Université de Sherbrooke

Développement, validation et nouvelles applications d'un modèle d'analyse des modes normaux basé sur la séquence et la structure de protéines

Par
Vincent Frappier
Programme de biochimie

Thèse présentée à la Faculté de médecine et des sciences de la santé
en vue de l'obtention du grade de philosophiae doctor (Ph.D.)
en biochimie

Sherbrooke, Québec, Canada
Mai, 2016

Membres du jury d'évaluation
Pr. Rafael Najmanovich, Programme de Biochimie
Pr. Pierre Lavigne, Programme de Biochimie
Pr. Jean-Guy Lehoux, Programme de Biochimie
Pr. Michelle Scott, Programme de Biochimie
Pr. Jean-Bernard Denault, Programme de Pharmacologie
Pr. Roberto A. Chica, Chemistry and Biomolecular Sciences, University of Ottawa

À mes parents, ma famille et Martine pour m'avoir encouragé et supporté

Nobody ever figures out what life is all about, and it doesn't matter. Explore the world. Nearly everything is really interesting if you go into it deeply enough
-Richard Feynman

Study hard what interests you the most in the most undisciplined, irreverent and original manner possible.
-Richard Feynman

What I cannot create, I do not understand.
- Richard Feynman

Certainly with our present knowledge, the person who attempts to estimate dihedral angles to an accuracy of one or two degrees does so at his own peril
- Martin Karplus

RÉSUMÉ

Développement, validation et nouvelles applications d'un modèle d'analyse des modes normaux basé sur la séquence et la structure de protéines

Par
Vincent Frappier
Programmes de biochimie

Thèse présentée à la Faculté de médecine et des sciences de la santé en vue de l'obtention du diplôme de philosophiae doctor (Ph.D.) en biochimie, Faculté de médecine et des sciences de la santé, Université de Sherbrooke, Sherbrooke, Québec, Canada, J1H 5N4

Les protéines existent sous différents états fonctionnels régulés de façon précise par leur environnement afin de maintenir l'homéostasie de la cellule et de l'organisme vivant. La prévalence de ces états protéiques est dictée par leur énergie libre de Gibbs alors que la vitesse de transition entre ces états biologiquement pertinents est déterminée par le paysage d'énergie libre. Ces paramètres sont particulièrement intéressants dans un contexte thérapeutique et biotechnologique, où leur perturbation par la modulation de la séquence protéique par des mutations affecte leur fonction. Bien que des nouvelles approches expérimentales permettent d'étudier l'effet de mutations en haut débit pour une protéine, ces méthodes sont laborieuses et ne couvrent qu'une fraction de l'ensemble des structures primaires d'intérêt. L'utilisation de modèles bio-informatiques permet de tester et générer *in silico* différentes hypothèses afin d'orienter les approches expérimentales. Cependant, ces méthodes basées sur la structure se concentrent principalement sur la prédiction de l'enthalpie d'un état, alors que plusieurs évidences expérimentales ont démontré l'importance de la contribution de l'entropie. De plus, ces approches ignorent l'importance de l'espace conformationnel protéique dicté par le paysage énergétique cruciale à son fonctionnement. Une analyse des modes normaux peut être effectuée afin d'explorer cet espace par l'approximation que la protéine est dans une conformation d'équilibre où chaque acide aminé est représenté par une masse régie par un potentiel harmonique. Les approches actuelles ignorent l'identité des résidus et ne peuvent prédire l'effet de mutations sur les propriétés dynamiques. Nous avons développé un nouveau modèle appelé ENCoM qui pallie à cette lacune en intégrant de l'information physique et spécifique sur les contacts entre les atomes des chaînes latérales. Cet ajout permet une meilleure description de changements conformationnels d'enzymes, la prédiction de l'effet d'une mutation allostérique dans la protéine DHFR et également la prédiction de l'effet de mutations sur la stabilité protéique par une valeur entropique. Comparativement à des approches spécifiquement développées pour cette application, ENCoM est plus constant et prédit mieux l'effet de mutations stabilisantes. Notre approche a également été en mesure de capturer la pression évolutive qui confère aux protéines d'organismes thermophiles une thermorésistance accrue.

Mots clés : Protéine, structure, dynamique conformationelle, analyse des modes normaux, mutation, stabilité protéique, entropie

TABLE DES MATIERES

LISTE DES FIGURES

LISTE DES ABRÉVIATIONS

AMN	Analyse des modes normaux/*Normal Mode Analysis*
ANM	*Anisostropic Network Model*
ENCoM	*Elastic Network Contact Model*
STeM	*Spring Generalised Tensor Model*
RMN	Résonance Magnétique Nucléaire

INTRODUCTION

La régulation protéique

Les protéines jouent un rôle essentiel dans les processus biologiques responsables de l'homéostasie de la cellule et des êtres vivants. Pourtant, elles ne sont que des polymères constitués de 20 acides aminés qui remplissent des rôles très diversifiés. Par exemple, le consortium « Gene Ontoloy » a recensé plus de 40 000 fonctions différentes (The Gene Ontology Consortium, 2000) qui peuvent être impliquées autant dans des rôles de métabolisme de petites molécules, que de signalisation cellulaire ou bien du maintien de l'intégrité cellulaire. Cette impressionnante polyvalence provient de la capacité des protéines à exister sous différents états (ou formes) qui à leur tour remplissent différentes fonctions (Klotz et Klotz, 1955). La définition des états dépend du contexte dans lequel un système biologique est étudié et de la résolution des approches expérimentales. Par exemple, la protéine de signalisation calmoduline peut exister sous plusieurs formes dont la représentation la plus simple serait une forme dénaturée et une forme native. Cependant, cette dernière peut être décomposée en deux autres états essentiels à son mécanisme d'action soit la formes fermée et et la forme ouverte (Gopalakrishna et Anderson, 1982). À son tour, la forme ouverte peut-être retrouvée sous plusieurs états, une forme liant du calcium, une forme qui interagit avec un partenaire protéique ou une combinaison des deux. Ainsi, la protéine calmoduline peut exister sous sept formes différentes caractérisées expérimentalement. Il existe probablement encore plus d'états, cependant les approches expérimentales actuelles ne peuvent les détecter ou bien ces états ne sont pas pertinents dans le contexte étudié. Tous ces différents états sont macroscopiques et ils ne représentent que les propriétés moyennes d'une population de molécules. De fait, un état macroscopique est constitué d'un continuum d'états microscopiques qui ne peuvent être définis par une forme distincte. À l'équilibre, les propriétés de l'état macroscopique sont stables alors que les états microscopiques sont en perpétuels changement. Par exemple, une solution d'eau est composée d'un ensemble de molécules qui possèdent chacune une énergie cinétique qui leur est propre alors que la température de la solution représente l'énergie cinétique moyenne de toutes ces molécules. Ainsi, la répartition de l'énergie cinétique dans chaque

molécule représente un état microscopique du système et la température une propriété de l'état macroscopique.

Les proportions des différents états de protéines dans une population sont régulées par leur environnement. Dans l'exemple précédent de la calmoduline, la fonction de cette protéine est de détecter la présence de calcium dans le cytosol de la cellule et de transmettre cette détection à d'autres protéines par des interactions protéine-protéine afin de moduler leur fonction. En l'absence de calcium, la forme native fermée de la calmoduline est prédominante et ne peut pas reconnaître ces partenaires protéiques. Une augmentation de la concentration de calcium dans la cellule va induire une augmentation de la proportion de la forme ouverte au détriment de la forme fermée par une liaison du calcium. Ce changement d'état va permettre à la calmoduline d'interagir avec ses différents partenaires d'interaction, transmettre la détection de calcium et remplir sa fonction. Ainsi, pour s'assurer du bon fonctionnement d'un être vivant, les différentes proportions des états protéiques doivent être régulées de façon précise en fonction du contexte cellulaire, afin que les bonnes fonctions soient exécutées au bon moment au bon endroit.

Fondement thermodynamique

D'un point de vue microscopique, le principe fondamental qui régit les proportions qu'une population de molécules va adopter est dicté par la loi de Maxwell-Boltzmann :

$$P(E_i) = e^{-e_i/k_B T}/Z$$

où P représente la population d'un état E_i d'énergie e_i, k_B la constante de Boltzmann, T la température absolue en Kelvin et Z la fonction de répartition sur tous les N états microscopiques:

$$Z = \sum_{i=1}^{N} e^{-e_i/k_B T}$$

Cette loi énonce que, à une température donnée, plus un état possède une énergie libre molaire favorable (ou basse) par rapport aux autres états, plus il sera fréquent dans une population à l'équilibre et que plus la température augmente, plus des états non favorables seront probables au détriment des états favorables. Cependant, par définition de cette équation, l'état favorable sera toujours plus probable qu'un état moins favorable. Ainsi, les valeurs d'énergie d'un état microscopique sont indépendantes de la température du système. D'un point de vue macroscopique, cet équilibre peut être représenté par la différence d'énergie libre de Gibbs (ΔG) entre deux états (A et B):

$$G_A - G_B = \Delta G = -RTlnK = -RTln(P_A/P_B)$$

Où R représente la constante des gaz parfaits, T la température absolue en Kelvin et K la constante d'équilibre ou le ratio des populations des deux états, P_A et P_B. Ainsi, lorsque la différence d'énergie est nulle, la proportion de chacun des états macroscopiques est égale :

$$K = e^{-\frac{\Delta G}{RT}}$$

$$K = e^{-\frac{0}{RT}} = 1$$

La différence d'énergie libre de Gibbs peut aussi être décrite par la différence d'enthalpie (ΔH) et d'entropie (ΔS) entre ces deux états:

$$\Delta G = \Delta H - T\Delta S = (H_A - H_B) - T(S_A - S_B)$$

L'enthalpie représente l'énergie nécessaire pour créer un système. Elle est la somme de l'énergie interne (U) et du produit de la pression (p) et du volume (V) d'un système :

$$H = U + pV$$

D'un point de vue microscopique, l'énergie interne est représentée par la somme des produits de l'énergie de chaque état (e_i) et de leur proportion ($P(E_i)$):

$$U = \sum_{i=1}^{N} P(E_i) e_i$$

Pour des systèmes biologiques, l'enthalpie est principalement représentée par les interactions entres les atomes : les liens covalents, les ponts hydrogène, les interactions électrostatiques et les forces d'interaction électrodynamiques approximées par l'énergie potentielle de Lennard-Jones. L'entropie quant à elle représente le désordre d'un système. Plus elle est élevée, plus le système est désordonné, c'est-à-dire plus le système occupe différents états distincts. Microscopiquement l'entropie (S) est décrite par les probabilités d'observer le système dans un état d'énergie E_i et la constante des gaz parfaits (R) :

$$S = -R \sum_{i=1}^{N} P(E_i) \ln P(E_i)$$

Pour un système biologique, l'entropie est composée des degrés de liberté et de mouvement d'un système. Par exemple les différentes conformations qu'une protéine peut adopter dans son état macroscopique (entropie conformationnelle) et les degrés de liberté de l'eau.

La différence d'énergie libre de Gibbs varie en fonction de la température. En effet, l'enthalpie et l'entropie d'un état macroscopique augmentent avec une variation de température (ΔT) selon la capacité calorifique à pression constante (C_P) :

$$\Delta S = C_P \frac{\Delta T}{T}$$

$$\Delta H = C_P \Delta T$$

La capacité calorifique à pression constante peut également varier en fonction de la température, cependant, dans un contexte biologique, elle est souvent considérée comme étant constante sur les intervalles de températures étudiées. La relation entre la capacité calorifique, l'entropie et l'enthalpie peut également être expliquée d'un point de vue microscopique. En effet, plus la température augmente, plus des états de hautes énergies deviennent probables, ce qui augmente l'énergie interne moyenne du système et par le fait même l'enthalpie. Également, le peuplement de ces états de hautes énergies augmente l'entropie du système en augmentant le nombre d'états observés.

Ainsi, plus un système est en mesure d'emmagasiner de l'énergie sous la forme de chaleur, plus il aura une capacité calorifique importante. D'un point de vue macroscopique, cette énergie est emmagasinée dans ces degrés de liberté, alors que d'un point de vue microscopique, elle est emmagasinée dans la dégénérescence des états. Pour des substances liquides ou à l'état gazeux, les principales sources de degré de liberté sont les mouvements de translation et de rotation des molécules alors que pour les substances solides, il s'agirait principalement des mouvements vibrationnels. En général, les liquides ont des capacités calorifiques plus élevées que les solides. Ainsi, une même substance possède des capacités calorifiques distinctes pour chacun de ses états. Par exemple, la glace possède une capacité calorifique deux fois plus faible que l'eau; malgré le fait que la glace peut emmagasiner de l'énergie vibrationnelle, l'eau peut emmagasiner d'avantage de chaleur dans l'énergie cinétique de ses particules.

Les propriétés thermodynamiques d'un système sont décomposables de façon additive en plusieurs sous-systèmes:

$$\Delta G = \Delta G_A + \Delta G_B = \Delta H_A + \Delta H_B - T(\Delta S_A + \Delta S_B)$$

Où A et B représentent deux sous-systèmes indépendants qui possèdent leurs propres entropie, enthalpie et capacité calorifique.

Cinétique et équilibre

Toutes ces propriétés thermodynamiques sont valides pour un système à l'équilibre, c'est-à-dire que macroscopiquement les propriétés du système sont constantes, alors que d'un point de vue microscopique les molécules sont constamment en train de changer d'état. Ces changements s'effectuent sur un paysage énergétique qui représente un continuum d'états microscopiques. Comme ce paysage est multidimensionnel, il existe alors différents chemins qui peuvent relier deux formes. Les chemins empruntés doivent généralement traverser une ou plusieurs barrières énergétiques qui dictent alors la vitesse de la transition entre ces deux états. D'un point de vue microscopique, en se référant à l'équation d'Arrhenius, cette vitesse (k) est dépendante de l'énergie de l'état de transition (e_a) et de la température du système (T) (Arrhenius, 1889; Laidler, 1984) :

$$k = e^{-e_a/(k_B T)}$$

Ainsi, plus l'énergie de l'état de transition est élevée et la température basse, plus la vitesse de transition sera lente, car l'état de transition sera moins probable. Macroscopiquement, cette réaction peut être interprétée selon la différence d'énergie libre de Gibbs de l'état de transition avec l'état initial (ΔG_a) selon :

$$k = e^{-\Delta G_a/(RT)}$$

Un catalyseur (tel qu'une enzyme) aura pour effet de diminuer cette énergie d'activation et ainsi accélérer un processus chimique.

Il est également possible d'emprisonner une population dans un puit énergétique en refroidissant rapidement un échantillon par cryogénie (Doster, 2010). Ce refroidissement devrait en temps normal diminuer la proportion d'états de haute énergie vers des états de basses énergies. Cependant, en raison de la basse température, l'état de transition devient très peu probable et le système prendra énormément de temps pour atteindre l'équilibre. Par exemple, à pression et température normales les diamants se transforment spontanément en

graphite, car cette forme est plus stable. Cependant, l'énergie de transition est extrêmement élevée et requiert une réorganisation complexe de liens covalents. Ainsi, la transition est lente, voire pratiquement nulle. C'est également le cas avec les formes agrégées de protéines qui sont généralement plus stables que les formes repliées. Par contre, la cinétique d'agrégation est longue et sur les périodes d'étude des phénomènes thermodynamiques on peut parfois les ignorer (Fowler *et al.*, 2005).

Lors de l'étude de systèmes biologiques, la forme des paysages énergétiques est aussi (sinon plus) importante que les énergies des états macroscopiques. Cette importance est entre autres exposée dans le paradoxe de Levinthal (Levinthal, 1969) où il est estimé que pour une protéine de 100 acides aminés, en ne considérant que les différentes conformations du squelette peptidique, il existe plus de 10^{300} conformations possibles et qu'il n'y a probablement pas eu suffisamment de temps dans l'univers pour toutes les explorer. Pour obtenir un repliement dans un temps pertinent au contexte biologique, il faut alors une configuration de surface qui guide le repliement de la protéine de façon cinétique à partir de la forme dénaturée (Dill et Chan, 1997). Également, la transition entre les différents états macroscopiques protéiques fonctionnels doit traverser une barrière énergétique qui dicte la vitesse de transition. Si cette barrière est trop importante, la réponse à l'environnement serait trop lente et mènerait probablement à une perte de fonction de la protéine. Par exemple, les enzymes doivent catalyser des réactions essentielles au métabolisme d'une cellule en alternant entre différentes formes fonctionnelles. Des états actifs et inactifs trop stables augmenteraient la différence d'énergie et ralentiraient la réaction à des niveaux non fonctionnels (Kiss *et al.*, 2010).

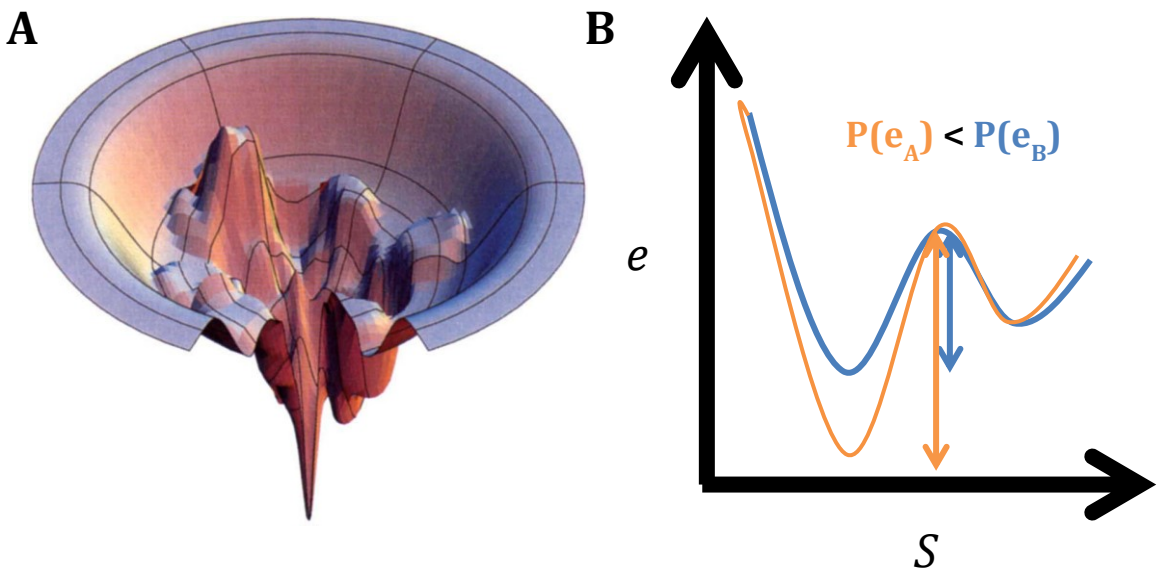


Figure 1.1 – Paysage d'énergie libre
A) Représentation du paysage d'énergie libre de repliement d'une protéine composée d'un gradient et de plusieurs minimums locaux. Il existe plusieurs chemins qui mènent à la forme native située dans le minimum global (Dill et Chan, 1997). B) Représentation des coordonnées de transition entre deux états macroscopiques pour deux protéines hypothétiques *A* (orange) et *B* (bleu). Une stabilisation de la forme *A* réduit la probabilité d'observer l'état de transition et diminue la vitesse de transition entre les deux états.

Repliement des protéines

Pour effectuer sa fonction primaire, une protéine doit généralement être dans sa forme repliée, ou dite native (Perry et Wetzel, 1984). Cette forme est caractérisée par des conformations compactes et ordonnées qui possèdent des structures secondaires distinctes telles que des hélices α et des feuillets β (Gao *et al.*, 2005; Kumar *et al.*, 2002; Fontana *et al.*, 1997). Les formes natives possèdent des caractéristiques matérielles et thermodynamiques qui s'apparentent plus aux solides qu'aux substances liquides (Cooper, 2000; Sturtevant, 1977; Cooper, 2005; Leitner, 2008; Fultz, 2010; Vitkup *et al.*, 2000). En effet, elles ont des capacités calorifiques semblables et conduisent la chaleur de façon similaire (Foley *et al.*, 2014). Plusieurs études ont démontré que les protéines se comportent de la même façon que de la matière condensée (Doster, 2010; Bée, 1988)

D'un point de vue thermodynamique, la différence d'énergie libre entre la forme repliée et dénaturée (ΔG_D) est représentée selon :

$$\Delta G_D = \Delta H_D - T\Delta S_D$$

Où ΔH_D et ΔS_D représentent la différence d'enthalpie et d'entropie, respectivement, entre la forme native et la forme dénaturée à une température T. Ainsi, lorsque les populations des deux formes sont équivalentes, à la température de dénaturation (T_M), cette équation est représentée par :

$$\Delta H_D = T_M \Delta S_D$$

En incluant l'effet de la différence de capacité calorifique (ΔC_P) entre la forme native et la forme dénaturée dans cette relation, nous obtenons l'équation de Gibbs–Helmholtz:

$$\Delta G_D(T) = \Delta H_D T_M (1 - \frac{T}{T_M}) - \Delta C_P (T_M - T + T \ln T/T_M)$$

Les protéines dénaturées ont plus de degrés de liberté que leur forme native. Ainsi la différence de capacité calorifique est positive. L'équation de Gibbs–Helmholtz représente alors une cloche inversée qui possède une température de dénaturation à la chaleur (processus endothermique) et au froid (processus exothermique) ainsi qu'une température à stabilité maximale (T_S). La différence de capacité calorifique module la courbature de façon négative, c'est-à-dire que plus elle sera élevée, plus petit sera l'intervalle de température où la protéine est majoritairement repliée.

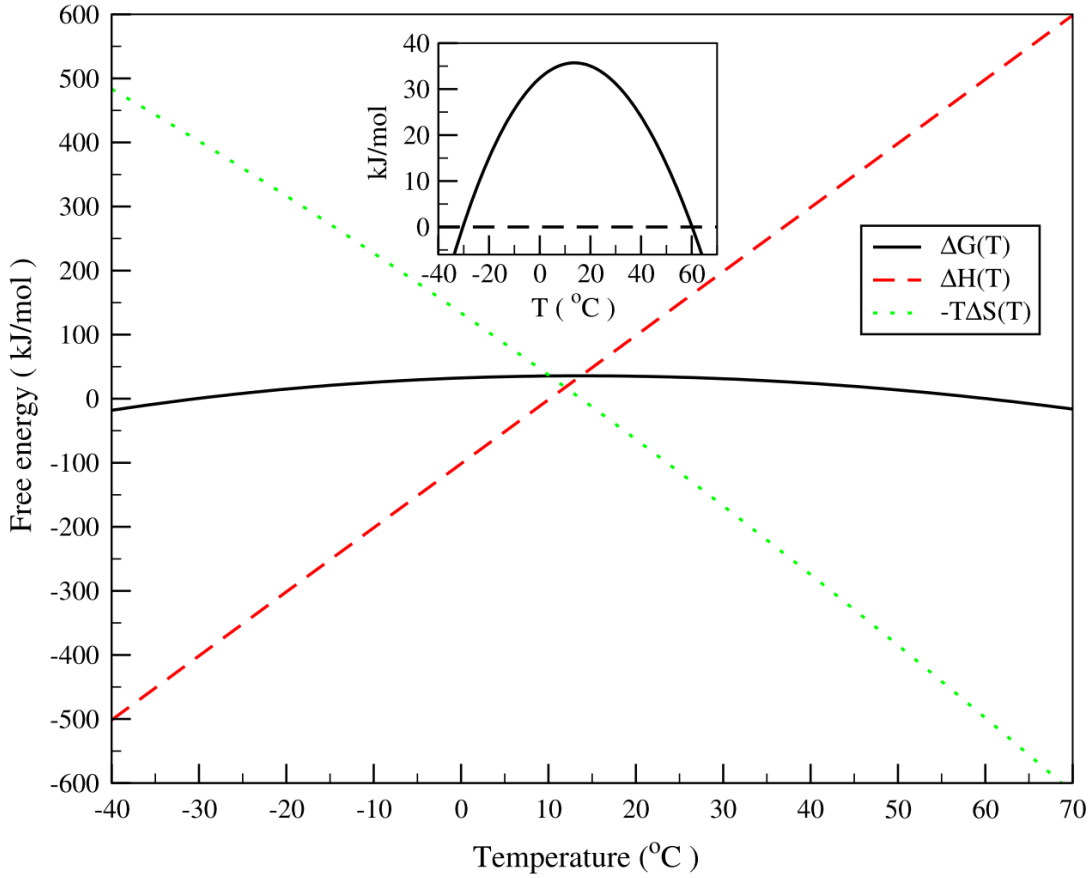


Figure 1.2 – Courbe de la stabilité protéique

L'énergie libre de Gibbs de la forme repliée varie en fonction de la température selon l'équation de Gibbs-Helmholtz. La différence d'enthalpie (rouge) et d'entropie (vert) confère une stabilité marginale à la protéine (noir). La courbature de cette courbe est dépendante de la différence de capacité calorifique, la protéine possède une énergie maximale et deux températures de dénaturation. Figure tirée de Dias *et al*. (2010)

Thermodynamiquement, le repliement des protéines est caractérisé par un gain entropique de l'eau (Tanford, 1997), une perte d'entropie conformationnelle de la protéine (He *et al.*, 2003), un gain enthalpique pour les interactions intra-protéine (Porter et Rose, 2011) et une perte enthalpique pour les interactions entre les atomes de la protéine et de l'eau (Bennion et Daggett, 2003).

Les protéines possèdent des atomes hydrophobes qui lorsqu'ils sont exposés au solvant, doivent être hydratées par des molécules d'eau qui s'organisent sous la forme de clathrates

(Némethy et Scheraga, 1962; Hummer *et al.*, 1998). Ces structures entraînent un gain enthalpique par la formation de ponts hydrogènes entre les molécules d'eau qui se font au détriment d'une perte entropique plus importante liée à la « solidification » de ces dernières. En se repliant, la protéine minimise l'exposition de ses surfaces apolaires au solvant par la création d'un cœur hydrophobe et maximise ainsi l'entropie de l'eau par la destruction des clathrates (Spolar *et al.*, 1989). Ces forces hydrophobes stabilisantes diminuent avec la température (Privalov, 1989). Ainsi, lors de la dénaturation protéique par le froid, l'entropie de l'eau ne peut surpasser le gain enthalpique et l'hydratation des résidus hydrophobes devient favorable entraînant une perte du cœur hydrophobe et de la structure de la protéine (Yang *et al.*, 2014; Dias, 2012; Dias *et al.*, 2010). Autrement dit, la formation de clathrates devient favorable à basse température. L'ajout de solvant organique diminue également les forces hydrophobes et cause la dénaturation des protéines (Vajpai *et al.*, 2013; Asakura *et al.*, 1978; Liang *et al.*, 2004) par une solubilisation des cœurs hydrophobes à la place des clathrates. La dénaturation de protéines par des agents chaotropes (p. ex., urée) met également en évidence les forces hydrophobes par une interaction directe de ces agents avec le cœur hydrophobe de la protéine et indirectement en perturbant les propriétés du solvant (Bennion et Daggett, 2003). L'agent chaotrope est alors en mesure de solubiliser les résidus hydrophobes à la place de l'eau, contrecarrant l'effet déstabilisant de la formation des clathrates qui ont également un effet entropique moins déstabilisant, car les agents chaotropes ont tendance à diminuer l'entropie de l'eau (Ball et Hallsworth, 2015). Les atomes hydrophiles des protéines perturbent également les propriétés de l'eau, mais de façon moins défavorable que les atomes hydrophobes. En effet, il est estimé que la coquille d'hydratation d'une protéine peut aller jusqu'à 10 Ångstrom (Ebbinghaus *et al.*, 2007) et l'eau présente dans cette coquille possède des caractéristiques distinctes par rapport à de l'eau pure.

Bien que les protéines dans leur forme repliée demeurent des entités dynamiques, leur forme dénaturée possède beaucoup plus de degrés de liberté que ces dernières. Ainsi, le repliement des protéines diminue leur entropie en restreignant le nombre de conformations accessibles. Cette perte entropique est plus importante pour le squelette peptidique que pour les chaînes latérales (Baxa *et al.*, 2014) qui peuvent continuer à explorer différents états

énergétiques similaires en adoptant différentes conformations (Syme *et al.*, 2010; Yu *et al.*, 1999; Creamer et Rose, 1992; Khechinashvili *et al.*, 2014). Les degrés de liberté accessibles au squelette peptidique sont représentés par deux angles dièdres qui sont généralement restreints à un espace représenté par le diagramme de Ramachandran. Cet espace peut être décomposé en deux principaux puits énergétiques qui représentent les brins β et les hélices α. Ces structures secondaires tendent à minimiser les encombrements stériques (Ramachandran *et al.*, 1963; Porter et Rose, 2011) et maximiser l'enthalpie par la formation de ponts hydrogènes au niveau du squelette peptidique, principalement dans les régions enfouies (Deechongkit, Nguyen, *et al.*, 2004; Porter et Rose, 2011; J. Gao *et al.*, 2009; Deechongkit, Dawson, *et al.*, 2004). Ces interactions représentent 66% de l'enthalpie intra-protéine (Baker et Hubbard, 1984).

L'ampleur des forces stabilisatrices enthalpiques responsables du repliement de protéines ne fait toujours pas consensus. Certaines approches ont démontré que la forme dénaturée crée en fait plus d'interactions enthalpiques avec le solvant que les interactions intra-protéine de la forme native (Matysiak *et al.*, 2012; Bennion et Daggett, 2003) et que c'est le fait de ne pas les créer qui serait déstabilisant pour cette dernière. Au niveau des chaînes latérales, un des rôles de ses interactions polaires serait de restreindre les conformations accessibles, car il n'existe qu'un ensemble de conformations qui peut satisfaire à ces interactions géométriquement très sensibles (Donald *et al.*, 2011). À haute température, la protéine se dénature par la perte de ses interactions enthalpiques et la forme dénaturée devient plus probable étant donné qu'elle possède une entropie plus élevée. Ainsi, la stabilité de la forme native peut être augmentée en diminuant l'entropie de la forme dénaturée par des liaisons covalentes disulfures (Vaz *et al.*, 2006; Perry et Wetzel, 1984; Arolas *et al.*, 2006; Zavodszky *et al.*, 2001) et par des charges électrostatiques similaires à proximité dans la structure primaire (Hendsch et Tidor, 1994; Xiao *et al.*, 2013).

La somme des forces responsables du repliement de protéines ne leur donne qu'une stabilité marginale. En effet, elle n'est que de 5 à 10 kcal/mol, ce qui représente l'équivalent de la formation d'un pont hydrogène (Privalov, 1989). Cependant, cette stabilité est très résistante aux perturbations, car toutes ces forces sont inter reliées. Par

exemple, la création d'une nouvelle interaction polaire pourrait diminuer l'enthalpie de la protéine, mais restreindre ses degrés de liberté et ainsi diminuer son entropie conformationnelle (Dunitz, 1995; Liu *et al.*, 2000; Cooper *et al.*, 2001). Ce phénomène de compensation enthalpie-entropie a par exemple été maintes fois observé p des interactions protéine-protéine. Ainsi, lors d'études de perturbation de stabilité protéique, il est important d'analyser l'ensemble des propriétés thermodynamiques du système afin de mieux comprendre l'effet des changements pour éventuellement mieux les moduler et les prédire.

Thermophiles

Les organismes thermophiles sont des êtres vivants qui vivent à des températures élevées pouvant aller jusqu'à 100 ℃. À cette tempérautre, les protéines des organismes mésophiles sont dénaturées et non-fonctionnelles. Pour s'adapter à ces environnements et augmenter leur thermorésistance, les principales stratégies thermodynamiques adoptées par ces organismes sont par ordre d'importance: augmenter leur stabilité maximale ($\uparrow \Delta G$), diminuer leur différence de capacité calorifique ($\downarrow \Delta C_P$) et augmenter leur température de stabilité maximale ($\uparrow T_S$) (Razvi et Scholtz, 2006). Ces effets sur la courbe de stabilité protéique sont démontrés dans la figure 1.3. Les mécanismes exacts qui confèrent ces propriétés aux thermophiles ne sont toujours pas bien compris et peu de modèles sont en mesure de prédire effectivement leurs propriétés thermodynamiques (Pucci et Rooman, 2015). D'un point de vue structurel et génomique, il a été observé que par rapport à leur homologue mésophile, les protéines thermophiles possèdent plus de résidus polaires en surface (Cambillau et Claverie, 2000), des cœurs hydrophobes plus compacts (Glyakina *et al.*, 2007), plus de ponts disulfures et de ponts salins (Beeby *et al.*, 2005) et à température égale, elles sont généralement plus rigides (Mamonova *et al.*, 2010). L'exposition de résidus polaires en surface a pour effet de diminuer la différence de capacité calorifique (Matthews, 1993; Myers *et al.*, 1995), mais l'importance de la différence observée entre les mésophiles et thermophiles ne peut pas être simplement expliquée par ce mécanisme. En effet, Robertson et Murphy (1997) ont démontré qu'en réalité, cette relation était principalement dépendante du nombre de résidus présents dans la protéine (Szilagyi et Zavodszky, 2000) et qu'elle ne peut décrire efficacement les différences de ΔC_P entre deux protéines de même taille, mais qui possèdent des séquences différentes (Uchiyama *et al.*,

2002; Loladze *et al.*, 2001). La création de ponts disulfures et de ponts salins aurait également pour effet de diminuer la différence de capacité calorifique par une rigidification de la forme dénaturée (Zavodszky *et al.*, 2001; Myers *et al.*, 1995).

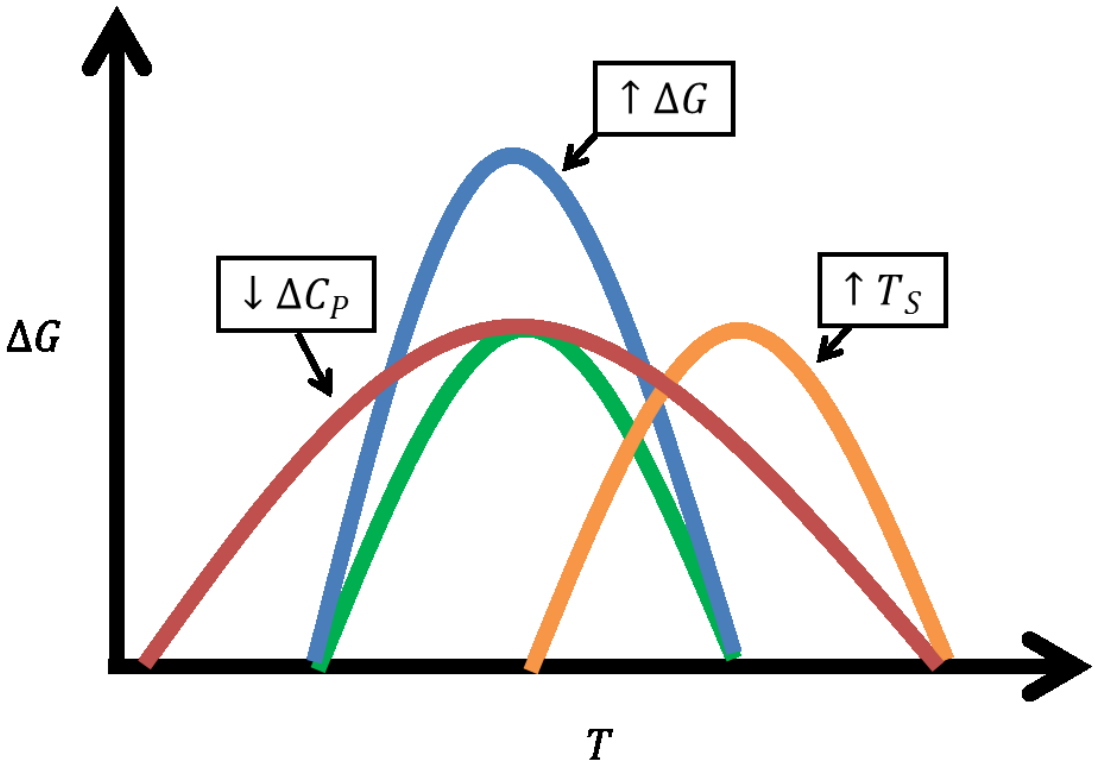


Figure 1.3 - Effet de la modulation des paramètres thermodynamiques sur la courbe de stabilité protéique
La résistance accrue à la chaleur des protéines thermophiles par rapport à leurs homologues mésophiles (vert) est principalement obtenue par une combinaison de la modulation de trois paramètres thermodynamiques : augmentation de la stabilité maximale (bleu), une diminution de la différence de capacité calorifique (rouge) et une augmentation de la température de stabilité maximale (orange).

Les formes natives des structures thermophiles sont en moyenne plus rigides et ce gain de rigidité devrait alors diminuer la différence de capacité calorifique. Cependant, de façon contre-intuitive, une augmentation de cette différence est observée expérimentalement. Deux hypothèses sont avancées pour expliquer ce phénomène, soit que l'effet de l'eau à haute température sur la capacité calorifique est sous-estimé (Zhou, 2002) ou que la forme

dénaturée conserve des structures résiduelles (Robic *et al.*, 2003). Cette rigidité accrue a également un impact sur la fonction enzymatique des protéines thermophiles en diminuant leur vitesse de réaction à température ambiante comparativement à leur homologue mésophile. Ces résultats sont en accord avec des approches de dynamique moléculaire (Wolf-Watz *et al.*, 2004; Elias *et al.*, 2014) et de résonance magnétique nucléaire (Gagné *et al.*, 2015) qui ont démontré le lien entre la vitesse de réaction et la dynamique de la protéine, suggérant ainsi que les protéines thermophiles sont plus rigides et que leur espace conformationnel est plus restreint à température égale.

Les mutations

Les propriétés thermodynamiques des protéines sont modulées par des mutations qui peuvent affecter leur fonction par la modification de la stabilité thermodynamique, l'efficacité enzymatique ou l'affinité pour des partenaires d'interactions (protéiques ou petites molécules). Ces modifications sont alors d'un grand intérêt pour mieux comprendre le fonctionnement des organismes vivants et éventuellement utiliser cette information pour les moduler dans leur application à un contexte biotechnologique.

Le « 1000 Genomes Project » a identifié plus de 500 000 mutations dans les régions codantes de protéines (McVean *et al.*, 2012), ce qui représente en moyenne 24 000 mutations par individu. Ces mutations sont responsables en partie de la diversité des êtres humains, mais aussi de leur prédisposition à certaines maladies (Sherry *et al.*, 2001). Par exemple, des mutations dans les gènes BRCA1 et BRCA2 peuvent augmenter le risque de développer un cancer du sein ou des ovaires (King *et al.*, 2003). Aussi, des mutations dans le gène suppresseur de tumeur p53 peuvent abolir son activité, moduler son interaction avec d'autres partenaires protéiques et ainsi permettre à un cancer de progresser plus facilement vers des formes plus agressives (Muller et Vousden, 2013). Certaines mutations dans des récepteurs transmembranaires couplés aux protéines G peuvent les activer de façon constitutive ou, inversement, abolir leur fonction (Parnot *et al.*, 2000; Feng *et al.*, 1998; Gether *et al.*, 1997; Barak *et al.*, 2001). Ces modulations peuvent être associées à certains types de cancers (O'Hayre *et al.*, 2013). Des mutations ont été identifiées pour favoriser l'agrégation de protéines et favoriser leur progression dans cetaines maladies dégénératives,

par la diminution de l'énergie nécessaire à la transition de la forme native vers la forme aggrégée. (Münch *et al.*, 2011; De Baets *et al.*, 2015). En général, les mutations impliquées dans certaines maladies brisent l'équilibre entre les formes fonctionnelles et les formes non fonctionnelles et dérégulent l'homéostasie de la cellule (Wettstein *et al.*, 2014).

Cependant, les mutations ne sont pas toutes délétères; elles peuvent être neutres en ayant peu d'impact sur la fonction de la protéine. En effet, certaines enzymes homologues peuvent avoir jusqu'à 60% de leur séquence mutée et conserver leur fonction (Tian et Skolnick, 2003). Cette robustesse provient entre autres du fait que certains types d'acides aminés ont des propriétés physicochimiques similaires (Kawashima *et al.*, 1999) et peuvent être substitués sans nécessairement avoir un effet drastique sur les propriétés de la structure. En effet, des alignements de séquences multiples de gènes homologues ont permis de dériver des matrices de substitutions basées sur la fréquence d'acides aminés à une position donnée. Il a été observé que certaines substitutions étaient plus fréquentes que d'autres et qu'elles corrélaient avec la similarité des propriétés des acides aminés impliqués dans la substitution (Henikoff et Henikoff, 1992). Cette résistance aux mutations provient également du fait que seulement une fraction des résidus est responsable du maintien de l'intégrité de la structure, des régions essentielles à la fonction et de la conservation des propriétés dynamiques (McLaughlin *et al.*, 2012; Ramanathan et Agarwal, 2011). Cette fraction peut être identifiée par la comparaison de séquences de protéines homologues et l'identification de résidus conservés ou de paires de résidus en coévolution (Morcos *et al.*, 2011; Hopf *et al.*, 2014). Ces tendances nous informent également sur les forces thermodynamiques qui guident l'évolution des protéines en explorant un ensemble de séquences liées à une fonction (Beadle et Shoichet, 2002; Marsh et Teichmann, 2014).

Certaines mutations peuvent être bénéfiques pour un organisme en lui conférant un avantage qui dépend du contexte dans lequel elles se produisent. Par exemple, une mutation qui augmente la vitesse d'une réaction catalisée par une enzyme pourrait également la déstabiliser. Cependant, si l'enzyme était déjà suffisamment stable, cette mutation est bénéfique alors qu'à l'inverse, si la protéine était peu stable, cette mutation pourrait faire en sorte que l'enzyme ne se replie plus et ne puisse plus effectuer sa fonction (Harms et

Thornton, 2013). À long terme ces mutations peuvent mener à l'apparition de nouvelles fonctions (Furnham *et al.*, 2015) alors qu'à court terme, elles peuvent conférer des résistances à certaines maladies ou certains environnements (Ruwende *et al.*, 1995). D'un point de vue thermodynamique, l'acquisition de nouvelles fonctions est généralement faite par un mécanisme d'échange stabilité-fonction. En effet, les résidus fonctionnels ou catalytiques ne sont généralement pas optimisés pour stabiliser la protéine et déstabilisent la forme repliée. Ainsi, les résidus non fonctionnels doivent compenser pour cette déstabilisation (Tokuriki *et al.*, 2008; Wang *et al.*, 2002).

La modulation de la séquence d'une protéine par des mutations est également exploitée dans un contexte biotechnologique. L'industrie pharmaceutique déploie énormément de ressources afin de développer de nouvelles molécules thérapeutiques peptidiques qui vont lier spécifiquement certaines protéines dans le but de moduler leur fonction dans le traitement de maladies (Thielges *et al.*, 2008; Fleishman *et al.*, 2011). Récemment, des modifications au niveau de la séquence des anticorps ont été effectuées afin d'obtenir des molécules de très hautes affinités et spécificités (Sirin *et al.*, 2016). Par exemple, la drogue Trastuzumab est un anticorps dirigé contre le récepteur HER2 qui bloque son activation et son effet anti-apoptotique. Cette drogue améliore le taux de réponse aux traitements de chimiothérapie et augmente le taux de survie de patientes atteintes d'un cancer du sein HER2 positif métastatique (Tan et Swain, 2003). Récemment de nouveaux traitements conjuguent des médicaments aux anticorps et peuvent être utilisés pour les transporter de façon spécifique vers certaines cellules, diminuant ainsi les effets secondaires et augmentant l'efficacité du traitement (Bouchard *et al.*, 2014).

En ingénierie des protéines, la séquence d'enzymes est modifiée afin d'augmenter leur thermorésistance (Guo *et al.*, 2014), augmenter leur vitesse de réaction (Khersonsky *et al.*, 2011) et moduler les réactions qu'elles catalysent (Otten *et al.*, 2010). Ces enzymes synthétiques ont des applications en chimie verte organique en remplaçant des catalyseurs énergétiquement coûteux et potentiellement dangereux pour l'environnement (Sheldon, 2012) par des enzymes hautement spécifiques qui améliorent l'efficacité des réactions (Bornscheuer *et al.*, 2012). Par exemple, certains groupes ont développé des souches de

levures qui produisent des quantitées industrielles de composés chimiques utilisés dans la synthèse de l'Artemisinin, une drogue utilisée contre la malaria (Ro *et al.*, 2006). Également, des enzymes ont été créées *de novo* afin de catalyser des réactions qu'aucune enzyme naturelle n'effectue (Privett *et al.*, 2012; Röthlisberger *et al.*, 2008).

Prédictions de l'effet des mutations sur la stabilité protéique

Les protéines sont des molécules complexes construites à partir de 20 acides aminés relativement simples. Pour en arriver à cette sophistication, chaque résidu d'une protéine doit interagir avec plusieurs autres résidus, créant ainsi pour chacune un contexte structurel et physicochimique pratiquement unique (Lu et Freeland, 2006). Sans surprise, les mutations causent des changements physico-chimiques dans ces environnements complexes et produisent des effets variés qu'aucune règle simple ne peut prédire. Heureusement, il existe plusieurs approches expérimentales qui permettent de mesurer l'effet de la modulation de la séquence d'un gène sur sa fonction et ses différents états thermodynamiques. Ces approches possèdent différents niveaux de précision et de vitesse d'exécution. En effet, des approches fondamentales permettent d'obtenir avec grande précision plusieurs paramètres thermodynamiques tels que l'enthalpie et l'entropie de dénaturation, alors que de récentes méthodes de hauts débits permettent l'évaluation de l'effet de plusieurs centaines de milliers de mutations sur un gène. Par exemple, l'utilisation de systèmes d'expression de protéines en surface de cellules ou phages permet de tester la force d'intéraction d'une protéine pour une dizaine de millions de séquences peptidiques uniques (Reich *et al.*, 2015). Cependant, malgré des avancements majeurs de ce type d'approches, l'étude de leur effet dans un contexte de hauts débits demeure encore laborieuse (Vasser *et al.*, 2004; Zuber *et al.*, 2015; Foight *et al.*, 2014; Verschueren *et al.*, 2013; Fowler et Fields, 2014) et il n'est toujours pas possible de tester exhaustivement toutes les possibilités de mutations. En effet, lors du design protéique, les possibilités de combinaisons de mutations augmentent exponentiellement avec le nombre de résidus à optimiser et le nombre de substitutions désiré. Par exemple, dans le développement de nouveaux peptides spécifiques qui interagissent avec des membres de la famille de protéines Bcl-2, le nombre de séquences possible pour un 23-mer est de 838,860,800,000,000,000,000,000,000,000 (20^{23}) alors que les approches expérimentales

actuelles permettent de tester au maximum 10^{12} séquences par expérience. Certaines approches tentent alors d'évaluer seulement de simples mutations et ensuite de combiner celles qui confèrent les propriétés désirées. Malheureusement, leur effet est rarement additif et est plutôt épistatique (Gong *et al.*, 2013; Figliuzzi *et al.*, 2015).

Ainsi, peu importe l'approche expérimentale utilisée, il demeure toujours impossible d'évaluer systématiquement toutes les mutations d'intérêt, autant dans un contexte thérapeutique que biotechnologique. En règle générale, une petite fraction des mutations retrouvées chez l'humain est pathogénique et, tout comme une fraction des mutations possibles dans le design protéique modulera de façon significative le phénotype recherché (Studer *et al.*, 2014; Fowler *et al.*, 2010). Alors, des outils bio-informatiques peuvent être utilisés pour guider la recherche et la sélection de mutations qui ont de hautes probabilités de conférer les caractéristiques voulues. Ces approches doivent être en mesure de prédire un phénotype à partir d'une séquence. Heureusement, plusieurs banques de données expérimentales ont été construites à partir de la littérature permettant ainsi de construire et de valider des modèles prédictifs par l'utilisation de leurs valeurs. Par exemple, la banque de données ProTherm a récolté l'effet sur la stabilité thermodynamique de plus de 12000 mutations retrouvées chez 740 protéines uniques possédant une structure connue (Kumar, 2006). La banque de données SKEMPI comprend 3047 mutations qui affectent l'affinité d'interaction protéine-protéine (Moal et Fernandez-Recio, 2012), alors que les banques de données ClinVar (Landrum *et al.*, 2014), dbSNP(Sherry *et al.*, 2001) et HGMD (Stenson *et al.*, 2014) sont composées de mutations associées à des maladies observées cliniquement. Finalement, le répertoire COSMIC comprend des mutations associées à certains types de cancers (Forbes *et al.*, 2015).

Les modèles de prédiction les plus simples sont basés sur la séquence de gènes et s'intéressent principalement au niveau de conservation des résidus mutés et du type de mutations. Par exemple, une mutation d'une position conservée par un acide aminé aux propriétés physicochimiques différentes sera souvent classée comme étant déstabilisante ou bien susceptible d'affecter la fonction d'un gène (Ng et Henikoff, 2006). Les modèles plus récents utilisent des approches d'apprentissage automatisé qui ont été entraînées à

discriminer des mutations pathogéniques des mutations neutres à partir des banques de données (Bendl *et al.*, 2014; Wettstein *et al.*, 2014; Thusberg *et al.*, 2011; Cooper et Shendure, 2011). Ces modèles sont généralement performants, mais on a de la difficulté à prédire l'effet des mutations se produisant à des positions non conservées (Kircher, 2014). En outre, ils fournissent peu d'information sur les déterminants structuraux ou les mécanismes thermodynamiques responsables de la pathogénicité. Ils ne peuvent également pas être appliqués en design protéique, car ils ne prédisent que des mutations néfastes.

Des approches plus complexes intègrent l'information de la structure protéique afin de considérer l'environnement local d'un résidu muté et ses interactions avec les atomes à proximité. Généralement, ces algorithmes utilisent une fonction de pointage qui va déterminer l'énergie potentielle de la structure de la forme sauvage et de la forme mutée pour ensuite comparer ces valeurs afin d'obtenir l'effet de la mutation sur la forme native. Les champs de force utilisés pour évaluer ces structures sont variés. Certains sont basés sur des principes physiques et de mécanique moléculaire (van Gunsteren et Mark, 1992) en intégrant les interactions électrostatiques, les forces de van der Waals, l'énergie covalente, l'énergie de rotation d'angle dièdre et l'énergie du repliement d'angle (Pearlman *et al.*, 1995). D'autres fonctions de pointage empirique sont inspirées de ces champs de force, mais elles sont cependant paramétrées afin de corréler avec des données expérimentales. Par exemple, la fonction de score FoldX (Schymkowitz *et al.*, 2005) utilise dix termes énergétiques; l'importance de chacun a été optimisée linéairement afin de minimiser l'erreur sur la prédiction de l'effet de 1 000 mutations sur la stabilité protéique. Certaines fonctions de score sont également basées sur le principe statistique de la loi de Maxwell-Boltzmann en inférant que plus un état est observé, plus il devrait être énergétiquement favorable. Les différents modèles se démarquent par leur définition d'un état qui est souvent représenté par la distance entre deux atomes. Ainsi, plus une distance est fréquente entre deux types d'atomes, plus elle devrait être favorable. Des récents modèles ont tenté d'utiliser des distances entre trois atomes pour définir un état, cependant les combinaisons possibles sont exponentiellement plus élevées et les statistiques dérivées sont moins robustes (Thompson *et al.*, 2014). Finalement, suite à des développements importants en intelligence artificielle, de nouveaux algorithmes basés sur des algorithmes d'apprentissage

automatisé tentent de détecter des tendances et des relations multidimensionnelles reliées à la stabilité protéique. Par exemple, le programme PoPMuSiC 2.0 utilise 15 termes qui décrivent l'impact des mutations sur les propriétés d'une structure protéique pour entraîner un réseau de neurones. Il ont utilisé cette information pour prédire la variation de l'énergie libre de Gibbs dans plus de 2600 mutations (Dehouck *et al.*, 2009).

La considération de plusieurs états (multistate) a démontré son efficacité dans la performance des prédictions de l'effet de mutations (Fischer *et al.*, 2014; Davey et Chica, 2014). Ces techniques utilisent (ou génèrent) plusieurs conformations d'une protéine et vont utiliser l'énergie moyenne ou minimale pour décrire l'énergie de cet état macroscopique. L'exploration de cet espace conformationnel permet des prédictions plus robustes en accommodant un plus grand nombre de mutations (Friedland *et al.*, 2009). Dans le même ordre d'idée, de nouvelles approches considèrent l'énergie de plusieurs états macroscopiques dans la prédiction des proportions de populations. Par exemple, en considérant des états dénaturés et natifs dans la prédiction de l'effet de mutations sur la stabilité protéique, Davey *et al.* (2015) ont été en mesure de prédire 10 nouvelles séquences stables de la protéine Gβ1 alors qu'en considérant l'énergie des formes non liées de complexes protéiques (Dehouck *et al.*, 2013) de meilleures prédictions ont été obtenues lors de l'optimisation d'un complexe synthétique protéique liant l'hémagglutinine.

La majorité des approches d'analyses structurales considère l'énergie d'une structure ou bien l'énergie minimale d'un ensemble de structures lors de l'évaluation de l'énergie de cette dernière. Ces approches considèrent ainsi que l'énergie d'un état macroscopique est représentée par une seule structure, alors qu'en réalité les protéines sont des entités dynamiques constituées de plusieurs conformations. Théoriquement, ces approches approximent alors que la structure évaluée représente la grande majorité de l'état macroscopique et que l'énergie calculée représente l'enthalpie de l'état macroscopique. Cependant, considérant les nombreuses démonstrations expérimentales de l'apport important de l'entropie à la stabilité protéique, il est surprenant qu'aucun algorithme ne la considère dans leur fonction de score. En effet, à l'exception d'une approximation de l'entropie de l'eau par le calcul des surfaces accessibles aux solvants ou bien d'un coût

entropique générique indépendant de la structure lors du repliement protéique (Doig et Sternberg, 1995), la contribution de l'entropie conformationnelle d'un état macroscopique est régulièrement négligée. Cette valeur est cependant difficilement validée et est computationnellement coûteuse à obtenir. En effet, son obtention s'effectue par une exploration complète de l'espace conformationnel d'un état macroscopique et de l'évaluation énergétique de chacune des conformations obtenues. Également, la validité des valeurs calculées est difficilement vérifiable. À l'exception de la Résonance Magnétique Nucléaire (RMN) qui donne de l'information sur les propriétés dynamiques d'une protéine sur plusieurs échelles de temps pour chacun des résidus, les autres approches expérimentales ne donnent que l'entropie totale du système. Il est donc impossible de différentier l'entropie de l'eau de l'entropie de la protéine.

D'un point de vue bioinformatique, l'approche privilégiée pour explorer l'espace conformationnel d'une protéine et potentiellement obtenir son entropie est l'utilisation de simulations de dynamique moléculaire. Il s'agit d'une approche itérative qui à l'aide d'un champ de force évalue les forces appliquées sur chacun des atomes d'une structure. En intégrant l'équation de Newton, il est possible d'obtenir le déplacement sur un court laps de temps pour tous ces atomes afin de recalculer les forces sur cette nouvelle conformation. Cette approche explore et cartographie alors la surface énergétique d'un système d'intérêt en rapportant les états accessibles à une température et un intervalle de temps donnés. Les conformations prédominantes sont alors, selon la loi de Maxwell-Boltzmann, les conformations les plus stables. En comparant la trajectoire d'une forme sauvage et mutée, il est possible d'obtenir la différence d'énergie entre ces deux formes (van der Kamp *et al.*, 2010). Cependant, les champs de force utilisés semblent être biaisés et vont parfois donner des résultats contradictoires. En effet, les simulations de dynamique moléculaire surestiment les ponts hydrogènes et sous évaluent l'hydration du squelette peptidique (Skinner *et al.*, 2014). Également, une évaluation de différents champs de force a démontré que chacun avait un biais importants sur la formation et une préférence pour différentes structures secondaires (Freddolino *et al.*, 2009) alors qu'une modulation des propriétés des molécules d'eau est nécessaire pour reproduire des valeurs dynamiques obtenus part le transfer d'énergie de resonance de Förster (FRET)(Best *et al.*, 2015). Finalement, pour

obtenir une bonne représentation de l'espace conformationnel et extraire une entropie, il faut simuler de longues trajectoires qui, sans ressources informatiques spécialisées, peuvent être difficiles à obtenir (Piana *et al.*, 2012). Ainsi, dans un contexte de haut débit où une évaluation de milliers de mutations est requise, l'utilisation de simulation de dynamique moléculaire n'est pas applicable. L'analyse des modes normaux (AMN) est une approche computationnelle alternative qui permet d'explorer l'espace conformationnel des protéines de façon approximative et rapide.

Analyse des modes normaux

Théorie

L'AMN est une approche qui consiste à étudier les modes de résonance d'un système. Chaque mode correspond à un mouvement et à une fréquence de résonance qui oscille autour d'une position d'équilibre. Généralement, ces modes sont classés de la fréquence la plus lente à la plus rapide et les modes les plus lents sont les plus coopératifs. Ils représentent alors des mouvements coordonnés d'un ensemble de degrés de liberté. Un des systèmes les plus simples d'AMN est représenté par une masse attachée à un plafond avec un ressort. En tirant la masse vers le bas et en la relâchant, le système oscillera de haut en bas autour de sa position initiale à une fréquence constante. En effet, un travail est effectué par la déformation du ressort et cette énergie potentielle est accumulée dans ce dernier. En relâchant la masse, le système va transformer cette énergie potentielle en énergie cinétique qui sera maximale au point d'équilibre (position initiale au repos). En raison de l'inertie accumulée, la masse va poursuivre sa trajectoire et compresser le ressort qui va réemmagasiner de l'énergie potentielle qui sera ensuite retransformée en énergie cinétique pour décompresser le ressort. En absence de friction, ce mouvement serait perpétuel alors qu'en réalité la friction transfère l'énergie du système oscillant dans son environnement, ce qui diminue graduellement l'amplitude du mouvement. Plus le ressort est étiré ou compressé, plus il y aura de l'énergie dans le système et plus grande sera l'amplitude du mouvement. L'énergie (ΔE) nécessaire pour déformer un ressort (Δx) de sa position initiale (x_0) à une nouvelle position (x) est une fonction quadratique dépendante de la constante du ressort (k) représentée par :

$$\Delta E(x) = k\Delta x^2/2 \ = k(x - x_0)^2/2$$

Le système vibre alors le long d'une surface énergétique quadratique et il s'agit d'un mouvement dit harmonique. L'intégration des statistiques de Maxwell-Boltzmann sur cette fonction est représentée par une distribution suivant une loi normale. Cette dernière représente la probabilité d'obtenir la masse à une position (x) en fonction de son énergie (E) et d'une température donnée (T):

$$P(x) = e^{-\Delta E(x)/k_B T}/Z$$
$$= e^{-k(x-x_0)^2/2k_B T}/Z$$
$$Z = \int e^{-k(x-x_0)^2/2k_B T} dx = \sqrt{\pi k_B T/2k}$$

En intégrant en fonction de toutes les conformations possibles et leur probabilité, l'entropie vibrationnelle (S_{Vib}) est définie selon:

$$S_{Vib} = -k_B \int P(x) \ln P(x) dx$$
$$S_{Vib} = k_B \ln(2\pi e k_B T/2k)/2$$

Si le poids de la masse (m) et la constante du ressort sont connus, il est possible analytiquement de prédire la fréquence et le mouvement de résonance du système. En effet, selon la loi de Hooke, la force (F) exercée par le ressort pour revenir à sa position d'équilibre lorsqu'il est déformé est dépendante de sa constante de ressort et de sa déformation:

$$F = -k\Delta x$$

Ainsi, selon la deuxième loi de Newton, la force peut être représentée en fonction du poids de la masse (m), de l'accélération ($\ddot{x}$) ou représentée en fonction du déplacement (x) dans le temps (t):

$$F = m\ddot{x} = m\frac{d^2x}{dt^2}$$

En combinant ces deux dernières équations, en intégrant en fonction du temps (par l'utilisation de nombres imaginaires et l'identité d'Euler) et en considérant qu'initialement le ressort était à son amplitude maximale (A), la position de la masse en fonction du temps est décrite par :

$$x(t) = Acos(\lambda t)$$

Où la fréquence de résonance (λ) est dépendante de la constance de ressort et de la masse :

$$\lambda = \sqrt{k/m}$$

Ainsi, la fréquence est indépendante de l'énergie donnée au système qui est représentée par l'amplitude du mouvement. Également, la friction du système avec son environnement n'affectera pas le type de mouvement, mais diminuera la fréquence de résonance selon un facteur d'amortissement. Ce facteur d'amortissement est principalement dicté par la viscosité du milieu.

L'obtention des modes normaux pour un système composé de multiples degrés de liberté suit le même raisonnement mathématique, mais utilise des vecteurs et des matrices pour représenter les différents masses et ressorts. Les différents degrés de liberté peuvent être n'importe quels paramètres contraints par un potentiel harmonique. Il pourrait autant s'agir de la distance entre deux masses, que leur position cartésienne ou bien la rotation autour

d'un angle dièdre formé par quatre masses. Pour un système qui possède N degrés de liberté, la loi de Hooke est alors représentée par :

$$\vec{F} = H\overrightarrow{\Delta \mathrm{x}}$$

Où $\vec{F}$ représente un vecteur des forces appliquées sur chacun des N degrés de liberté, $\overrightarrow{\Delta \mathrm{x}}$ le déplacement des N dégrées de liberté et H une matrice Hessienne symétrique de dimension N par N. Cette matrice représente l'interaction entre chaque degré de liberté et est composée des constantes des ressorts qui relient ces derniers et du poids des masses. Pour des degrés de liberté i et j, elle est construite selon :

$$H_{i,j} = \frac{d^2E}{dx_i dx_j} = \frac{-k_{i,j}}{\sqrt{M_i M_j}}$$

La diagonale de cette matrice est égale à la somme négative d'une rangée. En utilisant les mêmes transformations que le système à une masse, il est possible de transformer le système d'équations en problème de type Eigen :

$$H\overrightarrow{\Delta \mathrm{x}} = \lambda\overrightarrow{\Delta \mathrm{x}}$$

Ce système peut être facilement résolu par la décomposition de H en vecteurs propres (Q) et en valeurs propres (λ). Ils représentent respectivement les modes de résonance du système, soit les mouvements de résonance et leurs fréquences respectives:

$$H = Q\Lambda Q^{-1}$$

Les vecteurs propres sont les colonnes de Q et ils sont normalisés afin d'être unitaire. Les valeurs propres sont la diagonale de Λ. Les matrices sont réorganisées afin que les fréquences et mouvements de résonance les plus lentes soient les premières. Il y a autant de

solutions que de degrés de liberté présents dans le système, soit N modes. Ainsi, la modulation des positions du système ($\overrightarrow{\Delta x}$) dans le temps par un mode i sera une vibration représentée par le vecteur propre Q_i à une fréquence représentée par la valeur propre λ_i:

$$\overrightarrow{\Delta x}(t) = Q_i A sin(\lambda_i t)$$

Les masses vont vibrer de façon coopérative, c'est-à-dire qu'elles vont bouger de façon simultanée avec l'amplitude relative et la direction représentée dans le vecteur propre. Par exemple, un système de 3 masses connectées par 2 ressorts le long d'un axe possède 3 degrés de liberté et donc 3 modes de résonance. En considérant que chaque masse a le même poids ($M = 1$) et chaque ressort la même constante ($k = 1$), la matrice Hessienne sera représentée par :

$$H = \begin{bmatrix} 1 & -1 & 0 \\ -1 & 2 & -1 \\ 0 & -1 & 1 \end{bmatrix}$$

La décomposition de la matrice Hessienne donnera des valeurs propres de 0, 1 et 3 et les vecteurs propres correspondants sont $(0.58 \quad 0.58 \quad 0.58)$, $(-0.7 \quad 0 \quad 0.7)$, et $(0.4 \quad -0.8 \quad 0.4)$. Les vecteurs représentent un mouvement de résonance pour chaque degré de liberté qui dans cet exemple est le mouvement de chaque masse le long de l'axe représenté dans la figure 1.4

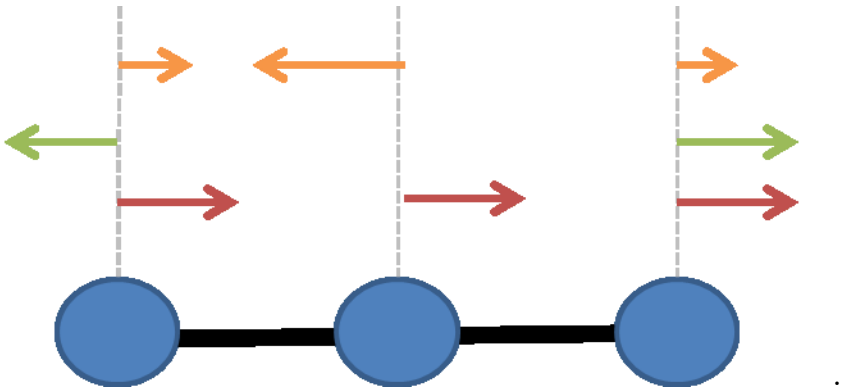


Figure 1.4 – Exemple de mouvements de résonance
Schéma d'un système de 3 masses de poids identique connectées par 2 ressorts de constante identique. Ce système possède 3 fréquences de résonance. Le mode le plus lent (rouge) est une fréquence non triviale qui n'affecte pas l'énergie potentielle du système, car aucun ressort ne subit de distorsion. Le deuxième (vert) et le troisième (orange) mouvement de résonance sont des mouvements coopératifs, c'est-à-dire que toutes les masses vont se déplacer le long de l'axe de façon simultanée selon l'amplitude relative des flèches.

Le premier mode possède une fréquence de résonance nulle et est considéré comme non trivial. Il ne requiert aucune déformation de ressort, car il représente la translation du système le long de l'axe. Les modes de résonance sont également orthogonaux, c'est-à-dire qu'il est impossible de reconstruire un vecteur propre à partir des autres vecteurs. Ainsi, tous les arrangements d'un système ($\vec{x}$) à partir de la configuration d'équilibre initiale ($\overrightarrow{x_0}$) sont décrits par une combinaison unique d'amplitude ($\vec{A}$) appliquée sur chacun des modes :

$$\vec{x} = \overrightarrow{x_0} + \sum_i^N A_i Q_i = \overrightarrow{x_0} + Q\vec{A}$$

L'obtention des amplitudes qui décrivent une conformation est obtenue par une résolution d'un système de N équations et N inconnus :

$$Q\vec{A} = \overrightarrow{\Delta x} = \vec{x} - \overrightarrow{x_0}$$
$$\vec{A} = Q^{-1}\overrightarrow{\Delta x}$$

Ce système peut être résolu par une décomposition en valeurs singulières de la matrice Q^{-1}. L'énergie potentielle nécessaire pour atteindre cette conformation est dépendante des valeurs propres et des amplitudes $\vec{A}$:

$$E(\vec{A}) = \sum_{i}^{N} \lambda_i {A_i}^2/2 = \vec{\lambda} \cdot \overrightarrow{A^2}/2$$

Ainsi, pour une même énergie, les modes les plus lents peuvent alors explorer une plus grande amplitude de mouvement que les modes les plus rapides. Étant donné que chaque conformation est décrite par un ensemble unique d'amplitudes, chaque conformation possède une énergie unique. En représentant les amplitudes par les conformations correspondantes et en intégrant selon les statistiques de Maxwell-Boltzmann de la même façon qu'un système à une masse, la probabilité d'obtenir une conformation $\vec{x}$ à une température donnée est représentée par la fonction de densité de probabilité (FDP) :

$$P(\vec{A}) = e^{-E(\vec{A})/k_B T}/Z$$

$$P(\vec{x}) = \frac{1}{\sqrt{(2\pi)^N |C|}} e^{-(\vec{x}-\overrightarrow{x_0})^T H(\vec{x}-\overrightarrow{x_0})/2k_B T}$$

Où C représente la matrice de covariance définie à partir de la pseudo inverse de la matrice Hessienne (H^{-1}) (Yuen, 2010) qui à son tour est obtenue à partir de la décomposition en valeurs et vecteurs propres :

$$C = k_B T H^{-1}$$

$$H^{-1} = \sum_{i,\lambda_i \neq 0}^{N} \lambda_i^{-1} Q_i Q_i^T$$

Les modes non triviaux qui possèdent une fréquence de résonance nulle ($\lambda_i = 0$) sont ignorés dans la formation de la matrice de covariance. Ainsi, la matrice Hessienne décrit la surface énergétique du système alors que la matrice de covariance décrit le comportement

moyen du système dans cet espace cartésien. Le déterminant de la matrice de covariance est obtenu à l'aide des valeurs propres de la matrice Hessienne:

$$|C| = (k_B T)^N / \prod_{i,\lambda_i \neq 0}^{N} \lambda_i$$

En considérant que chaque conformation $\vec{x}$ est un état microscopique du système de résonance, l'entropie vibrationnelle est définie selon (Ahmed et Gokhale, 1989; Karplus et Kushick, 1981) :

$$S_{Vib} = -k_B \int P(\vec{x}) \ln P(\vec{x}) d\vec{x}$$

$$\frac{S_{Vib}}{k_B} = \frac{N}{2}(1 + \ln(2\pi)) + \frac{1}{2} N \ln(k_B T) - \sum_{i,\lambda_i \neq 0}^{N} \ln \lambda_i$$

Ainsi, une augmentation de la température, une diminution des fréquences de résonance ou une augmentation du nombre de degrés de liberté augmente l'entropie du système. Étant donné que les modes les plus lents possèdent les courbatures énergétiques les plus faibles, ils contribuent de façon plus importante à l'entropie. La variation d'entropie en fonction de la température (T_1 à T_2) est indépendante des fréquences de résonance et seulement dépendant du nombre de degrés de liberté:

$$\frac{S_{Vib}(T_1) - S_{Vib}(T_2)}{k_B} = \frac{1}{2} N (\ln(k_B T_1) - \ln(k_B T_2))$$

$$\Delta S_{Vib}(\Delta T) = \frac{1}{2} N k_B \ln\left(\frac{T_1}{T_2}\right)$$

Alors que la différence d'entropie entre deux systèmes (A et B) est dépendante des fréquences de résonance et indépendante de la température :

$$\frac{\Delta S_{Vib,A,B}}{k_B} = \left(\frac{N}{2}(1+\ln(2\pi)) + \frac{1}{2}N\ln(k_BT) - \sum_{i,\lambda_i\neq 0}^{N}\ln\lambda_{i,A}\right) - \left(\frac{N}{2}(1+\ln(2\pi)) + \frac{1}{2}N\ln(k_BT) - \sum_{i,\lambda_i\neq 0}^{N}\ln\lambda_{i,B}\right)$$

$$\Delta S_{Vib,A,B} = \sum_{i,\lambda_i\neq 0}^{N}\ln\lambda_{i,B} - \sum_{i,\lambda_i\neq 0}^{N}\ln\lambda_{i,A}$$

La dérivation de ces paramètres thermodynamiques est vraie seulement si on considère que le système de vibration est en équilibre avec son environnement et suit le théorème de l'équipartition de l'énergie (Waterson, 1850). Ce théorème dicte que chaque degré de liberté d'un système possède en moyenne la même énergie. Alors, l'amplitude maximale des mouvements sur chacun des modes est dépendante de la température du système. À l'inverse, si on considère un régime quantique dans des conditions plus froides, chaque fréquence de résonance représente un état d'énergie discret dont la répartition d'énergie est distribuée selon les statistiques de Bose–Einstein. Les modes les plus lents seraient les modes les plus probables. L'entropie est alors définie selon (Tidor et Karplus, 1994) :

$$S_{Vib} = \sum_{i,\lambda_i\neq 0}^{N}\frac{h\lambda_i e^{-h\lambda_i/2k_BT}}{k_BT(1-e^{-h\lambda_i/2k_BT})} - \ln(1-e^{-\frac{h\lambda_i}{2k_BT}})$$

Où h représente la constante de Planck. Ainsi, en théorie, en abaissant la température, un système devrait seulement vibrer sur les modes les plus lents et les modes les plus rapides devraient être « gelés ». Plusieurs travaux concernant la transition vitreuse semblent appuyer cette hypothèse (Miyazaki *et al.*, 2000; Demmel *et al.*, 1997). En effet, en refroidissant subitement un échantillon de protéine avant qu'elle ne puisse se dénaturer, il est possible de voir un changement dans les propriétés dynamiques harmoniques de la protéine correspondant à une perte de certains modes de résonance (Daniel *et al.*, 2002). Cependant, ces conditions expérimentales ne représentent pas un contexte biologique et, probablement, le théorème de l'équipartition de l'énergie est plus approprié dans l'analyse des modes normaux chez les systèmes biologiques.

L'analyse des modes normaux ne se limite pas seulement à prédire les fréquences et les mouvements de résonance d'un système, mais aussi à comprendre comment certains degrés de liberté interagissent entre eux à partir de valeurs de covariance. En effet, la relation entre la matrice de covariance et la matrice Hessienne peut être utilisée afin de discriminer des covariances indirectes (LeVine et Weinstein, 2015). Par exemple, LeVine et Weinstein (2014) ont mis à jour un mécanisme d'allostérie en étudiant une matrice de covariance obtenue par une simulation dynamique moléculaire du gène *LeuT*. En observant la forme pseudo-inverse de cette matrice, ils ont identifié des positions fortement couplées qui correspondent à des résidus identifiés comme étant critiques à la fonction de la protéine. Également, l'analyse de la matrice de covariance peut être utilisée pour observer des tendances générales d'un système. Cette méthode est utilisée dans la décomposition des simulations de dynamique moléculaire en dynamique essentielle ou analyse des composantes principales (« Principal Component Analysis ») (Jollife, 1986). Ces approches observent la tendance générale d'une simulation et extrapolent des mouvements globaux. Cependant, pour être valide, la matrice de covariance doit être obtenue par l'étude d'un échantillon qui couvre un ensemble réaliste de l'espace configurationnel du système. Par exemple, l'étude d'un échantillon de quelques picosecondes de dynamique moléculaire laisserait croire que les seuls mouvements d'une protéine sont des réarrangements de chaînes latérales, alors qu'en réalité il peut y avoir des mouvements de larges amplitudes au niveau du squelette peptide.

Historique

Même si l'analyse des modes normaux est utilisée dans différents domaines, par exemple en sismologie (Gilbert et Dziewonski, 1975), océonagrahie (Arvelo et Zabal, 1997), génie civil (Green et Unruh, 2006) ou l'astronomie (van der Spoel *et al.*, 2015), nous allons nous concentrer sur les applications dans un contexte chimique et biologique. Les premières études d'analyse des modes normaux dans ces domaines consistaient principalement à observer de façon expérimentale ces vibrations par l'utilisation d'approches de spectroscopie. En effet, lorsqu'un système est excité par un photon de fréquence identique à une fréquence de résonance de la molécule, ce dernier est absorbé et ce signal peut être détecté par un appareil. Ainsi, en balayant un ensemble de fréquences continues, il est

possible d'obtenir de façon précise les fréquences de résonance d'un système. Cependant, ces approches ne donnent pas d'information sur le type de mouvements de résonance et des interactions intramoléculaires. Ainsi, des modèles théoriques ont été construits afin de prédire ces fréquences et de comprendre leur origine par la création de modèles d'analyse des modes normaux (Ermer et Lifson, 1974; Hayward et Henry, 1974). Par exemple, les fréquences de résonance du dioxyde de carbone sont connues ainsi que sa conformation chimique. En optimisant les forces d'interaction entre les atomes de carbone et d'oxygènes, il est possible d'avoir un modèle qui prédit avec 1.7% d'écart des valeurs expérimentales. Cependant, ces forces d'interaction dérivées entre les différents atomes sont difficilement transférables à d'autres systèmes ou molécules et utilisent seulement comme degrés de liberté la longueur des liens covalents et les angles dièdres, ignorants les interactions non covalentes. Ces problèmes ont été en partie abordés dans d'importants travaux menés par le groupe de Shneior Lifson en collaboration avec Michael Levitt (Warshel *et al.*, 1970), Arieh Warshel (Lifson, 1968) et Martin Karplus (Warshel et Karplus, 1972) afin de produire des champs de force généralisés ou *consistent force field*. Ces travaux leur ont notamment valu le prix Nobel de chimie en 2013. Ces premiers champs de force étaient composés de coordonnées internes qui représentaient l'énergie du système par la distance des liens covalents, les rotations autour d'angles dièdres, les repliements d'angle, les interactions de Coulomb et un potentiel Lennard-Jones. Ils ont initialement été paramétrés afin de prédire les fréquences de résonance de plusieurs petites molécules par l'analyse des modes normaux (Warshel *et al.*, 1970) et sont également utilisés lors de simulations de dynamique moléculaire (McCammon *et al.*, 1977). En effet, le champ de force CHARMM a été initialement développé par Martin Karplus et est régulièrement utilisé dans ces types de simulations (Brooks *et al.*, 1983).

Cependant, l'utilisation de ces coordonnées internes non cartésiennes complexifie l'analyse des modes normaux. Par exemple, la rotation autour d'un angle dièdre va affecter tous les atomes rattachés à ce dernier et la prédiction et l'impact de cette rotation sur les autres atomes sont complexes à caractériser de façon analytique. Il faut alors recourir à des approches numériques, où chaque degré de liberté est perturbé individuellement et l'impact de cette perturbation est évalué avec un champ de force. Il est alors possible de paramétrer

une constante de ressorts qui va décrire cette perturbation. Cette approche exige beaucoup de ressources lors de l'analyse de plus gros systèmes et n'est pas harmonique sur de grandes amplitudes, restreignant ainsi l'analyse à des mouvements locaux. Une partie des travaux d'Arieh Warshel contournent ces problèmes en transformant ces systèmes en coordonnée cartésiennes. Chaque mouvement des positions d'atomes devient indépendant des autres et l'estimation de la matrice Hessienne est alors simplifiée et peut être dérivée analytiquement (Warshel, 1970). L'inversion de la matrice Hessienne, essentielle à l'analyse des modes normaux, est également une des étapes limitantes dans l'analyse de systèmes plus complexes. En effet, il s'agit d'un problème d'ordre N^3, c'est-à-dire qu'il faut effectuer 8 fois plus d'étapes algorithmiques lorsque la grosseur de la matrice est doublée (Peters et Wilkinson, 1975). Ainsi, sans les ressources informatiques appropriées, l'AMN se limitait à de petites molécules d'une dizaine d'atomes. Avec l'avènement d'ordinateurs de plus en plus puissants et performants, l'AMN a pu être appliquée à de plus gros systèmes biologiques tels que des protéines. La première structure protéique à avoir été analysée est le glucagon (Tasumi *et al.*, 1982) par l'utilisation d'un champ de force ne considérant que les liens covalents comme degrés de liberté. Par la suite, le groupe de Michael Levitt utilisa des champs de force constants afin d'étudier les fréquences et les mouvements de résonance d'un ensemble de 16 protéines (Levitt *et al.*, 1985). Cependant, les différentes analyses sont computationnellement exigeantes, car malgré le fait que les structures résolues par cristallographie devraient représenter les conformations d'équilibre, selon les différents champs de force, elles ne sont pas dans leur conformation énergétique minimale locale. Ainsi, d'intenses protocoles de minimisation doivent être appliqués et ont pour effet de déformer les structures de leur conformation initiale. Afin de contourner ce problème, Monique Tirion proposa de simplifier l'approche en considérant que la structure est déjà dans sa conformation d'équilibre en utilisant un potentiel qui connecte avec un ressort tous les atomes présents dans un rayon prédéterminé variant de 1.1 à 2.5 Angströms (Tirion, 1996). La constante de ressort est la même pour toutes les paires d'interactions, il s'agit alors d'un modèle de réseaux d'élastiques (*elastic network model*), Malgré ces grandes simplifications, comparativement aux résultats obtenus par le groupe de Levitt avec son potentiel plus complexe, les modes les plus lents sont similaires et ils sont peu influencés par la distance utilisée pour connecter deux atomes. En revanche, les modes plus

rapides sont plus sensibles à la précision du potentiel et corrèlent peu avec la simplification du modèle. Quelques années plus tard, Tama *et al.* (2000) ont également démontré qu'il est possible de regrouper plusieurs atomes et considérer qu'ils ne sont qu'une seule masse sans affecter de façon significative les dynamiques des premiers modes. Finalement, le groupe d'Ivet Bahar a approximé que les mouvements de ces masses regroupées peuvent être considérés comme isotropes, réduisant de 66% le nombre de degrés de liberté. En effet, dans ce modèle, le seul degré de liberté pour un résidu est une déviation de sa position initiale, peu importe la direction de cette déviation, alors que dans un modèle anisotropique, il y a 3 degrés de liberté par masse, un pour chaque dimension cartésienne (Bahar *et al.*, 1998). Ces nombreuses approximations diminuent drastiquement le temps computationnel nécessaire pour effectuer une analyse des modes normaux. En utilisant un de ces modèles de faibles résolutions (*coarse-grained*), une exploration complète de l'espace conformationnel d'une structure d'environ 300 résidus peut-être faite en quelques secondes sur un ordinateur portable (Yang *et al.*, 2006). Les propriétés dynamiques de macromolécule de milliers de résidus peuvent être donc explorées avec ces approches qui seraient inaccessibles par dynamique moléculaire. Par exemple, un groupe a étudié une structure du complexe de la protéine chaperonne GroEL qui possède 8015 résidus (Schuyler et Chirikjian, 2005) et des structures de capside de virus constituées de plusieurs centaines de milliers de résidus (Tama et Brooks, 2005).

Intuitivement, la simplification du champ de force utilisé pour effectuer une AMN porteraient à croire que les mouvements prédits seraient de piètre qualité. Cependant, une récente comparaison des mouvements prédits par 30 simulations de dynamique moléculaire de longueur variant de 30 à 100 nanosecondes sur des protéines de 30 à 2542 résidus à l'espace conformationnel prédit par un modèle d'AMN de faible résolution a démontré que les mouvements généraux coopératifs étaient similaires (Rueda *et al.*, 2007). Les simulations de dynamique moléculaire ont exigé l'équivalent de 100 ans de temps calcul et ont dû être effectuées sur une grappe de calcul, alors que l'AMN n'a nécessité que quelques minutes. Des résultats similaires ont été observés pour la protéine MAP kinase p38 map kinase, où une AMN décrit les changements conformationnels observés lors de la liaison à différents partenaires moléculaires de façon plus efficace que des analyses de dynamique

moléculaire (Powell, 2011). Cependant, ces deux méthodes sont complémentaires. En effet, l'AMN est une approche approximative qui permet d'explorer rapidement des changements conformationnels importants au détriment de la perte de résolution de certains détails atomiques que les simulations de dynamique moléculaire peuvent résoudre. De nouvelles approches hybrides combinent ces deux avantages en générant des ensembles de structures par AMN qui sont par la suite simulés en parallèle par la dynamique moléculaire, couvrant ainsi une plus grande fraction de l'espace conformationnel d'une protéine (Tatsumi *et al.*, 2004). L'AMN est également utilisée dans des contextes où il y a des changements conformationnels importants. Par exemple, elle est utilisée pour raffiner des structures cristallines ou des structures obtenues par microscopie électronique (Tama *et al.*, 2004). Elle est aussi utilisée pour caractériser des changements conformationnels entre deux structures qui requièrent des mouvements coordonnés de grands nombres d'atomes. Par exemple, un ensemble de structures générées par AMN augmente le taux de prédiction lors de la recherche d'une pose d'interaction protéine-protéine (Moal et Bates, 2010) ou de petites molécules (May et Zacharias, 2008). Finalement, certains groupes utilisent l'AMN pour observer des réponses à des stress mécaniques ou à des modulations allostériques (Mitternacht et Berezovsky, 2011).

Problématique

Malgré les nombreux avantages offerts par la simplification des champs de force utilisés dans l'AMN, ils demeurent tous incapables de prédire l'effet de mutation. En effet, en regroupant les atomes des résidus à une seule masse située au niveau du squelette peptidique, l'information sur la nature des résidus est perdue. Ainsi, deux protéines distinctes au niveau de la séquence primaire, mais qui possèdent une conformation identique au niveau du squelette peptide, auront des dynamiques identiques lors de l'utilisation de modèle d'AMN. Ce résultat met en évidence une sévère lacune des modèles actuels en sachant qu'il existe plusieurs évidences expérimentales qui démontrent que certaines mutations peuvent affecter de façon significative les propriétés dynamiques d'une protéine et sa fonction (Preiswerk *et al.*, 2014; Jiménez-Osés *et al.*, 2014, Gagné et Doucet, 2013).

Les prochains chapitres décrivent le développement d'un nouveau modèle d'AMN appelé ENCoM qui introduit l'information de la structure primaire dans son potentiel, sa validation comparativement à d'autres modèles et ses nouvelles applications dans des problèmes qui ne peuvent être étudiés par les modèles actuels.

Objectifs

Objectif #1

Implémentation de la nouvelle méthode ENCoM par le développement d'un nouveau potentiel harmonique d'AMN de faible résolution qui introduit la nature des résidus dans les interactions de longue portée.

Objectif #2

Validation sur des tests classiques par la prédiction de facteurs B et par la prédiction de changements conformationnels.

Objectif #3

Validation de la modulation de la structure primaire sur les propriétés dynamiques protéiques par l'étude de structures provenant d'organismes thermophiles et mésophiles et par la prédiction de l'effet de mutations sur la stabilité protéique.

Objectif #4

Développement d'une interface en ligne afin de simplifier l'utilisation d'ENCoM par des utilisateurs non techniques lors de la prédiction de l'effet de mutations sur les propriétés dynamiques et pour la génération d'un ensemble de conformations réalistes.

ARTICLE 1

A Coarse-Grained Elastic Network Atom Contact Model and Its Use in the Simulation of Protein Dynamics and the Prediction of the Effect of Mutations

Auteurs de l'article: Vincent Frappier et Rafael Najmanovich

Statut de l'article: Accepté. Frappier V & Najmanovich RJ, *A coarse-grained elastic network atom contact model and its use in the simulation of protein dynamics and the prediction of the effect of mutations*. PLoS Comput Biol. 2014

Avant-propos: J'ai effectué l'ensemble du travail expérimental ainsi que la rédaction de l'article. Rafael Najmanovich a corrigé et revisé le manuscrit.

Résumé : L'analyse des modes normaux (AMN) est une méthode couramment utilisée lors de l'étude des propriétés dynamiques de structures de protéines. Deux propriétés spécifiques à ces analyses sont leur niveau de simplification lors de la représentation de la structure protéique et la complexité du potentiel énergétique utilisé. Ces choix influencent l'équilibre associé à la vitesse d'exécution de la méthode et sa précision. À une extrémité, des méthodes précises, mais lentes, vont considérer tous les atomes lors de simulations de dynamique moléculaire, alors qu'à l'autre extrême des modèles de réseaux d'élastiques utilisent une représentation simplifiée basée seulement sur la géométrie des Cα de la structure et sont agnostiques de la séquence. Dans cet article, nous présentons ENCoM, un « Elastic Network Contact Model » qui utilise une fonction énergétique basée sur les interactions non covalentes des atomes, permettant ainsi d'inclure l'effet de la nature des résidus sur la dynamique dans un contexte d'AMN. ENCoM est aussi rapide que les méthodes actuelles et est supérieure dans la génération d'ensembles de conformations. Dans cet article, nous présentons également une nouvelle application de l'AMN, soit la prédiction de l'effet de mutations sur la stabilité protéique. Les méthodes actuelles sont basées sur des approches d'apprentissage automatisé ou structurale qui ne considèrent que l'enthalpie, alors que l'utilisation d'ENCoM, basée sur des modes de vibrations, considère un score entropique. Il s'agit d'un nouveau domaine d'application pour l'AMN et il s'agit d'une nouvelle approche pour prédire l'effet de mutations. Nous avons comparé ENCoM à un grand ensemble de méthodes au niveau du pouvoir prédictif et de la robustesse. La précision d'ENCoM est comparable aux méthodes actuelles, lesquelles sont biaisées à prédire les mutations déstabilisantes alors qu'ENCoM est plus robuste pour prédire des mutations stabilisatrices.

Abstract

Normal mode analysis (NMA) methods are widely used to study dynamic aspects of protein structures. Two critical components of NMA methods are coarse-graining in the level of simplification used to represent protein structures and the choice of potential energy functional form. There is a trade-off between speed and accuracy in different choices. In one extreme one finds accurate but slow molecular-dynamics based methods with all-atom representations and detailed atom potentials. On the other extreme, fast elastic network model (ENM) methods with Cα only representations and simplified potentials that based on geometry alone, thus oblivious to protein sequence. Here we present ENCoM, an Elastic Network Contact Model that employs a potential energy function that includes a pairwise atom-type non-bonded interaction term and thus makes it possible to consider the effect of the specific nature of amino-acids on dynamics within the context of NMA. ENCoM is as fast as existing ENM methods and outperforms such methods in the generation of conformational ensembles. Here we introduce a new application for NMA methods with the use of ENCoM in the prediction of the effect of mutations on protein stability. While existing methods are based on machine learning or enthalpic considerations, the use of ENCoM, based on vibrational normal modes, is based on entropic considerations. This represents a novel area of application for NMA methods and a novel approach for the prediction of the effect of mutations. We compare ENCoM to a large number of methods in terms of accuracy and self-consistency. We show that the accuracy of ENCoM is comparable to that of the best existing methods. We show that existing methods are biased towards the prediction of destabilizing mutations and that ENCoM is less biased at predicting stabilizing mutations.

Author Summary

Normal mode analysis (NMA) methods can be used to explore potential movements around an equilibrium conformation by mean of calculating the eigenvectors and eigenvalues associated to different normal modes. Each normal mode represents a global collective, correlated and complex, form of motion of the entire protein. Any conformation around equilibrium can be represented as a weighted combination of normal modes. Differences in the magnitudes of the set of eigenvalues between two structures can be used to calculate

differences in entropy. We introduce ENCoM the first coarse-grained NMA method to consider atom-specific side-chain interactions and thus account for the effect of mutations on eigenvectors and eigenvalues. ENCoM performs better than existing NMA methods with respect to traditional applications of NMA methods but is the first to predict the effect of mutations on protein stability and function. Comparing ENCoM to a large set of dedicated methods for the prediction of the effect of mutations on protein stability shows that ENCoM performs better than existing methods particularly on stabilizing mutations. ENCoM is the first entropy-based method developed to predict the effect of mutations on protein stability.

Introduction

Biological macromolecules are dynamic objects. In the case of proteins, such movements form a continuum ranging from bond and angle vibrations, sub-rotameric and rotameric side-chain rearrangements [1], loop or domain movements through to folding. Such movements are closely related to function and play important roles in most processes such as enzyme catalysis [2], signal transduction [3] and molecular recognition [4] among others. While the number of proteins with known structure is vast with around 85K structures for over 35K protein chains (at 90% sequence identity) in the PDB database [5], our view of protein structure tends to be somewhat biased, even if unconsciously, towards considering such macromolecules as rigid objects. This is due in part to the static nature of images used in publications to guide our interpretations of how structural details influence protein function. However, the main reason is that most known structures were solved using X-ray crystallography [6] where dynamic proprieties are limited to b-factors and the observation of alternative locations. Despite this, it is common to analyze larger conformational changes using X-ray structures with the comparison of different crystal structures for the same protein obtained in different conditions or bound to different partners (protein, ligand, nucleic acid). It is particularly necessary to consider the potential effect of crystal packing [7,8] when studying dynamic properties using X-ray structures. Nuclear magnetic resonance (NMR) is a powerful technique that gives more direct information regarding protein dynamics [9,10]. Different NMR methodologies probe distinct timescales covering 15 orders of magnitude from 10^{-12} s side chain rotations via nuclear spin relaxation to 10^{3} s using real time NMR [10]. In practice, there is a limitation

on the size of proteins that can be studied (between 50–100 kDa) although this boundary is being continuously pushed [11] providing at least partial dynamic information on extremely large systems [12]. However, only a small portion (around 10%) of the available proteins structures in the Protein Data Bank (PDB) are the result of NMR experiments [5].

Molecular dynamic simulations numerically solve the classical equations of motion for an ensemble of atoms whose interactions are modeled using empirical potential energy functions [13–15]. At each time step the positions and velocities of each atom are calculated based on their current position and velocity as a result of the forces exerted by the rest of the system. The first MD simulation of a protein (Bovine Pancreatic Trypsin Inhibitor, BPTI) ran for a total 8.8 ps [16] followed by slightly longer simulations (up to 56 ps) [17]. A film of the latter can be seen online (http://youtu.be/_hMa6G0ZoPQ). Despite using simplified potentials and structure representations (implicit hydrogen atoms) as well as ignoring the solvent, these first simulations showed large oscillations around the equilibrium structure, concerted loop motions and hydrogen bond fluctuations that correlate with experimental observations. Nowadays, the latest breakthroughs in molecular dynamics simulations deal with biological processes that take place over longer timescales. For example, protein folding [18], transmembrane receptor activation [19,20] and ligand binding [21]. These simulations require substantial computer power or purpose built hardware such as Anton that pushes the current limit of MD simulations to the millisecond range [22]. Despite powerful freely available programs like NAMD [23] and GROMACS [24] and the raise of computational power over the last decade, longer simulations reaching timescales where most biological processes take place are still state-of-the-art.

Normal modes are long established in the analysis of the vibrational properties of molecules in physical chemistry [25]. Their application to the study of proteins dates back to just over 30 years [26–30]. These earlier Normal Mode Analysis (NMA) methods utilized either internal or Cartesian coordinates and complex potentials (at times the same ones used in MD). As with earlier MD methods their application was restricted to relatively small proteins. Size limitations notwithstanding, these early studies were sufficient to demonstrate the existence of modes representing concerted delocalized motions, showing a

facet of protein dynamics that is difficult to access with MD methods. Some simplifications were later introduced and shown to have little effect on the slowest vibrational modes and their utility to predict certain molecular properties such as crystallographic b-factors. These simplifications included the use of a single-parameter potential [31], blocks of consecutive amino acids considered as units (nodes) [32] [32] and the assumptions of isotropic [33] fluctuations in the Gaussian Network Model (GNM) or anisotropic fluctuations [34]. These approximations have drastically reduced the computational time required, thus permitting a much broader exploration of conformational space using conventional desktop computers in a matter of minutes. Of these, the most amply used method is the Anisotropic Network Model (ANM) [35,36]. ANM is often referred simply as an elastic network model; one should however bear in mind that all normal mode analysis methods are examples of elastic network models. ANM uses a simple Hook potential that connects every node (a point mass defined at the position of an alpha carbon), within a predetermined cut-off distance (usually 18 A). More recently, a simplified model, called Spring generalized Tensor Model (STeM), that uses a potential function with four terms (covalent bond stretching, angle bending, dihedral angle torsion and non-bonded interaction) has been proposed [37]. The normal mode analysis of a macromolecule produces a set of modes (eigenvectors and their respective eigenvalues) that represent possible movements. Any conformation of the macromolecule can in principle be reached from any other using a linear combination of amplitudes associated to eigenvectors.

It is essential however to not loose sight of the limitations of normal mode analysis methods. Namely, normal modes tell us absolutely nothing about the actual dynamics of a protein in the sense of the evolution in time of atomic coordinates. Plainly speaking, normal mode analysis is informative about the possible movements but not actual movements. Additionally, normal modes tell us of the possible movements around equilibrium. These two caveats clearly place normal mode analysis and molecular dynamics apart. First, molecular dynamics gives an actual dynamics (insofar as the potential is realistic and quantum effects can be ignored). Second, while the equilibrium state (or the starting conformation) affects the dynamics, one can explore biologically relevant timescales given sufficient computational resources to perform long simulations.

The vast majority of coarse-grained NMA models only use the geometry of the protein backbone (via Cα Cartesian position) disregarding the nature of the corresponding amino acid, in doing so a lot of information is lost. To our knowledge, there have been three independent attempts at expanding coarse-grained NMA models over the years to include extra information based on backbone and side-chain atoms. Micheletti et al. [38] developed the bGM model in which the protein is represented by Cb atoms for all residues except Glycine in addition to the Cα atoms. The Hamiltonian is a function exclusively of Cα and Cb distances. As a Gaussian model, bGM does not give information about directions of movement but only their magnitude and can be used solely to predict b-factors. As Cb atoms do not change position, this model cannot be used by definition to predict the effect of mutations, either on dynamics or stability. The authors report results on b-factor prediction comparable to GNM. Lopez-Blanco et al. [39] developed a NMA model in internal coordinates with three different levels of representation: 1. Heavy-atoms, 2. five pseudo- atoms (backbone: NH, Ca, CO and side-chain: Cb and one at the center of mass of the remaining side chain atoms) and 3. Cα representation. While the potential is customizable, the default potential uses a force constant that is distance dependent but atom type independent. The method is validated through overlap analysis on a dataset of 23 cases. The authors report no significant differences in overlap for the different representations. Lastly, Kurkcuoglu et al. [40] developed a method that mixes different levels of coarse-graining. The authors test the method on a single protein, Triosephosphate Isomerase with higher atomic representation for a loop and achieve better overlap values compared to Cα only representation. All the methods above, while adding more detail to the representation utilize force constants that are not atom-type dependent. Therefore, while less coarse-grained, all the methods above are still atom-type and amino-acid type agnostic.

By definition, irrespective of the level of coarse-graining, such models cannot account for the effect of mutations on protein dynamics or stability. It has been shown that different amino acids interact differently and that single mutations can have a high impact on protein function and stability [41–43]. Mutations on non-catalytic residues that participate into concerted (correlated) movements have been shown to disrupt protein function in NMR relaxation experiments [44–46]. Several cases have been documented of mutations that

don’t affect the global fold of the protein, but affect protein dynamics and disrupt enzyme function [47].

To overcome this limitation of coarse-grained NMA methods while maintaining the advantages of simplified elastic network models, we developed a new mixed coarse-grained NMA model called Elastic Network Contact Model (ENCoM). ENCoM employs a potential function based on the four bodies potential of STeM with an addition to take in consideration the nature and the orientation of side chains. Side-chain atomic contacts are used to modulate the long range interaction term with a factor based on the surface area in contact [48] and the type of each atom in contact. Additionally, we introduce a non-specific version of ENCoM (ENCoMns) where all interactions between atom types are the same. ENCoM and ENCoMns were validated with comparison to ANM, GNM and STeM with respect to the prediction of crystallographic b-factors and conformational changes, two properties conventionally used to test ENM methods. Moreover, we test the ability of ENCoM and ENCoMns to predict the effect of mutations with respect to protein stability and compare the ability of ENCoM and ENCoMns to a large number of existing methods specifically designed for the prediction of the effect of single point mutations on protein stability. Finally, we use ENCoM to predict the effect of mutations on protein function in the absence of any effects on protein stability.

Results

Correlation between experimental and predicted crystallographic b-factors

We utilized a dataset of 113 non-redundant high-resolution crystal structures [49] to predict b-factors using the calculated ENCoM eigenvectors and eigenvalues as described previously [35] (Equation 4). We compared the predicted b-factors using ENCoM, ENCoMns, ANM, STeM and GNM to the experimental Cα b- factors for the above dataset (Supplementary Table S1). For each protein we calculate the Pearson correlation between experimental and predicted values. The results in Figure 1 represent the bootstrapping average of 10000 iterations. We observe that while comparable, ENCoM (median=0.54) and ENCoMns (median=0.56) have lower median values than STeM (median=0.60) and GNM (median=0.59) but similar or higher than ANM (median=0.54). It should be noted

that it is possible to find specific parameter sets that maximize b-factor correlations beyond the values obtained with STeM and GNM (see methods). However we observe a trade-off between the prediction of b- factors on one side and overlap and the effect of mutations on the other (see methods). Ultimately we opted for a parameter set that maximizes overlap and the prediction of mutations with complete disregard to b-factor predictions. Nonetheless, as shown below, even the lower correlations obtained with ENCoM are sufficiently high to detect functionally relevant local variations in b-factors as a result of mutations. As GNM does not provide information on the direction direction of movements or the effect of mutations, it is not considered further in the present study.

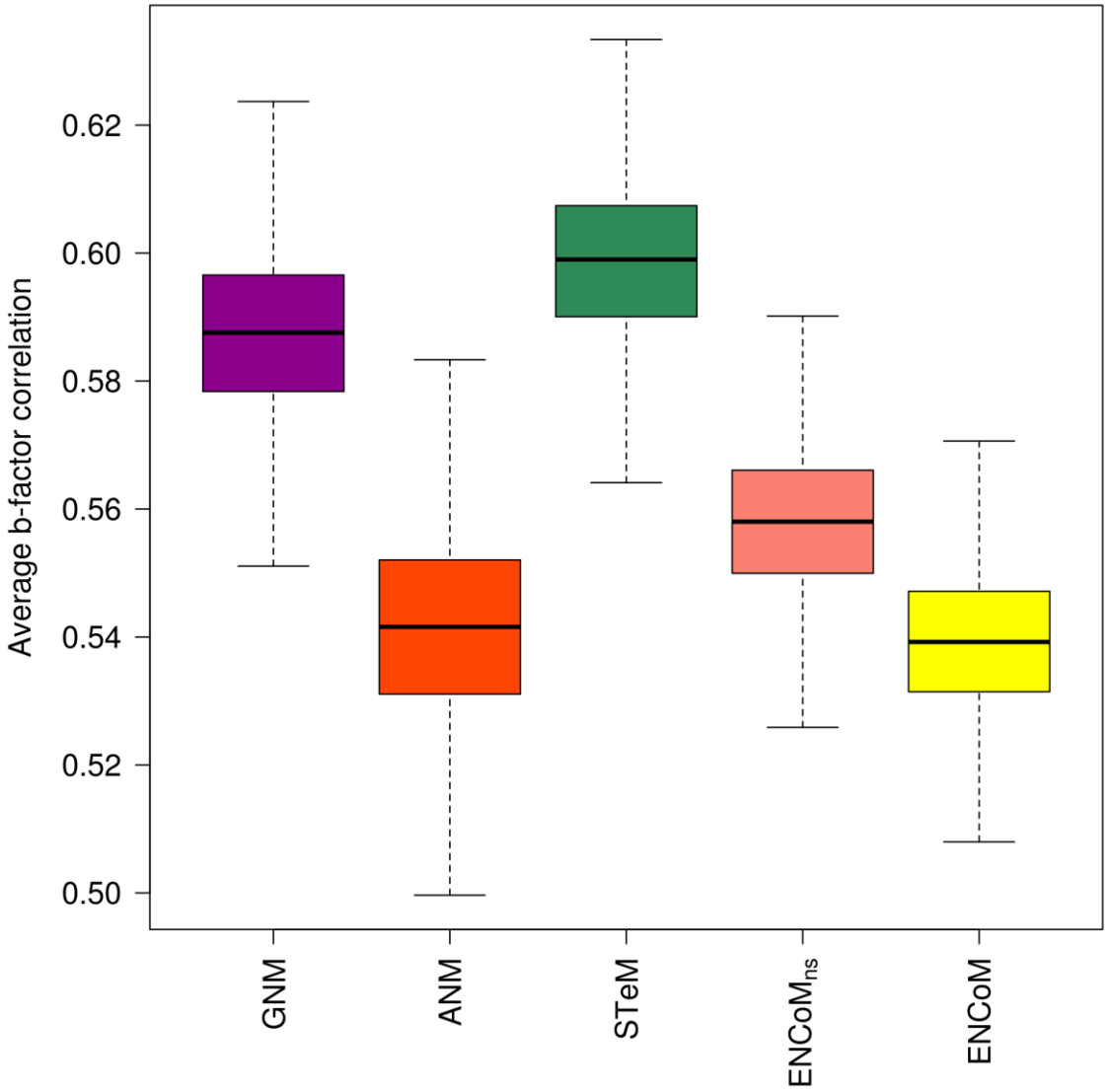


Figure 1 Correlation between predicted and experimental b-factors for different ENM models. Box plots represent the average correlations from a non-redundant dataset containing 113 proteins generated from 10000 resampling bootstrapping iterations. ANM and ENCoM have comparable correlations but lower than GNM and STeM. doi:10.1371/journal.pcbi.1003569.g001

Exploration of conformational space

By definition, any conformation of a protein can be described as a linear combination of amplitudes associated to the eigenvectors representing normal modes. It should be stressed that such conformations are as precise as the choice of structure representation used and correct within the quadratic approximation of the potential around equilibrium. Those limitations notwithstanding, one application of NMA is to explore the conformational space

of macromolecules using such linear combinations of amplitudes. Pairs of distinct protein conformations, often obtained by X-ray crystallography are used to assess the extent to which the eigenvectors calculated from a starting conformation could generate movements that could lead to conformational changes in the direction of a target conformation. Rather than an optimization to determine the amplitudes for a linear combination of eigenvectors, this is often simplified to the analysis of the overlap (Equation 5), i.e., the determination of the single largest contribution from a single eigenvector towards the target conformation. In a sense the overlap represents a lower bound on the ability to predict conformational changes without requiring the use of an optimization process.

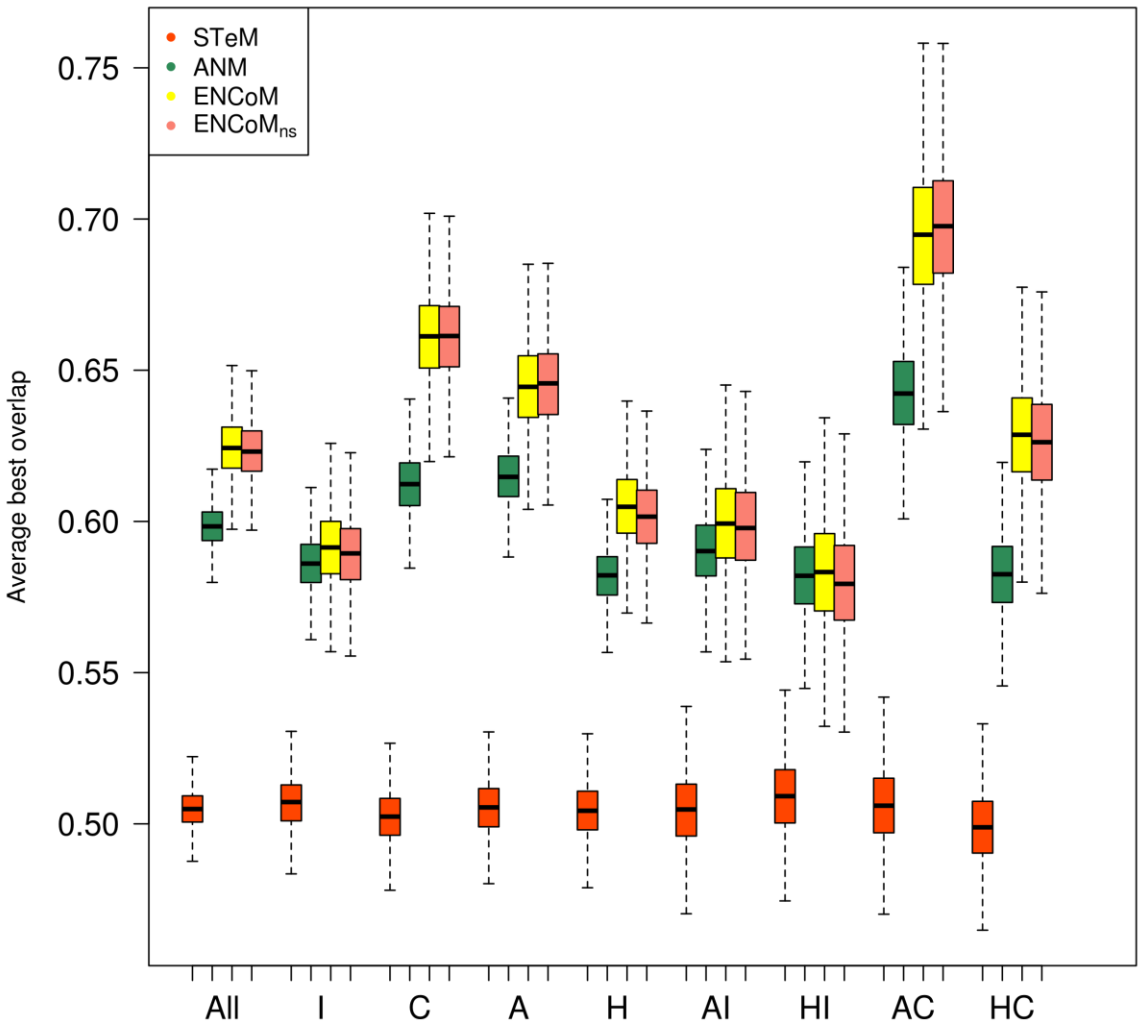


Figure 2. Prediction of domain motions. The best overlap found within the 10 slowest internal motion modes for different NMA models on domain movements. Legend: All (all cases, N=248), I (Independent movements, N=130), C (Coupled movements, N=117), A (apo form, N=124) and H (holo form, N=124). AI (apo form independent, N=130), HI (holo independent, N=130), AC (apo coupled, N=116) and HC (holo coupled, N=116). Box plots generated from 10000 resampling bootstrapping iterations. ENCoM/ENCoMns outperform ANM and STeM on all types of motion. doi:10.1371/journal.pcbi.1003569.g002

The analysis of overlap for ANM, STeM, ENCoM and ENCoMns was performed using the Protein Structural Change Database (PSCDB) [50], which contains 839 pairs of protein structures undergoing conformational change upon ligand binding. The authors classify those changes into seven types: coupled domain motions (59 entries), independent domain motions (70 entries), coupled local motions (125 entries), independent local motions (135 entries), burying ligand motions (104 entries), no significant motion (311 entries) and other type of motions (35 entries). The independent movements are movements that don't affect the binding pocket, while dependent movements are necessary to accommodate ligands in the pose found in the bound (holo) form. Burying movements are associated with a significant change of the solvent accessible surfaces of the ligand, but with small structural changes (backbone RMSD variation lower than 1 Å). Despite differentiating between types of movements based on the ligands, the ligands were not used as part of the normal mode analysis. Since side-chain movements associated to the burying movements cannot be predicted with coarse-grained NMA methods, we restrict the analysis to domain and loop movements [51] as these involve backbone movements amenable to analysis using coarse grained NMA methods. For practical purposes, in order to simplify the calculations in this large-scale analysis, NMR structures were not considered. It is worth stressing however that all NMA methods presented here don't have any restriction with respect to the structure determination method and can also be used with modeled structures. A total of 736 conformational changes, half representing apo to holo changes and the other half holo to apo (in total 368 entries from PSCDB) are used in this study (Supplementary Table S2).

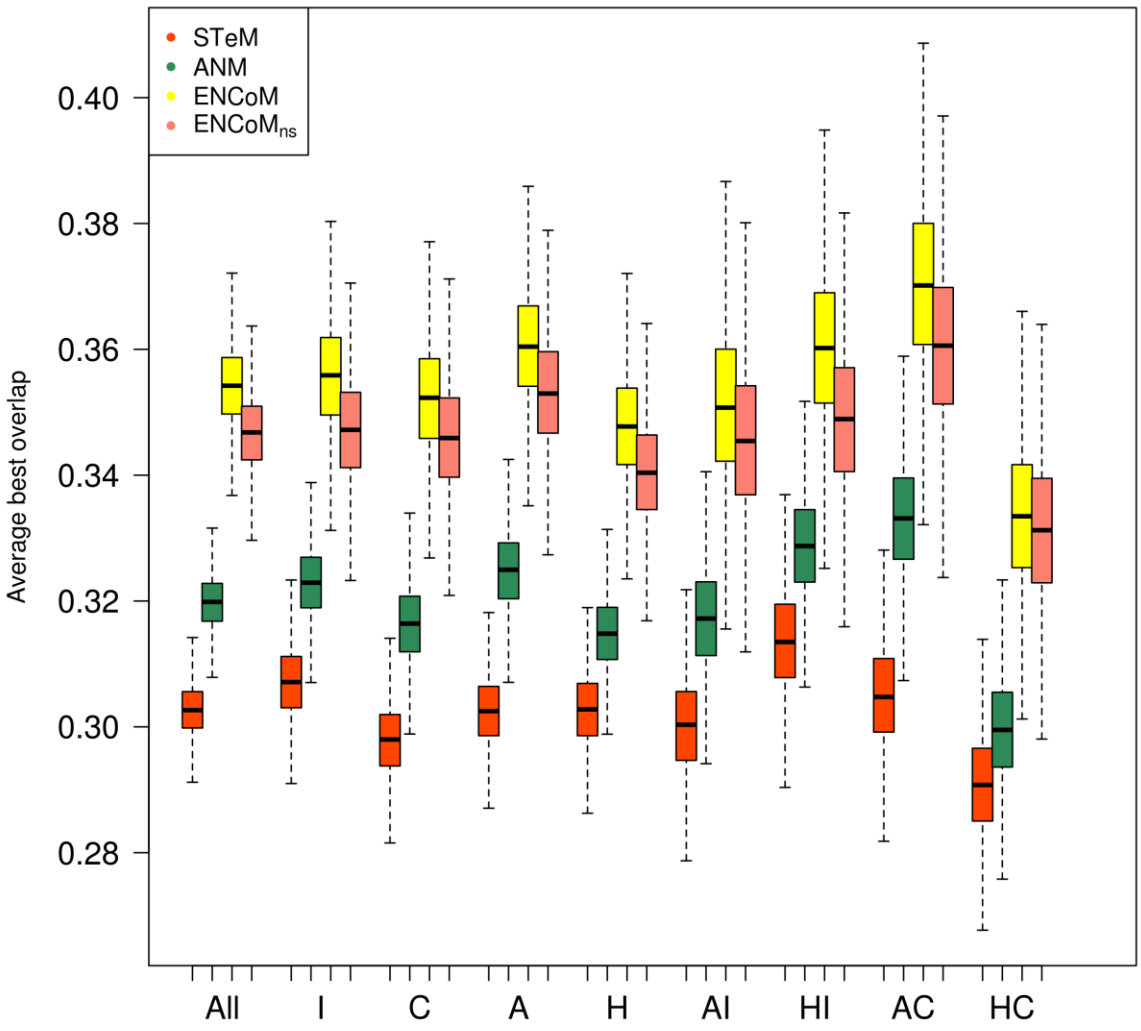


Figure 3. Prediction of loop conformational change. The best overlap found within the 10 slowest internal motion modes for different NMA models on loop movements. Acronyms are the same as in Figure 2. Number of cases: All (488), I (252), C (236), A (244) and H (244). AI (126), HI (126), AC (118) and HC (118). Box plots generated from 10000 resampling bootstrapping iterations. ENCoM/ENCoMns outperform ANM and STeM on all types of motion. The prediction of loop motions is much harder and here the difference between ENCoM and ENCoMns are more pronounced. doi:10.1371/journal.pcbi.1003569.g003

Overlap calculations were performed from the unbound (apo) form to the bound form (holo) and from the bound form to the unbound form. Bootstrapped results based onto the

best overlap found within the first 10 slowest modes [52,53] for the different types of conformational changes, domain or loop are shown in figures 2 and 3 respectively. In each case a set of box-plots represent the performance of the four methods being compared, namely STeM, ANM, ENCoM and ENCoMns. The left-most set of box-plots represents the average over all data while subsequent sets represent distinct subsets of the dataset as labeled. The first observation (comparing Figures 2 and 3) is that all tested NMA models show higher average overlaps for domain movements (Figure 2) than loop movements (Figure 3). This confirms earlier observations that NMA methods capture essential cooperative global (delocalized) movements associated with domain movements [51]. Loop movements on the other hand are likely to come about from a more fine tuned combination of normal mode amplitudes than what can be adequately described with a single eigenvector as measured by the overlap.

The second observation is that STeM performs quite poorly compared to other methods irrespective of the type of movement (domain or loop). This is somewhat surprising when one compares with ENCoM or ENCoMns considering how similar the potentials are. This suggests that the modulation of interactions by the surface area in contact (the bij terms in Equation 1) of the corresponding side-chains as well as the specific parameters used are crucial.

Focusing for a moment on domain movements (Figure 2), ENCoM/ENCoMns outperform all other methods for domain movements in general as well as for every sub category of types of motions therein. Independent movements show lower overlaps than coupled ones, a fraction of those movements may not be biologically relevant due to crystal packing. Interestingly, while there are no differences between the overlap for independent movements starting from the apo or holo forms, this is not the case for coupled movements. In this case (right-most two sets in Figure 2), it is easier to use the apo (unbound) form to predict the holo (bound) form, suggesting that the lower packing in the apo form (as this are frequently more open) generates eigenvectors that favor a more comprehensive exploration of conformational space.

Lastly, with respect to loop movements (Figure 3), while it is more difficult to obtain good overlaps irrespective of the method or type of structure used, overall ENCoM/ENCoMns again outperforms ANM. Some of the same patterns observed for domain movements are repeated here. For example, the higher overlap for coupled apo versus holo movements.

We observe that ENCoMns consistently performs almost as well as ENCoM irrespective of the type of motion used (all sets in Figures 2 and 3). As side-chain conformations in crystal structures tent to minimize unfavourable interactions, the modulation of interactions by atom types that differentiate ENCoM from ENCoMns plays as minor but still positive role.

Prediction of the effect of mutations

Normal mode resonance frequencies (eigenvalues) are related to vibrational entropy [54,55] (see methods). Therefore, it is reasonable to assume that the information contained in the eigenvectors can be used to infer differences in protein stability between two structures differing by a mutation under the assumption that the mutation does not drastically affect the equilibrium structure. A mutation may affect stability due to an increase in the entropy of the folded state by lowering its resonance frequencies, thus making more microstates accessible around the equilibrium.

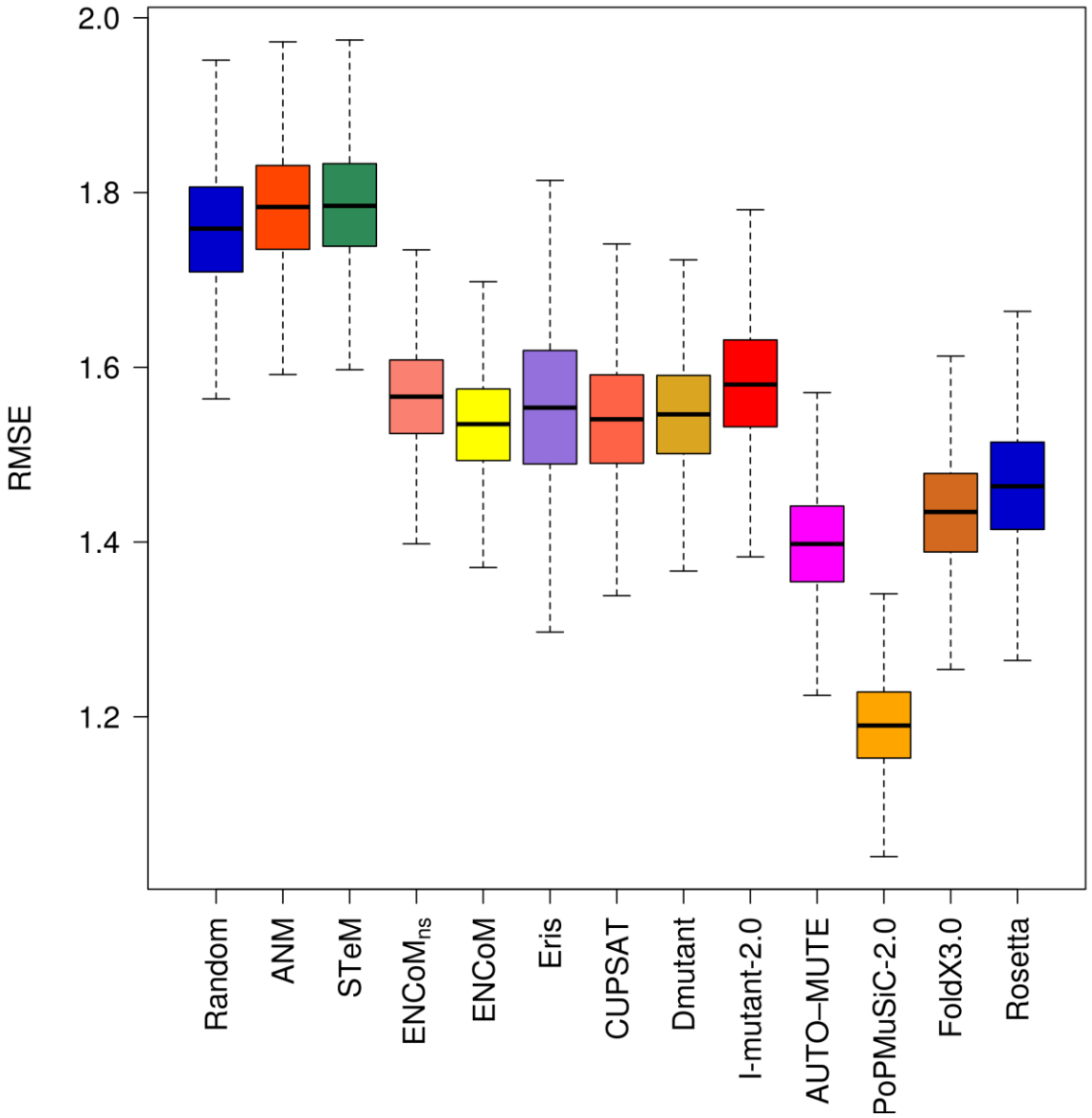


Figure 4. Root Mean Square Error (RMSE) on the prediction of the effect of mutations. RMSE of the linear regression through the origin between experimental and predicted variations in free energy variations (ΔΔG). Box plots generated from 10000 resampling bootstrapping iterations on the entire PoPMuSiC-2.0 dataset (N=303). With the caveat that the dataset is biased towards destabilizing mutations, ENCoM/ENCoMns predict the effect of mutations with similar overall RMSE as most other methods. doi:10.1371/journal.pcbi.1003569.g004

We utilize experimental data from the ProTherm database [56] on the thermodynamic effect of mutations to validate the use of ENCoM to predict protein stability. Here we benefit from the manual curation efforts previously performed to generate a non- redundant

subset of ProTherm comprising 303 mutations used for the validation of the PoPMuSiC-2.0 [57]. The dataset contains 45 stabilizing mutations ($\Delta\Delta G < -0.5$ kcal/mol), 84 neutral mutations ($\Delta\Delta G$ [-0.5,0.5] kcal/mol) and 174 destabilizing mutations ($\Delta\Delta G > 0.5$ kcal/mol) (Supplementary Table S2). Each protein in the dataset has at least one structure in the PDB database [58]. As we calculate the eigenvectors in the mutated form we require model structures of the mutants. We generate such models using Modeller [59] and are thus assume that the mutation does not drastically affect the structure. Mutations were generated using the mutated.py script from the standard Modeller software distribution. Modeller utilizes a two-pass minimization. The first one optimizes only the mutated residue, with the rest of the protein fixed. The second pass optimizes the non-mutated neighboring atoms. It is important to stress that our goal is to model the mutated protein as accurately as possible and thus using any method that unrealistically holds the backbone fixed to model the mutant form would be an unnecessary simplification. We observe a backbone RMSD for the whole protein of 0.01 ± 0.01 Å on average. Considering that RMSD is a global measure that could mask more drastic local backbone rearrangements, we also calculated the average maximum Cα displacement but with a value of 0.13 ± 0.12 Å we are confident that while not fixed, backbone rearrangements are indeed minimal.

In the present work we predict the effect of mutations (Equation 6) for ENCoM, ENCoMns, ANM and STeM and compare the results to existing methods for the prediction of the effect of mutations using the PoPMuSiC-2.0 dataset above. We compare our results to those reported by Dehouck et al. [57] for different existing techniques: CUPSAT, a Boltzman mean-force potential [60]; DMutant, an atom-based distance potential [61]; PoPMuSiC-2.0, a neural network based method [57]; Eris, a force field based approach [62]; I-Mutant 2.0, a support vector machine method [63]; and AUTO-MUTE, a knowledge-based four-body potential [64]. We used the same dataset to generate the data for FoldX 3.0, an empirical full atom force-field [65] and Rosetta [66], based on the knowledge based Rosetta energy function. A negative control model was build with a randomized reshuffling of the experimental data. Figure 4 presents RMSE results for each model. The raw data for the 303 mutations is available in Supplementary Table S3.

ANM and STeM are as good as the random model when considering all types of mutations together (Figure 4). This is not surprising as the potentials used ANM and STeM are exclusively geometry-based and are thus agnostic to sequence. ENCoMns, ENCoM, Eris, CUPSAT, DMutant and I-Mutant 2.0 give similar results and predict significantly better than the random model. AUTO-MUTE, FoldX 3.0, Rosetta and in particular PoPMuSiC-2.0 outperform all of the other models.

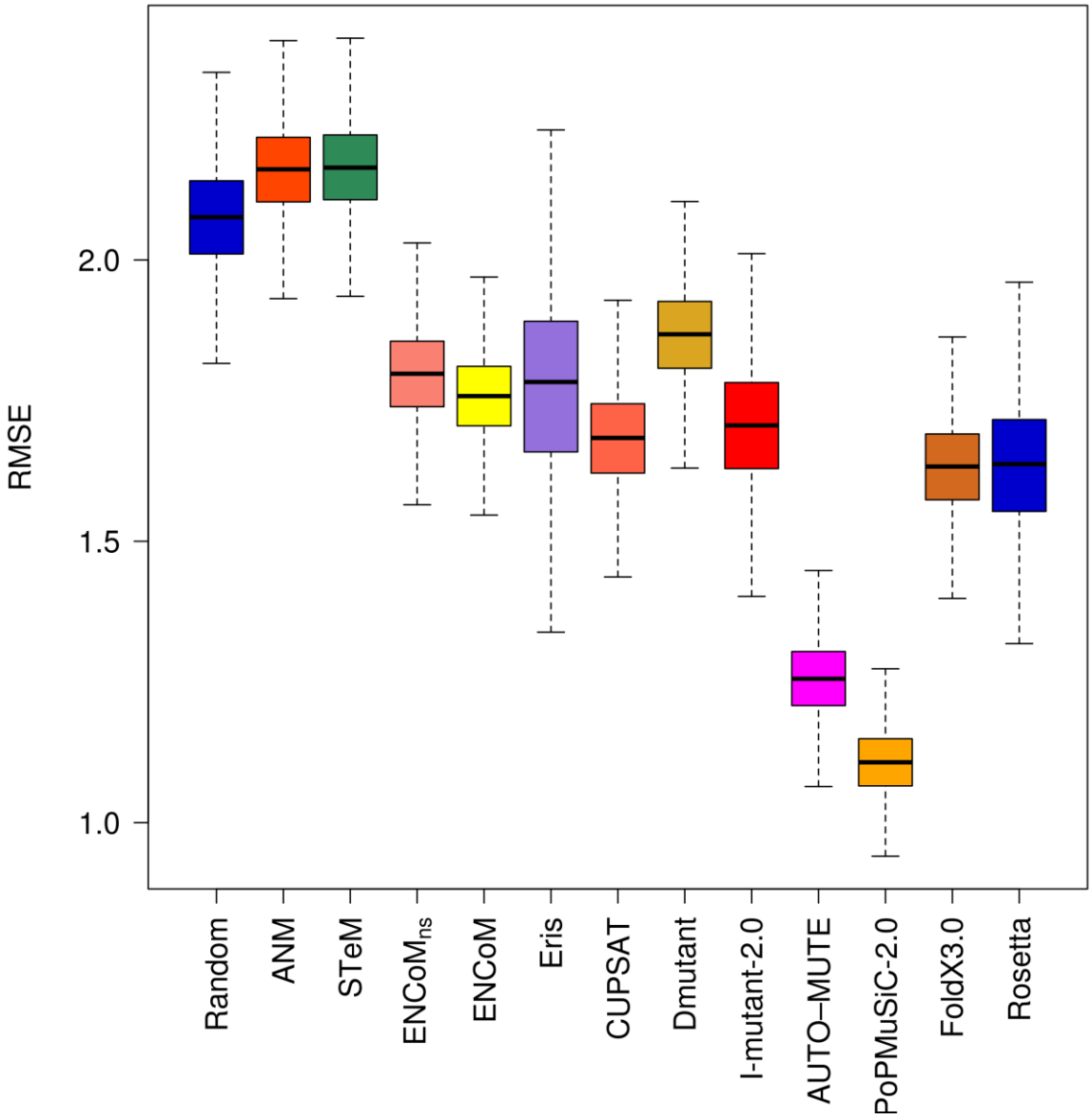


Figure 5. Root Mean Square Error (RMSE) on the prediction of the effect of destabilizing mutations. RMSE of the linear regression through the origin between

experimental and predicted variations in free energy variations (ΔΔG). Box plots generated from 10000 resampling bootstrapping iterations on the subset of destabilizing mutations (experimental ΔΔG > 0.5 kcal/mol) in the PoPMuSiC-2.0 dataset (N=174). The results for destabilizing mutations mirror to a great extent those for the entire dataset given that they represent 57% of the entire dataset. doi:10.1371/journal.pcbi.1003569.g005

The RMSE values for the subset of 174 destabilizing mutations (Figure 5) shows similar trends as the whole dataset with the exceptions of DMutant losing performance and PoPMuSiC-2.0 as well as AUTO-MUTE gaining performance compared to the others. It is important to stress that the low RMSE of PoPMuSiC-2.0 on the overall dataset is to a great extent due to its ability to predict destabilizing mutations.

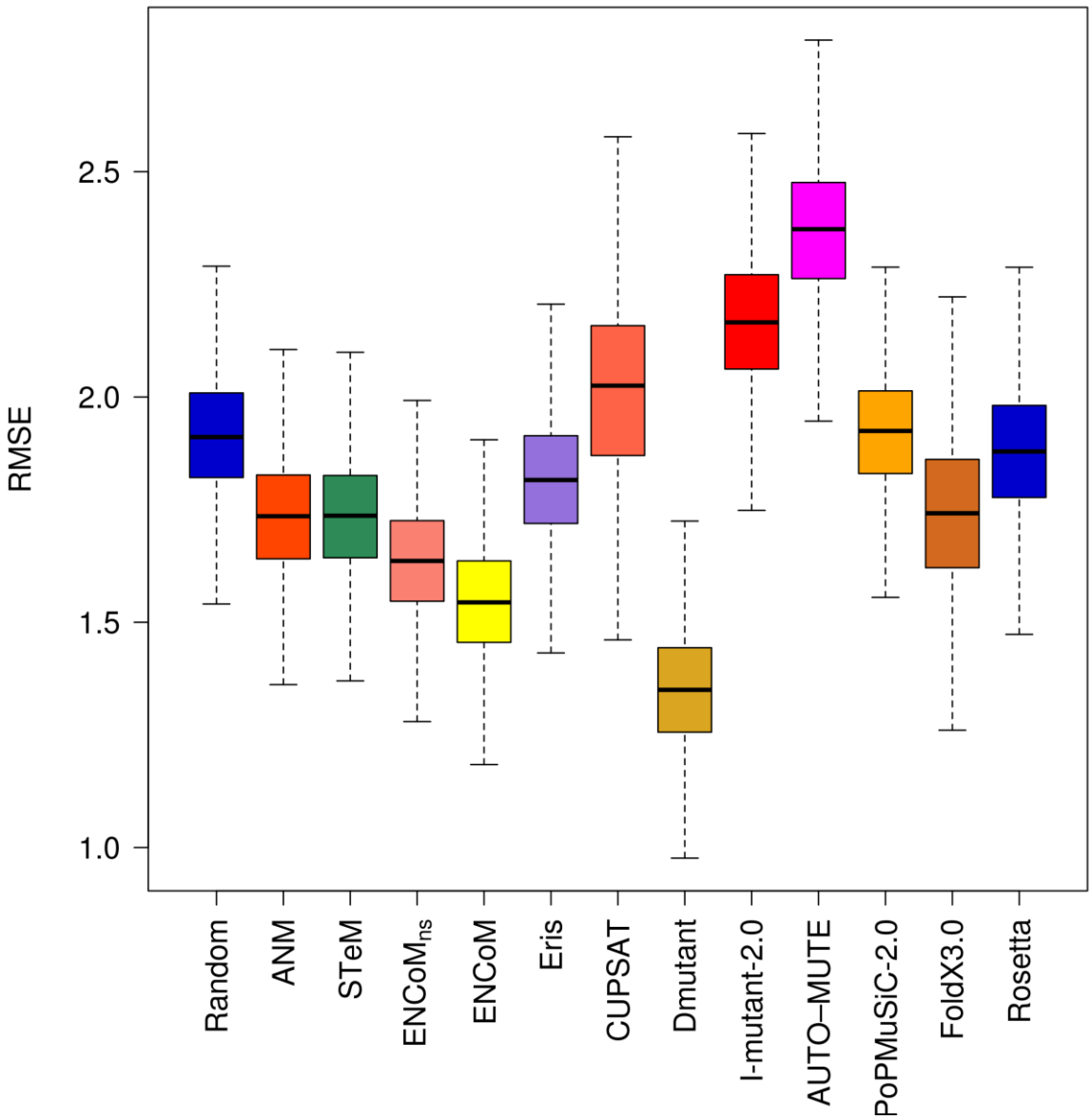


Figure 6. Root Mean Square Error (RMSE) on the prediction of the effect of stabilizing mutations. RMSE of the linear regression through the origin between experimental and predicted variations in free energy variations (ΔΔG). Box plots generated from 10000 resampling bootstrapping iterations on the subset of stabilizing mutations (experimental ΔΔG < -0.5 kcal/mol) in the PoPMuSiC-2.0 dataset (N=45). With the exception of DMutant and ENCoM, most methods are not substantially better than random and some are substantially worst for stabilizing mutations. doi:10.1371/journal.pcbi.1003569.g006

The subset of 45 stabilizing mutations (Figure 6) gives completely different results as those obtained for destabilizing mutations. AUTO-MUTE, Rosetta, FoldX 3.0 and PoPMuSiC-

2.0 that outperformed all of the models on the whole dataset or the destabilizing mutations dataset cannot predict better than the random model. This is also true for CUPSAT, I-Mutant 2.0 and Eris. ENCoM and DMutant are the only models with significantly better than random RMSE values for the prediction of stabilizing mutations.

ANM and STeM outperform all models on the neutral mutations (Figure 7). All other models fail to predict neutral mutations any better than random. While the accuracy of ANM and STeM to predict neutral mutations may seem surprising at first, it is in fact an artifact of the methodology. As the wild type or mutated structures are assumed to maintain the same general backbone structure, the eigenvectors/eigenvalues calculated with ANM or STeM will always be extremely similar for wild type and mutant forms. Any differences will arise as a result of small variations in backbone conformation produced by Modeller. As such, ANM and STeM predict almost every mutation as neutral, explaining their high success in this case.

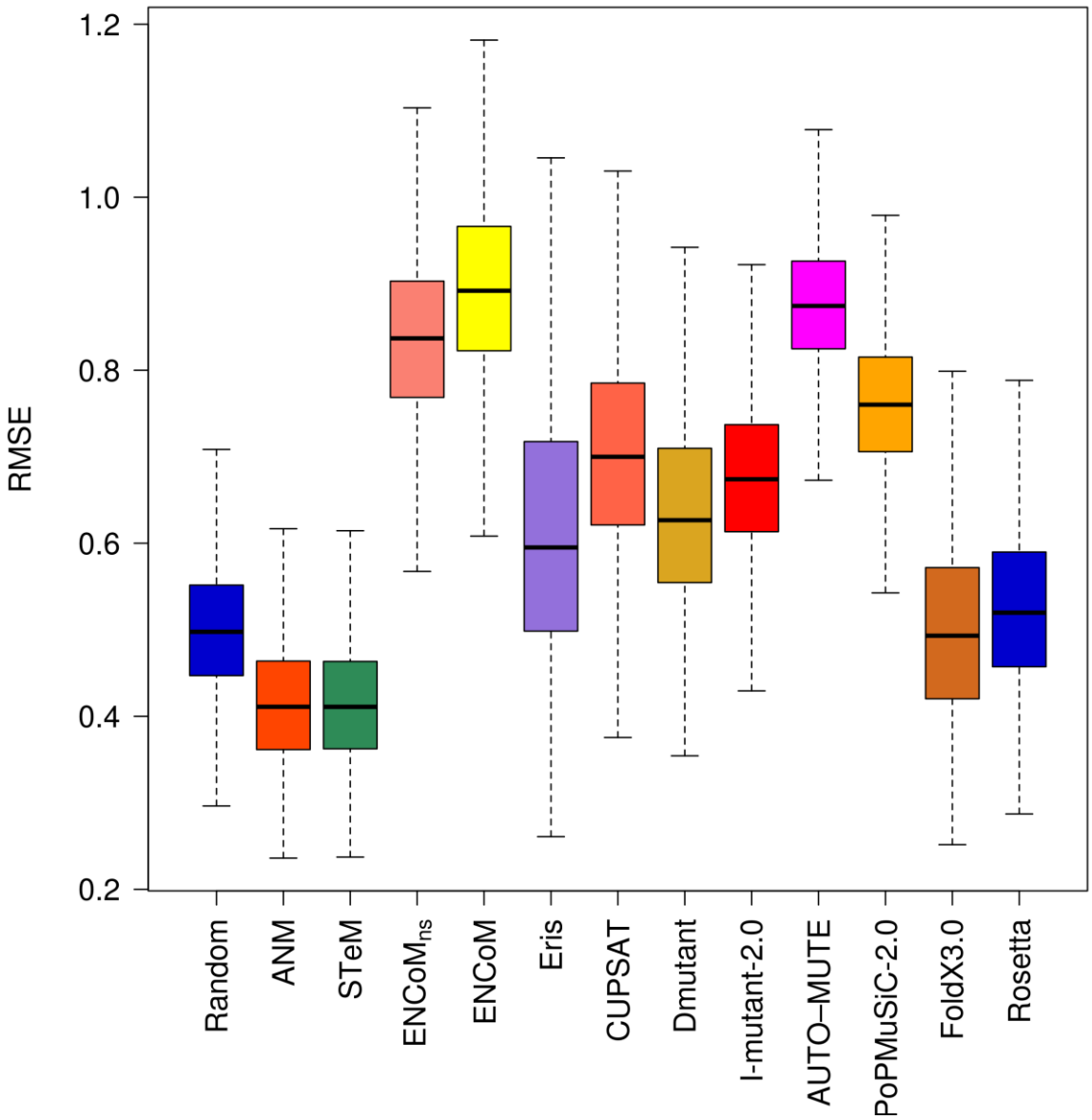


Figure 7. Root Mean Square Error (RMSE) on the prediction of the effect of neutral mutations. RMSE of the linear regression through the origin between experimental and predicted variations in free energy variations (ΔΔG). Box plots generated from 10000 resampling bootstrapping iterations on the subset of neutral mutations (experimental ΔΔG=[20.5,0.5] kcal/mol) in the PoPMuSiC-2.0 dataset (N=84). Apart from ANM and STeM that predict most mutations as neutral as an artifact, no method is better than the reshuffled random model at predicting neutral mutations. doi:10.1371/journal.pcbi.1003569.g007

At first glance, the comparison of ENCoMns and ENCoM could suggest that a large part of the effect observed come from a consideration of the total area in contact and not the

specific types of amino acids in contact. However, the side chains in contact are already in conformations that minimize unfavourable contacts to the extent that is acceptable in reality (in the experimental structure) or as a result of the energy minimization performed by Modeller for the mutant form given the local environments. The fact that ENCoM is able to improve on ENCoMns is the actual surprising result and points to the existence of frustration in molecular interactions [67].

Considering that none of the existing models can reasonably predict neutral mutations, the only models that achieve a certain balance in predicting both destabilizing as well as stabilizing mutations better than random and with low bias are ENCoM and DMutant.

The analysis of the performance of ENCoM in the prediction of different types of mutations in terms of amino acid properties shows that mutations from small (ANDCGPSV) to big (others) residues are the most accurately predicted followed by mutations between non-polar or aromatic residues (ACGILMFPWV). ENCoM performs poorly on exposed residues (defined as having more than 30% of the surface area exposed to solvent) (Figure 8).

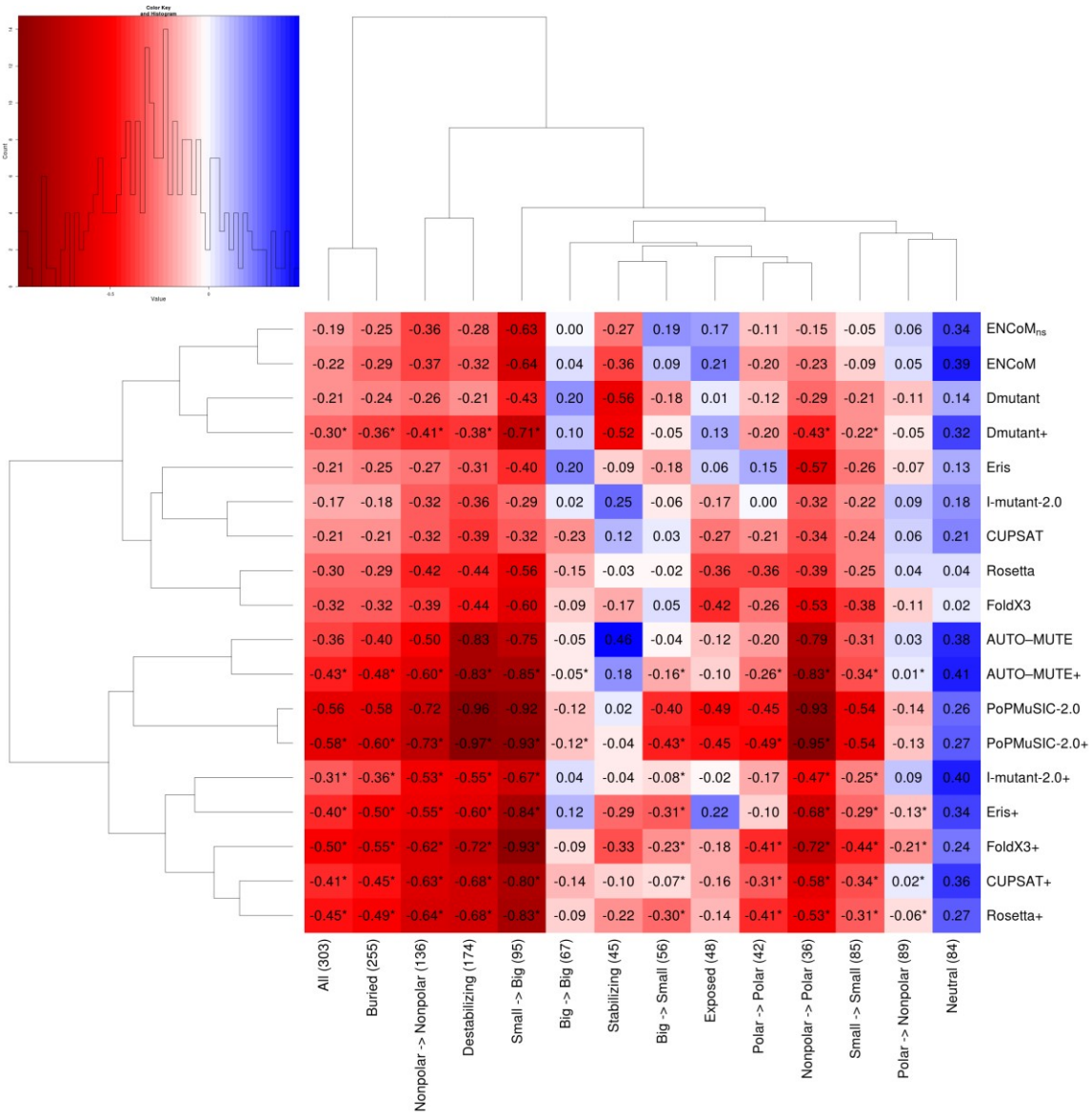


Figure 8. Average RMSE differences for different types of mutations, prediction methods and their combinations with ENCoM. We calculate the RMSE difference (Eq. 8) for different subsets of the data (columns) and different methods (rows). Methods followed by a '+' denote linear combinations of the named method with ENCoM. The heatmap values (shown within cells) denote the bootstrapped (10000 iterations) average RMSE difference with respect to random and are color coded according to the map on the upper left (lower values in red signify better predictions). For example the values in the leftmost column representing all data comes from the subtraction of the averages in the corresponding box plots in Figure 4 for the non-combined methods and the random model. Different subsets of types of mutations are shown according to certain properties of the amino-acids or residues involved: buried (less than 30% solvent-exposed surface area) or exposed (otherwise), small (A,N,D,C, G, P,S, and V) or big (otherwise), polar

(R,N,D,E,Q,H,K,S,T and Y) or non-polar including hydrophobic and aromatic (otherwise). Both methods and subsets of the data are clustered according to similarities in the RMSE difference profiles. Cases where the combination of a method with ENCoM is beneficial, i.e., the RMSE difference is lower than either method in isolation are denoted by a '*' next to their values within the cell. doi:10.1371/journal.pcbi.1003569.g008

Predictions based on linear combination of models

It may in principle be possible to find particular linear combinations of ENCoM and other methods that further improve predictions given the widely different (and potentially complementary) nature of the various approaches with respect to ENCoM. We performed linear regressions to find parameters involving ENCoM and each of the other methods in turn that maximize the RMSE difference between the combined models and random predictions (Eq. 8). When considering all types of mutations together, all mixed models perform better than either model individually (left-most column in Figure 8). By definition, a mixed model cannot perform worse than the better of the two models individually. The contribution of ENCoM to the improved performance of the combined model varies according to the model. The ratio of the relative contributions $\lambda_{ENCoM}=\lambda_i$ (in parenthesis), broadly classifies the methods into three categories: 1. Methods where ENCoM contributes highly, including I-Mutant (1.18 ± 0.25), DMutant (1.07 ± 0.31), CUPSAT (1.02 ± 0.09) and Eris (1.02 ± 0.11); 2. Methods where the contribution of ENCoM is smaller than that of the other method but still significant, including Rosetta (0.90 ± 0.11) and FoldX3 (0.89 ± 0.08); and finally methods where the addition of ENCoM have a small beneficial effect, including in this class Automute (0.69 ± 0.13) and especially PoPMuSIC (0.18 ± 0.10). The left-hand side dendrogram in Figure 8 clusters the methods according to their overall accuracy relative to random based on the entire profile of ΔRMSE predictions (Eq. 8) for different subsets of the data (columns) according to the type of mutations being predicted. This clustering of methods shows that the relative position of the methods is maintained throughout except for a small rearrangement due to changes in the predictions for CUPSAT. This result suggests that the contribution from the combination of ENCoM to other methods is uniform irrespective of the type of mutations studied.

Self-consistency in the prediction of the effect of mutations

One basic requirement for a system that predicts the effect of mutations on stability is that it should be self-consistent, both unbiased and with small error with respect to the prediction of the forward or back mutations as reported by Thiltgen et al. [68]. The authors built a non-redundant set of 65 pairs of PDB structures containing single mutations (called form A and form B) and utilized different models to predict the effect of each mutation going from the form A to form B and back. From a thermodynamic point of view, the predicted variation in free energy variation should be of the same magnitude for the forward or back mutations, $\Delta\Delta G_{A,B} = -\Delta\Delta G_{B,A}$. Using the Thiltgen dataset we performed a similar analysis for ENCoM, ENCoM$_{ns}$, ANM, STeM, CUPSAT, DMutant, PoPMuSiC-2.0 and a random model (Gaussian prediction with unitary standard deviation). For the remaining methods (Rosetta, Eris and I-Mutant) we utilize the data provided by Thiltgen. We removed three cases involving prolines as such cases produce backbone alterations. Furthermore, PoPMuSiC-2.0 failed to return results for five cases. The final dataset therefore contains 57 pairs (Supplementary Table S4). The CUPSAT and AUTO-MUTE servers failed to predict 25 and 32 cases respectively. As these failure rates are significant considering the size of the dataset, we prefer to not include these two methods in figures 9 and 10 (the remaining cases appear however in Supplementary Table S4).

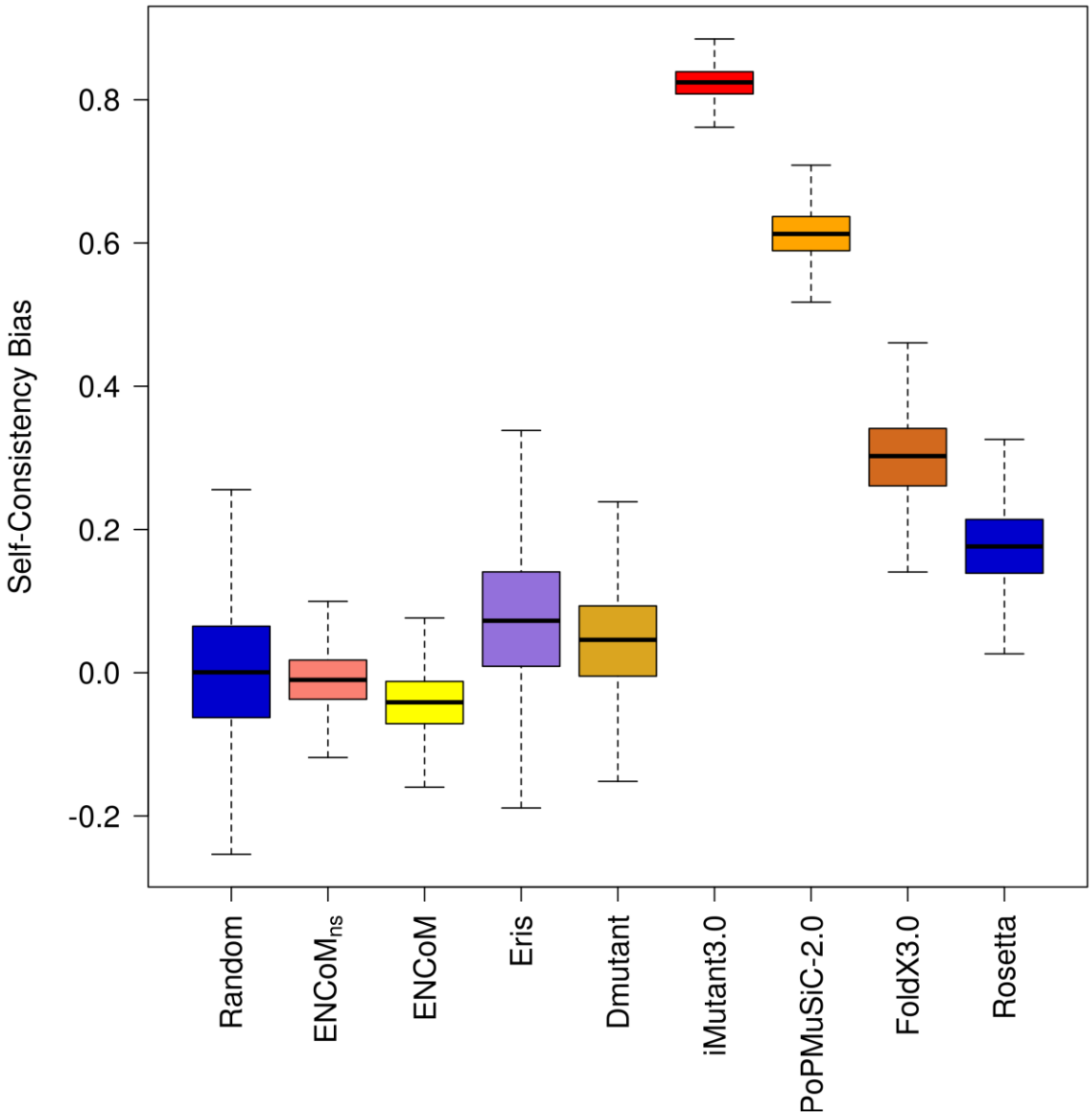


Figure 9. Self-consistency bias. The bias quantifies the tendency of a method to predict more accurately mutations in one direction than in the opposite. Machine learning based methods in particular show a high bias. ENCoM/ENCoMns have low bias. doi:10.1371/journal.pcbi.1003569.g009

The results in figure 9 show that compared to the random model (a positive control in this case), Rosetta and FoldX 3.0 show moderate bias while PoPMuSiC-2.0 and I-Mutant show significant bias. All biased methods are biased toward the prediction of destabilizing mutations (data not shown) in agreement with the results in Figure 3. DMutant, Eris, ENCoM and ENCoMns are the only models with bias comparable to that of the random

model (the positive control in this experiment). ENCoM, ENCoMns, and to a lesser extent Rosetta and DMutant have lower errors than the random model (Figure 10). All other methods display an error equal or higher than that of the random model. ENCoM and ENCoMns vastly outperform all the others models in terms of error. Lastly, STeM and ANM show low and moderate biases respectively and errors equivalent to random (data not shown) but as mentioned, these methods cannot be used for the prediction of mutations (other than neutral mutations as an artefact).

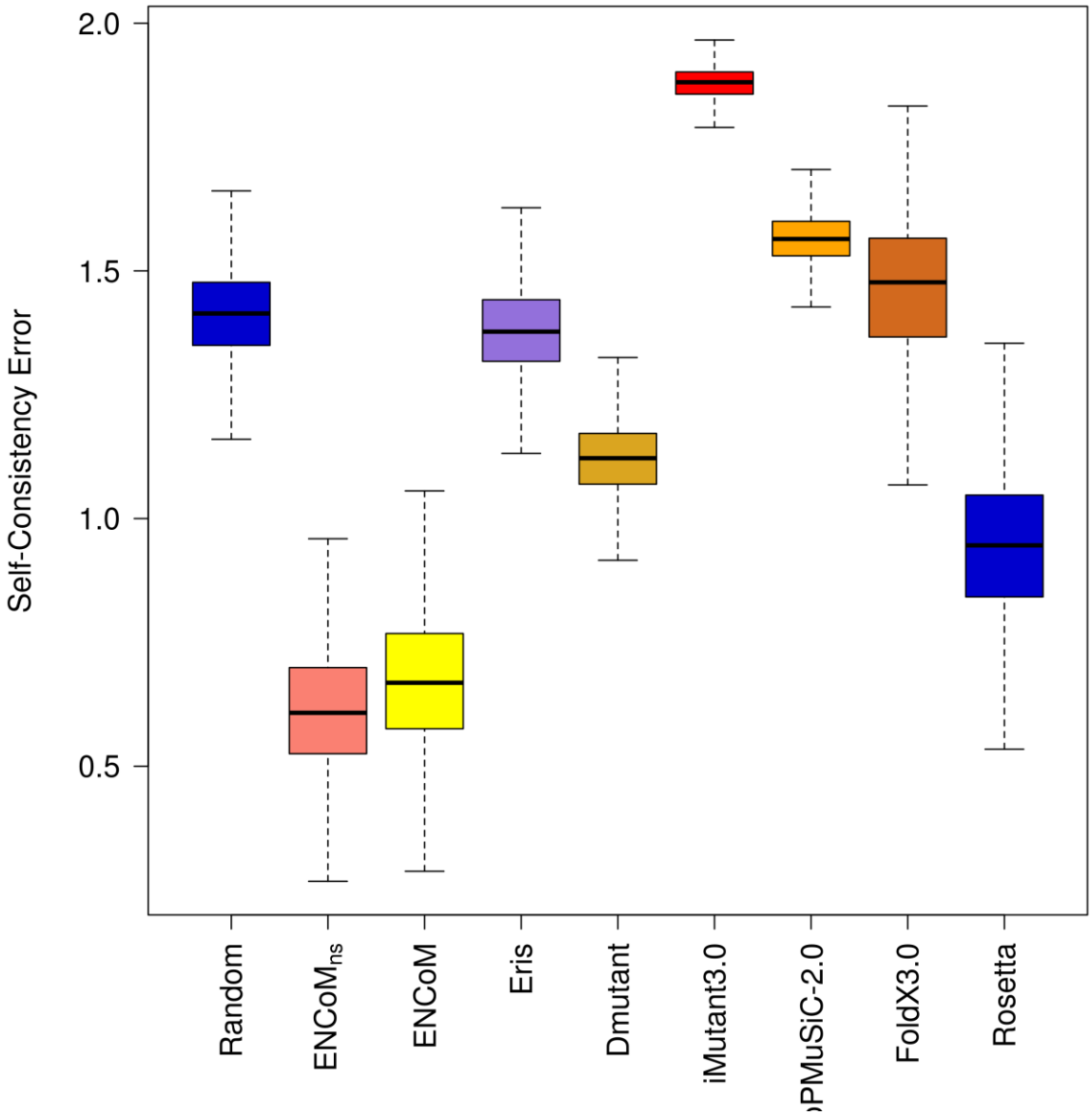


Figure 10. Self-consistency error. The error calculated the magnitude of the biases in the prediction of forward and back mutations. Box plots were generated from 10000 resampling bootstrapping iterations for the 57 proteins pairs in the Thiltgen dataset. ENCoM/ENCoMns are the methods with lowest self-consistency errors. doi:10.1371/journal.pcbi.1003569.g010

Prediction of NMR S2 order parameter differences

Mutations may not only affect protein stability but also protein function. While experimental data is less abundant, one protein in particular, dihydrofolate reductase (DHFR) from *E. coli*, has been widely used experimentally to understand this relationship [69,70]. Recently, Boehr et al. [47] have analyzed the effect of the G121V mutation on protein dynamics in DHFR by NMR spectroscopy. This mutation is located 15 A˚ away from the binding site but reduces enzyme catalysis by 200 fold with negligible effect on protein stability (0.70 kcal/mol). The authors evaluated, among many other parameters, the S2 parameter of the folic acid bound form for the wild type and mutated forms and identify the regions where the mutation affects flexibility. We calculated b-factor differences (Equation 4) between the folate-bound wild type (PDB ID 1RX7) and the G121V mutant (modeled with Modeller) forms of DHFR (Supplementary Table S5). We obtain a good agreement (Pearson correlation = 0.61) between our predicted b-factor difference and S2 differences (Figure 11). As mentioned earlier, the overall correlation of 0.54 in the prediction of b-factors (Figure 1) appears at least in this case to be sufficient to capture essential functional information.

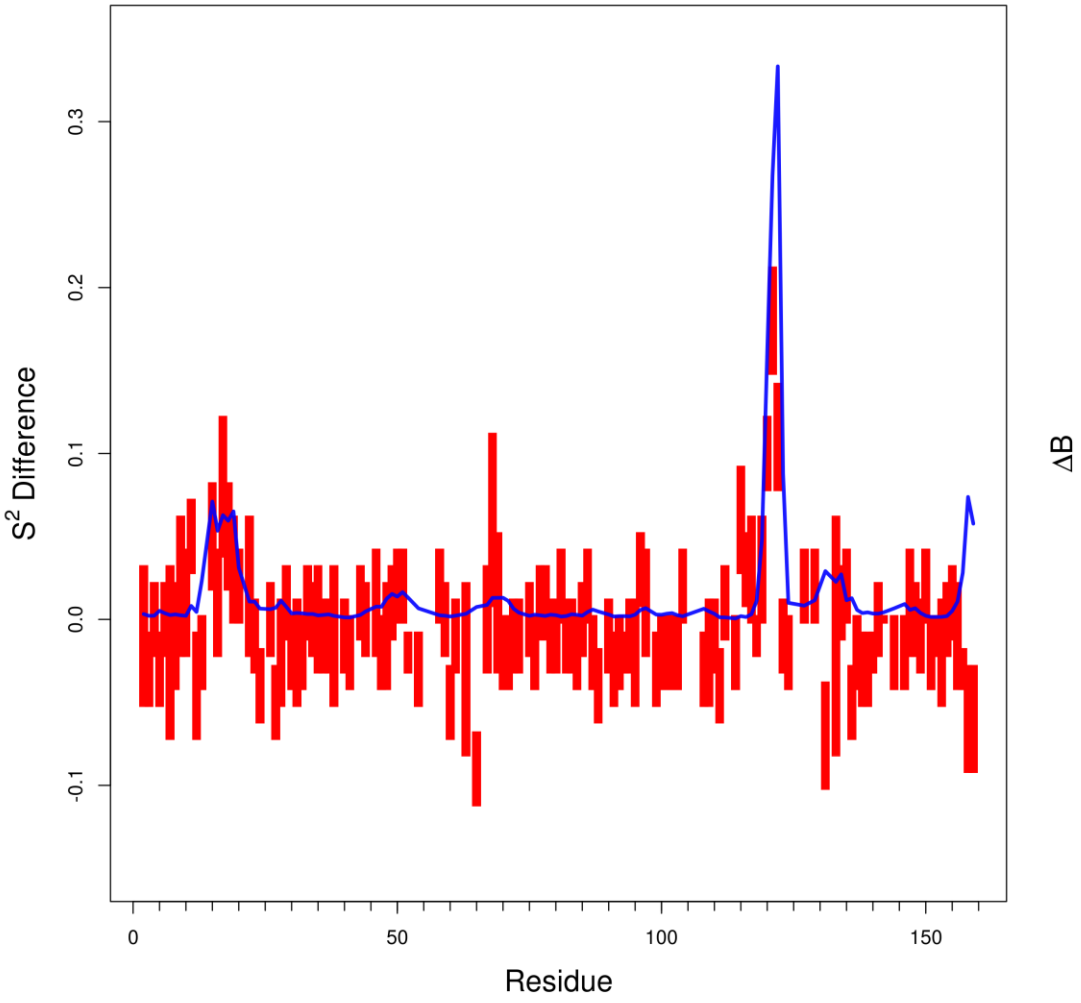


Figure 11. Prediction of NMR S2 order parameter differences for the G121V mutation on DHFR from E. Coli Experimental values (red, the bar represents the experimental error) are compared to the inverse normalized predicted b-factors differences showing a Pearson correlation coefficient of 0.6. doi:10.1371/journal.pcbi.1003569.g011

Discussion

Our results show that a small modification of the long-range interaction term in the potential energy function of STeM had an important positive impact on the model. This small change improves the method in comparison to existing NMA methods in the traditional areas such as the prediction of b-factors and conformational sampling (overlap) where coarse-grained normal mode analysis are applied. More importantly however, it

opens an entire new area of application to coarse-grained normal mode analysis methods. Specifically ENCoM is the first coarse-grained normal-mode analysis method that permits to take in consideration the specific sequence of the protein in addition to the geometry. This is introduced through a modification in the long-range interactions to account for types of atoms in contact modulated by their surface in contact. As a validation of the approach we explored the ability of the method to predict the effect of mutations in protein stability. In doing so we created the first entropy-based methodology to predict the effect of mutations on the thermodynamic stability of proteins. This methodology is entirely orthogonal to existing methods that are either machine learning or enthalpy based. Not only the approach is novel but also the method performs extremely favourably compared to other methods when viewed in terms of both error and bias.

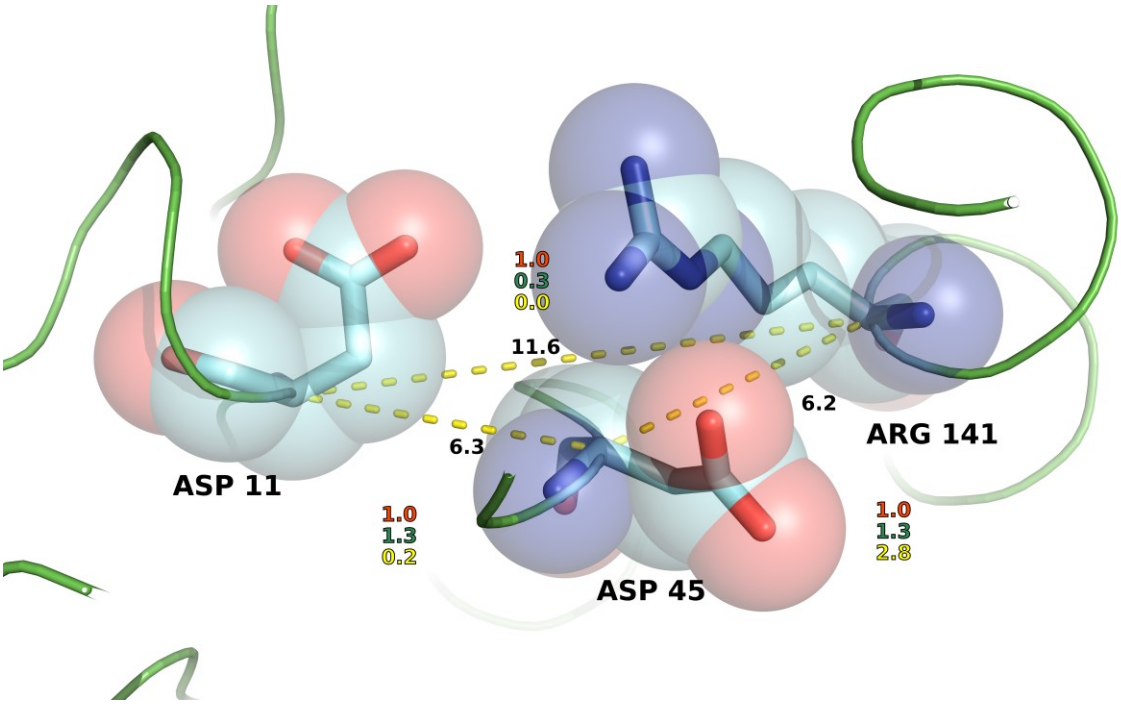


Figure 12. Illustration of the representation of inter-residue interactions by the different NMA methods. The figure shows three amino acids (D11, D45 and R141) from the M. tuberculosis ribose-5-phospate isomerase (PDB ID=2VVO). The distances between alpha carbons are shown with the yellow dotted lines and labeled in black. Interaction strengths relative to ANM (in red) are shown for STeM (green) and ENCoM (yellow). ANM treats all pairs as equal while STeM treat equally D11 and R141 with respect to D45

even though the interaction between D45 and R141 is much stronger by virtue of side-chain interactions, as correctly described in ENCoM. doi:10.1371/journal.pcbi.1003569.g012

As the approach taken in ENCoM is completely different from existing methods for the prediction of the effect of mutations on protein stability, a new opportunity arises to combine ENCoM with enthalpy and machine-learning methods. Unfortunately, we tried to create a naive method based on linear combinations of the predictions of ENCoM and the different methods presented without success, perhaps due to the large bias characteristic to the different methods.

To assess the relative importance of contact area and the modulation of interactions with atom types, we tested a model that has non-specific atom-type interactions (ENCoMns), this model is atom type insensitive, but is sensitive the orientation of side-chain atoms. While a large fraction of the observed effect can be attributed to surfaces in contact only, ENCoM is consistently better than ENCoMns, particularly at predicting destabilizing mutations where the possibility to accommodate unfavourable interactions is more restricted. We cannot however exclude the effect of the intrinsic difficulty in modeling destabilizing mutations. For stabilizing mutations, the near equivalence of ENCoM and ENCoMns may be explained in part by the successful energy minimization of the mutated side-chain performed by Modeller. ANM and STeM failed to predict the effect of mutations on the whole dataset. They were not expected to perform well because their respective potentials only take in account the position of alpha carbons (backbone geometry). As such ANM and STeM tend to predict mutations as neutral, explaining their excellent performance onto the neutral subset and failure otherwise. Our results suggest that surfaces in contact are essential in a coarse- grained NMA model to predict the effect of mutation and that the specific interactions between atom types is necessary to get more subtle results, particularly stabilizing mutations.

ENCoM is consistently better than ENCoMns in the prediction of loop or domain movements irrespective of the dependency of the coupling of this movement to ligand binding or the starting structure (apo or holo form) and both outperform ANM and STeM.

Our results corroborate previous work on a mix coarse- grained method adding a atomistic resolution to loops capable of improving the prediction of loop movements [40]. ENCoM performs considerably better than STeM throughout despite having very similar potentials, showing the importance of surfaces in contact in the prediction of movements. There is little difference between ENCoMns and ENCoM in the prediction of b-factors, but both perform worst than ANM, STeM and GNM. At least in the case of DHFR b-factor differences capture some essential characteristics of the system as calculated by NMR. However, one should be careful in placing too much emphasis on the validation of b-factor predictions using experimental data derived from crystals as these are affected to a great extent by rigid body motions within the crystal [71].

PoPMuSiC-2.0, AUTO-MUTE, FoldX 3.0 and Rosetta perform better than other models in the whole test dataset of mutations. However, the dataset consists of 15% stabilizing mutation, 57% of destabilizing and 28% of neutral mutations. When looking at each subset, machine learning or enthalpy based models failed to predict better than random on the stabilizing mutations subset. Biases in the dataset may have affected the training of machine-learning methods. For example the training set of PoPMuSiC-2.0 contains 2648 mutations in proportions that are similar to those in the testing set with 60%, 29% and 11% destabilizing, neutral and stabilizing mutations respectively. While it is true that most mutations tend to be destabilizing, if one is interested in detecting stabilizing mutations, a method over trained on destabilizing mutations will not meet expectations. Indeed, PoPMuSiC-2.0 and I-mutant the two machine learning based methods, have larger biases and errors than other methods in their predictions. Our method relies on a model structure of the mutant. As the modeling may fail to find the most stable side-chain conformation, it could have a bias toward giving slightly higher energies to the mutant. Notwithstanding this potential bias, ENCoM has the lowest error and bias. This may be a case where less is more as the coarse-grained nature of the method makes it also less sensitive to errors in modeling that may affect enthalpy-based methods to a greater extent. Finally, there is one more advantage in the approach taken in ENCoM. As the network model is a global connected model it considers indirectly the entire protein, while in existing enthalpy or machine-learning methods the effect of a mutation is calculated mostly from a local point of view.

The prediction of the thermodynamic effect of mutations is very important to understand disease-causing mutations as well as in protein engineering. With respect to human diseases, and particularly speaking of cancer mutations, one of the factors that may lead to tumour suppressor or oncogenic mutations is their effect on stability (the authors thank Gaddy Getz from the Broad Institute for first introducing us to this hypothesis). Specifically, destabilizing mutations in tumour suppressor genes or alternatively stabilizing mutations in oncogenes may be driver mutations in cancer. Therefore the prediction of stabilizing mutations may be very important to predict driver mutations in oncogenes. Likewise, in protein engineering, one major goal is that of improving protein stability with the prediction of stabilizing mutations. Such mutations may be useful not as the final goal (for purification or industrial purposes) but also to create a 'stability buffer' that permits the introduction of potentially destabilizing additional mutations that may be relevant to create the intended new function.

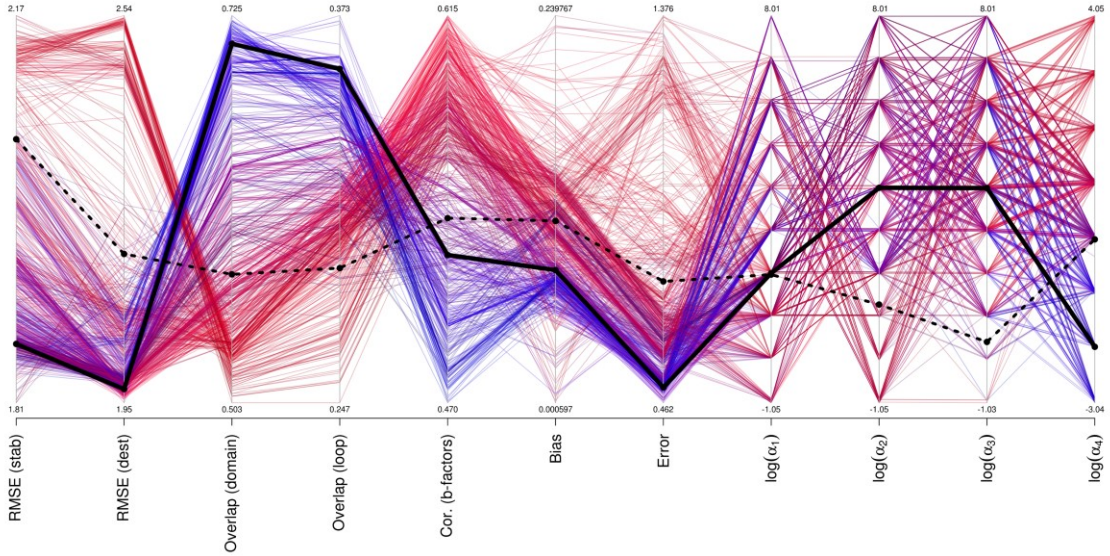


Figure 13. Performance of different parameter sets on the prediction of mutations, b-factors and motions. We present as a parallel plot the bootstrapped median RMSE for stabilizing and destabilizing mutation, average best overlap for domains and loop movements as well as self- consistency bias and errors. In the right-most four columns with include the logarithm of the 4 alpha variables. Different parameter sets are colored based on b-factors correlations (red gradient) or domain movement overlaps (blue gradient). The black line represent the specific set of parameters used in ENCoM while the dashed line represents the values for ENCoM using the set of alpha parameters employed in STeM.

There is a dichotomy in parameter space such that most sets of parameters are either good at predicting b-factors or overlap and mutations. doi:10.1371/journal.pcbi.1003569.g013

The work presented here is to our knowledge also the most extensive test of existing methods for the prediction of the effect of mutations in protein stability. The majority of methods tested in the present work fail to predict stabilizing mutations. However, we are aware that the random reshuffled model used may be too stringent given the excessive number of destabilizing mutations in the dataset. The only models that predict stabilizing as well as destabilizing mutations are ENCoM and DMutant, however ENCoM is the only method with low self-consistency bias and error.

While the contribution of side chain entropy to stability is well established [72,73], here we use backbone normal modes to predict stability. As a consequence of the relationship between normal modes and entropy, our results attest to the importance of backbone entropy to stability and increase our understanding of the overall importance of entropy to stability. The strong trend observed on the behaviour of different parameters sets with respect to the α4 parameter is very interesting. Lower values are associated with better predictions of conformational changes while higher values are associated with better b-factor predictions. One way to rationalize this observation is to consider that higher a4 values lead to a rigidification of the structure, adding constraints and restricting overall motion. Likewise, lower α4 values remove constraints and thus lead to higher overlap.

We used ENCoM to predict the functional effects of the G121V mutant of the *E. coli* DHFR compared toNMR data. This position is part of a correlated network of residues that play a role in enzyme catalysis but with little effect on stability. The mutation affects this network by disrupting the movement of residues that are far from the binding site. We can predict the local changes in S2 with ENCoM. As these predictions are based on b-factor calculations, this result shows that at least in this case, even with b-factor prediction correlation lower than ANM, STeM and GNM we can detect functionally relevant variations. Clearly, despite the greater performance of GNM, ANM or STeM in the calculation of b-factors, these methods cannot predict b-factor differences as a consequence

of mutations, as their predictions are the same for the two forms. While a more extensive study is necessary involving S2 NMR parameters, our results serve as an example against relying too heavily on crystallographic b-factors for the evaluation of normal mode analysis methods.

Methods

The fundamentals of Normal Mode Analysis (NMA) have been extensively reviewed [74,75]. The key assumption in NMA is that the protein is in an equilibrium state around which fluctuations can be described using a quadratic approximation of the potential via a Taylor series approximation. In equilibrium the force constants are summarized in the Hessian matrix *H* that contains the elements of the partial second derivatives of the terms of the potential with respect to the coordinates. The potential used in ENCoM is similar to that of STeM, a Go-like potential where the closer a conformation $\vec{R}$ is to the reference (in this case equilibrium) conformation $\overrightarrow{R_0}$, the lower the energy.

$$V_{ENCoM}\left(\vec{R}, \overrightarrow{R_0}\right) = \sum_{bonds} V_1(r, r_0) + \sum_{angles} V_2(\theta, \theta_0) + \sum_{dihedral} V_3(\emptyset, \emptyset_0) + \sum_{i-j<3} V_4\left(r, r_{ij,0}\right)$$

$$= \sum_{bonds} \alpha_1 (r - r_0)^2 + \sum_{angles} \alpha_2 (\theta - \theta_0)^2$$

$$+ \sum_{dihedral} \alpha_3 (\emptyset - \emptyset_0)^2 + \sum_{i-j<3} \left(\beta_{ij} + \alpha_4\right) \left(5 \left(\frac{r_{ij}}{r_{ij,0}} \right)^{12} - 6 \left(\frac{r_{ij}}{r_{ij,0}} \right)^{10} \right)$$

The principal difference between the potential above and that of STeM are the β_{ij}terms that modulate non-bonded interactions between amino acid pairs according to the surface area in contact. Specifically,

$$\beta_{ij} = \sum_{k}^{N_i} \sum_{l}^{N_j} \varepsilon T(k) T(l) S_{kl}$$

where $\varepsilon T(k)T(l)$ represents a pairwise interaction energy between atom types $T(k)$ and $T(l)$ of atom k and l respectively of amino acids i and j containing N_i and N_j atoms each. Finally, S_{kl} represent represent the surface area in contact between atoms k and l calculated analytically [48]. We utilize the atom types classification of Sobolev et al. [76] containing 8 atom types. A matrix with all the interaction between atom types set at the value 1 is used in the non-specific ENCoMns model. In Figure 12 we illustrate with the concrete case of a set of 3 amino acids (D11, D45 and R141) in M. tuberculosis ribose-5-phosphate isomerase (PDB ID 2VVO), the differences between ENCoM, STeM and ANM in terms of spring strengths associated to different amino acids pairs. D45 is equally distant from R141 and D11 (around 6.0 Å) and interacts with R141 but does not with D11. Likewise, D11 does not interact with R141 with a Cα distance of 11.6 A. ANM assigns equal strength to all three pairwise Cα springs (as their distances fall within the 18.0 Å threshold). STeM assigns equal spring strengths to the D11–D45 and D45–R141 pairs. Among the three methods, ENCoM is the only one to properly assign an extremely weak strength to the D11–R141 pair, a still weak but slightly stronger strength to the D11–D45 pair (due to their closer distance) and a very strong strength to the spring representing the D45–R141 interactions.

The hessian matrix can be decomposed into eigenvectors $\overrightarrow{E_N}$ and their associated eigenvalues λ_n. For a system with N amino acids (each represented by one node in the elastic network), there are 3N eigenvectors. Each eigenvector describes a mode of vibration in the resonance frequency defined by the corresponding eigenvalue of all nodes, in other words the simultaneous movement in distinct individual directions for each node (Cα atoms in this case). The first 6 eigenvectors represent rigid body translations and rotations of the entire system. The remaining eigenvectors represent internal motions. The eigenvectors associated to lower eigenvalues (lower modes) represent more global or cooperative movements while the eigenvectors associated to higher eigenvalues (fastest modes) represent more local movements. Any conformation of the protein can be described by a linear combination of different amplitudes $\vec{A} = \{A_i, \dots, A_{3N}\}$ of eigenvector $\overrightarrow{E_n}$:

$$\vec{R}(\vec{A}) = \overrightarrow{R_0} + \sum_{n=7}^{3N} (A_n, \overrightarrow{E_n})$$

The source code for ENCoM is freely available at http://bcb.med. usherbrooke.ca/encom.

In terms of running time, the computational cost of running ENCoM is only slightly higher than that of other methods representing the protein structure with one node per amino acid. The main bottleneck in terms of computational time is the diagonalization of the Hessian matrix. As this matrix is the same size for ENCoM, ANM, STeM and GNM by virtue of considering a single node per amino acid, all methods should in principle run equivalently. Differences occur due to pre-processing, in particular with ENCoM where this step is more involved due to the more detailed calculations involved in the measurement of surface areas in contact. Taking the dataset used for the prediction of b-factors as an example, we obtain an average running time of 23.8, 30.2 and 34 seconds on average for STeM, ANM and ENCoM respectively on an Intel Core i7 CPU Q 740 @ 1.73GHz laptop.

Parameterization of ENCoM

In order to obtain a set of a parameters to be used with ENCoM we performed a sparse exhaustive integer search of the logarithm of a parameters with $\alpha_i = [10^{-4}, 10^{8}]$ for $i = [1,4]$ to maximize the prediction ability of the algorithm in terms of overlap and prediction of mutations. In other words, we searched all combinations of 13 distinct relative orders of magnitude for the set of 4 parameters a. For each parameter set, we calculated the bootstrapped median RMSE (see below) Z-score sum for the prediction of stabilizing and destabilizing mutations, $\tilde{Z}_{\,mutations} = -(\tilde{Z}_{\,destabilizing} + \tilde{Z}_{\,stabilizing})$. Keeping in mind that lower RMSE values represent better predictions, the 2000 parameter sets (out of 28561 combinations) with highest Z mutations were then used to calculate Z-scores for overlap in domain and loop movements. As our goal is to obtain a parameter set that combines low RMSE and high overlap, we ranked the 1000 parameter sets according to $\tilde{Z} = \tilde{Z}_{\,mutations} + \tilde{Z}_{\,loop} + \tilde{Z}_{\,domain}$. The parameter set with highest Z is (α1, α2, α3, α4)~(10^2,10^4,10^4,10^{-2}) (solid black line in Figure 13). The optimization of the bootstrapped median is equivalent to a training procedure with leave-many-out testing. Given the

dichotomy in predicting the effect of mutations and overlap on the one hand and b-factors on the other, we provide, the following is the best parameter set observed for the prediction of b-factors $(\alpha 1, \alpha 2, \alpha 3, \alpha 4) \sim (10^3, 10^5, 10^3, 10^2)$ with average b-factor correlation of 0.61 ± 0:13.

The exploration of parameter space shows that there is a clear trade-off between the prediction of mutations (low RMSE), conformational sampling (high overlap) and b-factors (high correlations). Parameter sets that improve the prediction of b- factors are invariably associated with poor conformational changes (low overlap) associated to both domain and loop movements and variable RMSE for the prediction of mutations (red lines in Figure 13). On the other hand, parameter sets that predict poorly b-factors, perform better in the prediction of conformational changes and the effect of mutations (blue lines in Figure 13). The a parameters used in STeM, $(\alpha 1, \alpha 2, \alpha 3, \alpha 4) \sim (360,72,9.9,3.6)$ are arbitrary, taken without modifications from a previous study focusing on folding [77]. As expected, this set of parameters can be considerably improved upon as can be observed in Figure 13 (dashed line). The four right-most variables in the parallel coordinates plot in Figure 13 show the logarithm of the a parameters for each parameter set. Either class of parameter sets, better for b-factors (in red) or better for overlap/RMSE (in blue) come about from widely diverging values for each parameter across several orders of magnitude. There are however some patterns. Most notably for α 4, where there is an almost perfect separation of parameter sets around α 4=1. Interestingly, higher values of α 1 and α 2, associated with stronger constraints on distances and angles tend also to be associated to better overlap values. While it is likely that a better- performing set of parameters can be found, the wide variation of values across many orders of magnitude show that within certain limits, the method is robust with respect to the choice of parameters. This result justifies the sparse search employed.

Bootstrapping

Bootstrapping is a simple and general statistical technique to estimate standard errors, p-values, and other quantities associated with finite samples of unknown distributions. In particular, bootstrapping help mitigate the effect of outliers and offers better estimates in small samples. Bootstrapping is a process by which the replicates (here 10000 replicates) of the sample points are stochastically generated (with repetitions) and used to measure

statistical quantities. In particular, bootstrapping allows the quantification of error of the mean [78–82]. Explained in simple terms, two extreme bootstrapping samples would be one in which the estimations of the real distribution of values is entirely made of replicates of the best case and another entirely of the worst case. Some more realistic combination of cases in fact better describes the real distribution. Thus, bootstrapping, while still affected by any biases present in the sample of cases, helps alleviate them to some extent.

Predicted b-factors

One of the most common types of experimental data used to validate normal mode models is the calculation of predicted b- factors and their correlation to experimentally determined b- factors. B-factors measure how much each atom oscillates around its equilibrium position [6]. Predicted b-factors are calculated as previously described [35]. Namely, for a given Cα node (i), one calculates the sum over all eigenvectors representing internal movements (n=7 to 3N) of the sum of the squared ith component of each eigenvector in the spatial coordinates x,y and z normalized by the corresponding eigenvalues:

$$B_i = \sum_{n=7}^{3N} \frac{E_{n,i,x}^2 + E_{n,i,y}^2 + E_{n,i,z}^2}{\gamma_n}$$

We calculate the Pearson correlation between predicted and experimental b-factors for each protein and average random samples according to the bootstrapping protocol described above.

Overlap

The overlap is a measure that quantifies the similarity between the direction of movements described by eigenvectors calculated from a starting structure and the variations in coordinates observed between that conformation and a target conformation [52,53]. In other words, the goal of overlap is to quantify to what extent movements based on particular eigenvectors can describe another conformation. The overlap between the nth mode, O_n, described by the eigenvector $\vec{E}_n$n is given by :

$$O_n(\overrightarrow{E_n}, \vec{R}) = \frac{\sum_{i=1}^{3N} E_{n,i} r_i}{\sum_{i=1}^{3N} E_{n,i} \sum_{i=1}^{3N} r_i}$$

where $\vec{r}$ represent the vector of displacements of coordinates between the starting and target conformations. The larger the overlap, the closer one can get to the target conformation from the starting one though the movements defined entirely and by the nth eigenvector.

We calculate the best overlap among the first 10 slowest modes representing internal motions.

Evaluating the effect of mutations on dynamics

Insofar as the simplified elastic network model captures essential characteristics of the dynamics of proteins around their equilibrium structures, the eigenvalues obtained from the normal mode analysis can be directly used to define entropy differences around equilibrium. Following earlier work [54,55], the vibrational entropy difference between two conformations in terms of their respective sets of eigenvalues is given by:

$$\Delta S_{Vib,A\to B} = \ln\left(\frac{\prod_{n=7}^{3N_A}\lambda_{n,A}}{\prod_{n=7}^{3N_B}\lambda_{n,B}}\right)$$

In the present work the enthalpic contributions to the free energy are completely ignored. Therefore, in the present work we directly compare experimental values of ΔG to predicted ΔS values. In order to use the same nomenclature as the existing published methods, we utilize ΔΔG to calculate the variation of free energy variation as a measure of conferred stability of a mutation.

Root mean square error

A linear regression going through the origin is build between predicted ΔΔG and experimental ΔΔG values to evaluate the prediction ability of the different models. The use of this type of regression is justified by the fact that a comparison of a protein to itself (in the absence of any mutation) should not have any impact on the energy of the model and the model should always predict an experimental variation of zero. However, a linear regression that is not going through the origin would predict a value different from zero equal to the intercept term. In other words, the effect of two consecutive mutations, going from the wild type to a mutated form back to the wild type form (WT→M→WT) would not end with the expected net null change.

The accuracy of the different methods was evaluated using a bootstrapped average root mean square error of a linear regression going through the origin between the predicted and experimental values. We refer to this as RMSE for short and use it to describe the strength of the relationship between experimental and predicted data.

Self-consistency bias and error in the prediction of the effect of mutations

If one was to plot the predicted energies variation of $\Delta\Delta G_{A\rightarrow B}$ versus $\Delta\Delta G_{B\rightarrow A}$ and trace a line y=-x, the bias would represent a tendency of a model to have points not equally distributed above or below that line while the error would represent how far away a point is from this line. In other words, considering a dataset of forward and back predictions, the error is a measure of how the predicted ΔΔG differ and the bias how skewed the predictions are towards the forward or back predictions [68]. A perfect model, both self-consistent and unbiased, would have all the points in the line. Statistically, the measures of bias and error are positively correlated. The higher the error for a particular method, the higher the chance of bias.

Linear combination of models

We determined the efficacy of linear combinations involving ENCoM and any of the other models for the prediction of the effect of mutations on thermostability as follows. For a given bootstrap sample of the data, we rescaled the ΔΔ predictions of each model as follows:

$$\overrightarrow{\Delta\Delta G_{Experimental}} \propto \overrightarrow{\beta_i \Delta\Delta G_i} = \overrightarrow{\Delta\Delta G_i}$$

where the vector notation signifies all data points in the particular bootstrap sample and the index i represents each model as well as the random model. We then use singular value decomposition to determine the best parameter the normalized predicted $\overrightarrow{\Delta\Delta G_i}$ values to calculate the parameters that maximize the RMSE difference between the linear combination model and the random predictions as follows :

$$\Delta RMSE_i = RMSE_i - RMSE_{random}$$

where:

$$RMSE_{random} = \sqrt{\frac{\left(\sum_{Bootsrap\ sample}^{N} \Delta\Delta G_{experimental} - \Delta\Delta G_{random}\right)}{N}}$$

and :

$$RMSE_i = \sqrt{\frac{\left(\sum_{Bootsrap\ sample}^{N} \Delta\Delta G_{experimental} - \left(\Delta\Delta G_i - \frac{\gamma ENCoM}{\gamma_i} \Delta\Delta G_{ENCoM}\right)\right)}{N}}$$

The bootstrapped average $\Delta RMSE_i$ is then calculated from the 10000 bootstrap iterations. The relative contribution of ENCoM and the model i under consideration is given by the ratio of the parameters $\frac{\gamma ENCoM}{\gamma_i}$. It is interesting to note that this ratio could be seen as an effective temperature factor, particularly considering that predicted $\Delta\Delta G$ values are primarily enthalpic in nature for certain methods and entropy based in ENCoM.

Supporting Information

Table S1 Raw data for the calculation b-factor correlations. (BZ2)

Table S2 Raw overlap calculations for the different methods. (BZ2)

Table S3 Experimental and predicted ΔΔG values on the effect of mutations. (BZ2)

Table S4 Raw data for the calculation of self-consistency bias and error on the prediction of forward and back mutations. (BZ2)

Table S5 Raw DB data and experimental S2 NMR order parameter for the G121V DHFR mutant. (BZ2)

Acknowledgments

The authors would like to thank Profs. Pierre Lavigne and Jean-Guy LeHoux from the department of Biochemistry, Université de Sherbrooke for useful discussions throughout the development of the method. RJN is part of Centre de Recherche Clinique Etienne-Le Bel, a member of the Institute of Pharmacology of Sherbrooke, PROTEO (the Quebec network for research on protein function, structure and engineering) and GRASP (Groupe de Recherche Axé sur la Structure des Proteines).

Author Contributions

Conceived and designed the experiments: VF RJN. Performed the experiments: VF. Analyzed the data: VF RJN. Contributed reagents/ materials/analysis tools: VF RJN. Wrote the paper: VF RJN

References

1. Gaudreault F, Chartier M, Najmanovich R (2012) Side-chain rotamer changes upon ligand binding: Common, crucial, correlate with entropy and rearrange hydrogen bonding. Bioinformatics 28: i423–i430.

2. Wu Z, Xing J (2012) Functional roles of slow enzyme conformational changes in network dynamics. Biophys J 103: 1052–1059.

3. Vaidehi N (2010) Dynamics and flexibility of G-protein-coupled receptor conformations and their relevance to drug design. Drug Discov Today 15: 951– 957.
4. Mittag T, Kay LE, Forman-Kay JD (2010) Protein dynamics and conformational disorder in molecular recognition. J Mol Recognit 23: 105–116.
5. Berman HM, Coimbatore Narayanan B, Di Costanzo L, Dutta S, Ghosh S, et al. (2013) Trendspotting in the Protein Data Bank. FEBS Lett 587: 1036– 1045.
6. Rhodes G (2010) Crystallography Made Crystal Clear. Academic Press. 7. Eyal E, Gerzon S, Potapov V, Edelman M, Sobolev V (2005) The limit of accuracy of protein modeling: influence of crystal packing on protein structure. J Mol Biol 351: 431–442.
8. Garbuzynskiy SO, Melnik BS, Lobanov MY, Finkelstein AV, Galzitskaya OV (2005) Comparison of X-ray and NMR structures: is there a systematic difference in residue contacts between X-ray- and NMR-resolved protein structures? Proteins 60: 139–147.
9. Go
¨bl C, Tjandra N (2012) Application of Solution NMR Spectroscopy to Study Protein Dynamics. Entropy 14: 581–598.
10. Kleckner IR, Foster MP (2011) An introduction to NMR-based approaches for measuring protein dynamics. Biochim Biophys Acta 1814: 942–968.
11. Foster MP, McElroy CA, Amero CD (2007) Solution NMR of large molecules and assemblies. Biochemistry 46: 331–340.
12. Ruschak AM, Kay LE (2012) Proteasome allostery as a population shift between interchanging conformers. Proc Natl Acad Sci U S A 109: E3454–E3462.
13. Karplus M, Kuriyan J (2005) Molecular dynamics and protein function. Proc Natl Acad Sci USA 102: 6679–6685.
14. Karplus M, McCammon JA (2002) Molecular dynamics simulations of biomolecules. Nat Struct Biol 9: 646–652.
15. Karplus M, Petsko GA (1990) Molecular dynamics simulations in biology. Nature 347: 631–639.
16. McCammon JA, Gelin BR, Karplus M (1977) Dynamics of folded proteins. Nature 267: 585–590.
17. Levitt M (1981) Molecular dynamics of hydrogen bonds in bovine pancreatic trypsin inhibitor protein. Nature 294: 379–380.
18. Piana S, Lindorff-Larsen K, Shaw DE (2012) Protein folding kinetics and thermodynamics from atomistic simulation. Proc Natl Acad Sci U S A 109: 17845–17850.
19. Cabana J, Holleran B, Beaulieu M-E ` , Leduc R, Escher E, et al. (2013) Critical
hydrogen bond formation for activation of the angiotensin II type 1 receptor. Journal Of Biological Chemistry 288: 2593–2604.
20. Shan Y, Eastwood MP, Zhang X, Kim ET, Arkhipov A, et al. (2012) Oncogenic mutations counteract intrinsic disorder in the EGFR kinase and promote receptor dimerization. Cell 149: 860–870.
21. Shan Y, Kim ET, Eastwood MP, Dror RO, Seeliger MA, et al. (2011) How does a drug molecule find its target binding site? J Am Chem Soc 133: 9181–9183.
22. Shaw DE, Dror RO, Salmon JK, Grossman JP, Mackenzie KM, et al. (2009) Millisecond-scale molecular dynamics simulations on Anton. In: Proceedings of the Conference on High Performance Computing Networking, Storage and Analysis; 14–20 November 2009, Portland, Oregon, United States. ACM SC '09. Available: http://dx.doi.org/10.1145/1654059.1654126

23. Wang Y, Harrison CB, Schulten K, McCammon JA (2011) Implementation of Accelerated Molecular Dynamics in NAMD. Comput Sci Discov 4: 015002.
24. Pronk S, Pa´ll S, Schulz R, Larsson P, Bjelkmar P, et al. (2013) GROMACS 4.5:a high-throughput and highly parallel open source molecular simulation toolkit. Bioinformatics 29: 845–854.
25. Atkins P, de Paula J (2010) Physical Chemistry. 9 ed. W. H. Freeman. 26. Tasumi M, Takeuchi H, Ataka S, Dwivedi AM, Krimm S (1982) Normal vibrations of proteins: glucagon. Biopolymers 21: 711–714.
27. Go N, Noguti T, Nishikawa T (1983) Dynamics of a small globular protein in terms of low-frequency vibrational modes. Proc Natl Acad Sci USA 80: 3696– 3700.
28. Brooks B, KarplusM(1983) Harmonic dynamics of proteins: normal modes and fluctuations in bovine pancreatic trypsin inhibitor. Proc Natl Acad Sci USA 80: 6571–6575.
29. Levitt M, Sander C, Stern PS (1985) Protein normal-mode dynamics: trypsin inhibitor, crambin, ribonuclease and lysozyme. J Mol Biol 181: 423–447.
30. Brooks B, Karplus M (1985) Normal modes for specific motions of macromolecules: application to the hinge-bending mode of lysozyme. Proc Natl Acad Sci USA 82: 4995–4999.
31. Tirion M (1996) Large Amplitude Elastic Motions in Proteins from a Single-Parameter, Atomic Analysis. Phys Rev Lett 77: 1905–1908.
32. Tama F, Gadea FX, Marques O, Sanejouand YH (2000) Building-block approach for determining low-frequency normal modes of macromolecules. PROTEINS: Structure, Function and Genetics 41: 1–7.
33. Bahar I, Atilgan AR, Erman B (1997) Direct evaluation of thermal fluctuations in proteins using a single-parameter harmonic potential. Fold Des 2: 173–181.
34. Doruker P, Atilgan AR, Bahar I (2000) Dynamics of proteins predicted by molecular dynamics simulations and analytical approaches: application to alpha- amylase inhibitor. PROTEINS: Structure, Function and Genetics 40: 512–524.
35. Atilgan AR, Durell SR, Jernigan RL, Demirel MC, Keskin O, et al. (2001) Anisotropy of fluctuation dynamics of proteins with an elastic network model. Biophys J 80: 505–515.
36. Doruker P, Jernigan RL, Bahar I (2002) Dynamics of large proteins through hierarchical levels of coarse-grained structures. J Comput Chem 23: 119–127.
37. Lin T-L, Song G (2010) Generalized spring tensor models for protein fluctuation dynamics and conformation changes. BMC Struct Biol 10 Suppl 1: S3.
38. Micheletti C, Carloni P, Maritan A (2004) Accurate and efficient description of protein vibrational dynamics: comparing molecular dynamics and Gaussian models. Proteins 55: 635–645.
39. Lope´z-Blanco JR, Garzon JI, Chaco ´n P (2011) iMod: multipurpose normal mode analysis in internal coordinates. Bioinformatics 27: 2843–2850.
40. Kurkcuoglu O, Jernigan RL, Doruker P (2006) Loop motions of triosephosphate isomerase observed with elastic networks. Biochemistry 45: 1173–1182.
41. Wong KF, Selzer T, Benkovic SJ, Hammes-Schiffer S (2005) Impact of distal mutations on the network of coupled motions correlated to hydride transfer in dihydrofolate reductase. Proc Natl Acad Sci USA 102: 6807–6812.
42. Rod TH, Radkiewicz JL, Brooks CL (2003) Correlated motion and the effect of distal mutations in dihydrofolate reductase. Proc Natl Acad Sci USA 100: 6980– 6985.

43. Kumar MDS, Bava KA, Gromiha MM, Prabakaran P, Kitajima K, et al. (2006) ProTherm and ProNIT: thermodynamic databases for proteins and protein- nucleic acid interactions. Nucleic Acids Res 34: D204–D206.
44. Henzler-Wildman K, Kern D (2007) Dynamic personalities of proteins. Nature 450: 964–972.
45. Doucet N (2011) Can enzyme engineering benefit from the modulation of protein motions? Lessons learned from NMR relaxation dispersion experiments. Protein Pept Lett 18: 336–343.
46. Gagne ´ D, Charest L-A, Morin S, Kovrigin EL, Doucet N (2012) Conservation of flexible residue clusters among structural and functional enzyme homologues. Journal Of Biological Chemistry 287: 44289–44300.
47. Boehr DD, Schnell JR, McElheny D, Bae S-H, Duggan BM, et al. (2013) A Distal Mutation Perturbs Dynamic Amino Acid Networks in Dihydrofolate Reductase. Biochemistry 52: pp 4605–4619.
48. McConkey BJ, Sobolev V, EdelmanM(2002) Quantification of protein surfaces, volumes and atom-atom contacts using a constrained Voronoi procedure. Bioinformatics 18: 1365–1373.
49. Kundu S, Melton JS, Sorensen DC, Phillips GN (2002) Dynamics of proteins in crystals: comparison of experiment with simple models. Biophys J 83: 723–732.
50. Amemiya T, Koike R, Kidera A, Ota M (2012) PSCDB: a database for protein structural change upon ligand binding. Nucleic Acids Res 40: D554–D558.
51. Hinsen K (1998) Analysis of domain motions by approximate normal mode calculations. PROTEINS: Structure, Function and Genetics 33: 417–429.
52. Marques O, Sanejouand YH (1995) Hinge-bending motion in citrate synthase arising from normal mode calculations. PROTEINS: Structure, Function and Genetics 23: 557–560.
53. Ma J, Karplus M (1997) Ligand-induced conformational changes in ras p21: a normal mode and energy minimization analysis. J Mol Biol 274: 114–131.
54. Karplus M, Kushick JN (1981) Method for estimating the configurational entropy of macromolecules. Macromolecules (14): 325–332.
55. McQuarrie DA (1976) Statistical Mechanics. University Science Books. 56. Gromiha MM, Sarai A (2010) Thermodynamic database for proteins: features and applications. Methods Mol Biol 609: 97–112.
57. Dehouck Y, Grosfils A, Folch B, Gilis D, Bogaerts P, et al. (2009) Fast and accurate predictions of protein stability changes upon mutations using statistical potentials and neural networks: PoPMuSiC-2.0. Bioinformatics 25: 2537–2543.
58. Berman H, Henrick K, Nakamura H,Markley JL (2007) The worldwide Protein Data Bank (wwPDB): ensuring a single, uniform archive of PDB data. Nucleic Acids Res 35: D301–D303.
59. Eswar N, Webb B, Marti-Renom MA, Madhusudhan MS, Eramian D, et al. (2006) Comparative protein structure modeling using Modeller. Current protocols in bioinformatics 15: 5.6.1–5.6.30.
60. Parthiban V, Gromiha MM, Schomburg D (2006) CUPSAT: prediction of protein stability upon point mutations. Nucleic Acids Res 34: W239–W242.
61. Zhou H, Zhou Y (2002) Distance-scaled, finite ideal-gas reference state improves structure-derived potentials of mean force for structure selection and stability prediction. Protein Sci 11: 2714–2726.

62. Yin S, Ding F, Dokholyan NV (2007) Eris: an automated estimator of protein stability. Nat Methods 4: 466–467.
63. Capriotti E, Fariselli P, Casadio R (2005) I-Mutant2.0: predicting stability changes upon mutation from the protein sequence or structure. Nucleic Acids Res 33: W306–W310.
64. Masso M, Vaisman II (2010) AUTO-MUTE: web-based tools for predicting stability changes in proteins due to single amino acid replacements. Protein Engineering Design & Selection 23: 683–687.
65. Schymkowitz J, Borg J, Stricher F, Nys R, Rousseau F, et al. (2005) The FoldX web server: an online force field. Nucleic Acids Res 33: W382–W388.
66. Leaver-Fay A, Tyka M, Lewis SM, Lange OF, Thompson J, et al. (2011) ROSETTA3: an object-oriented software suite for the simulation and design of macromolecules. Methods in Enzymology 487: 545–574.
67. Ferreiro DU, Komives EA, Wolynes PG (2013) Frustration in Biomolecules. arXiv:1312.0867v1.
68. Thiltgen G, Goldstein RA (2012) Assessing predictors of changes in protein stability upon mutation using self-consistency. PLoS ONE 7: e46084.
69. Gekko K, Yamagami K, Kunori Y, Ichihara S, Kodama M, et al. (1993) Effects of point mutation in a flexible loop on the stability and enzymatic function of Escherichia coli dihydrofolate reductase. J Biochem 113: 74–80.
70. Swanwick RS, Shrimpton PJ, Allemann RK (2004) Pivotal role of Gly 121 in dihydrofolate reductase from Escherichia coli: the altered structure of a mutant enzyme may form the basis of its diminished catalytic performance. Biochemistry 43: 4119–4127.
71. Soheilifard R, Makarov DE, Rodin GJ (2008) Critical evaluation of simple network models of protein dynamics and their comparison with crystallographic B-factors. Phys Biol 5: 026008.
72. Zhang J, Liu JS (2006) On side-chain conformational entropy of proteins. PLoS Comput Biol 2: e168.
73. Doig AJ, Sternberg MJ (1995) Side-chain conformational entropy in protein folding. Protein Sci 4: 2247–2251.
74. Cui Q, Bahar I (2006) Normal Mode Analysis. CRC Press. 75. Skjaerven L, Hollup SM, Reuter N (2009) Normal mode analysis for proteins. Journal of Molecular Structure: THEOCHEM 898: 42–48.
76. Sobolev V, Wade R, Vriend G, Edelman M (1996) Molecular docking using surface complementarity. PROTEINS: Structure, Function and Genetics 25: 120–129.
77. Clementi C, Nymeyer H, Onuchic JN (2000) Topological and energetic factors: what determines the structural details of the transition state ensemble and ‘‘en- route’’ intermediates for protein folding? An investigation for small globular proteins. J Mol Biol 298: 937–953.
78. Boos DD (2003) Introduction to the Bootstrap World. Statistical Science 18: 168–174.
79. Hinkley DV (1988) Bootstrap Methods. Journal of the Royal Statistical Society Series B (Methodological) 50: 321–337.
80. Stine R (1989) An Introduction to Bootstrap Methods Examples and Ideas. Sociological Methods & Research 18: 243–291.
81. Efron B, Tibshirani R (1995) Cross-validation and the bootstrap: Estimating the error rate of a prediction rule. Technical Report No. 577, National Science Foundation.
82. Efron B (1979) Bootstrap Methods: Another Look at the Jackknife. The Annals of Statistics 7: 1–26.

ARTICLE 2

Vibrational entropy differences between mesophile and thermophile proteins and their use in protein engineering

Auteurs de l'article:Vincent Frappier et Rafael Najmanovich

Statut de l'article: Accepté, Frappier V & Najmanovich RJ. *Vibrational entropy differences between mesophile and thermophile proteins and their use in protein engineering.* Protein Sci. 2014

Avant-propos: J'ai effectué toutes les méthodes expérimentales, généré les figures, participé à l'analyse des résultats et participé à la rédaction du manuscrit. Rafael Najmanovich est responsable de l'écriture et la révision du manuscrit.

Résumé : Nous avons récemment introduit ENCoM, un *Elastic Network Atomic Contact Mode*l , comme étant le premier modèle d'AMN de faible résolution qui considère la nature des acides aminés et qui peut prédire l'effet de mutations sur la thermostabilité protéique à partir de changement d'entropie vibrationnelle. Dans cet article preuve de concept, nous avons utilisé des paires de structures de conformations identiques de protéines homologues d'organismes mésophiles et thermophiles afin d'évaluer la capacité à les différentier à partir de l'entropie vibrationnelle. Nous avons observé que dans environs 60% des cas, les protéines thermophiles sont plus rigides que leurs homologues mésophiles et que cette différence peut être utilisée pour guider un design protéique qui vise à augmenter la stabilité thermique à travers une série de mutations. Nous avons également observé que les mutations qui différencient les thermophiles de leurs homologues thermophiles contribuent de façon indépendante à la réduction de l'entropie vibrationnelle. Nous discutons également de l'application et des implications de cette méthode dans l'ingénierie des protéines.

Abstract
We recently introduced ENCoM, an elastic network atomic contact model, as the first coarse-grained normal mode analysis method that accounts for the nature of amino acids and can predict the effect of mutations on thermostability based on changes vibrational entropy. In this proof-of-concept article, we use pairs of mesophile and thermophile homolog proteins with identical structures to determine if a measure of vibrational entropy based on normal mode analysis can discriminate thermophile from mesophile proteins. We observe that in around 60% of cases, thermophile proteins are more rigid at equivalent temperatures than their mesophile counterpart and this difference can guide the design of proteins to increase their thermostability through series of mutations. We observe that mutations separating thermophile proteins from their mesophile orthologs contribute independently to a decrease in vibrational entropy and discuss the application and implications of this methodology to protein engineering.

Keywords
mesophiles; thermophiles; thermostability; protein engineering; normal mode analysis; vibrational entropy; flexibility

Introduction

Increasing the thermostability of proteins is an important component of protein engineering[1,2]. In a number of industrial applications, it is more desirable or necessary to work at higher temperatures. Additionally, thermostable proteins have longer shelf lives. Perhaps the most widespread use in research of an enzyme with higher thermal stability is the DNA polymerase from *Thermus aquaticus* or Taq polymerase for short[3] that replaced the earlier use of *E. coli* DNA polymerase in PCR. Increasing protein stability is also used as a general strategy in protein engineering to build in a stability buffer to offset unwanted changes in stability caused by the introduction of other alterations that are necessary to achieve the new function that is the objective of the optimization. The above design considerations affect also the development of biologics (therapeutic proteins).[4,5]

Thermophile and hyperthermophile organisms, thought to be among the earliest life forms,[6] provide us with some of the most thermostable proteins.[7] Currently, the record holder is *Geogemma barossii*, an obligatory lithoautotroph isolated from the active Finn black-smoker hydrothermal-vent at a depth of 2280 m.[8] *G. barossii* also known as strain 121, does not grow below 85C, can grow at 121 °C (106 °C optimal) and survives up to two hours at 130 C.[9,10] While little is known about the characteristics of *G. barossii* proteins, the question of what are the structural factors that lead to the higher thermal stability of thermophile proteins has been addressed numerous times.[11–25]

Perhaps the earliest attempt at understanding the differences between thermophile and mesophile proteins was performed by Perutz et al.[11] who studied a number of ferrodoxins and hemoglobins. By analyzing the potential effect of amino acid substitutions based on their positions in the known structures, the authors suggested that the amino acid substitutions observed in the thermophile proteins may lead to higher thermal stability through the creation of salt bridges and hydrogen bonds. Argos et al.[26] used a set of criteria based on a-helix and b- sheet preferences, hydrophobicity, bulkiness, polarity, and least required number of codon alterations to define a subset of 24 pairwise amino acid exchanges most likely to increase protein stability. Of these, the authors observed 9 among the most frequently occurring exchanges using a set of 15 sequences representing three proteins from different organisms. More recently, Sadeghi et al.[19] used a dataset of 60 mesophile/thermophile homolog pairs and obtained the percentage of amino acid exchanges from mesophiles to thermophiles. Using a cutoff of 5%, the most frequently seen exchanges confirm the earlier results of Argos et al.[26] as well as those of Perutz et al.[11] showing a slight increase in the number of salt bridges and hydrogen bonds. For a long time, it has been suggested that thermophile proteins show improved packing,[12–14] a result that was confirmed by Robinon-Rechavi et al.[18] studying the wealth of *Thermatoga maritima* proteins structures elucidated by the Joint Center for Structural Genomics. The authors showed that there are small but statistically significant differences in compactness as measured through contact order[27] between *T. maritima* proteins and their mesophile homologs. Although the increased thermostability of thermophiles appear to come from a

number of different mechanisms,[16] amino acids preferences can be used as a guiding principle in the manual design of thermostable proteins.

One of the factors differentiating thermo from mesophile proteins is increased compactness,[12–14,18] leading to the suggestion that thermophile proteins are more rigid. It is important to keep in mind though that this is based on studying protein structures at nonthermophile temperatures.[15] Excluding differences between mesophile and thermophile proteins that reflect their divergent evolution to account for factors other than the temperature differences, the properties of highly similar mesophile and thermophile proteins ought to be similar. In that respect, the dynamics of mesophile and thermophile proteins has been studied via molecular dynamics[28] where it was found that at room temperature thermophile proteins are more rigid but equally flexible at higher temperatures. This equivalence is known as the hypothesis of equivalent states.[13,14]

Radestock et al.[22] compared 19 mesophile/thermophile homolog protein pairs using constraint network analysis.[29–31] The method uses the FIRST approach[31] based on the detection of regions where flexibility is affected by existing energetic constraints from covalent bonds and nonbonded interactions (primarily hydrogen bonds using a simplified distance and angle dependent potential[32]). The increase in temperature is simulated through the removal of hydrogen bonds from the list of constraints according to their calculated entropy, producing a phase transition as a function of the mean coordination of residues[33] akin to an artificial temperature. This artificial melting temperature is observed to be higher for the thermophile protein compared to its mesophile homolog in 13 out of 19 cases.[22]

Normal mode analysis[34–37] of proteins has almost as long a history as molecular dynamics[38,39] in the study of proteins. Like molecular dynamics, the number of atoms being studied is a limiting factor given the complexity of the calculations involved. However, molecular dynamics and normal mode analysis are not equivalent methods, they are complementary to each other. Both methods can use the same potential energy functions. Setting aside the caveat of how well such functions represent reality, there is a fundamental difference between molecular dynamics and normal mode analysis. Whereas molecular

dynamics generates a trajectory in conformational space starting from an equilibrium structure, normal mode analysis determines a basis set of modes of movements that when combined with the appropriate choice of amplitudes, can generate any other configuration of the system. In other words, while molecular dynamics generates actual movements, normal mode analysis generates possible movements. The eigenvalues obtained by normal mode analysis can be directly used to calculate vibrational entropy differences.[40]

Over the years a number of strategies have been devised to simplify the calculations allowing the two techniques to treat larger systems, or in the case of molecular dynamics longer time scales, to address biologically relevant processes. For molecular dynamics for example among other strategies,[41] steered molecular dynamics applies an external force to drive the dynamics.[42] For normal mode analysis, a number of simplifications in the representation of the protein have allowed to tackle larger systems or perform large-scale comparisons due to the lower requirements in terms of computational time. The vast majority of simplifications for normal mode analysis involve representing amino acids through their Cα atoms[43,44] or using block representations.[45]

Our group recently introduced ENCoM as the first coarse-grained normal mode analysis method that accounts for the nature of amino acids present in the protein and not just the structure.[46] Prior to ENCoM, coarse-grained normal mode analysis methods ignored the nature of the amino acid sequence of the protein such that two proteins with the same structure would produce equivalent results. In particular, such methods could not account for the effect of mutations. ENCoM modulates the spring constants between amino acids via the surface area in contact[47] between atoms weighted by an atom-type pairwise potential. With this more realistic representation of side-chain interactions, ENCoM performs better than ANM[48] and other methods in terms of the prediction of conformational changes.[46] Furthermore, ENCoM was tested on a large nonredundant subset of the ProTherm database with 303 cases of point mutations with experimentally calculated ΔΔG. We used vibrational entropy differences calculated with ENCoM as an approximation for ΔΔG to predict experimental ΔΔG values for single mutants and compared ENCoM against eight existing methods to predict the effect of mutations on thermal stability. We showed

that ENCoM is less biased than existing methods and particularly good at predicting stabilizing mutations.[46] To date except for ENCoM, all existing methods for the prediction of the effect of mutations on thermostability are based on machine learning approaches and enthalpy calculations. ENCoM represent an entirely new way to predict thermostability based on vibrational entropy.

In this article, we use a curated extensive non-redundant dataset of mesophile/thermophile homolog pairs[24] to study the extent to which vibrational entropy variations calculated using normal mode frequencies can differentiate thermophile proteins from their mesophile homologs. We then proceed to study how different types of mutations affect vibrational entropy differences and how such entropy changes vary with the order in which mutations leading from mesophile to the thermophile are introduced in rubredoxin. Lastly, we perform all possible mutations in all positions for rubredoxin from *D. vulgaris* and rank the mutations observed in the thermophile homolog to determine if the methodology can be used to guide the selection of mutations for added thermal stability.

Results

Dataset

Out of the initial 373 proteins pairs,[24] 314 pairs were kept based on the criteria described in the method section. The mean RMSD between pairs is 1.32 ± 0.49 A, the average percentage of sequence identity is 42 ± 16 and the average sequence length is 165 ± 80 amino acids.

Vibrational entropy

Normal mode eigenvectors and eigenvalues are used to estimate vibrational entropy differences (ΔS_{vib}) as described in the methods section.[46] ENCoM predict that thermophilic proteins have statistically significant (P-value=0.03) smaller vibrational entropy in 186 thermophile proteins relative to their mesophilic counterparts. There is no correlation between the predicted difference in vibrational entropy of ENCoM with the RMSD between the pair (r=0.04, P-value=0.49), the sequence identity (r=-0.04, P- value=0.51) or the difference in sequence length (r=0.03, P-value=0.61). The P-values above represent the

statistical significance of the alternative hypothesis that the true correlation is different from 0, therefore the high P-values obtained mean that the correlation cannot be said to be different from 0, showing the statistical significance of the near null found correlations.

To assess the robustness of the predictions above and to what extent these are attributable to small structural differences, we generated for every PDB structure 50 models by homology modeling using Modeller with a flexible backbone and compared the vibrational entropy for each pair of ensembles using a Student t-test for each case in the database. ENCoM predicts that thermophilic proteins have less vibrational entropy on average than their mesophilic counterparts ($\Delta S_{vib} > 0$) in 195 cases, 104 with $\Delta S_{vib} < 0$ and 15 with no statistically significant differences (P-value 0.05). We obtain a Pearson correlation coefficient of 0.33 between the average ΔS_{vib} for the models and the ΔS_{vib} of the crystal structure only. In the 104 cases where $\Delta S_{vib} < 0$, it is likely that factors other than vibrational entropy contribute for the higher stability of the thermophile protein as previously observed.[49]

The above results for a dataset of 314 mesophile and thermophile homolog protein pairs were performed using a single mesophile homolog. It is possible however to perform such ΔS_{vib} calculations in cases were there is more than one mesophile homolog. We used a similar approach as above generating 100 models for each of 4 structures for rubredoxin from three mesophile species: *C. pasteurianum* (PDB ID 1IRO), *D. vulgaris* (PDB ID 8RXN) and *D. gigas* (PDB ID 1RDG) and one thermophile *P. furiosus* (PDB ID 1CAA). Rubredoxin is a small protein (51 amino acids) with an iron-sulphur cluster that is present in all homologs. We observe that the thermophile vibrational entropy is higher on average for the thermophile than any of its mesophile counterparts (Fig. 1) suggesting that the positive ΔS_{vib} differences observed in the large database are not the result of the particular mesophile species selected but a property conserved across mesophile homologs with respect to a thermophile.

Mutations affecting ΔS_{vib}

In this section, we are interested in what amino acid changes on average are more likely to lead to positive entropy changes ($\Delta S_{vib} > 0$) leading to a more rigid thermophile protein than its mesophile counterpart. For that purpose, we focus only on the 186 cases where $\Delta S_{vib} > 0$. Furthermore, we remove the 10% outliers at both extremes of ΔS_{vib} variation (see the methods section for justification) to obtain a dataset of 14,884 mutations comprising 99 mutations types (out of the 380 possible combinations) that produce statistically significant (P-value < 0.01) ΔS_{vib} variations. We analyzed if their mean ΔS_{vib} correlates with the rate at which the given residue is observed to mutate from mesophile to thermophile proteins. In other words, are mutations more frequently seen to occur between mesophiles and thermophiles more likely to affect flexibility? We observe a correlation of 0.54 when looking at the average effect of residues mutating to any residue and -0.56 when looking at any residue mutated to a given residue (Table I). For example, mutating an alanine to any residue was observed 1559 times while the inverse, any residue to alanine 1236 times (-20.6%). Replacing an alanine for other residues increases rigidity as measured by an average ΔS_{vib} of 0.35 for the 1559 single mutations, representing an increase in stability. Other amino acids, such as the aromatic amino acids tyrosine and tryptophan or the charged amino acid arginine produce the opposite effect, are more abundant in thermophiles, lead to increased rigidity as measured by ΔS_{vib} and consequently contribute to the increase in stability of the thermophile protein. Overall, this positive correlation between abundance and ΔS_{vib} means that residues that increase rigidity are found more often in the thermophile and residues that are known to increase flexibility are more often found in the mesophile proteins.

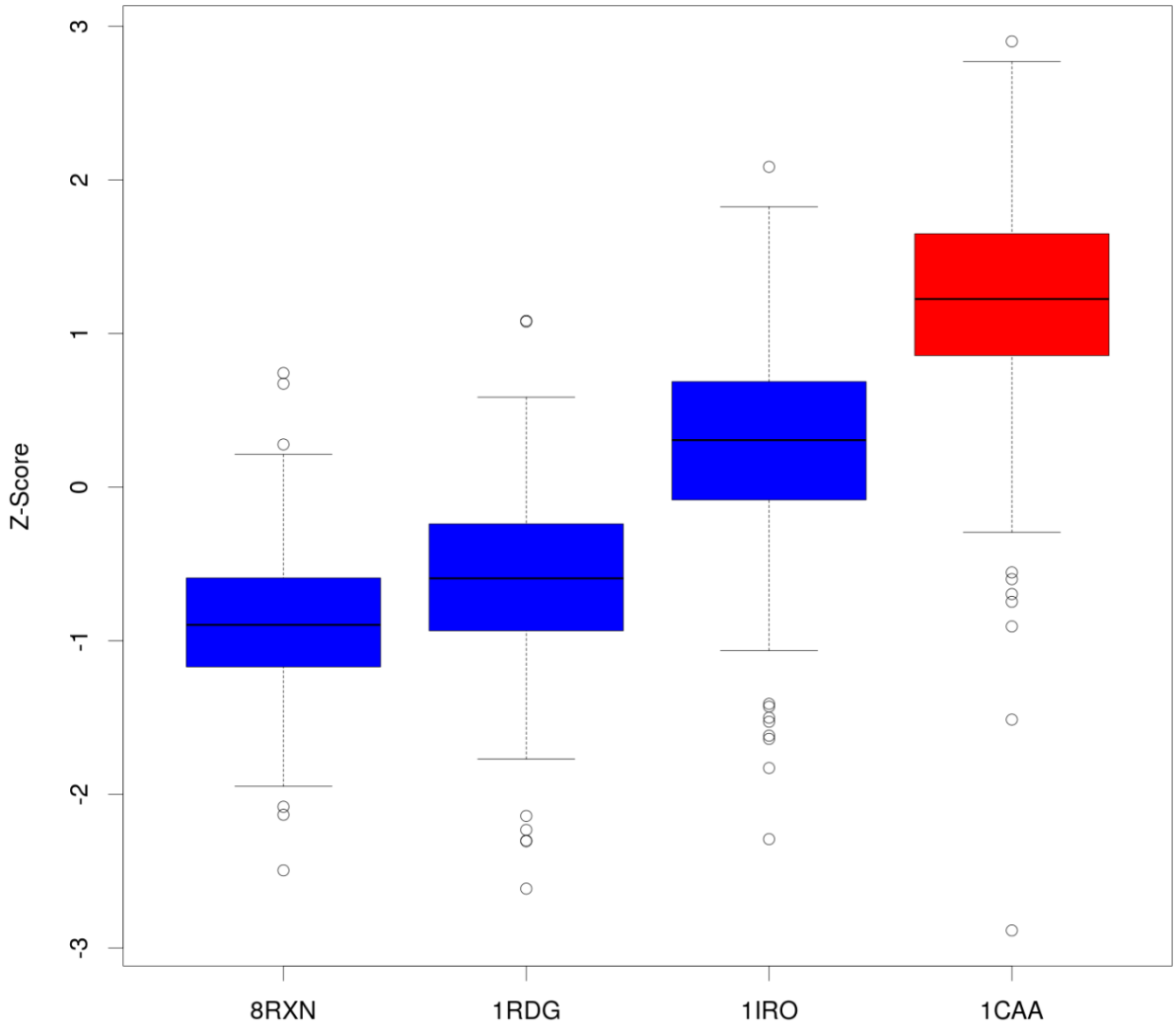


Figure 1. Average ΔS_{vib} Z-scores for rubredoxin homologs. The mesophile homologs from *C. pasteurianum* (PDB ID 1IRO), *D. vulgaris* (PDB ID 8RXN), and *D. gigas* (PDB ID 1RDG) in blue are compared with that of the thermophile *P. furiosus* (PDB ID 1CAA) in red. Bootstrapped differences are statistically significant with P-value < 0.001.

Not just particular amino acids are preferred when changing from mesophile to thermophile, but specific preferences in amino acid pairwise replacements can be observed. For example, replacing an alanine in the mesophile for any residue other than glycine

increases ΔS_{vib} (therefore stability). For alanine, the highest ΔS_{vib} is obtained when mutating to a tryptophan. The heatmap in Figure 2 shows clear preferences for particular amino acids with essentially the upper half triangle around the inverse diagonal discriminating between mutations that increase ΔS_{vib} and pairwise mutations in the lower right half triangle decreasing ΔS_{vib}, thus diminishing stability. Only pairwise combinations with statistically significant results are shown in the heatmap. The heatmap also includes a wildcard row (marked by *) that represents the data in Table I, that is, any amino acid in the mesophile mutated to some particular amino acid in the thermophile and a wildcard column, for mutations of particular amino acids in the mesophile to any amino acid in the thermophile. Despite not having sufficient data or no clear trend to assign average ΔS_{vib} values for every pair of amino acids, the data in Figure 2 can be used as a guide when seeking to introduce mutations into a protein in order to affect its rigidity and stability.

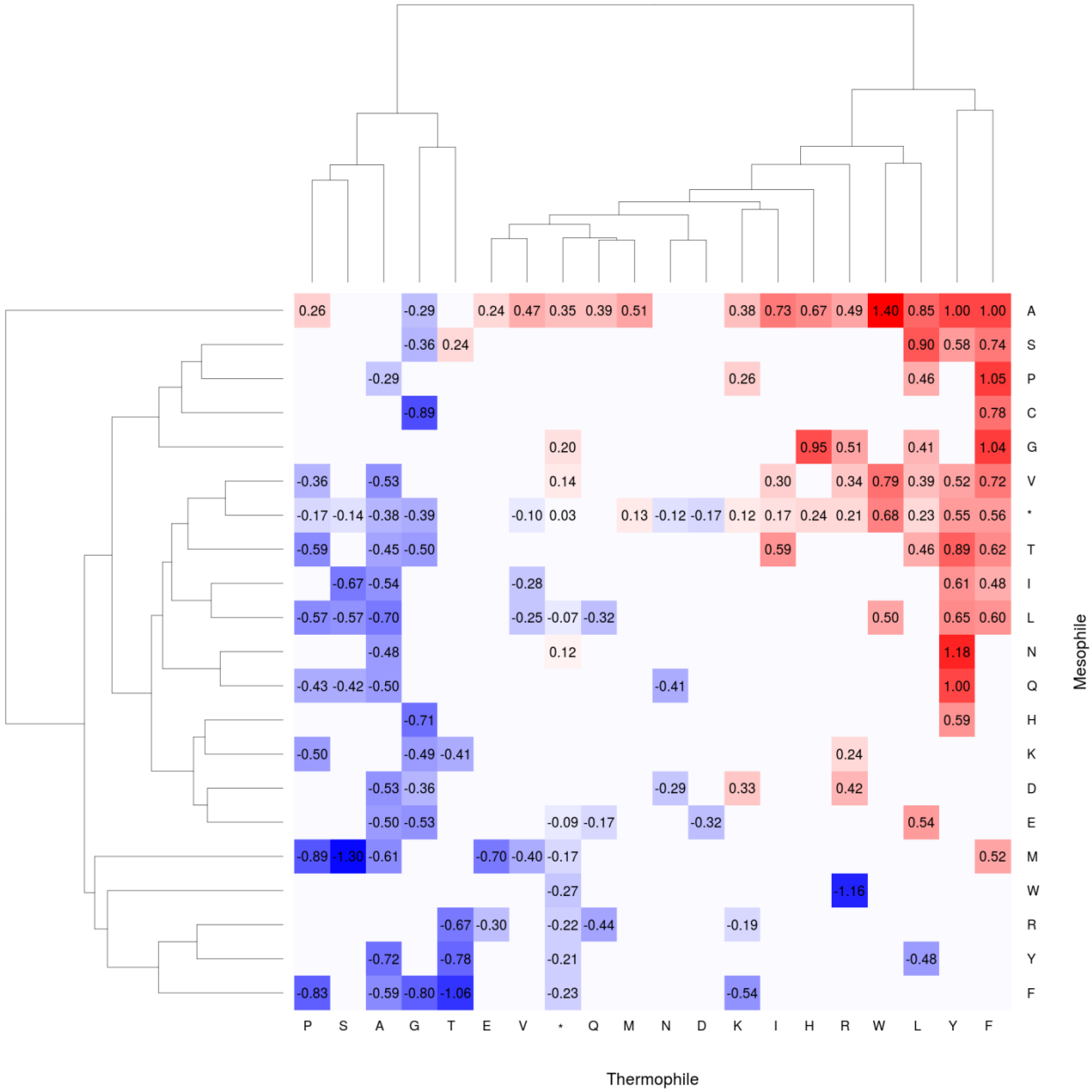


Figure 2. Heatmap of average ΔS_{vib} for pair-wise amino acid substitutions from mesophile to thermophile proteins. The top right half matrix around the inverse diagonal represent for the most part mutations that increase the stability of the thermophile protein. Missing values represent pairwise amino acid substitutions without statistically significant results. ΔS_{vib} values are scaled by 103 for visualization purposes.

Engineering mutations

For each of the mesophile homologs of rubredoxin described above we calculated the best and worst and most probable order of mutations to reach the thermophile sequence as described in the methods section. In Figure 3, we show a sequence alignment of the four

sequences with colors that indicate the ΔS_{vib} of each mutation. We observe that for the most part the mutations in each position have equivalent effects across the different sequences. For each of the mesophile homologs, the best ΔS_{vib} path shows a steep increase followed by a peak/plateau and a decrease [Fig. 4(A–C)]. The single most contributing mutation according to ΔS_{vib} is P15E. Curiously, the mutation Y4W is stabilizing for two of the three homologs but destabilizing for 1RDG.

Table I. *Average effect of single point mutations*

Amino-Acid	Number of cases				ΔS_{vib}	
	From	To	Difference[a]		From	To
A	1559	1236	-323	(-20.72)	0.35	-0.38
C	223	112	-111	(-49.78)	NS[b]	NS
D	852	734	-118	(-13.85)	NS	-0.17
E	916	1418	502	(54.80)	-0.09	NS
F	499	593	94	(18.84)	-0.23	0.56
G	730	667	-63	(-8.63)	0.20	-0.39
H	344	282	-62	(-18.02)	NS	0.24
I	951	1093	142	(14.93)	NS	0.17
K	859	1211	352	(40.98)	NS	0.12
L	1239	1314	75	(6.05)	-0.07	0.23
M	397	364	-33	(-8.31)	-0.17	0.13
N	634	531	-103	(-16.25)	0.12	-0.12
P	505	532	27	(5.35)	NS	-0.17
Q	718	432	-286	(-39.83)	NS	NS
R	729	899	170	(23.32)	-0.22	0.21
S	1001	776	-225	(-22.48)	NS	NS
T	990	740	-250	(-25.25)	NS	-0.14
V	1198	1285	87	(7.26)	0.14	-0.10
W	137	140	3	(2.19)	-0.27	0.68

Y	403	525	122	(30.27)	-0.21	0.55

[a] The number in parenthesis represents percent of change, e.g. -323/1559 for A.
[b] Statistically non-significant value

The most probable path bifurcates at times when more than one mutation of the same amino acid could be performed but in general, it also shows a steady increase. The ΔS_{vib} calculations here suggest that it may be possible to achieve the higher stability of the thermophile protein with less than the number of mutations observed. Additionally, the fact that different mutational pathways lead to approximately the same thermophile protein suggests that mutations have independent effects.

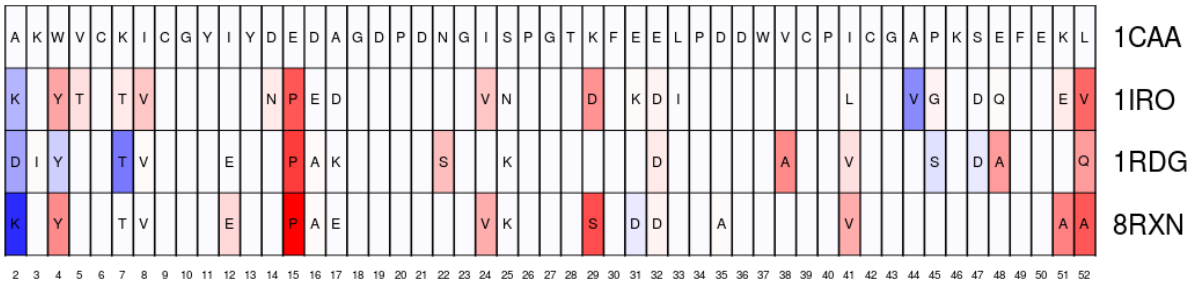


Figure 3. Alignment of rubredoxin homolog sequences showing ΔS_{vib} for single mutations. The color gradient is proportional to ΔS_{vib} with mutations predicted to increase the stability of the thermophile in red and those decreasing ΔS_{vib} in blue. Most mutations have equivalent effects across homologs. Unlabeled positions are unchanged with respect to the thermophile (1CAA).

Lastly, for rubredoxin from *D. vulgaris* (PDB ID 8RXN) we performed every one of the 969 possible single point mutations (i.e., 19 mutations per position) and calculated ΔS_{vib}. All mutations were assigned an overall rank (out of 969) and a position-specific rank (out of 19). In Table II, we present the ranks of the 17 mutations observed for rubredoxin between *D. vulgaris* and *P. furiosus* (PDB ID 1CAA). The top 3 mutations in terms of ΔS_{vib} have an overall rank within the top 10%. Furthermore, for the 13 or 8 out of 16 positions that change in *P. furiosus*, the position-specific rank is among the top 10 or top 5, respectively (Table II). The results above suggest that it is possible to find among top ranking mutations those that were naturally selected. Thus, it is possible to use ΔS_{vib} with

ENCoM to select mutations that increase rigidity and thermal stability in protein engineering.

Table II. ***Rank of observed mutations for D. vulgaris Rubredoxin***

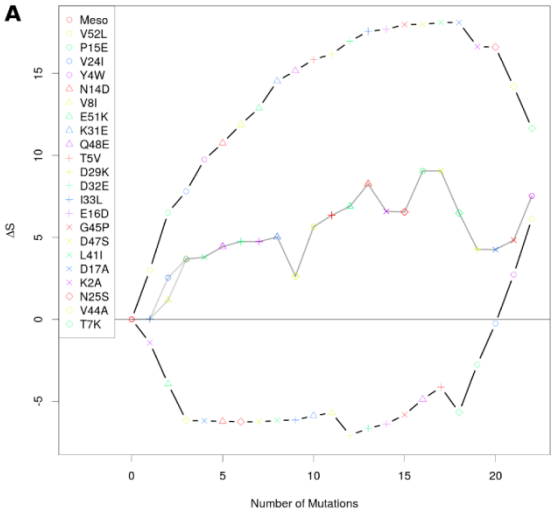
Mutation	ΔS_{vib}	Rank	
		Overall[a]	Position-specific[b]
K2A	-4.1781	914	19
Y4W	2.3298	137	1
T7K	-0.0181	620	8
V8I	0.0782	490	7
E12I	0.8392	295	6
P15E	5.2338	46	4
A16D	0.1984	449	5
E17A	0.0029	531	15.5
V24I	1.7099	190	7
S29K	3.5877	78	4
D31E	-0.2756	667	7
D32E	0.432	380	4
A35D	0.2632	425.5	3
V41I	1.6565	196	2
A51K	2.5215	125	2
A52L	3.4532	82	11

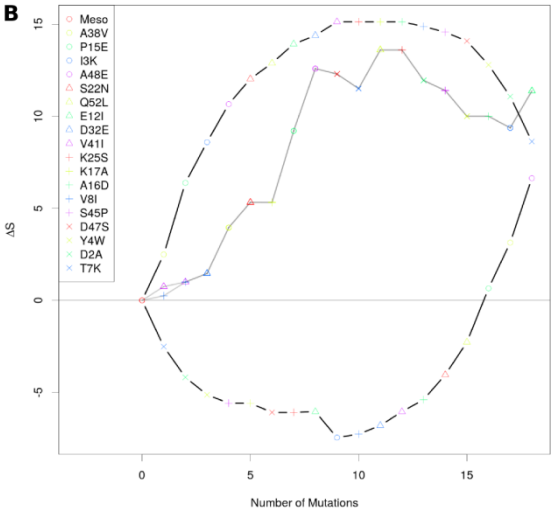
[a] Out of 969 performed single mutations (19 amino-acids in 51 positions).
[b] Out of 19 possibilities at the given position

Discussion

The number of factors contributing to the higher stability of thermophile proteins is as varied as the number of approaches used to understand them. From a thermodynamic point of view these led to three types of stability curves relative to a mesophile protein: an over increase in stability across the temperature range with associated increases in ΔS, a shift toward higher temperatures of the entire curve but with no change in the maximum or ΔS at

a given temperature and a flattening of the curve with associated decrease in ΔS. In all cases, the T_m of the thermophile will be higher than that of the mesophile.[50] In this article, we explore the use of normal mode analysis as implemented in ENCoM to understand the differences between mesophile and thermophile proteins in terms of entropy changes. ENCoM is unique in that as a normal mode analysis method it permits to take in consideration the nature of amino acids of a protein in addition to its structure and as such account for the effect of mutations. Being a normal mode analysis method it is possible to calculate vibrational entropy differences that can be used to differentiate mesophile and thermophile proteins and being a coarse-grained method it allows us to perform a large scale analysis. As such, this work offers a new perspective on the problem. When comparing mesophile and thermophile homolog protein pairs, we observe that in around 60% of cases there is a decrease in entropy (increase in rigidity) of the thermophile protein relative to the mesophile homolog. As a result of evolution, different mechanisms to maintain the thermal stability of proteins at higher temperatures were selected.[49] The decrease in entropy as measured using ENCoM is one such factor that is statistically significant in around 60% of cases. Given the temperature dependence of the vibrational entropy,[51] the entropy difference between mesophile and thermophile homologs should decrease at the normal higher temperature of thermophile organisms, in what is known as the hypothesis of corresponding states.[20] Using crystal structures or models gave similar results in term ΔS_{vib} and no correlation was observed between this value and RMSD or the sequence identity, suggesting that ΔS_{vib} observed between the mesophile and thermophile proteins are independent of the conformation used.


A
Meso
V52L
P15E
V24I
Y4W
N14D
V8I
E51K
K31E
Q48E
T5V
D29K
D32E
I33L
E16D
G45P
D47S
L41I
D17A
K2A
N25S
V44A
T7K
ΔS
Number of Mutations


B
Meso
A38V
P15E
I3K
A48E
S22N
Q52L
E12I
D32E
V41I
K25S
K17A
A16D
V8I
S45P
D47S
Y4W
D2A
T7K
ΔS
Number of Mutations

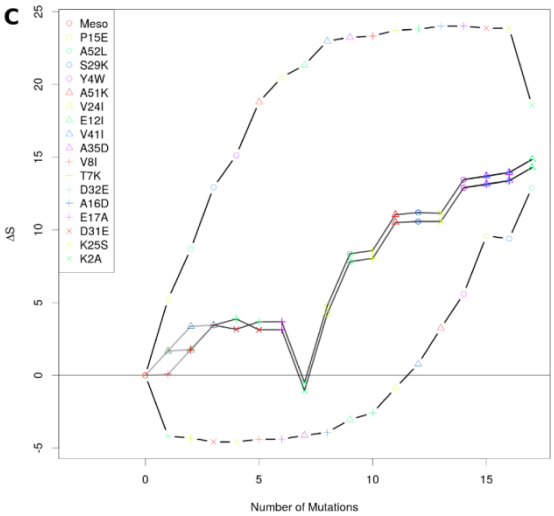

C
Meso
P15E
A52L
S29K
Y4W
A51K
V24I
E12I
V41I
A35D
V8I
T7K
D32E
A16D
E17A
D31E
K25S
K2A
ΔS
Number of Mutations

Figure 4. Mutational pathway for mesophile rubredoxin homologs to reach the thermophile sequence. Three pathways are presented, the best and worst pathways according to the ΔS_{vib} contributions of single mutations and the most probable (see methods). Each point from right to left represents a new structure with one extra mutation toward the thermophile protein. A: *C. pasteurianum* (PDB ID 1IRO), (B) *D. gigas* (PDB ID 1RDG), and (C) *D. vulgaris* (PDB ID 8RXN).

The analysis of ΔS_{vib} for single mutations shows distinct preferences in terms of amino acid substitutions that favor a decrease of vibrational entropy of the thermophile protein. Such substitutions can be used in protein engineering when trying to increase the thermal stability of proteins through modeling potential mutations and calculating ΔS_{vib} with ENCoM. When selecting mutations based on ΔS_{vib}, we observe a steady increase in ΔS_{vib} and a peak/plateau, suggesting that not all mutations affect the entropy and some of the thermostability effect may be obtained with a fraction of the mutations. We also observe no synergy from the order of the mutations. As different combinations of the order of performing mutations lead to the same final result, each mutation seems to contribute independently to the overall effect. The very good overall and position-specific rankings obtained for the mutations in rubredoxin from *D. vulgaris* are very promising and suggest that this strategy can be used in practice to increase the thermal stability of proteins.

Overall, the results shown here are encouraging from a protein-engineering standpoint, as it is possible to limit the search to a few top predicted residues

Material and Methods

Database

We used the extensively curated dataset of 373 pairs of mesophile and thermophile homolog proteins of Glyakina et al.[24] We removed NMR structures (10 structures) as well as structure pairs with sequence lengths differing by more than 9 aminoacids (as length has a direct effect on normal mode calculations). The structures were cropped on the basis of the structural alignment provided by the authors. RMSD and sequence identity were calculated for the cropped structures using TM-align in the I-TASSER software suite.[52] Heteroatoms and hydrogen atoms were removed from the structures, as the former would affect the normal mode calculations with ENCoM and the latter are not taken in account in

ENCoM. The final dataset contains 314 pairs (Supporting Information File 1). The mean RMSD between pairs is 1.32 ± 0.49 Å, the average percent sequence identity is 42 ± 16 and the average sequence length is 165 ± 80 amino acids.

All structures were rebuilt with Modeller using their own actual structure as template to generate a conformational ensemble for the given structure. A total of 50 models were generated for each protein.

Preparation of rubredoxin structures

All four rubredoxin structures (PDB IDs 8XRN, 1RDG, 1IRO, and 1CAA) were structurally aligned and verified using PyMol. Terminal residues that were not structurally aligned were removed in order to achieve the same sequence length for every structure. The average RMSD between all structures is 0.51 ± 0.05 Å. In working with a small number of structures (as opposed to the entire dataset above), we increased the size of the conformational ensemble generated by Modeller to 100 models for each sequence using each of the four crystal structures as templates to see if the results would be affected by the number of models. The results represent the average of 400 models for each sequence.

Determination of vibrational entropy

Normal mode analysis methods are uniquely suited to calculate vibrational entropy differences.[40] All-atom normal mode calculations are computationally expensive and thus cannot be used in large scale. While coarse- grained normal mode analysis methods are computationally fast, with the exception of ENCoM,[46] such methods cannot by definition predict the effect of mutations when these do not change the backbone conformation of the protein. Therefore, the ENCoM method is particularly suited to perform scale analyses as required for the analysis of vibrational entropy changes both in terms of the number of proteins and number of mutations. We used the ENCoM method to calculate vibrational entropy differences (ΔS_{vib}) as described earlier[40,46] and normalized by the number of modes to account for the effect of varying sequence lengths. Specifically,

$$\Delta S_{Vib,A\to B} = \frac{N_B}{N_A}\ln\left(\frac{\prod_{n=7}^{3N_A}\lambda_{n,A}}{\prod_{n=7}^{3N_B}\lambda_{n,B}}\right)$$

where N_A and N_B represent the number of amino acids in proteins A and B and $\lambda_{n,i}$ represents the nth normal mode (the first 6 correspond to rotational and translational degrees of freedom) for protein i. The smaller ΔS_{vib}, the higher the flexibility of the thermophile relative to the mesophile protein. Likewise, the smaller ΔS_{vib}, the smaller the contribution to the stability of the thermophile protein relative to the mesophile. The ENCoM method is available for download or online use at http://bcb.med.usherbrooke.ca/encom.

Engineering protocol

Homolog pairs of mesophile and thermophile rubredoxin protein sequences were aligned using TMalign to identify the mutations that differentiate each pair. Starting from the mesophile protein, every single mutation was generated using the Modeller Mutated function to produce one structure per mutant. For one of the homologs (*D. vulgaris* PDB ID 8RXN), modeling with a flexible backbone lead to drastic conformational changes in the C-terminus of the protein that are not observed in the thermophile. This modeling artefact did not occur for the other homologs. Thus, for consistency we modelled all structures with a fixed backbone for all three homologs. Considering the number of mutations between homologs with minimal differences in the structures (average RMSD 0.51 A) we feel that the restriction to use a fixed backbone in this experiment does not affect the results. The ΔS_{vib} relative to the mesophile was evaluated for every structure as described above. The mutations with highest ΔS_{vib} was selected and the process was repeated for the remaining mutations in turn until the thermophile sequence was reached. The same process was performed for the worst mutations in terms of ΔS_{vib}. This protocol produces an upper and lower bound on ΔS_{vib} and selects a particular order of performing the mutations that separates a thermophile protein from its mesophile counterpart. In addition, we also use the percent amino-acid replacements between mesophile and thermophile proteins of Sadeghi et al.[19] to choose at each step the most probable mutation to introduce. We call this mutational pathway as the most probable pathway.

Effect of mutations

We performed every mutation observed between mesophile and thermophile proteins in the 186 cases with $\Delta S_{vib} > 0$ as an individual mutation using the Mutated function of Modeller, in this case with a flexible backbone. To be included in our analysis, a type of mutation (e.g., alanine to arginine) must appear in the dataset more than 5 times and have a consistent ΔS_{vib} effect (P-value < 0.01). A mutation type may have a large or small effect on entropy differences as a result of artefacts in the methodology (modeling errors) or due to its effect on a number of other molecular properties of the protein as discussed in the introduction. As such, we remove the extremes (top and bottom) 5% ΔS_{vib} outlier mutations when trying to identify the effect on ΔS_{vib} most often associated to different types of mutations.

Acknowledgments

RJN is part of Centre de Recherche Clinique Etienne-Le Bel, the Institute of Pharmacology of Sherbrooke, PROTEO (the Québec network for research on protein function, structure and engineering) and GRASP (Groupe de Recherche Axe sur la Structure des Protéines).

References

1. Samish I, Macdermaid CM, Perez-Aguilar JM, Saven JG (2011) Theoretical and computational protein design. Annu Rev Phys Chem 62:129–149.
2. Kiss G, Çelebi-Ölçüm, N, Moretti R, Baker D, Houk KN (2013) Computational enzyme design. Angew Chem Int Ed Engl 52:5700–5725.
3. Ghadessy FJ, Ong JL, Holliger P (2001) Directed evolution of polymerase function by compartmentalized self-replication. Proc Natl Acad Sci USA 98:4552–4557.
4. Cauchy M, D'Aoust S, Dawson B, Rode H, Hefford MA (2002) Thermal stability: a means to assure tertiary structure in therapeutic proteins. Biologicals 30:175– 185.
5. Chennamsetty N, Voynov V, Kayser V, Helk B, Trout BL (2009) Design of therapeutic proteins with enhanced stability. Proc Natl Acad Sci USA 106:11937– 11942.
6. Fujiwara S (2002) Extremophiles: developments of their special functions and potential resources. J Biosci Bioenegry 94:518–525.
7. Stetter KO (1999) Extremophiles and their adaptation to hot environments. FEBS Lett 452:22–25.
8. Delaney JR, Kelley DS, Mathez EA, Yoerger DR, Baross J, Schrenk MO, Tivey MK, Kaye J, Robigou V (2001) "Edifice Rex" Sulfide Recovery Project: Analysis of submarine hydrothermal, microbial habitat. Eos Trans Am Geophys Union 82:67–73.
9. Miroshnichenko ML, Bonch-Osmolovskaya EA (2006) Recent developments in the thermophilic microbiology of deep-sea hydrothermal vents. Extremophiles 10:85– 96.

10. Kashefi K, Lovley DR (2003) Extending the upper temperature limit for life. Science 301:934–934.
11. Perutz MF, Raidt H (1975) Stereochemical basis of heat stability in bacterial ferredoxins and in haemoglobin A2. Nature 255:256–259.
12. Menendez-Arias L, Argos P (1989) Engineering protein thermal stability. Sequence statistics point to residue substitutions in alpha-helices. J Mol Biol 206:397–406.
13. Jaenicke R, Zavodszky P (1990) Proteins under extreme physical conditions. FEBS Lett 268:344–349.
14. Jaenicke R (1991) Protein stability and molecular adaptation to extreme conditions. Eur J Biochem 202: 715–728.
15. Scandurra R, Consalvi V, Chiaraluce R, Politi L, Engel P (1998) Protein thermostability in extremophiles. Biochimie 80:933–941.
16. Szilagyi A, Zavodszky P (2000) Structural differences between mesophilic, moderately thermophilic and extremely thermophilic protein subunits: results of a comprehensive survey. Structure 8:493–504.
17. Demirjian D, Moris-Varas F, Cassidy C (2001) Enzymes from extremophiles. Curr Opin Chem Biol 5:144–151.
18. Robinson-Rechavi M, Godzik A (2005) Structural genomics of thermotoga maritima proteins shows that contact order is a major determinant of protein thermostability. Structure 13:857–860.
19. Sadeghi M, Naderi-Manesh H, Zarrabi M, Ranjbar B (2006) Effective factors in thermostability of thermophilic proteins. Biophys Chem 119:256–270.
20. Radestock S, Gohlke H (2008) Exploiting the link between protein rigidity and thermostability for data- driven protein engineering. Eng Life Sci 8:507–522.
21. Mamonova TB, Glyakina AV, Kurnikova MG, Galzitskaya OV (2010) Flexibility and mobility in mesophilic and thermophilic homologous proteins from molecular dynamics and FoldUnfold method. J Bioinform Comput Biol 8:377–394.
22. Radestock S, Gohlke H (2011) Protein rigidity and thermophilic adaptation. Proteins 79:1089–1108.
23. Sterpone F, Melchionna S (2012) Thermophilic proteins: insight and perspective from in silico experiments. Chem Soc Rev 41:1665–1676.
24. Glyakina AV, Garbuzynskiy SO, Lobanov MY, Galzitskaya OV (2007) Different packing of external residues can explain differences in the thermostability of proteins from thermophilic and mesophilic organisms. Bioinformatics 23:2231–2238.
25. Taylor TJ, Vaisman II (2010) Discrimination of thermophilic and mesophilic proteins. BMC Struct Biol 10:S5.
26. Argos P, Rossmann MG, Grau UM, Zuber H, Frank G, Tratschin JD (1979) Thermal stability and protein structure. Biochemistry 18:5698–5703.
27. Plaxco KW, Simons KT, Baker D (1998) Contact order, transition state placement and the refolding rates of single domain proteins. J Mol Biol 277:985–994.
28. Lazaridis T, Lee I, Karplus M (1997) Dynamics and unfolding pathways of a hyperthermophilic and a mesophilic rubredoxin. Protein Sci 6:2589–2605.
29. Pfleger C, Rathi PC, Klein DL, Radestock S, Gohlke H (2013) Constraint network analysis (CNA): a python software package for efficiently linking biomacromolecular structure, flexibility, (thermo-)stability, and function. J Chem Inf Model 53:1007–1015.
30. Kr€ uger DM, Rathi PC, Pfleger C, Gohlke H (2013) CNA

web server: rigidity theory-based thermal unfolding simulations of proteins for linking structure, (thermo-)stability, and function. Nucl Acids Res 41:gkt292–W348.
31. Jacobs DJ, Rader AJ, Kuhn LA, Thorpe MF (2001) Protein flexibility predictions using graph theory. Proteins 44:150–165.
32. Dahiyat BI, Gordon DB, Mayo SL (1997) Automated design of the surface positions of protein helices. Protein Sci 6:1333–1337.
33. Rader AJ, Hespenheide BM, Kuhn LA, Thorpe MF (2002) Protein unfolding: rigidity lost. Proc Natl Acad Sci USA 99:3540–3545.
34. Tasumi M, Takeuchi H, Ataka S, Dwivedi AM, Krimm S (1982) Normal vibrations of proteins: glucagon. Biopolymers 21:711–714.
35. Go N, Noguti T, Nishikawa T (1983) Dynamics of a small globular protein in terms of low-frequency vibrational modes. Proc Natl Acad Sci USA 80:3696–3700.
36. Brooks B, Karplus M (1983) Harmonic dynamics of proteins: normal modes and fluctuations in bovine pancreatic trypsin inhibitor. Proc Natl Acad Sci USA 80: 6571–6575.
37. Levitt M, Sander C, Stern PS (1985) Protein normal- mode dynamics: trypsin inhibitor, crambin, ribonuclease and lysozyme. J Mol Biol 181:423–447.
Frappier 38. McCammon JA, Gelin BR, Karplus M (1977) Dynamics of folded proteins. Nature 267:585–590.
39. Levitt M (1981) Molecular dynamics of hydrogen bonds in bovine pancreatic trypsin inhibitor protein. Nature 294:379–380.
40. Karplus M, Kushick JN (1981) Method for estimating the configurational entropy of macromolecules. Macromolecules 14:325–332.
41. Adcock SA, McCammon JA (2006) Molecular dynamics: survey of methods for simulating the activity of proteins. Chem Rev 106:1589–1615.
42. Park S, Schulten K (2004) Calculating potentials of mean force from steered molecular dynamics simulations. J Chem Phys 120:5946–5961.
43. Atilgan AR, Durell SR, Jernigan RL, Demirel MC, Keskin O, Bahar I (2001) Anisotropy of fluctuation dynamics of proteins with an elastic network model. Biophys J 80:505–515.
44. Doruker P, Jernigan RL, Bahar I (2002) Dynamics of large proteins through hierarchical levels of coarsegrained structures. J Comput Chem 23:119–127.
45. Tama F, Gadea FX, Marques O, Sanejouand YH (2000) Building-block approach for determining low-frequency normal modes of macromolecules. Proteins 41:1–7.
46. Frappier V, Najmanovich RJ (2014) A coarse-grained elastic network atom contact model and its use in the simulation of protein dynamics and the prediction of the effect of mutations. PLoS Comput Biol 10: e1003569.
47. McConkey B, Sobolev V, Edelman M (2003) Discrimination of native protein structures using atom-atom contact scoring. Proc Natl Acad Sci USA 100:3215– 3220.
48. Doruker P, Atilgan AR, Bahar I (2000) Dynamics of proteins predicted by molecular dynamics simulations and analytical approaches: application to alpha- amylase inhibitor. Proteins 40:512–524.
49. Kumar S, Nussinov R (2001) How do thermophilic proteins deal with heat? Cell Mol Life Sci 58:1216–1233.
50. Beadle BM, Baase WA, Wilson DB, Gilkes NR, Shoichet BK (1999) Comparing the thermodynamic stabilities of a related thermophilic and mesophilic enzyme. Biochemistry 38:2570–2576.
51. McQuarrie DA (1976) Statistical Mechanics. Harper Collins, New York.

52. Zhang Y, Skolnick J (2005) TM-align: a protein structure alignment algorithm based on the TM-score. Nucleic Acids Res 33:2302–2309

ARTICLE 3

ENCoM server: exploring protein conformational space and the effect of mutations on protein function and stability

Auteurs de l'article: Vincent Frappier, Matthieu Chartier et Rafael Najmanovich

Statut de l'article: Publié, Frappier V, Chartier M, Najmanovich RJ, *ENCoM Server: Exploring protein conformational space and the effect of mutations on protein function and stability*, Nucleic Acids Res 2015

Avant-propos: J'ai écrit une majeure partie du manuscrit avec Matthieu Chartier et Rafael Najmanovich. J'ai effectué la majorité des méthodes expérimentales à l'exception de l'interface web qui a été créée par Matthieu.

Résumé : ENCoM est un modèle d'analyse des modes normaux de faible résolution qui comparativement aux autres approches, considèrent la nature des résidus. L'inclusion de cette information a été démontrée comme étant bénéfique à l'exploration de l'espace conformationnel et a permis d'utiliser pour la première fois une approche d'AMN afin de prédire l'effet de mutations sur la stabilité protéique et sur les propriétés dynamiques à partir d'un changement d'entropie vibrationnelle. Dans cet article, nous présentons un serveur web qui permet à des utilisateurs non experts d'utiliser ENCoM afin de prédire l'effet de mutations et de générer un ensemble de conformation réaliste. Ce serveur est accessible à: http://bcb.med.usherbrooke.ca/encom.

ABSTRACT

ENCoM is a coarse-grained normal mode analysis method recently introduced that unlike previous such methods is unique in that it accounts for the nature of amino acids. The inclusion of this layer of information was shown to improve conformational space sampling and apply for the first time a coarse-grained normal mode analysis method to predict the effect of single point mutations on protein dynamics and thermostability resulting from vibrational entropy changes. Here we present a web server that allows non-technical users to have access to ENCoM calculations to predict the effect of mutations on thermostability and dynamics as well as to generate geometrically realistic conformational ensembles. The server is accessible at: http://bcb.med.usherbrooke.ca/encom.

INTRODUCTION

Proteins are dynamic objects with movements ranging from sub-rotameric side-chain movements to domain movements intrinsically associated to their function. Among the main computational techniques to study protein dynamics are molecular dynamics (MD) and normal mode analysis (NMA).The following properties are common to both techniques: (i) both can be used to explore the conformational space; (ii) may use the same force fields and the accuracy of the simulation depends on the quality of the potential; (iii) both techniques are as exact descriptions of the dynamics as the level of detail of the representation of the protein structure and the force field used permits. The major difference between MD and NMA is that the former produces an actual trajectory in conformational space while the later produces a basis set of movements described as a set of normal modes (Eigenvectors) and associated frequencies (Eigenvalues) with which individual points in conformational space can be sampled. Every possible conformational change of a protein from the starting equilibrium structure can be described as a linear combination of all Eigenvectors modulated by specific amplitudes. Therefore, NMA produces the set of possible movements whereas MD provides an actual trajectory. The modes associated to the slowest frequencies are the most energetically accessible.

Coarse-grained NMA methods use reduced amino acids representations, for example, one point mass per amino acid. Different levels of simplification in the representation of protein

structures exist. However, with the exception of ENCoM (1) and VAMM (2), coarse-grained NMA models do not account for the nature of amino acids and are therefore sequence agnostic (3–5). For example, in the widely used Anisotropic Network Model (ANM) (4), all residues within a given distance threshold (usually 18 Å), are connected by springs with equal spring constants independent of the nature of the amino acids involved or even if they are actually interacting or not. Our group recently introduced ENCoM, a coarse-grained NMA method that accounts for the nature of amino-acids through the inclusion of a pairwise atom-type term proportional to the surface area in contact between heavy-atoms in the potential. The more realistic representation of intramolecular interactions in ENCoM results in more accurate predictions of conformational changes in terms of calculated squared overlap (the extent that movement in a given normal mode direction drives the structure from a starting state toward a target one). Specifically, compared to ANM using the PSCDB (6), a database for protein structural change upon ligand binding, we obtain an average increase in squared overlap of 28% on 117 coupled domain movements and 60% on 236 cases of coupled loop movements (1).

As the first coarse-grained NMA method to account for the type and extent of pairwise atomic interactions, ENCoM can be used to calculate vibrational entropy differences as a result of mutations (1). ENCoM was compared (1) to several existing methods, notably FoldX3.0 (7), Rosetta (8), DMutant (9) and PoPMusic (10), Eris (11), CUPSAT (12), I-Mutant (13) and AUTO-MUTE (14) on a data set of 303 manually curated mutations (10). Although not the best overall predictive method when considering both stabilizing and destabilizing mutations together, ENCoM proved to be the most self-consistent and least biased method. ENCoM and DMutant had the best performance on the subset of 45 stabilizing mutations compared to other methods whose predictions in this case were as good or worse than a random model. Classic coarse-grained NMA models predicted every mutation as neutral since by definition two identical structures irrespective of their sequences will generate identical sets of Eigenvectors and Eigenvalues. Lastly, ENCoM can predict local variations in dynamics. As a proof of concept, ENCoM was used to predict the effect of the G121V mutation in dihydrofolate reductase (DHFR) from *E. Coli*. Although this mutation has only a modest effect on protein stability and is 15 Å away from

the binding site, it disrupts catalytic activity 200 fold through allosteric effects. ENCoM calculations are consistent with experimentally observed variations in NMR S2 values (15) for the effect of this mutation with a Pearson correlation of 0.6 (1). More recently, ENCoM was used to compare a large data set of structurally identical ortholog mesophile/thermophile protein pairs to show that thermophile proteins are on average more rigid than their mesophile counterparts and that it is possible to use vibrational entropy differences to guide the selection of mutations observed between such proteins with potential uses in protein engineering (16).

The level of technical skill required to use NMA methods is a barrier to their adoption even when using freely available packages such as ENCoM. A number of existing webservers make NMA methods more accessible. These are NOMAD-ref (17), Elnemo (18), webnm@ (19), AD-ENM (20), Promode Elastic (21), ANM 2.0 (22) and iMODS (23). However, with the exception of iMODS that performs NMA in dihedral angle space, existing servers are based on Cartesian space NMA and are of limited use. In particular, some methods do not generate actual structures and those that do, provide structures that do not respect the most basic geometric constraints of bond angles and distances. Furthermore, many servers provide structures derived from single normal modes. Thus at the very least, the structures provided by existing servers require further regularization to represent geometrically plausible structures. Lastly, as a consequence of the limitations of the sequence-agnostic coarse-grained NMA models on which they are all based, none of the existing servers can be used to predict the effect of mutations on dynamics or stability. Here we present a simple, yet powerful, web-server that given a protein structure (Figure 1A) uses ENCoM to calculate coarse-grained normal modes (Figure 1B) to predict the effect of mutations on protein stability and dynamics (Figure 1C) and generate realistic, geometrically correct conformational ensembles based on the uniform exhaustive sampling of accessible modes (Figure 1D).

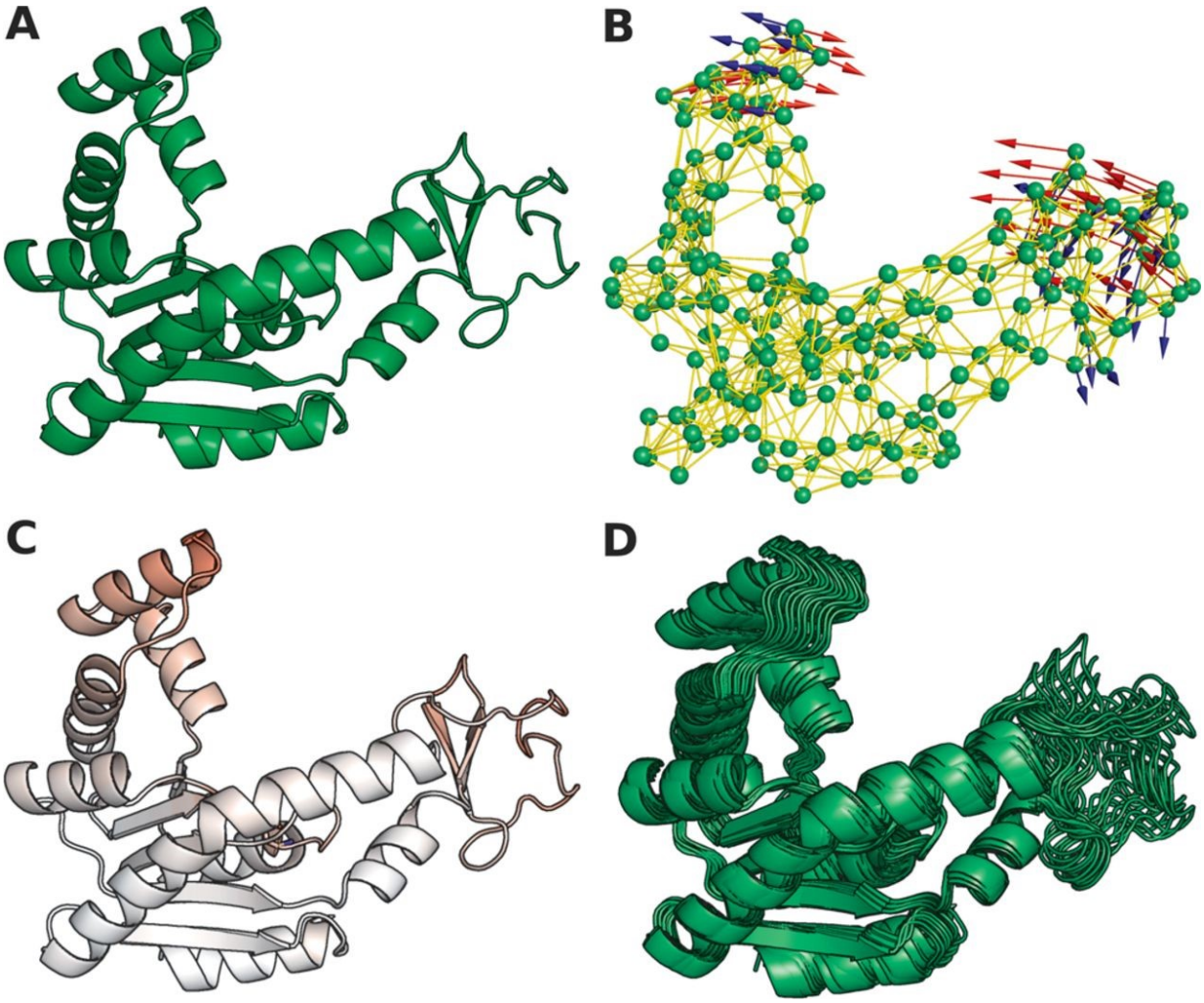


Figure 1. - Example of using the ENCoM web-server to analyze the unbound form of adenylate kinase (PDB ID 4AKE, Panel A). (B) The protein structure is represented as an elastic network model, where amino acids are represented by masses (green spheres) and interactions by springs (yellow stick). (C) The mutation K13Q causes an increase in flexibility in the regions marked in red. (D) The eigenvectors representing the two slowest non-trivial modes (shown in red and blue in panel B) were used to generate 49 conformations.

IMPLEMENTATION

The ENCoM server interface is split into two input forms, depending on what type of analysis the user wants to perform: prediction of the effect of mutations or conformational sampling. In either case, users can import their own structures in PDB format or use existing PDB accession codes. Every structure analyzed with ENCoM in the server implementation is stripped of heteroatoms and hydrogen atoms and the user can choose all

chains or any subset thereof. ENCoM takes in consideration inter-chain contacts for all selected chains. When using PDB accession codes, it is important to note that the ENCoM server uses the asymmetric unit, thus it is important to specify the correct chains to be included in the calculations or alternatively download the biological unit independently from a PDB repository and use this as input. Upon the submission of a job the user receives a link to a results page that is maintained for 7 days where a log shows the level of completion of the job whilst the job is running and the results once finished. If the user provides an email, a notification email is sent when the job starts and once the calculations are completed. Run time may vary from minutes to several hours and is estimated by comparing the input parameters (size of protein and number of mutations or conformations to be evaluated) to the time taken by jobs with similar parameters. Thus, as the ENCoM server usage increases, running time estimates will improve.

Effect of mutations on thermal stability

The mutation interface is built to allow high-throughput analysis of single-point mutations with the implementation of a simple command line interface. The user inputs the residue number or range of residues to mutate, the chain and the type of mutations. Whereas a large number of mutations may be specified at once in the input, the ENCoM server is restricted to the analysis of single mutations and models each mutation in turn as a single-point mutation. The ENCoM server utilizes the same methodology employed in our recent papers (1,16), namely we use the Modeller (24) 'mutated' function to model each mutant and ENCoM to predict the $\Delta\Delta G$ with respect to the starting structure. Particular mutations that cannot be successfully modeled are skipped. As some mutations may severely disrupt protein structure, it is necessary to keep in mind that the quality of the resulting models directly affects the calculations. While minor backbone or side-chain rearrangements are unlikely to have major effects on the analysis, if a Modeller-generated structure is entirely wrong, for example modeling a mutant that in reality cannot generate a stable fold, the results will clearly be meaningless.

As stabilizing mutations are more rare, the possibility to input a large number of mutations at once (even all 19 possibilities in each position) in the ENCoM server web interface

simplifies the search for such mutations that are particularly relevant in the context of protein design. With the exception of PoPMusic, all other web-servers dedicated to the prediction of the effect of mutations on stability require that users input every position to be mutated (and the nature of the corresponding residue in the wild type) manually, making the task of exploring the effect of mutations unnecessarily time consuming.

In our previous work (1), we tested linear combinations of ENCoM with other methods, but did not test combination that did not involve ENCoM. Here, we perform all possible method combinations between ENCoM, Eris, FoldX, Rosetta and DMutant as these were shown in our previous work to be the best performing, less biased methods on both stabilizing and destabilizing mutations. Some methods were excluded due to significant number of errors (CUPSAT and AUTO-MUTE), strong biases toward predicting mutation as destabilizing (I-Mutant and PoPMusic) and that had no predictive power (ANM and STeM). In order to determine which combination of methods is best, the 303 mutations from the Dehouck *et al.* data set (10) were used to create 10 000 bootstrapping samples. For each sample we obtained by linear regression the parameters α and β (Equation 1) that minimized the root mean square error (RMSE) between experimental and predicted $\Delta\Delta G$ values on the whole data set or the subset of stabilizing ($\Delta\Delta G < 0.5$ kcal/mol) and destabilizing ($\Delta\Delta G > 0.5$ kcal/mol) mutations as previously done (1) according to the following equation:

$$\Delta\Delta G = \alpha\Delta\Delta G_A + \beta\Delta\Delta G_B \quad (1)$$

where A and B represent different prediction methods. In the case of ENCoM, $\Delta\Delta G$ is approximated by the calculated $\Delta\Delta S_{vib}$. Variance partitioning analysis (25) was performed using the 'varpart' function in R to quantify synergy between methods. Synergistic method combinations should have low shared variance and high individual variance. The results in Table 1 represent the median parameters, RMSE, error and variance of the bootstrapped ensemble.

Table 1. RMSE and variance for different linear combinations of methods

Model		Parameters		Average bootstrapped RMSE			Variance		
A	**B**	**α**	**β**	**All**	**Stabilizing**	**Destabilizing**	**A**	**B**	**Shared**
ENCoM	FoldX	-1.12	0.38	1.24 ± 0.14	1.45 ± 0.39	1.40 ± 0.17	0.12	0.24	0.05
ENCoM	Rosetta	-1.14	0.46	1.32 ± 0.16	1.70 ± 0.41	1.43 ± 0.21	0.14	0.15	0.04
FoldX	DMutant	0.34	0.50	1.33 ± 0.15	1.42 ± 0.40	1.54 ± 0.20	0.14	0.06	0.15
FoldX	Eris	0.35	0.14	1.36 ± 0.18	1.61 ± 0.40	1.54 ± 0.26	0.16	0.01	0.12
FoldX	Rosetta	0.31	0.20	1.37 ± 0.17	1.67 ± 0.44	1.55 ± 0.23	0.10	0.00	0.18
DMutant	Rosetta	0.59	0.43	1.37 ± 0.16	1.60 ± 0.41	1.55 ± 0.24	0.10	0.07	0.11
ENCoM	Eris	-1.30	0.48	1.37 ± 0.20	1.63 ± 0.36	1.52 ± 0.35	0.16	0.12	0.01
DMutant	Eris	0.72	0.28	1.40 ± 0.19	1.48 ± 0.37	1.63 ± 0.36	0.14	0.06	0.06
Rosetta	Eris	0.41	0.15	1.43 ± 0.18	1.87 ± 0.41	1.56 ± 0.28	0.07	0.02	0.11
ENCoM	DMutant	-0.81	0.68	1.49 ± 0.17	1.40 ± 0.40	1.77 ± 0.21	0.03	0.07	0.14

Table 1 confirms our previous results showing that ENCoM has a high degree of synergy (high individual variance and low shared variance) with existing methods, that is, the combination of ENCoM to other methods is beneficial. Furthermore, combinations involving ENCoM are more beneficial than those between other methods (with the exception of the combination with DMutant). The best predictive method is a combination of ENCoM and FoldX, with the best (lowest) median RMSE values across bootstrapping samples for all mutations combined or destabilizing ones and third best median RMSE on stabilizing mutations.

Based on the results above, the final predicted ΔG values correspond to the linear combination of the predictions by vibrational entropy based ENCoM calculations and the enthalpy-based FoldX3.0 beta 6 (26) (Supplementary Figure S1). Our previous results (1) showed that most methods are biased toward the prediction of destabilizing mutations and are at best equal to random in predicting stabilizing mutations. The results in Table 1

extend our previous results in that it tests combinations of methods not involving ENCoM and confirm that the combination of ENCoM and FoldX is the best overall.

Effect of mutations on dynamics

We also predict the effect of mutations on the flexibility of individual residues via the calculation of predicted b-factors using ENCoM alone (1). For each modeled mutant we provide a coordinate file and a PyMOL script to color residues as a function of the predicted effect of the mutation on the flexibility of the entire protein with respect to the wild type (Supplementary Figure S1). Residues are colored in a blue (less flexible) to red (gain of flexibility) relative to the wild type. The color range is scaled relative to the maximum absolute difference of predicted b-factor differences between wild type and all mutants across residues or three times the standard deviation of their absolute difference, whichever is smaller. Thus, the color assignment can only be used to compare different mutations that were run as part of the same job. Predicted b-factors are provided in graphical form in the web interface and numerically in a .csv file.

Benefits of a combined analysis of stability and dynamics

The dynamic properties of individual residues are closely related to protein function. The rigidification of an enzyme may lead to a decrease in its activity (15,27,28). The increase in rigidity is often associated to an increase in thermal stability and numerous studies have analyzed the balance between protein stability and protein function (29–33). Some of the best examples come from thermophilic enzymes that display decreased activity and increased rigidity at room temperature (34). Considering that existing methods for the prediction of the effect of mutations on stability do not provide any insight into their effect on dynamics, mutations predicted with such methods need to be further analyzed or risk achieving increased stability at the expense of disrupting important aspects of dynamics. Therefore, the combined analysis of the effect of mutations on stability and dynamics with the ENCoM server may help users perform a more informed choice of mutations for protein engineering applications and understand the effect of mutations in broader terms. We provide detailed results for each mutation requested as well as a summary of the results for the top 25 stabilizing mutations with the contribution of each scoring function to the

prediction of stability and the effect of mutation on the b-factors (Supplementary Figure S1). If the input includes more than two mutated residues and two types of mutations, a heatmap is generated representing the predicted ΔΔG of every position and mutant type. This heatmap may help identify hotspots affecting stability.

Generation of conformational ensembles

Considering proteins in terms of conformational ensembles rather than single rigid structures has been shown to be beneficial in a number of different applications: small molecule docking (35–37), protein–protein docking (38,39) and protein modeling and design (40). Normal modes provide a means to generate conformational ensembles. However, care needs to be taken in the details of how the exploration of the relevant modes is performed. Existing NMA web-servers for the generation of conformational ensembles do not sample uniformly even the most accessible modes and as such provide a biased and incomplete image of the conformational landscape for a given protein structure. In particular, looking at the motions of any one mode individually does not represent the real situation where any one conformation is the linear combination of all modes. Furthermore, the models returned are the result of Cartesian translations of groups of atoms along an individual mode, thus do not respect even the most basic geometric constraints of bond angles and distances. We address these problems by performing a uniform sampling of modes selected by the user and rebuilding every conformation using Modeller with the ENCoM generated structure as a template as described below. The conformational sampling interface works as follows. First, normal modes are calculated using the input PDB structure. Second, considering that Eigenvectors are orthonormal, an ensemble of amplitudes is found for each Eigenvector that respects the maximum RMSD distortion from the input and minimum distortion between conformations selected by the user. A new position for each atom in the protein structure is calculated from the translation of its Cartesian coordinates along the direction of movement associated to each Eigenvector by an amount defined by the selected amplitude. Thus, each combination of amplitudes for the selected normal modes generates a different structure. However, some combinations of amplitudes would effectively create a final structure with RMSD larger than the maximum chosen by the user, such combinations are removed. With the methodology described

above, the ENCoM server generates an exhaustive and uniform sampling of the selected modes. Users should notice that the first six normal modes represent the translation and rotation of the entire structure and do not represent internal movements. The first mode representing internal movements is the 7th mode but users may choose as starting mode any other subsequent mode.

Depending on the initial parameters, it might be impossible to have conformations that are a combination of all the modes. The number of generated structures grows combinatorially with the number of modes. To decrease computational time, the analysis is limited to at most 350 conformations. If the choice of parameters exceeds this limit, the server automatically increases the minimum RMSD between conformations in order produce at most 350 structures, thus guaranteeing a uniform sampling of conformational space within the choice of normal modes selected by the user. Lastly, every conformation generated by ENCoM is rebuilt using Modeller with the normal-mode-generated structure serving as template. This last step maps the geometrically non-realistic conformations generated by the linear combination of amplitudes for the different normal modes into the closest possible geometrically plausible structure. In addition to obtaining a structure with correct backbone angles and distances, Modeller will repackage side-chains into the new conformation. Every conformation rebuilt is provided as an individual PDB file.

In order to provide a convenient way to visualize all the generated conformations, the set of conformations is ordered in a way to minimize the RMSD between consecutive conformations to yield a smooth trajectory. It is important to note that this trajectory represents a rough morphing between structures and not dynamics in conformational space. Mathematically, this reordering is essentially equivalent to the NP hard traveling salesman problem. However, given that this trajectory is provided only for visualization purposes, we utilize a simple heuristics to find an approximate solution. Simply, starting from every structure we generate the path that minimizes the sum RMSD values and choose that with the lowest sum. The visualization pathway is provided as a single PDB file containing consecutive states representing the ordering of individual conformations described above. We also provide a text file with the pairwise RMSD between every conformation and the

amplitude applied on each normal mode employed to build each provided model. Lastly, we also produce individual PDB files with consecutive states demonstrating each individual normal mode of movement for each of the first 20 internal modes (Supplementary Figure S2).

CONCLUSIONS

The ENCoM server offers easy access to powerful NMA-based predictions of the effect of mutations on thermostability and dynamics of single residues and generates comprehensive, geometrically realistic conformational ensembles.

SUPPLEMENTARY DATA

Supplementary Data are available at NAR Online.

ACKNOWLEDGEMENT

R.J.N. is part of the CR-CHUS, Institute of Pharmacology of Sherbrooke, PROTEO (the Québec network for research on protein function, structure and engineering) and GRASP (Groupe de Recherche Axe sur la Structure des Protéines). The authors would like to thank Dr Luis Serrano for giving his permission to use FoldX within the ENCoM server.

FUNDING

V.F. is the recipient of a PhD fellowship from the Fonds de Recherche du Québec - Nature et Technologies (FRQ-NT); M.C. is the recipient of a PhD fellowship from the Natural Sciences and Engineering Research Council of Canada (NSERC). NSERC Discovery Grant RGPIN-2014–05766. Funding for open access charge: NSERC.

Conflict of interest statement.

None declared

REFERENCES

1. Frappier,V. and Najmanovich,R.J. (2014) A coarse-grained elastic network atom contact model and its use in the simulation of protein dynamics and the prediction of the effect of mutations. PLoS Comput. Biol., 10, e1003569.
2. Korkut,A. and Hendrickson,W.A. (2009) A force field for virtual atom molecular mechanics of proteins. Proc. Natl. Acad. Sci. U.S.A., 106, 15667–15672.
3. Doruker,P., Atilgan,A.R. and Bahar,I. (2000) Dynamics of proteins predicted by molecular dynamics simulations and analytical approaches: application to alpha-amylase inhibitor. Proteins, 40, 512–524.
4. Atilgan,A.R., Durell,S.R., Jernigan,R.L., Demirel,M.C., Keskin,O. and Bahar,I. (2001) Anisotropy of fluctuation dynamics of proteins with an elastic network model. Biophys. J., 80, 505–515.
5. Lin,T.-L. and Song,G. (2010) Generalized spring tensor models for protein fluctuation dynamics and conformation changes. BMC Struct. Biol., 10(Suppl. 1), S3.
6. Amemiya,T., Koike,R., Kidera,A. and Ota,M. (2012) PSCDB: a database for protein structural change upon ligand binding. Nucleic Acids Res., 40, D554–D558.
7. Schymkowitz,J., Borg,J., Stricher,F., Nys,R., Rousseau,F. and Serrano,L. (2005) The FoldX web server: an online force field. Nucleic Acids Res., 33,W382–W388.
8. Leaver-Fay,A., Tyka,M., Lewis,S.M., Lange,O.F., Thompson,J., Jacak,R., Kaufman,K., Renfrew,P.D., Smith,C.A., Sheffler,W. et al. (2011) ROSETTA3: an object-oriented software suite for the simulation and design of macromolecules. Methods Enzymol., 487, 545–574.
9. Zhou,H. and Zhou,Y. (2002) Distance-scaled, finite ideal-gas reference state improves structure-derived potentials of mean force for structure selection and stability prediction. Protein Sci., 11, 2714–2726.
10. Dehouck,Y., Grosfils,A., Folch,B., Gilis,D., Bogaerts,P. and Rooman,M. (2009) Fast and accurate predictions of protein stability changes upon mutations using statistical potentials and neural networks: PoPMuSiC-2.0. Bioinformatics, 25, 2537–2543.
11. Yin,S., Ding,F. and Dokholyan,N.V. (2007) Eris: an automated estimator of protein stability. Nat. Methods, 4, 466–467.
12. Parthiban,V., Gromiha,M.M. and Schomburg,D. (2006) CUPSAT: prediction of protein stability upon point mutations. Nucleic Acids Res., 34,W239–W242.
13. Capriotti,E., Fariselli,P. and Casadio,R. (2005) I-Mutant2.0: predicting stability changes upon mutation from the protein sequence or structure. Nucleic Acids Res., 33,W306–W310.
14. Masso,M. and Vaisman,I.I. (2010) AUTO-MUTE: web-based tools for predicting stability changes in proteins due to single amino acid replacements. Protein Eng. Des. Sel., 23, 683–687.
15. Boehr,D.D., Schnell,J.R., McElheny,D., Bae,S.-H., Duggan,B.M., Benkovic,S.J., Dyson,H.J. andWright,P.E. (2013) A DistalMutation Perturbs Dynamic Amino Acid Networks in Dihydrofolate Reductase. Biochemistry , 52, 4605–4619.
16. Frappier,V. and Najmanovich,R.J. (2015) Vibrational entropy differences between mesophile and thermophile proteins and their use in protein engineering. Protein Sci., 24, 474–483.

17. Lindahl,E., Azuara,C., Koehl,P. and Delarue,M. (2006) NOMAD-Ref: visualization, deformation and refinement of macromolecular structures based on all-atom normal mode analysis. Nucleic Acids Res., 34,W52–W56.
18. Suhre,K. and Sanejouand,Y.-H. (2004) ElNemo: a normal mode web server for protein movement analysis and the generation of templates for molecular replacement. Nucleic Acids Res., 32,W610–W614.
19. Hollup,S.M., Salensminde,G. and Reuter,N. (2005) WEBnm@: a web application for normal mode analyses of proteins. BMC Bioinformatics, 6, 52.
20. Zheng,W. and Doniach,S. (2003) A comparative study of motor-protein motions by using a simple elastic-network model. Proc. Natl. Acad. Sci. U.S.A., 100, 13253–13258.
21. Wako,H. and Endo,S. (2011) Ligand-induced conformational change of a protein reproduced by a linear combination of displacement vectors obtained from normal mode analysis. Biophys. Chem., 159, 257–266.
22. Eyal,E., Lum,G. and Bahar,I. (2015) The anisotropic network model web server at 2015 (ANM 2.0). Bioinformatics, doi:10.1093/bioinformatics/btu847.
23. Lop´ ez-Blanco,J.R., Aliaga,J.I., Quintana-Ort´ ı,E.S. and Chac´ on,P.
(2014) iMODS: internal coordinates normal mode analysis server. Nucleic Acids Res., 42,W271–W276.
24. Webb,B. and Sali,A. (2014) Comparative Protein Structure Modeling Using MODELLER. Curr. Protoc. Bioinform., 47, 5.6.1–5.6.32.
25. Mood,A.M. (1971) Partitioning variance in multiple regression analyses as a tool for developing learning models. Am.Educ.Res.J., 8, 191–202.
9. Zhou,H. and Zhou,Y. (2002) Distance-scaled, finite ideal-gas reference state improves structure-derived potentials of mean force for structure selection and stability prediction. Protein Sci., 11, 2714–2726.
10. Dehouck,Y., Grosfils,A., Folch,B., Gilis,D., Bogaerts,P. and Rooman,M. (2009) Fast and accurate predictions of protein stability changes upon mutations using statistical potentials and neural networks: PoPMuSiC-2.0. Bioinformatics, 25, 2537–2543.
11. Yin,S., Ding,F. and Dokholyan,N.V. (2007) Eris: an automated estimator of protein stability. Nat. Methods, 4, 466–467.
12. Parthiban,V., Gromiha,M.M. and Schomburg,D. (2006) CUPSAT: prediction of protein stability upon point mutations. Nucleic Acids Res., 34,W239–W242.
13. Capriotti,E., Fariselli,P. and Casadio,R. (2005) I-Mutant2.0: predicting stability changes upon mutation from the protein sequence or structure. Nucleic Acids Res., 33,W306–W310.
14. Masso,M. and Vaisman,I.I. (2010) AUTO-MUTE: web-based tools for predicting stability changes in proteins due to single amino acid replacements. Protein Eng. Des. Sel., 23, 683–687.
15. Boehr,D.D., Schnell,J.R., McElheny,D., Bae,S.-H., Duggan,B.M., Benkovic,S.J., Dyson,H.J. andWright,P.E. (2013) A DistalMutation Perturbs Dynamic Amino Acid Networks in Dihydrofolate Reductase. Biochemistry , 52, 4605–4619.
16. Frappier,V. and Najmanovich,R.J. (2015) Vibrational entropy differences between mesophile and thermophile proteins and their use in protein engineering. Protein Sci., 24, 474–483.
17. Lindahl,E., Azuara,C., Koehl,P. and Delarue,M. (2006) NOMAD-Ref: visualization, deformation and refinement of macromolecular structures based on all-atom normal mode analysis. Nucleic Acids Res., 34,W52–W56.

18. Suhre,K. and Sanejouand,Y.-H. (2004) ElNemo: a normal mode web server for protein movement analysis and the generation of templates for molecular replacement. Nucleic Acids Res., 32,W610–W614.
19. Hollup,S.M., Salensminde,G. and Reuter,N. (2005) WEBnm@: a web application for normal mode analyses of proteins. BMC Bioinformatics, 6, 52.
20. Zheng,W. and Doniach,S. (2003) A comparative study of motor-protein motions by using a simple elastic-network model. Proc. Natl. Acad. Sci. U.S.A., 100, 13253–13258.
21. Wako,H. and Endo,S. (2011) Ligand-induced conformational change of a protein reproduced by a linear combination of displacement vectors obtained from normal mode analysis. Biophys. Chem., 159, 257–266.
22. Eyal,E., Lum,G. and Bahar,I. (2015) The anisotropic network model web server at 2015 (ANM 2.0). Bioinformatics, doi:10.1093/bioinformatics/btu847.
23. Lop´ ez-Blanco,J.R., Aliaga,J.I., Quintana-Ort´ ı,E.S. and Chac´ on,P.
(2014) iMODS: internal coordinates normal mode analysis server. Nucleic Acids Res., 42,W271–W276.
24. Webb,B. and Sali,A. (2014) Comparative Protein Structure Modeling Using MODELLER. Curr. Protoc. Bioinform., 47, 5.6.1–5.6.32.
25. Mood,A.M. (1971) Partitioning variance in multiple regression analyses as a tool for developing learning models. Am.Educ.Res.J., 8, 191–202.
26. Guerois,R., Nielsen,J.E. and Serrano,L. (2002) Predicting changes in the stability of proteins and protein complexes: a study of more than 1000 mutations. J. Mol. Biol., 320, 369–387.
27. Eisenmesser,E.Z.,Millet,O., Labeikovsky,W., Korzhnev,D.M., Wolf-Watz,M., Bosco,D.A., Skalicky,J.J., Kay,L.E. and Kern,D. (2005) Intrinsic dynamics of an enzyme underlies catalysis. Nature, 438, 117–121.
28. Doucet,N. (2011) Can enzyme engineering benefit from the modulation of protein motions? Lessons learned from NMR relaxation dispersion experiments. Protein Pept. Lett., 18, 336–343.
29. Giver,L., Gershenson,A., Freskgard,P.O. and Arnold,F.H. (1998) Directed evolution of a thermostable esterase. Proc. Natl. Acad. Sci. U.S.A., 95, 12809–12813.
30. Shoichet,B.K., Baase,W.A., Kuroki,R. and Matthews,B.W. (1995) A relationship between protein stability and protein function. Proc. Natl. Acad. Sci. U.S.A., 92, 452–456.
31. Bloom,J.D.,Meyer,M.M.,Meinhold,P., Otey,C.R., MacMillan,D. and Arnold,F.H. (2005) Evolving strategies for enzyme engineering. Curr. Opin. Struct. Biol., 15, 447–452.
32. van den Burg,B. and Eijsink,V.G.H. (2002) Selection of mutations for increased protein stability. Curr. Opin. Biotechnol., 13, 333–337.
33. Fields,P.A. (2001) Review: Protein function at thermal extremes: balancing stability and flexibility. Comp. Biochem. Physiol. A Mol. Integr. Physiol., 129, 417–431.
34. Wolf-Watz,M., Thai,V., Henzler-Wildman,K., Hadjipavlou,G., Eisenmesser,E.Z. and Kern,D. (2004) Linkage between dynamics and catalysis in a thermophilic-mesophilic enzyme pair. Nat. Struct. Mol. Biol., 11, 945–949.
35. Floquet,N.,Marechal,J.-D., Badet-Denisot,M.-A., Robert,C.H., Dauchez,M. and Perahia,D. (2006) Normal mode analysis as a prerequisite for drug design: Application to matrix metalloproteinases inhibitors. FEBS Lett., 580, 5130–5136.
36. Abagyan,R., Rueda,M. and Bottegoni,G. (2009) Consistent improvement of cross-docking results using binding site ensembles generated with elastic network normal modes. J. Chem. Inf.Model, 49, 716–725.

37. Sperandio,O.,Mouawad,L., Pinto,E., Villoutreix,B.O., Perahia,D. and Miteva,M.A. (2010) How to choose relevant multiple receptor conformations for virtual screening: a test case of Cdk2 and normal mode analysis. Eur. Biophys. J., , 39, 1365–1372.
38. Dobbins,S.E., Lesk,V.I. and Sternberg,M.J.E. (2008) Insights into protein flexibility: The relationship between normal modes and conformational change upon protein-protein docking. Proc. Natl. Acad. Sci. U.S.A., 105, 10390–10395.
39. Torchala,M.,Moal,I.H., Chaleil,R.A.G., Fernandez-Recio,J. and Bates,P.A. (2013) SwarmDock: a server for flexible protein-protein docking. Bioinformatics, 29, 807–809.
40. Smith,C.A. and Kortemme,T. (2008) Backrub-like backbone simulation recapitulates natural protein conformational variability and improves mutant side-chain prediction. J. Mol. Biol., 380, 742–756.

DISCUSSION

Champ de force

ENCoM et les différents modèles d'AMN de faible résolution se distinguent par leurs différents potentiels qui décrivent les interactions entre chaque degré de liberté d'un système. Le modèle le plus populaire jusqu'à présent est l'« Anisotropic Network Model » qui utilise un champ de force relativement simple en connectant toutes les masses situées dans un rayon prédéterminé par des ressorts aux constantes uniformes. Cette propriété est incompatible avec les champs de force contemporains qui ont démontré que les structures protéiques sont contraintes par des interactions covalentes, des repliements d'angle, des rotations autour d'angle dièdre et des interactions de longues portées. Le potentiel de STeM (et en partie d'ENCoM) pallie cette lacune par l'intégration de tous ces termes dans son potentiel, à l'exception des interactions coulombiennes qui ne sont pas harmoniques. Malheureusement, ce type de potentiel est seulement valide dans un contexte où tous les atomes sont considérés et ne s'applique pas à un modèle de faible résolution où les résidus sont représentés par une seule masse. En effet, la relation entre 2 résidus séparés par 2 autres résidus dans la séquence primaire n'est pas caractérisée par une rotation d'angle dièdre autour des masses centrales, cette rotation s'applique plutôt à deux résidus consécutifs dans la séquence. Cependant, ces termes permettent de générer des modèles plus complexes qui capturent des interactions distinctes de résidus à proximité dans la séquence. En effet, des interactions covalentes décrivent les interactions $i, i+1$, les repliements d'angle les interactions $i, i+2$, les rotations d'angle des interactions $i, i+3$ alors que les interactions de longues portées décrivent des interactions $i, i > 3$. Ce gain de complexité est cohérent avec les résultats obtenus par un autre groupe qui, en inversant des matrices de covariances de trajectoires de RMN, a obtenus des matrices Hessiennes « parfaites » (Lezon et Bahar, 2010). Ces matrices possèdent des constantes d'interaction spécifique aux structures secondaires et à la distance dans la séquence primaire qui, par exemple, connectent fortement des résidus d'hélices alpha séparées par un tour d'hélice. À la lumière de ces résultats, il serait intéressant d'ajouter un terme dans la fonction de potentiel d'ENCoM qui considère la nature des structures secondaires. En effet, en se

référant au diagramme de Ramachandran, certaines structures secondaires contraignent beaucoup plus les conformations accessibles du squelette peptidique et devraient diminuer la flexibilité de ces régions. Une plus grande constante de ressort pourrait être utilisée pour caractériser ce phénomène.

Malgré l'utilisation de termes plus complexes et réalistes, le potentiel de la méthode STeM n'est toujours pas en mesure de distinguer les différents types de résidus impliqués dans les interactions ni l'orientation de leur chaîne latérale. Le champ de force d'ENCoM qui est inspiré de ce dernier inclut cette information par la modulation du terme d'interaction de longue portée en fonction des surfaces en contact entre différents types d'atomes des résidus. Cette modification qui englobe les interactions de tous les atomes d'un résidu permet d'obtenir une résolution atomique tout en conservant la rapidité de calcul associée aux modèles de faible résolution. Bien qu'il existe plusieurs fonctions énergétiques qui caractérisent la force d'interaction entre deux résidus, nous avons opté pour dériver notre propre alogrithme afin de maintenir une flexibilité plus importante dans la paramétrisation du potentiel d'ENCoM. Bien que plus complexes à calculer, les surfaces en contact semblent plus performantes que les distances entre atomes utilisés par ces autres algorithmes de pointages. En effet, elles se sont montrées comme étant plus robustes dans la prédiction de pose de petites molécules (Gaudreault et Najmanovich, 2015) et elles détectent mieux la densité autour des atomes (Ying *et al.*, 2015), une propriété suggérée comme étant liée au pouvoir prédictif de l'AMN. Finalement, les surfaces en contact réussissent à reproduire un effet bouclier qui masque l'interaction entre deux résidus à proximité, mais séparée par un autre résidu. Par exemple, les potentiels classiques vont connecter des résidus séparés par une dizaine d'Ångstroms, traversant ainsi plusieurs couches d'atomes avoisinants.

La recherche exhaustive de 28,561 différentes combinaisons de valeurs pour les paramètres a démontré la robustesse du potentiel d'ENCoM. En effet, la contribution de chacun des termes a été modifiée par des facteurs 10 sans nécessairement affecter drastiquement la capacité de prédiction d'ENCoM lors de divers tests. C'est également le cas pour la matrice d'interaction qui caractérise les interactions entre les différents types d'atomes. Dans les

ordres de grandeur testés, elle semble peu influencer le pouvoir prédictif d'ENCoM et sa spécificité confère une augmentation marginale au pouvoir prédictif. La robustesse du champ de force suggère que les déterminants majeurs de l'AMN proviennent de la conformation de la protéine et ensuite des interactions entre les différents degrés de liberté. Le potentiel d'ENCoM est validé et paramétré pour une application aux structures protéiques. Éventuellement, on pourrait ajouter des termes pour les macromolécules d'ADN et d'ARN afin d'augmenter son champ d'applications. Certains groupes ont déjà démontré que l'AMN avait des applications intéressantes sur ce type de molécules (Dykeman et Twarock, 2010).

Validation

Les différents champs de force utilisés lors de l'AMN vont prédire différents mouvements harmoniques et différentes entropies vibrationnelles. Classiquement, elles étaient comparées au facteur B (*b-factors*) des structures cristallines (Eyal *et al.*, 2006) et plus récemment à la prédiction de changement conformationnel entre deux ou plusieurs structures d'une protéine cristallisée dans différentes conditions (Na *et al.*, 2014). Étant donné qu'ENCoM intègre l'information de la séquence, ce modèle est également évalué sur sa capacité à prédire l'effet de mutations sur les propriétés dynamiques d'une structure. Chacune des validations possède ses forces, faiblesses et biais caractéristiques de l'approche expérimentale. Ainsi, une bonne compréhension de ces méthodes est cruciale à la mise en perspective des performances d'ENCoM ainsi qu'à la découverte de possibles amélioration ou applications. De plus, bien qu'en temps normal l'AMN prédit des mouvements vibrationnels, les nombreuses simplifications apportées aux champs de force modulent la validité de cette théorie. Ainsi, la corrélation avec différentes valeurs expérimentales dynamiques aide à mieux comprendre la signification de ces prédictions.

Facteurs B

ENCoM a premièrement été validé par sa capacité à prédire des facteurs B (ou facteur de Debye–Waller). Ce facteur est attribué pour chaque atome d'une structure cristalline et il est modélisé selon :

$$B = 8\pi^2\langle r^2\rangle$$

où $\langle r^2\rangle$ représente la déviation moyenne quadratique de la position de l'atome. Ils correspondent alors à la diagonale de la matrice de covariance obtenue par une AMN. Cependant, plusieurs données expérimentales remettent en question la validité des facteurs B. En effet, par leur définition théorique, il s'agit d'un paramètre qui ne considère que des mouvements harmoniques isotropes dépendant de la température alors qu'en réalité l'incertitude de la position des atomes est assujettie à l'empaquetage des cristaux, aux contacts entre les treillis, aux mouvements anharmoniques, aux mouvements anisotropiques, aux conformations alternatives, aux mouvements de corps rigides et à la température des cristaux. Il n'est alors pas rare d'observer des profils de facteurs B différents pour une même protéine dans la même conformation, mais cristallisée dans différentes conditions ou par différents groupes.

La quasi-totalité (95%) des structures cristallines est obtenue par cryogénie (Garman, 2003). Ces températures extrêmes altèrent leurs propriétés dynamiques par une densification de la structure et par une modulation non linéaire et non proportionnelle des facteurs B (Tilton *et al.*, 1992). En dessous du point de vitrification, qui est typiquement de 180 Kelvins, les propriétés du solvant présent dans le cristal sont altérées et la variation des facteurs B de la protéine n'est plus dépendante de la température et demeure constante (Teeter *et al.*, 2001). En effet, des études de dispersement inélastiques d'électron (Paciaroni *et al.*, 2002), des approches RMN (Lee et Wand, 2001), de capacité calorifique (Miyazaki *et al.*, 2000) et des dynamiques moléculaires (Joti *et al.*, 2005) ont également mis en évidence ce phénomène. La protéine serait alors emprisonnée dans un puits énergétique qui limite l'amplitude des mouvements par la formation d'une cage de solvant (Kim *et al.*, 2011). Ce phénomène est amplifié en fonction de la concentration et de la viscosité de ce dernier. De plus, l'effet de la vitrification n'influence pas uniformément les facteurs B et affecte plus fortement les atomes exposés au solvant (Tilton *et al.*, 1992). Ainsi, la corrélation des facteurs B prédits par un modèle AMN serait alors influencée par la température de cristallisation et également par le contenu en solvant du cristal. Aucun groupe n'a cependant tenté de valider ce biais par AMN. Il pourrait facilement être vérifié

en comparant les performances d'un modèle AMN sur une même protéine, mais cristallisé à différente température (Fraser *et al.*, 2011). En temps normal, les performances devraient être supérieures à température pièce où les facteurs B ne sont pas influencés par la vitrification.

Également, des analyses mathématiques ont démontré que les facteurs B sont principalement dominés par des mouvements de corps rigides (« rigid body motion ») qui par définition ne sont pas décrits par l'AMN. Il s'agit en effet de mouvements de translation et rotation de la protéine et représentent alors les modes non triviaux ignorés dans la formation de la matrice de covariance utilisée pour prédire les facteurs B. En corrigeant ces biais, les performances d'AMN augmentent en moyenne de 50% (Soheilifard *et al.*, 2008). Cependant, cette correction ne représente pas un contexte biologique et ne génère pas nécessairement des modèles plus réalistes. La paramétrisation d'ENCoM a observé ce phénomène dans la dichotomie des performances de la prédiction de facteur B et les autres critères d'évaluation. En effet, une augmentation de l'importance des interactions de longues portées a pour effet de rigidifier la structure et de la transformer en un bloc inélastique qui imite alors des mouvements de corps rigides.

Les facteurs B considèrent que les mouvements dans la structure cristalline sont isotropes, c'est-à-dire qu'ils ne possèdent aucune direction, alors que les mouvements prédits par l'AMN sont principalement anisotropes. Sans surprise les facteurs B anisotropes retrouvés dans des structures de hautes résolutions sont beaucoup mieux décrits par des modèles AMN (Yang *et al.*, 2009) que des facteurs B isotropes. Finalement, les facteurs B ne décrivent qu'une petite partie des propriétés dynamiques d'une structure. Ils ne sont que la diagonale de la matrice de covariance et ils ne peuvent pas directement fournir d'information sur la corrélation entre les différents degrés de liberté. Cette covariance est pourtant essentielle au contexte biologique en élucidant par exemple des mécanismes allostériques de régulation (Selvaratnam *et al.*, 2011). Ainsi, une bonne corrélation avec les facteurs B n'indique pas nécessairement que les mouvements prédits par le modèle AMN représentent entièrement les propriétés dynamiques du cristal ou de la protéine. Une comparaison avec des valeurs obtenues par RMN est beaucoup plus appropriée et

représentative de conditions biologiques. En effet, pour une même protéine, les modèles d'AMN corrèlent beaucoup mieux (25%) avec les propriétés dynamiques obtenues par RMN que par cristallographie (Yang *et al.*, 2007).

Malgré les nombreuses limites associées aux facteurs B, leurs valeurs sont partiellement en accord avec d'autres mesures expérimentales et capturent des propriétés dynamiques biologiquement importantes (Parthasarathy et Murthy, 2000; Reetz *et al.*, 2006). Par exemple, autant les valeurs structurelles (Andrec *et al.*, 2007) que dynamiques (Buck *et al.*, 1995) corrèlent avec des données RMN qui décrivent un contexte beaucoup plus réaliste. Les structures cristallines composent 90% de toutes les structures déposées dans la Protein Data Bank (Berman *et al.*, 2000) et représentent alors une source importante, diversifiée et non négligeable d'information qui devrait être utilisée pour paramétrer les modèles d'AMN. Cependant, une corrélation parfaite des facteurs B avec les propriétés dynamiques prédites par un modèle n'est pas essentielle, mais un certain minimum est requis. À l'inverse, une corrélation parfaite n'est pas souhaitée, car elle pourrait tenter d'imiter les biais causés par la cristallographie qui ne représente pas nécessairement des mouvements harmoniques biologiquement importants. Par exemple, la modulation des paramètres du champ de force d'ENCoM peut générer de meilleurs résultats de corrélation qui se fait au détriment de la prédiction de changements conformationnels et de l'effet des mutations sur la stabilité protéique. Somme toute, ENCoM possède des pouvoirs prédictifs similaires à l'AMN et légèrement inférieurs à STeM.

Chevauchement

Il est estimé que les données expérimentales de facteurs B cristallographiques représentent des mouvements de l'ordre de la pico et de la nano seconde (Clore et Schwieters, 2006), ne représentent qu'un état macroscopique et fournissent peu d'information sur les mouvements plus lents qui explorent différents états biologiquement pertinents (Henzler-Wildman et Kern, 2007). Afin d'étudier les performances des modèles AMN sur la prédiction de propriétés dynamiques se produisant sur des échelles de temps supérieures ou égales à la microseconde, on étudie leur capacité à décrire des changements conformationnels entre deux états macroscopiques.

Ce changement conformationnel est décrit par un vecteur cartésien qui est comparé individuellement aux modes prédits par l'AMN sous la forme d'un chevauchement (« overlap »). Une valeur unitaire représente une description parfaite du changement conformationnel par un mode alors qu'une valeur de zéro indique que le mode est orthogonal à la distorsion. En général, les mouvements de domaines coopératifs sont mieux décrits que les mouvements de boucle locale et les modes les plus lents décrivent mieux les changements entre différents états macroscopiques. Certains groupes de chercheur émettent l'hypothèse que ces phénomènes proviennent de pressions évolutives qui favoriseraient des états de transitions ou des états biologiques de faible énergie (Echave, 2008). En effet, pour une même énergie, les modes les plus lents peuvent effectuer un plus grand déplacement qu'un mode plus rapide, couvrant ainsi un plus grand espace conformationnel et pour une distorsion équivalente, atteint un état de transition de plus faible énergie. Ces états seraient alors plus probables et faciliteraient le changement conformationnel entre deux formes par une vitesse de transition plus élevée et biologiquement plus efficace. Ce principe est également en accord avec le modèle proposé par Slater il y a presque 60 ans qui postule que les réactions chimiques se produisent le long des modes normaux (Slater, 1948, 1953). En effet, ces processus requièrent des mouvements coordonnés précis qui ne peuvent etre expliqués par des mouvements aléatoires thermiques.

Comme tous les modes de résonances sont orthogonaux, tout changement conformationnel peut être décrit par un ensemble unique d'amplitude appliqué sur chacun des modes. En fonction des fréquences de résonance de chaque mode et de leur amplitude, il est possible d'obtenir l'énergie associée à ce changement conformationnel sur la surface énergétique décrite par l'AMN. Cette énergie est aussi véridique que le potentiel utilisé, c'est-à-dire un potentiel harmonique ayant un état macroscopique unique. Cette approximation n'est pas nécessairement vraie et il est probablement plus réaliste de considérer un mouvement anharmonique qui relie deux ou plusieurs états macroscopiques. Ainsi, l'évaluation d'une AMN par le chevauchement d'un seul mode à partir d'un état macroscopique fait abstraction de ce principe. Une approche qui considère tous les mouvements des deux structures et de leur fréquence de résonance serait alors plus appropriée, telle qu'une convolution des fonctions de densités de probabilité des deux états macroscopiques. En

effet, cette nouvelle fonction représente la probabilité de la différence de conformation entre les deux états. Ainsi, la valeur de densité pour le vecteur nul de cette fonction représente la probabilité que les deux états macroscopiques adoptent les mêmes conformations qui sont alors des états de transition. Il s'agit donc de la probabilité de transition. Un modèle d'AMN qui décrit bien l'espace conformationnel de chacune des structures et leur transition possédera des probabilités plus élevées de transition. De plus, la nouvelle moyenne de cette convolution représente l'état intermédiaire le plus probable et cette conformation pourrait être utilisée dans un contexte de design protéique afin de diminuer la barrière énergétique de transition dans des approches d'optimisation *multistate* .

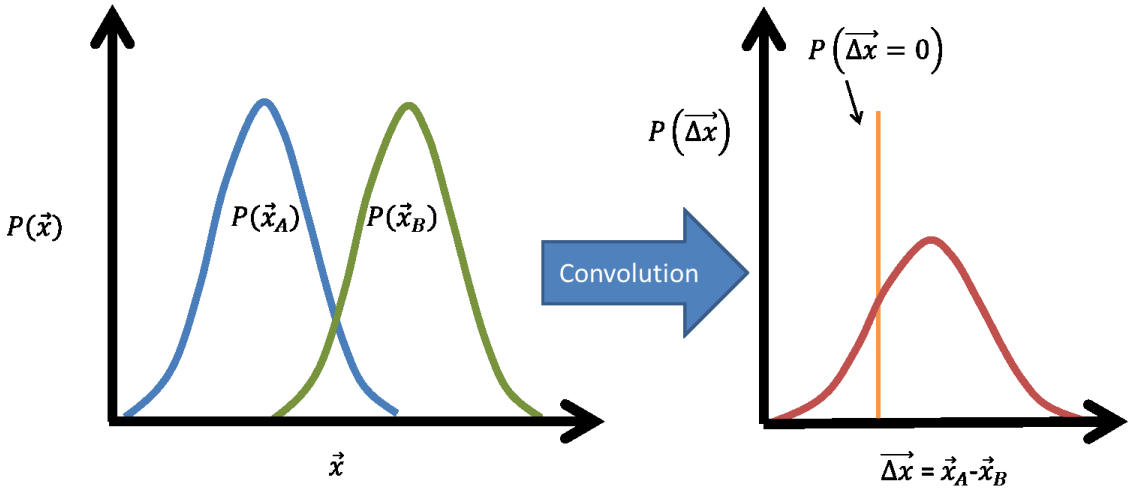


Figure 5.1 – Convolution de deux fonctions de densité de probabilité
Une protéine possède deux états macroscopique (A et B) de conformation $\vec{x}_A$ et $\vec{x}_B$. En considérant un potentiel harmonique, les propriétés dynamiques de ces conformations peuvent être représentées par les fonctions de densité de probabilité $P(\vec{x}_A)$ et $P(\vec{x}_B)$. La convolution de ces fonctions représente la fonction de densité de probabilité de la différence de conformation entre ces deux états ($\overrightarrow{\Delta x}$). La probabilité au vecteur nul représente la proportion de l'état macroscopique A et B qui possède la même conformation.

Les protéines opèrent principalement leur changement de conformation par des rotations autour d'angles dièdres du squelette peptidique, alors que les mouvements prédits par l'AMN sont des translations d'atomes dans l'espace cartésien. Ces distorsions sont irréalistes et énergiquement défavorables, principalement au niveau de l'étirement et de la compression des liens covalents. Pour pallier en partie ce problème, nous avons décidé de reconstruire les structures prédites par AMN avec le logiciel « Modeller » qui possède un potentiel plus réaliste et plus robuste (Eswar *et al.*, 2008). Ce processus s'apparente à une minimisation de la structure et génère des modèles physiquement réalistes qui peuvent par la suite être utilisés par d'autres outils bio-informatiques. Ces outils voient leur pouvoir prédictif augmenter lors de la prise en considération d'un ensemble de modèles générés à partir d'AMN, autant dans la prédiction de poses de petites molécules (Dietzen *et al.*, 2012) ou d'interactions protéine-protéine (Sankar *et al.*, 2015) que dans le design protéique (Friedland *et al.*, 2009). Certaines approches ont également démontré que l'utilisation de ces structures pour générer de nouvelles AMN permettait de mieux explorer la surface énergétique (Schuyler *et al.*, 2009) et même reproduire des repliements de protéines (Williams et Toon, 2010). Cette approche itérative considère alors que l'espace conformationnel possède plusieurs états macroscopiques intermédiaires. Afin d'éliminer le problème de distorsion de liens covalents, d'autres modèles AMN utilisent des coordonnées internes de rotation d'angle dièdre comme degrés de liberté (Frezza et Lavery, 2015; Kamiya *et al.*, 2003). Cependant, l'harmonicité de ces modèles est valide sur de petites amplitudes et nécessite une dérivation numérique de la matrice Hessienne, ce qui complexifie son application à des contextes de hauts débits.

Considérant que la méthode de validation des modèles d'AMN n'utilise qu'un état macroscopique, qu'elle ignore les mouvements anharmoniques et qu'elle crée des distorsions de liens covalents, une description parfaite de changements conformationnels par quelques mouvements de résonance n'est pas nécessairement souhaitée afin d'éviter d'introduire un biais dans le modèle. Cependant, tout comme les facteurs B, une performance adéquate demeure utile, surtout dans la description de mouvements se produisant sur de longues échelles de temps. En effet, il existe peu de méthodes qui peuvent les décrire efficacement, car la dynamique moléculaire exige trop de ressources

informatiques pour explorer en profondeur ces mouvements (Powell, 2011) et les modèles générés à partir de contraintes provenant de RMN ne peuvent détecter efficacement ces structures plutôt rares (Bakan et Bahar, 2009).

Comparativement aux autres méthodes testées, ENCoM décrit mieux les changements de conformations suite à la liaison de ligand, autant au niveau des domaines que des mouvements de boucles. Ainsi, étant donné qu'il décrit de façon plus réaliste l'espace conformationnel, son utilisation dans la génération de nouvelles conformations augmente les probabilités d'observer des états biologiquement pertinents. Cependant, cette recherche d'états de façon efficace demeure ardue. Bien qu'elle puisse être décrite par une distorsion sur une dizaine de modes, la sélection des amplitudes et des modes à utiliser est contrainte à une dimensionnalité exponentielle qui limite la recherche à quelques dimensions. Ainsi, la plupart des serveurs d'AMN génèrent des modèles qui utilisent seulement une fréquence de résonance, ce qui biaise la représentation de l'espace conformationnel. Le serveur ENCoM quant à lui va exhaustivement générer des conformations en utilisant quelques modes et choisissant automatiquement des paramètres qui limitent le nombre de conformations. Alternativement, en postulant que les états biologiques d'intérêt ont des énergies plus favorables, des algorithmes d'optimisation pourraient être utilisés afin de chercher des combinaisons de modes et amplitudes qui minimisent une fonction de score provenant d'un potentiel plus réaliste et de structures minimisées. En effet, la génération de modèles par AMN est relativement rapide, cependant la minimisation de ces structures et l'évaluation de leur énergie sont prohibitives. Ainsi, une recherche exhaustive ne serait pas efficace et demanderait beaucoup de temps de calculs. Les algorithmes d'optimisation et de recherche permettraient d'aborder en partie ce problème.

Mutations

À moins d'une modification importante de la structure du squelette peptidique, les approches actuelles d'AMN prédiront qu'une mutation n'aura pas d'effet sur la surface énergétique d'une protéine, alors qu'en réalité elles affectent leurs propriétés dynamiques et leur stabilité. Cependant, en intégrant l'information des chaînes latérales dans son potentiel, ENCoM est en mesure de détecter des modulations de la séquence qui modifiera les

mouvements prédits par une perturbation de la matrice Hessienne. Ainsi, à partir des fréquences de résonance prédites sur une structure mutante et sauvage, il est possible d'obtenir une différence d'entropie vibrationnelle:

$$\Delta S_{Vib,A,B} = \sum_{i,\lambda_i \neq 0}^{N} \ln \lambda_{i,B} - \sum_{i,\lambda_i \neq 0}^{N} \ln \lambda_{i,A}$$

Ces différences prédites par ENCoM corrèlent de façon significative avec des variations d'énergie libre de Gibbs de 303 mutations (Frappier et Najmanovich, 2014). Il s'agit de la première description dans la littérature de la prédiction de l'effet de mutations à partir d'un score entropique dérivé d'un modèle d'AMN. Comparativement à des méthodes spécifiquement développées pour cette application, sur un même ensemble de mutations, ENCoM performe aussi bien que la plupart d'entre elles et performe légèrement moins bien que deux approches populaires : FoldX et Rosetta. Cependant, les méthodes d'apprentissage automatisé testées ont des pouvoirs prédictifs largement supérieurs à toutes les autres méthodes. Toutefois, une analyse plus approfondie des résultats met en évidence un biais important dans l'interprétation des mesures de performance. En effet, plus de la moitié des mutations dans le jeu de données de tests sont déstabilisantes. Ainsi un modèle qui performe bien sur ce type de perturbations sera artificiellement meilleur. Lorsque l'analyse est restreinte aux mutations stabilisantes, à l'exception d'ENCoM et DMutant, les performances des modèles diminuent de façon drastique et performent aussi bien, sinon pire, qu'un contrôle négatif aléatoire. Ces mutations favorables sont d'un grand intérêt dans l'industrie biotechnologique étant donné qu'elles confèrent les caractéristiques recherchées dans le design protéique. En effet, une résistance importante à la dénaturation corrèle avec de plus hautes expressions de protéines, une augmentation de leur durée de vie dans des conditions semi-dénaturantes et augmente leur capacité à rester active dans des solvants non aqueux (Ferdjani *et al.*, 2011). Pour des molécules utilisées dans un contexte thérapeutique, des protéines plus stables possèdent généralement des temps de demi-vie plus élevés et une meilleure efficacité thérapeutique (D. Gao *et al.*, 2009). En ingénierie des protéines, les mutations qui confèrent des gains de fonctions sont souvent déstabilisantes, ainsi en augmentant leur stabilité avec des mutations stabilisantes, ces protéines sont plus

facilement modifiables (Sikosek et Chan, 2014) et offrent un meilleur gabarit initial dans un design rationnel (Bloom *et al.*, 2006).

La diminution de performance sur un ensemble de données moins biaisé a également été observée lors de l'évaluation de toutes les mutations simples provenant de la protéine TEM-1 (Figliuzzi *et al.*, 2015) où l'approche PoPMuSiC corrèle à 0.14 avec les valeurs expérimentales, soit pratiquement le tiers de ce qui est rapporté dans la publication originale du modèle. Cette corrélation est équivalente à des performances obtenues par un modèle naïf provenant d'une matrice de substitutions BLOSUM (Henikoff et Henikoff, 1992) qui ne possède aucune information sur le contexte structural ou le contexte de la séquence. Le même genre de phénomène a été observé lors du dernier concours « Critical Assessment of PRediction of Interactions » (CAPRI) (Moretti *et al.*, 2013), où 22 groupes ont eu à prédire l'effet qu'auront 2900 mutations sur des interactions protéine-protéine. Seulement 2 groupes ont été meilleurs qu'une matrice BLOSUM et seulement 3 groupes ont été plus performants dans la découverte de mutations stabilisantes. Ces résultats suggèrent que l'ensemble de mutations utilisé pour valider ENCoM est biaisé et qu'une validation plus robuste doit être effectuée sur des ensembles de données plus vastes représentant mieux toutes les mutations possibles chez un gène.

Afin de contrôler en partie le biais des jeux de données dans les métriques d'évaluation de performance, nous avons opté pour une déviation moyenne des valeurs prédites aux valeurs expérimentales en considérant une équation linéaire qui traverse l'origine. Cette approche contraste avec ce qui est majoritairement utilisé dans la littérature, soit la corrélation entre les valeurs prédites et expérimentales de variation d'énergie libre de Gibbs. La corrélation est biaisée étant donné que ce modèle ne prend en considération que la covariance des valeurs et ignore leurs valeurs absolues. Ainsi, une méthode qui est systématiquement biaisée vers des mutations déstabilisantes pourrait obtenir une bonne corrélation, aussi longtemps qu'elle classe les mutations stabilisatrices comme étant les moins déstabilisantes. Cependant, ce modèle serait physiquement et mathématiquement erroné, étant donné qu'une mutation neutre (ou même une absence de mutations) sera automatiquement prédite comme déstabilisante. Également, le biais des approches

prédictives est mis en lumière par la méthodologie proposée par (Thiltgen et Goldstein, 2012) en utilisant des paires de structures cristallines dont la séquence diffère seulement par une mutation. Ainsi, la mutation de chaque structure vers l'autre devrait suivre la relation :

$$\Delta\Delta G_{A\rightarrow B} = -\Delta\Delta G_{B\rightarrow A}$$

À l'exception d'ENCoM, les méthodes testées dévient de cette équation et indiquent qu'elles sont biaisées ou inconstantes. Cet effet est encore plus prononcé pour les approches d'apprentissage automatisé qui prédisent toutes mutations comme étant déstabilisantes, alors que seulement une des deux provenant d'une paire de structure devrait l'être. Ces résultats concordent avec une récente étude qui a également démontré que ces approches sont surentraînées, en ayant une corrélation moyenne de 0.88 sur des mutations déstabilisantes alors que lorsqu'elles sont testées sur un ensemble de mutations stabilisatrices, leurs corrélations moyennes atteignent 0.03 (Fang, 2015). Ces modèles mathématiques complexes sont composés de plusieurs centaines (sinon des milliers) de paramètres optimisés afin de minimiser l'erreur sur la prédiction de valeurs expérimentales. Cette complexité permet alors de capturer des relations non linéaires multidimensionnelles, mais également elle expose ces modèles au phénomène de surapprentissage (*overfitting*). Ce biais se produit lorsque la complexité du modèle est trop grande et qu'il décrit alors le bruit de l'ensemble de données au lieu du vrai signal. Pour contrer ce phénomène, il faut soit réduire la complexité du modèle et ses performances ou bien augmenter la quantité de données expérimentales utilisée pour le paramétrer.

Il existe également un biais dans les ensembles de données utilisées pour paramétrer les approches, ce qui diminue leurs pouvoirs de généralisation. En effet, afin de pouvoir être appliquées dans différents contextes, les valeurs utilisées pour l'entraînement doivent représenter l'application voulue. Ainsi un modèle qui est utilisé pour prédire l'effet de toutes les mutations simples possibles sur une structure, doit avoir été paramétré sur un ensemble de mutations aléatoires qui représente statistiquement la composition moyenne en acide aminés d'une protéine, la proportion de résidus exposés au solvant, la proportion de structures secondaires et les différents niveaux de conservation d'acides aminés.

Malheureusement, les jeux de données expérimentales utilisés proviennent de la littérature et sont des mutations rationnellement choisies par des expérimentateurs qui sont intéressés par certains types de mutation et qui ne sont alors pas représentatifs du hasard. En analysant les données de ProTherm, on observe que respectivement 28 % et 16% des mutations impliquent des résidus d'alanine et de valine, alors qu'en réalité ces acides aminés ne composent que 9% et 8% des protéines. De plus, les mutations retrouvées dans cette banque de données sont principalement déstabilisantes avec un effet moyen 1.60 kcal/mol (Wickstrom *et al.*, 2012), alors qu'en général la plupart des mutations sont neutres (Firnberg *et al.*, 2014; Thyagarajan et Bloom, 2014). Finalement, les mutations très déstabilisantes (> 5 kcal/mol) sont difficilement caractérisables par les approches expérimentales étant donné qu'il existe peu de protéine dans la forme native en condition non dénaturante ou elles ne sont tout simplement pas solubles. Les approches d'apprentissage automatisé pourront difficilement prédire ces types de mutations étant donné qu'elles n'auront jamais eu d'exemples pour apprendre leur déterminant. Ainsi, des contrôles importants doivent être effectués afin de limiter tous ces biais et créer des modèles généralisables et applicables dans différents contextes. Tous ces biais sont également vrais pour les méthodes empiriques lorsque leurs termes énergétiques sont paramétrés. Cependant, le danger de surentraînement est moindre étant donné que le nombre de paramètres à optimiser est nettement plus petit, se limitant à une dizaine de termes. ENCoM quant à lui est encore moins sensible à ces problèmes, étant donné que seulement 4 termes ont été paramétrés et qu'ils sont optimisés sur plusieurs tests indépendants, soit l'effet de mutations, les changements conformationnels et les facteurs B. La matrice d'interaction des 8 types d'atomes n'a pas été paramétrée lors de la création d'ENCoM et est dérivée de façon empirique (Sobolev *et al.*, 1996). Dans les ordres de grandeurs testés, elle semble avoir peu d'impact sur la performance et la capacité de généralisation d'ENCoM. Elle est ainsi peu assujettie à un effet de biais.

Également, la robustesse d'ENCoM peut provenir de son potentiel plus permissif qui est moins sensible aux interactions négatives. Ainsi, des mutations stabilisantes qui nécessitent des réorganisations précises et importantes de l'environnement local sont sensibles aux mauvaises modélisations en raison de la complexité des réarrangements. Bien qu'il existe

une conformation d'énergie favorable dans des potentiels plus sensibles et complexes, cette dernière n'est jamais explorée et ces mutations seront alors considérées comme déstabilisantes. Étant donné qu'ENCoM ne peut détecter d'encombrement stérique ou bien d'interaction coulombienne, ces mutations auront des effets moins drastiques et seront tempérées. De plus, il a été récemment démontré que les surfaces en contact étaient moins sensibles aux préconfigurations de site de liaison lors de la prédiction de pose de petites molécules (Gaudreault et Najmanovich, 2015), suggérant qu'une partie de la robustesse d'ENCoM provient de cette représentation. Finalement, les approches enthalpiques ne prédisent que des effets locaux, alors qu'ENCoM prend en considération toute la protéine dans l'évaluation de l'entropie vibrationnelle, pouvant ainsi détecter des effets allostériques compensatoires en considération le contexte global de la structure (Tzeng et Kalodimos, 2012). De cette façon, une même mutation aura un effet différent si elle se produit dans une boucle à proximité de la surface qu'à l'intérieur de la protéine. Cette hypothèse pourrait facilement être validée en testant les performances d'ENCoM lorsque des régions distales de la structure protéique sont artificiellement enlevées de l'analyse. Il est alors attendu que les performances devraient diminuer lorsque le contexte global est ignoré.

La robustesse d'ENCoM est également observée lors de la comparaison de la flexibilité de paires de structures d'homologues mésophiles et thermophiles. En effet, pour une même structure, plusieurs valeurs d'entropie ont été obtenues, une provenant de la structure cristalline et 100 autres provenant de modèles par homologie basée sur cette structure. Ces modèles possèdent différentes configurations de chaîne latérale et de squelettes peptides qui affectent les interactions du champ de force. Malgré tout, les énergies entropiques des structures cristallines corrèlent à 33% avec la moyenne des valeurs d'un ensemble de modèles par homologie et le signe de la différence d'énergie a été conservé dans tous les 314 cas comparés. Également, en considérant qu'il est généralement accepté que les protéines des thermophiles sont plus rigides que leurs homologues mésophiles à une même température, ENCoM a été en mesure d'identifier pratiquement les deux tiers des cas de paires de structures qui respectaient cette relation, et ce même si elles possèdent des structures initiales diverses. La moyenne de différence de conformation (RMSD) était de 1.32 Å alors que la différence d'identité de séquence est de 42%. Ces résultats suggèrent

qu'ENCoM est robuste à la conformation initiale et au nombre de mutations. Nous n'avons cependant pas de valeur de référence et il faudrait idéalement évaluer ces structures avec d'autres modèles afin de voir à quel point notre méthode est robuste sur ce type de test. Il serait par contre impossible d'utiliser les approches d'apprentissage automatisé étant donné qu'elles ne peuvent prédire que des mutations simples.

La valeur prédite par ENCoM ne représente en théorie que l'entropie vibrationnelle du squelette peptidique et alors ne représente qu'une partie des forces qui confèrent une stabilité aux protéines. En effet, malgré les bons pouvoirs prédictifs de l'AMN, cette performance provient seulement d'une comparaison avec les changements d'énergie libre de Gibbs qui est aussi influencée par l'énergie de la forme dénaturée, du solvant, de l'enthalpie de la forme native, l'entropie des chaînes latérales, l'entropie conformationnelle et la capacité calorifique. De façon contre-intuitive l'entropie prédite par ENCoM corrèle de façon négative avec l'énergie libre de Gibbs, c'est-à-dire qu'une rigidification de la protéine entraîne une stabilisation et *vice* et *versa*. Si l'effet de la mutation était purement entropique, on devrait observer l'effet inverse. Cependant, les différents paramètres thermodynamiques de la stabilité protéique sont inter-reliés et il n'est pas rare d'observer des phénomènes de compensation d'entropie, d'enthalpie et de capacité calorifique (Speedy, 2003). Ainsi, afin de valider les valeurs prédites par ENCoM, il faudrait avoir des données expérimentales qui représentent seulement l'entropie vibrationnelle ou bien des mutations qui affectent peu les autres termes. En approximant que l'entropie de la forme dénaturée est peu influencée lors d'une mutation, le paramètre thermodynamique le plus approprié pour valider ENCoM serait des variations d'entropie. Cependant, les valeurs retrouvées dans ProTherm et la littérature sont majoritairement des variations d'énergie libre de Gibbs et peu de mutations ont été caractérisées complètement afin de définir la courbe de Gibbs-Helmholtz (Pucci *et al.*, 2016). Cette équation permettrait d'obtenir la différence d'entropie entre les formes native et dénaturée à une température donnée. Récemment, un groupe a publié tous les paramètres thermodynamiques de 100 designs de l'enzyme Adénylate kinase (Howell *et al.*, 2014), permettant ainsi une éventuelle comparaison des valeurs d'ENCoM à ces nombreux paramètres thermodynamiques. En théorie, l'AMN devrait corréler avec les valeurs d'entropie et dans une certaine mesure

avec les capacités calorifiques. Dans le même ordre d'idées, il serait intéressant d'observer la description de ces termes par les autres méthodes décrites étant donné qu'elles sont supposées représenter principalement l'enthalpie des structures. Ainsi, une description totale des paramètres thermodynamiques permettrait de mieux comprendre et prédire comment des mutations modulent la courbe de Gibbs-Helmholtz et ainsi aider à générer des protéines thermorésistantes plus efficacement. En effet, une augmentation de l'énergie de Gibbs n'est pas nécessairement reflétée par une augmentation de la température de dénaturation.

Comme plusieurs autres fonctions de pointage (Sirin *et al.*, 2016), ENCoM prédit difficilement l'effet de mutation se produisant à la surface de la protéine. Cette performance n'est pas surprenante étant donné qu'une modulation de ces résidus affectera peu la matrice Hessienne et la plupart des mutations seront prédites comme étant neutres. Bien évidemment, une mutation de surface va moduler la stabilité protéique par une modification des interactions avec le solvant qui affecte les propriétés dynamiques. En effet, des approches de dynamiques moléculaires ont démontré que les protéines ne possèdent pas les mêmes propriétés dynamiques dans l'eau que dans d'autres solvants ou le vide (Levy *et al.*, 2001). Plusieurs stratégies pourraient être utilisées afin d'intégrer la contribution de l'eau dans le potentiel d'ENCoM. Une approche naïve et simple consisterait à construire une cage d'eau autour de la structure et considérer que chaque molécule est une masse qui interagit avec la protéine. Cependant, le solvant est très mobile et peut adopter différentes configurations aléatoires qui pourraient affecter la robustesse des prédictions. Afin de réduire cette incertitude, une approche plus constante consisterait à ajouter un nouveau terme à la matrice Hessienne basé sur les surfaces accessibles au solvant d'un résidu. Ce terme maintiendrait plus fortement les résidus en place étant donné qu'ils devront déployer de l'énergie afin de déplacer les molécules d'eau environnantes. Une approche similaire pourrait également être utilisée pour représenter les membranes des cellules qui sont reconnues pour affecter les propriétés des protéines (Chachisvilis *et al.*, 2006). Cet ajout serait cohérent avec le phénomène des facteurs d'amortissements qui modulent les fréquences de résonance d'un système en fonction de la viscosité de l'environnement.

Relation stabilité, dynamique et fonction

L'énergie d'un état macroscopique n'est pas une finalité à la fonction biologique. En effet, la forme de la surface énergétique est autant, sinon plus, importante que l'énergie des états, car elle confère aux protéines leurs propriétés dynamiques essentielles à leur mécanisme fonctionnel. Par exemple, pour certaine enzyme, l'efficacité enzymatique est influencée par la vitesse des mouvements de la protéine (Gagné *et al.*, 2015, Gagné et Doucet, 2013) et est alors maximale à proximité (-5 à 5 ℃) de la température de dénaturation étant donné que ces mouvements sont plus rapides (Howell *et al.*, 2014). Cette relation est également observée chez les protéines d'organismes thermophiles qui ont été démontrées par dynamique moléculaire et par RMN comme étant plus rigides que leurs homologues mésophiles et sont alors moins efficaces à température pièce. ENCoM capture en partie cette relation lors de la comparaison de la flexibilité de paires de structures de protéines homologues provenant d'organismes mésophiles et thermophiles en observant qu'environ les deux tiers sont plus rigides. De plus, les types de mutations prédites par ENCoM comme étant les plus susceptibles de rigidifier la protéine, sont les types de mutations les plus souvent observées entre les séquences de mésophiles et thermophiles. Ces résultats suggèrent qu'ENCoM est en mesure de capturer et de prédire cette pression évolutive de perte de flexibilité. En se référant à la théorie de l'AMN où une augmentation de la température augmente l'entropie du système, il serait alors facile de déterminer la différence de température nécessaire afin que les deux structures possèdent la même entropie vibrationnelle. Cette différence devrait en théorie correspondre à la différence de température de croissance des organismes des deux structures.

L'entropie vibrationnelle totale d'une protéine ne reflète pas nécessairement les propriétés dynamiques nécessaires à la fonction enzymatique. En effet, cette entropie peut provenir de régions non essentielles qui ne participent pas au mécanisme d'action alors que seulement un petit groupe de résidus peuvent être responsables de la vitesse de réaction (Gagné *et al.*, 2012). En observant la diagonale de la matrice de covariation, ENCoM peut prédire la flexibilité de chacun des résidus et a été en mesure de prédire l'effet allostérique de la mutation G121V de la DHFR. Cette mutation affecte peu la stabilité de l'enzyme, mais diminue drastiquement son efficacité enzymatique en modulant les propriétés dynamiques

du site de liaison malgré le fait qu'elle se trouve à plus de 15 Ångstroms du site enzymatique. Ce résultat n'est qu'une preuve de concept et une validation plus robuste doit être faite. La validation la plus directe consisterait à utiliser des perturbations de valeurs de RMN ou bien des facteurs B lors de mutations. En effet, il existe dans la banque de données PDB des paires de structures de conformations similaires différenciées par quelques mutations (Eyal *et al.*, 2001). En temps normal, ces mutations affectent les propriétés dynamiques et les facteurs B de ces structures. Tout en demeurant critique face aux biais concernant ces valeurs expérimentales, ENCoM pourrait être validé dans sa capacité à prédire ces variations. Également, différentes enzymes provenant de différentes espèces possèdent des efficacités enzymatiques différentes qui sont dictées par leur séquence. Ces valeurs sont regroupées dans la banque de données BRENDA (Chang *et al.*, 2009). En considérant que l'efficacité enzymatique est fortement liée à la dynamique de résidus à proximité du site catalytique, ENCoM pourrait être validé sur sa capacité à prédire ces différences, où une faible efficacité enzymatique pourrait correspondre à une faible flexibilité. En effet, une structure plus rigide, pourrait atteindre plus difficilement son état de transition essentielle à la vitesse de réaction. Cependant, les valeurs de cette banque sont difficilement généralisables étant donné qu'elles sont obtenues par plusieurs groupes différents. Alternativement, des valeurs provenant d'expérience de design protéique seraient probablement de meilleure qualité et représenteraient un contexte plus réaliste (Chen *et al.*, 2012).

Dans le même ordre d'idées, en considérant que les fonctions sont conservées dans différents gènes homologues et qu'elles sont en partie dépendantes des propriétés dynamiques de résidus, on s'attendrait à ce que les protéines possèdent des propriétés dynamiques conservées. ENCoM devrait être en mesure de capturer cette conservation et d'observer si elle correspond à différents déterminants structuraux. Des mutations pathogéniques pourraient également affecter ces propriétés dynamiques et perturber la fonction de la protéine (Callaway, 2010). Ainsi, ENCoM pourrait être validé dans sa capacité à prédire des mutations qui aboliraient la fonction protéique en se basant sur les perturbations des dynamiques conservées. Une méthodologie similaire pourrait être utilisée dans le design protéique d'enzymes, où la stabilité de la protéine veut être augmentée sans

nécessairement affecter l'efficacité enzymatique. Ainsi, la sélection de mutations à tester se ferait sur deux critères, soit la capacité à augmenter la stabilité protéique tout en perturbant le moins possible les propriétés dynamiques conservées. Cette recherche devra se faire en considérant une limite de Pareto (He *et al.*, 2012; Moal *et al.*, 2013), c'est-à-dire que ces propriétés sont inter-reliés et qu'à un certains point, il est impossible d'augmenter un des deux paramètres sans diminuer l'autre.

Finalement, une dernière validation pourrait être faite en évaluant la capacité d'ENCoM à prédire l'effet allostérique de liaison de ligands. En effet, cette régulation peut s'effectuer en modulant la conformation de la forme active ou bien en perturbant les propriétés dynamiques du site de liaison (Nussinov et Tsai, 2015, 2013; Nussinov *et al.*, 2014). Il existe déjà une banque de données appelée ASD qui a compilé toutes les structures subissant l'effet de liaison de ligands allostériques (Daily et Gray, 2009). Ainsi, les dynamiques de formes liée et non liée pourraient être comparées.

CONCLUSIONS

En conclusion, cette thèse présente le développement d'ENCoM, un nouveau modèle d'AMN de basse résolution qui, en contraste avec d'autres approches similaires, inclut la nature des résidus dans son potentiel harmonique. Ce nouvel ajout permet de mieux décrire le paysage énergétique d'une structure protéique et comparativement à ANM et STeM, de mieux prédire des changements conformationnels entre deux états macroscopiques. Également, l'inclusion de la séquence permet à ENCoM de détecter une différence d'entropie vibrationnelle lors de mutations qui corrèle avec des valeurs expérimentales de variation de l'énergie libre de Gibbs de repliement de protéines et des modulations de propriétés dynamiques allostériques obtenues par RMN. Comparativement à plusieurs approches populaires (FoldX, Rosetta, PoPMuSiC, etc.), cette nouvelle application d'AMN est moins biaisée et plus robuste. ENCoM prédit de façon plus efficace l'effet de mutations stabilisantes et complémente de façon synergique d'autres fonctions de score enthalpique. ENCoM a aussi été en mesure de capturer une pression évolutive qui rigidifie les protéines provenant d'organismes thermophiles et leur confère une résistance à la dénaturation par la chaleur. Ces principes pourraient éventuellement être utilisés dans un contexte biotechnologique lors du design protéique. Finalement, cette thèse présente le développement d'un serveur qui permet à des utilisateurs non technique d'utiliser ENCoM afin de prédire l'effet de mutations dans un contexte de haut débit et générer un ensemble de conformations réalistes.

REMERCIMENTS

J’aimerais premièrement remercier mes trois codirecteurs, Rafael, Pierre et Jean-Guy, de m’avoir soutenu et encouragé tout au long de ce projet de recherche. D’avoir été patient envers mon entêtement et mes nombreuses questions, où je pouvais débarquer dans leur bureau pour poser une question et finalement discuter pendant des heures. De m’avoir laissé une certaine liberté afin que je puisse explorer différents concepts et avenues de recherche, mais également d’avoir été critique envers mes travaux afin de me forcer à me poser les bonnes questions scientifiques et de revenir à des concepts fondamentaux afin que je puisse progresser en tant que chercheur.

J’aimerais aussi remercier les membres du laboratoire Najmanovich passé et présent avec qui j’ai eu beaucoup de plaisir à travailler. J’aimerais spécialement remercier Francis et Matthieu avec qui j’ai eu le plaisir de partager 6 belles années d’étude graduée. Francis pour avoir été un modèle de persévérance et d’éthique de travail et Matthieu pour les nombreux tours du CHUS afin de discuter de « science ».

J’aimerais particulièrement remercier ma conjointe Martine qui continue de m’encourager et surtout me supporter dans tous mes projets.

J’aimerais finalement remercier mes parents pour m’avoir poussé à faire ce que j’aime dans la vie.

LISTE DES RÉFÉRENCES

Ahmed,N.A. and Gokhale,D. V. (1989) Entropy Expressions and Their Estimators for Multivariate Distributions. *IEEE Trans. Inf. Theory*, **35**, 688–692.

Andrec,M. *et al.* (2007) A large data set comparison of protein structures determined by crystallography and NMR: statistical test for structural differences and the effect of crystal packing. *Proteins*, **69**, 449–465.

Arolas,J.L. *et al.* (2006) Folding of small disulfide-rich proteins: clarifying the puzzle. *Trends Biochem. Sci.*, **31**, 292–301.

Arrhenius,S. (1889) Über die Reaktiongeschwindigkeit bei der Inversion von Rohzucker durch Säuren. *Z. Phys. Chem.*, **4**, 226–248.

Arvelo,J.I. and Zabal,X.A. (1997) Reverberation rejection via modeforming with a vertical line array. *IEEE J. Ocean. Eng.*, **22**, 541–547.

Asakura,T. *et al.* (1978) Stabilizing effect of various organic solvents on protein. *J. Biol. Chem.*, **253**, 6423–6425.

De Baets,G. *et al.* (2015) Increased Aggregation Is More Frequently Associated to Human Disease-Associated Mutations Than to Neutral Polymorphisms. *PLoS Comput. Biol.*, **11**, e1004374.

Bahar,I. *et al.* (1998) Correlation between native-state hydrogen exchange and cooperative residue fluctuations from a simple model. *Biochemistry*, **37**, 1067–1075.

Bakan,A. and Bahar,I. (2009) The intrinsic dynamics of enzymes plays a dominant role in determining the structural changes induced upon inhibitor binding. *Proc. Natl. Acad. Sci. U. S. A.*, **106**, 14349–14354.

Baker,E.N. and Hubbard,R.E. (1984) Hydrogen bonding in globular proteins. *Prog. Biophys. Mol. Biol.*, **44**, 97–179.

Ball,P. and Hallsworth,J. (2015) Water structure and chaotropicity: their uses, abuses and implications for biology. *Phys. Chem. Chem. Phys.*, **17**, Ahead of Print.

Barak,L.S. *et al.* (2001) Constitutive arrestin-mediated desensitization of a human vasopressin receptor mutant associated with nephrogenic diabetes insipidus. *Proc.*

Natl. Acad. Sci. U. S. A., **98**, 93–98.

Baxa,M.C. *et al.* (2014) Loss of conformational entropy in protein folding calculated using realistic ensembles and its implications for NMR-based calculations. *Proc. Natl. Acad. Sci. U. S. A.*, **111**, 15396–401.

Beadle,B.M. and Shoichet,B.K. (2002) Structural bases of stability-function tradeoffs in enzymes. *J. Mol. Biol.*, **321**, 285–296.

Bée,M. (1988) Quasielastic Neutron Scattering. 437.

Beeby,M. *et al.* (2005) The genomics of disulfide bonding and protein stabilization in thermophiles. *PLoS Biol.*, **3**, 1549–1558.

Bendl,J. *et al.* (2014) PredictSNP: Robust and Accurate Consensus Classifier for Prediction of Disease-Related Mutations. *PLoS Comput. Biol.*, **10**, e1003440.

Bennion,B.J. and Daggett,V. (2003) The molecular basis for the chemical denaturation of proteins by urea. *Proc. Natl. Acad. Sci. U. S. A.*, **100**, 5142–7.

Berman,H. *et al.* (2000) The Protein Data Bank. *Nucleic Acids Res.*, **28**, 235–242.

Best,R.B. *et al.* (2015) Quantitative Interpretation of FRET Experiments via Molecular Simulation: Force Field and Validation. *Biophys. J.*, **108**, 2721–2731.

Bloom,J.D. *et al.* (2006) Protein stability promotes evolvability. *Proc. Natl. Acad. Sci. U. S. A.*, **103**, 5869–5874.

Bornscheuer,U.T. *et al.* (2012) Engineering the third wave of biocatalysis. *Nature*, **485**, 185–194.

Bouchard,H. *et al.* (2014) Antibody-drug conjugates—a new wave of cancer drugs. *Bioorg. Med. Chem. Lett.*, **24**, 5357–63.

Brooks,B.R. *et al.* (1983) CHARMM: A program for macromolecular energy, minimization, and dynamics calculations. *J. Comput. Chem.*, **4**, 187–217.

Buck,M. *et al.* (1995) Structural determinants of protein dynamics: analysis of 15N NMR relaxation measurements for main-chain and side-chain nuclei of hen egg white lysozyme. *Biochemistry*, **34**, 4041–55.

Callaway,E. (2010) Mutation-prediction software rewarded. *Nat. News*.

Cambillau,C. and Claverie,J.M. (2000) Structural and genomic correlates of hyperthermostability. *J. Biol. Chem.*, **275**, 32383–32386.

Chachisvilis,M. *et al.* (2006) G protein-coupled receptors sense fluid shear stress in

endothelial cells. *Proc. Natl. Acad. Sci. U. S. A.*, **103**, 15463–15468.

Chang,A. *et al.* (2009) BRENDA, AMENDA and FRENDA the enzyme information system: New content and tools in 2009. *Nucleic Acids Res.*, **37**.

Chen,M.M.Y. *et al.* (2012) Comparison of random mutagenesis and semi-rational designed libraries for improved cytochrome P450 BM3-catalyzed hydroxylation of small alkanes. *Protein Eng. Des. Sel.*, **25**, 171–178.

Clore,G.M. and Schwieters,C.D. (2006) Concordance of residual dipolar couplings, backbone order parameters and crystallographic B-factors for a small ??/?? protein: A unified picture of high probability, fast atomic motions in proteins. *J. Mol. Biol.*, **355**, 879–886.

Cooper,A. (2005) Heat capacity effects in protein folding and ligand binding: A re-evaluation of the role of water in biomolecular thermodynamics. In, *Biophysical Chemistry*., pp. 89–97.

Cooper,A. (2000) Heat capacity of hydrogen-bonded networks: An alternative view of protein folding thermodynamics. *Biophys. Chem.*, **85**, 25–39.

Cooper,A. *et al.* (2001) Heat does not come in different colours: Entropy-enthalpy compensation, free energy windows, quantum confinement, pressure perturbation calorimetry, solvation and the multiple causes of heat capacity effects in biomolecular interactions. *Biophys. Chem.*, **93**, 215–230.

Cooper,G.M. and Shendure,J. (2011) Needles in stacks of needles: finding disease-causal variants in a wealth of genomic data. *Nat. Rev. Genet.*, **12**, 628–40.

Creamer,T.P. and Rose,G.D. (1992) Side-chain entropy opposes a-helix formation but rationalizes experimentally determined helix-forming propensities (a-helix/protein folding/protein engineering). *Nat. Sci.*, **89**, 5937–5941.

Daily,M.D. and Gray,J.J. (2009) Allosteric communication occurs via networks of tertiary and quaternary motions in proteins. *PLoS Comput. Biol.*, **5**, e1000293.

Daniel,R.M. *et al.* (2002) The dynamic transition in proteins may have a simple explanation. *Faraday Disc.*, **122**, 163–169.

Davey,J.A. *et al.* (2015) Prediction of Stable Globular Proteins Using Negative Design with Non-native Backbone Ensembles. *Structure*, **23**, 2011–2021.

Davey,J.A. and Chica,R.A. (2014) Improving the accuracy of protein stability predictions

with multistate design using a variety of backbone ensembles. *Proteins Struct. Funct. Bioinforma.*, **82**, 771–784.

Deechongkit,S., Nguyen,H., *et al.* (2004) Context-dependent contributions of backbone hydrogen bonding to beta-sheet folding energetics. *Nature*, **430**, 101–105.

Deechongkit,S., Dawson,P.E., *et al.* (2004) Toward assessing the position-dependent contributions of backbone hydrogen bonding to ??-sheet folding thermodynamics employing amide-to-ester perturbations. *J. Am. Chem. Soc.*, **126**, 16762–16771.

Dehouck,Y. *et al.* (2013) BeAtMuSiC: Prediction of changes in protein-protein binding affinity on mutations. *Nucleic Acids Res.*, **41**, W333–W339.

Dehouck,Y. *et al.* (2009) Fast and accurate predictions of protein stability changes upon mutations using statistical potentials and neural networks: PoPMuSiC-2.0. *Bioinformatics*, **25**, 2537–2543.

Demmel,F. *et al.* (1997) Vibrational frequency shifts as a probe of hydrogen bonds: Thermal expansion and glass transition of myoglobin in mixed solvents. *Eur. Biophys. J.*, **26**, 327–335.

Dias,C.L. *et al.* (2010) The hydrophobic effect and its role in cold denaturation. *Cryobiology*, **60**, 91–99.

Dias,C.L. (2012) Unifying microscopic mechanism for pressure and cold denaturations of proteins. *Phys. Rev. Lett.*, **109**, 048104.

Dietzen,M. *et al.* (2012) On the applicability of elastic network normal modes in small-molecule docking. *J. Chem. Inf. Model.*, **52**, 844–856.

Dill,K. a and Chan,H.S. (1997) From Levinthal to pathways to funnels. *Nat. Struct. Biol.*, **4**, 10–19.

Doig, a J. and Sternberg,M.J. (1995) Side-chain conformational entropy in protein folding. *Protein Sci.*, **4**, 2247–2251.

Donald,J.E. *et al.* (2011) Salt bridges: Geometrically specific, designable interactions. *Proteins Struct. Funct. Bioinforma.*, **79**, 898–915.

Doster,W. (2010) The protein-solvent glass transition. *Biochim. Biophys. Acta - Proteins Proteomics*, **1804**, 3–14.

Dunitz,J.D. (1995) Win some, lose some: enthalpy-entropy compensation in weak intermolecular interactions. *Chem. Biol.*, **2**, 709–712.

Dykeman,E.C. and Twarock,R. (2010) All-atom normal-mode analysis reveals an RNA-induced allostery in a bacteriophage coat protein. *Phys. Rev. E - Stat. Nonlinear, Soft Matter Phys.*, **81**, 1–10.

Ebbinghaus,S. *et al.* (2007) An extended dynamical hydration shell around proteins. *Proc. Natl. Acad. Sci. U. S. A.*, **104**, 20749–20752.

Echave,J. (2008) Evolutionary divergence of protein structure: The linearly forced elastic network model. *Chem. Phys. Lett.*, **457**, 413–416.

Elias,M. *et al.* (2014) The universality of enzymatic rate-temperature dependency. *Trends Biochem. Sci.*, **39**, 1–7.

Ermer,O. and Lifson,S. (1974) Normal mode analysis of alkenes by the consistent force field. *J. Mol. Spectrosc.*, **51**, 261–272.

Eswar,N. *et al.* (2008) Protein structure modeling with MODELLER. *Struct. Proteomics High-Throughput Methods*, **426**, 145–159.

Eyal,E. *et al.* (2006) Anisotropic network model: Systematic evaluation and a new web interface. *Bioinformatics*, **22**, 2619–2627.

Eyal,E. *et al.* (2001) MutaProt: a web interface for structural analysis of point mutations. *Bioinformatics*, **17**, 381–382.

Fang,J. (2015) Drug Designing : Open Access Reliability of Machine Learning Based Algorithms for Designing Protein Drugs with Enhanced Stability. *Drug Des. Open Access*, **4**, 4–5.

Feng,Y.H. *et al.* (1998) Mechanism of constitutive activation of the AT1 receptor: Influence of the size of the agonist switch binding residue Asn111. *Biochemistry*, **37**, 15791–15798.

Ferdjani,S. *et al.* (2011) Correlation between thermostability and stability of glycosidases in ionic liquid. *Biotechnol. Lett.*, **33**, 1215–1219.

Figliuzzi,M. *et al.* (2015) Coevolutionary landscape inference and the context-dependence of mutations in beta-lactamase TEM-1. *Mol. Biol. Evol.*, **33**, msv211.

Firnberg,E. *et al.* (2014) A comprehensive, high-resolution map of a Gene's fitness landscape. *Mol. Biol. Evol.*, **31**, 1581–1592.

Fischer,M. *et al.* (2014) Incorporation of protein flexibility and conformational energy penalties in docking screens to improve ligand discovery. *Nat. Chem.*, **6**, 575–583.

Fleishman,S.J. *et al.* (2011) Computational design of proteins targeting the conserved stem region of influenza hemagglutinin. *Science*, **332**, 816–21.

Foight,G.W. *et al.* (2014) Designed BH3 peptides with high affinity and specificity for targeting Mcl-1 in cells. *ACS Chem. Biol.*, **9**, 1962–1968.

Foley,B.M. *et al.* (2014) Protein thermal conductivity measured in the solid state reveals anharmonic interactions of vibrations in a fractal structure. *J. Phys. Chem. Lett.*, **5**, 1077–1082.

Fontana,A. *et al.* (1997) Probing the partly folded states of proteins by limited proteolysis. *Fold. Des.*, **2**, R17–26.

Forbes,S.A. *et al.* (2015) COSMIC: Exploring the world's knowledge of somatic mutations in human cancer. *Nucleic Acids Res.*, **43**, D805–D811.

Fowler,D.M. *et al.* (2010) High-resolution mapping of protein sequence-function relationships. *Nat. Methods*, **7**, 741–6.

Fowler,D.M. and Fields,S. (2014) Deep mutational scanning: a new style of protein science. *Nat. Methods*, **11**, 801–7.

Fowler,S.B. *et al.* (2005) Rational design of aggregation-resistant bioactive peptides: reengineering human calcitonin. *Proc. Natl. Acad. Sci. U. S. A.*, **102**, 10105–10.

Frappier,V. and Najmanovich,R.J. (2014) A Coarse-Grained Elastic Network Atom Contact Model and Its Use in the Simulation of Protein Dynamics and the Prediction of the Effect of Mutations. *PLoS Comput. Biol.*, **10**, e1003569.

Fraser,J. *et al.* (2011) Accessing protein conformational ensembles using room-temperature X-ray crystallography. *Proc. Natl. Acad. Sci. U. S. A.*, **108**, 16247–16252.

Freddolino,P.L. *et al.* (2009) Force field bias in protein folding simulations. *Biophys. J.*, **96**, 3772–3780.

Frezza,E. and Lavery,R. (2015) Internal normal mode analysis (iNMA) applied to protein conformational flexibility. *J. Chem. Theory Comput.*, **11**, 5503–5512.

Friedland,G.D. *et al.* (2009) A correspondence between solution-state dynamics of an individual protein and the sequence and conformational diversity of its family. *PLoS Comput. Biol.*, **5**, e1000393.

Fultz,B. (2010) Vibrational thermodynamics of materials. *Prog. Mater. Sci.*, **55**, 247–352.

Furnham,N. *et al.* (2015) Large-Scale Analysis Exploring Evolution of Catalytic

Machineries and Mechanisms in Enzyme Superfamilies. *J. Mol. Biol.*, **428**, 253–267.

Gagné,D. *et al.* (2012) Conservation of flexible residue clusters among structural and functional enzyme homologues. *J. Biol. Chem.*, **287**, 44289–44300.

Gagné,D. *et al.* (2015) Perturbation of the Conformational Dynamics of an Active-Site Loop Alters Enzyme Activity. *Structure*, **23**, 2256–2266.

Gagné,D. and Doucet,N. (2013) Structural and functional importance of local and global conformational fluctuations in the RNase A superfamily. *FEBS J.*, **280**, 5596–5607.

Gao,D. *et al.* (2009) Thermostable variants of cocaine esterase for long-time protection against cocaine toxicity. *Mol. Pharmacol.*, **75**, 318–23.

Gao,J. *et al.* (2009) Localized thermodynamic coupling between hydrogen bonding and microenvironment polarity substantially stabilizes proteins. *Nat. Struct. Mol. Biol.*, **16**, 684–90.

Gao,X. *et al.* (2005) High-throughput limited proteolysis/mass spectrometry for protein domain elucidation. *J. Struct. Funct. Genomics*, **6**, 129–134.

Garman,E. (2003) 'Cool' crystals: Macromolecular cryocrystallography and radiation damage. *Curr. Opin. Struct. Biol.*, **13**, 545–551.

Gaudreault,F. and Najmanovich,R.J. (2015) FlexAID: Revisiting Docking on Non-Native-Complex Structures. *J. Chem. Inf. Model.*, **55**, 1323–1336.

Gether,U. *et al.* (1997) Constitutively Active G Protein-coupled Receptor. 2587–2591.

Gilbert,F. and Dziewonski, a. M. (1975) An Application of Normal Mode Theory to the Retrieval of Structural Parameters and Source Mechanisms from Seismic Spectra. *Philos. Trans. R. Soc. A Math. Phys. Eng. Sci.*, **278**, 187–269.

Glyakina,A. V *et al.* (2007) Different packing of external residues can explain differences in the thermostability of proteins from thermophilic and mesophilic organisms. *Bioinformatics*, **23**, 2231–8.

Gong,L.I. *et al.* (2013) Stability-mediated epistasis constrains the evolution of an influenza protein. *Elife*, **2013**, 1–19.

Gopalakrishna,R. and Anderson,W.B. (1982) Ca2+-induced hydrophobic site on calmodulin: Application for purification of calmodulin by phenyl-Sepharose affinity chromatography. *Biochem. Biophys. Res. Commun.*, **104**, 830–836.

Green,D. and Unruh,W.G. (2006) The failure of the Tacoma Bridge: A physical model.

Am. J. Phys., **74**, 706–716.

van Gunsteren,W.F. and Mark,A.E. (1992) Prediction of the activity and stability effects of site-directed mutagenesis on a protein core. *J. Mol. Biol.*, **227**, 389–395.

Guo,J. *et al.* (2014) Thermal adaptation of dihydrofolate reductase from the moderate thermophile geobacillus stearothermophilus. *Biochemistry*, **53**, 2855–2863.

Harms,M.J. and Thornton,J.W. (2013) Evolutionary biochemistry: revealing the historical and physical causes of protein properties. *Nat. Rev. Genet.*, **14**, 559–71.

Hayward,R.J. and Henry,B.R. (1974) Anharmonicity in polyatomic molecules: A local-mode analysis of the XH-stretching overtone spectra of ammonia and methane. *J. Mol. Spectrosc.*, **50**, 58–67.

He,H.P. *et al.* (2003) ^{29}Si and ^{27}Al MAS NMR study of the thermal transformations of kaolinite from North China. *Clay Miner.*, **38**, 551–559.

He,L. *et al.* (2012) A divide-and-conquer approach to determine the Pareto frontier for optimization of protein engineering experiments. *Proteins Struct. Funct. Bioinforma.*, **80**, 790–806.

Hendsch,Z.S. and Tidor,B. (1994) Do salt bridges stabilize proteins? A continuum electrostatic analysis. *Protein Sci.*, **3**, 211–226.

Henikoff,S. and Henikoff,J.G. (1992) Amino acid substitution matrices from protein blocks. *Proc. Natl. Acad. Sci. U. S. A.*, **89**, 10915–10919.

Henzler-Wildman,K. and Kern,D. (2007) Dynamic personalities of proteins. *Nature*, **450**, 964–972.

Hopf,T.A. *et al.* (2014) Sequence co-evolution gives 3D contacts and structures of protein complexes. *Elife*, **3**, e03430.

Howell,S.C. *et al.* (2014) Understanding thermal adaptation of enzymes through the multistate rational design and stability prediction of 100 adenylate kinases. *Structure*, **22**, 218–229.

Hummer,G. *et al.* (1998) The pressure dependence of hydrophobic interactions is consistent with the observed pressure denaturation of proteins. *Proc. Natl. Acad. Sci.*, **95**, 1552–1555.

Jiménez-Osés,G. *et al.* (2014) The role of distant mutations and allosteric regulation on LovD active site dynamics. *Nat. Chem. Biol.*, **10**, 431–6.

Jollife,I. (1986) Principal Component Analysis Springer-Verlag, New York.

Joti,Y. *et al.* (2005) Protein boson peak originated from hydration-related multiple minima energy landscape. *J. Am. Chem. Soc.*, **127**, 8705–8709.

Kamiya,K. *et al.* (2003) Algorithm for normal mode analysis with general internal coordinates. *J. Comput. Chem.*, **24**, 826–841.

van der Kamp,M.W. *et al.* (2010) Dynameomics: A Comprehensive Database of Protein Dynamics. *Structure*, **18**, 423–435.

Karplus,M. and Kushick,J.N. (1981) Method for estimating the configurational entropy of macromolecules. *Macromolecules*, **14**, 325–332.

Kawashima,S. *et al.* (1999) AAindex: Amino acid index database. *Nucleic Acids Res.*, **27**, 368–369.

Khechinashvili,N.N. *et al.* (2014) The entropic nature of protein thermal stabilization. *J. Biomol. Struct. Dyn.*, **32**, 1396–1405.

Khersonsky,O. *et al.* (2011) Optimization of the in-silico-designed Kemp eliminase KE70 by computational design and directed evolution. *J. Mol. Biol.*, **407**, 391–412.

Kim,C.U. *et al.* (2011) Protein dynamical transition at 110 K. *Proc. Natl. Acad. Sci.*, **108**, 20897–20901.

King,M.C. *et al.* (2003) Breast and ovarian cancer risks due to inherited mutations in BRCA1 and BRCA2. *Science (80-.).*, **302**, 643–646.

Kircher,M. (2014) A general framework for estimating the relative pathogenicity of human genetic variants. *Nat. g*, **46**, 310–315.

Kiss,G. *et al.* (2010) Evaluation and ranking of enzyme designs. *Protein Sci.*, **19**, 1760–1773.

Klotz,I.M. and Klotz,T. a (1955) Oxygen-carrying proteins: a comparison of the oxygenation reaction in hemocyanin and hemerythrin with that in hemoglobin. *Science*, **121**, 477–480.

Kumar,M.D.S. (2006) ProTherm and ProNIT: thermodynamic databases for proteins and protein-nucleic acid interactions. *Nucleic Acids Res.*, **34**, D204–D206.

Kumar,S. *et al.* (2002) Maximal stabilities of reversible two-state proteins. *Biochemistry*, **41**, 5359–5374.

Laidler,K.J. (1984) The development of the Arrhenius equation. *J. Chem. Educ.*, **61**, 494–

498.

Landrum,M.J. *et al.* (2014) ClinVar: Public archive of relationships among sequence variation and human phenotype. *Nucleic Acids Res.*, **42**, 980–985.

Lee, a L. and Wand, a J. (2001) Microscopic origins of entropy, heat capacity and the glass transition in proteins. *Nature*, **411**, 501–504.

Leitner,D.M. (2008) Energy flow in proteins. *Annu. Rev. Phys. Chem.*, **59**, 233–259.

LeVine,M. V. and Weinstein,H. (2015) AIM for allostery: Using the ising model to understand information processing and transmission in allosteric biomolecular systems. *Entropy*, **17**, 2895–2918.

LeVine,M. V. and Weinstein,H. (2014) NbIT - A New Information Theory-Based Analysis of Allosteric Mechanisms Reveals Residues that Underlie Function in the Leucine Transporter LeuT. *PLoS Comput. Biol.*, **10**, e1003603.

Levinthal,C. (1969) How to fold graciously. *Mössbauer Spectrosc. Biol. Syst. Proc.*, **24**, 22–24.

Levitt,M. *et al.* (1985) Protein normal-mode dynamics: Trypsin inhibitor, crambin, ribonuclease and lysozyme. *J. Mol. Biol.*, **181**, 423–447.

Levy,Y. *et al.* (2001) Solvent effects on the energy landscapes and folding kinetics of polyalanine. *Proc. Natl. Acad. Sci. U. S. A.*, **98**, 2188–2193.

Lezon,T.R. and Bahar,I. (2010) Using entropy maximization to understand the determinants of structural dynamics beyond native contact topology. *PLoS Comput. Biol.*, **6**, 1–12.

Liang,Z.X. *et al.* (2004) Evidence for increased local flexibility in psychrophilic alcohol dehydrogenase relative to its thermophilic homologue. *Biochemistry*, **43**, 14676–14683.

Lifson,S. (1968) Consistent Force Field for Calculations of Conformations, Vibrational Spectra, and Enthalpies of Cycloalkane and n-Alkane Molecules. *J. Chem. Phys.*, **49**, 5116.

Liu,L. *et al.* (2000) A study on the enthalpy-entropy compensation in protein unfolding. *Biophys. Chem.*, **84**, 239–251.

Loladze,V. V *et al.* (2001) Heat capacity changes upon burial of polar and nonpolar groups in proteins. *Protein Sci.*, **10**, 1343–52.

Lu,Y. and Freeland,S. (2006) On the evolution of the standard amino-acid alphabet. *Genome Biol*, **7**, 102.

Mamonova,T.B. *et al.* (2010) Flexibility and Mobility in Mesophilic and Thermophilic Homologous Proteins From Molecular Dynamics and Foldunfold Method. *J. Bioinform. Comput. Biol.*, **08**, 377–394.

Marsh,J.A. and Teichmann,S.A. (2014) Parallel dynamics and evolution: Protein conformational fluctuations and assembly reflect evolutionary changes in sequence and structure. *BioEssays*, **36**, 209–218.

Matthews,B.W. (1993) Structural and genetic analysis of protein stability. *Annu. Rev. Biochem.*, **62**, 139–160.

Matysiak,S. *et al.* (2012) Role of hydrophobic hydration in protein stability: A 3D water-explicit protein model exhibiting cold and heat denaturation. *J. Phys. Chem. B*, **116**, 8095–8104.

May,A. and Zacharias,M. (2008) Protein-ligand docking accounting for receptor side chain and global flexibility in normal modes: Evaluation on kinase inhibitor cross docking. *J. Med. Chem.*, **51**, 3499–3506.

McCammon,J.A. *et al.* (1977) Dynamics of folded proteins. *Nature*, **267**, 585–90.

McLaughlin,R.N. *et al.* (2012) The spatial architecture of protein function and adaptation. *Nature*, **491**, 138–42.

McVean,G.A. *et al.* (2012) An integrated map of genetic variation from 1,092 human genomes. *Nature*, **491**, 56–65.

Mitternacht,S. and Berezovsky,I.N. (2011) Binding leverage as a molecular basis for allosteric regulation. *PLoS Comput. Biol.*, **7**, e1002148.

Miyazaki,Y. *et al.* (2000) Low-temperature heat capacity and glassy behavior of lysozyme crystal. *J. Phys. Chem. B*, **104**, 8044–8052.

Moal,I.H. *et al.* (2013) Scoring functions for protein-protein interactions. *Curr. Opin. Struct. Biol.*, **23**, 862–867.

Moal,I.H. and Bates,P.A. (2010) SwarmDock and the use of normal modes in protein-protein Docking. *Int. J. Mol. Sci.*, **11**, 3623–3648.

Moal,I.H. and Fernandez-Recio,J. (2012) SKEMPI: a Structural Kinetic and Energetic database of Mutant Protein Interactions and its use in empirical models.

Bioinformatics, **28**, 2600–2607.

Morcos,F. *et al.* (2011) Direct-coupling analysis of residue coevolution captures native contacts across many protein families. *Proc. Natl. Acad. Sci. U. S. A.*, **108**, E1293–301.

Moretti,R. *et al.* (2013) Community-wide evaluation of methods for predicting the effect of mutations on protein-protein interactions. *Proteins Struct. Funct. Bioinforma.*, **81**, 1980–1987.

Muller,P. a J. and Vousden,K.H. (2013) P53 Mutations in Cancer. *Nat. Cell Biol.*, **15**, 2–8.

Münch,C. *et al.* (2011) Prion-like propagation of mutant superoxide dismutase-1 misfolding in neuronal cells. *Proc. Natl. Acad. Sci. U. S. A.*, **108**, 3548–53.

Myers,J.K.K. *et al.* (1995) Denaturant m values and heat capacity changes: relation to changes in accessible surface areas of protein unfolding. *Protein Sci.*, **4**, 2138–48.

Na,H. *et al.* (2014) Generalized spring tensor models for protein fluctuation dynamics and conformation changes. *Adv. Exp. Med. Biol.*, **805**, 107–135.

Némethy,G. and Scheraga,H.A. (1962) Structure of Water and Hydrophobic Bonding in Proteins. I. A Model for the Thermodynamic Properties of Liquid Water. *J. Chem. Phys.*, **36**, 3382–3400.

Ng,P.C. and Henikoff,S. (2006) Predicting the effects of amino acid substitutions on protein function. *Annu Rev Genomics Hum Genet*, **7**, 61–80.

Nussinov,R. *et al.* (2014) Principles of allosteric interactions in cell signaling. *J. Am. Chem. Soc.*, **136**, 17692–17701.

Nussinov,R. and Tsai,C.J. (2015) Allostery without a conformational change? Revisiting the paradigm. *Curr. Opin. Struct. Biol.*, **30**, 17–24.

Nussinov,R. and Tsai,C.-J. (2013) Allostery in disease and in drug discovery. *Cell*, **153**, 293–305.

O'Hayre,M. *et al.* (2013) The emerging mutational landscape of G proteins and G-protein-coupled receptors in cancer. *Nat Rev Cancer*, **13**, 412–424.

Otten,L.G. *et al.* (2010) Enzyme engineering for enantioselectivity: from trial-and-error to rational design? *Trends Biotechnol.*, **28**, 46–54.

Paciaroni,A. *et al.* (2002) Effect of the environment on the protein dynamical transition: a neutron scattering study. *Biophys. J.*, **83**, 1157–1164.

Parnot,C. *et al.* (2000) Systematic identification of mutations that constitutively activate the angiotensin II type 1A receptor by screening a randomly mutated cDNA library with an original pharmacological bioassay. *Proc. Natl. Acad. Sci. U. S. A.*, **97**, 7615–7620.

Parthasarathy,S. and Murthy,M.R. (2000) Protein thermal stability: insights from atomic displacement parameters (B values). *Protein Eng.*, **13**, 9–13.

Pearlman,D.A. *et al.* (1995) AMBER, a package of computer programs for applying molecular mechanics, normal mode analysis, molecular dynamics and free energy calculations to simulate the structural and energetic properties of molecules. *Comput. Phys. Commun.*, **91**, 1–41.

Perry,L.J. and Wetzel,R. (1984) Disulfide bond engineered into T4 lysozyme: stabilization of the protein toward thermal inactivation. *Science*, **226**, 555–7.

Peters,G. and Wilkinson,J.H. (1975) On the stability of Gauss-Jordan elimination with pivoting. *Commun. ACM*, **18**, 20–24.

Piana,S. *et al.* (2012) Protein folding kinetics and thermodynamics from atomistic simulation. *Proc. Natl. Acad. Sci. U. S. A.*, **109**, 17845–17850.

Porter,L.L. and Rose,G.D. (2011) Redrawing the Ramachandran plot after inclusion of hydrogen-bonding constraints. *Proc. Natl. Acad. Sci. U. S. A.*, **108**, 109–113.

Powell,D.E. (2011) Introduction to the special education issue. *Hum. Pathol.*, **42**, 761–762.

Preiswerk,N. *et al.* (2014) Impact of scaffold rigidity on the design and evolution of an artificial Diels-Alderase. *Proc. Natl. Acad. Sci. U. S. A.*, **111**, 8013–8.

Privalov,P. (1989) Thermodynamic Problems Of Protein Structure. *Annu. Rev. Biophys. Biomol. Struct.*, **18**, 47–69.

Privett,H.K. *et al.* (2012) Iterative approach to computational enzyme design. *Proc. Natl. Acad. Sci.*, **109**, 3790–3795.

Pucci,F. *et al.* (2016) High-quality thermodynamic data on the stability changes of proteins upon single-site mutations. 1–25.

Pucci,F. and Rooman,M. (2015) Towards an accurate prediction of the thermal stability of homologous proteins. *J. Biomol. Struct. Dyn.*, **1102**, 1–23.

Ramachandran,G.N. *et al.* (1963) Stereochemistry of polypeptide chain configurations. *J. Mol. Biol.*, **7**, 95–99.

Ramanathan,A. and Agarwal,P.K. (2011) Evolutionarily conserved linkage between

enzyme fold, flexibility, and catalysis. *PLoS Biol.*, **9**, e1001193.

Razvi,A. and Scholtz,J.M. (2006) Lessons in stability from thermophilic proteins. *Protein Sci.*, **15**, 1569–1578.

Reetz,M.T. *et al.* (2006) Iterative saturation mutagenesis on the basis of b factors as a strategy for increasing protein thermostability. *Angew. Chemie - Int. Ed.*, **45**, 7745–7751.

Reich,L. *et al.* (2015) SORTCERY - A High-Throughput Method to Affinity Rank Peptide Ligands. *J. Mol. Biol.*, **427**, 2135–2150.

Ro,D. *et al.* (2006) Production of the antimalarial drug precursor artemisinic acid in engineered yeast. *Nature*, **440**, 3–6.

Robertson,A.D. and Murphy,K.P. (1997) Protein Structure and the Energetics of Protein Stability. *Chem. Rev.*, **97**, 1251–1268.

Robic,S. *et al.* (2003) Role of residual structure in the unfolded state of a thermophilic protein. *Proc. Natl. Acad. Sci. U. S. A.*, **100**, 11345–11349.

Röthlisberger,D. *et al.* (2008) Kemp elimination catalysts by computational enzyme design. *Nature*, **453**, 190–195.

Rueda,M. *et al.* (2007) Thorough Validation of Protein Normal Mode Analysis: A Comparative Study with Essential Dynamics. *Structure*, **15**, 565–575.

Ruwende,C. *et al.* (1995) Natural selection of hemi- and heterozygotes for G6PD deficiency in Africa by resistance to severe malaria. *Nature*, **376**, 246–9.

Sankar,K. *et al.* (2015) Distributions of experimental protein structures on coarse-grained free energy landscapes. *J. Chem. Phys.*, **143**, 243153.

Schuyler,A.D. *et al.* (2009) Iterative cluster-NMA: A tool for generating conformational transitions in proteins. *Proteins Struct. Funct. Bioinforma.*, **74**, 760–776.

Schuyler,A.D. and Chirikjian,G.S. (2005) Efficient determination of low-frequency normal modes of large protein structures by cluster-NMA. *J. Mol. Graph. Model.*, **24**, 46–58.

Schymkowitz,J. *et al.* (2005) The FoldX web server: An online force field. *Nucleic Acids Res.*, **33**, W382–W388.

Selvaratnam,R. *et al.* (2011) Mapping allostery through the covariance analysis of NMR chemical shifts. *Proc. Natl. Acad. Sci. U. S. A.*, **108**, 6133–6138.

Sheldon,R. a. (2012) Fundamentals of green chemistry: efficiency in reaction design.

Chem. Soc. Rev., **41**, 1437.

Sherry,S.T. *et al.* (2001) dbSNP: the NCBI database of genetic variation. *Nucleic Acids Res.*, **29**, 308–11.

Sikosek,T. and Chan,H.S. (2014) Biophysics of protein evolution and evolutionary protein biophysics. *J. R. Soc. Interface*, **11**, 20140419.

Sirin,S. *et al.* (2016) AB-Bind: Antibody binding mutational database for computational affinity predictions. *Protein Sci.*, **25**, 393–409.

Skinner,J.J. *et al.* (2014) Benchmarking all-atom simulations using hydrogen exchange. *Proc. Natl. Acad. Sci. U. S. A.*, **111**, 15975–80.

Slater,N.B. (1948) Aspects of a Theory of Unimolecular Reaction Rates. *Proc. R. Soc. A Math. Phys. Eng. Sci.*, **194**, 112–131.

Slater,N.B. (1953) The Theoretical Rate of Isomerization of Cyclopropane. *Proc. R. Soc. A Math. Phys. Eng. Sci.*, **218**, 224–244.

Sobolev,V. *et al.* (1996) Molecular docking using surface complementarity. *Proteins Struct. Funct. Genet.*, **25**, 120–129.

Soheilifard,R. *et al.* (2008) Critical evaluation of simple network models of protein dynamics and their comparison with crystallographic B-factors. *Phys. Biol.*, **5**, 026008.

Speedy,R.J. (2003) Kauzmann ' s paradox and the glass transition. *Biophys. Chem.*, **105**, 411–420.

van der Spoel,E. *et al.* (2015) Association analysis of insulin-like growth factor-1 axis parameters with survival and functional status in nonagenarians of the Leiden Longevity Study. *Aging (Albany. NY).*, **7**, 956–963.

Spolar,R.S. *et al.* (1989) Hydrophobic effect in protein folding and other noncovalent processes involving proteins. *Proc Natl Acad Sci U S A*, **86**, 8382–8385.

Stenson,P.D. *et al.* (2014) The Human Gene Mutation Database: Building a comprehensive mutation repository for clinical and molecular genetics, diagnostic testing and personalized genomic medicine. *Hum. Genet.*, **133**, 1–9.

Studer,R. a *et al.* (2014) Stability-activity tradeoffs constrain the adaptive evolution of RubisCO. *Proc. Natl. Acad. Sci. U. S. A.*, **111**, 2223–8.

Sturtevant,J.M. (1977) Heat capacity and entropy changes in processes involving proteins.

Proc. Natl. Acad. Sci. U. S. A., **74**, 2236–40.

Syme,N.R. *et al.* (2010) Comparison of entropic contributions to binding in a ‘ hydrophilic’ versus ‘hydrophobic’ ligand-protein interaction. *J. Am. Chem. Soc.*, **132**, 8682–8689.

Szilagyi,A. and Zavodszky,P. (2000) Structural differences between mesophilic, moderately thermophilic and extremely thermophilic protein subunits: Results of a comprehensive survey. *Structure*, **8**, 493–504.

Tama,F. *et al.* (2000) Building-block approach for determining low-frequency normal modes of macromolecules. *Proteins Struct. Funct. Genet.*, **41**, 1–7.

Tama,F. *et al.* (2004) Normal mode based flexible fitting of high-resolution structure into low-resolution experimental data from cryo-EM. *J. Struct. Biol.*, **147**, 315–326.

Tama,F. and Brooks,C.L. (2005) Diversity and identity of mechanical properties of icosahedral viral capsids studied with elastic network normal mode analysis. *J. Mol. Biol.*, **345**, 299–314.

Tan,A.R. and Swain,S.M. (2003) Ongoing adjuvant trials with trastuzumab in breast cancer. *Semin. Oncol.*, **30**, 54–64.

Tanford,C. (1997) How protein chemists learned about the hydrophobic factor. *Protein Sci.*, **6**, 1358–1366.

Tasumi,M. *et al.* (1982) Normal vibrations of proteins: Glucagon. *Biopolymers*, **21**, 711–714.

Tatsumi,R. *et al.* (2004) A hybrid method of molecular dynamics and harmonic dynamics for docking of flexible ligand to flexible receptor. *J. Comput. Chem.*, **25**, 1995–2005.

Teeter,M.M. *et al.* (2001) On the nature of a glassy state of matter in a hydrated protein: Relation to protein function. *Proc. Natl. Acad. Sci. U. S. A.*, **98**, 11242–11247.

The Gene Ontology Consortium (2000) Gene Ontology: tool for the unification of biology. *Nat. Genet.*, **25**, 25–29.

Thielges,M.C. *et al.* (2008) Exploring the energy landscape of antibody-antigen complexes: Protein dynamics, flexibility, and molecular recognition. *Biochemistry*, **47**, 7237–7247.

Thiltgen,G. and Goldstein,R.A. (2012) Assessing Predictors of Changes in Protein Stability upon Mutation Using Self-Consistency. *PLoS One*, **7**, e46084.

Thompson,J.J. *et al.* (2014) Application of information theory to a three-body coarse-

grained representation of proteins in the PDB: Insights into the structural and evolutionary roles of residues in protein structure. *Proteins*, **82**, 3450–65.

Thusberg,J. *et al.* (2011) Performance of mutation pathogenicity prediction methods on missense variants. *Hum. Mutat.*, **32**, 358–368.

Thyagarajan,B. and Bloom,J.D. (2014) The inherent mutational tolerance and antigenic evolvability of influenza hemagglutinin. *Elife*, **2014**, 1–26.

Tian,W. and Skolnick,J. (2003) How well is enzyme function conserved as a function of pairwise sequence identity? *J. Mol. Biol.*, **333**, 863–882.

Tidor,B. and Karplus,M. (1994) The contribution of vibrational entropy to molecular association. The dimerization of insulin. *J. Mol. Biol.*, **238**, 405–414.

Tilton,R.F. *et al.* (1992) Effects of temperature on protein structure and dynamics: X-ray crystallographic studies of the protein ribonuclease-A at nine different temperatures from 98 to 320 K. *Biochemistry*, **31**, 2469–2481.

Tirion,M.M. (1996) Large Amplitude Elastic Motions in Proteins from a Single-Parameter, Atomic Analysis. *Phys. Rev. Lett.*, **77**, 1905–1908.

Tokuriki,N. *et al.* (2008) How protein stability and new functions trade off. *PLoS Comput. Biol.*, **4**, e1000002.

Tzeng,S.-R. and Kalodimos,C.G. (2012) Protein activity regulation by conformational entropy. *Nature*, **488**, 236–240.

Uchiyama,S. *et al.* (2002) Thermodynamic characterization of variants of mesophilic cytochrome c and its thermophilic counterpart. *Protein Eng.*, **15**, 455–462.

Vajpai,N. *et al.* (2013) High-pressure NMR reveals close similarity between cold and alcohol protein denaturation in ubiquitin. *Proc. Natl. Acad. Sci. U. S. A.*, **110**, E368–76.

Vasser,M. *et al.* (2004) Letters To Nature. *Nature*, **429**, 2–6.

Vaz,D.C. *et al.* (2006) Enthalpic and entropic contributions mediate the role of disulfide bonds on the conformational stability of interleukin-4. *Protein Sci.*, **15**, 33–44.

Verschueren,E. *et al.* (2013) Protein-peptide complex prediction through fragment interaction patterns. *Structure*, **21**, 789–797.

Vitkup,D. *et al.* (2000) Solvent mobility and the protein ‘glass’ transition. *Nat. Struct. Biol.*, **7**, 34–38.

Wang,X. *et al.* (2002) Evolution of an antibiotic resistance enzyme constrained by stability and activity trade-offs. *J. Mol. Biol.*, **320**, 85–95.

Warshel,A. (1970) Consistent Force Field Calculations. II. Crystal Structures, Sublimation Energies, Molecular and Lattice Vibrations, Molecular Conformations, and Enthalpies of Alkanes. *J. Chem. Phys.*, **53**, 582.

Warshel,A. *et al.* (1970) Consistent force field for calculation of vibrational spectra and conformations of some amides and lactam rings. *J. Mol. Spectrosc.*, **33**, 84–99.

Warshel,A. and Karplus,M. (1972) Calculation of ground and excited state potential surfaces of conjugated molecules. I. Formulation and parametrization. *J. Am. Chem. Soc.*, **94**, 5612–5625.

Waterson,J.J. (1850) On the Physics of Media That Are Composed of Free and Perfectly Elastic Molecules in a State of Motion. *Philos. Trans. R. Soc. London.*, **5**, 604.

Wettstein,S. *et al.* (2014) Linking genotypes database with locus-specific database and genotype-phenotype correlation in phenylketonuria. *Eur. J. Hum. Genet.*, **23**, 302–309.

Wickstrom,L. *et al.* (2012) The linear interaction energy method for the prediction of protein stability changes upon mutation. *Proteins*, **80**, 111–25.

Williams,G. and Toon,A.J. (2010) Protein folding pathways and state transitions described by classical equations of motion of an elastic network model. *Protein Sci.*, **19**, 2451–2461.

Wolf-Watz,M. *et al.* (2004) Linkage between dynamics and catalysis in a thermophilic-mesophilic enzyme pair. *Nat. Struct. Mol. Biol.*, **11**, 945–9.

Xiao,S. *et al.* (2013) Rational modification of protein stability by targeting surface sites leads to complicated results. *Proc. Natl. Acad. Sci.*, **110**, 11337–11342.

Yang,C. *et al.* (2014) A fully atomistic computer simulation study of cold denaturation of a β-hairpin. *Nat. Commun.*, **5**, 5773.

Yang,L. *et al.* (2009) Comparisons of experimental and computed protein anisotropic temperature factors. *Proteins Struct. Funct. Bioinforma.*, **76**, 164–175.

Yang,L.W. *et al.* (2007) Insights into Equilibrium Dynamics of Proteins from Comparison of NMR and X-Ray Data with Computational Predictions. *Structure*, **15**, 741–749.

Yang,L.W. *et al.* (2006) oGNM: Online computation of structural dynamics using the Gaussian Network Model. *Nucleic Acids Res.*, **34**, W24–W31.

Ying,S. *et al.* (2015) Point Cluster Analysis Using a 3D Voronoi Diagram with Applications in Point Cloud Segmentation. *ISPRS Int. J. Geo-Information*, **4**, 1480–1499.

Yu,Y.B. *et al.* (1999) The measure of interior disorder in a folded protein and its contribution to stability. *J. Am. Chem. Soc.*, **121**, 8443–8449.

Yuen,K. (2010) A Relationship between the Hessian and Covariance Matrix for Gaussian Random Variables. *Bayesian Methods Struct. Dyn. Civ. Eng.*, 0–5.

Zavodszky,M. *et al.* (2001) Disulfide bond effects on protein stability: designed variants of Cucurbita maxima trypsin inhibitor-V. *Protein Sci.*, **10**, 149–60.

Zhou,H.-X. (2002) Toward the physical basis of thermophilic proteins: linking of enriched polar interactions and reduced heat capacity of unfolding. *Biophys. J.*, **83**, 3126–33.

Zuber,J. *et al.* (2015) Identification of destabilizing and stabilizing mutations of Ste2p, a G protein-coupled receptor in saccharomyces cerevisiae. *Biochemistry*, **54**, 1787–1806.

ANNEXES

Annexe A – Données Supplémentaires de l'article 1

Tableaux supplémentaires

Tableau A.1- Raw data for the calculation b-factor correlations

Afin d'alléger ce document, le tableau A.1 est retrouvé à l'adresse suivante : http://journals.plos.org/ploscompbiol/article?id=10.1371/journal.pcbi.1003569 ou au DOI: http://dx.doi.org/10.1371/journal.pcbi.1003569

Tableau A.2 - Raw overlap calculations for the different methods.

Afin d'alléger ce document, le tableau A.2 est retrouvé à l'adresse suivante : http://journals.plos.org/ploscompbiol/article?id=10.1371/journal.pcbi.1003569 ou au DOI: http://dx.doi.org/10.1371/journal.pcbi.1003569

Tableau A.3 - Experimental and predicted ΔΔG values on the effect of mutations.

Afin d'alléger ce document, le tableau A.3 est retrouvé à l'adresse suivante : http://journals.plos.org/ploscompbiol/article?id=10.1371/journal.pcbi.1003569 ou au DOI: http://dx.doi.org/10.1371/journal.pcbi.1003569

Tableau A.4 - Raw data for the calculation of self-consistency bias and error on the prediction of forward and back mutations.

Afin d'alléger ce document, le tableau A.4 est retrouvé à l'adresse suivante : http://journals.plos.org/ploscompbiol/article?id=10.1371/journal.pcbi.1003569 ou au DOI: http://dx.doi.org/10.1371/journal.pcbi.1003569

Tableau A.5 - Raw DB data and experimental S2 NMR order parameter for the G121V DHFR mutant.

Afin d'alléger ce document, le tableau A.5 est retrouvé à l'adresse suivante : http://journals.plos.org/ploscompbiol/article?id=10.1371/journal.pcbi.1003569 ou au DOI: http://dx.doi.org/10.1371/journal.pcbi.1003569

Annexe B – Données Supplémentaires de l'article 2

Tableaux supplémentaires

Tableau B.1- Raw data for entropy calculation

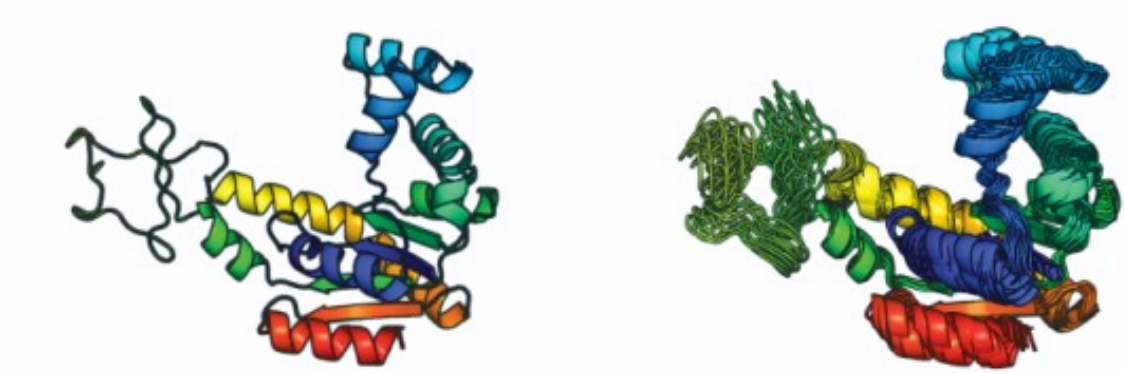
Afin d'alléger ce document, le tableau B.1 est retrouvé à l'adresse suivante : http://onlinelibrary.wiley.com/doi/10.1002/pro.2592/abstract ou au DOI: http://dx.doi.org/10.1002/pro.2592

Annexe C – Données Supplémentaires de l'article 3

Figures supplémentaires

#	WT Residue	Res. #	Chain	Mutation	ENCoM ΔkCal/mol desc asc	FoldX3.0 ΔkCal/mol desc asc	Combined ΔkCal/mol desc asc	GIF	PNG	Pymol Session	PDB
1	ALA	11	A	GLN	-0.93	1.51	0.58	GIF	PNG	PML	PDB
2	ALA	11	A	LEU	-0.51	0.89	0.38	GIF	PNG	PML	PDB
3	GLY	10	A	LEU	-0.39	1.29	0.90	GIF	PNG	PML	PDB
4	LYS	13	A	LEU	-0.29	-0.17	-0.45	GIF	PNG	PML	PDB
5	GLY	10	A	GLN	-0.24	1.79	1.54	GIF	PNG	PML	PDB
6	GLY	14	A	LEU	-0.12	-0.07	-0.19	GIF	PNG	PML	PDB
7	GLY	10	A	ALA	-0.10	1.86	1.76	GIF	PNG	PML	PDB
8	GLY	12	A	GLN	-0.06	0.97	0.91	GIF	PNG	PML	PDB
9	GLY	14	A	GLN	-0.05						
10	GLY	12	A	LEU	-0.05						
11	THR	15	A	LEU	-0.02						
12	GLY	14	A	ALA	0.00						
13	THR	15	A	GLN	0.01						
14	GLY	12	A	ALA	0.02						
15	LYS	13	A	GLN	0.11						
16	ALA	11	A	GLY	0.27						
17	THR	15	A	ALA	0.29						
18	LYS	13	A	ALA	0.63	-0.01	0.62	GIF	PNG	PML	PDB
19	THR	15	A	GLY	0.72	0.25	0.97	GIF	PNG	PML	PDB
20	LYS	13	A	GLY	1.03	0.47	1.49	GIF	PNG	PML	PDB

Figure S1. Results for the effect of mutations on Adenylate kinase (PDB ID 4AKE). Each amino acid from positions 10 to 15 in chain A was mutated to Q, A, L or G. The weighted predictions of ENCoM and FoldX individually or combined are shown in the respective columns are shown for every mutant. The effect of each mutation on protein flexibility can be quickly assessed graphically by hovering over the PNG or GIF links as shown. Users can download each image individually as well a PyMOL script used to generate the image and the modeled mutated structures in PDB format.



The models in PDB format are available in the models folder of Results.zip.

Amplitude of each mode in each model.

Model	Mode_7	Mode_8
model_1.pdb	0.000	0.000
model_2.pdb	14.629	0.000
model_3.pdb	29.257	0.000
model_4.pdb	0.000	14.629
model_5.pdb	14.629	14.629

Figure S2. Screen shot of the results displayed for conformational sampling in the web interface. Adenylate kinase (4AKE) conformations were generated using the first 2 slowest non-trivial modes with a maximum RMSD distortion of 2.0 Å and a step of 1.0 Å RMSD distortion per conformation per mode. An image showing all the generated conformations and a GIF showing the reordered model trajectory is displayed. A table of the amplitudes applied to each mode on each model is shown.

Annexe D – Protocoles utilisés avec Rosetta dans le premier article (chapitre 2)

Nous avons utilisé la version 3.4 de Rosetta et utilisé la même méthodologie retrouvée dans l'exemple de l'Eglin C présent dans les dossiers du code source. La structure cristalline sauvage a premièrement été relaxée selon :

« minimize_with_cst.default.linuxgccrelease -in:file:l lst -database ~/minirosetta_database/ -in:file:fullatom -ddg::out_pdb_prefix minimize_with_cs »

Ensuite, l'effet de mutations a été étudié avec la fonction suivante :

«ddg_monomer.default.macosgccrelease -in:file:l lst -ddg::weight_file soft_rep -ddg::iterations 5 -ddg::dump_pdbs true -ddg::mut_file mutations.multiples.txt -database ~/minirosetta_database/ -ddg::local_opt_only false -ddg::min_cst false -ddg::mean true -ddg::min -ignore_unrecognized_res »