Residue Contact with Dynamic Time Warping and Least Squares Adjustment for Protein Structure Alignment

Sean J.S. Lee
Department of Bioinformatics
Asia University

Han C.W. Hsiao
Department of Bioinformatics
Asia University
hwhsiao@asia.edu.tw

Jeffrey J.P. Tsai
Department of Computer Science
University of Illinois at Chicago

Abstract

Protein structure alignment is one of important issues in protein study. In general, such task can be divided into two categories, i.e. global and local structure alignment. In this paper, the proposed residue contact is extracted from original protein structure. A hybrid approach combining dynamic time warping and least squares adjustment is proposed for global alignment of protein 3D structures in an iterative fashion, where dynamic time warping is responsible for coarse alignment of two structures and least squares adjustment handles the fine matching of amino acid residues. The residuals of matched pairs are utilized to calculate the weights to accelerate the convergence of coarse-to-fine matching. The matched amino acid residues are transformed to the same coordinate system for calculating RMSD value. The preliminary results have demonstrated the effectiveness and efficiency of the proposed approach.

Keywords: protein structure alignment, dynamic time warping, least squares adjustment, coarse-to-fine matching

1. Introduction

Comparing three-dimensional protein structures is one of the most important issues in structural proteomics and is helpful in solving the problems of protein folding, drug design, motif finding, etc. The task of structure alignment is in general performed globally or locally. To carry out global alignment, two structures are generally aligned by affine transformation so as to calculate root mean square deviation (RMSD) value of three-dimensional coordinates, while local alignment aims at matching substructures with maximum local similarity between two protein structures. For example, Taylor and Orengo [13] proposed a method using double dynamic programming for global alignment problem. Subsequently, their approach [9] was applied to local alignment problem by using the torsion (phi and psi) angles and solvent accessibility to facilitate the alignment task. However, evaluating the structural environment of a residue is difficult. Later, Hiroike and Toh [6] proposed a method to construct a structural environment, which was robust against circular permutation. Akutsu and Horimoto [1] proposed a novel approach to multiple local structure alignment by integrating physicochemical characteristics and structural information of protein sequences to form a number of numeric profiles. These profiles were recoded back to some alphabetic sequences for the following local alignment. Lehtonen et al. [8] developed a tool to identify automatically regions with locally structural similarity in unrelated proteins containing different folds, as well as to define more global similarities from homologous protein structures. Zemla [15] proposed the LGA method for both local and global alignments in sequence independent and sequence dependent modes. It also took into account both local and global structure superposition without providing pre-assigned residue correspondence. Recently, Standley et al. [12] proposed an approach based on maximizing the number of spatially equivalent residues and realigning the structure using dynamic programming based on proximity of residues in the superposition. In this paper, the proposed residue contact is designed to describe protein structure. Following the previous work [7], the alignment is based on the dynamic time warping and least squares adjustment. The matched amino acid residues are transformed to the same coordinate system for calculating RMSD. In the following section, the proposed alignment framework is introduced. Section 3 gives some experimental results and discussion. The last section concludes the work with future improvement.

2009 Conference on Information Technology and Applications in Outlying Islands

108
2. Protein Structure Alignment

On the basis of coarse-to-fine matching strategy, the proposed approach consists of three major phases, i.e., feature extraction from protein structures, dynamic time warping for coarse alignment of protein structures, and least squares adjustment for fine matching of amino acid residues. Note that the extracted features as well as the original 3D coordinates of protein structures are utilized simultaneously for structural analysis. The advantage of using both original and extracted data is twofold. The extracted features are capable of representing the local characteristic that is more invariant to different coordinate systems, whereas the original 3D coordinates are considered for fine matching and accuracy calculation. The framework of protein structural alignment is illustrated in Figure 1, and the steps are given in more detail in the following.

![Figure 1. Flowchart of protein structure alignment.](image)

2.1 Weighted Residue Contact

Distance between any two amino acid residues is in general considered as intuitive information for protein structure analysis. Residue contact can be thought of as one of essential features derived from this concept to describe local characteristic along a protein structure. Given a specified radius, as shown in Figure 2, the original residue contact of an amino acid structure is simply the number of other amino acid residues falling into this range (or sphere centered at that amino acid) in a three-dimensional space by calculating the Euclidean distance $d_{ij}$ between amino acid residues $i$ and $j$.

A modified version of the residue contact $r_{cij}$ of amino acid residue $i$ is weighted by a Gaussian function, i.e.

$$r_{cij} = \sum_{j=1}^{m} e^{-\frac{d_{ij}}{2\sigma^2}}$$

where $\sigma$ is a specified radius and $m$ is the number of amino acid residues in a protein structure, as shown in Figure 2. In order to reveal the local characteristic, apparently, the parameter $\sigma$ should not be assigned a too small or large value. Empirically, $\sigma$ is assigned 5Å to have a better performance.

2.2 Dynamic Time Warping

Once the profile of the weighted residue contact is calculated for all protein structures, the alignment is carried out firstly by dynamic time warping [10] [14]. Two similar sequences of four-dimensional vectors are expected to be aligned through the introduction of expansion and contraction. Let two protein structures $\mathbf{x}$ and $\mathbf{y}$ with length of $M$ and $N$ be denoted as $\mathbf{x} = \{x_1, \ldots, x_M\}$ and $\mathbf{y} = \{y_1, \ldots, y_N\}$, respectively. An $M \times N$ matrix $\mathbf{M}$ is then constructed, where the matrix element $m_{ij}$ designates the normalized Manhattan distance between $x_i$ and $y_j$ and will be described later. Let $\mathbf{P}$ be a warping path consisting of a sequence of elements $P_k$ defined below, where $k = 1, 2, \ldots, K$ and $\min(M, N) \leq K \leq M+N-1$. The relationship between two consecutive elements $P_k$ and $P_{k+1}$ is as follows,

$$P_k = m_{i,j} \rightarrow P_{k+1} = m_{i',j'}$$

where $0 \leq (i - p) \leq 1, 0 \leq (j - q) \leq 1$ \hspace{1cm} (2)

$$i, p \leq M \text{ and } j, q \leq N$$

Therefore, the goal is to find an optimal warping path having the minimum accumulated distance $D_{ij}$ by evaluating the recursive equation

$$D_{ij} = m_{ij} + \min(D_{i-1,j-1}, D_{i-1,j}, D_{i,j-1})$$

2009 Conference on Information Technology and Applications in Outlying Islands
2.3 Least Squares Adjustment

Although dynamic time warping is a very advantageous approach relevant to sequence matching problems and has been employed in a wide variety of applications, the outcome of protein structure alignment using its standard version may not be satisfactory. The reason partially lies in the ignorance of adjacency between residues. In other words, the optimal warping path does not take into account the fact that any little difference between two local structures that should be matched will result in a deviation from a perfectly matched path. It is because the algorithm always tries to find a path with the minimum distance (or difference). It turns out that the aligned structures may contain one-to-many matching. To remedy the potential problem mentioned above, the matched result is regarded as a rough alignment, which provides a number of one-to-one matched pairs of points for solving the transformation parameters. In general, it suffices to transform protein structures by rotations and translations. Without loss of generality, suppose there are \( N_m \) pairs of one-to-one matched points extracted from the warping path \( P \). The transformation \( z = T(y) \) can be expressed as

\[
\begin{bmatrix}
    z_{x_1} \\
    z_{x_2} \\
    \vdots \\
    z_{x_{s}}
\end{bmatrix}
= \begin{bmatrix}
    r_{11} & r_{12} & r_{13} & \cdots & r_{1t} \\
    r_{21} & r_{22} & r_{23} & \cdots & r_{2t} \\
    \vdots & \vdots & \vdots & \ddots & \vdots \\
    r_{s1} & r_{s2} & r_{s3} & \cdots & r_{st}
\end{bmatrix}
\begin{bmatrix}
    y_{t1} \\
    y_{t2} \\
    \vdots \\
    y_{ts}
\end{bmatrix}
+ \begin{bmatrix}
    t_1 \\
    t_2 \\
    \vdots \\
    t_s
\end{bmatrix}
\tag{4}
\]

where \( s = 1, 2, \ldots, N_m \). The optimal transformation parameters are solved by least squares adjustment with redundant observations, which are the matched pairs of points. The energy function is defined as

\[
E = \sum_{s=1}^{N_m} \sum_{t=1}^{4} w_{st} \left( y_{st} - z_{st} \right)^2
\tag{5}
\]

After minimization of the energy function, the normal equations are constructed in a matrix form as \( AX = B \), where \( A \) is the 12x12 design matrix, \( B \) is the column vector of observation and \( X \) is the column vector of 12 parameters. Hence, the parameters solved at iteration \( t \) can be expressed as

\[
X_t = (A^\top A)^{-1}(A^\top B)
\tag{6}
\]

The structure of the test protein is then transformed into the coordinate system of the reference protein using the tentatively solved transformation parameters that provides an approximation between two protein structures for the subsequent fine matching. The result of fine matching is evaluated by a root mean square deviation (RMSD) given as

\[
RMSD = \sqrt{\frac{1}{N_m} \sum_{s=1}^{N_m} \left( x_{st} - z_{st} \right)^2}
\tag{7}
\]

The convergence criterion of the least squares adjustment is to examine if the relative change of the RMSD at the end of iteration \( t \) is within an insignificant tolerance \( \varepsilon \), that is

\[
\frac{|RMSD_{t+1} - RMSD_t|}{RMSD_t} \leq \varepsilon
\tag{8}
\]

where the tolerance \( \varepsilon \) is given as \( 10^{-3} \) in this study.

2.4 Normalized Manhattan Distance Matrix

The weighted residue contact profile can be pre-processed for each protein structure. To perform structural alignment for each pair of proteins, an initial distance matrix \( M \) is obtained by calculating dot product for each position combination. More specifically, the matrix element \( m_{ij} \) is defined as

\[
m_{ij} = 1 - \mathbf{r}_i \cdot \mathbf{r}_j \tag{9}
\]

where \( \mathbf{r}_i = [r_{C_2,i}, r_{C_3,i}, r_{C_4,i}, r_{C_5,i}, r_{C_6,i}]^\top \) indicates a vector retrieved from a sliding window of size \( 5 \) centered at residue \( i \) on the weighted residue contact profile of the first protein, while \( \mathbf{r}_j = [r_{C_2,j}, r_{C_3,j}, r_{C_4,j}, r_{C_5,j}, r_{C_6,j}]^\top \) indicates another vector retrieved from a sliding window centered at residue \( j \) on the profile of the second protein. Zero-padding is done on both ends of all profile to facilitate the calculation.

Subsequent to coarse alignment using the initial distance matrix, searching the optimal path is based on the normalized Manhattan distance defined as

\[
d_{i,j} = \sum_{k=1}^{N} w_k \left| x_{ki} - z_{kj} \right|
\]

where \( w_k = \frac{1}{\sigma_{k_i}} \) and \( \sigma_{k_i} = \frac{1}{N_m - 1} \sum_{i=1}^{N_m} (x_{ki} - z_{ki})^2 \)

Hence, the Manhattan distance between any pair of \( x_i \) and \( y_j \) from two proteins respectively is normalized by four weights \( w = [w_1, w_2, w_3, w_4] \), which is defined as the inverse of the standard deviations of the matched pairs. As two coordinate systems are becoming closer, the weights increasingly reflect the importance of the matched pairs.

3. Results and Discussion

A program has been implemented in MATLAB on a laptop computer equipped with a Pentium 1.6 GHz processor and 1GB RAM. All protein structures in this study are obtained from the Protein Data Bank [2]. Nine pairs of protein structures with various types of similarity measures [3] [9] [11] were used here for demonstration, that is globally similar, locally similar, and difficult to be aligned. Each pair of protein structures with PDB IDs, protein lengths (in parenthesis), sequence identity, the final RMSD value, the number of aligned residues \( N_m \), and computation time are provided in Table 1. An additional pair of proteins with completely different structures is arbitrarily selected as well for comparison. These aligned results are also compared with those by the incremental combinatorial
extension (CE) method [11]. In both of the globally and locally similar cases, the proposed weighted residue contact profile has shown to be representative feature, which provides a good initial condition so that alignment task can converge quickly. In the third case that is difficult to be aligned, the result may not be so perfect. It is possibly due to locally large dissimilarity between two structures by visual inspection. The more the matched residues, the larger the RMSD value is. Two proteins in the fourth case are totally different and should not be aligned correctly. It is expected that the hybrid approach still tries to find a best match but the final RMSD value is obviously higher. Figure 3 presents four pairs of protein structures with global similarity, where columns (c) shows the aligned results and the convergence profiles are shown in column (d). The experimental results are competitive in comparison with that of CE [11] listed in Table 1, and some cases are even better. The profiles in column (d) show that the alignment task converges quite well.

4. Conclusion

In this paper, the weighted residue contact profile has been proposed to be one useful feature for protein structure alignment. The experimental results demonstrate that such feature is able to present local characteristic of protein structure and provides a better initial condition in the stage of coarse matching. The overall alignment process can thus converge faster. Dynamic time warping is employed for approximate alignment. Least squares adjustment provides a chance to refine the matched result, which then feeds back to the step of approximate alignment in an iterative fashion. It is expected that two similar structures can be optimally matched in only few iterations. Many pairs of protein structures with various levels of similarity are aligned for demonstration, and the preliminary results are satisfactory. Nevertheless, the propose approach is not capable of handling ambiguous cases, e.g. one α-helix structure aligned with a protein having four α-helix structures. Moreover, it may not be successful if two protein structures are roughly similar but rather different locally. Future work of this study could be in the direction of elastic matching.

Acknowledgement

This work is supported in part under grant number NSC96-2221-E-468-011-MY3 from the National Science Council, Taiwan and Asia University.

References

Table 1. Comparison of structure alignments for 10 protein pairs

<table>
<thead>
<tr>
<th>Test</th>
<th>Reference</th>
<th>Identity (%)</th>
<th>Time (sec)</th>
<th>Proposed RMSD(Å)</th>
<th>Proposed N_m</th>
<th>CE RMSD(Å)</th>
<th>CE N_m</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Global</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>3DFR (162)</td>
<td>1DHFa (186)</td>
<td>27.8</td>
<td>1.603</td>
<td>1.890</td>
<td>153</td>
<td>1.7</td>
<td>158</td>
</tr>
<tr>
<td>1ATPe (350)</td>
<td>2CPKe (350)</td>
<td>100.0</td>
<td>6.036</td>
<td>0.367</td>
<td>334</td>
<td>0.4</td>
<td>336</td>
</tr>
<tr>
<td>1CDKa (350)</td>
<td>1CMKe (351)</td>
<td>100.0</td>
<td>7.418</td>
<td>1.837</td>
<td>337</td>
<td>2.1</td>
<td>343</td>
</tr>
<tr>
<td>1BPI (58)</td>
<td>1BUNb (61)</td>
<td>34.5</td>
<td>0.192</td>
<td>1.549</td>
<td>52</td>
<td>1.5</td>
<td>55</td>
</tr>
<tr>
<td>Local</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>3ICB (75)</td>
<td>4CPV (109)</td>
<td>24.6</td>
<td>0.444</td>
<td>3.038</td>
<td>64</td>
<td>3.4</td>
<td>65</td>
</tr>
<tr>
<td>2ASR (142)</td>
<td>2BRD (247)</td>
<td>7.7</td>
<td>1.985</td>
<td>3.136</td>
<td>97</td>
<td>4.3</td>
<td>117</td>
</tr>
<tr>
<td>Difficult</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>1BGEb (175)</td>
<td>2GMFa (127)</td>
<td>12.1</td>
<td>2.088</td>
<td>3.509</td>
<td>76</td>
<td>3.9</td>
<td>107</td>
</tr>
<tr>
<td>1CEWi (108)</td>
<td>1MOLa (94)</td>
<td>17.3</td>
<td>0.549</td>
<td>2.843</td>
<td>57</td>
<td>2.3</td>
<td>81</td>
</tr>
<tr>
<td>1FXIa (96)</td>
<td>1UBQ (76)</td>
<td>9.4</td>
<td>0.458</td>
<td>3.189</td>
<td>44</td>
<td>3.8</td>
<td>64</td>
</tr>
<tr>
<td>Different</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>1XWM (217)</td>
<td>1Q3B (262)</td>
<td>13.8</td>
<td>6.954</td>
<td>7.934</td>
<td>103</td>
<td>5.3</td>
<td>80</td>
</tr>
</tbody>
</table>

Figure 3. Four pairs of protein structures with global similarity are drawn by PyMol [4]. Columns (a) and (b) represent the test and reference structures, respectively. Column (c) shows the aligned results, where the test and reference proteins are drawn in black and gray, respectively. Convergence profiles are shown in column (d).
Figure 4. Two pairs of protein structures with local similarity.

Figure 5. Three pairs of protein structures identified to be difficult to be aligned.