Application of Genetic Algorithms to Modeling and Design Proteins

Luis P. B. Scott
CMCC - Universidade Federal do ABC , Rua Catequese, 242 – Jardim, CEP: 09090-400 - Santo André – SP, Brasil, Phone: (55) (11) 4437-1600, Fax: (55) (11) 4437-1600, email: luis.scott@ufabc.edu.br

The information for life is stored by a four-letters alphabet in the genes (DNA). Proteins are, among others, the macromolecules that perform all important task in organism as catalysis of biochemical reactions, transport, recognition. The three-dimensional structure (tertiary structure) of proteins determines their function. This fact has been described as the determination of the second genetic code [1,2,3]. Potentials derived from *ab initio* principles or statistical potentials based on structural databases have been used in the simulations which are performed through a variety of methods such as molecular dynamics, monte carlo simulations, genetic algorithms, neural networks, simulated annealing to predict the secondary and tertiary structure of proteins and to optimize the conformation of macromolecules [4]. This work discusses the use of genetic algorithms (GA) to design new sequences to the hydrophobic core of a protein, Cytochrome b_{562} and predict the tertiary structure. Starting from the known PDB structure of its backbone, which is maintained fixed, the side chains of the hydrophobic core are allowed to adopt the conformations present in the rotamer library built from a structural database. The atoms of the side chains forming the core interact via van der Waals energy. The rapidly growing sequence-structure gap has enticed theoreticians to solve simplified prediction problems The structure prediction of the hydrophobic core of ***Cytochrome B_{562}*** has been object of study by means of Automata Network along with the rotamer library of Ponder and Richards [5]. The obtained conformation for the native sequence of the core was in the 13% lowest energy sequences out of 170 proposed sequences, from which 22 sequences had energies below the native energy. In the present work we revisit this study by means of genetic algorithms and the same rotamer library to optimize the side chains of the hydrophobic core of ***Cytochrome B_{562}*** and compare with their results. The genetic algorithm showed good performance and was able to find sequences whose structures are very similar to the core structure and energy slightly lower than the native sequence. The algorithm convergence is typically observed after 70 or 80 generations. A fundamental question regarding the implementation of GA is the effect of a larger initial population of conformations. Will the lowest energy sequences show higher identity to the native one as the starting population increases? That is, a larger search in the conformation space of the rotamers would result in low energy sequences more similar (identity) to native. The GA generates low energies sequences with higher identity to the native and also a higher number of structures which are more similar to the native It was noted that an increase in the starting population lead to a better agreement between structures of the alternative sequences for the core and the native core. Concomitantly to the agreement in the structure, the identity also increased, thus correlating sequence and structure.

1. Baldi P. et al., Exploiting the past and the future in protein secondary structure prediction: Bioninformatics. 15:11: 937-946 (1999).
2. Dill et al., Principles of Protein folding – A perspective form simple exact models. Protein Science. 4:561-602 (1995).
3. Kollman, P. A.; Djam Y. and Lee, M. R., State of the art in studying protein Folding and protein structure prediction using molecular dynamics methods. J. Molecular Graphics and Modeling. Vol 19:I1:146-149 (2001).
4. Desjarlais, J. R and Handel T. M., Side-chain backbone flexibility in protein core design. Journal of Molecular Biology. vol. 290:I:July:305-318 (1999).
5. Ponder, J. W.; Richards F. M., Tertiary templates for protein use packing criteria in the enumeration of allowed sequences for different structural classes. J. Molecular Biology, v. 193:775-791 (1987).