université PARIS-SACLAY
ÉCOLE DOCTORALE
INTERFACES
Approches Interdisciplinaires : Fondements, Applications et Innovation

Titre : Computational protein design : un outil pour l'ingénierie des protéines et la biologie synthétique

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

Résumé : Le « Computational protein design » ou CPD est la recherche 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 dans de notre laboratoire depuis plusieurs années, avec le logiciel Proteus qui a plusieurs succès à son actif. Notre approche utilise un modèle énergétique basé sur la physique et s'appuie sur la différence d'énergie entre l'état plié et l'état déplié de la protéine. 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. Nous avons comparé trois méthodes stochastiques pour l'exploration de l'espace de la séquence : le REMC, le Monte Carlo simple et une heuristique conçue pour le CPD, le « Multistart Steepest Descent » ou MSD. Ces comparaisons portent sur neuf protéines de trois familles de structures : SH2, SH3 et PDZ. En utilisant les techniques d'exploration ci-dessus, nous avons été en mesure d'identifier la conformation du minimum global d'énergie ou GMEC pour presque tous les tests dans lesquels jusqu'à 10 positions de la chaîne polypeptidique étaient libres de muter (les autres conservant leurs types natifs). Pour les tests avec 20 positions libres de muter, le GMEC a été identifié dans 2/3 des cas. Globalement, le REMC et le MSD donnent de très bonnes séquences en termes d'énergie, souvent identiques ou très proches du GMEC. Le MSD a obtenu les meilleurs résultats sur les tests à 30 positions mutables. Le REMC avec huit répliques et des paramètres optimisés a donné le plus souvent le meilleur résultat lorsque toutes les positions peuvent muter. De plus, comparé à une énumération exacte des séquences de faible énergie, le REMC fournit un échantillon de séquences de grande diversité.

Dans la seconde partie de ce travail, nous avons testé notre modèle pour la conception de domaines PDZ. Pour l'état plié, nous avons utilisé deux variantes d'un modèle de solvant GB. La première utilise une frontière diélectrique protéine/solvant effective moyenne ; la seconde, plus rigoureuse, utilise une frontière exacte qui fluctue le long de la trajectoire MC. 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 maximisation d'une fonction de vraisemblance afin de reproduire les fréquences d'acides aminés des domaines PDZ naturels. Les séquences conçues par Proteus ont été comparées aux séquences naturelles. Nos séquences sont globalement similaires aux séquences Pfam, au sens des scores BLOSUM40, avec des scores particulièrement élevés pour les résidus au cœur de la protéine. La variante de GB la plus rigoureuse donne toujours des séquences similaires à des homologues naturels modérément éloignés et l'outil de reconnaissance de plis Superfamily appliqué à ces séquences donne une reconnaissance parfaite. 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 Rosetta présentent moins de mutations que les séquences 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: Computational Protein Design, or CPD is the search for the amino acid sequences compatible with a targeted protein structure. The goal is to design a new function and/or add a new behavior. CPD has been developed in our laboratory for several years, with the software Proteus which has several successes to its credit. Our approach uses a physics-based energy model, and relies on the energy difference between the folded and unfolded states of the protein. During this thesis, we enriched Proteus on several points, including the addition of a Monte Carlo exploration method with Replica Exchange or REMC. We compared extensively three stochastic methods for the exploration of sequence space: REMC, plain Monte Carlo and a heuristic designed for CPD: Multistart Steepest Descent or MSD. These comparisons concerned nine proteins from three structural families: SH2, SH3 and PDZ. Using the exploration techniques above, we were able to identify the Global Minimum Energy Conformation, or GMEC for nearly all the test cases where up to 10 positions of the polypeptide chain were free to mutate (the others retaining their native types). For the tests where 20 positions were free to mutate, the GMEC was identified in 2/3 of the cases. Overall, REMC and MSD give very good sequences in terms of energy, often identical or very close to the GMEC. MSD performed best in the tests with 30 mutating positions. REMC with eight replicas and optimized parameters often gave the best result when all positions could mutate. Moreover, compared to an exact enumeration of the low energy sequences, REMC provided a sample of sequences with a high sequence diversity.

In the second part of this work, we tested our CPD model for PDZ domain design. For the folded state, we used two variants of a GB solvent model. The first used a mean, effective protein/solvent dielectric boundary; the second one, more rigorous, used an exact boundary that flucutated over the MC trajectory. To characterize the unfolded state, we used a set of amino acid chemical potentials or reference energies. These reference energies were determined by maximizing a likelihood function so as to reproduce the amino acid frequencies in natural PDZ domains. The sequences designed by Proteus were compared to the natural sequences. Our sequences are globally similar to the Pfam sequences, in the sense of the BLOSUM40 scores, with especially high scores for the residues in the core of the protein. The more rigorous GB variant always gives sequences similar to moderately distant natural homologues and perfect recognition by the the Superfamily fold recognition tool. Our sequences were also compared to those produced by the Rosetta software. The quality, according to the same criteria as before, was very similar, but the Rosetta sequences exhibit fewer mutations than the Proteus sequences.

Université Paris-Saclay
Espace Technologique / Immeuble Discovery
Route de l'Orme aux Merisiers RD 128 / 91190 Saint-Aubin, France