A new framework for computational protein design through cost function network optimization

Seydou Traore, David Allouche, Isabelle André, Simon de Givry, George Katsirelos, Thomas Schiex, Sophie Barbe

A New Framework for Computational Protein Design

through Cost Function Network Optimization

Seydou TRAORÉ[1-3], David ALLOUCHE[4], Isabelle ANDRÉ[1-3], Simon DE GIVRY[4], George KATSIRELOS[4], Thomas SCHIEX[4*] and Sophie BARBE[1-3*]

[1] Université de Toulouse;INSA,UPS,INP; LISBP, 135 Avenue de Rangueil, F-31077 Toulouse, France
sophie.barbe@insa-toulouse.fr

[2] INRA, UMR792, Ingénierie des Systèmes Biologiques et des Procédés, F-31400 Toulouse

[3] CNRS, UMR5504, F-31400 Toulouse, France

[4] Mathématiques et Informatique Appliquées de Toulouse, UR 875, INRA, F-31320 Castanet Tolosan, France
thomas.schiex@toulouse.inra.fr

Keywords Structural biology, combinatorial optimization, computational protein design.

The engineering of tailored proteins with desired properties holds great interest for applications ranging from medecine, biotechnology and synthetic biology and nanotechnologies. In recent years, structure-based computational protein design (CPD) approaches have demonstrated their potential to adequately capture fundamental aspects of molecular recognition and interactions and have already enabled the successful (re)design of several enzymes (see for example [1]).

One of the main challenges of CPD lies in the exponential size of the conformational and sequence protein space that has to be explored which rapidly grows out of reach of computational approaches. In the simplest form, the CPD problem assumes a fixed protein backbone and, for each type of amino acid considered at a given position, allows the side-chains to move only among a set of discrete and low-energy conformations, called rotamers. CPD is thus formulated as an optimization problem which consists in choosing combinations of rotamers at designable specified positions such that the fold has minimum energy (global minimum energy conformation or GMEC). This problem has been proven to be NP-hard [2]. If several meta-heuristic methods have been applied to it, there also exist methods which solve the GMEC exactly. The most usual is based on the Dead-End Elimination (DEE) theorem [3]. Such exact methods offer several advantages. First, they ensure that discrepancies between CPD predictions and experimental results come exclusively from the inadequacies of the biophysical model and not from the algorithm. Next, because provable methods can determine that the optimum is reached, they may actually stop before meta-heuristic approaches. Finally, empirical studies on solving the GMEC problem reported that the accuracy of meta-heuristic approaches tend to degrade as the problem size increases [4].

In this work, we show that a recent combinatorial optimization technique, defined in the field of "Cost Function Networks" (or Weighted Constraint satisfaction, [5]) can push CPD beyond the limit of usual tools. We propose a new design strategy which starts from a PDB structure, selects mutable and flexible residues identified on the basis of the functional and structural knowledge on the target protein, in particular the amino acid burial in the structure which is captured by the solvation radius [6] and exploits energy fields to reach a final CPD instance. A CPD instance defines an exponential space of possible sequences with associated conformations (choice of rotamer) and the ability to compute the energy of any sequence-conformation configuration using a pairwise energy matrix.

Following this methodology, we have built 35 design cases, involving free proteins, or proteins bound to a cofactor, a ligand or a protein. The studied systems have all been extracted from previously published papers about protein engineering, *in silico* protein design or protein structural studies. The size of the design cases include from 3 to 119 mutable residues and encompass spaces from 4.10^{26} to 2.10^{249}.

For each of these cases, we tried to identify the GMEC using either Osprey 2.0 (a common open source CPD software [7]), or modeled the problem as a Cost Function Network and solved using toulbar2 (a dedicated open source CFN solver[1]) or reformulated it as an Integer Linear programming problem (ILP) and solved it using the commercial IBM ILOG Cplex ILP solver. Within a time-out of 100 hours per design,

[1] Toulbar2 is available at https://mulcyber.toulouse.inra.fr/projects/toulbar2.

Osprey solved 18 cases, CPLEX solved 27 cases and toulbar2 solved 30 instances. No instance unsolved by toulbar2 could be solved by other approaches. With a shorter time-out of 10 minutes, these numbers reduced to 11, 13 and 30, showing the efficiency of the CFN approach.

In practice, finding the GMEC is not always sufficient and a gap-free ensemble of solutions close to (and including) the GMEC is sought in order to design a larger library of protein mutants to be tested experimentally. Indeed, the energetic model used is only an approximation and the GMEC may not be the actual most stable configuration. Furthermore, the most stable configuration may be so stable that it loses the original function of the parental wild type protein. We therefore compared the capabilities of Osprey and toulbar2 for generating the set of all suboptimal solutions within a 2 kcal/mol threshold of the optimum. Among the set of 30 cases for which a GMEC could be previously identified, Osprey 2.0 was able to produce such an ensemble for just one of the simplest design cases (taking around 37 hours), while toulbar2 successfully produced all ensembles for the 30 cases, taking less than 7 hours in all cases.

The produced ensembles could contain up to 8.10^8 different conformations, but never represented more than 3.10^5 different sequences (and often much less). We more thoroughly analyzed 4 cases for which the number of sequences was below 300. Their energy was lower than the wild type model by as much as 20 kcal/mol. We expected that mutations would favor the introduction of bulkier amino acids in order to fill up the free space in the core of proteins. However, changes in amino acid sizes were subtle, probably because of the lack of molecular flexibility in the underlying model and the discretization generated by rotamers. After submitting the best conformation of each unique sequence to energy minimization, we observed important decreases in energy (up to 60 kcal/mol) but the superposition of the structures before and after minimization only showed slight rearrangements of side chains and backbone. This clearly indicates that slight geometrical adjustments can significantly lower model energies. This minimization also often changed the GMEC, showing the importance of ensemble generation and post-minimization based re-ranking.

Beyond providing a design framework and computational tools to facilitate the optimization of highly combinatorial design cases, our approach also has the potential to speedup methods that integrate more flexibility as long as they reduce to the same type of optimization problems [8]. This is even more important as this considerably expands the size of the search space or may require solving a large number of GMEC instances

Acknowledgments

This work was supported by the “Agence Nationale de la Recherche”, references ANR 10-BLA-0214 and ANR-12-MONU-0015-03. We thank the Computing Center of Region Midi-Pyrénées (CALMIP, Toulouse, France) and the GenoToul Bioinformatics Platform of INRA-Toulouse for providing computing resources and support. S. Traoré was supported by a grant from the INRA and the Region Midi-Pyrénées.

References

[1] Khare,S.D. et al. Computational redesign of a mononuclear zinc metalloenzyme for organophosphate hydrolysis. *Nat. Chem. Biol*., 8, 294–300. 2012.

[2] Pierce,N.A. and Winfree,E. (2002) Protein Design is NP-hard. Protein Engineering, 15, 779–782.

[3] Desmet,J. et al. The dead-end elimination theorem and its use in protein sidechain positioning. *Nature*, 356, 539–542. 1992.

[4] Voigt,Christopher A. et al. Trading accuracy for speed: A quantitative comparison of search algorithms in protein sequence design. *Journal of Molecular Biology*, 299, 789–803, 2000.

[5] Allouche,D. et al. Computational Protein Design as a Cost Function Network Optimization Problem. In *Proc. of CP*. 2012.

[6] Archontis,G. and Simonson,T. A residue-pairwise Generalized Born scheme suitable for protein design calculations. *J. Phys. Chem. B*, 109, 22667–22673. 2005.

[7] P. Gainza et al. OSPREY: Protein design with ensembles, flexibility, and provable algorithms. *Methods Enzymol.* 2013;523:87-107.

[8] Hallen,M.A. et al. Dead-end elimination with perturbations (DEEPer): a provable protein design algorithm with continuous sidechain and backbone flexibility. *Proteins*, 81, 18–39. 2013.