Feature Discovery in a Protein Database using a Memetic Algorithm-Based Solution

[1] Denise Fukumi Tsunoda, [2] Heitor Silvério Lopes
[1,2] Laboratório de Bioinformática / CPGEI
Centro Federal de Educação Tecnológica do Paraná (CEFET-PR)
Av. 7 de setembro, 3165 – 80230-901 Curitiba (PR), Brazil
[1] **denise@cpgei.cefetpr.br,** [2] **hslopes@cpgei.cefetpr.br**

ABSTRACT

Proteins can be grouped into families (and these families into superfamilies) according to some features such as hydrophobicity, composition, structure or function, with the objective of establishing the common biological functions.

This paper presents the results of the use of a system that was conceived to discover features (sequences of amino acids) that occur very often in proteins of a given class (family) but rarely occur in proteins of other classes. These features (or motifs) can be used for the classification of new proteins. In order to find such features, the system consists of a hybrid Genetic Algorithm (GA) developed specifically for discovering identical or similar amino acids sequences extracted from the Protein Data Bank – PDB (release 102, oct. 2002). This kind of GA is also known as Memetic Algorithm (MA) [1], since it uses global (mutation, crossover) and local search (expansion) operators.

Methods usually employed for the alignment of protein sequences use substitution matrices such as PAM or BLOSUM that provides probabilities for all possible exchanges between amino acids. Dayhoff and coworkers [2] studied residue replacements in closely related proteins and constructed the PAM (for "point accepted mutation") model of molecular evolution. Another approach for estimating target frequencies and the corresponding log-odds matrices was proposed by Henikoff and Henikoff [3]. They examined multiple alignments of distantly related protein regions. In our work, both PAM250 and BLOSUM62 matrices were used.

The system generates a random initial population of individuals, each one representing a possible feature (motif). The quality of the candidate solution represented by the individual (motif) is calculated as follows. First, for each class of proteins, the relative frequency of the motif in that class is calculated. This can be made using a substitution matrix to align the possible solution and the proteins of the database (either PAM or BLOSUM or none). Then, the system computes a measure of the class-discrimination ability of the individual (how good this feature is for an automatic protein classification system). If the result has a minimum pre-defined quality, the expansion operator is used to expand the motif length. This operator starts by locating the position of the feature in the first protein of the associated class. Then it adds the immediately preceding amino acid (of the protein) to the motif and recomputes its quality. If the result is not worst than the previous one, the amino acid is added to the individual. This process is repeated until the quality becomes lower than the previous one. At this point the same process is applied for the subsequent amino acids.

Two subsets of the PBD – enzymes and membrane proteins – were used to evaluate the performance of the system. The enzymes dataset (EZ) has 900 proteins divided into 6 subfamilies. The membrane proteins dataset (MP) has 304 proteins divided into 5 subfamilies. The system uses a well-known 5-fold cross validation procedure: the dataset is divided into 5 mutually-exclusive partitions. One of the partitions is used as the test set and the others as the training set. The average accuracy rate over 5 executions (5 test sets) is the accuracy rate reported as the final result. Experiments were done using PAM250, BLOSUM62 and no substitution matrix. Results are normalized in the range [0..1], where 1 is the best possible discrimination. Using the EZ dataset, the system obtained accuracy rates between 0.69 and 0.79 and for the MP those rates ranged between 0.50 and 0.67.

Results demonstrate that the use of the expansion operator allowed the MA to obtain longer motifs with equivalent fitness than the short sequences found in previous experiments. Also, for the datasets used, accuracy rates indicate that it is possible to use simple motifs to characterize families of proteins. Furthers experiments shall be done aiming to combine motifs and obtain even more efficient classifications.

REFERENCES

[1] Moscato, P., **On Evolution, Search, Optimization, Genetic Algorithms and Martial Arts**: Towards Memetic Algorithms. Caltech Concurrent Computation Program, C3P Report 826, 1989.

[2] Dayhoff, M.O. and Schwartz, R.M. Matrices for detecting distant relationships. **Atlas of Protein Sequence and Structure**, vol. 5, pp. 353-358, 1978.

[3] Henikoff, S. and Henikoff, J.G. Amino acid substitution matrices from protein blocks. **Proc. Natl. Acad. Sci.**, vol. 89, pp. 10915-10919, 1989.