A SIMPLE ALIGNMENT ALGORITHM FOR THREE-DIMENSIONAL PROTEIN STRUCTURES

TATSUYA AKUTSU
Department of Computer Science, Gunma University
1-5-1 Tenjin, Kiryu, Gunma 376, Japan [a]

1 Introduction

Alignment is a basic and important tool for discovery of motifs. A lot of methods have been proposed for alignment of DNA or amino acid sequences. Recently, several studies have been done for alignment of three-dimensional protein structures.[1−5] However they do not seem to be sufficient from a viewpoint of computation time and the quality of the obtained alignment. Thus we have developed a simple and fast alignment algorithm for three-dimensional protein structures. The proposed algorithm is based on a well-known technique: *iterative improvement*. In this algorithm, first a rough alignment is found applying rms(*root mean square*)-fitting [3] to small fragments, and then an improvement procedure is repeatedly applied so that a good alignment is obtained. Note that a similar algorithm using iterative improvement was already proposed.[3] However our algorithm is different from their algorithm: a graph matching technique is used in our algorithm while a heuristic procedure is used in their algorithm. Thus our algorithm seems to be more general than their algorithm.

2 Algorithm

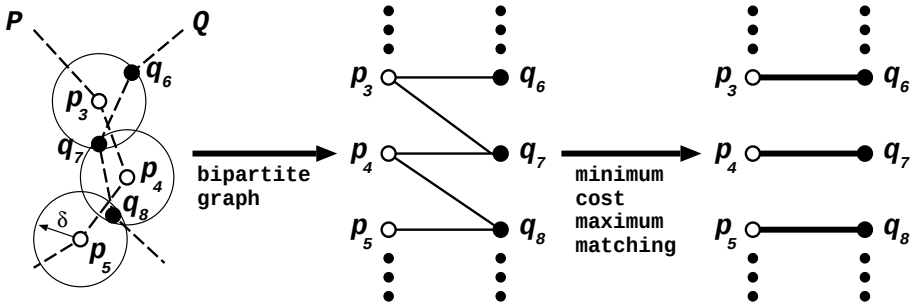
We assume that each three-dimensional structure is input as a sequence of points (Cα atoms) in three-dimensional space as usual. Let $P = (\boldsymbol{p}_1, \cdots, \boldsymbol{p}_m)$ and $Q = (\boldsymbol{q}_1, \cdots, \boldsymbol{q}_n)$ be two input sequences ($m \leq n$), where $\boldsymbol{p}_i$ (resp. $\boldsymbol{q}_i$) denotes a point in three-dimensional space. We call a correspondence $M = (\langle \boldsymbol{p}_{i_1}, \boldsymbol{q}_{j_1} \rangle, \cdots, \langle \boldsymbol{p}_{i_k}, \boldsymbol{q}_{j_k} \rangle)$ between P and Q an *alignment* if $i_1 < i_2 < \cdots < i_k$ and $j_1 < j_2 < \cdots < j_k$ hold. For an alignment M, $M(P)$ denotes a subsequence $(\boldsymbol{p}_{i_1}, \cdots, \boldsymbol{p}_{i_k})$ of P. P_i denotes a fragment $(\boldsymbol{p}_i, \cdots, \boldsymbol{p}_{i+L-1})$ of P, where we use $L = 15$ in the current version. $M(Q)$ and Q_i are defined in a similar way. Then the following procedure describes (a basic version of) the alignment algorithm, where $\delta, \delta_1, \delta_2, L_1$ are constants depending on inputs, and $d_{rms}(X, Y)$ denotes the rms distance between sequences X and Y of the same length.

[a] **e-mail:** akutsu@cs.gunma-u.ac.jp

(1) Let $M_0 = \{\}$.
(2) For all pairs of fragments (P_i, Q_j), do (3)-(9).
(3) If $d_{rms}(P_i, Q_j) > \delta_1$, test the next pair.
(4) Translate and rotate P applying *rms*-fitting to P_i and Q_j.
(5) Compute an alignment M between P and Q using a graph matching.
(6) If $|M| > L_1$ and $d_{rms}(M(P), M(Q)) < \delta_2$, repeat (7)-(8) several times.
(7) Translate and rotate P applying *rms*-fitting to $M(P)$ and $M(Q)$.
(8) Compute an alignment M between P and Q using a graph matching.
(9) If $|M| > |M_0|$ and $d_{rms}(M(P), M(Q)) < \delta_2$, then let $M_0 = M$.
(10) If $M_0 \neq \{\}$ then output M_0. Otherwise output "there is no good alignment".

Note that steps (7)-(8) correspond to an iterative improvement procedure. In steps (5) and (8), an alignment M between P and Q is computed using a graph matching technique in the following way (see Figure). From P and Q, we construct a bipartite graph $G(P, Q; E)$ where E is a set of edges between P and Q. $\langle \boldsymbol{p}_i, \boldsymbol{q}_j \rangle \in P \times Q$ is contained in E if $|\overline{\boldsymbol{p}_i \boldsymbol{q}_j}| < \delta$. Moreover a cost is associated with each edge ($cost(\boldsymbol{p}_i, \boldsymbol{q}_j) = |\overline{\boldsymbol{p}_i \boldsymbol{q}_j}|$). Then M is a *minimum cost maximum matching* of $G(P, Q; E)$ under the condition that M is an alignment.

Details of the algorithm and the experimental results comparing with other algorithms will be presented at the conference.



References

1. T. Akutsu, *Proc. 27th Hawaii Int. Conf. Syst. Sci.* **5**, 225–234 (1994).
2. R. Nussinov and H. J. Wolfson, *Proc. Natl. Acad. Sci. (USA)* **88**, 10495–10499 (1991).
3. S. T. Rao and M. G. Rossmann, *J. Molecular Biology* **76**, 241–256 (1973).
4. W. R. Taylor and C. A. Orengo, *J. Molecular Biology* **208**, 1–22 (1989).
5. G. Vriend and C. Sander, *PROTEINS: Structure, Function, and Genetics* **11**, 52–58 (1991).