Bilevel optimization and its bicriteria approximation in computational protein design

Samuel Buchet, Marianne Defresne, Simon de Givry, Manon Ruffini, Thomas Schiex

Université Fédérale de Toulouse, ANITI, INRAE, UR 875, 31326 Toulouse, France
contact author: thomas.schiex@inrae.fr

Mots-clés : *cost function network, discrete optimization, negative multistate design, bilevel optimization*

Proteins are the main active molecules of Life. While natural proteins play many roles, as enzymes or antibodies for example, there is a need to go beyond the repertoire of natural proteins to produce engineered proteins that precisely meet application requirements, in terms of function, stability, activity or other protein capacities. Computational Protein Design (CPD) aims at designing new proteins using full-atom molecular or coarse-grain models.

CPD methods aim at finding a sequence of amino acids that folds into a target 3D structure, that corresponds to the desired properties and functions. A general formulation of this problem being highly intractable, simplifying assumptions have been made : the target protein structure (or backbone) is often assumed to be rigid, the continuous space of flexibility of amino acids side-chains is represented as a discrete set of conformations called rotamers and the atomic forces that control the protein stability are represented as a decomposable energy function, defined as the sum of terms involving at most two bodies (amino acids). The problem of design is then reduced to a purely discrete optimization problem : given a rigid backbone, one must find a combination of discrete side-chain natures and conformations (rotamers) that minimizes the energy. The resulting sequence and associated side-chain conformations define the Global Minimum Energy Conformation (GMEC) for the target backbone.

Unfortunately, the rigid backbone and discrete rotamer simplifications ignore protein flexibility, even though proteins are known to not remain in a single position. They rather evolve in a set of low energy conformations. To integrate flexibility in the design model, multistate design requires an ensemble of backbone conformations as the description of the input structure. This favors the design of flexible proteins that move slightly around their *ideal* conformation, or perform large conformational changes. Negative multistate design allows for both desired and undesired structures as an input. The sought sequence is aimed at stabilizing the positive (desired) states while destabilizing the negative (undesired) states. This enables for example the modeling of affinity improvement : the desired state is the bound structure, and the undesired state the unbound structure. Target specificity can also be modeled : complexes with the protein bound to the desired and undesired ligands are the positive and negative states respectively. Negative design is a challenging NP^{NP}-complete problem [8].

Constraint programming is a discrete optimization technology focused on constraint satisfaction/feasibility problems, approaching optimization by iteratively solving feasibility problems. Its generalization into weighted constraint programming (*aka* cost function networks) focuses on optimization problems, which also allows it to tackle machine learning problems defined on discrete probabilistic graphical models (Markov random fields, Bayesian networks).

A cost function network factorizes a complex function in local functions on discrete variables. Each function returns a cost for any assignment of its variables. Constraints are represented as functions with costs in $\{0, \top\}$ where $\top$ is an upper bound cost associated with forbidden assignments. The goal is to find a non-forbidden assignment of all variables that minimizes the sum of all the local functions. This problem is NP-hard. Several techniques have been developed to solve it using exact branch and bound methods relying on soft local consistencies [2, 3].

Mutable position	26	28	100
N_{flex}	18	18	21
D	3135	2122	2653
iCFN	103.37	81.50	62.07
toulbar2	**8.52**	**17.23**	**11.28**

FIG. 1 – Comparison of CPU-times (in seconds) for iCFN [4] and toulbar2 with bilevel optimization, for solving the negative design, with one positive and one negative state. N_{flex} is the number of flexible positions, besides the one mutable residue. D is the maximum domain size.

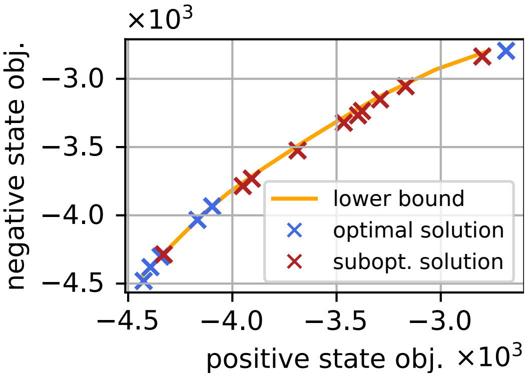

FIG. 2 – Bicriteria optimization on another negative protein design formulation.

The single-state protein design problem can be easily cast as a cost function network [1, 7, 6]. In negative design, the problem can be formulated as a bilevel query on a cost function network. Existing minimization mechanisms, such as branch-and-bound and local consistencies, were adapted to bilevel optimization in the CFN solver toulbar2 (see Fig. 1).[1]

Alternatively, coarse-grain models (side-chain conformations are precompiled by the model) allow to formulate the negative design as a bicriteria optimization problem where the criteria are cost function networks representing the desired and undesired states respectively. As a first tested approach, a linear combination of these networks allows to find a subset of pareto solutions by exploiting the existing methods in toulbar2 (Fig. 2). This approach allows to reduce the problem complexity from NP^{NP} to NP. From a wider perspective, this problem can also be tackled as a global constraint on the secondary network controlling its maximum (and/or minimum) total cost. These developments have a broad application interest and solving these problems will require upper bounding techniques such as Lagrange relaxation [5].

Références

[1] D Allouche, J Davies, S de Givry, G Katsirelos, T Schiex, S Traoré, I André, S Barbe, S Prestwich, and B O'Sullivan. Computational protein design as an optimization problem. *Artificial Intelligence*, 212 :59–79, 2014.

[2] M. Cooper, S. de Givry, M. Sanchez, T. Schiex, M. Zytnicki, and T. Werner. Soft arc consistency revisited. *Artificial Intelligence*, 174(7–8) :449–478, 2010.

[3] Martin C. Cooper, Simon de Givry, and Thomas Schiex. Graphical models : Queries, complexity, algorithms (tutorial). In *37th International Symposium on Theoretical Aspects of Computer Science (STACS-20)*, volume 154 of *LIPIcs*, pages 4 :1–4 :22, Montpellier, France, 2020.

[4] Mostafa Karimi and Yang Shen. icfn : an efficient exact algorithm for multistate protein design. *Bioinformatics*, 34(17) :i811–i820, 2018.

[5] Tahrima Rahman, Sara Rouhani, and Vibhav Gogate. Novel upper bounds for the constrained most probable explanation task. *Advances in Neural Information Processing Systems*, 34 :9613–9624, 2021.

[6] M. Ruffini. *Models and Algorithms for Computational Protein Design.* PhD thesis, University Toulouse 3 Paul Sabatier, 2021.

[7] Manon Ruffini, Jelena Vucinic, Simon de Givry, George Katsirelos, Sophie Barbe, and Thomas Schiex. Guaranteed diversity and optimality in cost function network based computational protein design methods. *Algorithms*, 14(6) :168, 2021.

[8] Jelena Vucinic, David Simoncini, Manon Ruffini, Sophie Barbe, and Thomas Schiex. Positive multistate protein design. *Bioinformatics*, 36(1) :122–130, 2020.

[1] `https://github.com/toulbar2/toulbar2`