Real-Coded Genetic Algorithm to Reveal Biological Significant Sites of Remotely Homologous Proteins

Sung-Joon Park and Masayuki Yamamura

Interdisciplinary Graduate School of Science and Engineering
Tokyo Institute of Technology
R1-418, 4259 Nagatsuta, Midori, Yokohama, 226-8502, Japan
{park,my}@es.dis.titech.ac.jp
http://www.es.dis.titech.ac.jp/

1 Introduction

Discovering biological importance from protein structures needs to utilize heuristic approaches. Since three-dimensional(3D) protein structures play a crucial role in biological reactions, comparing new proteins with well-studied proteins is vital for understanding the native functions. A few residues forming in geometrically close positions activate such biological functions.

Our goal is to find biological significant sites from protein pairs. Since a pair belonging different protein families has lower global similarity but similar function (so called *remotely homologous*), we develop a GA-based alignment tool to emphasize small regions of protein pairs geometrically similar.

2 Method

The proposed Real-coded GA, GSA (Genetic Structural Alignment), optimizes an isometric transformation consisting of Euler's angle $\mathcal{R}$ for rotation and a translation vector $\overrightarrow{T}$ for superposition. $\mathcal{R}$ and $\overrightarrow{T}$ code individuals in GSA, i.e. six-dimensional function optimization. When the pair of a query protein (Stc_1) and a reference (Stc_2) are given ($length(Stc_1) \leq length(Stc_2)$), the 1st C$\alpha$[1] atom of Stc_1 connects to the last Cα of Stc_2. The structures move to the absolute origin. A 3D rectangle containing Stc_1 and Stc_2 is defined, and $\overrightarrow{T}$s of the initial population are plotted in the rectangle. A $\overrightarrow{T}$ defines the position of the 1st Cα of Stc_1, and a $\mathcal{R}$ of an individual rotates Stc_1; the 1st Cα is the origin of rotation.

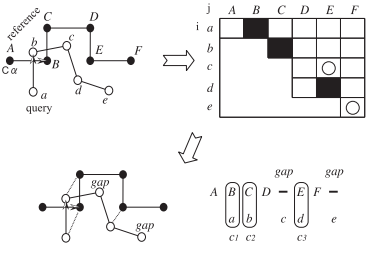

Fig. 1. Estimation of equivalent Cα atom pairs in fitness function.

[1] α-carbon. The side chain of an amino acid links to a Cα. The backbone of a protein is consecutive Cα atoms.

E. Cantú-Paz et al. (Eds.): GECCO 2003, LNCS 2724, pp. 1602–1603, 2003.

UNDX [1] generates two sets of six-dimensional real numbers for two children at a cross time by using the positional relationship of three parents. The best offspring and a randomly selected member replace the parents.

Estimating equivalent Cα atoms is prerequisite to evaluate individuals. If d_{ij} between the ith Cα of Stc_1 transformed by a individual and the jth of Stc_2 is the nearest pair and less than δ, it is added to equivalent set $\mathcal{C}$. Otherwise, the ith atom corresponds to a gap (Fig. 1). Once $\mathcal{C} = \{c_1, c_2, \ldots, c_z\}$ and the number of gaps g are defined, the fitness function f evaluates the individual;

$$s = \sum_{i=1}^{z} \mathrm{e}^{\varepsilon \times d_{c_i}} \qquad (1) \qquad\qquad f = \frac{s + 1.0}{g + 1.0} \qquad (2)$$

3 Results and Conclusion

We adjusted parameters for GSA and finally set to generation=3000, population=50, cross-time=100, UNDX α=0.5, β=0.3, ε=-0.8, δ=2.24. Fig. 2 shows distribution of $P_1 = \frac{z}{length(Stc_1)} \times 100$ and $P_2 = \frac{z'}{z} \times 100$, where z' is the number of equivalent pairs being less than 0.5 in distance. The protein pairs in Fig. 2 have less than 30% sequence identity.

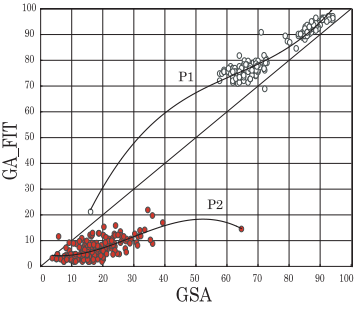

Fig. 2. Distribution of equivalences.

GA_FIT [2], a SGA using dynamic programming(DP), has found more z than GSA. It is conspicuous, however, that GSA possesses a great number of z' in the equivalent set. Such geometrically conserved Cα atoms in the protein pair have a significant possibility to present similar biological function. DP-based methods can find topological similarity from global protein structures. On the other hand, a few Cα atoms are lost for superimposing global structures.

The backbone of a protein is rigorously changed by the amino acid side chains, but the backbone is conservative for keeping its function [3]. For this paradox, comparing proteins has to consider both 3D coordinates of the backbone and chemical properties. We intend to add extra information to the fitness function for finding biologically reliable alignments.

References

1. Ono, I., Kobayashi, S.: A Real-coded Genetic Algorithm for function optimization using unimodal normal distribution crossover. Proc. the 7th ICGA (1997) 246–253
2. May, A.C.W., Johnson, M.S.: Improved genetic algorithm-based protein structure comparisons: pairwise and multiple superpositions. Protein Engng. **8** (1995) 873–882
3. Chothia, C., Lesk, A.M.: The relation between the divergence of sequence and structure in proteins. EMBO J. **5** (1986) 823–826