Ab Initio Protein Structure Prediction with a Dipeptide-assembly Evolutionary Algorithm

Andrea Bazzoli
Department of Information Technologies,
University of Milan,
via Bramante 65,
26013 Crema (CR), Italy
bazzoli@dti.unimi.it

Giorgio Colombo
Istituto di Chimica del Riconoscimento Molecolare,
CNR,
via Mario Bianco 9,
20131 Milano, Italy
g.colombo@icrm.cnr.it

Andrea G. B. Tettamanzi
Department of Information Technologies,
University of Milan,
via Bramante 65,
26013 Crema (CR), Italy
tettamanzi@dti.unimi.it

Here is described an evolutionary algorithm that predicts the native structure of single-chain proteins by minimizing a fitness function on a discretized conformational space. Predictions are made *ab initio*, i.e., without taking any known protein structure as a starting template for the search.

The computational model of protein considers only the heavy atoms of the backbone, and reduces non-Glycine side chains to a single virtual C^β, which is located along a fixed direction but whose distance from C^α depends on amino acid type. Bond lengths and angles are set to ideal values and peptide units are assumed to be planar, limiting the conformational degrees of freedom of an n-residue protein to the sequence $\{\psi_0, \phi_1, \ldots, \psi_{n-2}, \phi_{n-1}, \psi_{n-1}\}$ of backbone dihedral angles.

Candidate dihedral sequences are generated by assembling native (ψ, ϕ) pairs, that were extracted from a database of known protein structures and grouped by dipeptide type into $20 \times 20 = 400$ sets. Each set, representable as in Fig. 1, contains the only (ψ, ϕ) pairs allowed by the assembly procedure to a specific dipeptide type, and can actually be thought to provide the local conformations that, in nature, are most likely for that type. The last dihedral angle, ψ_{n-1}, needed to locate the C-terminal oxygen, is restrained to only 64 values uniformly distributed in $[-180°, 180°)$.

Due to the finite cardinality of the sampling sets used in the assembly procedure, dihedral sequences are encoded as strings of integer genes (chromosomes). At each generation of the evolutionary algorithm, these are subjected to one-point crossover with probability 0.7 and to single-gene blind mutation with probability 0.001; The population size is 800 and the total number of generations is 1600.

The fitness of a chromosome, a numerical evaluation of the structure it encodes, is expressed as a linear combination of three terms, measuring the amount of atomic collisions in the structure, its amino acid contact energy, and its radius of gyration, respectively. The weights of the three terms were chosen after numerous experiments on a training set of 12 proteins, which were aimed at maximizing the correlation between fitness and RMSD (Root Mean Square Deviation) to the native structure. Interestingly, the highest correlation is achieved when the weight of the radius of gyration is far greater than the other two, in a combination that biases the evolutionary algorithm toward compact structures which are also sterically feasible. Even so, the algorithm is unable to capture the native secondary structure, especially in the case of beta strands, and cannot therefore compete with current state-of-the-art predictors.

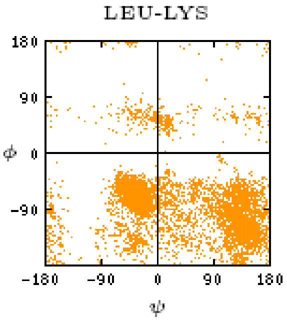

Figure 1: Representation of the (ψ, ϕ) sampling set for dipeptide type Leucine-Lysine. As opposed to the classical Ramachandran plot, ψ values are reported on the horizontal axis and ϕ values on the vertical axis, emphasizing the fact that, for each (ψ, ϕ) pair, ψ precedes ϕ along the polypeptide chain.

REFERENCES

[1] S. Schulze-Kremer. Genetic algorithms for protein tertiary structure prediction. In *Parallel Problem Solving from Nature II*, pages 391–400. North Holland, 1992.

[2] S. Sun. Reduced representation model of protein structure prediction: statistical potential and genetic algorithms. *Protein Science*, 2:762–785, 1993.

Categories and Subject Descriptors

G.3 [**Probability and Statistics**]: Probabilistic Algorithms; J.3 [**Life and Medical Sciences**]: Biology and Genetics

General Terms

Experimentation

Keywords

Evolutionary Algorithms, Protein Structure Prediction

GECCO'07 July 7–11, 2007, London, England, United Kingdom
ACM 978-1-59593-697-4/07/0007.