 1FI2, 1FVI, 1FYE, 1G44, 1G57, 1G88, 1GA7, 1GC1, 1GC6, 1GCU, 1GDO, 1GL4, 1GNH, 1GTP, 1GWZ, 1H2R, 1H5W, 1HAV, 1I49, 1I6O, 1I78, 1I9B, 1IE9, 1IG3, 1II5, 1IIS, 1IM0, 1IM8, 1INI, 1INL, 1J97, 1J9L, 1JCL, 1JE0, 1JK0, 1JK7, 1JLS, 1JLX, 1JP3, 1JUK, 1JWB, 1JYK, 1KPG, 1NFD, 1QEX, 1QHL, 1QLA, 1QO2, 1QQ9, 1QQQ, 1QRE, 1QSF, 1QSG, 1QTF, 1SCJ, 1TL2, 1VID, 1XO1, 2BC2, 2PHK, 2TMK, 3CD2, 3CLA, 3CSM, 3HDH, 3MAG, 3SEB, 3TDT, 4FAP, 6GSV et 9GAF,

Protéines dimériques Les chaînes A et B de 1ABR, 1AI7, 1AUI, 1BLX, 1BPL, 1DKF, 1EFV, 1EUD, 1F3V, 1HDM, 1ITB, 1KSG, 1LOT, 1N52, 1NPE, 1NW9, 1O5M, 1PHN, 1QGE, 1RKE, 1RTJ, 1SMP, 1STF et 1UBT.

Le degré d'exposition, $f(\mathrm{R})$ d'un résidu R est un nombre compris entre zéro et un, donné par

$$f(\mathrm{R}) = \frac{\mathrm{ASA}(\mathrm{R})}{\mathrm{ASA}_{\mathrm{ref}}(\mathrm{R})},$$

où les valeurs $\mathrm{ASA}_{\mathrm{ref}}$ sont calculées à partir de tripeptides GXG et sont données dans le tableau V.8. Plus $f(\mathrm{R})$ est élevé, plus le résidu concerné est exposé au solvant. Cinq intervalles d'exposition vont être envisagés : de 0 à 20 %, de 20 % à 40 %, de 40 % à 60 %, de 60 % à 80 % et de 80 % à 100 %. La figure V.19 présente pour chacun des cinq intervalles susmentionnés le rapport, r^{HP} entre le nombre de résidus hydrophobes et celui de résidus polaires.

ALA	119.448	GLN	198.677	LEU	197.610	SER	129.473
ARG	250.467	GLU	195.305	LYS	223.716	THR	153.380
ASN	160.894	GLY	92.241	MET	207.096	TRP	279.732
ASP	163.046	HIS	211.854	PHE	241.284	TYR	240.779
CYS	146.209	ILE	197.642	PRO	155.957	VAL	168.663

Tab. V.8 : ASA de référence (en Å^2) pour chaque acide aminé. La molécule sonde de solvant est une sphère de rayon 1,6 Å.

34. Merci à Guillaume Launay pour son aide.

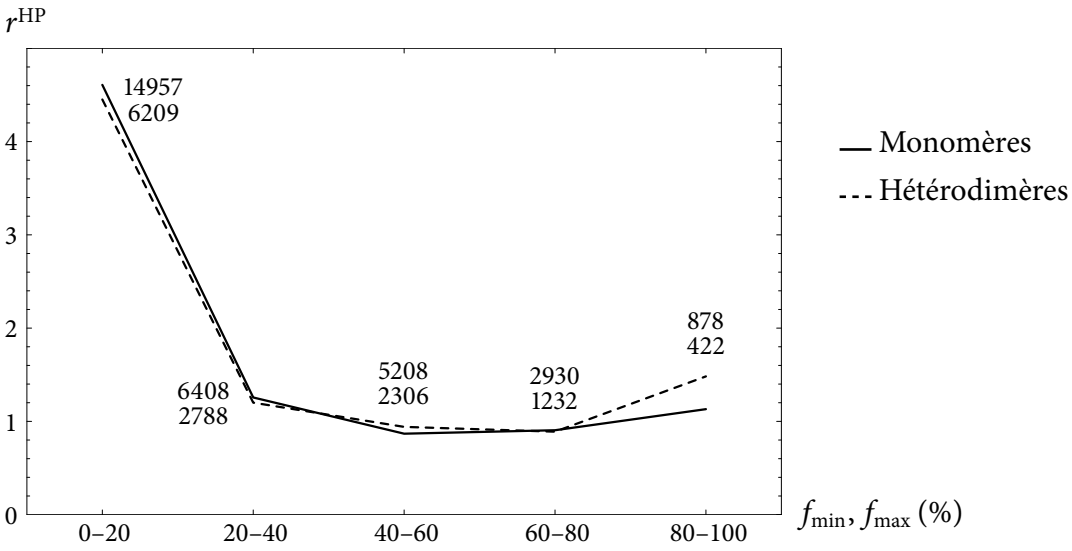


Fig. V.19 : Les nombres à proximité de chaque point indiquent le nombre de résidus pris en compte (nombre supérieur : monomères, nombre inférieur : hétérodimères).

La différence de composition hydrophobe entre les protéines monomériques et dimériques est faible : elle est limitée à la plus faible fraction (2 %) des résidus très exposés ($f(\mathrm{R}) > 80$ %). On peut même aller plus loin dans la conclusion : la composition *en acide aminé* est très similaire entre les deux ensembles de protéines que nous avons étudiés : cf. figure V.20.

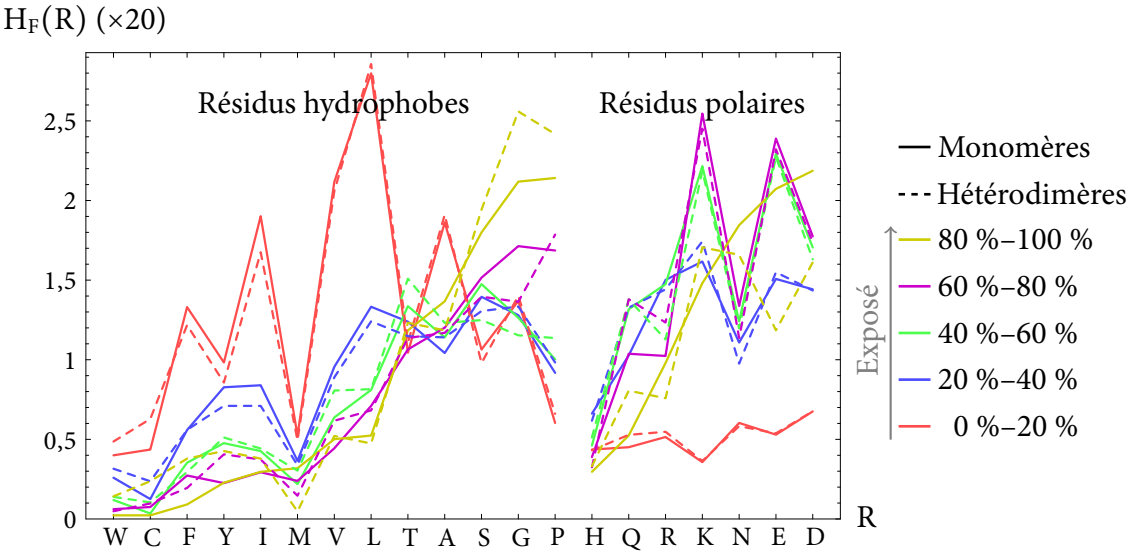


Fig. V.20 : Fréquence de chaque acide aminé selon son appartenance à l'un des cinq intervalles d'exposition. L'axe des ordonnées est mis à l'échelle d'un facteur 1/20.

Chapitre VI

Article

Prusias — *Et que si je lui laisse un jour une couronne,*
Ma tête en porte trois, que sa valeur me donne.
…et ma confusion,
[...] Sans cesse offre à mes yeux cette vue importune,
Que qui m'en donne trois, peut bien m'en ôter une,
Qu'il n'a qu'à l'entreprendre, et peut tout ce qu'il veut.
Juge, Araspe, où j'en suis, s'il veut tout ce qu'il peut.

Corneille, *Nicomède*

Dans ce chapitre, nous reproduisons l'article que nous avons soumis avec Thomas Simonson à la revue *Biophysical Journal*.

Josselin Noirel et Thomas Simonson, « Neutral evolution of protein-protein interactions : A computational study using toy models », *Biophysical Journal*.

Neutral evolution of protein–protein interactions: a computational study using toy models

Josselin Noirel and Thomas Simonson*

Laboratoire de Biochimie (UMR 7654 CNRS), Departement of Biology,
Ecole Polytechnique, 91128 Palaiseau, France.
*Corresponding author: thomas.simonson@polytechnique.fr

Abstract

Protein–protein interactions are central to cellular organization, and must have appeared at a very early stage of evolution. To better understand their emergence, we consider a very simple model of protein structure and evolution and determine the effect of including an explicit selection for protein–protein interactions. Viable alleles are assumed to all have the same fitness, following the neutral theory of evolution. Using a 2D lattice representation of the protein dimer structures and binary, hydrophobic/hydrophilic sequences, exact calculations are performed. Results do not depend too strongly on these assumptions, since a more realistic, 3D, off-lattice representation gives results in qualitative agreement with the 2D one. The present model selects for protein–protein interactions by requiring that the cellular concentration of a specific, functional dimer must be above a chosen threshhold. The evolutionary dynamics then lead to a steady state population that is enriched in sequences that are good dimerizers, well beyond the minimal ability needed to survive. Correspondingly, sequences close to the viability threshold (the edge of the neutral network) are depleted, because they are subject to a larger proportion of lethal mutations. The steady state sequences should then provide an increased resistance to environmental change, or adaptability, and an increased ability to evolve and invent new interacting pairs of proteins.

1 Introduction

Modern genomics and molecular biology have transformed our understanding of molecular evolution. The diversity of modern proteins is illustrated by the millions of known gene sequences and thousands of known protein structures. In vitro evolution and computer simulations allow us to go beyond natural sequences, and explore protein structure and plasticity within a much larger space [1]. It has become clear that proteins are remarkably robust with respect to accidental or designed mutations, retaining structure and function in many cases. This has helped renew interest in theories of evolution that explore the role of random drift, in contrast to positive selection, such as the neutral evolution theory of Kimura [2, 3] and the RNA evolution models of Eigen and Schuster, which led to the quasispecies concept [4]. Models of molecular evolution have also been renewed by advances in the protein folding field. Lattice models first revealed, for example, that fast folding protein sequences tend to adopt highly "designable" structures, which are shared by many different sequences and are robust with respect to amino acid mutations [5–7].

To model evolution at the molecular level requires that sequence and structure space be explored together. By using very simple models, a precise mapping can be defined between genotype, phenotype, and fitness [4]. Once this mapping is known, an evolutionary population dynamics can be performed to predict the statistics of gene frequencies. A correct understanding and modelling of these three aspects—sequence space exploration, structure space sampling, and population dynamics—is critical, as demonstrated theoretically and experimentally by the emergence of the so-called 'quasi-species' concept [4].

In recent years, both on-lattice and off-lattice models of protein structure have been employed in evolutionary models [8–11]. Important aspects of the Darwinist [12] and Neutralist theories [8, 9] have been explored. Very simple models suggest ways, for example, in which increased evolvability can emerge spontaneously in an evolving population [13–15]. In a neutral evolutionary model that allows only point mutations, the population dynamics have the form of a random diffusion through the set of viable sequences. Such models usually distinguish between the set of viable sequences and a surrounding space of non-viable sequences. With such a model, the steady state has the remarkable property that alleles with a high tolerance of mutations are overrepresented, compared to a random selection of viable sequences. There is a corresponding depletion in sequences that have a low tolerance of mutations, because they are subject to a larger proportion of lethal mutations. In effect, the most robust, mutationally-tolerant sequences form a basin, or "funnel" in sequence space [8, 16], whereas the sequences at the outer "edge" of the viable set are more fragile. Thus, a form of evolvability arises purely from the geometry of sequence space. Sequence funneling was recently observed experimentally by directed evolution [17].

Because protein functionality is very complex, evolutionary models usually assume that protein structure can be used as a proxy for function: proteins that adopt the correct structure are assumed viable [11, 18]. More recently, explicit models of functionality have been introduced, involving the ability of the protein to bind a small ligand [19–22]. Here, previous models are extended to take into account explicitly the essential role of protein–protein interactions. Indeed, most proteins must interact with other proteins to function, and protein–protein interactions are a key source of epistasis, or coupling between genes [23]. The network of protein–protein interactions has been studied exhaustively for several organisms, and some of its topological properties established [24, 25]. Its complexity is thought to correlate with the overall complexity of an organism.

We model the neutral evolution of two proteins, coupled by a selection criterion that requires the formation of a specific protein–protein interaction. Only neutral evolution through point mutations is considered. This mutation mechanism, though simple, is nevertheless important for the evolution of individual protein domains. More complex events like recombination, essential for the creation or rearrangement of entire domains in higher organisms, are neglected here. Protein structure is represented through two simple, very different models: a two-dimensional lattice model and a three-dimensional off-lattice model. These models are so simple that they are sometimes referred to as 'toy' models. Nevertheless, these and similar models have been used extensively in the past to study both protein folding and the evolution of individual proteins [7, 11], and have been shown to provide useful qualitative insights.

Only limited studies of protein–protein pairs have been reported [26]. Here, a functional coupling between proteins is introduced. In this case, two proteins must not only fold, but specifically associate to perform a vital function. Using the 2D lattice description of the structures, the chemical equations for dimerization can be solved exactly. The viable sequences are then enumerated and the evolutionary dynamics characterized. Under conditions of moderate selection, neutral evolution is found to increase the functional effectiveness of the two proteins: the steady state population is enriched in alleles coding for proteins that readily dimerize. Using the 3D off-lattice description, a similar effect is observed. This result is analogous to the monomeric result described above: the monomeric steady state is enriched in mutationally robust sequences. In both the monomeric and dimeric cases, sequences in the core of the neutral network are overpopulated, while sequences at the edge are depleted. In the dimeric case, the steady state enrichment corresponds to an enhanced functional ability. The enhancement emerges from a neutral model that requires only a minimal ability to function, through the funneled shape of the network of viable sequences.

2 Materials and methods

2.1 The evolutionary model and its properties

Following [4, 9] and others, we first assume that all genes evolve independently, and we focus arbitrarily on one of them. In a second step, below, we will consider co-evolution of two interacting proteins. For now, the single gene of interest is assumed viable if the corresponding protein folds into its correct conformation. The S sequences that adopt this conformation are all assumed to be equally fit. We assume evolution can only occur through point mutations; i.e., changes of a single amino acid. The complete set of viable sequences defines a graph, containing S nodes. Each node represents a viable sequence; links between nodes represent point mutations. The graph may not be fully connected; i.e., it may be impossible to connect two viable sequences by a series of point mutations. If the entire population starts out with the same, 'native' protein sequence, then future evolution will only explore the corresponding, connected subgraph. Therefore, we can assume without loss of generality that there is only one connected graph, which is referred to as a 'neutral network'.

The following, discrete-time, evolutionary model is analyzed [4, 9]. For simplicity, we describe it in detail for the present, single gene case. The case of interacting proteins is considered further on. Between times t and $t+1$, an individual with allele i has a probability α of undergoing a point mutation and a probability β of dying. A new individual has a probability γp_i of appearing spontaneously by birth; with probability $1-(\alpha+\beta+\gamma)$, the individual continues unchanged. After each generation, populations are rescaled to maintain a constant total. The probability to find a given individual with allele i at t is denoted p_i; the change between t and $t+1$ is denoted δp_i. In the limit of a large population, these probabilities follow the equation:

$$\delta p_i = -(\alpha+\beta-\gamma)p_i + \frac{\alpha}{M}\sum_{k\sim i} p_k, \tag{1}$$

where $k \sim i$ means that k and i are neighbors in the neutral network. The $S \times S$ adjacency matrix C [25] is defined by: $C_{ij} = 1$ if $i \sim j$ and zero otherwise; the vector of allele probabilities is $p = (p_1, p_2, \cdots, p_S)$. Eq. (1) can be rearranged into the following vector form:

$$\delta p = \frac{\alpha}{M}\left(C - \nu I\right) p. \tag{2}$$

It is easy to show that ν is the mean number of neighbors of the alleles in the network: $\nu = \sum_{i=1}^{S} \nu_i p_i$, where ν_i is the number of neighbors of allele i. Eq. (2) describes the flow of population within the neutral network. The first term in parentheses on the right represents alleles flowing into a given node i; the second term represents alleles flowing out of i, taking into account the fraction of viable and lethal mutations. An important property of Eq. (2) is that there is a single stable steady state. The

steady state probability vector, $p = p^{ss}$, is an eigenvector of C, associated with the largest eigenvalue, $\nu = \nu^{ss}$. Remarkably, the steady state can be shown to be not only stable with respect to small fluctuations, but globally stable. A detailed proof will be published elsewhere; see [4] for a detailed treatment of related mathematical models.

2.2 Structural models: 2D lattice model and 3D off-lattice model

Two physical models of a protein are considered. The first treats the "protein" as a chain of $L = 25$ beads, or amino acids, which can be either polar (P) or hydrophobic (H). Acceptable conformations occupy a two-dimensional, 5×5, square lattice. The energy is

$$E = \sum_{i<j} e_{ij} \Delta_{ij}, \tag{3}$$

where $\Delta_{ij} = 1$ if the beads i, j are neighbors on the lattice and zero otherwise, and the interaction coefficients depend on the type (P or H) of each bead. The values $e_{\mathrm{HH}} = -2.3$, $e_{\mathrm{HP}} = -1$, and $e_{\mathrm{PP}} = 0$ are used, following [9, 27], to favor compact conformations with a hydrophobic core. A particular sequence is considered to fold if its lowest energy conformation is unique (i.e., non-degenerate). It is viable if it folds into the same conformation as the other sequences of the native neutral network. The protein chain is considered to have a direction (even though the energy function does not); e.g., the sequences HPP and PPH are different. Sequence exploration is done by exhaustive enumeration.

With the 2D model, we can calculate exactly the folding temperature T_f of each structure and sequence. T_f is the temperature at which the native conformation is populated 50% of the time; it is straightforward to compute it numerically from Boltzmann's law and the energy spectrum of the 1081 possible conformations. For a protein dimer, we define T_f as the minimum of the folding temperatures of the two separate partners.

The second physical model is a three-dimensional, off-lattice model [10, 28]. Two proteins are considered: the 69-residue SH3 domain of Vav and the 57-residue SH3 domain of Grb2. These two form a protein–protein complex (PDB accession number 1gcq). For each one, the experimental, 3D structure is considered, along with over 1200 "decoy" structures, whose backbone geometries are taken from completely different proteins [29, 30]. The sidechains are built assuming the most common rotamer for each amino acid type [31]. For each protein, 100 additional decoys, with more native-like structures, were produced by molecular dynamics *in vacuo* at 310 Kelvins. Amino acids interact through Eq. (3), with $\Delta_{ij} = 1$ if they have two nonhydrogen atoms within 4.5 Å of each other; $= 0$ otherwise. The amino acids are divided into two classes: H = {LVIMCASTPGFWY} and P = {EDNQKRH}. The

energy parameters are e_{HH} = -8.5, e_{HP} = 9.0, e_{PP} = -3.5. These parameters were optimized to discriminate experimental protein structures from large sets of decoys [29, 30]. With this model, we first enumerate sequences that are viable as monomers; i.e., they fold into the desired, native structure. For this step, the monomer sequence space is explored by a Monte Carlo method [10]. A "move" consists in a random point mutation, which is accepted if the desired, functional structure has a sufficiently low energy, compared to the non-functional, decoy structures. Specifically, the functional fold must have an energy gap (energy difference from the lowest decoy) and a Z-score (energy difference from the average decoy, in standard deviation units) as large as those of the starting sequence. The starting sequence is slightly different from the native sequence. It is obtained by minimizing the latter through several thousand Monte Carlo moves. A trajectory of one hundred million mutations is then performed. For each accepted mutation, we also explore systematically its nearest "neighbors" (all its single mutations), thus generating a large, representative subset of the relevant neutral network in monomeric sequence space [10].

Once the neutral network has been constructed (either for a 2D or a 3D protein), the steady state distribution of alleles is computed by an iterative, shifted power method [32], which yields the eigenvector of C corresponding to the largest eigenvalue.

2.3 Interacting genes: the 2D on-lattice case

In a second step, an evolutionary scenario with interacting genes is explored. We describe first the 2D lattice case. It is assumed that a vital function can only be performed when two proteins A, B not only fold, but specifically dimerize. The sequences that are viable as monomomers are first obtained by the procedure described above. In addition, the two proteins, because of their square-lattice structure, can form ten homodimers AA, BB and 16 heterodimers AB, just one of which is functional. Inter-protein interactions are described by Eq. (3). In addition, dimerization is opposed by a constant entropic penalty, ϵ. We consider here only sequences that are known to form viable monomers (see above), so that their unfolded conformations are unstable and can be neglected. There are then 38 possible chemical species, whose equilibrium concentrations are obtained by solving the system:

$$\begin{aligned} [AA_I] &= a_I[A]^2, \quad 1 \le I \le 10 \\ [BB_J] &= a_J[B]^2, \quad 1 \le J \le 10 \\ [AB_K] &= c_K[A][B], \quad 1 \le K \le 16 \end{aligned} \tag{4}$$

with fixed total concentrations $[A]_{tot}$ and $[B]_{tot}$. Here, a_I, b_J, c_K are equilibrium constants; for example $a_I = \exp(-\Delta E_I/kT)$, where k is Boltzmann's constant, T the temperature, and ΔE_I is the association free energy of the dimer AA_I. The

a_I, b_J, c_K depend on the sequences of A, B through the association free energies. The chemical equations (4) can be reduced to a fourth-order polynomial and solved either analytically or numerically (which is faster). An A, B pair with particular sequences is considered viable if the functional dimer has an equilibrium concentration greater than a chosen threshold δ. The functional dimer is the one that minimizes the dimerization energy, averaged over all the A, B sequences that fold (into their designated native conformations).

2.4 Interacting genes: the 3D off-lattice case

We now turn to the 3D, off-lattice case. We consider the Grb2:Vav complex. In contrast to the lattice case above, there are far too many possible dimer structures for an exact enumeration to be done. Instead, we consider a limited set of dimeric decoy structures. These were generated by a docking procedure described in detail elsewhere [30]. Briefly, we start from the two separate proteins, with their native sequences, positioned randomly with respect to each other. They are then docked together with a molecular mechanics energy function [33], using restrained energy minimization. In an initial phase, the restraint consists in a harmonic spring that pulls their centers of mass together. In a second phase, the restraint corresponds to an electrostatic contrast introduced artificially between the two proteins: charges on one are slightly increased; charges on the other are slightly decreased. The pair is energy-minimized, allowing for limited intra-protein deformations. Structures that involve too large a deformation of either partner (rms deviation of more than 3.5 Å with respect to the starting monomer conformation) are discarded. Overall, we produced a total of 1695 decoys, of which 912 were discarded because they lead to a lower interaction energy than the native structure. This is due to the simplicity of the energy function. We are left with 783 decoys (compared to 35 non-native structures in the 2D dimer case).

To determine the viable dimer sequences, we start from the sequences that are viable as monomers, generated by the Monte Carlo method described above. Because the number of accepted monomer sequences is very high, we only keep one sequence per HP profile, picked arbitrarily from the available sequences. There were a total of 31469 and 29667 profiles for Grb2 and Vav, respectively. We then consider the ability to dimerize, by comparing the energy of the native dimer structure to the energy of all the decoy structures. For a given pair of Vav and Grb2 sequences and a given dimer structure, the energy is obtained by threading the sequence onto the dimer structure. Sidechains are positioned in their most common rotamer, as described above for the monomer case. For a given sequence pair, a Z-score is calculated for each dimeric structure. The Z-score is defined as the energy of the structure relative to the average energy, measured in standard deviation units. A pair of sequences is considered to form a viable dimer if the native structure has a

sufficiently low Z-score. Specifically, its Z-score should be among the top k values, where k is an integer between 3 and 10. Choosing $k = 3$, for example, means that for a sequence pair to be viable, at most two decoys should have a lower Z-score than the native dimer structure. By varying k, we can explore different stringencies for the selection criterion. This is analogous to varying δ in the 2D lattice case. With $k = 10$, there are 470,334 viable pairs of HP profiles. To test the viability of the sequences pairs required several weeks of CPU time using ten computer processors. Once the viable sequences are known, the steady state is computed as for the monomer case.

3 Results

3.1 Sequence diversity and the pressure to dimerize

Two structural models of a protein were considered in this work: a 2D and a 3D model. With either model, acceptable sequence pairs are those that not only fold, but also form a functional dimer with a sufficient cellular concentration (see Methods). In this section, we consider how the selective pressure to dimerize affects the sequence diversity.

2D on-lattice proteins Monomer evolution has been extensively studied with the 2D lattice model [11]. The model gives 1,081 monomer conformations, for 33,554,432 sequences, of which 12,386,286 fold [9]. Ten conformations are especially robust towards mutations, or "designable", with neutral networks of 40,000—68000 sequences. A pair of such proteins can adopt well over one billion possible sequence pairs ($40{,}000^2$). For these pairs, the selection stringency is characterized by the fractional population δ required for the functional dimer AB. A value of $\delta = 0.1$, for example, means that at chemical equilibrium, the dimer must be present at least 10% of the time. For reasons of computational cost, the analysis is limited to 16 2D dimers. They all involve monomeric neutral networks of about 10,000 sequences. The largest monomeric neutral networks (40,000—68,000 sequences) are too large to allow complete dimerization studies.

Fig. 1 shows the effect of the selection criterion on a typical dimer. For $\delta = 0$, all the pairs of sequences formed from the viable monomeric sequences of A and B are viable. As δ increases, alleles that dimerize poorly are increasingly eliminated, and the number of viable sequence pairs decreases rapidly. This decrease is accompanied by a fragmentation of the dimer's neutral network into smaller, disconnected pieces, as shown in Fig. 2. Interestingly, there is always one very large connected component, along with a number of much smaller components. The existence of a single large component implies that many sequences can be explored even though only point mutations are allowed.

The sequences eliminated by selection are those with too few hydrophobic residues at the functional interface. Indeed, Fig. 1 shows that the typical sequence, averaged over the steady state population, has an interface that is increasingly hydrophobic (darker) as δ increases. The neutral network for the pair (Fig. 1, lower panels) is increasingly depleted. This is seen by the decreasing number of red dots going from left to right in the lower panels of Fig. 1. Despite this depletion, the viable sequences of A and B remain very diverse: the red dots are not grouped in one part of their respective neutral networks, but are widely distributed throughout the network.

Another, more quantitative measure of sequence diversity is given by the network diameters. The diameter of a neutral network is defined as the largest number of point mutations separating any two viable alleles [25]. In Fig. 3A, the neutral network diameters in the absence (D) and presence (D') of selection for dimerization are shown as a histogram. We consider each 2D protein in turn, with its neutral network of sequences (1081 networks in all). The dimerization condition (when applied) requires that this protein dimerize specifically with another, particular protein (not shown), chosen arbitrarily. The dimer concentration threshold for viability was set to $\delta = 0.2$. Although the networks shrink when the dimerization condition is applied (many alleles are no longer viable), the diameters shrink very little: D' is typically only 1–2 units (amino acids) smaller than D. Similarly, the "distance" between two protein folds can be defined as the number of mutations needed to convert one fold into the other. Fig. 3B shows that for the 2D model, the distances between folds increase only slightly (by 1–2 amino acids) under the dimerization condition. In fact, the sequence diversity is such that for moderate values of δ, and for typical pairs of 2D proteins, almost every sequence in the neutral network of A has at least one B sequence with which it can form a viable dimer.

3D off-lattice proteins The second structural model is the three-dimensional, off-lattice model [10, 28]. For Grb2 and Vav, it gives 31469 and 29667 different HP profiles. These profiles lead to almost 10^9 possible pairs of HP profiles. The selection for dimerization is determined by the Z-score of the native, functional structure, compared to the Z-score of the decoys. For the functional structure to be populated at least 10% of the time in the cell, there cannot be more than 9 alternate structures of lower energy. Therefore, the rank k we require for the native Z-score was varied from 1 to 10. With $k = 1$, only 537 sequence pairs were viable. With $k =$ 3–10, there were between 67,291 and 470,334 viable pairs of HP profiles. The latter value represents just $\frac{1}{2000}$th of all possible pairs. The same reduction is seen in the 2D case with a δ of about 0.25 (Fig. 2; Table 1). The rather small, viable, 3D fraction is related to the larger size of the 3D dimer interface. The Grb2:Vav complex involves 14–16 amino acids on each partner. A reduction factor of $\frac{1}{2000}$ for

the number of sequence profiles can be obtained by fixing the profile (H or P) of just 11 positions in the dimer (since $2^{11} \approx 2000$), or 5–6 positions on each monomer. Fixing 11 positions appears reasonable with respect to the size and diversity of typical protein–protein interfaces [34].

Similar to the 2D case, most (70%) of the Grb2 sequences have at least one Vav sequence with which they are able to form a viable dimer. The Vav sequences are less diverse: only 8% of the monomeric sequences survive when $k = 10$. This may be an indication of insufficient sequence sampling during the Monte Carlo simulation of the Vav monomer. Longer (and expensive) simulations are needed to test this further. However, the cost of the present calculations is already close to the limit of what is feasible (weeks of CPU time to construct the dimeric neutral network using $\sim$10 recent processors).

3.2 Independence between mutational robustness and dimerization ability

The selective pressure to expose hydrophobic residues might be expected to correlate with a lower mutational robustness n (since the dimerizing sequences are more constrained). Similarly, the folding temperature T_f (which is a measure protein stability; see Methods), might be expected to decrease as the stringency of selection increases, since exposing hydrophobic residues tends to lower stability. In fact, when one considers the viable sequences of A and B, their ability to dimerize is actually not correlated with either n (Fig. 4A) or the folding temperature T_f (Fig. 4B). Sequences with very diverse values of n and T_f have the same ability to dimerize, as measured by the cellular concentration $[AB]_{func}$ of the functional dimer. The data in Fig. 4A,B correspond to a representative dimer, made of a particular pair of protein structures. For the 16 2D dimer structures we analyzed, the correlation coefficients range from -0.026 to 0.042 for n (respectively, -0.035 to 0.100 for T_f). Analyzing the viable 2D protein sequences, we find that in fact, dimerization can be enhanced without increasing the number of exposed hydrophobic residues. Instead, hydrophobic residues can be moved to the interface region from another part of the protein surface. Within the simple 2D model, this operation has very little effect on the protein stability and folding temperature, which explains that T_f and $[AB]_{func}$ are uncorrelated. Since T_f and n are known to be strongly correlated [16], n must also be uncorrelated with $[AB]_{func}$.

For the Grb2:Vav dimer, results are similar (Fig. 4C). Here, the mutational robustness n is plotted against the dimerization energy Z-score. The data correspond to a dimerization selection threshold of $k = 10$, for a total of 470,334 viable dimer sequence profiles. The correlation coefficient between n and Z is 0.046.

The independence between n or T_f is likely to hold qualitatively for real proteins.

For a given dimer interface AB, we expect that the interface sequences will be rather contrained [34], whereas a wider range of amino acid types may be found on the remaining parts of the surfaces and in the proteins' core, leading to a wide range of protein stabilities. However, a systematic analysis of sequence conservation and protein stability in families of dimerizing proteins would be needed to make this statement quantitative. Thermodynamic data is scarce, and such an analysis is beyond the scope of this study.

3.3 The steady state is enriched in functional alleles

Previous studies of single protein evolution have revealed an enrichment in mutational robustness in the steady state [8, 9, 16]. Sequences in the core of the neutral network are overpopulated, while those at the edge of the network, with fewer connections, are underpopulated. This steady state enrichment is preserved under the dimerization constraint, as shown in Fig. 5A. The extent of enrichment is similar to the pure monomeric case; see Table 1 for illustrative, numerical values for a particular complex. A similar enrichment is observed for the 3D proteins (Fig. 5B). The agreement between the 2D and 3D models provides encouraging evidence that this behavior does not depend on model details. The dimer folding temperature is also enriched in the steady state (Fig. 5C). This is consistent with the known correlation between n and T_f.

In a similar way, the cellular concentration of the functional dimer is enriched in the steady state (Fig. 5D). In other words, the alleles that are good dimerizers are overpopulated. Thus, neutral evolution leads not only to increased mutational robustness, but to increased concentrations of the functional species present in the average cell. This result was harder to anticipate, partly because the concentration of functional dimer is not correlated or closely-related to either n or T_f. The enrichment in n arises because highly-connected sequences are grouped in the middle of the neutral network. In effect, the enrichment arises because n varies in a smooth, continuous manner over the network, so that robust sequences are close to other robust sequences. But we saw above that the sequences satisfying the dimerization threshold are widely distributed throughout the underlying monomeric network (Fig. 1). Therefore, it was not obvious ahead of time that dimerization ability would vary sufficiently smoothly and continuously.

In Fig. 6, we define an enrichment factor for dimerization ability, $\Phi([AB]) = \langle [AB] \rangle_{ss} / \langle [AB] \rangle_{random}$, where $\langle [AB] \rangle_{ss}$ is the cellular concentration of the functional dimer averaged over the steady state sequences, and $\langle [AB] \rangle_{random}$ is the concentration averaged over all the viable sequences, regardless of their population. Typical values of $\Phi([AB])$ are greater than 1, corresponding to enrichment. The enrichment in functional species is strongest when the selection criterion is only moderately stringent: $\Phi([AB]) \approx 1.2$–1.3 when $\delta \approx 0.01$–0.10. As δ increases, selection be-

comes more stringent and the set of viable sequences is increasingly depleted (Fig. 1). Enrichment then decreases. A model with larger proteins and/or a more complex amino acid alphabet would probably allow a greater enrichment, extending to higher selection stringencies.

For the 3D model, it is harder to characterize the enrichment (if any) in steady state dimerization ability, because there are too many possible structures and the dimer concentrations cannot be computed. Nevertheless, the steady state populations of the viable alleles are available, so that we can compare the typical Z-scores in the steady state population and a random population. This is done in Table 2. An enrichment factor $\Phi(Z)$ is defined in the same way as $\Phi([AB])$, above. As the selection threshold k increases from 3 to 10, $\Phi(Z)$ first increases from 1.07 to 1.18 ($k = 4$ or 5), then decreases to 1.04 ($k = 9$ or 10). As in the 2D case, the enrichment is maximal for an intermediate selection stringency. The maximum enrichment factor is roughly comparable in the 2D and 3D cases, even though the measures of dimerization ($[AB]_{func}$ and Z) are obviously different. Again, the qualitative 2D-3D agreement is encouraging.

3.4 Functional alleles form a sequence funnel

The enrichment in functional species is strongest for sequence pairs near the "prototype" pair, defined as the most populated pair in the steady state [9]. This is illustrated in Fig. 7A for the 2D proteins. The mean concentration $[AB]_{func}$ of the functional dimer is plotted for each viable sequence pair, as a function of its distance from the prototype pair (for a representative dimer and a few values of the selection threshold, δ). The concentration $[AB]_{func}$ varies widely, but the mean value drops off rapidly and smoothly as one moves away from the prototype sequence. Similar behavior is seen for other 2D dimers. Thus, the sequences responsible for the functional enrichment are grouped in the center of the neutral network, forming a basin, or funnel in sequence space.

With the 3D model, dimerizing ability is measured by the Z-score. We saw above that the corresponding enrichment factor, $\Phi(Z)$, was slightly smaller than $\Phi([AB])$ in the 2D case. Nevertheless, a basin of good dimerizers is also seen with the 3D model, although the funnel shape is somewhat less pronounced (Fig. 7B). The funnel shape flattens out at a distance of about 6–7 from the prototype sequence. A small number of very good dimerizers are actually found outside the basin, at distances of 13–14 from the prototype.

4 Discussion

Protein–protein interactions are central to cellular organization, and must have appeared at a very early stage of evolution. To better understand their effects, we considered here two simple, "toy" models of protein structure and evolution, and determined the effect of explicitly selecting for protein–protein interactions. By employing a 2D, lattice representation of protein structure and binary, hydrophobic/hydrophilic sequences, exact calculations could be performed. The 3D, off-lattice model gives a similar qualitative picture. For example, the 3D model also predicts that mutationally robust sequences are overrepresented in the steady state, in agreement with the well-known result of lattice models [8, 16, 35].

Here, a functional coupling between pairs of genes was added to two previous evolutionary models: the two proteins of interest must associate in order to function. The steady state enrichment in mutational robustness is preserved under this additional constraint. Sequence diversity remains very large when dimerization is required, even though only a fraction of alleles survive under these more selective conditions. The sequence diversity is reflected, for example, by the wide range of protein robustnesses and folding temperatures that can lead to the same dimerization ability. It is somewhat unexpected that as the pressure to dimerize is increased and more and more sequence pairs are eliminated, the viable sequences continue to be largely grouped in a single, continuous network (Fig. 2), instead of splitting into many small, disconnected networks. A single, continuous network makes it easier to explore sequence space, since single mutations can be used more extensively, rather than large hops involving several mutations at a time.

The present treatment corresponds to neutral evolution, in the sense that it treats all viable alleles as equally fit. The model has a flat fitness plateau—the neutral network, surrounded by a sea of non-viable sequence pairs. Evolution takes the form of a random diffusion throughout the neutral network. This neutral picture should be in qualitative accord with real proteins. Neutral mutations are very common in proteins, as shown by the sequence diversity associated with modern protein folds. The neutral model predicts that the probability for a protein to retain its native fold decreases exponentially with the number of mutations, at least for the first few mutations; this prediction agrees with recent experimental observations [36]. The proportions of tolerated mutations computed here for the individual "proteins" are also comparable to those of several real protein folds [37].

Our model selects for a minimal level of functional ability, determined by the chosen dimer concentration threshold, δ, or the Z-score threshold, k. The steady state dynamics then lead to a population that is enriched in sequences that are good dimerizers, well beyond the minimal ability needed to survive. In other words, the functionality of the typical sequence pairs has been enhanced by the evolutionary dynamics: adaptive evolution has occurred. The adaptation arises because of the

plateau form of the neutral network and because of the funneled organization of the sequences within the network. This result is analogous to the enrichment in mutational tolerance for the single protein models. Nevertheless, we believe it was not obvious that a similar result would hold for dimer functionality. Indeed, the dimer sequences are widely dispersed throughout the monomeric networks, whereas the steady state enrichment in a given property (mutational tolerance, dimerization ability) is closely related to its continuity over the set of viable sequences. Little or no correlation is seen between the "monomeric" properties, n and T_f, and the dimerization ability. In addition, while the functional enrichment first increases with δ, it then decreases for larger values of δ. Back-of-the-envelope predictions for dimers are difficult because of the complex chemical equilibria involved (see Methods).

The timescale of the present model is set by the mutational probability per unit time, α in Eq. (2). The experimental timescale to fix point mutations within a population (the "molecular clock") depends on the organism, the population size, and the protein. For example, highly expressed proteins tend to evolve more slowly [38]. For eubacteria with a generation time of minutes, a neutral mutation should appear in a typical protein about $N = 10^1$–10^3 times a day within a large colony [39]. The given range corresponds to different colony sizes (10^5–10^7 individuals); it can be expanded if one considers longer generation times or artificially accelerated mutation rates (e.g., in the presence of chemical mutagens). This range for N should encompass two extreme regimes, identified in computer simulations [40]. At high mutation rates ($N \gg 100$), the population is expected to sample the steady state, so that the enrichment pheneomena predicted above should be visible. At low mutation rates ($N \ll 100$), the population is expected [40] to behave like a random sample drawn from the neutral network, so that no enrichment should be observed. Thus, the role of the steady state dynamics in elevating the average functionality could be experimentally tested by comparing two such colonies.

Some of our predictions could be also tested by analyzing experimental protein sequences. The weak correlation between dimerizing ability and protein stability is in accord with our knowledge of protein–protein interfaces. Typical protein–protein interfaces have a few amino acids forming a central hydrophobic patch; small, polar, mutational "hotspots" are also frequent [34]. The evolutionary contraints on these local surface patches should have a limited effect on other surface and core regions, so that a large range of protein stabilities can be achieved despite the constraints. An empirical analysis of the stabilities of dimerizing proteins could shed further light. Conversely, it would be interesting to compare the dimerization abilities of very stable proteins, such as those of thermophilic organisms.

From the present models, the alleles that are populated in the steady state are enhanced in their functional ability. This should allow an increased resistance to

environmental change, or adaptability [17]. Indeed, a strong dimer is more likely to be preserved under a change in the surrounding temperature or pH, for example. They should also provide an increased ability to evolve, invent new functions, or acquire new binding partners. Indeed, after a gene duplication event, a protein A that starts out as a strong dimerizer will be better able to explore mutations that allow it to co-evolve with its partner B, or to dimerize with other, existing, homologues of B. This effect, which arises from a very simple, minimal model of protein evolution, should lead to an enhanced ability to create new interacting pairs of proteins, and could have played a role in the early emergence of protein–protein interaction networks.

References

[1] Zhao, H., Chockalingam, K., and Chen, Z. Directed evolution of enzymes and pathways for industrial biocatalysis. *Curr. Opin. Biotech. 13* (2002), 104–110.

[2] Kimura, M. *The neutral theory of molecular evolution.* Cambridge University Press, 1983.

[3] Ohta, T. Near-neutrality in evolution of genes and gene regulation. *Proc. Natl. Acad. Sci. USA 99* (2002), 16134–16137.

[4] Eigen, M., McCaskill, J., and Schuster, P. The molecular quasispecies. In *Advances in Chemical Physics, vol. 75*, I. Prigogine and S. Rice, Eds. Wiley, New York, 1989, pp. 149–263.

[5] Sali, A., Shakhnovich, E., and Karplus, M. How does a protein fold? *Nature 369* (1994), 248–251.

[6] England, J. L., Shakhnovich, B. E., and Shakhnovich, E. I. Natural selection of more designable folds: a mechanism for thermophilic adaptation. *Proc. Natl. Acad. Sci. USA 100*, 15 (2003), 8727–8731.

[7] Wolynes, P. Energy landscapes and solved protein folding problems. *Phil. Trans. Roy. Soc. A 363* (2005), 453–464.

[8] Bornberg-Bauer, E., and Chan, H. S. Modeling evolutionary landscapes: Mutational stability, topology, and superfunnels in sequence space. *Proc. Natl. Acad. Sci. USA 96* (1999), 10689–10694.

[9] Xia, Y., and Levitt, M. Roles of mutation and recombination in the evolution of protein thermodynamics. *Proc. Natl. Acad. Sci. USA 99* (2002), 10382–10387.

[10] Bastolla, U., Porto, M., Roman, E., and Vendruscolo, M. Lack of self-averaging in neutral evolution of proteins. *Phys. Rev. Lett. 89* (2002), 208101–1 – 208101–4.

[11] Xia, Y., and Levitt, M. Simulating protein evolution in sequence space and structure space. *Curr. Opin. Struct. Biol. 14* (2004), 202–207.

[12] Williams, P. D., Pollock, D. D., and Goldstein, R. A. Selective advantage of recombination in evolving populations: A lattice model study. *Int. J. Mod. Phys. 17* (2005), 75–90.

[13] Kauffman, S. A. *The origins of order. Self-organization and selection in evolution.* Oxford University Press, New York, 1993.

[14] Earl, D. J., and Deem, M. W. Evolvability is a selectable trait. *Proc. Natl. Acad. Sci. USA 101* (2004), 11531–11536.

[15] Tiana, G., Shakhnovich, B. E., Dokholyan, N. V., and Shakhnovich, E. I. Imprint of evolution on protein structures. *Proc. Natl. Acad. Sci. USA 101*, 9 (2004), 2846–2851.

[16] Taverna, D. M., and Goldstein, R. A. Why are proteins so robust to site mutations? *J. Mol. Biol. 315*, 3 (2002), 479–484.

[17] Bloom, J. D., Labthavikul, S. T., Otey, C. R., and Arnold, F. H. Protein stability promotes evolvability. *Proc. Natl. Acad. Sci. USA 103* (2006), 5869–5874.

[18] Chan, H. S., and Bornberg-Bauer, E. Perspectives on protein evolution from simple exact models. *Applied Bioinformatics 1* (2002), 121–144.

[19] Bloom, J. D., Wilke, C. O., Arnold, F. H., and Adami, C. Stability and the evolvability of function in a model protein. *Biophys. J. 86* (2004), 2758–2764.

[20] Williams, P. D., Pollock, D. D., and Goldstein, R. A. Evolution of functionality in lattice proteins. *J Molec. Graph. Model. 19* (2001), 150–156.

[21] Blackburne, B. P., and Hirst, J. D. Evolution of functional model proteins. *J. Chem. Phys. 115* (2001).

[22] Blackburne, B. P., and Hirst, J. D. Population dynamics simulations of functional model proteins. *J. Chem. Phys. 123* (2005), 154907.

[23] Gavin, A., and Superti-Furga, G. Protein complexes and proteome organization from yeast to man. *Curr. Opin. Chem. Biol. 7* (2003), 21–27.

[24] Mika, S., and Rost, B. Protein-protein interactions more conserved within species than across species. *PLOS Comp. Biol. 2* (2006), 698–709.

[25] Albert, R., and Barabasi, A. L. Statistical mechanics of complex networks. *Rev. Mod. Phys. 74* (2002), 47–97.

[26] Tiana, G., Provasi, D., and Broglia, R. A. Role of bulk and of interface contacts in the behaviour of lattice model dimeric proteins. *Phys. Rev. Lett. 67* (2003).

[27] Li, H., Helling, R., Tang, C., and Wingreen, N. S. Why do proteins look like proteins? *Science 273* (1996), 666–669.

[28] Ueda, Y., Taketomi, H., and Go, N. Studies on protein folding, unfolding, and fluctuations by computer simulation. 2. 3-dimensional lattice model of lysozyme. *Biopolymers 17* (1978), 1531–1548.

[29] Bastolla, U., Farwer, J., Knapp, E., and Vendruscolo, M. How to guarantee optimal stability for the most representative structures in the Protein Data Bank. *Proteins 44* (2001), 79–96.

[30] Launay, G., Mendez, R., Wodak, S., and Simonson, T. Recognizing protein–protein interfaces with reduced amino acid alphabets. *BMC Bioinformatics submitted* (2007), 0000.

[31] Tuffery, P., Etchebest, C., Hazout, S., and Lavery, R. A new approach to the rapid determination of protein side chain conformations. *J. Biomol. Struct. Dyn. 8* (1991), 1267.

[32] Press, W., Flannery, B., Teukolsky, S., and Vetterling, W. *Numerical Recipes.* Cambridge University Press, Cambridge, 1986.

[33] Brooks, B., Bruccoleri, R., Olafson, B., States, D., Swaminathan, S., and Karplus, M. Charmm: a program for macromolecular energy, minimization, and molecular dynamics calculations. *J. Comp. Chem. 4* (1983), 187–217.

[34] Wodak, S. J., and Janin, J. Structural basis of macromolecular recognition. *Adv. Prot. Chem. 61* (2002), 9–73.

[35] Xia, Y., and Levitt, M. Funnel-like organization in sequence space determines the distributions of protein stability and folding rate preferred by evolution. *Proteins 55*, 1 (2004), 107–114.

[36] Bloom, J., Silberg, J., Wilke, C., Drummond, D., Adami, C., and Arnold, F. Thermodynamic prediction of protein neutrality. *Proc. Natl. Acad. Sci. USA 102* (2005), 606–611.

[37] Guo, H., Choe, J., and Loeb, L. Protein tolerance to random amino acid change. *Proc. Natl. Acad. Sci. USA 101* (2004), 9205–9210.

[38] Drummond, D., Bloom, J., Adami, C., Wilke, C., and Arnold, F. Why highly expressed proteins evolve slowly. *Proc. Natl. Acad. Sci. USA 102*, 40 (2005), 14338–14343.

[39] Elena, S., Cooper, V., and Lenski, R. Punctuated evolution caused by selection of rare beneficial mutations. *Science 272* (1996), 1802–1804.

[40] Nimwegen, E. v., Crutchfield, J. P., and Huygen, M. Neutral evolution of mutational robustness. *Proc. Natl. Acad. Sci. USA 96* (1999), 9716–9720.

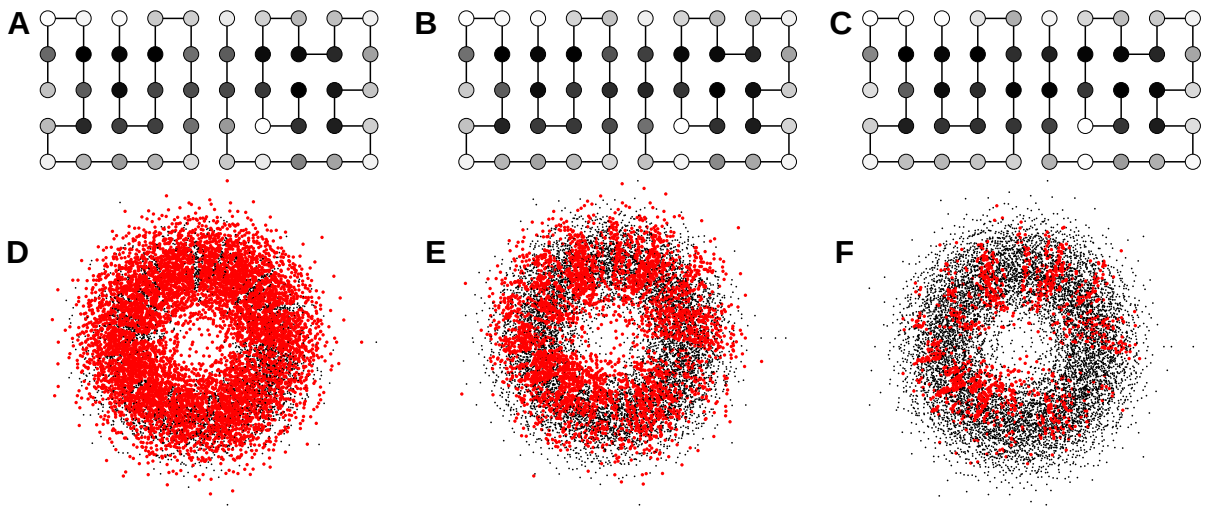


Figure 1: **An example of a 2D dimer. A)** Low selective pressure for dimerization: $\delta = 0.04$ (i.e., only sequences that lead to a protein fraction of at least 4% engaged in the functional dimer are viable). The amino acids are colored according to the mean sequence in the steady state population (hydrophobic: dark; hydrophilic: light). **B)** The same dimer under a moderate selective pressure for dimerization: $\delta = 0.1$ This leads to a more hydrophobic interface. **C)** The same dimer with $\delta = 0.2$. **D)** The corresponding neutral network for one of the protein partners when $\delta = 0.04$. Black dots represent viable monomer sequences; red dots represent sequences that survive under the dimerization condition. Connections between sequences are omitted for clarity. **E)** Idem, $\delta = 0.1$. **F)** Idem, $\delta = 0.2$.

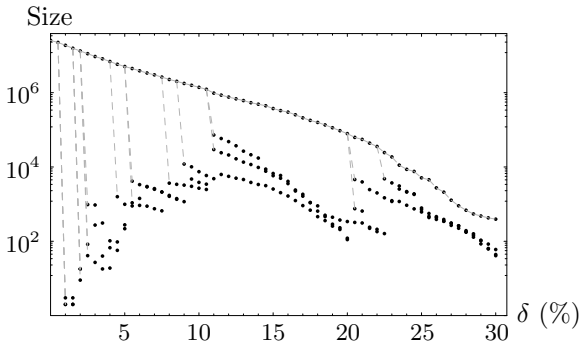


Figure 2: **Size of neutral network components.** For a representative 2D dimer, the size of the four largest components as a function of the selective pressure δ. As δ increases, there are fewer viable sequence pairs, but there is always a single connected component that is much larger than the other, small components. Dashed vertical lines are visual aids to show how the small components progressively break off from the largest one.

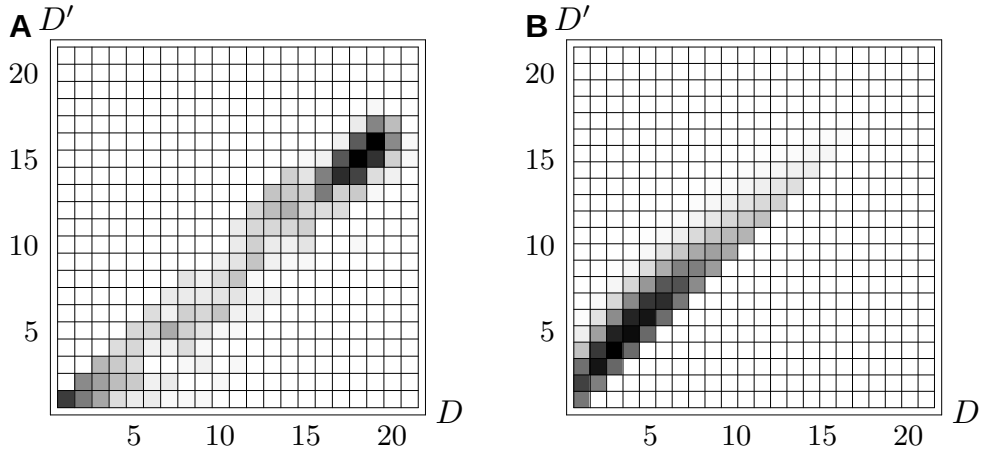


Figure 3: **Diversity of viable genotypes in the neutral networks. A)** The neutral network diameters in the absence (D) and presence (D') of selection for dimerization, shown as a 2D histogram. We consider each 2D protein in turn, with its neutral network of sequences (1081 networks in all). The dimerization condition (when applied) requires that this protein dimerize specifically with another, particular protein (not shown), chosen arbitrarily. The dimer concentration threshold for viability was set to $\delta = 0.2$. The diameter represents the largest "distance" between any two sequences in the neutral network (the number of amino acid mutations that separate them) [25]. Although the networks shrink when a dimerization condition is applied (many alleles are no longer viable), the diameters shrink very little: D' is typically only 1–2 units (amino acids) smaller than D. **B)** The distance between neutral networks in the absence (d) and presence (d') of selection for dimerization, shown as a 2D histogram. The dimerization selection criterion is the same as above. Under the dimerization selection criterion, the networks shrink and their distances therefore increase ($d' > d$), but only by 1–2 units.

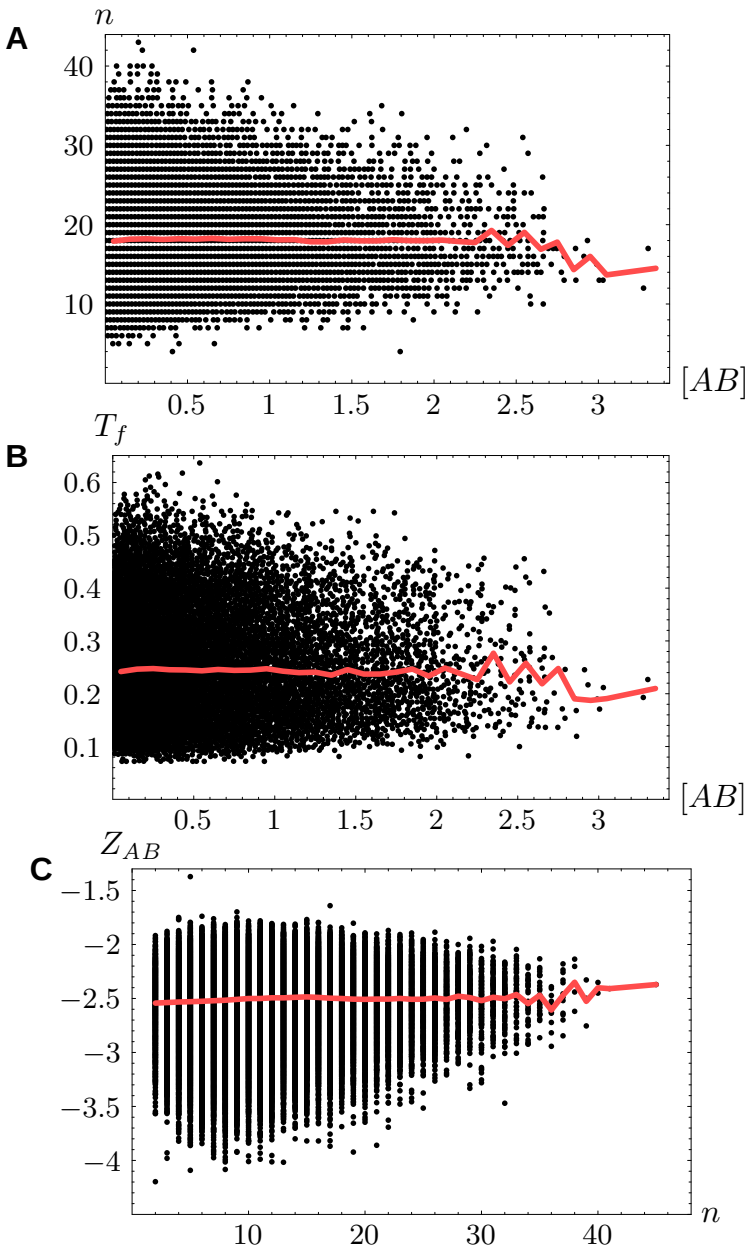


Figure 4: **Absence of correlation of n and T_f with dimerization ability. A)** For a representative 2D protein dimer, we show the mutational robustness n of each viable sequence pair, versus the cellular concentration $[AB]_{func}$ of the functional dimer at chemical equilibrium for that pair. The red line represents the average over the distribution of n values for each value of $[AB]_{func}$. n and $[AB]_{func}$ are seen to be uncorrelated. **B)** For the same 2D dimer, we show the folding temperature of each viable sequence pair as a function of $[AB]_{func}$. **C)** For the Grb2:Vav 3D dimer, we show the Z value (which measures the dimerization ability) of each viable sequence pair as a function of n. The two are seen to be uncorrelated, as in the 2D case.

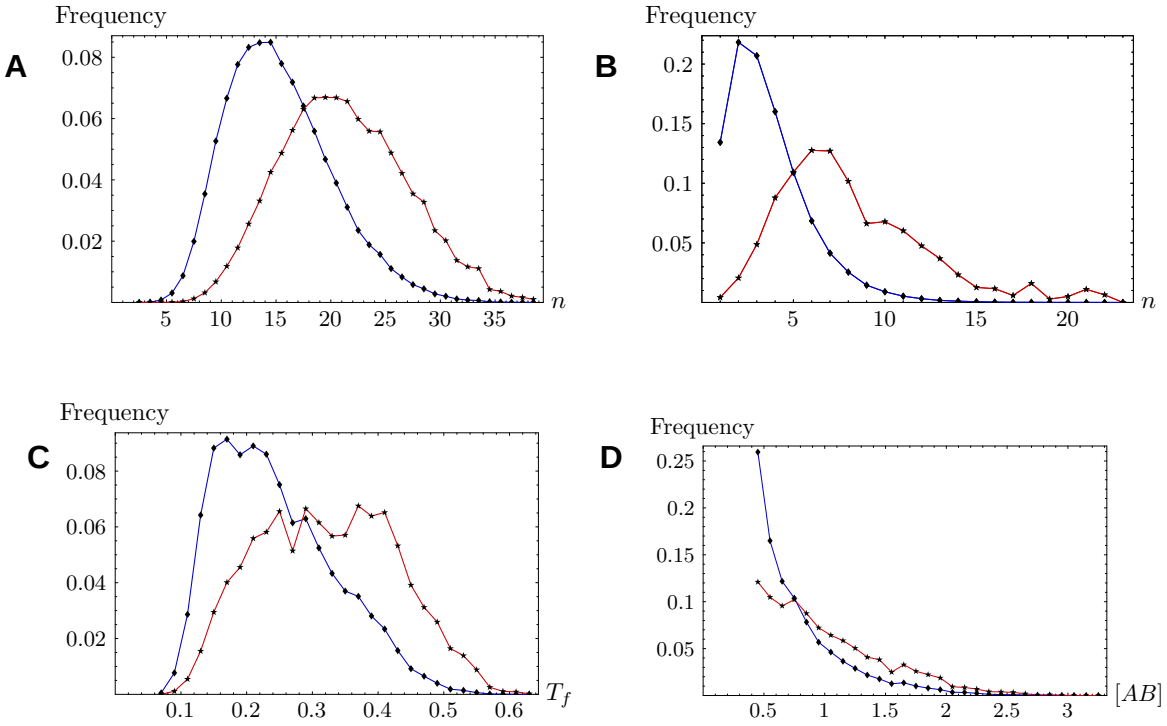


Figure 5: **The population dynamics enhance robustness and functionality.** **A)** The distribution of mutational robustness n for a given 2D protein dimer. Red: the steady state population. Blue: a population drawn randomly from the neutral network. **B)** Idem for the 3D, off-lattice Grb2:Vav 3D dimer. **C)** The folding temperature distribution for the same 2D dimer; red: steady state population; blue: random population. **D)** The distribution of the equilibrium concentration $[AB]_{func}$ of the functional dimer, for the same 2D dimer; red: steady state population; blue: random population.

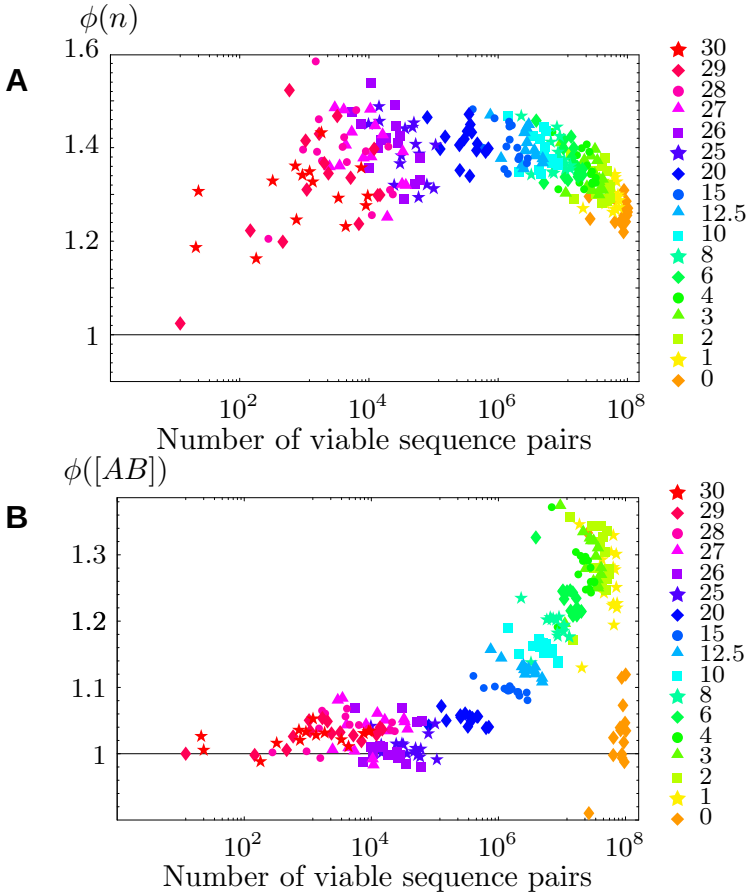


Figure 6: **The population dynamics enhance robustness and functionality. A)** $\Phi(n) = \langle n \rangle_{ss} / \langle n \rangle_{random}$ measures the enrichment in mutational robustness due to the steady state dynamics. Each colored point corresponds to a 2D protein pair subjected to a dimerization condition. Each color corresponds to a particular level of the selection stringency, indicated by the value of the concentration threshold δ (in %, legend on right). Data are shown for a selection of 15 dimers. **B)** $\Phi([AB]) = \langle [AB] \rangle_{ss} / \langle [AB] \rangle_{random}$ measures the enrichment in dimerization ability due to the steady state dynamics: $\langle [AB] \rangle_{ss}$ is the cellular concentration of the functional dimer averaged over the steady state sequences; $\langle [AB] \rangle_{random}$ is the value averaged over the viable sequences, regardless of their population. A value of $\Phi([AB])$ greater than 1 indicates an enrichment of the steady state population in sequences that readily dimerize. Colors indicate the selection stringency. Data are shown for a selection of 10 dimers.

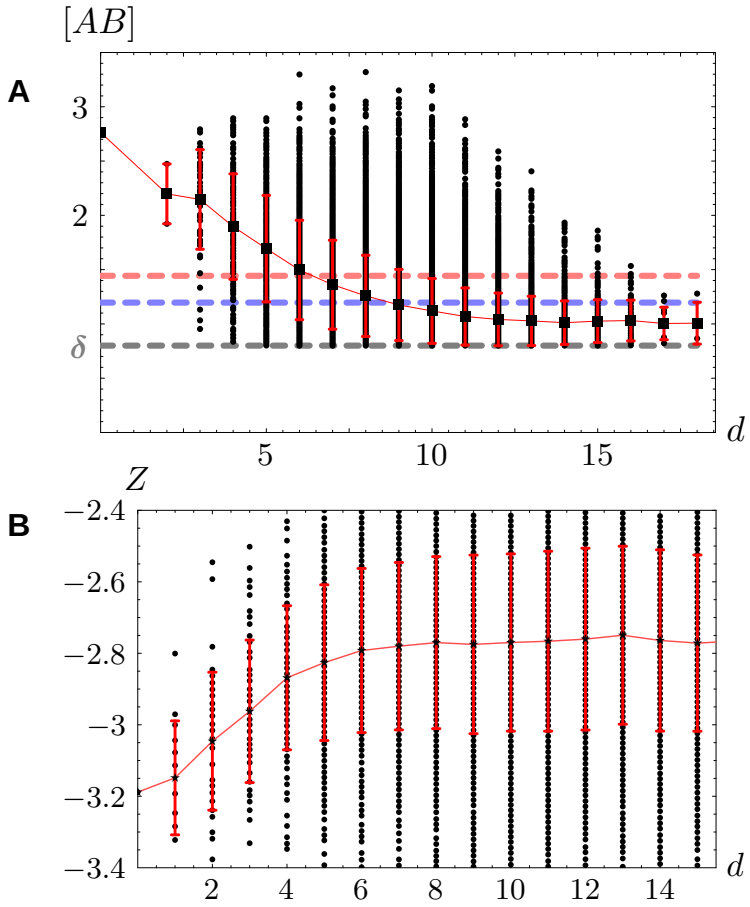


Figure 7: **Emergence of a "functional funnel" in sequence space. A)** 2D dimer: The mean concentration $[AB]_{func}$ of the functional dimer as a function of the distance of each viable sequence pair from the prototype pair (the most populated pair in the steady state). Data are shown for a representative dimer and a selection threshold of $\delta = 0.08$ (grey dashed horizontal line). There are no sequence pairs below the dashed line, because such sequences are not viable, by definition. Red curve: the mean value for each distance. Red vertical bars indicate the standard deviation at each distance. Red dashed horizontal line: overall steady state average. Blue dashed horizontal line: average over a random set of sequences. The concentration $[AB]_{func}$ varies widely, but the mean value drops off rapidly and smoothly as one moves away from the prototype sequence pair. **A)** Similar representation for the 3D Grb2:Vav dimer: the dimerization Z score as a function of the distance from the prototype sequence.

δ	$\langle n \rangle^A_{random}$	$\langle n \rangle^B_{random}$	$\langle n \rangle_{random}$	$\langle n \rangle^A_{ss}$	$\langle n \rangle^B_{ss}$	$\langle n \rangle_{ss}$	Survival #	Survival %
0.00	9.18	8.96	18.14	11.48	11.84	23.32	96,108,582	100
0.01	8.47	8.36	16.84	10.95	11.39	22.34	68,290,887	71
0.02	8.03	7.96	16.00	10.61	11.00	21.61	47,695,260	50
0.03	7.67	7.63	15.31	10.30	10.71	21.01	34,273,488	36
0.04	7.43	7.42	14.85	10.10	10.57	20.66	25,785,086	27
0.06	7.06	7.01	14.07	9.69	9.99	19.68	14,825,259	15
0.08	6.71	6.68	13.39	9.38	9.69	19.07	8,983,407	9.3
0.10	6.45	6.42	12.87	9.08	9.40	18.48	5,522,233	5.8
0.15	5.86	5.78	11.64	8.32	8.53	16.86	1,626,428	1.7
0.20	5.18	5.02	10.20	7.48	7.52	15.01	353,538	0.37
0.25	4.43	4.16	8.60	6.21	6.20	12.41	47,192	0.05

Table 1: Mutational robustness as a function of the functional constraint (measured by δ). Data is shown for a representative 2D dimer. The two partners each have about ten thousand viable sequences. The mean mutational robustness for randomly chosen sequences is $\langle n \rangle_{random}$. The robustness averaged over the steady state is $\langle n \rangle_{ss}$. These values correspond to the overall mutational robustness of the AB dimer. The separate contributions of each partner, A and B, are also shown. The number and percentage of viable sequence pairs are shown.

k	viable sequences	$\langle n \rangle_{ss}$	$\Phi(Z)$
1	537	14.13	1.05
2	30801	4.65	1.07
3	67291	3.50	1.08
4	109954	3.07	1.18
5	157903	3.07	1.18
6	211133	3.01	1.06
7	269022	3.00	1.07
8	331852	2.96	1.07
9	398776	2.99	1.04
10	470334	3.00	1.04

Table 2: Steady state enrichment factor $\Phi(Z)$ of the dimerization ability (measured by the Z-score) for the Grb2:Vav 3D dimer as a function of the functional constraint (measured by k). The number of viable sequence profile pairs and the mean mutational robustness are also shown.

Conclusion

A question arose as to whether we were covering the field that it was intended we should fill with this manual.

Richard R. Donnelley, *Proceedings, United Typothetæ of America*

D'après Darwin, l'évolution des populations est l'œuvre de la sélection positive. L'évolution qu'il envisageait était l'*évolution phénotypique*. Au niveau moléculaire, c'est la théorie de l'évolution neutre qui prévaut : de nombreuses mutations sont neutres et leur effet est imperceptible pour la sélection naturelle. La sélection naturelle n'est pas absente mais sa composante majeure est la sélection négative qui exclut les mutations les plus délétères. La compréhension des mécanismes de l'évolution, des gènes aux individus en passant par les macromolécules biologiques, implique donc de prendre en compte une double correspondance génotype-phénotype, phénotype-*fitness*.

Des modèles simples d'ARNt et de protéines permettent d'établir *ab initio* un modèle de cette correspondance. De la structure induite par cette correspondance dans l'espace des génotypes dépendent la dynamique et la convergence de l'évolution des populations. Cela a été démontré théoriquement et expérimentalement par l'existence des quasi-espèces. L'objet de travaux précédents et des nôtres a été de montrer qu'au niveau moléculaire, l'organisation des génotypes est un ensemble de séquences de fonctionnalité équivalente interconnectées en « réseaux neutres ». Les réseaux neutres sont organisés en *superfunnel* caractérisés par un noyau de séquences densément connectées entouré par un anneau de séquences moins connectées. Les séquences du noyau sont plus robustes aux mutations et sont plus stables thermodynamiquement.

Nous avons montré dans le chapitre « Évolution des protéines monomériques » que cette organisation émergeait dans un modèle de protéine sur réseau en utilisant diverses

matrices d'énergie typiques de protéines ou non. Nous avons également vu que cette organisation était préservée lorsque des structures tridimensionnelles hors-réseau, plus réalistes, étaient envisagées. Ces mêmes structures permirent de mettre en évidence que le *superfunnel* persistait lorsque des alphabets plus complets étaient considérés. Nos résultats soutiennent donc les présomptions de BORNBERG-BAUER et CHAN quand ils affirment :

> In light of these considerations, the generality of our conclusion may transcend the two [lattice] models studied here, and may also be independent of questions regarding what model contact interactions and chain representations are more protein like [21, p. 10693].

Encore fallait-il prouver que leur observations ne dépendaient pas principalement des simplifications des conformations sur réseau et surtout de l'alphabet HP.

De ces réseaux neutres peut émerger un comportement adaptif : la population a tendance à peupler préférentiellement le noyau des séquences robustes et stables. Cet effet a pu être mis en évidence à l'aide d'un modèle de population infinie dont nous avons analysé les limites. D'après nos résultats, il ne peut survenir que si le nombre de mutants produits par génération est suffisant ($M\mu \gg 1$, avec les notations de la section « Effet de la taille de la population », p. 121). Nous avons également émis l'hypothèse que cette condition était fortement liée à la nécessité d'un polymorphisme génétique dans la population pour que l'évolution élimine les génotypes peu robustes aux mutations. En revanche, l'apparition du comportement adaptatif ne dépend pas de la taille du réseau, bien qu'elle puisse dépendre de sa topologie[35] et que la taille du réseau puisse affecter la *vitesse* de convergence vers l'état stationnaire. Lorsque la condition $M\mu \gg 1$ n'est pas remplie, la population effectue une marche aléatoire sur le réseau neutre. La probabilité qu'une séquence particulière du réseau neutre apparaisse, dans ce cas, est une distribution uniforme, après un temps long. Lorsque $M\mu$ est faible, on retrouve les conditions d'application du modèle de SELLA et HIRSH qui prédisent effectivement une distribution équiprobable de la population [163].

Nous avons défendu l'idée que les interactions protéine-protéine constituaient un modèle de fonctionnalité pertinent. Traditionnellement la génétique des populations s'est développée sur la sélection des gènes (*gene centricity*) plutôt que des individus (*organism centricity*), pourtant ce sont bien les individus qui sont sujets à la sélection naturelle. Les modèles structuraux bâtis jusqu'à présent sont encore loin de pouvoir décrire la complexité de l'appareil téléonomique de MONOD. La modélisation des interactions protéine-protéine est le liant de cette complexité et un élément important des interactions épistatiques, il est

35. Mais comme nous l'avons mentionné plus haut, la topologie des réseaux neutres semble universelle.

par conséquent nécessaire de s'attacher à en dénouer les mécanismes et les conséquences. Ce travail représente un premier pas dans cette direction.

L'extension de ces modèles monomériques à des protéines dimériques maintient les conclusions tirées à partir des réseaux neutres monomériques. La sélection négative est importante dans les réseaux neutres dimériques (les hyper-réseaux neutres) mais les génotypes viables sont encore dans une grande majorité connectés. L'évolution neutre augmente encore, en moyenne, la robustesse mutationnelle et la stabilité thermodynamique. Nous observons également une augmentation moyenne de l'aptitude à dimériser due à une organisation en « *superfunnel* fonctionnel ». Le facteur d'enrichissement de la robustesse mutationnelle est accru à mesure que la contrainte fonctionnelle augmente. Cela est contraire à la relation entre le facteur d'enrichissement et la taille des réseaux neutres tirée des études monomériques. Nous avons proposé que ce phénomène ne pouvait se manifester que par l'existence d'interactions épistatiques entre les deux gènes codant pour les protéines monomériques. Cet effet pourrait en partie rendre compte de la difficulté à mettre en évidence une perte de vitesse évolutive pour les protéines engagées dans de nombreuses interactions protéine-protéine. Le modèle de dimère tridimensionnel hors-réseau est, par ailleurs, en excellent accord avec les observations récentes de conservation des résidus à l'interface.

Il a été observé que la recombinaison pouvait être un élément important à la fois dans la découverte phénotypique [35] et dans le peuplement préférentiel du centre du *superfunnel* [203]. Ce deuxième effet démontre la capacité de la recombinaison à contrecarrer l'accumulation de mutations délétères (hypothèse de Muller [140]). Des données expérimentales peuvent s'interpréter à la lumière de ces résultats : si ces conclusions sont exactes, les gènes dans lesquels de nombreuses recombinaisons surviennent doivent être plus robustes aux mutations et donc évoluer plus rapidement. Les gènes de *D. melanogaster* and *D. simulans* situés dans des régions chromosomiques possédant de faibles niveaux de recombinaison possèdent des taux d'évolution faibles [11]. Les résultats de Cui *et al.* et de Xia et Levitt se fondent sur un modèle monomérique. Puisque les séquences des dimères s'organisent en *superfunnel* fonctionnel, nous prédisons que la fonctionnalité, l'aptitude à dimériser, se trouverait elle aussi augmentée par l'intégration de la recombinaison. Des études préliminaires menées sur le réseau $\mathbf{s}_{10}$ (470 334 génotypes) obtenu avec le complexe tridimensionnel appuient cette conjecture. Si nous mesurons la fonctionnalité par le Z-score moyen, nous rappelons que $\langle \mathrm{Z} \rangle_r = -2{,}51$, $\langle \mathrm{Z} \rangle_s = -2{,}61$. Une simulation menée avec une population de 1 000 individus subissant 12 mutations et 62 recombinaisons par génération (correspondant à des taux de mutation et de recombinaison de 10^{-4} et $5 \cdot 10^{-4}$ par résidu et par génération), selon la méthode décrite dans la référence [203], permet d'attendre des valeurs inférieures à $-2{,}75$ en se limitant à la plus grande composante connexe (cf. figure 1).

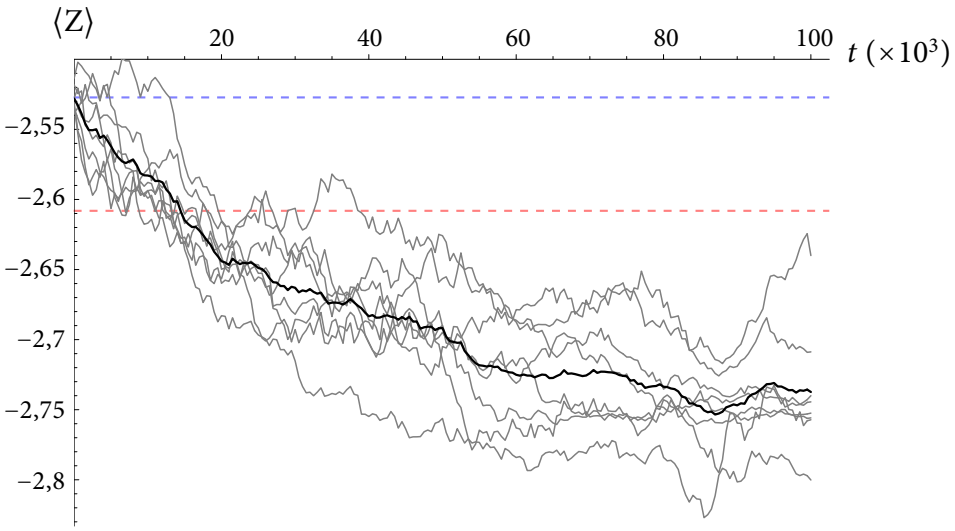


FIG. 1 : Évolution du Z-score moyen dans huit simulations (en gris) et de la moyenne des huit simulations (en noir) en fonction du temps (en nombre de générations). Les valeurs moyennes attendues dans le cas d'une distribution uniforme des allèles et à l'état stationnaire sous l'effet des mutations ponctuelles seules sont indiquées par des lignes pointillées, respectivement bleue et rouge.

Le comportement adaptatif de l'évolution neutre repose sur l'inégalité du nombre de connexions (en d'autres termes, de la fraction de mutations neutres) parmi les séquences. Un réseau neutre « régulier » où tous les nœuds possèdent le même nombre de connexions ne peut pas donner naissance à une adaptation. Ces observations ne rendent compte que du fait que la robustesse mutationnelle est sélectionnée par l'évolution. Le papier le plus démonstratif à ce sujet est celui de WILKE *et al.* où ont été identifiés deux régimes possibles d'évolution : *survival of the fittest* et *survival of the flattest*. Lorsque le taux de mutation est important, les génotypes robustes aux mutations sont favorisés.

Considérons deux individus. Le génotype de l'un d'entre eux est plus robuste aux mutations que celui de l'autre. La sélection de génotypes robustes est possible parce que la progéniture du génotype le moins robuste souffrira plus de mutations délétères. Mais alors, l'évolution est-elle vraiment neutre ? Comme nous l'avons annoncé en introduction, le *fitness* d'un individu est le produit de la probabilité qu'il parvienne à l'âge adulte et de sa fertilité. Cela a été rappelé par KIMURA dans une de ses dernières synthèses de la théorie neutre [95, p. 370],

> I would like to add here that by 'selectively neutral' I mean selectively equivalent: namely, mutant forms can do the job equally well in terms of survival *and reproduction of individuals possessing them*[36].

36. C'est nous qui soulignons.

Cette citation semble faire précisément écho à l'*Origine des espèces* de Darwin qui définit [36] :

> I should premise that I use this term [struggle for existence] in a large and metaphorical sense, including dependence of one being on another, and including (which is more important) *not only the life of the individual, but success in leaving progeny*[37].

La non-neutralité est également présente dans la formule donnant le *fitness* d'une séquence possédant n_i connexions dans son réseau neutre (équation 27, p. 56) :

$$w_i = 1 + \mu\,(n_i - \langle n \rangle)/\ell.$$

L'évolution peut donc être neutre, si le taux de mutation μ est proche de zéro. Mais lorsque le taux de mutation est faible, le comportement adaptatif ne peut justement pas se manifester puisqu'il dépend de la valeur de $M\mu$. Le comportement adaptatif que nous observons est donc bien lié à des différences de *fitness*. La prétendue neutralité n'est que *fonctionnelle*, pas *évolutive*, du moins pas comme l'entend Kimura. Un point de vue légèrement différent est exposé par Nei pour qui la neutralité définie par Kimura ($|N\,s| \ll 1$, avec les notations de l'introduction) est trop stricte. Selon Nei, il est vraisemblable qu'une mutation délétère présentant une infériorité $s = -0{,}001$ dans une population de $N = 10^6$ individus n'ait pas de pertinence biologique [142]. Nos résultats contredisent son analyse.

L'importance de la prise en compte des interactions protéine-protéine a été soulignée plus haut. La prise en considération de deux protéines est une avancée modeste, mais elle n'en est pas moins décisive. Nous citons Ewens, qui note dans son introduction du traitement de deux gènes,

> So far […] we have assumed that the fitness of any individual depends on his genetic constitution at a single locus. This is of course only an initial simplification […]. Although such a 'two-locus' theory may often be little more realistic than 'single-locus' behavior, it does allow at least two advances to be made. First, some assessment can be made of the accuracy of approximating two-locus behavior and measurements by combining two-single locus results. Second, *no assessment of the evolutionary importance of linkage between loci can be made without at least a two-locus analysis*[37] [50, p. 67].

Les études structurales des effets de la recombinaison se sont concentrées sur des recombinaisons intragéniques. Elles n'ont, par construction, pas pu s'intéresser aux événements intergéniques pourtant plus fréquents, chez les Eucaryotes tout du moins. Or l'intérêt de

37. C'est nous qui soulignons.

la recombinaison dans l'évolution, bien qu'il soit admis, est un sujet controversé. On trouvera une excellente revue de ce débat dans l'article de FELSENSTEIN [52]. La recombinaison est souhaitable car elle peut contrecarrer la roue à rochet[38] de Muller [140], augmenter la vitesse d'évolution [151,152] (voir aussi référence [34] pour une approche théorique). Elle peut délier deux allèles dont l'un serait délétère et l'autre avantageux, permettant ainsi aux allèles avantageux de se fixer plus facilement et aux allèles délétères d'être éliminés plus efficacement. MAYNARD SMITH et d'autres ont cependant exprimé quelques doutes [122, p. 242]. Notamment la possibilité que la recombinaison puisse aussi délier des allèles qui sont favorables lorsqu'ils agissent en synergie. La modélisation de la dimérisation fournit un modèle *ab initio* qui permettrait d'inclure les recombinaisons intra- et intergéniques et d'étudier leurs effets.

Les deux modèles que nous avons présentés dans cette thèse tentent de rendre compte de l'évolution « macroscopique » par une description simple de la nature biochimique des protéines et de leur fonction : leur capacité à se replier et à interagir de façon fonctionnelle. Bien qu'ils demandent à être confirmés, et en dépit de leurs différences, ces modèles, donnent une image cohérente de la façon dont les protéines semblent évoluer.

38. L'accumulation de mutations fut conçue par MULLER comme un processus irréversible car les mutations inverses sont très peu probables. L'expression de « roue à rochet » (*Muller's ratchet*) a été introduite pour illustrer ce processus : chaque mutation fait passer un cran de la roue. La recombinaison permet de reconstituer des séquences vierges de mutations à partir de divers fragments présents dans la population.

Remerciements

…à tous les gens de bien qui ne boivent pas d'eau

Offenbach, *Les contes d'Hoffman*

L'immense et compliqué palimpseste de ce mémoire

Adapté de Baudelaire

Mes remerciements vont avant tout aux membres du jury, MM. Richard Lavery, Marc Delarue, Denis Couvet, Nicolas Lartillot et Yves-Henri Sanejouand, pour avoir accepté de juger ce travail, pour m'avoir judicieusement conseillé pour améliorer le présent manuscrit et pour m'avoir apporté d'utiles critiques. J'adresse toute ma gratitude à Thomas Simonson, pour la direction de ma thèse et ses conseils qui ont, entre autres, permis de rendre intelligibles et le manuscrit et la présentation orale de mon travail, Pierre Plateau pour m'avoir accueilli au laboratoire, Sylvain Blanquet et Yves Mechulam pour m'avoir intégré au département de biologie.

Contrairement à ce que l'on pourrait croire, les remerciements ne sont pas nécessairement la partie la plus facile à rédiger. La refonte de cette partie fait de ces remerciements des remerciements « dignes de ceux d'une thèse de médecine » mais si je me demande qui, de près ou de loin, a participé à la réalisation de ce travail, tant de noms me viennent à l'esprit... Ceux qui ont contribué à ma formation ; ceux qui ont prêté main forte à la confection ou à la relecture du manuscrit de thèse ; ceux qui ont été à mes côtés pour me soutenir ; ceux, enfin, avec qui j'ai partagé des moments de joie quotidiennement, hebdomadairement ou annuellement. Finalement, la façon la plus simple de noircir cette ou ces pages eût été de procéder alphabétiquement, ou pire de se contenter d'un « merci à qui de droit ». Mais alors, François aurait vraisemblablement médit « Ils ne sont vraiment pas drôles, tes remerciements ! » Et il aurait eu, comme souvent, raison. Heureusement que je t'ai, François. François : puisque je

t'ai sous le coude, j'en profite pour te remercier abondamment, à la hauteur de la « dette » — si tant est qu'il existe des dettes en amitié — que je te dois.

Je dois aux personnes qui ont hanté, en même temps que moi, les couloirs du laboratoire de biochimie de l'École polytechnique, d'avoir passé trois ans dans une excellente ambiance. Je remercie donc Alexey, Annick, Catherine, Christine, Christine (la bioinformaticienne), David, Emmanuelle, Françoise, Guillaume (le petit), Ioana, Marc, Marcel, Michel, Michou, Pascal et Romary pour leur bonne humeur au quotidien. Je félicite en particulier Pierre et Yves d'avoir réussi à me guider dans mes premières expériences d'enseignement et l'enseignement de mes premières expériences (ou presque). J'espère qu'ils n'auront pas remarqué une baisse de niveau trop inquiétante chez les élèves pendant ma période d'activité. J'ai pris un vif plaisir à enseigner, en grande partie, parce que les élèves ont été adorables. La biologie moléculaire me fait un peu moins peur maintenant. Merci aussi à Catherine T. pour ses consultations à l'œil et à Laura pour qui j'ai un petit faible.

Les jeunes ne sont pas en reste : Anne, Caroline, Laure, Laurent, Lolo, Marie, Sandra et Thanh (exilée dans la jungle parisienne). Que seraient mes journées devenues sans nos débats tumultueux sur « l'orvet est-il un lézard sans pattes ou un serpent manchot ? » ou « le chichi est-il un donut qui a perdu son trou ? », sans nos petits et grands (dé)tours aux abords du lac. J'hésite à mentionner Agathe dont j'attends toujours la danse des chouquettes. Lolo, mille mercis pour avoir toujours volé à mon secours quand les pipettes et les eppendorfs se mutinaient en salle de TP. Dans le filon de la bioinformatique, je fais un petit traitement de faveur pour Guillaume, Damien et Éric qui ont une bonne part de responsabilité dans le plaisir que j'ai eu à faire ma thèse (et à goûter des parties baby-foot ou de flipper). J'ai passé également d'excellents moments avec quelques physiciens voisins (Driss et Yves).

Je sais gré à Alfonso et Mariel qui ont été, pendant des mois difficiles, une béquille infaillible : mi agradecimiento supera todas las palabras que pudieran caber en estas páginas, muchísimas gracias por todo. *Je salue également la horde d'Espagnols et d'hispanophones qui m'ont beaucoup apporté (notamment, un peu de fluidité dans la langue de Cervantès) : Anne-Marie, Esther* (en Madrid), *Guillermo, Javier, Selene.*

Une pensée pour mes ex-Pasteuriens et ex-Pasteuriens Canada Dry : Agathe et Cissé, Anne-Gaëlle et Pierre-Damien, Élodie et Pedro, Martial ; les aînés aussi, Christophe et Nathalie. Cette période reste pour moi comme un âge d'or. Ne serait-ce que parce que la cantine de Pasteur est quand même nettement supérieure à celle de l'X ! L'épigraphe de ce chapitre leur est destiné : ce sont des gens de bien, des vrais ! Ma gratitude va à Arnaud Blondel qui m'a beaucoup appris et Michel Goldberg pour son accueil cordial au sein de l'équipe de biologie cellulaire. Ludovic a été une autre rencontre importante pour moi, ces années-là, celles du DEA. Je tiens à dédier cet ouvrage au Pr S. Hazout et plus généralement aux professeurs qui m'ont enseigné le goût des choses.

L'École normale a été l'occasion de rencontrer des amis, dont je ne me lasse pas. François, Clotilde, Julien (le littéraire), Julien (le mathématicien) et Alex'-la-grimpeuse, Julien (le physicien) et Isabelle, Olivier (le {biolog|économ}iste) et Élodie, Olivier (le chimiste) et Émilie.

L'ENS facilite les rapports entre disciplines. Elle m'a donné l'opportunité inappréciable de côtoyer des personnes sans qui ce travail ne serait pas ce qu'il est. Je dois énormément à

Éric Brunet que mes équations d'évolution ont intrigué et qui a pris le temps de m'ouvrir les yeux sur un bon nombre de subtilités qui m'avaient échappé dans l'évolution neutre des populations finies. Je suis reconnaissant à Bernard Derrida de m'avoir éclairé sur la vitesse d'évolution des populations monomorphes et polymorphes. Je remercie la communauté des news *de l'ENS dont j'ai bénéficié (l'astuce de l'équation* 41 *leur est due). Par ailleurs, Richard Goldstein s'est toujours montré très disponible pour me communiquer des articles auxquels il a participé.*

Il y a encore un Julien (le « géo-quelque-chosien ») qui a toujours prodigué conseils et réconfort avec beaucoup de sagesse. Merci à Jérémy pour les nombreuses réponses qu'il a apportées à mes questions. J'adresse aussi ma sympathie aux camarades de fctt*, malheureusement trop nombreux pour en donner la liste ici. Une classe plus hétéroclite est formée par les personnes connues, ou* re*-connues par hasard — d'une certaine façon, par* random drift *—, Emmanuelle, par exemple, Jean-Laurent, un voisin bien sympathique ou Sophie, ma colocataire, dont la bonne humeur quotidienne a été salvatrice. Je suis heureux d'avoir retrouvé la Montpellieraine Catherine et ses parents, de m'être cassé le nez sur Roland dans un train fusant vers le Grand Est. Mes médecines préférées, aux effets secondaires délicieux — l'effet* tanguero*, par exemple —, sont sans aucun doute possible Anne, Florence et Oana.*

Mon père s'est montré un relecteur hors pair[39]*, ma mère une conseillère hors du commun, mes frère et sœur ont toujours été un appui sans égal. Je les remercie chaleureusement, eux et le reste de ma famille, pour la tendresse dont ils m'ont entouré et la confiance qu'ils m'ont toujours accordée. Corentin et Marine, ces adorables diablotins, m'apportent beaucoup de joie. Je souhaite aussi mentionner la famille d'Artois et Vigneau pour ses attentions et sa bonté pendant cinq années qui comprennent en grande partie celles du doctorat, avec une pensée particulière pour Manou et une certaine Florence d'Artois.*

Je puis maintenant dire aux rapides années :
Passez ! passez toujours ! je n'ai plus à vieillir !
Allez-vous-en avec vos fleurs toutes fanées ;
J'ai dans l'âme une fleur que nul ne peut cueillir !

Votre aile en le heurtant ne fera rien répandre
Du vase où je m'abreuve et que j'ai bien rempli.
Mon âme a plus de feu que vous n'avez de cendre !
Mon cœur a plus d'amour que vous n'avez d'oubli !

39. Les erreurs qui subsisteraient — et il en subsiste, *naturellement* — seraient à lui imputer, et je ne pourrais nullement en être tenu responsable.

Bibliographie

Zambia, for instance, has refused to import transgenic corn. But [transgenic] cotton has faced no such trade barriers. The obvious reason is that people tend not to eat their shirts. Another explanation for the difference between our attitudes toward corn and cotton is that humans prefer their food, particularly staples, to stay relatively the same, while they actively seek out mutations in their clothing.

James Gorman, *New York Times*, Saturday, May 27, 2006

[1] Babajide, A., I. L. Hofacker, M. J. Sippl et P. F. Stadler. Neutral networks in protein space:. *Folding and Design*, 2(5), p. 261–269, 1997.

[2] Bajaj, M. et T. Blundell. Evolution and the tertiary structure of proteins. *Annual Reviews of Biophysics and Bioengineering*, 13, p. 453–492, 1984.

[3] Barabasi, A. L. et R. Albert. Emergence of scaling in random networks. *Science*, 286(5439), p. 509–512, 1999.

[4] Bartlett, G. J., N. Borkakti et J. M. Thornton. Catalysing new reactions during evolution: Economy of residues and mechanism. *J. Mol. Biol.*, 331, p. 829–860, 2003.

[5] Bastolla, U., J. Farwer, E. W. Knapp et M. Vendruscolo. How to guarantee optimal stability for most representative structures in the Protein Data Bank. *Proteins*, 44, p. 79–96, 2001.

[6] Bastolla, U., M. Porto, H. E. Roman et M. Vendruscolo. Lack of self-averaging in neutral evolution of proteins. *Phys. Rev. Lett.*, 89(20), 2002.

[7] ————. Connectivity of neutral networks, overdispersion, and structural conservation in protein evolution. *Journal of Molecular Evolution*, 56(3), p. 243–

254, 2003.

[8] BENNETT, M. J., M. P. SCHLUNEGGER et D. EISENBERG. 3D domain swapping: A mechanism for oligomer assembly. *Protein Science*, 4(12), p. 2455–2468, 1995.

[9] BERGER, B. et T. LEIGHTON. Protein folding in the hydrophobic-hydrophilic (HP) model is NP-complete. *Journal of Computational Biology*, 5(1), p. 27–40, 1998.

[10] BERGSON, H. *L'Évolution créatrice*. Quadrige (PUF), Paris, 8 edition, 1998. Première parution en 1907.

[11] BIERNE, N. et A. EYRE-WALKER. The genomic rate of adaptive amino acid substitution in *Dropophila*. *Molecular Biology and Evolution*, 21, p. 1350–1360, 2004.

[12] BLACKBURNE, B. P. et J. D. HIRST. Evolution of functional model proteins. *J. of Chem. Phys.*, 115(4), July 2001.

[13] ————. Three-dimensional functional model proteins: Structure, function and evolution. *J. of Chem. Phys.*, 119, p. 3453–5460, 2003.

[14] ————. Population dynamics simulations of functional model proteins. *J. of Chem. Phys.*, 123, 2005.

[15] BLOOM, J. D. et C. ADAMI. Apparent dependence of protein evolutionary rate on number of interactions is linked to biases in protein-protein interactions data sets. *BMC Evol. Biol.*, 3, p. 21, 2003.

[16] BLOOM, J. D., D. A. DRUMMOND, F. H. ARNOLD et C. O. WILKE. Structural determinants of the rate of protein evolution in yeast. *Mol. Biol. Evol.*, 23(9), p. 1751–1761, 2006.

[17] BLOOM, J. D., S. T. LABTHAVIKUL, C. R. OTEY et F. H. ARNOLD. Protein stability promotes evolvability. *Proc. Natl. Acad. Sci.*, 103(15), p. 5869–5874, 2006.

[18] BLOOM, J. D., C. O. WILKE, F. H. ARNOLD et C. ADAMI. Stability and the evolvability of function in a model protein. *Biophysical Journal*, 86, p. 2758–2764, 2004.

[19] BOGAN, A. A. et K. S. THORN. Anatomy of hot spots in protein interfaces. *J. Mol. Biol.*, 280, p. 1–9, 1998.

[20] BORNBERG-BAUER, E. How are model protein structures distributed in sequence space? *Biophysical Journal*, 73, p. 2393–2403, 1997.

[21] BORNBERG-BAUER, E. et H. S. CHAN. Modeling evolutionary landscapes: Mutational stability, topology, and superfunnels in sequence space. *Proc. Natl. Acad. Sci.*, 96, p. 10689–10694, 1999.

[22] BOWIE, J. U., R. LUTHY et D. EISENBERG. A method to identify protein sequences that fold into a known three-dimensional structure. *Science*, 253, p. 164–170, 1991.

[23] Brenner, S. E., C. Chothia et T. J. Hubbard. Population statistics of protein structures: Lessons from structural classifications. *Curr. Opin. Struct. Biol.*, 7(3), p. 369–376, 1997.

[24] Brooks, B. R., R. E. Bruccoleri, B. D. Olafson, D. J. States, S. Swaminathan et M. Karplus. A program for macromolecular energy, minimization, and dynamics calculations. *J. Computational Chemistry*, 4, p. 187–217, 1983.

[25] Buchler, N. E. G. et R. A. Goldstein. Effect of alphabet size and foldability requirements on protein structure designability. *Proteins*, 34, p. 113–124, 1999.

[26] Caffrey, D. R., S. Somaroo, J. D. Hugues, J. Mintseris et E. S. Huang. Are protein-protein interfaces more conserved in sequence than the rest of the protein surface? *Protein Science*, 13, p. 190–202, 2004.

[27] Chamary, J. V., J. L. Parmley et L. D. Hurst. Hearing silence: non-neutral evolution at synonymous sites in mammals. *Nature Reviews Genetics*, 7(2), p. 98–108, 2006.

[28] Chan, H. S. et K. A. Dill. Origins of structure in globular proteins. *Proc. Natl. Acad. Sci.*, 87(16), p. 6388–6392, 1990.

[29] ————. Sequence space soup of proteins and copolymers. *Journal of Chemical Physics*, 95, p. 2393–2403, 1991.

[30] ————. Comparing folding codes for proteins and polymers. *Proteins*, 100, p. 9238–9257, 1996.

[31] Chan, H. S. Folding alphabets. *Nature Structural Biology*, 6(11), p. 994–996, 1999.

[32] Chan, H. S. et E. Bornberg-Bauer. Perspectives on protein evolution from simple exact models. *Applied Bioinformatics*, 1(3), p. 121–144, 2002.

[33] Chothia, C. Proteins. One thousand families for the molecular biologist. *Nature*, 357, p. 543–544, 1992.

[34] Cohen, E., D. A. Kessler et H. Levine. Recombination dramatically speeds up evolution of finite populations. *Phys. Rev. Lett.*, 94(9), p. 098102, 2005.

[35] Cui, Y., W. H. Wong, E. Bornberg-Bauer et H. S. Chan. Recombinatoric exploration of novel folded structures: A heteropolymer-based model of protein evolutionary landscapes. *Proc. Natl. Acad. Sci.*, 99(809-814), 2002.

[36] Darwin, C. *On The Origin of Species by Means of Natural Selection, or The Preservation of Favoured Races in the Struggle for Life*. 1859.

[37] Delano, W. L. The PyMOL molecular graphics system. `http://www.pymol.org/`, 2002.

[38] DILL, K. A. Theory for the folding and stability of globular proteins. *Biochemistry.*, 24(6), p. 1501–1509, 1985.

[39] DIMA, R. I. et D. THIRUMALAI. Exploring protein aggregation and self-propagation using lattice models: Phase diagram and kinetics. *Protein Science*, 11(1036-1049), 2002.

[40] DOKHOLYAN, N. V. et E. I. SHAKHNOVICH. Understanding hierarchical protein evolution from first principles. *J. Mol. Biol.*, 312(1), p. 289–307, 2001.

[41] DOOLITTLE, R. F. et B. BLOMBAECK. Amino-acid sequence investigations of fibrinopeptides from various mammals: Evolutionary implications. *Nature*, 202, p. 147–152, 1963.

[42] DRUMMOND, D. A., J. D. BLOOM, C. ADAMI, C. O. WILKE et F. H. ARNOLD. Why highly expressed proteins evolve slowly. *Proc. Natl. Acad. Sci.*, 102(40), p. 14338–14343, 2005.

[43] DUNBAR, A. Y., Y. KAMADA, G. J. JENKINS, E. R. LOWE, S. S. BILLECKE et Y. OSAWA. Ubiquitination and degradation of neutral nitric-oxide synthase *in vitro*: dimer stabilization protects the enzyme from proteolysis. *Mol. Pharmacol.*, 66, p. 964–969, 2004.

[44] EIGEN, M. Selforganization of matter and the evolution of biological macromolecules. *Naturwissenschaften*, 58, p. 465–523, 1971.

[45] ELCOCK, A. H. et J. A. MCCAMMON. Identification of protein oligomerization states by analysis of interface conservation. *Proc. Natl. Acad. Sci.*, 98, p. 2990–2994, 2001.

[46] ELDREDGE, N. et S. J. GOULD. Punctuated equilibria: An alternative to phyletic gradualism. In SCHOPF, T. J. M., editor, *Models in Paleobiology*, pages 82–115, San Francisco, California, 1972. Freeman, Cooper and Co.

[47] EMBERLY, E. G., N. S. WINGREEN et C. TANG. Designability of alpha-helical proteins. *Proc. Natl. Acad. Sci.*, 99, p. 11163–11168, 2002.

[48] ENGLAND, J. L., B. E. SHAKHNOVICH et E. I. SHAKHNOVICH. Natural selection of more designable folds: a mechanism for thermophilic adaptation. *Proc. Natl. Acad. Sci.*, 100(15), p. 8727–8731, 2003.

[49] ENGLAND, J. L. et E. I. SHAKHNOVICH. Structural determinant of protein designability. *Phys. Rev. Lett.*, 90(21), p. 218101, 2003.

[50] EWENS, W. J. *Mathematical population genetics.* Springer-Verlag, 1979.

[51] FAN, K. et W. WANG. What is the minimum number of letters required to fold a protein? *J. Mol. Biol.*, 328, p. 921–926, 2003.

[52] Felsenstein, J. The evolutionary advantage of recombination. *Genetics*, 78, p. 737–756, 1974.

[53] ———— *Theoretical evolutionary genetics*. 1978 [2005]. `http://evolution.genetics.washington.edu/pgbook/pgbook.html`.

[54] Feynman, R. P. *Lectures on Physics*. Pearson/Addison-Wesley, 1963 [1977, 2006].

[55] Finkelstein, A. V., A. M. Gutin et A. Y. Badretdinov. Why are the same protein folds used to perform different functions? *FEBS Letters*, 325, p. 23–8, 1993.

[56] Fisher, R. A. On the dominance ratio. *Proc. R. Soc. Edinburgh*, 42, p. 321–341, 1922.

[57] ———— *The genetical theory of Natural Selection*. Dover Publications, 1930.

[58] Fontana, W. Modelling « evo-devo » with RNA. *BioEssays*, 24, p. 1164–1177, 2002.

[59] Fontana, W. et P. Schuster. Continuity in evolution: On the nature of transitions. *Science*, 280(5368), p. 1451–1455, 1998.

[60] Franz, S. et L. Peliti. Error threshold in simple landscapes. *J. Phys. A*, 30, p. 4481–4487, 1997.

[61] Fraser, H. B., A. E. Hirsh, L. M. Steinmetz, C. Scharfe et M. W. Feldman. Evolutionary rate in the protein interaction network. *Science*, 296, p. 750–752, 2002.

[62] Gavin, A. C. et G. Superti-Furga. Protein complexes and proteome organization from yeast to man. *Curr. Opin. Chem. Biol.*, 7(1), p. 21–27, 2003.

[63] Gillespie, J. H. The molecular clock may be an episodic clock. *Proc. Natl. Acad. Sci.*, 81, p. 8009–8013, 1984.

[64] ———— Natural selection and the molecular clock. *Molecular Biology and Evolution*, 3(2), p. 138–155, 1986.

[65] Go, N. et H. Taketomi. Respective roles of short- and long-range interactions in protein folding. *Proc. Natl. Acad. Sci.*, 75, p. 559–563, 1978.

[66] Goldstein, R. A., Z. A. Luthey-Schulten et P. G. Wolynes. Optimal protein-folding codes from spin-glass theory. *Proc. Natl. Acad. Sci.*, 89, p. 4918–4922, 1992.

[67] Gould, S. J. et N. Eldredge. Punctuated equilibria: The tempo and mode of evolution reconsidered. *Paleobiology*, 3, p. 115–151, 1977.

[68] GOVINDARAJAN, S. et R. A. GOLDSTEIN. The foldability landscape of model proteins. *Biopolymers*, 42, p. 427–438, 1997.

[69] GOVINDARAJAN, S., R. RECABARREN et R. A. GOLDSTEIN. Estimating the total number of protein folds. *Proteins: Structure, Function and Genetics*, 35(4), p. 408–414, 1999.

[70] GREEN, S. M., A. K. MEEKER et D. SHORTLE. Contributions of the polar, uncharged amino acids to the stability of staphylococcal nuclease: evidence for mutational effects on the free energy of the denatured state. *Biochemistry*, 31(25), p. 5717–5728, 1992.

[71] GUHAROY, M. et P. CHAKRABARTI. Conservation and relative importance of residues across protein-protein interfaces. *Proc. Natl. Acad. Sci.*, 102(43), p. 15447–15452, 2005.

[72] GUTTERIDGE, A. et J. M. THORNTON. Understanding nature's catalytic toolkit. *Trends in Biochemical Sciences*, 30(11), p. 622–629, 2005.

[73] H J WHITFIELD, J., R. G. MARTIN et B. AMES. Classification of aminotransferase (C gene) mutants in the histidine operon. *J. Mol. Biol.*, 21, p. 335–355, 1966.

[74] HAHN, M. W., G. C. CONANT et A. WAGNER. Molecular evolution in large genetic networks: Does connectivity equal constraint? *J. Mol. Evol.*, 2004.

[75] HALDANE, J. B. S. The cost of natural selection. *Journal of Genetics*, 55, p. 511–524, 1957.

[76] HALL, B. G. The EVOLVEAGENE software.

[77] ———— Simple and accurate estimation of ancestral protein sequence. *Proc. Natl. Acad. Sci.*, 103(14), p. 5431–5436, April 2006.

[78] HARRISON, P. M., H. S. CHAN, S. B. PRUSINER et F. E. COHEN. Conformational propagation with prion-like characteristics in a simple model of protein folding. *Protein Science*, 10, p. 819–835, 2001.

[79] HATTORI, T., N. OHOKA, Y. INOUE, H. HAYASHI et K. ONOZAKI. C/EBP family transcription factors are degraded by the proteasome but stabilised by forming dimer. *Oncogene*, 22, p. 1273–1280, 2003.

[80] HELLINGA, H. W. Rational protein design: Combining theory and experiment. *Proc. Natl. Acad. Sci.*, 94, p. 10015–10017, 1997.

[81] HEROLD, M. et K. KIRSCHNER. Reversible dissociation and unfolding of aspartate aminotransferase from *Escherichia coli*: characterization of a monomeric intermediate. *Biochemistry*, 29(7), p. 1907–1913, 1990.

[82] HIRST, J. D. The evolutionary landscape of functional model proteins. *Protein Engineering*, 12(9), p. 721–726, 1999.

[83] Huelsenbeck, J. P. et K. A. Dyer. Bayesian estimation of positively selected sites. *Journal of Molecular Evolution*, 58(6), p. 661–672, 2004.

[84] Huynen, M. A. Exploring phenotype space through neutral evolution. *Journal of Molecular Evolution*, 43(3), p. 165–169, 1996.

[85] Jaramillo, A., L. Wernisch, S. Héry et S. Wodak. Folding free energy function selects native-like protein sequences in the core but not on the surface. *Proc. Natl. Acad. Sci.*, 99(21), p. 13554–13559, 2002.

[86] Ji, Y.-Y., Y.-Q. Li, J.-W. Mao et X.-W. Tang. Model study of prionlike folding behavior in aggregated proteins. *Physical Review E*, 72, 2005.

[87] Jones, D. T. Theoretical approaches to designing novel sequences to fit a given fold. *Current Opinion in Biotechnology*, 6, p. 452–459, 1995.

[88] Jones, S. et J. M. Thornton. Principles of protein-protein interactions. *Proc. Natl. Acad. Sci.*, 93, p. 13–20, 1996.

[89] Jukes, T. H. Dr. Best, insulin, and molecular evolution. *Canadian Journal of Biochemistry*, 57, p. 455–458, 1979.

[90] Kajander, T., P. C. Kahn, S. H. Passila, D. C. Cohen, L. Lehtiö, W. Adolfsen, J. Warwicker, U. Schell et A. Goldman. Buried charged surface in proteins. *Structure*, 8, p. 1203–1214, nov 2000.

[91] Kauffman, S. A. et S. Levin. Towards a general theory of adaptive walks on rugged landscapes. *Journal of Theoretical Biology*, 128, p. 11–45, 1987.

[92] Kimura, M. Stochastics processes and distribution of gene frequencies under natural selection. *Cold Spring Harbor Symposia on Quantitative Biology*, 20, p. 33–53, 1955.

[93] ———— Optimum mutation rate and degree of dominance as determined by the principle of minimum genetic load. *Journal of Genetics*, 57, p. 21–34, 1960.

[94] ———— *The neutral theory of molecular evolution*. Cambridge University Press, 1983. Réimpression en 1986.

[95] ———— The neutral theory of molecular evolution: A review of recent evidence. *Jpn Journal of Genetics*, 66(367-386), 1991.

[96] ———— Recent development of the neutral theory from the Wrightian tradition of theoretical population genetics. *Proc. Natl. Acad. Sci.*, 88, p. 5969–5973, 1991.

[97] ———— *Population genetics, molecular evolution, and the neutral theory—Selected papers*. The University of Chicago Press, 1994. Edited and with introductory Essays by Takahata, with a foreword of Crow.

[98] KIMURA, M. et J. F. CROW. The number of alleles that can be maintained in a finite population. *Genetics*, 69, p. 725–738, 1964.

[99] KIMURA, M. et T. OHTA. The average number of generations until fixation of a mutant gene in a finite population. *Genetics*, 61, p. 763–71, 1969.

[100] KING, J. L. et T. H. JUKES. Non-darwinian evolution. *Science*, 1969.

[101] KLEINMAN, C. L., N. RODRIGUE, C. BONNARD, H. PHILIPPE et N. LARTILLOT. A maximum likelihood framework for protein design. *BMC Bioinformatics*, 7(326), 2006.

[102] KOEHL, P. et M. LEVITT. Protein topology and stability define the space of allowed sequences. *Proc. Natl. Acad. Sci.*, 99, p. 1280–1285, 2002.

[103] KRAUSE, A. et M. VINGRON. A new paradigm in sequence database searching and clustering. *Bioinformatics*, 14, p. 430–438, 1998.

[104] KUMAR, S. et S. SUBRAMANIAN. Mutation rates in mammalian genomes. *Proc. Natl. Acad. Sci.*, 99(2), p. 803–808, 2002.

[105] LANGLER, E. S. *et al.* Initial sequencing and analysis of the human genome. *Nature*, 409(6822), p. 860–921, 2001.

[106] LAUNAY, G., R. MENDEZ, S. WODAK et T. SIMONSON. Recognizing protein-protein interfaces with empirical portentials and reduced amino-acid alphabets. *Submitted to J. Mol. Biol.*, 2006.

[107] LEVY, Y., A. CAFLISCH, J. N. ONUCHIC et P. G. WOLYNES. The folding and dimerization of hiv-1 protease: evidence for a stable monomer from simulations. *J. Mol. Biol.*, 340(1), p. 67–79, 2004.

[108] LEVY, Y., S. S. CHO, J. N. ONUCHIC et P. G. WOLYNES. A survey of flexible protein binding mechanisms and their transition states using native topology based energy landscapes. *J. Mol. Biol.*, 346(4), p. 1121–1145, 2005.

[109] LEVY, Y., P. G. WOLYNES et J. N. ONUCHIC. Protein topology determines binding mechanism. *Proc. Natl. Acad. Sci.*, 101(2), p. 511–516, 2004.

[110] LI *et al.* A map of the interactome network of the metazoan *C. elegans*. *Science*, 303(5657), p. 540–543, 2004.

[111] LI, H., C. TANG et N. S. WINGREEN. Designability of protein structures: a lattice-model study using the miyazawa-jernigan matrix. *Proteins*, 49, p. 403–412, 2002.

[112] LI, H., R. HELLING, C. TANG et N. S. WINGREEN. Why do proteins look like proteins? *Science*, 273, p. 666–669, 1996.

[113] LI, H., C. TANG et N. S. WINGREEN. Nature of driving force for protein folding: A result from analyzing the statistical potential. *Phys. Rev. Lett.*, 79(4), p. 765–768, 1997.

[114] Li, H., C. Tang et N. S. Wingreen. Are protein folds atypical? *Proc. Natl. Acad. Sci.*, 95, p. 4987–4990, 1998.

[115] Li, T., K. Fan, J. Wang et W. Wang. Reduction of protein sequence complexity by residue grouping. *Protein Engineering*, 16(5), p. 323–330, 2003.

[116] Li, W.-H., T. Gojobori et M. Nei. Pseudogenes as a paradigm of neutral evolution. *Nature*, 292, p. 237–239, 1981.

[117] Li, Z. R., X. Han et G. R. Liu. Protein designability analysis in sequence principal component space using 2D lattice model. *Computer Methods and Programs in Biomedicine*, 76, p. 21–29, 2004.

[118] Ma, B., T. Elkayam, H. Wolfson et R. Nussinov. Protein-protein interactions: structurally conserved residues distinguish between binding sites and exposed protein surfaces. *Proc. Natl. Acad. Sci.*, 100(10), p. 5772–5777, 2003.

[119] Margoliash, E. Primary structure and evolution of cytochrome *c*. *Proc. Natl. Acad. Sci.*, 50, p. 672–679, 1963.

[120] Mateu, M. G., M. M. Sanchez Del Pino et A. R. Fersht. Mechanism of folding and assembly of a small tetrameric protein domain from tumor suppressor p53. *Nature Structural Biology*, 6(2), p. 191–198, 1999.

[121] Maynard Smith, J. Natural selection and the concept of protein space. *Nature*, 225, p. 563–564, 1970.

[122] ———— *Evolutionary genetics*. Oxford University Press, 1997. First published in 1988.

[123] Meeker, A. K., B. Garcia-Moreno et D. Shortle. Contributions of the ionizable amino acids to the stability of staphylococcal nuclease. *Biochemistry*, 35(20), p. 6443–6449, 1996.

[124] Melin, R., H. Li, N. S. Wingreen et C. Tang. Designability, thermodynamic stability, and dynamics in protein folding: a lattice model study. *Journal of Chemical Physics*, 110, p. 1252, 1999.

[125] Mendez, R., R. Leplae, M. F. Lensink et S. J. Wodak. Assessment of CAPRI predictions in rounds 3-5 shows progress in docking procedures. *Proteins*, 60(2), p. 150–169, 2005.

[126] Meyerguz, L., D. Kempe, J. Kleinberg et R. Elber. The evolutionary capacity of protein structures. In *RECOMB '04: Proceedings of the eighth annual international conference on Research in computational molecular biology*, pages 290–297, New York, NY, USA, 2004. ACM Press.

[127] Milla, M. E. et R. T. Sauer. P22 Arc repressor: Folding kinetics of a single-domain, dimeric protein. *Biochemistry*, 33(5), p. 1125–1133, 1994.

[128] MILLER, B. G. et R. WOLFENDEN. Catalytic proficiency: The unusual case of OMP decarboxylase. *Annual Reviews of Biochemistry*, 71, p. 847–885, 2002.

[129] MILLER, D. W. et K. A. DILL. Ligand binding to proteins: The binding landscape model. *Protein Science*, 6, p. 2166–2179, 1997.

[130] MILLER, J., C. ZENG, N. S. WINGREEN et C. TANG. Emergence of highly designable protein-backbone conformations in an off-lattice model. *Proteins: Structure, Function and Genetics*, 47, p. 506–512, 2002.

[131] MINTSERIS, J. et Z. WENG. Structure, function, and evolution of transient and obligate protein-protein interactions. *Proc. Natl. Acad. Sci.*, 102(31), p. 10930–10935, 2005.

[132] MIRNY, L. et E. I. SHAKHNOVICH. Protein folding: From lattice to all-atom models. *Annual Review of Biophysics and Biomolecular Structure*, 30, p. 361–396, 2001.

[133] MIYATA, T., S. MIYAZAWA et T. YASUNAGA. Two types of amino acid substitutions in protein evolution. *Journal of Molecular Evolution*, 12(3), p. 219–236, 1979.

[134] MIYATA, T. et T. YASUNAGA. Rapidly evolving mouse α-globin-related pseudogene and its evolutionary history. *Proc. Natl. Acad. Sci.*, 78, p. 450–453, 1981.

[135] MIYAZAWA, S. et R. L. JERNIGAN. Estimation of effective interresidue contact energies from protein crystal structures: Quasi-chemical approximation. *Macromolecules*, 18, p. 534–552, 1985.

[136] ————————. Residue-residue potentials with a favorable contact pair term and an unfavorable high density term, for simulation and threading. *J. Mol. Biol.*, 256(3), p. 623–644, 1996.

[137] MONOD, J. *Le Hasard et la nécessité. Essai sur la philosophie naturelle de la biologie moderne*. Seuil, 1970.

[138] MORGAN, T. H. *The Scientific Basis of Evolution*. W. W. Norton, New York, 1932.

[139] MUKAI, T. The genetic structure of natural populations of drosophila melanogaster. I. Spontaneous mutation rate of polygenes controlling viability. *Genetics*, 50, p. 1–19, 1964.

[140] MULLER, H. J. The relation of recombination to mutational advance. *Mutat. Res.*, 106, p. 2–9, 1964.

[141] MURPHY, L. R., A. WALLQVIST et R. M. LEVY. Simplified amino acid alphabets for protein fold recognition and implications for folding. *Protein Engineering*, 13(3), p. 149–152, 2000.

[142] NEI, M. Selectionism and neutralism in molecular evolution. *Molecular Biology and Evolution*, 22(12), p. 2318–2342, 2005.

[143] Neidigh, J. W., R. M. Fesinmeyer et N. H. Andersen. Designing a 20-residue protein. *Nature Structural Biology*, 9, p. 425–430, 2002.

[144] Neuwald, A. F., J. S. Liu, D. J. Lipman et C. E. Lawrence. Extracting protein alignment models from the sequence database. *Nucleic Acids Research*, 25, p. 1665–1677, 1997.

[145] Nishida, M., K. Nagata, Y. Hachimori, M. Horiuchi, K. Ogura, V. Mandiyan, J. Schlessinger et F. Inagaki. Novel recognition mode between vav and grb2 sh3 domains. *EMBO Journal*, 20, p. 2995–3007, 2001.

[146] Nussinov, R. et A. B. Jacobson. Fast algorithm for predicting the secondary structure of single-stranded RNA. *Proc. Natl. Acad. Sci.*, 77(11), p. 6309–6313, 1980.

[147] Ohta, T. Slightly deleterious mutant substitution in evolution. *Nature*, 246, p. 96–98, 1973.

[148] ———— Mutation pressure as the main cause of molecular evolution and polymorphism. *Nature*, 252, p. 351–354, 1974.

[149] Ohta, T. et J. H. Gillespie. Development of neutral and nearly neutral theories. *Theoretical Population Biology*, 49(2), p. 128–142, 1996.

[150] Orengo, C. A., D. T. Jones et J. M. Thornton. Protein superfamilies and domain superfolds. *Nature*, 372, p. 631–634, 1994.

[151] Pál, C., B. Papp et L. D. Hurst. Does the recombination rate affect the efficiency of purifying selection? the Yeast genome provides a partial answer. *Mol. Biol. Evol.*, 18(12), p. 2323–2326, 2001.

[152] Pál, C., B. Papp et M. J. Lercher. An integrated view of protein evolution. *Nature Reviews Genetics*, 7, p. 337–348, May 2006.

[153] Perutz, M. F. et H. Lehmann. Molecular pathology of human haemoglobin. *Nature*, 219(157), p. 902–909, 1968.

[154] Popper, K. *La connaissance objective*. Bibliothèque philosophique. Aubier, 1991. Traduction de *Objective knowledge*, 1972.

[155] Rajamani, D., S. Thiel, S. Vajda et C. Camacho. Anchor residues in protein-protein interactions. *Proc. Natl. Acad. Sci.*, 101(31), p. 11287–11292, 2004.

[156] Riddle, D. S., J. V. Santiago, S. T. Bray-Hall, N. Doshi, V. P. Grantcharova, Q. Yi et D. Baker. Functional rapidly folding proteins from simplified amino acid sequences. *Nature Structural Biology*, 4(10), p. 805–809, 1997.

[157] Rodrigue, N., N. Lartillot, D. Bryant et H. Philippe. Site interdependence attributed to tertiary structure in amino acid sequence evolution. *Gene*, 347, p. 207–217, 2004.

[158] ROSE, G. D. et T. P. CREAMER. Protein folding: Predicting predicting. *Proteins*, 19, p. 1–3, 1994.

[159] ROSE, G. D. et R. WOLFENDEN. Hydrogen bonding, hydrophobicity, packing, and protein folding. *Annual Review of Biophysics and Biomolecular Structure*, 22, p. 381–415, 1993.

[160] SALISBURY, F. B. Natural selection and the complexity of the gene. *Nature*, 224, p. 342–343, 1969.

[161] SANEJOUAND, Y.-H. Protein functional dynamics: computational approaches. In *Energy localisation and transfer in crystals, biomolecules and josephson arrays*, volume 22 of *Advanced Series in Nonlinear Dynamics*, pages 273–300, 2004. Thierry Dauxois, Anna Litvak-Hinenzon, Robert MacKay, and Anna Spanoudaki (Editors).

[162] SCHUSTER, P., W. FONTANA, P. F. STADLER et I. L. HOFACKER. From sequences to shapes and back—A case-study in RNA secondary structures. *Proc. of the Royal Society of London B*, 255, p. 279–284, 1994.

[163] SELLA, G. et A. E. HIRSH. The application of statistical physics to evolutionary biology. *Proc. Natl. Acad. Sci.*, 102(27), p. 9541–9546, 2005.

[164] SHAKHNOVICH, E. I. Theoretical studies of protein-folding thermodynamics and kinetics. *Curr. Opin. Struct. Biol.*, 7, p. 29–40, 1997.

[165] ———— Folding by association. *Nature Structural Biology*, 6(2), p. 99–102, 1999.

[166] SHORTLE, D., W. E. STITES et A. K. MEEKER. Contributions of the large hydrophobic amino acids to the stability of staphylococcal nuclease. *Biochemistry*, 29(35), p. 8033–8041, 1990.

[167] SIMPSON, G. G. Organisms and molecules in evolution. *Science*, 146, p. 1535–1538, 1964.

[168] SUEOKA, N. On the genetic basis of variation and heterogeneity of DNA base composition. *Proc. Natl. Acad. Sci.*, 48, p. 582–592, 1962.

[169] TAKAHATA, N. On the overdispersed molecular clock. *Genetics*, 116, p. 168–179, 1987.

[170] TAKETOMI, H., Y. UEDA et N. GO. Studies on protein folding, unfolding and fluctuations by computer simulation. I. The effect of specific amino acid sequence represented by specific inter-unit interactions. *Int. J. Pept. Prot. Res.*, 7, p. 445–459, 1975.

[171] TAVERNA, D. M. et R. A. GOLDSTEIN. The distribution of structures in evolving protein populations. *Biopolymers*, 53, p. 1–8, 2000.

[172] Taverna, D. M. et R. A. Goldstein. Why are proteins marginally stable? *Proteins*, 46, p. 105–109, 2002.

[173] ————. Why are proteins so robust to site mutations? *J. Mol. Biol.*, 315(3), p. 479–484, 2002.

[174] Teichmann, J. P. S. A., T. Hubbard et C. Chothia. Intermediate sequences increase the detection of homology betwenn sequences. *J. Mol. Biol.*, 273, p. 349–354, 1997.

[175] Thom, G., A. C. Cockroft, A. G. Buchanan, C. J. Candotti, E. S. Cohen, D. Lowne, P. Monk, C. P. Shorrock-Hart, L. Jermutus et R. R. Minter. Probing a protein-protein interaction by *in vitro* evolution. *Proc. Natl. Acad. Sci.*, 103(20), p. 7619, 7624 2006.

[176] Thomas, P. D. et K. A. Dill. Local and nonlocal interactions in globular proteins and mechanisms of alcohol denaturation. *Protein Science*, 2, p. 2050–2065, 1993.

[177] ————. Statistical potentials extracted from protein structures: How accurate are they? *J. Mol. Biol.*, 257, p. 457–469, 1996.

[178] Tiana, G., D. Provasi et R. A. Broglia. Role of bulk and of interface contacts in the behaviour of lattice model dimeric proteins. *Phys. Rev. Lett.*, 67, 2003.

[179] Tiana, G. et R. A. Broglia. Design and folding of dimeric proteins. *Proteins: Structure, Function and Genetics*, 49, p. 82–94, 2002.

[180] Tiana, G., B. E. Shakhnovich, N. V. Dokholyan et E. I. Shakhnovich. Imprint of evolution on protein structures. *Proc. Natl. Acad. Sci.*, 101(9), p. 2846–2851, 2004.

[181] Trinquier, G. et Y.-H. Sanejouand. New protein-like properties of cubic lattice models. *Phys. Rev. E*, 59(1), p. 942–946, 1999.

[182] Tuffery, P., C. Etchebest, S. Hazout et R. Lavery. A new approach to the rapid determination of protein side chain conformations. *Journal of biomolecular structure and dynamics*, 8(6), p. 1267–1289, 1991.

[183] Valdar, W. S. et J. M. Thornton. Conservation helps to identify biologically relevant crystal contacts. *J. Mol. Biol.*, 313, p. 399–416, 2001.

[184] van Nimwegen, E., J. P. Crutchfield et M. Huygen. Neutral evolution of mutational robustness. *Proc. Natl. Acad. Sci.*, 96, p. 9716–9720, August 1999.

[185] Vanderzande, C. *Lattice Models of Polymers*. Cambridge University Press, 1998.

[186] Vendruscolo, M. et E. Domany. Pairwise contact potentials are unsuitable for protein folding. *Journal of Chemical Physics*, 2006.

[187] VENTER, J. C. *et al.* The sequence of the human genome. *Science*, 291(1304-1351), 2001.

[188] WANG, J. et W. WANG. A computational approach to simplifying the protein folding alphabet. *Nature Structural Biology*, 6, p. 5811–5816, 1999.

[189] WANG, Z.-X. How many fold types of protein are there in nature? *Proteins*, 26, p. 186–191, 1996.

[190] WARD, R. D. Relationship between enzyme heterozygosity and quaternary structure. *Biochem. Genet.*, 15, p. 123–135, 1977.

[191] WERNISCH, L., S. HERY et S. J. WODAK. Automatic protein design with all atom force-fields by exact and heuristic optimization. *J. Mol. Biol.*, 301, p. 713–736, 2000.

[192] WHITTLE, C.-A. et M. O. JOHNSTON. Moving forward in determining the causes of mutations: The features of plants that make them suitable for assessing the impact of environmental factors and cell age. *Journal of Experimental Botany*, 57, p. 1847–1855, 2006.

[193] WILKE, C. O., J. L. WAND, C. OFRIA, R. E. LENSKI et C. ADAMI. Evolution of digital organisms at high mutation rates leads to survival of the flattest. *Nature*, 412, p. 331–333, 2001.

[194] WILKE, C. O. Adaptive evolution on neutral network. *Bulletin of Mathematical Biology*, 63(4), p. 715–730, 2001.

[195] ———— Molecular clock in neutral protein evolution. *BMC Genetics*, 5, 2004.

[196] WILLIAMS, P. D., D. D. POLLOCK et R. A. GOLDSTEIN. Evolution of functionality in lattice proteins. *Journal of Molecular Graphics and Modelling*, 19, p. 150–156, 2001.

[197] ————. Selective advantage of recombination in evolving populations: A lattice model study. *International Journal of Modern Physics C*, 17(1), p. 75–90, 2006.

[198] WODAK, S. J. et J. JANIN. Structural basis of macromolecular recognition. *Advances in Protein Chemistry*, 61, p. 9–73, 2002.

[199] WRIGHT, S. Evolution in mendelian populations. *Genetics*, 16, p. 97–159, 1931.

[200] ———— The roles of mutation, inbreeding, crossbreeding, and selection in evolution. In *Proceedings of the Sixth International Congress on Genetics*, pages 355–366, 1932.

[201] ———— *Evolution and the Genetics of Populations: Vol 3. Experimental results and Evolutionary Deductions*. University of Chicago Press, 1977.

[202] Xia, Y. et M. Levitt. Funnel-like organization in sequence space determines the distributions of protein stability and folding rate preferred by evolution. *Proteins*, 55(1), p. 107–114, 2004.

[203] ————————. Roles of mutation and recombination in the evolution of protein thermodynamics. *Proc. Natl. Acad. Sci.*, 99(16), p. 10382–10387, August 2002.

[204] ————————. Simulating protein evolution in sequence space and structure space. *Curr. Opin. in Struct. Biol.*, 14, p. 202–207, 2004.

[205] Xu, D., C. J. Tsai et R. Nussinov. Mechanism and evolution of protein dimerization. *Protein Science*, 7(3), p. 533–544, 1998.

[206] Yahyanejad, M., M. Kardar et C. Tang. Structure space of model proteins: A principal component analysis. *J. Chem. Phys.*, 118, p. 4277–4284, 2003.

[207] Zeldovich, K. B., I. N. Berezovsky et E. I. Shakhnovich. Physical origins of protein superfamilies. *J. Mol. Biol.*, 357, p. 1335–1343, 2006.

[208] Zhang, Y., I. A. Hubner, A. K. Arakaki, E. I. Shakhnovich et J. Skolnick. On the origin and highly likely completeness of single-domain protein structures. *Proc. Natl. Acad. Sci.*, 103, p. 2605–2610, 2006.

[209] Zuckerkandl, E. Evolutionary processes and evolutionary noise at the molecular level. II. A selectionist model for random fixations in proteins. *Journal of Molecular Evolution*, 26(7), p. 269–311, 1976.

[210] Zuckerkandl, E. et L. Pauling. Molecules as documents of evolutionary history. *Journal of Theoretical Biology*, 8, p. 357–366, 1965.