université PARIS-SACLAY

NNT : 2017SACLXXXX

ÉCOLE POLYTECHNIQUE
UNIVERSITÉ PARIS-SACLAY

# Computational Protein Design : un outil pour l'ingénierie des protéines et la biologie synthètique

Thèse de doctorat de l'Université Paris-Saclay
préparée à l' École Polytechnique

École doctorale n°573
INTERFACES: approches interdisciplinaires, fondements, applications et innovation
Spécialité de doctorat: Les sciences du vivant

Thèse présentée et soutenue à Palaiseau, le 20 Décembre 2017 par

**David Mignon**

Composition du Jury :

Jean-François Gibrat

Thèse de doctorat

université
PARIS-SACLAY

Thèse de doctorat

| | |
|---|---|
| Directeur de Recherche (INRA)<br>responsable scientifique de l'IFB | Rapporteur |
| Yves-Henri Sanejouand<br>Directeur de recherche (Université de Nantes) | Rapporteur |
| Alain Denise<br>Professeur (Université Paris-Sud) | Examinateur |
| Sophie Barbe<br>Chargée de recherche (INSA) | Examinateur |
| Julien Bigot<br>Chercheur (CEA) | Examinateur |
| Thomas Simonson<br>Professeur (École Polytechnique) | Directeur |

**Titre :** Computational Protein Design : un outil pour l'ingénierie des protéines et la biologie synthètique

**Mots clés :** modélisation moléculaire, conception de protéine par ordinateur, Proteus, Monte Carlo, domaine PDZ

**Résumé :** Le CPD ou « Computational protein design » est la recherche par modélisation moléculaire des séquences d'acides aminés compatibles avec une structure protéique ciblée. L'objectif est de concevoir une fonction nouvelle et/ou d'ajouter un nouveau comportement. Le CPD est en développement au sein de notre laboratoire depuis quelques années, avec le logiciel Proteus qui a plusieurs succès à son actif. Au cours de cette Thèse, nous avons enrichi Proteus sur plusieurs points, avec notamment l'ajout d'une méthode d'exploration Monte Carlo avec échange de répliques ou REMC. Une série de comparaisons entre trois méthodes stochastiques de Proteus ont été effectuées : le REMC, le Monte Carlo et une heuristique conçue pour le CPD, le « Multistart Steepest Descent » ou MSD. Ces comparaisons portent sur neuf protéines de trois domaines (SH2, SH3 et PDZ). Nous avons fixé le type de plusieurs acides aminés de nos protéines afin de restreindre l'espace de recherche. Ainsi, à l'aide de techniques de l'optimisation combinatoire, la séquence et la conformation qui minimisent notre fonction d'énergie (le GMEC) sont déterminées pour tous les tests avec moins de 10 positions de la chaîne polypeptidique laissées libres et jusqu'à environ deux tiers des tests avec 20 positions libres. Globalement, le REMC et le MSD donnent de très bonnes séquences en termes d'énergie, avec souvent un accord au GMEC lorsqu'il est connu. Le MSD domine sur les tests à 30 positions actives. Mais le REMC avec huit marcheurs et des paramètres optimisés est plus souvent le meilleur sur les tests tout actif. De plus, comparé à une énumération exacte des séquences sous optimales, le REMC fourni un échantillon de séquences de très bonne diversité. Dans la seconde partie de ce travail, nous avons paramétré notre modèle pour la conception de domaines PDZ. Notre approche du CPD est fondée sur la Physique ; notre fonction d'énergie se base sur la différence entre l'état replié de la protéine et son état déplié. Pour l'état replié, nous avons utilisé un modèle de solvant GB/NEA avec une constante diélectrique égale à 8, puis deux modèles de solvant, le GB/NEA et un nouveau modèle, le GB/FDB avec une constante diélectrique égale à 4. Pour caractériser l'état déplié, nous utilisons un ensemble de potentiels chimiques d'acide aminé ou énergies de références. Ces énergies de références sont déterminées par une procédure de maximisation de la vraisemblance qui permet de reproduire la composition en acides aminés d'un ensemble d'homologues naturels. Les séquences conçues par Proteus sont comparées aux séquences naturelles. Nos séquences sont globalement similaires aux séquences Pfam, au sens des scores BLOSUM40, avec de très bons scores pour les résidus au cœur de la protéine. Le modèle FDB donne toujours des séquences similaires à des homologues naturels modérément éloignés et l'outil de reconnaissance de pli Superfamily appliqué à ces séquences donne d'excellents résultats. Nos séquences ont également été comparées à celles du logiciel Rosetta. La qualité, selon les mêmes critères que précédemment, est très comparable. Mais les séquences de Rosetta restent beaucoup plus proches de la séquence nati ve que celles de Proteus.

**Title :** Computational protein design: a tool for protein engineering and synthetic biology

**Keywords :** molecular modeling, computational protein design, Proteus, Monte Carlo, PDZ domain

**Abstract :** The CPD or Computational Protein Design is the molecular modeling search of the amino acid sequences compatible with a targeted protein structure. The goal is to design a new function and/or add a new behaviour. The CPD has been developed in our laboratory for several years, with the software Proteus which has several successes to its credit. During this thesis, we have enriched Proteus on several points, including the addition of a Monte Carlo exploration method with Replica Exchange or REMC. A series of comparisons of three Proteus stochastic methods have been performed: REMC, Monte Carlo and a heuristic designed for CPD, Multistart Steepest Descent or MSD. These comparisons concern nine proteins from three domains (SH2, SH3 and PDZ). We have set the type of several amino acids of our proteins in order to restrict the search space. Thus, using combinatorial optimization techniques, the sequence and conformation that minimize our energy function (GMEC) is determined for all tests with less than 10 positions of the polypeptide chain left free and up to about two thirds of the tests with 20 free positions. Overall, the REMC and the MSD give very good sequences in terms of energy, often with an agreement with the GMEC when it is known. The MSD dominates the tests at 30 active positions. But the REMC with eight walkers and optimized parameters is most often the best on all active tests. Moreover, compared to an exact enumeration of the sub-optimal sequences, the REMC provides a sample of sequences of very good diversity. In the second part of this work, we have parameterized our model for PDZ domain design. Our CPD's approach is rooted in Physics; our energy function is based on the difference between the folded state of the protein and its unfolded state. For the folded state, we used a GB/NEA solvent model with a dielectric constant of 8, then two solvent models, GB/NEA and a new model, GB/FDB with a dielectric constant equal to 4. To characterize the unfolded state, we use a set of amino acid chemical potentials or reference energies. These reference energies are determined by a maximization of likelihood procedure which allows the amino acid composition of a set of natural homologues to be reproduced. The sequences designed by Proteus are compared to the natural sequences. Our sequences are globally similar to the Pfam sequences, in the sense of the BLOSUM40 scores, with high scores for the residues in the core of the protein. The FDB model always gives sequences similar to moderately distant natural homologues and the Superfamily fold recognition tool applied to these sequences gives excellent results. Our sequences were also compared to those of the Rosetta software. The quality, according to the

same criteria as before, is very comparable. But the Rosetta sequences remain much closer to the native sequence than those of Proteus.