Biologie Aujourd'hui **211 (3)**, 239-244 (2017)

https://doi.org/10.1051/jbio/2017030

Disponible en ligne :
www.biologie-journal.org

ARTICLE

De la variabilité des séquences à la prédiction structurale et fonctionnelle : modélisation de familles de protéines homologues

Pierre Barrat-Charlaix et Martin Weigt*

Sorbonne Universités, UPMC Université Paris 06, CNRS, Biologie Computationnelle et Quantitative, Institut de Biologie Paris Seine, 75005 Paris, France

Reçu le 20 novembre 2017

Résumé– Grâce au séquençage de nouvelle génération, le nombre de génomes séquencés augmente rapidement, fournissant notamment de nombreux exemples de la variabilité des séquences de protéines homologues. Cet article traite de modèles probabilistes fondés sur les données pour les séquences protéiques, qui sont capables d'extraire une multitude d'informations à partir de données séquentielles, dont (i) des caractéristiques structurelles telles que les contacts inter-résidus formés dans la protéine repliée, (ii) les interfaces d'interaction et (iii) les effets phénotypiques des substitutions d'acides aminés dans les protéines.

Mots clés : analyse de séquences de protéines, co-évolution, familles de protéines homologues, modélisation statistique de séquences protéiques

Abstract – From sequence variability to structural and functional prediction: modeling of homologous protein families. Thanks to next-generation sequencing, the number of sequenced genomes grows rapidly, providing in particular ample examples for the sequence variability between homologous proteins. This article discusses data-driven probabilistic sequence models, which are able to extract a multitude of information from sequence data alone, including (i) structural features like residue-residue contacts, which are formed in the folded protein, (ii) protein-protein interaction interfaces and (iii) phenotypic effects of amino-acid substitutions in proteins.

Keywords: protein sequence analysis, coevolution, homologous protein families, data-driven probabilistic protein sequences

Introduction

Les protéines sont essentielles dans presque tous les processus cellulaires. Au cours de l'évolution, les substitutions d'acides aminés provoquent des changements dans leurs séquences primaires. Cependant, ces changements ne peuvent pas être complètement aléatoires : seules les substitutions conservant la fonction biologique d'une protéine sont acceptées par la sélection naturelle. Puisque la fonction d'une protéine repose fortement sur sa structure tridimensionnelle (3D), cette dernière doit également être conservée pendant l'évolution. Nous nous trouvons dans une situation apparemment paradoxale : d'une part, deux homologues, c'est-à-dire deux protéines d'ascendance évolutive commune, peuvent différer sur plus de 70 ou 80 % de leurs acides aminés, mais gardent des structures tridimensionnelles et des fonctions biologiques très similaires. D'autre part, seulement quelques mutations aléatoires peuvent déstabiliser le repliement d'une protéine ou perturber sa fonction biologique.

Les expériences de caractérisation structurelle et fonctionnelle détaillées des protéines sont chronophages, et leurs résultats souvent incertains. Parmi les plus de 90 millions de séquences protéiques distinctes rassemblées dans la base de données Uniprot (http://www.uniprot.org/), seules un peu plus de 550 000 (0,6 %) ont des annotations manuelles vérifiées (UniProt Consortium, 2015). Tandis que le nombre total d'entrées d'Uniprot augmente de façon exponentielle grâce au séquençage de nouvelle génération, la base de données Swiss-Prot, c'est-à-dire la partie d'Uniprot collectant des protéines annotées manuellement, reste de taille presque constante (les annotations existantes sont régulièrement revues et mises à jour, mais très peu de protéines sont nouvellement annotées manuellement). La situation est encore plus dramatique lorsqu'on parle de structures protéiques

*Auteur correspondant : martin.weigt@upmc.fr

déterminées expérimentalement. La banque de données sur les protéines (PDB – http://www.rcsb.org/) contient des structures pour environ 42 000 (0,05 %) séquences protéiques distinctes (Berman *et al.*, 2000).

Il ne faut toutefois pas désespérer : la richesse des données sur les séquences offre en fait une opportunité sans précédent pour le développement et l'application d'approches de modélisation statistique guidées par les données. En fait, les 90 millions d'entrées Uniprot sont, dans une large mesure, classées automatiquement en familles de séquences homologues. La très importante base de données Pfam (http://pfam.xfam.org/) répertorie près de 17 000 familles de domaines de protéines de haute qualité ; une grande partie d'entre elles rassemblent entre 10^3 et 10^6 séquences distinctes (Finn *et al.*, 2014). Nous pouvons considérer chacune de ces familles comme un échantillon de séquences ayant une structure et une fonction partagées, selon les arguments présentés dans le premier paragraphe. Il est donc très important de développer des approches informatiques pour caractériser la variabilité des séquences à l'intérieur de chaque famille et de la mettre en relation avec les propriétés biologiques communes des protéines membres de la famille.

Dans cette revue, nous montrerons quelques développements récents dans la modélisation coévolutionnaire des séquences protéiques. Ces dernières années, cette modélisation a permis de prédire des structures protéiques inconnues pour des centaines de familles de protéines, chacune contenant des milliers de séquences distinctes, et de concevoir des outils prometteurs pour prédire les interactions protéine-protéine et les effets de mutations dans la séquence des protéines.

Modélisation de la conservation des résidus et de la coévolution des paires de résidus

Il est bien connu et admis que la similarité et la variabilité des séquences de protéines homologues contiennent des informations précieuses. Le point de départ de toutes les méthodes statistiques examinées dans cet article réside donc dans les alignements multiples de séquences (MSA) pour les familles de protéines homologues, chaque ligne d'un alignement contenant une séquence appartenant à la famille et chaque colonne assignant un résidu à une position (Durbin *et al.*, 1998).

La plus importante propriété des alignements dans ce contexte est la conservation des résidus fonctionnellement et structurellement importants entre des séquences homologues. Pour donner des exemples, les sites actifs ou les résidus enfouis profondément dans le noyau de la protéine sont généralement sujets à moins de variation que la plupart des résidus en surface de la protéine. Statistiquement, la conservation des résidus est prise en compte par ce que l'on appelle des *modèles de profil*, qui analysent la composition en acides aminés de chaque colonne individuellement (Durbin *et al.*, 1998). Ces modèles – prenant en partie la forme de Chaînes de Markov Cachées (HMM) permettant de traiter les insertions et les délétions d'acides aminés (Eddy, 1998) – sont à la base des méthodes numériques les plus performantes pour l'analyse de séquences : ils permettent de créer de multiples alignements de séquences (Edgar & Batzoglou, 2006), de détecter l'homologie entre les protéines (Söding, 2004), et de détecter les résidus fonctionnellement importants. La conservation structurelle et fonctionnelle entre les protéines homologues est à la base de nombreux outils bioinformatiques pour l'annotation automatique de séquences et la prédiction de structure 3D (Biasini *et al.*, 2014 ; Webb & Sali, 2014).

Mathématiquement, la variabilité de la colonne i d'un alignement est prise en compte par les quantités $f_i(a)$, donnant la fraction des protéines alignées portant l'acide aminé a à la position i, pour chacun des 20 acides aminés naturels {A, C, D, ..., W, Y}. En général, une séquence alignée $(a_1,\ldots,a_L)$ peut être caractérisée par la probabilité $P(a_1,\ldots,a_L) = f_1(a_1) \times f_2(a_2) \times \ldots \times f_L(a_L)$, qui multiplie les fractions spécifiques à chaque site sur toutes les positions de la séquence.

Bien qu'appartenant aux approches les plus fructueuses en bio-informatique, les modèles de profil montrent des limites importantes. En regardant chaque résidu individuellement, ils ne peuvent pas saisir les relations entre les résidus – comme des contacts inter-résidus dans des protéines individuelles ou entre des protéines en interaction, ou les effets épistatiques des mutations. Celles-ci sont liées à la coévolution des résidus, visible par les corrélations des acides aminés présents à différentes positions d'une protéine. Une raison intuitive peut être énoncée facilement : la plupart des substitutions d'acides aminés à une seule position sont délétères, par exemple *via* une réduction de la stabilité thermodynamique ou de l'activité biochimique de la protéine. Dans certains cas, ces effets délétères peuvent être compensés par des substitutions appropriées d'autres résidus en contact dans la structure – on dit que les résidus co-évoluent (De Juan *et al.*, 2013).

La coévolution des résidus est reflétée par l'occupation conjointe de paires de colonnes de l'alignement par des acides aminés, c'est-à-dire par la fraction $f_{ij}(a,b)$ de séquences alignées ayant l'acide aminé a en position i et l'acide aminé b en position j, pour toutes les 20×20 combinaisons possibles. Bien que de tels changements corrélés d'acides aminés aient déjà été étudiés il y a environ 25 ans (Göbel *et al.*, 1994 ; Neher, 1994), leur utilisation pour la prédiction de la structure de la protéine est toujours restée limitée. Cela a changé au cours de la dernière décennie, en raison de l'abondance récente des données séquentielles et, de manière tout aussi importante, du développement de nouvelles approches numériques fondées sur l'apprentissage de modèles statistiques globaux et inspirées par la physique statistique des systèmes complexes désordonnés (Weigt *et al.*, 2009 ; De Juan *et al.*, 2013).

Un $f_{ij}(a,b)$, qui est significativement différent de $f_i(a) \times f_j(b)$, indique une présence corrélée des acides aminés aux deux positions correspondantes, et pourrait être un bon prédicteur des contacts dans la structure

tertiaire. Malheureusement, ce n'est pas le cas, en raison des effets de transitivité de la corrélation : si une position i est corrélée à une position j, et j à k, on s'attend à des corrélations entre i et k même sans relation directe entre les deux. L'idée de base de l'*analyse par couplage direct (Direct Coupling Analysis – DCA)* (Weigt *et al.*, 2009 ; Morcos *et al.*, 2011 ; Ekeberg *et al.*, 2013) et d'autres méthodes récentes apparentées (Jones *et al.*, 2012 ; Kamisetty *et al.*, 2013) est donc de démêler les corrélations directes et indirectes et d'en déduire les couplages directs dans le cadre d'une approche de modélisation statistique globale. À chaque séquence est donnée la probabilité

$$P(a_1,\ldots,a_L)=\frac{1}{Z}\exp\left\{\sum_{i<j}^{L}J_{ij}(a_i,a_j)+\sum_{i}^{L}h_i(a_i)\right\},$$

où les $J_{ij}(a,b)$ sont les couplages dits directs mesurant la quantité de coévolution entre les sites i et j, et les $h_i(a)$ sont des champs locaux ou biais mesurant la conservation du site de la position des résidus i. Les valeurs numériques de ces paramètres doivent être déterminées de telle sorte que le modèle soit cohérent avec les statistiques empiriques $f_i(a)$ et $f_{ij}(a,b)$ extraites des données originales, c'est-à-dire l'alignement des séquences d'une famille de protéines homologues.

L'inférence exacte de ces paramètres est difficile à calculer. Plusieurs approximations ont donc été développées, y compris des approches en champ moyen (Weigt *et al.*, 2009 ; Morcos *et al.*, 2011) ou gaussiennes (Jones *et al.*, 2012), ou de maximisation de la pseudo-vraisemblance (Ekeberg *et al.*, 2013 ; Kamisetty *et al.*, 2013). Ces approximations sont suffisamment efficaces pour analyser des centaines de familles de protéines. Très récemment, des méthodes plus précises ont été proposées. Leur coût de calcul plus élevé les rend seulement aptes à la modélisation très précise de quelques familles de protéines (Sutto *et al.*, 2015 ; Haldane *et al.*, 2016).

Quelle est la signification de ces modèles ? Que pouvons-nous apprendre des paramètres ? Les trois sections suivantes traitent des principaux domaines d'application des méthodes coévolutionnaires globales.

Prédiction de la structure tridimensionnelle des protéines

Au cours du processus de repliement, la protéine prend une forme compacte et les acides aminés distants le long de la chaîne polypeptidique unidimensionnelle entrent en contact spatial direct. La prédiction de ces contacts à partir de la séquence seule est l'une des tâches les plus importantes dans la prédiction de la structure des protéines. Les méthodes coévolutionnaires sont la partie centrale de certains des prédicteurs de contact les plus fructueux (Jones *et al.*, 2015 ; Wang *et al.*, 2017).

Une première étape est l'observation que deux résidus i et j présentant de forts couplages coévolutifs – représentés par de grands paramètres $J_{ij}(a,b)$ – ont une forte probabilité d'être en contact (Weigt *et al.*, 2009). Ceci est représenté sur la figure 1 (en bas à gauche), où sont montrés les trente premiers contacts prédits dans l'inhibiteur de trypsine (5pti). Cependant, DCA est une méthode non supervisée, et sa précision reste souvent limitée. En particulier, pour de petits alignements contenant moins d'environ 1000 séquences, l'important bruit statistique rend les prédictions imprécises.

Les meilleurs prédicteurs de contact ajoutent donc une deuxième étape d'apprentissage supervisé, combinant des scores coévolutionnaires avec d'autres caractéristiques comme des prédictions de la structure secondaire et de l'accessibilité au solvant, ou les potentiels d'interaction statistique des acides aminés. Selon l'évaluation des deux dernières compétitions CASP (*Critical Assessment of protein Structure Prediction* – http://predictioncenter.org/), ces méthodes sont actuellement les meilleurs prédicteurs de contact entre résidus (Jones *et al.*, 2015 ; Wang *et al.*, 2017).

Les contacts prédits peuvent être intégrés dans une simulation moléculaire de la protéine afin de prédire la structure tertiaire (Marks *et al.*, 2011 ; Sulkowska *et al.*, 2012). Parmi les réussites les plus remarquables dans ce contexte, on trouve la prédiction de la structure active d'une kinase sensorielle bactérienne (la première structure prévue par DCA confirmée expérimentalement) (Dago *et al.*, 2012), la prédiction de grandes protéines membranaires (Hopf *et al.*, 2012 ; Nugent & Jones, 2012) et, très récemment, la prédiction de plus de 500 modèles structuraux représentatifs pour les familles de protéines Pfam sans structures 3D connues (Ovchinnikov *et al.*, 2017).

Détection des interactions protéine-protéine et assemblage de complexes protéiques

L'analyse en couplages directs a été initialement développée dans le contexte des interactions protéine-protéine (Weigt *et al.*, 2009), où des questions allant bien au-delà de la prédiction de contact et de la biologie structurale peuvent être posées. Étant donné deux alignements correspondant à deux familles de protéines homologues, nous pouvons poser des questions telles que :

Les deux familles interagissent-elles, c'est-à-dire contiennent-elles un nombre important de paires de protéines qui interagissent ?

Si oui, quelles protéines spécifiques interagissent ? Cette question est particulièrement intéressante dans le contexte des familles avec un grand nombre de paralogues, qui ont des fonctions biologiques similaires mais sont liés à des processus biologiques distincts.

Si l'on sait que deux protéines interagissent, peut-on dire comment ? Peut-on prédire les interfaces, et quelles paires de résidus seront en contact entre les deux protéines ?

Toutes ces questions ont, au moins dans des cas anecdotiques, été discutées en utilisant des modèles coévolutionnaires (Weigt *et al.*, 2009 ; Schug *et al.*, 2009 ; Hopf *et al.*, 2014 ; Ovchinnikov & Kamisetty, 2014 ; Bitbol *et al.*, 2016 ; Feinauer *et al.*, 2016 ; Gueudré

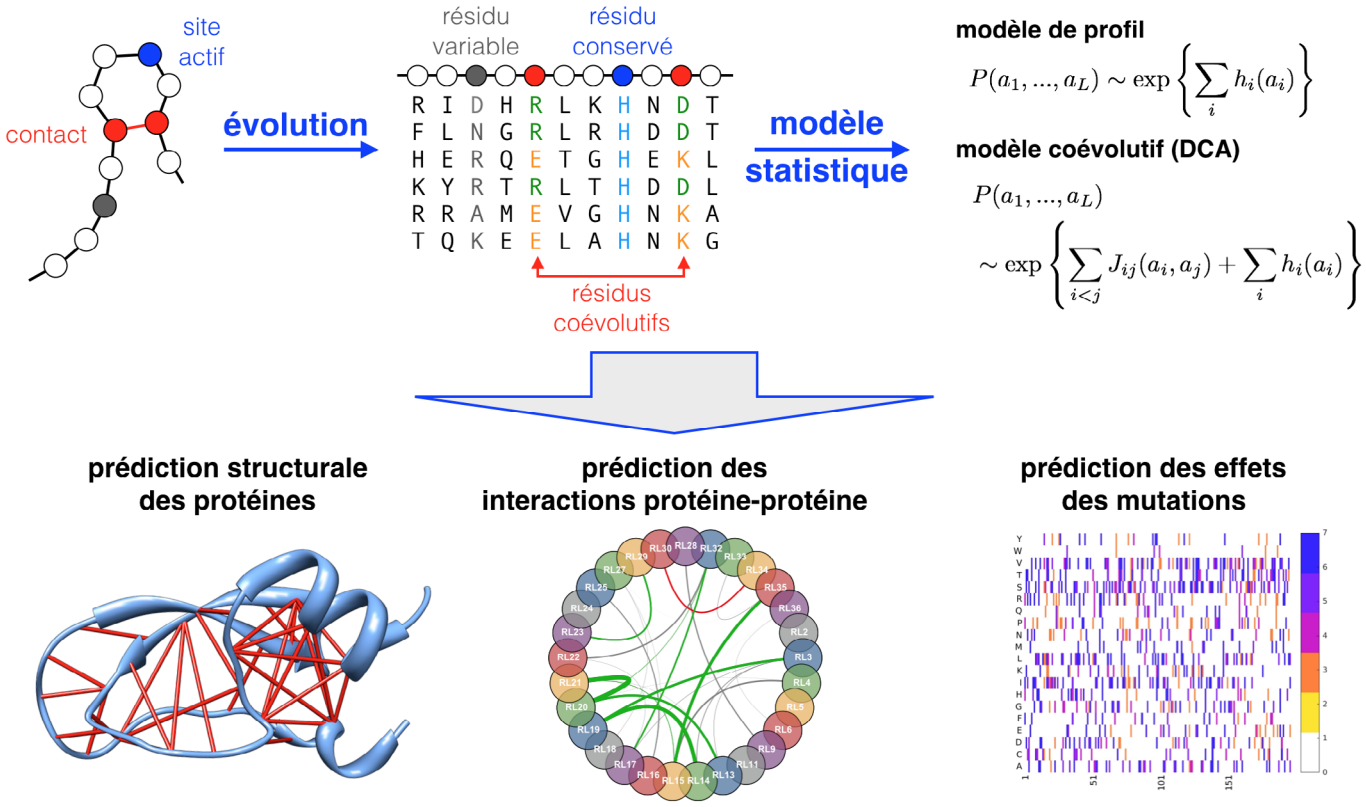

Figure 1. La boîte supérieure montre le contexte général et la stratégie de l'approche de modélisation – Les contraintes évolutives structurelles et fonctionnelles sont reflétées par des conservations et corrélations caractéristiques dans des alignements multiples de séquences de protéines homologues. Ceux-ci sont capturés par des modèles de séquences statistiques, comme les modèles de profils (conservation seulement) et les modèles coévolutionnaires (conservation et coévolution). Ces modèles sont utilisés pour extraire une multitude d'informations à partir de données de séquence, comprenant des contacts résiduels dans les protéines (lignes rouges en bas à gauche), des interactions protéine-protéine (lignes vertes en bas au centre, la ligne rouge est une fausse prédiction, les lignes grises sont des interactions non détectées) et des effets mutationnels dans les protéines.

et al., 2016 ; Uguzzoni *et al.*, 2017). Cependant, de nombreuses interrogations persistent, concernant notamment les deux premières questions, et d'importants progrès sont à prévoir au cours des prochaines années.

Prédiction des effets mutationnels dans les protéines

La compréhension des effets des changements d'acides aminés dans une protéine présente un intérêt énorme : elle peut concerner la distinction entre des variantes pathogènes et des polymorphismes neutres et est centrale pour comprendre l'évolution de la pharmacorésistance ou de la virulence. De nombreux outils ont été proposés pour évaluer ces effets, basés sur la stabilité structurale des protéines, sur l'information évolutive, ou sur des combinaisons des deux. Notre modèle coévolutionnaire probabiliste permet le calcul d'un score statistique, qui est donné par le logarithme du rapport des probabilités de la séquence mutante et de la séquence naturelle :

$$\Delta L = log\left(\frac{P(séquence\ mutante)}{P(séquence\ naturelle)}\right).$$

On s'attendrait alors à ce que des mutations bénéfiques conduisent à des probabilités plus élevées, et donc des scores positifs, et des mutations délétères à des scores négatifs. Selon une analyse intuitive, les acides aminés rarement observés à certaines positions de la protéine au cours de l'évolution sont susceptibles d'être délétères. Au-delà de cette intuition simple, déjà utilisée dans les prédicteurs d'effets mutationnels basés sur le profil de séquence, le modèle DCA tente de capturer la dépendance contextuelle des mutations grâce aux couplages coévolutifs. Des études sur des protéines uniques (Mann *et al.*, 2014 ; Morcos *et al.*, 2014 ; Figliuzzi *et al.*, 2015 ; Hopf *et al.*, 2017) et des mutations pathogènes humaines (Feinauer & Weigt, 2017 ; Hopf *et al.*, 2017) et une analyse à l'échelle du génome de la bactérie modèle *Escherichia coli* (Couce *et al.*, 2017) ont permis d'établir que les méthodes coévolutionnaires améliorent la précision des prédictions. Pour le moment, ces modèles ne sont pas supervisés et n'utilisent pas d'informations au-delà d'alignement de séquences de protéines homologues. Comme dans le cas de la prédiction des contacts structuraux entre les résidus, une méthode supervisée, prenant en compte les connaissances *a priori* sur les effets mutationnels (par exemple, des expériences de

mutagenèse ou des données biomédicales liées aux maladies mais aussi des informations structurelles), devrait permettre d'affiner la précision des prédictions.

Discussion et perspectives

Les méthodes coévolutionnaires ont subi d'importants changements au cours des dernières années. Ces méthodes sont maintenant capables de prédire les contacts résidus-résidus à l'intérieur des protéines et entre les protéines en interaction avec une précision sans précédent. À leur tour, ces contacts prédits peuvent être utilisés pour prédire des structures tertiaires et quaternaires actuellement inconnues.

Bien que l'utilisation de ces méthodes en biologie structurale soit bien établie à l'heure actuelle, d'autres champs d'application sont tout juste en train d'émerger. À notre avis, deux des plus prometteurs se rapportent aux domaines suivants. Premièrement, les méthodes coévolutionnaires peuvent être utilisées pour détecter les interactions protéine-protéine entre les familles de protéines et prédire simultanément les contacts entre les protéines des deux familles, qui peuvent être utilisés pour l'assemblage algorithmique de complexes de protéines. Ceci est intéressant car les deux prédictions − c'est-à-dire quelles protéines interagissent, et comment − sont purement basées sur des informations concernant la séquence, offrant un énorme potentiel d'application dans la biologie structurale des systèmes.

Deuxièmement, nous pensons que l'inférence des paysages mutationnels gagnera en importance au cours des prochaines années. Cela concerne d'abord le voisinage de la séquence d'une protéine d'intérêt : la prédiction des effets mutationnels est primordiale dans, par exemple, la détection de variantes pathogènes d'une séquence ou dans la recherche liée à l'émergence d'une résistance aux médicaments chez les bactéries ou dans le cancer. À une échelle plus lointaine, on peut s'attendre à ce que les méthodes coévolutives soient à l'origine de nouvelles possibilités pour la conception de protéines guidée par l'évolution (Socolich *et al.*, 2005).

Il convient également de mentionner que les approches actuelles de modélisation coévolutionnaire présentent des limites évidentes. Elles sont, par exemple, basées sur l'hypothèse selon laquelle une famille de protéines est équivalente à un échantillon de séquences générées indépendamment, alors que les protéines réelles ont des relations phylogénétiques hiérarchiques. On peut s'attendre à ce que la précision déjà remarquable de certains outils basés sur la coévolution augmente encore lorsque de telles limitations seront prises en compte et que les méthodes de calcul seront adaptées en conséquence.

Références

Berman, H.M., Westbrook, J., Feng, Z., Gilliland, G., Bhat, T. N., Weissig, H., Shindyalov, I.N., Bourne, P.E. (2000). The Protein Data Bank. *Nucleic Acids Res*, 28, 235-242.

Biasini, M., Bienert, S., Waterhouse, A., Arnold, K., Studer, G., Schmidt, T., Kiefer, F., Gallo Cassarino, T., Bertoni, M., Bordoli, L., Schwede, T. (2014). SWISS-MODEL: modelling protein tertiary and quaternary structure using evolutionary information. *Nucleic Acids Res*, 42, W1, W252-W258.

Bitbol, A.F., Dwyer, R.S., Colwell, L.J., Wingreen, N.S. (2016). Inferring interaction partners from protein sequences. *Proc Nat Acad Sci USA*, 113, 12180-12185.

Couce, A., Caudwell, L.V., Feinauer, C., Hindré, T., Feugeas, J. P., Weigt, M., Lenski, R.E., Schneider, D., Tenaillon, O. (2017). Mutator genomes decay, despite sustained fitness gains, in a long-term experiment with bacteria. *Proc Nat Acad Sci USA*, 114, E9026-E9035.

Dago, A.E., Schug, A., Procaccini, A., Hoch, J.A., Weigt, M., Szurmant, H. (2012). Structural basis of histidine kinase autophosphorylation deduced by integrating genomics, molecular dynamics, and mutagenesis. *Proc Nat Acad Sci USA*, 109, E1733-E1742.

De Juan, D., Pazos, F., Valencia A. (2013). Emerging methods in protein co-evolution. *Nat Rev Genetics*, 14, 249-261.

Durbin, R., Eddy, S.R., Krogh, A., Mitchison, G., Biological sequence analysis: probabilistic models of proteins and nucleic acids, Cambridge University Press, 1998.

Eddy, S. R. (1998). Profile hidden Markov models. *Bioinformatics*, 14, 755-763.

Edgar, R.C., Batzoglou, S. (2006). Multiple sequence alignment. *Curr Opin Struct Biol*, 16, 368-373.

Ekeberg, M., Lövkvist, C., Lan, Y., Weigt, M., Aurell, E. (2013). Improved contact prediction in proteins: using pseudolikelihoods to infer Potts models. *Phys Review*, 87, 012707.

Feinauer, C., Szurmant, H., Weigt, M., Pagnani, A. (2016). Inter-protein sequence co-evolution predicts known physical interactions in bacterial ribosomes and the Trp operon. *PLoS One*, 11, e0149166.

Feinauer, C., Weigt, M. (2017). Context-aware prediction of pathogenicity of missense mutations involved in human disease. *arXiv preprint arXiv:1701.07246*

Figliuzzi, M., Jacquier, H., Schug, A., Tenaillon, O., Weigt, M. (2015). Coevolutionary landscape inference and the context-dependence of mutations in beta-lactamase TEM-1. *Mol Biol Evol*, 33, 268-280.

Finn, R.D., Bateman, A., Clements, J., Coggill, P., Eberhardt, R. Y., Eddy, S.R., Heger, A., Hetherington, K., Holm, L., Mistry, J., Sonnhammer, E.L., Tate, J., Punta, M. (2014). Pfam: the protein families database. *Nucleic Acids Res*, 42, D222-230.

Göbel, U., Sander, C., Schneider, R., Valencia, A. (1994). Correlated mutations and residue contacts in proteins. *Proteins*, 18, 309-317.

Gueudré, T., Baldassi, C., Zamparo, M., Weigt, M., Pagnani, A. (2016). Simultaneous identification of specifically interacting paralogs and interprotein contacts by direct coupling analysis. *Proc Nat Acad Sci USA*, 113, 12186-12191.

Haldane, A., Flynn, W.F., He, P., Vijayan, R.S., Levy, R.M. (2016). Structural propensities of kinase family proteins from a Potts model of residue co-variation. *Protein Sci*, 25, 1378-1384.

Hopf, T.A., Colwell, L.J., Sheridan, R., Rost, B., Sander, C., Marks, D.S. (2012). Three-dimensional structures of membrane proteins from genomic sequencing. *Cell*, 149, 1607-1621.

Hopf, T.A., Schärfe, C.P., Rodrigues, J.P., Green, A.G., Kohlbacher, O., Sander, C., Bonvin, A.M., Marks, D.S. (2014). Sequence co-evolution gives 3D contacts and structures of protein complexes. *Elife*, 3, e03430.

Hopf, T.A., Ingraham, J.B., Poelwijk, F.J., Schärfe, C.P., Springer, M., Sander, C., Marks, D.S. (2017). Mutation effects predicted from sequence co-variation. *Nat Biotechnol*, 35, 128-135.

Jones, D.T., Buchan, D.W., Cozzetto, D., Pontil, M. (2012) PSICOV: precise structural contact prediction using sparse inverse covariance estimation on large multiple sequence alignments. *Bioinformatics*, 28, 184-190.

Jones, D.T., Singh, T., Kosciolek, T., Tetchner, S. (2015). MetaPSICOV: combining coevolution methods for accurate prediction of contacts and long range hydrogen bonding in proteins. *Bioinformatics*, 31, 999-1006.

Kamisetty, H., Ovchinnikov, S., Baker, D. (2013). Assessing the utility of coevolution-based residue–residue contact predictions in a sequence-and structure-rich era. *Proc Nat Acad Sci USA*, 110, 15674-15679.

Mann, J.K., Barton, J.P., Ferguson, A.L., Omarjee, S., Walker, B.D., Chakraborty, A., Ndung'u, T. (2014). The fitness landscape of HIV-1 gag: advanced modeling approaches and validation of model predictions by *in vitro* testing. *PLoS Comput Biol*, 10, e1003776.

Marks, D.S., Colwell, L.J., Sheridan, R., Hopf, T.A., Pagnani, A., Zecchina, R., Sander, C. (2011). Protein 3D structure computed from evolutionary sequence variation. *PLoS One*, 6, e28766.

Morcos, F., Pagnani, A., Lunt, B., Bertolino, A., Marks, D.S., Sander, C., Zecchina, R., Onuchic, J.N., Hwa, T., Weigt, M. (2011). Direct-coupling analysis of residue coevolution captures native contacts across many protein families. *Proc Nat Acad Sci USA*, 108, E1293-E1301.

Morcos, F., Schafer, N.P., Cheng, R.R., Onuchic, J.N., Wolynes, P.G. (2014). Coevolutionary information, protein folding landscapes, and the thermodynamics of natural selection. *Proc Nat Acad Sci USA*, 111, 12408-12413.

Neher, E. (1994). How frequent are correlated changes in families of protein sequences? *Proc Nat Acad Sci USA*, 91, 98-102.

Nugent, T., Jones, D.T. (2012). Accurate de novo structure prediction of large transmembrane protein domains using fragment-assembly and correlated mutation analysis. *Proc Nat Acad Sci USA*, 109, E1540-E1547.

Ovchinnikov, S., Kamisetty, H., Baker, D. (2014). Robust and accurate prediction of residue–residue interactions across protein interfaces using evolutionary information. *Elife*, 3, e02030.

Ovchinnikov, S., Park, H., Varghese, N., Huang, P.S., Pavlopoulos, G.A., Kim, D.E., Kamisetty, H., Kyrpides, N.C., Baker, D. (2017). Protein structure determination using metagenome sequence data. *Science*, 355, 294-298.

Schug, A., Weigt, M., Onuchic, J.N., Hwa, T., Szurmant, H. (2009). High-resolution protein complexes from integrating genomic information with molecular simulation. *Proc Nat Acad Sci USA*, 106, 22124-22129.

Socolich, M., Lockless, S.W., Russ, W.P., Lee, H., Gardner, K. H., Ranganathan, R. (2005). Evolutionary information for specifying a protein fold. *Nature*, 437, 512-518.

Söding, J. (2004). Protein homology detection by HMM-HMM comparison. *Bioinformatics*, 21, 951-960.

Sułkowska, J.I., Morcos, F., Weigt, M., Hwa, T., Onuchic, J.N. (2012). Genomics-aided structure prediction. *Proc Nat Acad Sci USA*, 109, 10340-10345.

Sutto, L., Marsili, S., Valencia, A., Gervasio, F.L. (2015). From residue coevolution to protein conformational ensembles and functional dynamics. *Proc Nat Acad Sci USA*, 112, 13567-13572.

Uguzzoni, G., John Lovis, S., Oteri, F., Schug, A., Szurmant, H., Weigt, M. (2017). Large-scale identification of coevolution signals across homo-oligomeric protein interfaces by direct coupling analysis. *Proc Nat Acad Sci USA*, 114, E2662-E2671.

UniProt Consortium. UniProt: a hub for protein information. (2015). *Nucleic Acids Res*, 43, D204–212.

Wang, S., Sun, S., Li, Z., Zhang, R., Xu, J. (2017). Accurate de novo prediction of protein contact map by ultra-deep learning model. *PLoS Comput Biol*, 13, e1005324.

Webb B., Sali A. Protein Structure Modeling with MODELLER, in: Kihara D. (ed.), Protein Structure Prediction. Methods in Molecular Biology (Methods and Protocols), vol 1137, Humana Press, New York, 2014.

Weigt, M., White, R.A., Szurmant, H., Hoch, J.A., Hwa, T. (2009). Identification of direct residue contacts in protein-protein interaction by message passing. *Proc Nat Acad Sci USA*, 106, 67-72.

Citation de l'article : Barrat-Charlaix, P. et Weigt, M. (2017) De la variabilité des séquences à la prédiction structurale et fonctionnelle : modélisation de familles de protéines homologues. *Biologie Aujourd'hui*, **211**, 239-244