Une nouvelle fonction de score pour l'amarrage protéine-protéine fondée sur les diagrammes de Voronoï

Julie Bernauer, Jérôme Azé, Joël Janin, Anne Poupon

Une nouvelle fonction de score pour l'amarrage protéine-protéine fondée sur les diagrammes de Voronoï

Julie Bernauer[1], Jérôme Azé[2], Joël Janin[3], Anne Poupon[1]

[1] Equipe Génomique Structurale, IBBMC, Bât 430, Université Paris-Sud, 91405 Orsay Cedex
julie.bernauer@ibbmc.u-psud.fr
anne.poupon@ibbmc.u-psud.fr
[2] Equipe Bioinformatique, LRI, Bât 490, Université Paris-Sud, 91405 Orsay Cedex
aze@lri.fr
[3] Laboratoire d'Enzymologie et Biochimie Structurales, Bât 34, CNRS, 91198 Gif-sur-Yvette.
janin@lebs.cnrs-gif.fr

Résumé: *Dans cet article, nous décrivons l'interaction protéine-protéine dans le cadre du modèle de l'énergie stochastique (Random Energy Model) de la physique statistique. Nous simulons en amarrant les deux protéines partenaires, le processus d'association de deux protéines formant un complexe. Nous obtenons les spectres d'énergie d'un jeu de complexes protéine-protéine de structure tridimensionnelle connue en effectuant l'amarrage dans des orientations aléatoires et en calculant un score pour les complexes modèles ainsi générés. Nous utilisons une représentation très simplifiée de la structure où chaque acide aminé est remplacé par sa cellule de Voronoï, et nous appliquons le programme d'apprentissage par algorithme génétique ROGER à un ensemble de paramètres mesurés sur cette représentation, afin de dériver une fonction de score appropriée. En considérant les scores obtenus comme des énergies d'interaction, nous obtenons le spectre d'énergie de chaque complexe. Il s'approche d'une distribution gaussienne qui nous permet de calculer des paramètres physiques comme la température de transition vitreuse et la température de transition spécifique du système.*

Mots-clé: interaction protéine-protéine, diagramme de Voronoï, amarrage protéine-protéine

1 Introduction

La reconnaissance spécifique d'une macromolécule biologique par une autre est un processus fondamental dans tous les domaines de la biologie, et les biologistes structuraux s'efforcent de lui donner une base physico-chimique au niveau atomique. Dans le cas de l'ADN, la double hélice et la complémentarité des bases nucléiques constituent une solution élégante découverte par Watson et Crick il y a une cinquantaine d'années. Mais pour les protéines, nous n'avons pas réponse générale de cette nature. Les protéines sont conçues pour interagir de manière spécifique avec des objets d'une grande diversité, allant de l'ion métallique aux petites molécules organiques et aux protéines, à l'ADN et l'ARN, chacun avec un mode de reconnaissance différent. La reconnaissance protéine-protéine est le processus de reconnaissance moléculaire le plus hétérogène, et l'un des plus étudiés, notamment grâce à l'accumulation des données provenant du séquençage systématique des génomes et à des expériences de génétique ou de biochimie conduite à l'échelle de génomes entiers [1-3].

La simulation numérique est un moyen parmi bien d'autres d'approcher la reconnaissance spécifique. Les algorithmes d'amarrage (*docking*) prennent deux molécules définies par leurs coordonnées atomiques et cherchent des conformations favorables pour leur association. Pour les protéines, le premier algorithme d'amarrage est dû à Wodak et Janin [4, 5]. Les deux protéines sont décrites par un ensemble de centroïdes de résidus d'acides aminés liés de manière rigide, ce qui réduit à six le nombre de degrés de liberté du système. De nombreux autres algorithmes ont été mis au point ces dernières années pour réaliser la même tâche, et nous renvoyons le lecteur aux revues [6-8] qui les décrivent. Une application majeure des algorithmes d'amarrage est la prédiction de la structure de complexes à partir de celles de leurs composantes. Les performances de ces procédures font l'objet d'une expérience d'évaluation appelée CAPRI (*Critical Assessment of PRedicted Interactions*) [9], où il s'agit de prédire à l'aveugle la structure de complexes qui ont fait l'objet d'une étude expérimentale non encore publiée. Les résultats de CAPRI suggèrent que les algorithmes existants explorent efficacement l'espace des conformations, mais qu'il reste difficile de donner un score aux solutions. La solution "native" – celle qui correspond à la structure expérimentale - est souvent

noyée dans les faux positifs. La recherche de fonctions de score qui représentent correctement la physique du phénomène et éliminent les faux positifs est donc un champ très actif. Ceci peut se faire de manière empirique en examinant les caractéristiques des complexes protéine-protéine de structure connue et déposée dans la Protein Data Bank (PDB) [10]. Nous utilisons ici un modèle simplifié des protéines où nous remplaçons la description atomique par un ensemble de polyèdres de Voronoï construits autour de chaque acide aminé. Nous définissons un ensemble de 84 paramètres que nous mesurons sur la représentation de Voronoï de 79 complexes protéine-protéine: d'abord sur leur structure expérimentale, puis sur des modèles générés par amarrage des deux partenaires de chaque complexe dans des orientations aléatoires. Nous utilisons le programme DOCK [11] pour l'amarrage et le programme d'apprentissage par algorithme génétique ROGER [12, 13] pour déduire une fonction de score des valeurs de ces paramètres.

Au lieu de prédire la structure de complexes, nous utilisons ici l'algorithme d'amarrage DOCK et la fonction de score générée par ROGER pour générer des spectres d'énergie d'interaction que nous analysons à l'aide du modèle de l'énergie stochastique (*Random Energy Model* ou REM) de la physique statistique [14]. Le REM a un grand nombre d'applications dans l'étude des protéines, y compris pour l'étude de la reconnaissance spécifique [15-17]. Nous obtenons un spectre d'énergie en amarrant deux protéines dans des orientations aléatoires et en considérant le score des solutions obtenues comme une énergie d'interaction. Même si le score n'est pas suffisant, dans l'état actuel, pour discriminer les solutions natives des faux positifs, nous constatons que le REM décrit correctement sa distribution statistique, ce qui nous permet de déterminer des paramètres physiques comme la température de transition vitreuse ou la température de transition spécifique de chaque système.

2 Méthodes et résultats

2.1 Une représentation de l'interaction protéine-protéine fondée sur la tessellation de Voronoï

Soit un ensemble de nœuds $\{p_i\}$ dans l'espace à trois dimensions. La cellule de Voronoï du nœud p_i est l'ensemble des points de l'espace qui sont plus proches de p_i que de tout autre nœud. Les cellules de Voronoï sont des polyèdres convexes qui peuvent être non bornés. Chaque point de l'espace appartient soit à une cellule, soit à une face commune entre deux cellules: les cellules de Voronoï forment donc une tessellation. Pour des raisons de complexité algorithmique, la tessellation de Voronoï se construit de préférence à partir de son dual, la triangulation de Delaunay implémentée dans la bibliothèque CGAL (*Computational Geometric Algorithm Library*) [18]. Cette construction incrémentale aléatoire permet d'atteindre la complexité optimale qui est $O(n^2)$ en trois dimensions.

La structure d'une protéine peut se décrire par une tessellation de Voronoï dont les nœuds sont les atomes [19], ou comme nous le faisons ici, les centres géométriques des chaînes latérales des acides aminés (plus le C_α) [20]. On doit y ajouter des nœuds qui représentent des molécules de solvant pour fermer les cellules de Voronoï des résidus qui sont à la surface. Pour ce faire, nous plaçons sur un réseau cubique des sphères de diamètre 6,5 Å comme indiqué dans la référence [20]. Le diamètre choisi donne aux cellules de solvant un volume moyen égal à la moyenne des volumes des cellules des acides aminés. Cette représentation décrit correctement l'empilement atomique et d'autres propriétés structurales des protéines [20-22]. Son application aux complexes protéine-protéine permet les définitions suivantes :

- Deux résidus sont voisins au sens de Voronoï si leurs cellules partagent une face ;
- Un résidu appartient à l'intérieur de la protéine si tous ses voisins sont des résidus de la même protéine ;
- Un résidu appartient à la surface de la protéine si l'un au moins de ses voisins est de type solvant ;
- Un résidu appartient à l'interface protéine-protéine si l'un au moins de ses voisins appartient à l'autre protéine ;
- Un résidu de l'interface appartient au cœur de l'interface si aucun de ses voisins n'est de type solvant ;
- L'ensemble des faces partagées par les deux protéines constitue l'interface ;

2.2 Echantillonnage systématique des modèles issus de l'amarrage

Sur la base d'études déjà publiées [23, 24], nous avons sélectionné dans la PDB 79 fichiers qui contiennent des complexes protéine-protéine de structure cristallographique connue et illustrent la reconnaissance spécifique : 1a2k 1acb 1agr 1ahw 1aip 1ak4 1ao7 1atn 1avw 1avz 1bql 1bth 1bvk 1cbw 1cgi 1dan 1dee 1dfj 1dhk 1dkg 1dvf 1efu 1eo8 1fbi 1fc2 1fin 1fle 1fq1 1fss 1gg2 1gla 1got 1gua 1hez1 1hez2 1hia 1hwg 1iai 1igc 1jhl 1kb5 1mah 1mda 1mhh 1mkw 1mlc 1nca 1nfd 1nmb 1nsn 1osp 1ppf 1qfu 1seb 1spb 1stf 1tab 1tbq 1tco 1tgs 1toc 1tx4 1udi 1ugh 1vfb 1wej 1wq1 1ycs 2btf 2jel 2kai 2pcc 2sic 2trc 2vir 3hfl 3hfm 3hhr 4htc.

Nous avons séparé les composantes de chacun de ces 79 complexes et utilisé DOCK [11] pour les reconstituer dans des orientations aléatoires. DOCK utilise cinq angles et une distance pour décrire les degrés de liberté d'une composante par rapport à l'autre. Pour chaque combinaison des cinq angles, DOCK détermine la distance qui amène les deux protéines en contact et à l'aide de l'algorithme de Wodak et Janin [5], il construit ce que nous appellerons ici un modèle d'amarrage. Le modèle natif étant la structure cristallographique, l'amarrage dans toute autre orientation génère un modèle non-natif. Nous couvrons l'ensemble de l'espace des solutions en échantillonnant les cinq angles par pas de 10°, ce qui conduit pour chaque complexe à 18.10^6 modèles parmi lesquels nous sélectionnons un jeu de modèles de manière aléatoire.

2.3 Les paramètres du score

La fonction qui nous servira à attribuer un score aux modèles d'amarrage comporte 84 variables dérivées de la tessellation de Voronoï. Ces variables concernent l'interface et appartiennent à six classes :

P1 L'aire de l'interface (1 paramètre) ;

P2 Le nombre de résidus dans le cœur de l'interface (1 paramètre) ;

P3 La fréquence de chaque type de résidu à l'interface (20 paramètres)

P3 Le volume de Voronoï moyen correspondant (20 paramètres) ;

P4 La fréquence des paires de résidus en contact, calculée pour chaque type de paire (21 paramètres) ;

P5 La distance moyenne entre les centroides de résidus en contact, calculée par type de paire (21 paramètres).

Pour réduire le nombre de paramètres des catégories P5 et P6, les 20 acides aminés sont classés en six catégories physico-chimiques: (ILMV), 5FYW), (HKR), (DE), (NQ), (AGSTCP).

Nous avons mesuré les valeurs de ces paramètres, d'une part sur les modèles natifs des 79 complexes, et d'autre part, sur les modèles non-natifs produits par DOCK. Ces valeurs sont comparées dans la figure 1. En effet, il n'est possible de générer une fonction de score discriminante qu'à partir de descripteurs prenant des valeurs suffisamment différentes sur les deux catégories. On peut voir que c'est le cas des descripteurs que nous avons choisis. À quelques exceptions près, les différences sont statistiquement significatives. Sans surprise, on constate que les interfaces natives sont plus hydrophobes que les non-natives, et que les paires de résidus favorables d'un point de vue physico-chimique, notamment celles qui comportent un résidu hydrophobe, sont mieux représentées dans les interfaces natives. Enfin, les paramètres de volume montrent que les interfaces non-natives ont des empilements incorrects: les volumes des résidus polaires ou chargés sont trop grands, ceux des résidus hydrophobes et neutres sont trop petits. Les différence observées sur les distances vont dans le même sens: les résidus hydrophobes ou neutres sont trop près les uns des autres dans les interfaces non-natives, alors que les résidus polaires ou chargés sont trop loin.

La meilleure qualité de l'empilement des interfaces natives résulte en partie de réarrangements locaux des chaînes latérales lors de la formation du complexe. Pour tester l'effet de ces réarrangements sur nos paramètres, nous avons généré des modèles "quasi-natifs" de certains des complexes. Pour cela, nous utilisons les structures cristallographiques des deux partenaires cristallisés indépendamment et nous les amarrons avec DOCK en échantillonnant chaque angle dans un intervalle de 5° de part et d'autre de sa valeur native par pas de 1°. Les modèles ainsi obtenus ont une orientation correcte ou presque. Les résidus présents à l'interface sont ceux de l'interface native, mais les chaînes latérales peuvent avoir des conformations différentes. Nous avons constaté que les valeurs des paramètres changent peu par rapport aux modèles natifs,

et qu'ils restent significativement différents de ceux des non-natifs. La figure 2 donne les distances moyennes entre centroïdes de paires de résidus en contact à l'interface (paramètres P6) qui sont mesurées sur les modèles natifs, quasi-natifs et non-natifs. La distance aromatique-aromatique est presque deux fois plus faible dans les non-natifs que les natifs ou quasi-natifs. Cela suggère que les résidus aromatiques qui sont à l'interface ont en général leur chaîne latérale dans la bonne conformation avant la formation du complexe.

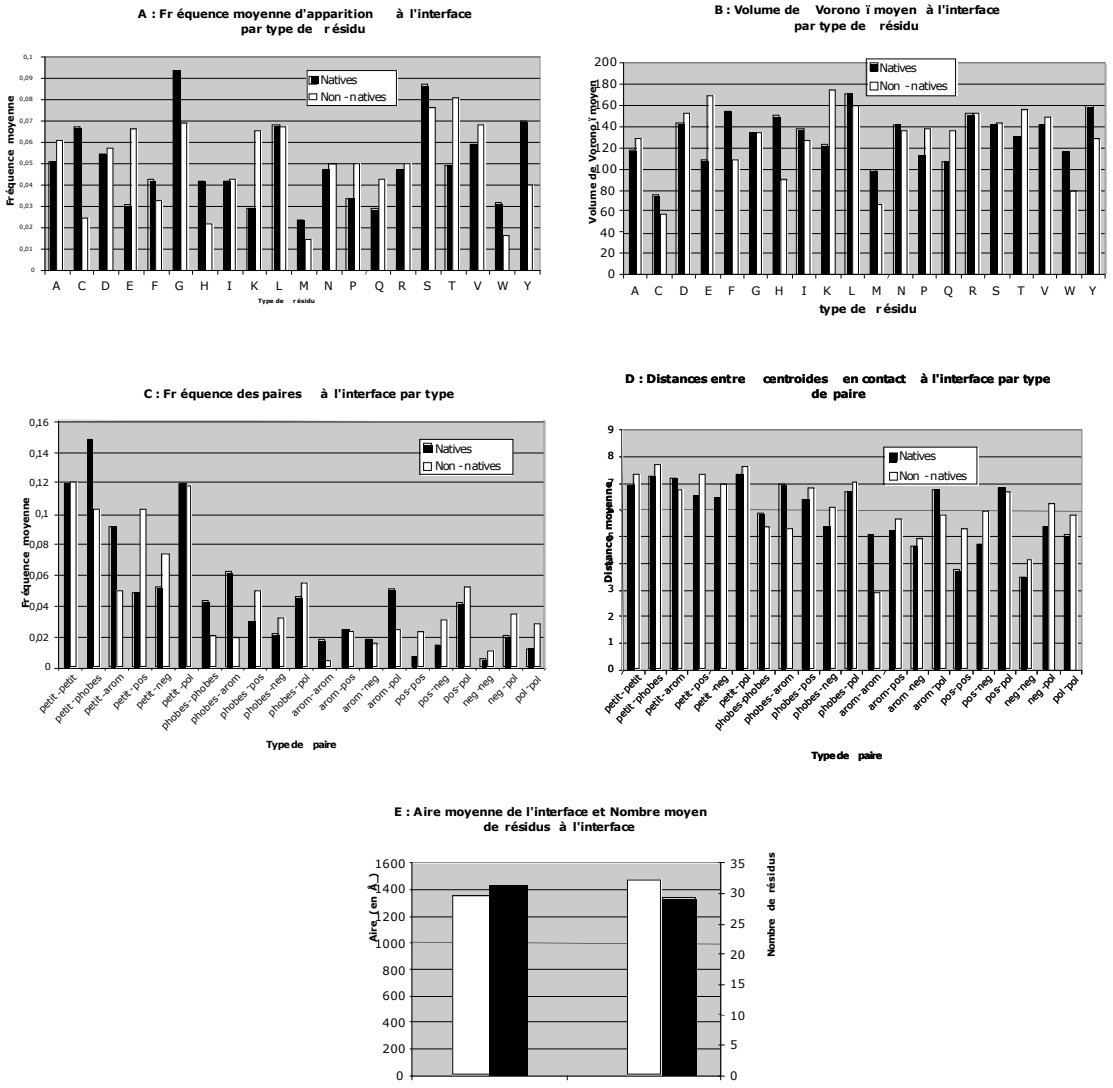

Figure 1 : Les paramètres utilisés dans l'apprentissage Histogrammes des valeurs mesurées sur les complexes natifs et non natifs du jeu d'apprentissage. A : paramètres P3, B : paramètres P4, C : paramètres P5, D : paramètres P6. Pour le calcul des paramètres P5 et P6, les acides aminés sont regroupés en six catégories : hydrophobes (phobes, IVLM), aromatiques (arom, YFW), chargés positivement (pos, HKR), chargés négativement (neg, DE), polaires (pol, NQ) et petits (AGSTCP).

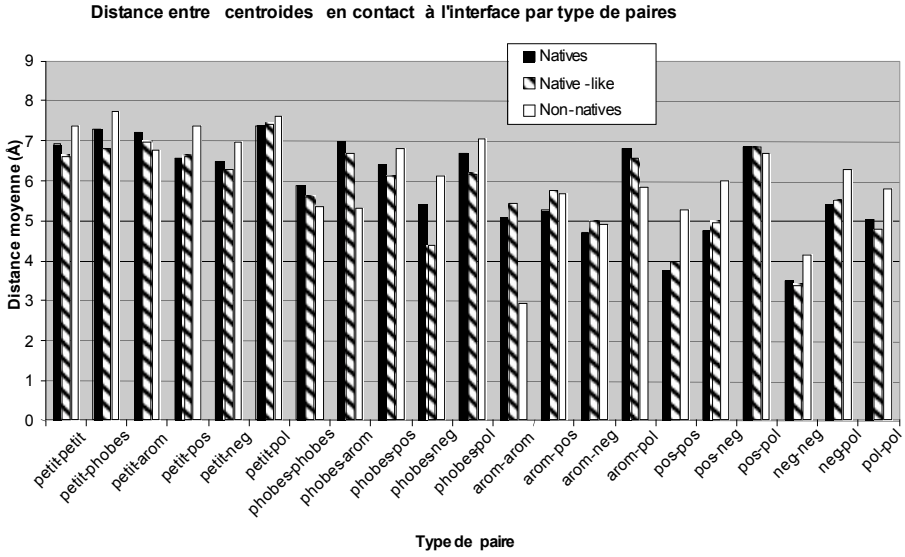

Figure 2 : Distances moyennes entre les centroides des résidus en contact à l'interface (paramètres P6), par type de paire. Les paramètres P6 ont été ici calculés pour les structures cristallographiques des 79 complexes (Natives), les solutions non-natives générées par DOCK (Non-natives), et des solutions générées par DOCK, à partir des structures cristallographiques individuelles des partenaires, avec des valeurs d'angles très proches de celles de la structure cristallographique (Native-like).

2.4 Calcul de la fonction de score par ROGER

La fonction de score que nous utilisons est de la forme :

$$(1) \qquad f(x)=\sum_i w_i \, | x_i - c_i |$$

Le programme ROGER [12,13] détermine les poids w_i et les valeurs de centrage c_i. ROGER (*ROc based GEnetic learneR*) est conçu pour maximiser l'aire sous la courbe de ROC (*Area Under the ROC Curve* ou AUC). La courbe de ROC (*Receiver Operating characteristics Curve*), empruntée au domaine du traitement du signal, représente le taux de vrais positifs en fonction du taux de faux positifs. Dans l'hypothèse idéale où il y a 100% de vrais positifs et aucun faux positif, la courbe de ROC est une fonction en escalier et l'AUC vaut 1. Dans le cas où la sélection est aléatoire, les taux de vrais et de faux positifs sont égaux et l'AUC vaut 0,5. Comme l'objectif d'un algorithme d'apprentissage est de maximiser le taux de vrais positifs tout en minimisant celui des faux positifs, l'AUC est une bonne mesure globale de l'efficacité de l'apprentissage. Son optimisation est un problème NP-complexe et ROGER utilise un algorithme génétique pour le résoudre.

Notre jeu d'apprentissage contient les valeurs des 84 paramètres décrits ci-dessus que nous avons mesurées sur les 79 complexes cristallographiques et sur 8400 modèles non-natifs de ces complexes (Figure 3). Ces modèles sont générés par DOCK en échantillonnant les angles par pas de 10° et regroupant les solutions par la procédure géométrique décrite dans [11]. Chaque groupe est alors représenté par la structure moyenne, et 8400 groupes sont sélectionnés aléatoirement. Les paramètres ne peuvent pas tous se mesurer sur chaque solution. Par exemple, si tel ou tel type d'acide aminé n'est pas présent à l'interface d'un modèle donné, les catégories P3, P4, P5 et P6 ont des valeurs manquantes. Comme ROGER n'est pas conçu pour traiter des jeux comportant des données manquantes, nous remplaçons celles-ci par la valeur médiane du paramètre dans la catégorie correspondante du jeu d'apprentissage (natif ou non-natif).

Nous avons fait tourner ROGER sur ce jeu avec un modèle non linéaire dans une stratégie d'évolution (20+200) où 20 parents sont sélectionnés de manière déterministe dans la génération précédente, auxquels viennent s'ajouter 200 descendants générés par mutation auto-adaptative et mutations croisées à un taux de 60%. Ce processus est exécuté 21 fois, menant à 21 fonctions de score, correspondant chacune à la meilleure fonction obtenue pour chaque exécution. Les résultats sont évalués en effectuant une validation croisée en 10 groupes, menant à 210 fonctions de score. Pour attribuer un score unique à chaque modèle, les 210 fonctions sont évaluées et une sélection est effectuée, des feuilles à la racine, dans l'arbre associée à la validation

croisée en 10 groupes, en prenant d'abord la valeur médiane des 21 exécutions indépendantes pour chaque groupe, puis en prenant la médiane des 10 valeurs obtenues.

2.5 Relation entre énergie et entropie dans le modèle de l'énergie stochastique

Nous considérons chaque modèle d'amarrage comme étant un état du système et son score comme étant l'énergie E de l'interaction entre les deux composants protéiques dans cet état. Nous analysons la distribution de E comme indiqué dans [17] en comptant les états d'énergie comprise entre E et $E+dE$. Leur nombre étant $m(E)$, l'entropie s'écrit:

$$(2) \qquad S(E) = k_B \ln [\, m(E) \,]$$

L'état natif possède une énergie E_0 que nous prendrons égale à zéro par commodité. Comme il est unique, $S(E_0) = 0$. Nous noterons Δ la différence entre l'énergie de l'état natif et l'énergie de l'état non-natif de plus basse énergie. Le nombre total d'états, natif compris, est :

$$(3) \qquad N = 1 + \int_{\Delta}^{+\infty} m(E).dE$$

À l'équilibre thermodynamique et à la température T, tous ces états coexistent avec une abondance relative $n(E)$ qui suit la loi de Boltzmann :

$$(4) \qquad n(E) = m(E) \exp\left(-\frac{E}{k_B T}\right)$$

Nous pouvons écrire la fonction de partition Z du système sous la forme:

$$(5) \qquad Z = 1 + r = 1 + \int_{\Delta}^{+\infty} n(E).dE$$

L'état natif contribue 1 à Z, les états non-natifs, r. Comme la spécificité de l'interaction implique $r \ll 1$, la différence d'énergie Δ doit être grande par rapport à l'énergie thermique $k_B T$. La condition $r=1/2$ définit une température T_S, que nous avons appelée la *température de transition spécifique* [17]: en dessous de T_S, l'état natif domine, au dessus de T_S, les états non-natifs sont majoritaires. T_S peut s'obtenir en traçant la tangente à la courbe $S(E)$ qui passe par l'origine. La tangente au point $E = \Delta$ définit une autre température caractéristique notée T_C. C'est la température critique du système, également appelée *température de transition vitreuse* dans l'étude des verres de spin [14]. En dessous de T_C, seuls les états d'énergie proche de Δ entrent en compétition avec l'état natif. Au-dessus de T_C, beaucoup d'états d'énergie plus élevée sont peuplés. T_S et T_C se calculent facilement en approchant par une expression analytique le spectre d'énergie. Le REM suppose que cette expression analytique est gaussienne [14].

2.6 Le spectre d'énergie d'un complexe protéine-protéine

À partir d'un complexe entre l'hémaglutinine du virus de la grippe et le fragment FAB d'un anticorps monoclonal [entrée PDB 1eo8, 25], nous avons généré des modèles d'amarrage à l'aide de DOCK en échantillonnant les angles par pas de 10°. Nous avons sélectionné aléatoirement 10.000 de ces modèles et calculé leur score. La Figure 3 décrit la distribution de ces scores. ROGER est conçu pour affecter le score le plus faible possible aux états natifs, et le plus élevé possible aux états non-natifs. Dans le cas de ce complexe, l'état natif a effectivement le score le plus faible (0,258) et il y a un écart significatif avec les états non-natifs dont les scores se situent entre 0,3 et 1,5. Les scores non-natifs ont une moyenne de 1,01, mais leur distribution est légèrement asymétrique: il y a moins d'états en-dessous qu'au-dessus de la moyenne, et la valeur médiane est 1,02. Néanmoins, la figure 2 montre que la partie gauche de l'histogramme des énergies se modélise bien par une gaussienne.

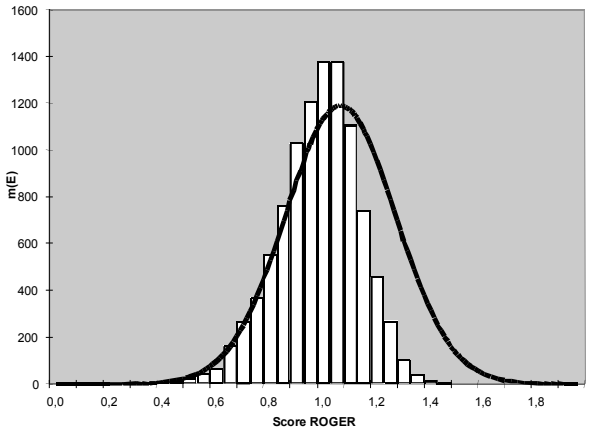

Figure 3 : Spectre d'énergie du complexe hémaglutinine-anticorps (PDB 1eo8). 10.000 solutions d'amarrage ont été sélectionnées aléatoirement et leurs scores sont représentés par un histogramme. Le score de l'état natif est 0,258. La partie gauche du spectre est approchée par une gaussienne (R^2=0,994).

Si l'on admet que le score ROGER représente l'énergie d'interaction entre l'hémaglutinine et l'anticorps, la figure 3 donne le spectre d'énergie de ce complexe et nous pouvons utiliser l'équation (2) pour calculer son entropie. Dans la figure 4 qui décrit l'entropie en fonction de l'énergie, la partie gauche de la courbe *S(E)* est approchée par une parabole équivalente à la gaussienne de la figure 3. La courbe dévie de la parabole pour des scores supérieurs à 1 mais cette région ne contient que des états de haute énergie et l'équation (4) montre qu'ils ne sont peuplés qu'à très haute température. La parabole coupe l'axe des abscisses en deux points, Δ et Δ'. La différence Δ-Δ' est la largeur du spectre. Si l'on admet que la distribution est gaussienne et que l'énergie de l'état natif est 0 (alors que sa valeur réelle est 0,258), Δ-Δ' est le *gap* (l'écart d'énergie) qui sépare l'état natif et l'état non-natif de plus basse énergie. Dans cette hypothèse, la pente de la tangente à la parabole passant par l'origine est $1/T_S$, et la pente de la tangente au point E=Δ est $1/T_C$. Nous obtenons pour la température de transition spécifique T_S une valeur 2,3 fois supérieure à celle de la température de transition vitreuse T_C.

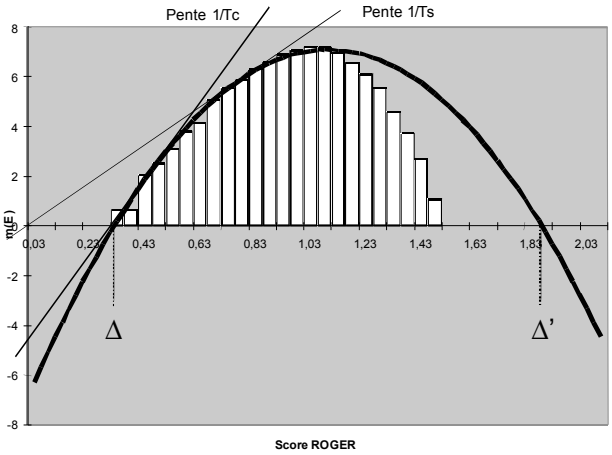

Figure 4 : Courbe énergie-entropie du complexe hémaglutinine-anticorps 1eo8. Les points sont déduits de l'histogramme de la figure 3, la ligne pleine est la parabole qui correspond à la gaussienne de cette figure. Elle coupe l'axe des abscisses aux points =0,258 et Δ'=1,87. La tangente au point E=Δ a une pente $1/T_C$, et la tangente à la parabole qui passe par l'origine a une pente $1/T_S$. Dans le REM [17], T_C = 0,116 est la température de transition vitreuse et T_S = 0,05 la température de transition spécifique.

Nous avons analysé de la même manière les autres complexes du jeu d'apprentissage. Pour chacun, nous

avons calculé les scores ROGER de 4000 états non-natifs sélectionnés aléatoirement. Sept complexes se comportent comme 1eo8 : leur état natif a le score le plus bas, et un gap significatif le sépare de l'état non-natif de plus basse énergie. Dans les autres complexes, l'état natif n'est pas celui de plus basse énergie: en moyenne, il se classe en 157ème position en termes de score ROGER. Après avoir normalisé *m(E)* pour avoir *N*=10.000 comme dans le cas de 1eo8, nous avons tracé l'histogramme des énergies et la courbe énergie-entropie de chaque complexe. Dans tous les cas sauf un, la partie gauche de la courbe *S(E)* se décrit correctement par une parabole ($R^2 > 0.98$) dont on peut déduire les paramètre physiques du REM. Ceux qui figurent dans le tableau 1 sont des valeurs moyennes. Les paraboles sont centrées sur des valeurs de *E* proches de 1 et ont une largeur entre 1 et 3. Comme la température de transition spécifique ne peut se calculer que si l'état natif est l'état de plus basse énergie, nous avons supposé dans le Tableau 1 comme dans la figure 4 que l'état natif a un score de zéro. Dans ces conditions le rapport sans dimension T_S/T_C, va de 1,7 à 3,4 avec une moyenne de 2.5 et un écart-type relativement faible.

Tableau 1 : Paramètres physiques dérivés des spectres d'énergie des complexes protéine-protéine. **a**: valeur moyenne et l'écart type des paramètres obtenus sur 4000 modèles d'amarrage de 70 complexes protéine-protéine excluant 1efu, 1fc2, 1nmb, 1qfu, 2btf, 2kai, 3hfl, 3hhr et 4htc pour des raisons techniques. **b**: les valeurs de Δ, Δ', T_S et T_C ont été obtenus en modélisant par une parabole la courbe entropie-énergie de chaque complexe et en faisant l'hypothèse que l'état natif à un score *E*=0.

Paramètre[a]	Moyenne	Ecart type
Rang du modèle natif	157	309
Score du modèle natif	0,52	0,15
Score non natif moyen	0,88	0,12
Différence d'énergie Δ[b]	0,29	0,08
Centre de la parabole	0,97	0,27
Largeur du spectre Δ-Δ'	1,36	0,57
T_S/T_C	2,5	0,4

3. Discussion

3.1 Le score ROGER

Nous avons défini une fonction de score pour les modèles d'amarrage en appliquant le programme d'apprentissage ROGER à un jeu de paramètres mesurés sur le diagramme de Voronoï de 79 complexes protéine-protéine. Le score ainsi obtenu ne discrimine pas efficacement l'état natif d'un complexe des états non-natifs de basse énergie que génère l'amarrage de ses composantes dans des orientations aléatoires. De fait, l'état natif n'a la plus basse énergie que dans 10% des cas testés. Même si à ce stade, il n'est pas directement utilisable pour la prédiction de la structure des complexes, le score ROGER peut être utilisé, pour représenter l'énergie d'interaction dans le cadre du modèle d'énergie stochastique REM. La plupart des paramètres qui interviennent dans ce score ont une signification physique. Ils contribuent à l'enthalpie libre d'association ΔG_a du complexe et donc à sa stabilité en solution. Par exemple, l'aire de l'interface (paramètre P1) est proportionnelle à la contribution de l'effet hydrophobe à ΔG_a [26-28]. Les volumes de Voronoï des résidus (paramètres P3) décrivent l'empilement atomique qui détermine l'énergie des interactions de Van der Waals. La composition en acide aminés du cœur de l'interface est différente de celle du reste de la surface de chacun des partenaires. Les paramètres P4 expriment donc une préférence pour l'interface dont on sait qu'elle est corrélée à la solubilité dans l'eau des acides aminés et à leur enthalpie libre de désolvatation [29-31]. Enfin, les paramètres des classes P5 et P6 reflètent le rôle des interactions électrostatiques et Van der Waals à l'interface [32,33].

L'échelle des contributions à ΔG_a est différente pour chaque paramètre et l'une des tâches de l'algorithme d'apprentissage est de la pondérer de manière appropriée. Le faible pouvoir discriminant du score peut avoir plusieurs origines. Premièrement, il se peut que ROGER n'ait pas trouvé les poids les mieux adaptés en raison du trop petit nombre d'états natifs dans le jeu d'apprentissage. Deuxièmement, la représentation simplifiée des protéines que nous avons choisie efface tous les détails atomiques. Enfin, nos paramètres ne représentent qu'une partie de la physique du système. Nous avons supposé que les protéines s'assemblent à la manière de corps rigides et ignoré les coûts enthalpique et entropique d'éventuels changements de

conformation qui ont lieu dans de nombreux complexes de notre liste. Ils pourraient en théorie être pris en compte dans l'analyse, mais comme la nature et l'amplitude des changements que l'on observe lorsque deux protéines s'associent sont très variables d'un système à l'autre [23,34], un jeu de données beaucoup plus grand serait nécessaire pour qu'un algorithme d'apprentissage puisse en tirer un score performant.

3.2 Comment simuler l'association de deux molécules

Nous avons considéré la distribution des scores ROGER comme étant le spectre des énergies d'interaction et la quantité définie à l'équation (2) comme étant une entropie. Ceci peut se justifier en considérant que l'amarrage simule la collision aléatoire de deux molécules diffusant librement en solution. Cette collision crée une paire de contact qui peut se dissocier ou évoluer vers le complexe stable selon la région de la surface des deux protéines qui forme le contact et l'orientation relative des molécules au moment de la collision [35-40]. Une paire de contact qui se transforme en complexe stable constitue l'état natif, une paire qui se dissocie rapidement, un état non-natif. Les vitesses d'association que l'on mesure expérimentalement sur des systèmes anticorps-antigène ou enzyme-inhibiteur indiquent qu'une collision sur 10^3 à 10^6 est productive en l'absence d'interaction électrostatique à longue distance, [35, 40-42]. Ceci suggère que l'état natif est en compétition avec 10^3 à 10^6 états non-natifs. Les 10.000 états non-natifs du complexe hémaglutinine-antigène que nous avons générés sont dans cette fourchette. Quatre des paramètres angulaires de DOCK représentent la latitude et la longitude des régions de la surface de chacun des deux partenaires qui sont en contact après l'amarrage; le cinquième angle est une rotation autour de l'axe reliant les centres des deux molécules [4,5]. Comme un échantillonnage de chacun de ces cinq angles par pas de 36° génère environ 10.000 modèles, la valeur de S donnée par l'équation (2) est le coût entropique d'une sélection à 36° près du positionnement et de l'orientation des deux régions de la surface qui entrent en contact lors de la collision. Ce pas correspond à un déplacement de 4 Å sur la surface d'une protéine de rayon moyen 20 Å, une valeur plausible pour la portée des interactions locales qui discriminent entre les modes d'association natif et non-natifs [43]. Cette valeur détermine la largeur du puits énergétique qui mène à la formation d'un complexe stable [44-46].

3.3 Interaction spécifique et non spécifique

ROGER est conçu pour donner un score le plus proche possible de 0 aux solutions natives, et un score le plus grand possible aux solutions non natives. Les scores que nous obtenons pour le complexe hémaglutinine-anticorps vont de 0,25 à 3 (figure 3), et la parabole qui décrit la courbe entropie-énergie a une largeur de 1,6. D'autres complexes donnent des valeurs similaires. Le point E=Δ' correspond à des modèles d'amarrage ou l'interface a une aire presque nulle et où l'énergie d'interaction doit être proche de 0. Le point E=Δ correspond à l'état non-natif de plus basse énergie. Le modèle correspondant contient beaucoup de contacts atomiques, mais les interactions locales spécifiques sont absentes. L'empilement des protéines dans les cristaux donne des exemples d'association non-spécifique de ce type. Si la plupart des interfaces créées par les contacts cristallins ont une aire faible [48], certaines sont comparables en taille aux interfaces étudiées ici. Ces grandes interfaces cristallines ont des propriétés physico-chimiques qui diffèrent des interfaces spécifiques. En moyenne, elles sont moins hydrophobes et moins compactes, et elles contiennent moins de liaisons hydrogène [49]. Leur composition en acides aminés est similaire à celle du reste de la surface protéique et différente de celle des interfaces spécifiques. Les paramètres qui contribuent au score ROGER prennent en compte certaines de ces propriétés. Néanmoins, la discrimination entre interface cristalline et interface spécifique est un problème non trivial [49-52].

L'enthalpie libre des interactions cristallines est impossible à mesurer, mais comme la valeur moyenne de ΔG_a est de l'ordre de 10 à 20 kcal.mol^{-1} pour les complexes spécifiques, nous pouvons supposer qu'une unité de score ROGER représente environ 10 à 20 kcal.mol^{-1}. Dans ce cas, la valeur moyenne de citée dans le tableau 1 équivaut à 3 à 6 kcalmol^{-1}, soit 5 à 10 fois l'énergie thermique à 300K. Pour le complexe 1eo8, cela donne une valeur de T_S de l'ordre de 600 à 1200 K qui garantit que l'état natif est l'espèce dominante à 300K. Cependant, T_C est 2,5 fois inférieure à T_S. Si T_C est inférieure à la température ambiante, les modes non-natifs d'association qui entrent en compétition avec le mode natif ne sont pas des modes de basse énergie, mais un large échantillon appartenant à la moitié gauche du spectre. Dans ce cas, le modèle à deux états dans lequel tous les états non-natifs ont la même énergie est inapproprié, et le REM est une meilleure description des propriétés d'équilibre du système.

4. Conclusion et perspectives

La représentation par un modèle de Voronoï est bien adaptée à une première analyse de l'interaction protéine-protéine et utile pour simuler des collisions, mais elle ne tient pas compte de la structure atomique détaillée et le score que nous en tirons ne peut rendre compte avec précision de la spécificité des interactions. Cependant, il satisfait les hypothèses du modèle de l'énergie stochastique et nous permet de représenter des événements qui ont lieu en solution *in vitro* et *in vivo* et impliquent des modes non natifs d'association protéine-protéine. Ces modes contribuent au processus physique de la reconnaissance, même dans les cas où seul l'état natif est biologiquement pertinent.

Remerciements

Ce travail a été en partie financé par le programme EIDIPP, dans le cadre de l'ACI IMPBio.

Références bibliographiques

[1] Marcotte EM, Pellegrini M, Ng HL, Rice DW, Yeates TO, Eisenberg D. Detecting protein function and protein-protein interactions from genome sequences. *Science* 1999, 285:751-713.
[2] Janin J, Wodak SJ. Protein modules and protein-protein interaction : towards a global view. *Adv. Prot. Chem.* 2003, 61:1-8.
[3] Janin J, Séraphin B. Genome-wide studies of protein-protein interaction. *Curr. Op. Struc. Bio*. 2003, 13:383-388.
[4] Wodak S, Janin J. Computer analysis of protein-protein interactions *J. Mol. Biol*. 1978, 124:323.
[5] Janin J, Wodak SJ. Reaction pathway for the quaternary structure change in hemoglobin *Biopolymers* 1985, 24:509-526.
[6] Halperin I, Ma B, Wolfson H, Nussinov R. Principles of docking: An overview of search algorithms and a guide to scoring functions. *Proteins* 2002, 47:409-443.
[7] Wodak SJ, Janin J. The structural basis of macromolecular recognition. *Adv. Prot. Chem.* 2003, 61:9-73.
[8] Smith GR, Sternberg MJ. Prediction of protein-protein interactions by docking methods. *Curr. Op. Struc. Bio.* 2002, 12:28-85.
[9] Janin J, Henrick K, Moult J, Eyck LT, Sternberg MJ, Vajda S, Vakser I, Wodak SJ. CAPRI: a Critical Assessment of PRedicted Interactions. *Proteins* 2003, 52:2-9.
[10] Berman HM, Westbrook J, Feng Z, Gilliland G, Bhat TN, Weissig H, Shindyalov IN, Bourne PE*Nucl. Acid Res*. 28:235-242, 2000
[11] Cherfils J, Duquerroy S, Janin J. Protein-protein recognition analyzed by docking simulation *Proteins* 1991, 11:271-280.
[12] Roche M, Aze J, Kodratoff Yet al., in *ROC Analysis in Artificial Intelligence, 1st International Workshop, ROCAI-2004, Valencia, Spain, August 22, 2004*, edited by Hernandez-Orallo J, Ferri C, Lachiche N, Flach PA, 2004, p. 81.
[13] Sebag M, Azé J, Lucas N. *ROC-Based Evolutionary Learning: Application to Medical Data Mining*, 2004.
[14] Derrida B. Random energy model: an exactly solvable model of disordered systems. *Phys. Rev.* B24:2613-2626, 1981.
[15] Onuchic JN, Luthey-Schulten Z, Wolynes PG. Theory of protein folding: the energy landscape perspective. *Annu Rev Phys Chem.* 1997, 48:545-600.
[16] Lancet D, Sadovsky E, Seidemann E. Probability model for molecular recognition in biological receptor repertoires: significance to the olfactory system. *Proc. Nat. Acad. Sci. USA* 90:3715-3719, 1993
[17] Janin J. Quantifying biological specificity: the statistical mechanics of molecular recognition. *Proteins* 1996, 25:438-445 .
[18] Boissonnat JD, Cazals F, Da F, et al., in *Symposium on Computational Geometry*, p. 421, 1999.
[19] Richards FM. The interpretation of protein structures: total volume, group volume distributions and packing density. *J. Mol. Biol.* 1974, 82:1-14.
[20] Soyer A, Chomilier J, Mornon JP, Jullien R, Sadoc JF. Voronoi tessellation reveals the condensed matter character of folded proteins. *Phys Rev Lett* 2000, 85:3532-5.
[21] Angelov B, Sadoc JF, Jullien R, Soyer A, Mornon JP, Chomilier J. Nonatomic solvent-driven Voronoi tessellation of proteins: an open tool to analyze protein folds. *Proteins.* 2002, 49:446-56.
[22] Poupon A. Voronoi and Voronoi-related tessellations in studies of protein structure and interaction. *Curr. Op. Struc. Bio.* 14:233-241, 2004.
[23] Lo Conte L, Chothia C, Janin J. The Atomic Structure of Protein-Protein Recognition Sites. *J. Mol. Biol.* 1999, 285:2177-2198.
[24] Chakrabarti P, Janin J. Dissecting protein-protein recognition sites. *Proteins* 2002, 47:334-343.

[25] Fleury D, Daniels RS, Skehel JJ, Knossow M, Bizebard T. Structural evidence for recognition of a single epitope by two distinct antibodies. *Proteins*. 2000, 40:572-8.
[26] Chothia C, Janin J. Principles of protein-protein recognition. *Nature* 1975, 256:705-708.
[27] Honig B, Nicholls A. Classical electrostatics in biology and chemistry. Science. 1995 268:1144-9.
[28] Young L, Jernigan RL, Covell DG. *Protein Sci.* 3:717-729, 1994
[29] Jones S, Thornton JM. Principles of protein-protein interactions *Proc. Nat. Acad. Sci. USA* 93, 13-20, 1997.
[30] Tsai CJ, Lin SL, Wolfson HJ, Nussinov R. Studies of protein-protein interfaces: a statistical analysis of the hydrophobic effect. *Protein Sci.* 6:53-64, 1997
[31] Zhang C, Vasmatzis G, Cornette JL, DeLisi C. Determination of atomic desolvation energies from the structures of crystallized proteins.*J. Mol. Biol.* 267:707-726, 1997
[32] Miyazawa S, Jernigan RL. Residue-residue potentials with a favorable contact pair term and an unfavorable high packing density term, for simulation and threading. J Mol Biol. 1996, 256:623-44.
[33] Ofran Y, Rost B Analysing six types of protein-protein interfaces. *J Mol Biol.* 2003 325:377-87.
[34] Betts MJ, Sternberg MJ. An analysis of conformational changes on protein-protein association: implications for predictive docking. *Protein Eng*. 1999, 12:271-83.
[35] Northrup SH, Erickson HP. Kinetics of protein-protein association explained by Brownian dynamics computer simulation. *Proc. Nat. Acad. Sci. USA* 1992, 89:3338-3342.
[36] Janin J. The kinetics of protein-protein recognition. *Proteins* 1997, 28:153-161.
[37] Gabdoulline RR, Wade RC. Protein-protein association: investigation of factors influencing association rates by brownian dynamics simulations. *J Mol Biol.* 2001, 306:1139-1155.
[38] Gabdoulline RR, Wade RC. Biomolecular diffusional association. *Curr Opin Struct Biol.* 2002, 12:204-13.
[39] Selzer T, Schreiber G. New insights into the mechanism of protein-protein association. *Proteins*. 2001 45:190-8.
[40] Schreiber G. Kinetic studies of protein-protein interactions. *Curr Opin Struct Biol.* 2002 12:41-7.
[41] Camacho CJ, Kimura SR, DeLisi C, Vajda S. Kinetics of desolvation-mediated protein-protein binding. *Biophys J.* 2000, 78:1094-105.
[42] Schlossauer M, Baker D. Realistic protein-protein association rates from a simple diffusional model neglecting long-range interactions, free energy barriers and landscape ruggedness. *Protein Sci.* 2004, 13 :1660-1669.
[43] Camacho CJ, Vajda S. Protein docking along smooth association pathways. *Proc Natl Acad Sci USA*. 2001, 98:10636-41.
[44] Camacho CJ, Vajda S. Protein-protein association kinetics and protein docking. *Curr Opin Struct Biol.* 2002, 12:36-40.
[45] Kumar S, Ma B, Tsai CJ, Sinha N, Nussinov R. Folding and binding cascades: dynamic landscapes and population shifts. *Protein Sci.* 2000 9:10-9.
[46] Zhang C, Chen J, DeLisi C. Protein-protein recognition: exploring the energy funnels near the binding sites. Proteins. 1999, 34:255-67.
[47] Rajamani D, Thiel S, Vajda S, Camacho CJ. Anchor residues in protein-protein interactions. *Proc Natl Acad Sci USA*. 2004 101:11287-92.
[48] Rodier F, Janin J. Protein-protein interaction at crystal contacts. *Proteins* 1995, 23:580-587.
[49] Bahadur RP, Chakrabarti P, Rodier F, Janin J. A dissection of specific and non-specific protein-protein interfaces. *J. Mol. Biol.* 2004, 336:943-955.
[50] Jones S, Thornton JM. Analysis of protein-protein interaction sites using surface patches.*J Mol Biol.* 1997 272:121-32.
[51] Ponstingl H, Henrick K, Thornton JM. Discriminating between homodimeric and monomeric proteins in the crystalline state. *Proteins.* 2000, 41:47-57.
[52] Mintseris J, Weng Z. Atomic contact vectors in protein-protein recognition. *Proteins.* 2003, 53:629-39.