Approximate Counting with Deterministic Guarantees for Binding Affinity Computation

Clément Viricel[1,2], David Simoncini[1], David Allouche[1]
Simon de Givry[1], Sophie Barbe[2], Thomas Schiex[1]

1 - Unité MIAT UR 875, INRA, F-31320 Castanet Tolosan, France,
2 - LISBP, INSA, UMR INRA 792/CNRS 5504, F-31400 Toulouse, France

Proteins are polymer chains of amino-acids. Natural evolution, by means of amino-acid sequence variations (mutations, recombinations and duplications), have fashioned an array of proteins with functions ranging from catalysis, signaling to recognition and repair [4]. However, for many applications in biotechnology, nanotechnology, green chemistry and medicine, there is an ever-increasing demand for proteins endowed with specific properties/functions which are not known to exist in nature. To help protein engineering tailoring desired proteins, structure-based Computational Protein Design (CPD) has emerged. CPD aims at rationally designing amino-acid sequences that fold into a given three-dimensional structure and that will bestow the designed protein with targeted properties. Most existing methods aimed at solving CPD problem are energy-minimization based stochastic meta-heuristics. Exact deterministic methods for energy minimization traditionally use Dead End Elimination (DEE) and A* enumeration [7]. More recently, Cost Function Network (CFN) relying on Depth First Branch and Bound (DFBB) and Local Consistency (LC) [1, 9] have been shown to outperform them for energy minimization. Here, we are interested in the computational design of proteins with the best possible affinity for a given partner (other protein, a small molecule,...) which is essential for large range of applications.

One can measure the potential of interaction between two molecules by means of the binding affinity constant $K_A \propto \frac{Z^C}{Z^A.Z^B}$ with C the complex formed by protein A and ligand B and Z its respective partition function and try to design proteins maximizing this affinity for interesting partners. The partition function of a protein with sequence S is defined as $Z^S = \sum_{\ell \in \Lambda^S} e^{-N \frac{E^S(\ell)}{k_B T}} = \sum_{\ell \in \Lambda^S} p^S(\ell)$ where $E^S(\ell)$ is the energy of the protein in conformation ℓ. By exploiting pairwise decomposable energy fields and discrete libraries of conformations, this problem can be reduced to the computation of the normalizing constant of a Markov Random Field where discrete variables represent conformation of each amino-acid and additive potential functions represent terms of the decomposed energy. This normalizing constant contains an exponential number of terms. The Boltzmann distribution leads to terms with sharp changes in magnitude, where most significant terms correspond to low energies. Algorithms exist that either offer approximations of Z with probabilistic guarantees [6, 3] or exact computations by reduction to the #P-complete #-SAT problem [8]. Our aim is to provide a deterministic algorithm that computes a lower approximation $\hat{Z}^S$ with a guarantee that $\frac{Z^S}{1+\varepsilon} \leq \hat{Z}^S \leq Z^S$.

This is already achieved by the K^* algorithm [5] that uses a combination of two algorithms DEE and A^*. DEE is a local dominance analysis that prune strongly dominated rotamers. A* is a best-first tree search algorithm. A* has the nice property that it produces conformation in decreasing order of energy (thus in increasing order of probability). K^* accumulates the most important probabilities first and stops when the desired level of approximation is reached.

We propose instead a family of algorithm called Z^* [10] that relies on optimization lower bounds developed for solving Cost Function networks (Weighted CSP) based on local consistencies [2] (instead of DEE), related to convergent message passing bounds based on linear programming in MRF [11]. We also use a polynomial space Depth First Branch and Bound tree-search algorithm

	K^*		Z_0^* (Ub_0)		Z_1^* (Ub_1)	
PDB ID	nodes	time	nodes	time	nodes	time
1AMU	6.45	1278	0.0845	0.5	23%	30%
1TP5	∞	∞	3.19	31	51%	47%
1B74	∞	∞	5.50	35	41%	35%
2Q2A	∞	∞	39.9	596	56%	43%

Table 1: For each system, we give the number of nodes ($\times 10^6$) and user cpu-time in minutes. We used $\varepsilon = 10^{-3}$. The last row is the purcentage reduction of explored node and time cost.

instead of the exponential space A^*. Soft local consistencies reformulate energies, trying to bring as much pairwise energies to unary and zero-ary energies. The DFBB algorithm performs counting instead of minimization and uses a simple dedicated dynamic pruning condition based on an invariant that guarantees to compute an ε-approximation of Z.

If we have a lower bound lb on the energy of all conformations, $p_{lb} = e^{-lb/k_bT}$ is an upper bound on the probability of all conformations. Given a partial conformation ℓ, we can conclude that $Ub_0(\ell) = N \cdot p_{lb}$ (with N the number of possible complete conformations refining ℓ) is an upper bound on the mass of probability. We can easily tighten this bound by taking into account unary energies in the reformulated model as $Ub_1(\ell) = p_{lb} \times \prod(\sum_{a \in \Lambda_{S_i}} e^{-E_i(a)/k_bT})$. During tree search, by accumulating probability masses of found conformations in a running estimation $\hat{Z}$, upper bounds of pruned probability masses in U and enforcing the invariant $U < \varepsilon \hat{Z}$ to decide when to prune, we have the deterministic guarantee to have an ε-approximation of Z. To compare K^* with z^* using our two bounds, we examined the binding affinity of different protein/ligand complexes using our approach (denoted as Z^*) with the two described bounds and K^* as implemented in the CPD software OSPREY[5] (Table 1). We observe that Z^* is able to compute ε-approximations of the partition function much faster than K^*.

References

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

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

[3] Ermon, S., Gomes, C. P., Sabharwal, A., and Selman, B. Taming the curse of dimensionality: Discrete integration by hashing and optimization. *arXiv preprint arXiv:1302.6677* (2013).

[4] Fersht, A. *Structure and mechanism in protein science: a guide to enzyme catalysis and protein folding*. WH. Freemean and Co., New York, 1999.

[5] Georgiev, I., Lilien, R. H., and Donald, B. R. The minimized dead-end elimination criterion and its application to protein redesign in a hybrid scoring and search algorithm for computing partition functions over molecular ensembles. *Journal of computational chemistry 29*, 10 (July 2008), 1527–42.

[6] Hazan, T., Maji, S., and Jaakkola, T. On sampling from the Gibbs distribution with random maximum a-posteriori perturbations. In *Advances in Neural Information Processing Systems* (2013), pp. 1268–1276.

[7] Leach, A. R., and Lemon, A. P. Exploring the conformational space of protein side chains using dead-end elimination and the A* algorithm. *Proteins 33*, 2 (Nov. 1998), 227–39.

[8] Sang, T., Beame, P., and Kautz, H. A. Performing bayesian inference by weighted model counting. In *AAAI* (2005), vol. 5, pp. 475–481.

[9] Traoré, S., Allouche, D., André, I., de Givry, S., Katsirelos, G., Schiex, T., and Barbe, S. A new framework for computational protein design through cost function network optimization. *Bioinformatics 29*, 17 (2013), 2129–2136.

[10] Viricel, C., Simoncini, D., Allouche, D., de Givry, S., Barbe, S., and Schiex, T. Approximate counting with deterministic guarantees for affinity computation. In *Modelling, Computation and Optimization in Information Systems and Management Sciences*. Springer, 2015, pp. 165–176.

[11] Wainwright, M. J., and Jordan, M. I. Graphical models, exponential families, and variational inference. *Foundations and Trends® in Machine Learning 1*, 1-2 (2008), 1–305.