Genetic Structural Alignment :
Using Genetic Algorithm to Detect High Structural Similarities in Two Proteins

Sung-Joon Park

Dept. of Computational Intelligence and Systems Science,

Tokyo Institute of Technology

park@es.dis.titech.ac.jp

Abstract

The comparison of proteins as three-dimensional coordinate structure is called the structural alignment. Many effective algorithms have been developed and utilized which are based on dynamic programming and root square mean deviation. The results of these methods are evaluated for average of square distances between pairs of Cα atoms. In consequence, the most important sites in functions of proteins can be ignored in the obtained alignments. Moreover, they are only searching local minima. In this thesis, a novel alignment method is proposed that is on the basis of the Real-coded Genetic Algorithm : *Genetic Structural Alignment*(GSA). GSA can align much more importance on conserved and active sites with a effective fitness function. The results of two experiments are described and shown, in regard to GA and GSA. It is reported that GSA found a novel alignment result on Ca^{2+}-binding proteins. Finally, an idea for the multiple structural alignment by suggested GSA is also described.

1 Introduction

The amino acid is a molecule made of hydrogen, carbon, nitrogen, oxygen, and other atoms. These amino acids are connected into a chain by peptide bonds. When the amino acid group (NH_2) of one amino acid and the carboxyl group (COOH) of another amino acid react, a water molecule is removed and the two amino acids are connected. This connection is called a peptide bond and protein has a polypeptide bond in which n amino acids are connected together by peptide bond. There exist 20 kinds of side chains in all organisms, illustrated as R_n in Figure 1. The side chains identify the component amino acids. The 20 amino acid side chains are organized by the general properties and chemical structures. These side chains fall into the following chemical classes, i.e. six aliphatic, three aromatic, two sulfur-containing, two alcohols, three bases, two acids, and two amides.

The part linked to the side chain is a Cα atom. The Cα atom chain makes every protein have a backbone. A backbone can be used to explain a protein structure. Proteins can be folded

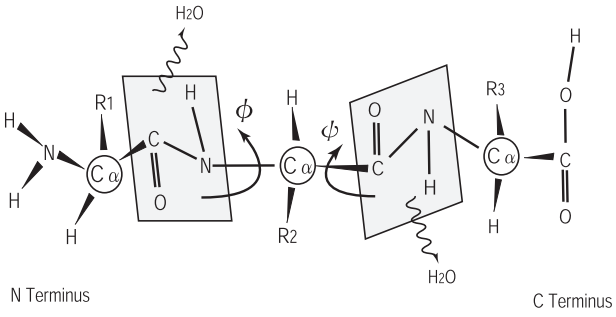


Figure 1: Polypeptide Chain

in three-dimensions because N-Cα bond and Cα-C bond may rotate. These rotation angles are known as ϕ and ψ. These angles can make local conformation regularities in protein structure, e.g. α-helix, β-strand. There exist four fundamental levels of structure, which are called primary, secondary, tertiary, and quaternary structures. Primary structure is the sequence of covalently linked amino acid residues. Secondary structure is the local conformation. The combination of secondary structures is called tertiary structure. Also, quaternary structure is the as-

sociate of two or more tertiary structures. Secondary structure or higher level structures fold with interaction between primary structures. For these reasons, many different proteins could exist in nature. However, we know from previous studies that three-dimensional structures are strongly maintained rather than their primary structures in protein evolution [Chothia 86, Rozwarski 94, Tsukihara 82]. That is, although the amino acids were mutated, the backbone structure of protein would be conserved. In other words, we can predict evolutionary similarities in sets of proteins, or the most functionally important sites, through the comparison of these structures.

The sequence alignment is a basic method of detecting phylogenetic relationships in molecular biology. Since dynamic programming approach (DP-matching) was proposed [Needleman 70], many different algorithms have been developed and utilized [Alexandrov 92, Altschul 90, Chellapilla 99, Gotoh 82, Taylor 87, Zhu 98]. Two fundamental purposes are achieved by the sequence alignment, i.e. an estimate of the evolutionary relation and a recognition of the functional sites in those sequences.

It is limited to the similarities of primary structure. On the other hands, a systematic comparison of proteins as a three-dimensional structure is called the structural alignment. The results of the structural alignment are extended to structure similarities of high levels, and hence one of the most important techniques of all. Thus a number of effective and rapid algorithms have been proposed [Akutsu 95, Gerstein 96, Holm 93, Madej 95, Rossmann 76, Taylor 89]. In the most existing algorithms, the sequence alignment methods are based on and extend them. They are also adopted the root mean square deviation (RMSd), the Monte Carlo method, distance matrices, and so on. However, it is most likely that they have two problems. Firstly, they are all local searches, i.e. the possible transfer positions and orientations can be ignored in the existing methods.

Secondly, the results obtained by these methods are evaluated for average of square distances between equivalent Cα atoms. In consequence, the obtained alignments could be failed in biological active sites.

This thesis proposes a novel algorithm, called GSA : *Genetic Structural Alignment*, to the structural alignment to detect high similarities in two proteins by using Real-coded GA. This new approach can resolve above-mentioned problems, that is, a global search can be realized and put much more importance on conserved and active sites with a effective fitness function.

In section 2 the alignment problem is defined and briefly review what the existing algorithms are. In section 3 GA for GSA is designed as the function optimization problem and also performances are observed with experiments. In section 4 the results of GSA are summarized with experiments. In section 5 it is reported that GSA found a new alignment pattern. Furthermore, an idea for the multiple structural alignment by suggested GSA is described in section 6.

2 Preliminaries

This section first explains the definition of the alignment problem in two proteins, which is called the pairwise alignment, and then surveys the existing algorithms of it briefly.

2.1 Definitions of Alignment Problem

2.1.1 Sequence Alignment

In protein evolution, the mutations as an insertion or deletion are denoted by '–', which is called gap. Suppose the followings are given in this problem:

- A fixed set of characters(corresponding to the 20 amino acids) $\sum$ has no gap, and also a set $\sum'$ has a gap.
- Let s and s' be a member of $\sum$ and $\sum'$, respectively.
- If Seq_k is a sequence, then we denote $\mathrm{Seq}_k = (s_{k1}, s_{k2}, \ldots, s_{kl})$ with s, where k indicates the kth sequence, l is the length of the sequence.

Seq_k contains at least one character. We denote an aligned sequence $\mathrm{Seq}'_k = (s'_{k1}, s'_{k2}, \ldots, s'_{kl})$ with s' obtained from Seq_k. The sequence alignment is defined as a transformation that satisfies the following properties:

Def. 1

$$\begin{aligned} \mathrm{Seq}_1 &= (s_{11}, s_{12}, \ldots, s_{1m}) \\ \mathrm{Seq}_2 &= (s_{21}, s_{22}, \ldots, s_{2n}) \\ \mathrm{Seq}'_1 &= (s'_{11}, s'_{12}, \ldots, s'_{1l}) \\ \mathrm{Seq}'_2 &= (s'_{21}, s'_{22}, \ldots, s'_{2l}) \end{aligned}$$

i. $n \leq l \leq (m+n)$

ii. $\mathrm{Seq}_1 = \mathrm{Seq}'_1 - \mathrm{GAP} \neq \emptyset$

iii. $\mathrm{Seq}_2 = \mathrm{Seq}'_2 - \mathrm{GAP} \neq \emptyset$

iv. If $\mathrm{Seq}'_1 = (\ldots, s_{1i}, \ldots, s_{1j}, \ldots, s_{1l})$, then $i < j < l$

v. If $\mathrm{Seq}'_2 = (\ldots, s_{2i}, \ldots, s_{2j}, \ldots, s_{2l})$, then $i < j < l$,

where '$\mathrm{Seq}'_1 - \mathrm{GAP}$' means removing gaps from sequence. m, n, and l are lengths of sequences($m \leq n$), respectively.

2.1.2 Structural Alignment

In the structural alignment, we use a vector of atom instead of a character. Let ev be a vector of the position of Cα atom. The backbone with Cα atoms is used in most alignment algorithms. When two backbones were compared, they had been superposed in three-dimensional space. Then the sequence alignment would be obtained with these geometric positions.

The Stc_k is given as the kth structure which consists of elements ev_{ki} for $(1 \leq i \leq m)$, where m is the number of Cα atoms at the structure. The structural alignment is defined as a transformation of the first structure, e.g. Stc_1, that satisfy the following properties:

Def. 2

$$\begin{aligned} \mathrm{Stc}_1 &= (ev_{11}, ev_{12}, \ldots, ev_{1m}) \\ \mathrm{Stc}_2 &= (ev_{21}, ev_{22}, \ldots, ev_{2n}) \\ \mathrm{Stc}' &= (ev'_1, ev'_2, \ldots, ev'_m) \end{aligned}$$

i. $\mathrm{Stc}' = \mathcal{R} \times \mathrm{Stc}_1 + t$

ii. $C = \{c_1, c_2, \ldots, c_i, c_{i+1}, \ldots, c_z\}$

iii. If $c_i = \{ev'_j, ev_{2k}\}$, then $c_{i+1} = \{ev'_{j'}, ev_{2k'}\}$ $(j' > j,\ k' > k)$

iv. $1 \leq z \leq m$.

where Stc' is an alignment, $\mathcal{R}$ is a rotation matrix, t is a vector of translation, and C is a set of equivalent atoms. The z became a matched length M.

2.2 Previous Works

2.2.1 Root Mean Square Deviation

In General, the root mean square deviation (RMSd) has been used as a measure of structural similarity in molecular biology. From now on, the RMSd means the average of square distances of all equivalent atoms and the RMS-fitting means the transformation in three-dimensions. For the defining RMSd, we use symbols in Def.2. RMSd is defined by

$$\mathrm{d}_{\mathrm{rmsd}}(C) = \min_{T} \sqrt{\frac{1}{z} \sum_{i=1}^{z} \|T(c_{i1}) - c_{i2}\|^2}, \qquad (1)$$

where the minimum is taken from all transformation T, and T is denoted by $\mathcal{R} \times c_{i1} + t$. c_{i1} and c_{i2} indicate ev'_j and ev_{2k} in Def.2(iii), respectively. Since $t = \mathbf{O}$($\mathbf{O}$ is the zero vector), for instance, $\mathcal{R}$ can be computed in $O(z)$ times as the following:

$$\mathcal{R} = (A^t A^{\frac{1}{2}}) A^{-1}. \qquad (2)$$

Here A is a matrix with $A_{ij} = \sum_{k=1}^{z} (c_{k1})_i (c_{k2})_j$, A^t is the translation matrix, and $A^{\frac{1}{2}} \times A^{\frac{1}{2}} = A$, A^{-1} is the inverse matrix. Although a set C must be established before performing the equation(1), this problem is avoided with the dynamic programming.

2.2.2 Dynamic Programming

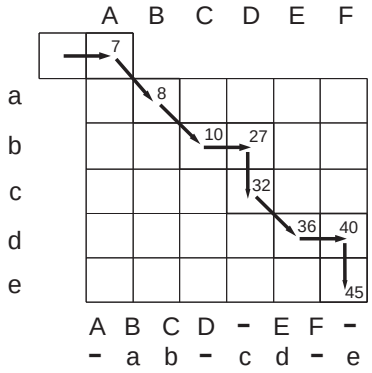


Figure 2: Dynamic Programming

The dynamic programming is a general method used to find the optimal alignment. The basic principle for alignment is either to maximize the number of M between the two proteins or to minimize the number of mismatched atom pairs, i.e. the number of gaps. Figure 2 shows the alignment of two short sequences. The arrows in Figure 2 become the maximum-match pathway. In this case,

pathway consists of three diagonal steps and five horizontal or vertical steps. The diagonal step indicates where the characters matched in two sequences. Also, the horizontal and vertical arrows signify insertions of gaps. The gap penalty factor could be a function of the size and direction of the gap. As such, the algorithm is defined by the following recurrence equation:

$$\begin{aligned} S_{ij} &= Max\{ \\ &\quad S_{i-1,j-1} + D_{ij}; \\ &\quad S_{i,j-1} + g; \\ &\quad S_{i-1,j} + g; \\ &\}, \end{aligned} \tag{3}$$

where S is any element in the matrix of dynamic programming, D is a score matrix and g is a gap penalty. However, This is not the case for the sequence alignment. This D can be converted into a similarity matrix s in the structural alignment by application of the following formula:

$$s_{ij} = f(dist_{ij}). \tag{4}$$

Where s_{ij} and $dist_{ij}$ are similarity and distance between ith atom in the first structure and jth atom in the second one. In the dynamic programming, it requires memory space for a k-dimensional array and calculation time in the k-th power of the number of atoms, where k is the number of proteins.

2.2.3 Existing Algorithms

Recently there has been an explosion of methods for the structural alignment. Rossmann,M.G. et al. proposed an iterative improvement method by using RMSd. Taylor,W.R. and co-works developed the SSAP(sequential structure alignment program), where are sequence alignment technique is applied. Gerstein,M. et al. are studying an iterative dynamic programming, it includes the Monte Carlo method(YSAS,yale structure alignment server). Holm,L. et al. have been proposed a new algorithm, which is based on the combination of distance matrix and Monte Carlo method. Akutsu,T. applied bipartite graph matching method, which is called Stralign. Madej,T. et al. proposed a new algorithm by a comparison of pairs of secondary structure elements (SSE's). These existing algorithms provide the most important databases of proteins for biologists, e.g. CATH (class architecture topology homology), SCOP (structural classification of proteins), DALI (distance matrix alignment), etc.

These algorithms fall into two categories, the iterative superposition and the Monte Carlo optimization. In the iterative superposition algorithms, DP-matching technique is used to identify the best align atom pairs, and RMS-fitting is adopted to rotation and translation. Although they are very fast, the result is aligned roughly when the complexity of structures is increased[Gibrat 96]. This problem can be caused by their local search that ignored variation of rotation angles and transfer positions. The Monte Carlo algorithms iteratively explore a series of shifts in the alignment of each fragments and extension by addition of new aligned atom pairs. They are also adopted DP-matching and RMSd.

However, RMSd is unclear that either is better than the other. Furthermore, the criterion by RMSd will failed to align functional sites of proteins. That is, RMS-fitting rotates the structure in average of spatial positions.

3 Genetic Structural Alignment(GSA)

In this section, a novel alignment algorithm by the name of GSA is proposed with experimental results. The 'Genetic' in GSA means that is greatly inspired by genetic phenomena in nature.

3.1 Approach

Real-coded GA is adopted as the core of GSA algorithm. GA is fit for the structural alignment in several points. First, GA can compare structures in possible spatial positions, i.e. a global search can be realized by populations. Second, although the complexity of structure was increased, with GA a higher precision can be obtained. Third, flexible object function can be designed. In addition, we can construct a fitness function which includes means of atom pairs.

In this thesis, the structural alignment is addressed as a function optimization problem to maximize the fitness function f. A chromosome is designed by six-dimensional real numbers. Crossover operator and generation alternation model is adopted UNDX (unimodal distribution crossover)[Ono 97] and MGG (minimal generation gap)[Satoh 96, Yamamura 97], respectively.

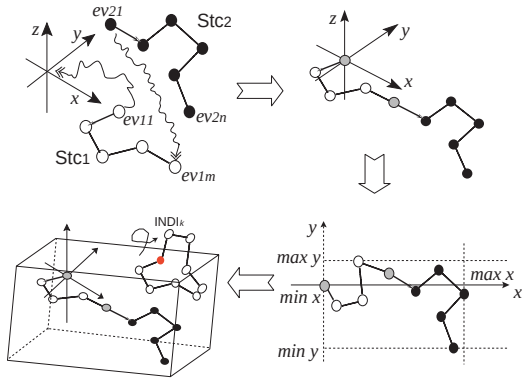


Figure 3: Generate Method for Initial Population

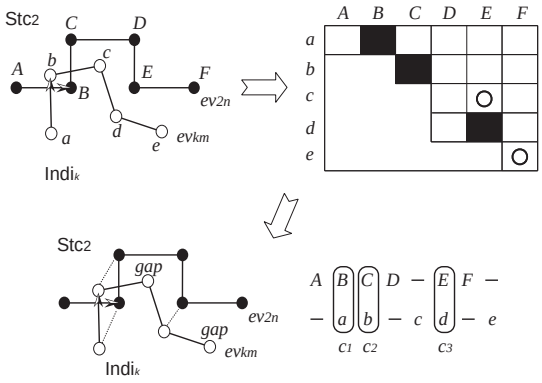


Figure 4: Estimate of Equivalent Atoms

3.2 Design for GA

3.2.1 Encoding

When we considered that we would explore in three-dimensional geometric spaces, we can represent an individual INDI_k by six parameters real numbers, i.e. three rotation angles and three-dimensional translation vector of the first atom. Figure 1 shows a generating method for ev_{k1} in an initial population. Stc_1 is moved to the origin in the first step. Next step, ev_{1m} and ev_{21} are connected, and then we can determine a possible individuals space, presented as a box($min(x, y, z)$, $max(x, y, z)$) in Figure 3. Then an initial population is generated in this box, while an individual INDI_k is rotated in which a fixed ev_{k1} is the origin. We can represent possible positions that the existing methods have never considered. A chromosome in an initial population could be encoded with α, β, γ, and x, y, z as defined by:

$$\begin{aligned} -\pi \leq \quad & (\alpha, \beta, \gamma) \quad \leq +\pi \\ min\ x \leq \quad & x \quad \leq max\ x \\ min\ y \leq \quad & y \quad \leq max\ y \\ min\ z \leq \quad & z \quad \leq max\ z. \end{aligned} \tag{5}$$

3.2.2 Crossover and Generation Alternation Model

We employed UNDX as a crossover operator, it is a powerful crossover method in characteristics preservation. Using normal distribution UNDX generates six-dimensional real numbers for two children in a determined area with three parents. In regard to rotation angles, we use complementary angles because 2π presents an equivalent position in geometric spaces. The complementary angles are defined by a narrow angle between two parents who have a difference angles more than π.And the MGG is adopted as a generation alternation model.

3.2.3 Mutation

Two individuals, who have survived in MGG, are likely to be mutated in a mutation probability. When we assume that INDI_k is mutated, the mutation procedure is described as follows:

- randomly select an atom ev_{2r} in Stc_2 where r is a position of selected atom.
- ev_{k1} is superimposed on ev_{2r} by force.
- INDI_k is rotated with $\alpha = \pi, \beta = 0, \gamma = 0$.

3.2.4 Fitness Function

The estimate of equivalent atoms is required for evaluation of individuals. As shown in Figure 4, we compute distances between each atom in the INDI_k and every atom in the Stc_2(exception of determined atoms). When we found the nearest atom pair(i,j), we check their distance $dist_{ij}$. If $dist_{ij}^2 \leq \delta$ then, added to a set of equivalent pairs, illustrated as ■ in Figure 4. If $dist_{ij}^2 > \delta$ then, it becomes a gap, illustrated as ◯ in Figure 4. Where δ is a constant for the gap.

For a member $c_i(a, b)$ of $C = \{c_1, c_2, \ldots, c_z\}$ (see Def.2), let $dist_{c_i}$ be $\|a - b\|$. And $g = m - z$ for a gap penalty, where m is the length of INDI_k. Hence we define the fitness function f as follows:

$$\begin{aligned} s &= \sum_{i=1}^{z} \exp(\epsilon \times dist_{c_i}^2) \\ f &= \frac{s + 1.0}{g + 1.0}\ , \end{aligned} \tag{6}$$

where $\epsilon (< 0)$ is a constant for the similarity. In an ideal α-helix, each amino acid residue keeps 1.54Å(1Å=0.1nm), i.e. the distance of Cα–Cα

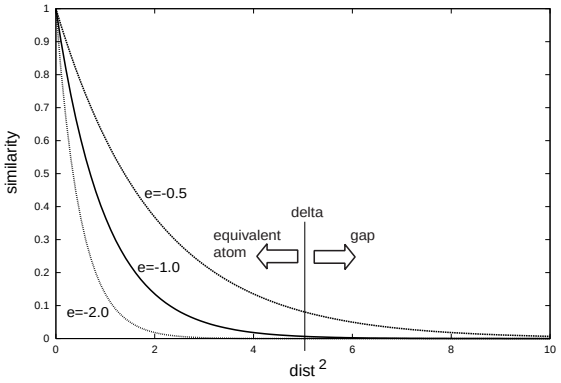


Figure 5: $\exp(\epsilon \times dist^2_{c_i})$

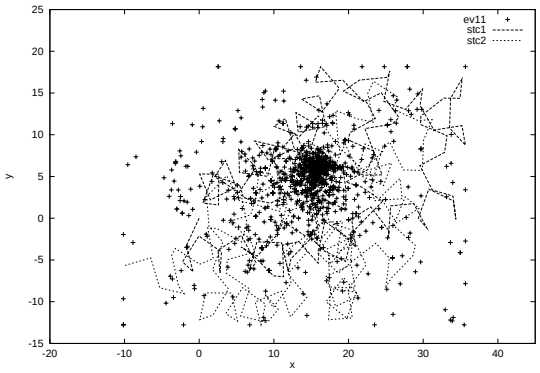


Figure 6: Distribution of All ev_{11}(1ecd vs. 1mbs)

bond. In contrast, each residue in β-strand accounts for 3.2–3.4Å. Thus, δ is taken 5.0(approximately 2.24^2), which is a midway value in the distances of Cα–Cα bonds. The ϵ has an important effect in this fitness function f. That is, the distance between a pair of atoms is reflected their similarity by ϵ (see Figure 5).

The conserved and active sites in similar proteins should locate in a close geometric position. The fitness function f is designed to realize a special emphasis on such the important sites. We can find the other functions for this approach. GA has adaptability, one of features, to any function. This flexibility in GA is an advantage over the other methods.

In section 4 we determine ϵ with experiments. The ϵ resembles the 'Enzymes', which are extremely effective as biological catalysts. In the fitness function f, ϵ might act on closely located atom pairs(ϵ is an initial letter of the 'Enzyme').

3.3 Experimental Results

In order to observe the behavior of designed GA, the comparison with two sets of proteins are performed as follows.

3.3.1 Materials

For the parameters in the experiments we assumed that the population size was 50 and the number of applying crossovers was 100, the generation size was 3000. α and β in the UNDX were set to 0.5 and 0.3, respectively. These parameters were fixed in this thesis. The mutation operator was not used in this experiment in order to confirm the effectiveness of the proposed encoding and the crossover. For the time being, ϵ is set to -0.5 arbitrarily.

The test case for this experiment was prepared from PDB [1] [Berman 00] as shown in Table 1. Two sets of proteins are used, myoglobin(PDB code:1mbs) and hemoglobin(PDB code:1ecd), which have a high structural similarity, and ubiquitin(PDB code:1ubq) and ferredoxin(PDB code:4fxc) with low one, see attached papers pp18-19.

All experiments have been done using SUN ULTRA SPARC–296Mhz workstation, and GA was implemented C language. The response time for the pair of 1mbs and 1ecd is approximately 2200 seconds.

3.3.2 Results

Encoding and Crossover, Generation Alternation Model

First of all, 1ecd and 1mbs are used as the test case. Stc_1 and Stc_2 are determined by length of structures, 1ecd and 1mbs consist of 136 and 153 residues, respectively. 1ecd became Stc_1 because the length of 1ecd is shorter than 1mbs. Figure 6 shows all the generated individuals by GA, as dotted with ev_{11}. The optimal superimposed position is known from their conformations. This position became the focus of searches as shown in Figure 6.

To observe the rotation angles, a comparative experiment is performed where two children are generated with complementary angles and non-complementary angles. These behaviors can be represented with the positions of ev_{12}, because they are rotated in a fixed ev_{11} as the origin. Figure 7(a) and Figure 7(b) show the distribution of all individuals generated with two different angle systems. Although two systems had no peculiari-

[1]PDB : Protein Data Bank(http://www.rcsb.org/pdb/) Proteins are identified by their 4 character PDB codes.

ties, their fitness curves, as shown in Figure 8(a) and Figure 8(b), indicated a gap in initial generations.

When GA is performed with high structural similarity proteins, this gap deserved to pay no attention. However, the extend of this gap is increased when we compare with the low similarity proteins(see Figure 8(b)). In other words, the complement angle system is possible to effectively search in limitative generation sizes. Figure 9 shows the results of 10 trials with the complementary angles. The MGG is deduced from these consequences that it succeeded in the maintaining a diversity of population.

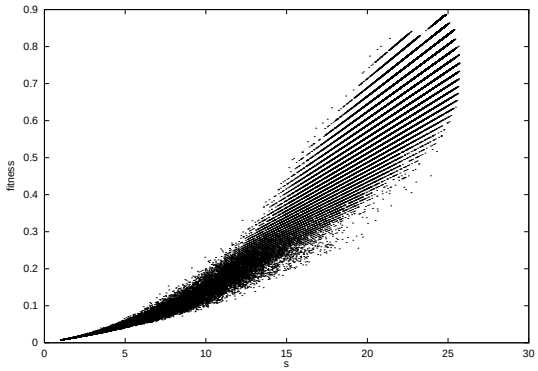


Figure 10: Distribution of All Individuals Fitness (1ecd vs. 1mbs)

Fitness Function

In Figure 8 and Figure 9, the average of fitness fluctuates where the nearest of best fitness curves in the letter of generations. The number of gaps g will increases by crossover between great closely individuals, because the distance of equivalent atoms can be taken longer than δ. For this reason, the average of fitness never converged into the best fitness with this degree of generation sizes. However, it gives a chance of improvement to a best individual, from the viewpoint of evolutional computation. The interaction between sum of similarities s and the number of gaps g in the fitness function is represented in Figure 10, the generation is proportional x axis. As shown in Figure 10, there are several lines, called *evolution roots*. The broken evolution roots are local minima and the efficiently UNDX are reconfirmed by these roots. Furthermore, there are no great variations of g in the latter of generation, and the gap penalty is stricter with a fitness of individuals there.

Moreover, the exponential curves indicate the effects of s with the exponential function. The s is a powerful function that the closed pairs is conserved into individuals. In contrast, the existing methods by RMSd attach importance to the average of square distances. Due to such as transformation by RMSd, the nearest pair in the biological conserved and active sites becomes a senseless atom for the protein.

3.4 Summary of Experiments

In this section, GSA was proposed with experimental results. The structural alignment problem was considered as a function optimization problem. Then, the new encoding method and the fitness function f were applied for GSA, and they have been confirmed by investigation of experiments.

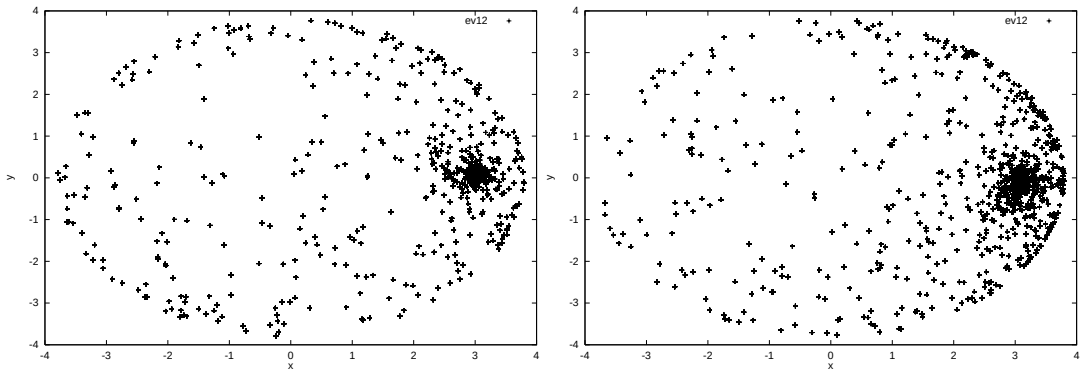
4 Experimental Results of GSA

The ϵ determined has not yet been fully considered. In this section, the reasonable ϵ is found by experimental results. Then it is compared that the proposed GSA in section 3 and the other methods. The other test proteins are prepared not to be partial in conformation of proteins. Furthermore, the results of experiments and response times are shown in this section.

4.1 Materials

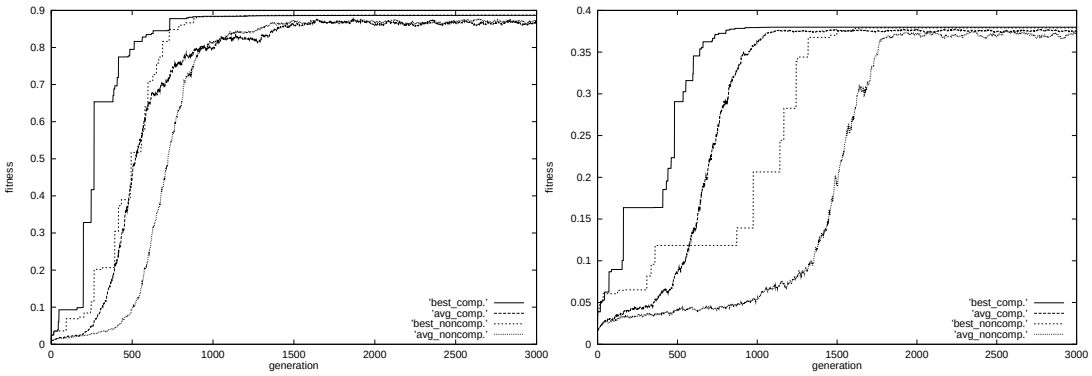
The globin proteins in Table 1 are typical of the globular proteins. Their functions are known completely[Moran 94]; myoglobin was the first protein to have its tertiary structure determined.

Oxygen binds to the heme prosthetic group, in the hemoglobin and myoglobin. However, hemoglobin, which transports oxygen in the blood of vertebrates, is a tetramer: myoglobin, which stores oxygen and facilitates its diffusion within muscle, is a monomer. Myoglobin accounts for about 8% of the total protein in the muscles of diving mammals, such as seals and whales can store large amounts of oxygen by myoglobin. Leg hemoglobin, a monomeric protein found in leguminous plants, has a structure much the same as mammals myoglobin. The interior of globin molecule is composed almost entirely of nonpolar residues, with the exception of two Histidines(H: One-letter code) , the 64th and 93rd residues in the case of myoglobin. These exceptional residues



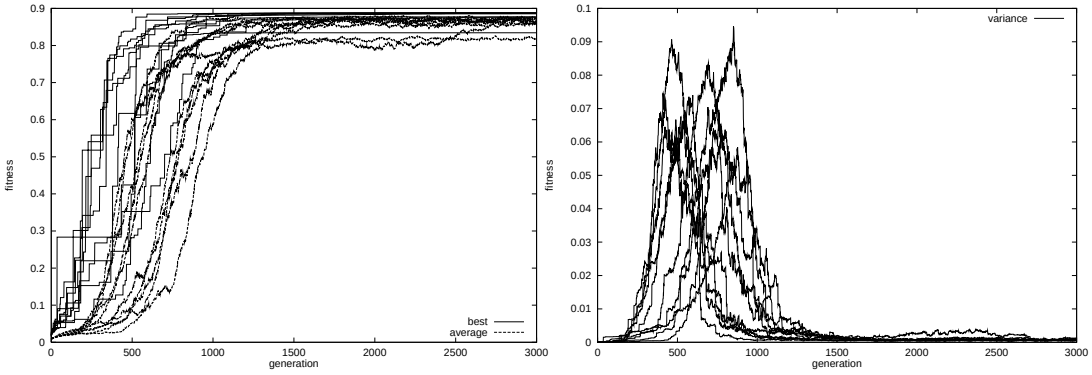
(a) Complementary Angles (b) Non-Complementary Angles

Figure 7: Distribution of All ev_{12}(1ecd vs. 1mbs)



(a) 1ecd vs. 1mbs (b) 1ubq vs. 4fxc

Figure 8: Performance Curves in Two Angle Systems



(a) Best and Average (b) Variance

Figure 9: Performance Curves in 10 trials (1ecd vs. 1mbs)

Table 1: Test Proteins:Globin Proteins And Low Similarities Proteins

PDB code	Residues	Source	Compound	Classification
1ecd	136	Chironomous Thummi Thummi	Hemoglobin	Oxygen Transport
4hhbA	141	Human(Homo Sapiens)	Hemoglobin	Oxygen Transport
1bijA	141	Homo Sapiens	Hemoglobin	Oxygen Transport
1rvwA	141	Homo Sapiens	Hemoglobin	Oxygen Transport
2hheA	141	Human(Homo Sapiens)	Hemoglobin	Oxygen Transport
1hhoA	141	Human(Homo Sapiens)	Hemoglobin	Oxygen Transport
1fdhA	141	Human Fetus(Homo Sapiens)	Hemoglobin	Oxygen Transport
1habC	141	Homo Sapiens	Hemoglobin	Oxygen Transport
1bz0C	141	Homo Sapiens	Hemoglobin	Oxygen Transport/Storage
1cohC	141	Human(Homo sapiens)	Hemoglobin	Oxygen Transport
1babA	143	Human(Homo Sapiens)	Hemoglobin	Oxygen Transport
1hgbC	146	Human (Homo Sapiens)	Hemoglobin	Oxygen Transport
1mbs	153	Common Seal(Phoca Vitulina)	Myoglobin	Oxygen Transport
5mbn	153	Sperm Whale(Physeter catodon)	Myoglobin	Oxygen Storage
1hsy	153	Horse(Equus Caballus)	Myoglobin	Oxygen Transport
1hrm	153	Horse(Equus Caballus)	Myoglobin	Oxygen Transport
1wla	153	Equus Caballus	Myoglobin	Oxygen Transport
1xch	153	Equus Caballus	Myoglobin	Oxygen Transport
1yma	153	Horse(Equus Caballus)	Myoglobin	Oxygen Transport
2gdm	153	Lupinus Luteus L.	Leghemoglobin	Oxygen Transport
1lh3	153	Yellow Lupin(Lupinus Luteus L)	Leghemoglobin	Oxygen Transport
1lh1	153	Yellow Lupin(Lupinus Luteus L)	Leghemoglobin	Oxygen Transport
2lh1	153	Yellow Lupin(Lupinus Luteus L)	Leghemoglobin	Oxygen Transport
1gdj	153	Yellow Lupin(Lupinus Luteus L)	Leghemoglobin	Oxygen Transport
1ubq	76	Human (Homo sapiens)	Ubiquitin	Chromosomal Protein
4fxc	98	Spirulina Platensis	Ferredoxin	Electron Transport

(Capital Character 'A' in PDB code : Chain 'A')

are important functional units of globin proteins, i.e. the heme plane is flanked by the 64th Histidine and the 93rd one. Moreover, the 43rd Phenylalanine(F) and the 68th Valine(V) contribute to the hydrophobic environment of the oxygen-binding site.

GSA is compared with YSAS[2] and Stralign[3], which are an iterative improvement methods. YSAS includes a combination of the DP-matching and the Monte Carlo method. It provides real-time alignment on the Internet. On the other hand, using a bipartite graph matching technique instead of DP-matching in Stralign(δ = 5.0, 15 fragment length and 100 maximum test times).

Here, the new symbols described in Table 2, *accuracy symbols*, represent the aligned accuracies. These symbols can denote close atom pairs, particularly in $*$ and | ,#. Their distances set on about 0.7Å steps. It is extremely interesting how many atom pairs are there within about 1.22Å.

[2]http://bioinfo.mbb.yale.edu/Align/

[3]http://www.hgc.ims.u-tokyo.ac.jp/service/tooldoc/stralign/intro.html

Table 2: Accuracy Symbols

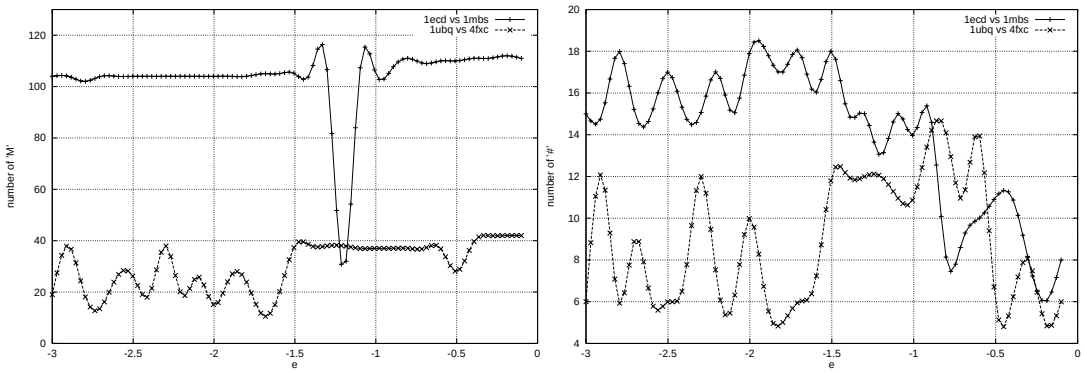
no.	symbol	$dist_{ij}^2$
1	empty	$> \delta(gap)$
2	$\cdot$	$>$ 1.5 and different character
3	:	$>$ 1.5 and same character
4	$*$	$\leq$ 1.5
5	|	$\leq$ 1.0
6	#	$\leq$ 0.5

Thus these accuracy symbols can show the biological active sites.

The mutation operator described in section 3 is used with 0.01 mutation probability.

4.2 Estimate ϵ

The ϵ in fitness function f can quicken detection of higher numbers of the close atom pairs. In other words, the involvement with distance and similarities of two atoms is determined by ϵ. As shown



(a) The Number of 'M' (b) The Number of '#'

Figure 11: Typical Reactions of ϵ

in Figure 5, if a weaker value is taken (towards nearly zero), it might never reflect close pair of atoms to the fitness function f. It might evaluate only nearer atom pairs and these atoms are conserved in individuals when ϵ was a stronger value (negative direction). Figure 11 shows typical ϵ reactions against aligned results, i.e. the number of M in Figure 11(a) and the number of # in Figure 11(b). In the case of easy structural alignment (1ecd vs. 1mbs), we can obtain a static M with the exception of one ϵ. Also, we can observe the proportional number of # with force of ϵ. However, they fluctuated in stronger ϵ values when the complexity of structure(1ubq vs. 4fxc) was increased.

The GSA has purposes that detect higher numbers of # and realize the global search independent to the specific structures. Therefore, we have to take the highest # points in two curves, i.e. approximately -2.0 and -0.8 in Figure 11(b). However, it is taken appropriate -0.8 as the ϵ in consideration of the number of M.

4.3 Results

4.3.1 Response Times

The experiments used 15 pairs of globular proteins; globin family and 10 pairs of complex conformations proteins. All experiments ran on a SUN ULTRA SPARC-296Mhz workstation. Figure 12 shows the response times in these experiments, x axis is averages of lengths of two proteins and y axis is the response times.

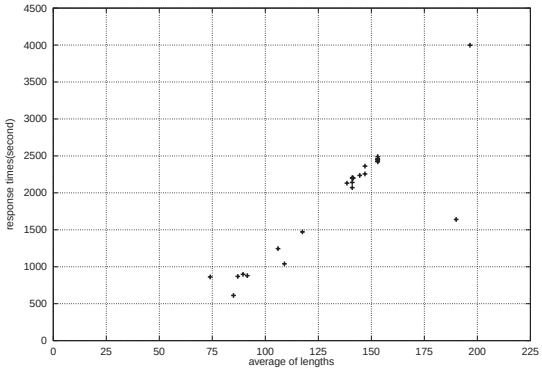


Figure 12: Response Times

4.3.2 Easy Structural Alignment

In this experiment, the globular proteins (Table 1) are compared by YSAS, Stralign, and GSA. Table 3 shows the results of structural alignment. In addition, the accuracy of results can represent with distances of equivalent atoms. Figure 13 shows that their aligned patterns are similar to the biological conserved sites of myoglobin(1mbs). the 64th and the 93rd Histidines, the heme is flanked, have # symbols in three different methods. It was confirmed that the 87th Histidine residue of hemoglobin (1ecd), the heme is bonded, was also closely connected with myoglobin, the figure is omitted.

4.3.3 Harder Structural Alignment

Over three-quarters of residues of a globular protein are in the α-helix. Thus, RMS-fitting method can align well. New test proteins are prepared not to be partial in conformations. They consist of α-helices and β-sheets combinations or only β-sheets. The results of structural alignment are described in Table 4.

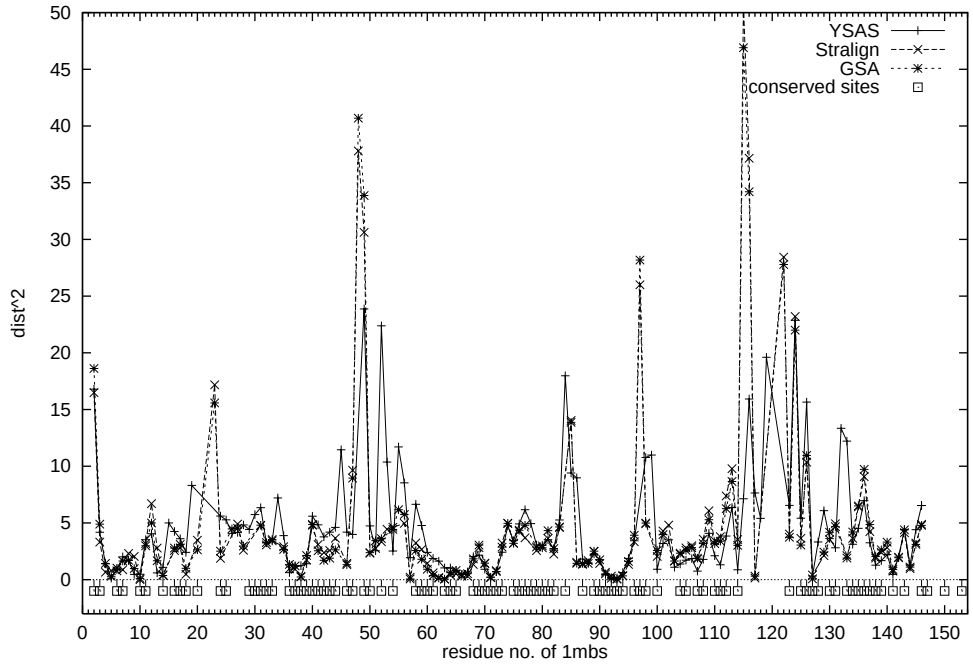


Figure 13: Distances of Equivalent Atoms and Conserved Sites

4.4 Discussion

As shown in Table 3, there are no wide differences between three different methods. We cannot distinguish that either is higher in accuracy with the number of # and M or the sum of '$* + j + \#$'. It is possible to find a good method in Table 4. In other words, using the RMSd and DP-matching methods are adequate to compare between the simplest conformation proteins. However, their accuracies deteriorate into 'not make any importance to functions of protein' when they are compared with complex structures. The number of ' · ' and ' : ' of YSAS and Stralign in Table 4 are higher than GSA. These symbols are senseless to the biological important sites, because they imply distant positions of equivalent atoms. Thus, it is appreciated that the existing methods fail to catch the nearest pair of atoms. In contrast, GSA makes a success of detecting high structural similarity pairs with independence on the conformations. Furthermore, GSA found a novel structural alignment in Ca^{2+}-binding proteins, complex conformation proteins.

5 Ca^{2+}-binding Proteins

3ICB and 5CPV(3ICB, intestinal Ca^{2+}-binding protein; 5CPV, carp parvalbumin) have binding sites for two calcium ions. The 14th to 27th and the 54th to 65th residues(EF-hand) in 3ICB are Ca^{2+} binding sites, the 51st to 62nd (CD-hand) and the 90th to 101st residues(EF-hand) in 5CPV, respectively. There exist EF-hands common to two proteins, that is, the results of structural alignment must detect the EF-hand active sites.

As shown in Figure 14, the alignment results of YSAS and Stralign indicate that there are many conservational residues in calcium-binding sites. However, GSA obtained an aligned pattern different from the other methods. It indicates that the EF-hand sites are similar in both structures much more than YSAS and Stralign. The similarities of CD-hand are also maintained. The first atom Lysine, senseless atom for Ca^{2+}-binding proteins, corresponds to different atom from YSAS and Stralign. We can deduce from this fact that GSA efficiently generates and rotates a structure as a rigid body. The positions of Lysine and results of structural alignments are shown in Figure 15.

6 Multiple GSA

It is generally believed that the prediction of an ancestral protein from given sets of proteins can find extremely significant biological knowledge. The effective and utilize methods have not been developed because of harder degree of difficulties.

The GSA can treat given proteins as a population, i.e. considering a protein as an indi-

Table 3: Results of Easy Structural Alignment

pair	method	RMSd	M	·	:	∗	|	#
1babA_1bz0C	YSAS	4.43306	142	0	0	1	2	139
	Stralign	0.25993	141	0	0	1	1	139
	GSA	0.26556	141	0	0	1	4	136
1fdhA_1cohC	YSAS	0.32632	142	0	1	0	4	137
	Stralign	0.31406	141	0	1	0	4	136
	GSA	0.31499	141	0	1	0	4	136
1rvwA_1habC	YSAS	0.40772	142	0	0	2	4	136
	Stralign	0.37769	141	0	0	2	4	135
	GSA	0.37972	141	0	0	2	3	136
2hheA_1hgbC	YSAS	0.35693	142	0	0	0	6	136
	Stralign	0.33632	141	0	0	0	5	136
	GSA	0.33787	141	0	0	0	6	135
1bijA_1hhoA	YSAS	0.57296	124	0	0	2	14	108
	Stralign	0.82160	136	3	7	5	12	109
	GSA	0.61952	134	0	7	6	13	108
4hhbA_5mbn	YSAS	7.33431	136	44	15	12	26	39
	Stralign	1.44108	140	43	16	16	27	38
	GSA	1.28604	130	36	14	20	21	39
1mbs_5mbn	YSAS	1.40223	144	12	48	23	34	27
	Stralign	1.48817	149	15	53	15	37	29
	GSA	1.25917	138	18	41	21	31	27
4hhbA_1mbs	YSAS	6.92031	135	61	21	11	23	19
	Stralign	1.74165	136	62	20	11	25	18
	GSA	1.49387	122	53	19	18	13	19
1hsy_1lh3	YSAS	2.31578	140	78	9	19	18	16
	Stralign	2.26041	138	88	14	9	18	9
	GSA	1.40499	103	48	6	15	17	17
1yma_1lh1	YSAS	2.38389	140	80	10	20	17	13
	Stralign	2.26061	138	91	14	7	16	10
	GSA	1.43810	105	48	8	14	16	19
1wla_2lh1	YSAS	2.67754	142	87	15	13	15	12
	Stralign	2.25637	138	87	15	9	18	9
	GSA	1.45423	111	52	7	10	21	21
1ecd_1mbs	YSAS	2.16801	137	83	20	13	11	10
	Stralign	1.96364	133	81	19	13	14	6
	GSA	1.57554	111	59	17	11	16	8
1hrm_2gdm	YSAS	2.17660	138	71	13	16	28	10
	Stralign	2.23722	140	103	15	5	11	6
	GSA	1.39203	108	47	6	12	22	21
1ecd_4hhbA	YSAS	7.78449	125	79	13	8	15	10
	Stralign	2.22607	129	92	20	8	3	6
	GSA	1.49441	96	50	3	8	17	18
1xch_1gdj	YSAS	3.03139	148	95	16	16	15	6
	Stralign	2.18654	140	81	14	17	15	13
	GSA	1.40113	112	46	5	12	34	15

Stralign:Fragment Length= 15, $\delta = 5.0$, Maximum Test Times= 100
GSA:Population= 50, Crossover Times= 100, Generation= 3000,
Mutation Prob.= 0.01, UNDX($\alpha = 0.5,\ \beta = 0.3$), $\delta = 5.0$,$\epsilon = -0.8$

Table 4: Results of Harder Structural Alignment

pair	method	RMSd	M	·	:	∗	|	#
1b11_1idaA	YSAS	0.98733	91	12	1	5	25	48
(113_99)	Stralign	2.09623	98	63	10	9	10	6
	GSA	0.97205	93	19	1	5	10	58
1buhB_1cksB	YSAS	0.47447	52	0	0	0	7	45
(70_78)	Stralign	2.60051	25	20	1	2	2	0
	GSA	0.63509	56	0	3	2	6	45
3icb_5cpv	YSAS	4.51080	57	23	12	4	10	8
(75_109)	Stralign	1.77822	58	24	11	6	8	9
	GSA	1.30172	46	13	6	4	5	18
1ubq_4fxc	YSAS	2.80691	65	47	2	7	6	3
(76_98)	Stralign	2.81187	64	46	4	5	6	3
	GSA	1.25622	37	14	2	5	2	14
2pkaX_1hvrB	YSAS	3.74044	53	44	5	1	2	1
(80_99)	Stralign	2.44886	30	24	5	0	0	1
	GSA	0.96766	22	5	0	2	3	12
1tig_1xvaA	YSAS	2.66371	77	54	5	10	5	3
(88_292)	Stralign	2.44049	75	54	6	5	6	4
	GSA	1.45020	38	19	1	2	6	10
1vscB_1bp3B	YSAS	8.08298	165	151	13	1	0	0
(196_197)	Stralign	3.01190	87	74	2	6	5	0
	GSA	1.01278	25	5	0	3	7	10
1ytfC_7rsa	YSAS	5.45598	40	33	3	3	1	0
(46_124)	Stralign	2.87897	38	28	1	5	3	1
	GSA	1.40295	24	9	2	3	1	9
1az5_1ang	YSAS	4.27268	65	58	2	4	1	0
(95_123)	Stralign	2.97083	55	52	1	0	2	0
	GSA	1.21145	16	6	0	2	1	7
1kul_1ttaB	YSAS	3.19986	70	61	6	2	1	0
(108_127)	Stralign	2.91203	70	52	8	2	6	2
	GSA	1.04946	17	4	0	5	2	6

Stralign:Fragment Length= 15, $\delta = 5.0$, Maximum Test Times= 100
GSA:Population= 50, Crossover Times= 100, Generation= 3000,
Mutation Prob.= 0.01, UNDX($\alpha = 0.5,\ \beta = 0.3$), $\delta = 5.0, \epsilon = -0.8$

vidual. It makes us image an inverse evolution. This inverse evolution means that current proteins evolve into the past (towards to their ancestral protein). For example, Figure 16(attached paper) shows evolution roots in the comparison between T4Glutaredoxin and E.coli. There are stronger local minima and some roots most likely connected with other proteins. If GSA extended to MGSA(*Multiple Genetic Structural Alignment*), it might shows local minima the same as GSA. Although they are assessed as the negative phenomena in the most optimization problems, these local minima are applied as the positive phenomena in MGSA. Figure 17(attached paper) shows an idea for MGSA in the abstract. An initial population evolves inversely into the peak. Then, we can obtain a best individual and some local minima (Figure 17(a)). The best individual can describe the structure of ancestral protein, because the fitness function of GSA is designed that it conserves the emphatic atoms in structures. As such, the phylogenetic tree can be constructed by these results (Figure 17(b)).

However, there exist many problems in this idea. First of all, what is the effective fitness function? How can we represent the inverse evolution? When should we consider to be converged? Therefore, we must consider these problems as one of the further works in this thesis.

7 Conclusion

The structural alignment must be considered to the biological important sites. In this sense, many

```
                                                                    c:CD-hand
                                                                    f:EF-hand
YSAS                                                          ccccccccccc
3icb ------------------------------------KSPEELKGIFEKYAAKEGDPNQL
                                          ::.*..#|.#|... .   |##|
5cpv AFAGVLNDADIAAALEACKAADSFNHKAFFAKVGLTSKSADDVKKAFAIIDQDK--SGFI
     cccc                                                cccc  cccc
     cccc                            fffffffffff
3icb SKEELKLLLQTEFPSLLKGPS-TLDELFEELDKNGDGEVSFEEFQVLVKKISQ
     #..:::::  ..          ..|*.#..**..::::.##||||..:|::
5cpv EEDELKLFLQNFKADARALTDGETKTFLKAGDSDGDGKIGVDEFTALVKA---
     cccc                            fffffffffffff

Stralign                                                      ccccccccccc
3icb ------------------------------------KSPEELKGIFEKYAAKEGDPNQL
                                          ::.|..#|.#|... ..  *###
5cpv AFAGVLNDADIAAALEACKAADSFNHKAFFAKVGLTSKSADDVKKAFAIID-QD-KSGFI
     cccc                                                c cc ccccc
     cccc                            fffffffffffff
3icb SKEELKLLLQ-TE-FPSLLKGPSTLDELFEELDKNGDGEVSFEEFQVLVKKISQ
     #..:::::. | ..          ..**.#..**..::::.##||||..:*:
5cpv EEDELKLF-LQNFKADARALTDGETKTFLKAGDSDGDGKIGVDEFTALVK---A
     cccc                            fffffffffffff

GSA                                                                 cccccc
3icb ------------------------------------KSPEELKGIFE--KYAAKEGD
                                          .   :  :.   .  .
5cpv AFAGVLNDADIAAALEACKAADSFNHKAFFAKVGLTSKSADD--VKKAFAIIDQ--D--K
                                                          cc  c  c
     cccccccc                        fffffffffffff
3icb PNQLSKEELKLLLQT-E-FPSLLKGPSTLDELFEELDKNGDGEVSFEEFQ-VLVKKISQ
     ###|#..|::: .|. .          ..*#*##|######*.##|###* . :
5cpv SGFIEEDELKL-FLQNFKADARALTDGETKTFLKAGDSDGDGKIGVDEFTAL-V---KA
     cccccccc                        fffffffffffff
```

Figure 14: Accuracies of Three Different Methods : Ca^{2+}-binding Protein(3icb vs. 5cpv)

N : N terminus
C : C terminus
k : Lysine

black line : 3icb
gray line : 5cpv

calcium binding 1 : 14 - 27
calcium binding 2 : 54 - 65 (EF-hand)
calcium binding 1 : 51 - 62 (CD-hand)
calcium binding 2 : 90 - 101(EF-hand)

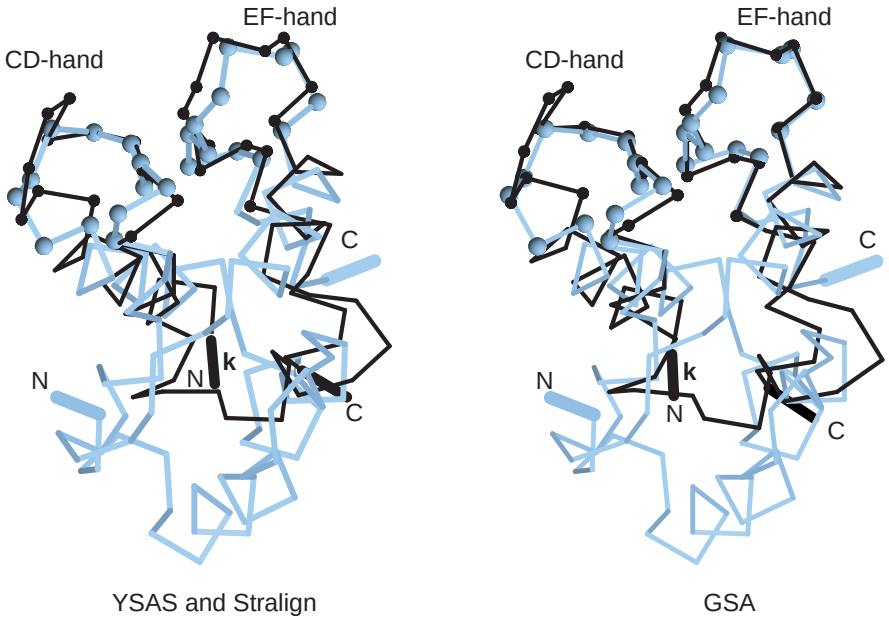


Figure 15: Result of Structural Alignment : Ca^{2+}-binding Protein(3icb vs. 5cpv)

existing methods are inadequate. Using RMSd, RMS-fitting, and DP-matching are extremely depending on the protein conformations. Such the iterative improvement approaches cannot find any novel alignment in complex conformation proteins. For these reasons, the existing methods have essential problems to their algorithms. First, they explore only local spaces by RMS-fitting and DP-matching. Second, RMSd is unclear which is a best result. Third, the most important active sites, i.e. for the protein still survives in nature, cannot be aligned with any combination of these methods.

In this thesis, a novel alignment method GSA was proposed with several experimental results. It is confirmed that GSA could be a global search and fitness function f conserved important sites into the individuals. GSA is compared with YSAS and Stralign, traditional methods. From the results of these tests, we reconfirmed that GSA has the ability to detect high similarities in two proteins. In addition, GSA found a novel alignment pattern within Ca^{2+}-binding proteins.

Nevertheless, the analyses of results in this thesis are insufficient. The parameters in fitness function f have not been fully considered. These problems and the realization of MGSA are further works.

Acknowledgment

The author is grateful to Professor S. Kobayashi, Associate Professor H. Kita and Associate Professor M. Yamamura for useful discussions and helpful advices. He is also indebted to Mr. D.I. Eglin, PRIMUS Co.,Ltd., for amendments of the manuscript. The author gratefully acknowledge the support, encouragement, and patience of his parents.

References

[Akutsu 95] Akutsu, T.: Protein structure alignment using a graph matching technique, in *GIW95*, pp. 1–8 (1995).

[Alexandrov 92] Alexandrov, N. N.: Local multiple alignment by consensus matrix, *CABIOS*, Vol. 8, No. 4, pp. 339–345 (1992).

[Altschul 90] Altschul, S. F., et al.: Basic local alignment search tool, *J.Mol.Biol.*, 215, pp. 403-410 (1990).

[Berman 00] Berman, H. M., et al.: The Protein Data Bank, *Nucleic Acids Research*, 28, pp. 235-242 (2000).

[Chellapilla 99] Chellapilla, K. and Fogel, G. B.: Multiple sequence alignment using evolutionary programming, *CEC99*, Vol. 1, pp. 445–452 (1999).

[Chothia 86] Chothia, C. and Lesk, A. M.: The relation between the divergence of sequence and structure in proteins, *EMBO J.*, Vol. 5, No. 4, pp. 823–826 (1986).

[Gerstein 96] Gerstein, M. and Levitt, M.: Using iterative dynamic programming to obtain accurate pairwise and multiple alignments of protein structures, in *ISMB96*, pp. 59–67 (1996).

[Gibrat 96] Gibrat, J. F., et al.: Surprising similarities in structure comparison, *Current Opinion in Structure Biology*, 6, pp. 377-385 (1996).

[Gotoh 82] Gotoh, O.: An improved algorithm for matching biological sequences, *J.Mol.Biol.*, 162, pp. 705-708 (1982).

[Holm 93] Holm, L. and Sander, C.: Protein structure comparison by alignment of distance matrices, *J.Mol.Biol.*, 233, pp. 123-138 (1993).

[Madej 95] Madej, T., et al.: Threading a database of protein cores, *Proteins*, 23, pp. 356-369 (1995).

[Moran 94] Moran, L. A., et al.: *Biochemistry*, Neil Patterson Publisher/Prentice-Hall (1994).

[Needleman 70] Needleman, S. B. and Wunsch, C. D.: A general method applicable to the search for similarities in the amino acid sequence of two proteins, *J.Mol.Biol.*, 48, pp. 443-453 (1970).

[Ono 97] Ono, I. and Kobayashi, S.: A Real-coded Genetic Algorithm for Function Optimization Using Unimodal Normal Distribution Crossover, in *Proc.7th Int.Conf. Genetic Algorithms*, pp. 246–253 (1997).

[Rossmann 76] Rossmann, M. G. and Argos, P.: Exploring structural homology of proteins, *J.Mol.Biol.*, 105, pp. 75-95 (1976).

[Rozwarski 94] Rozwarski, D. A., et al.: Structural comparisons among the short-chain helical cytokines, *Structure*, Vol. 2, No. 3, pp. 159–173 (1994).

[Satoh 96] Satoh, H., et al.: Minimal Generation Gap Model for GAs Considering Both Exploration and Exploitation, in *Proc. IIZUKA96*, pp. 494–497 (1996).

[Taylor 87] Taylor, W. R.: Multiple sequence alignment by a pairwise algorithm, *CABIOS*, Vol. 3, No. 2, pp. 81–87 (1987).

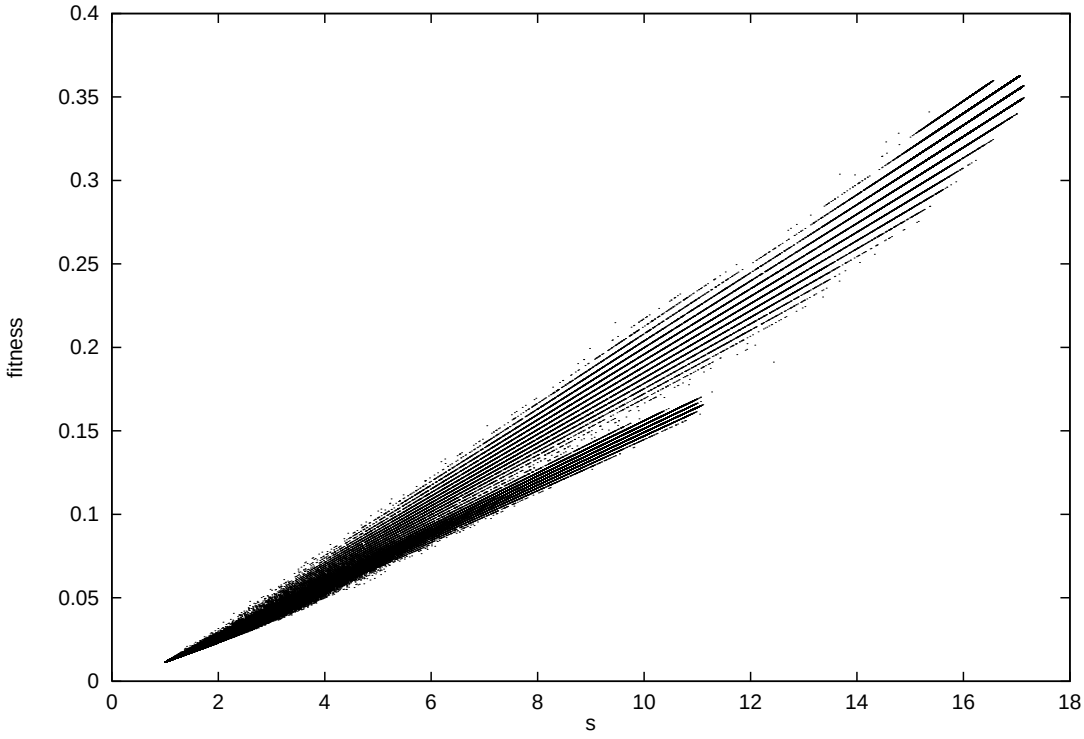
[Taylor 89] Taylor, W. R. and Orengo, C. A.: Protein structure alignment, *J.Mol.Biol.*, 208, pp. 1-22 (1989).

[Tsukihara 82] Tsukihara, T., et al.: STRUCTURE-function relationship of [2Fe-2S] ferredoxins and design of a model molecule, *BioSystems*, 15, pp. 243-257 (1982).

[Yamamura 97] Yamamura, M., et al.: A Markov Analysis of Generation Alternation Models on Minimal Deceptive Problems, in *Proc. of Frontiers in EAs* (1997).

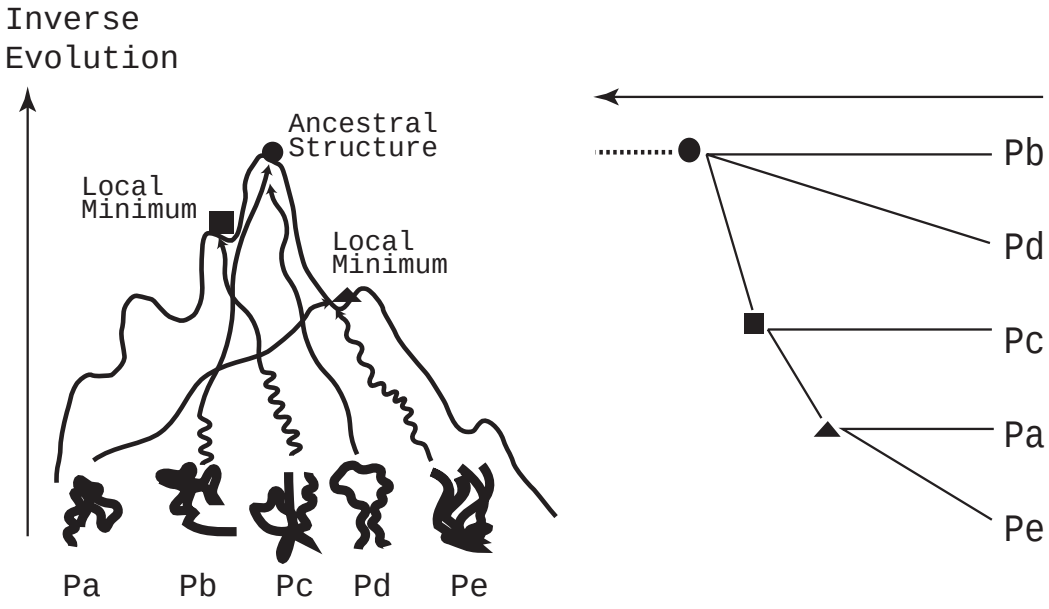
[Zhu 98] Zhu, J., et al.: Bayesian adaptive sequence alignment algorithms, *Bioinformatics*, Vol. 14, No. 1, pp. 25–39 (1998).

APPENDIX



T4Glutaredoxin(1aba) vs. E.coli(1dsbA)

Figure 16: Stronger Local Minimum



(a) MGSA (b) Phylogenetic Tree

Figure 17: An Idea for Multiple GSA

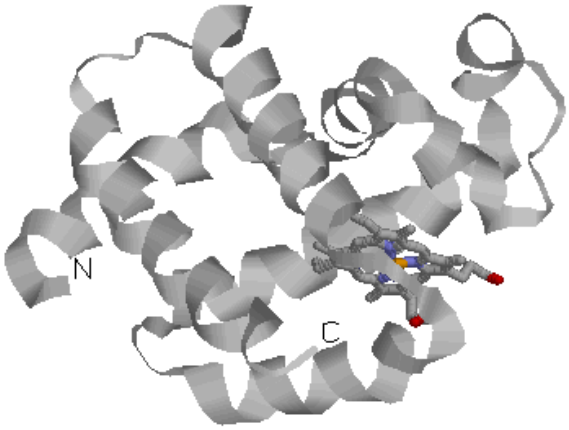


Figure 18: Structure of 1ecd.pdb

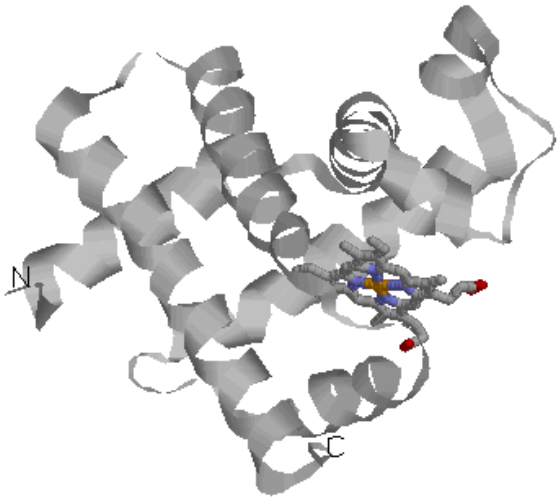


Figure 19: Structure of 1mbs.pdb

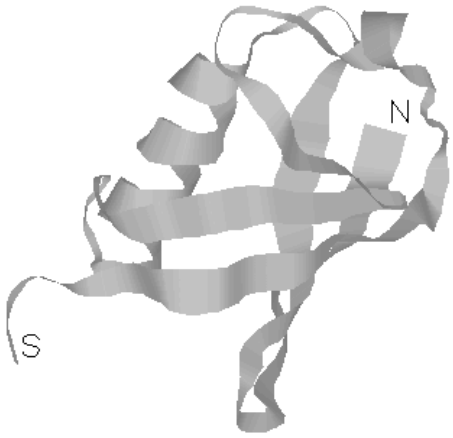


Figure 20: Structure of 1ubq.pdb

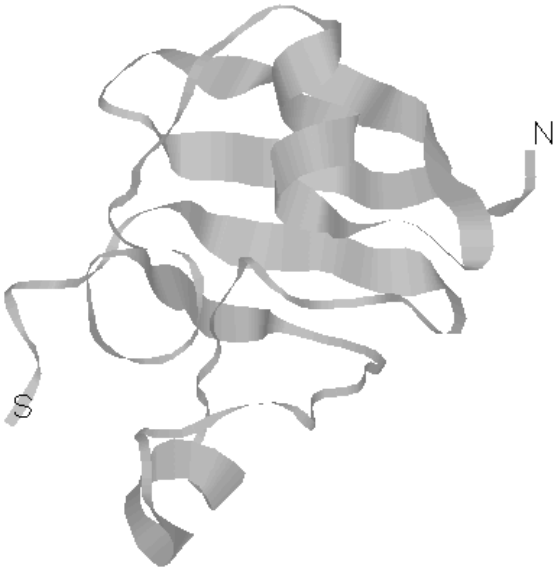


Figure 21: Structure of 4fxc.pdb