Acta Crystallographica Section A
Foundations of Crystallography
ISSN 0108-7673

Received 6 June 2003
Accepted 16 February 2004

A strict solution for the optimal superimposition of protein structures

Chuanbo Chen and Qishen Li*

College of Computer Science and Technology, Huazhong University of Science and Technology, Wuhan, People's Republic of China. Correspondence e-mail: li_qishen@yahoo.com

Existing methods for the optimal superimposition of one vector set on another in the comparison of parts or the whole of related protein molecules are based on the precondition that the centroids of the two sets are coincident. As a result, the translation components of the transformation are artificially removed from the superimposition process. This is obviously not strict in the mathematical sense. The theorem presented in this paper is a strict solution for the optimal superimposition of two vector sets, which is in fact the problem of the weighted optimal rigid superimposition of two vector sets. Examples show its advantages compared with the method of simply coinciding the centroids of the two vector sets for the translation transformation.

1. Introduction

The comparison of protein structures is an important problem in the study of bioinformatics. The optimal superimposition of one vector set on another is very useful in the comparison of parts or the whole of related protein molecules for finding a rigid combination transformation of rotation and translation, and its solution has attracted the attention of a number of authors, notably McLachlan (1972, 1979, 1982), Kabsch (1976, 1978), Diamond (1976, 1988), Ferro & Hermans (1977), Lesk (1986), Kaindl & Steipe (1997) and Steipe (2002). All of their methods achieve the optimal rotational superimposition by taking as a precondition the coincidence of the centroids of two coordinate sets. That is, the translation components are simply obtained by this precondition, and no further attention is given to them. This kind of disposal is not strict for this problem, because the rotation and translation components of the transformation are correlative in a mathematical sense for achieving the optimal superimposition.

The problem of the optimal superimposition of two protein structures is the weighted optimal rigid superimposition of two vector sets, which are the atom coordinate vectors of protein molecules. The theorem given in this paper is a strict solution for this problem in the mathematical sense, which gives not only the strict transformation parameters but also the minimal value of the mean squared deviation of the optimal superimposition. This work is derived by elaborating on Umeyama's work (Umeyama, 1991), which is applied in the field of computer vision.

2. Representation of the superimposition problem

The structure of a protein molecule consisting of n atoms can be described using n three-dimensional coordinate vectors such as (x_1, y_1, z_1), $(x_2, y_2, z_2), \ldots, (x_n, y_n, z_n)$. Therefore, the problem of protein structure optimal superimposition can be stated as follows.

Given two sets X_0 and Y_0 of n three-dimensional column vectors a_k and b_k ($k = 1, 2, \ldots, n$), find a 3×3 orthogonal rotation matrix R with determinant +1 and a 3×1 translation vector t which convert the coordinates x_{ik} ($i = 1, 2, 3$) to

$$x'_{ik} = \sum_j R_{ij} x_{jk} + t_i \tag{1}$$

and minimize the function

$$e^2(R, t) = (1/n) \sum_{ik} w_k (y_{ik} - x'_{ik})^2. \tag{2}$$

Here w_k is a weight assigned to the kth atom. This equation can be converted to

$$\begin{aligned} e^2(R, t) &= (1/n) \sum_k \| w_k^{1/2} y_k - w_k^{1/2} (R x_k + t) \|^2 \\ &= (1/n) \sum_k \| \tilde{y}_k - (R \tilde{x}_k + w_k^{1/2} t) \|^2 \\ &= (1/n) \| Y - (RX + t w^T) \|^2, \end{aligned} \tag{3}$$

where $\tilde{x}_k = w_k^{1/2} x_k$, $\tilde{y}_k = w_k^{1/2} y_k$, X and Y are two sets consisting of n-column vectors $\tilde{x}_k$ and $\tilde{y}_k$, $w = (w_1^{1/2}, w_2^{1/2}, \ldots, w_n^{1/2})^T$. Thus, the problem of the optimal superimposition of two protein structures is converted to obtain the rotation and translation parameters R and t of the transformation between two vector sets X and Y.

3. Theorem of a strict solution for optimal superimposition

Umeyama (1991) proposed a very useful theorem for finding the similarity transformation parameters (rotation, translation and scaling) that give the least mean squared error between two point sets. In this section, we derive a valuable theorem by introducing weights to the data sets and cancelling the scaling

Table 1
Comparison of RMSDs derived from the CCM and the NCCM.

Since the superimposition method requires the same number of pairs of vector sets, N residues are selected to be superimposed from the beginning of the N-terminal of the protein. When only one α-carbon atom of each residue is used for superimposition, the number of atom pairs in the superimposition sequence is N. When four atoms (nitrogen, α-carbon, carbon and oxygen) of each residue are used, the number of atom pairs is $4 \times N$. In this experiment, three pairs of protein molecules are superimposed by the CCM aand NCCM, respectively. The values of the RMSDs listed in the table are then calculated with pairs of the transformed vector sets.

Atoms used	PDB ID of protein pairs	Superimposition method	RMSD (Å)		
			$N = 10$	$N = 50$	$N = 100$
Only α-carbon	1lyz *versus* 2lzm	CCM	7.78	2.72	2.22
		NCCM	1.26	1.69	1.48
	1hfc *versus* 1mnc	CCM	4.05	2.92	1.95
		NCCM	1.03	0.52	0.37
	1ulb *versus* 2ctc	CCM	3.49	1.67	1.76
		NCCM	1.34	1.45	1.25
Nitrogen, α-carbon, carbon and oxygen	1lyz *versus* 2lzm	CCM	3.87	1.35	1.10
		NCCM	0.58	0.83	0.73
	1hfc *versus* 1mnc	CCM	2.03	1.46	0.97
		NCCM	0.49	0.25	0.18
	1ulb *versus* 2ctc	CCM	1.83	0.82	0.87
		NCCM	0.64	0.72	0.62

component of the transformation in Umeyama's theorem, which can give a strict solution for the above superimposition problem; all of the rotation and translation transformation parameters are obtained by means of the least-squares residual.

Theorem 1. Let $X = \{x_1, x_2, \ldots, x_n\}$ and $Y = \{y_1, y_2, \ldots, y_n\}$ be corresponding point sets in m-dimensional space. The minimum value ε^2 of the mean squared deviation

$$e^2(R, t) = (1/n)\sum_{i=1}^{n} \|Y - (RX + tw^T)\|^2 \qquad (4)$$

of these two point sets with respect to the transformational parameters [R (rotation) and t (translation)] of the optimal rigid superimposition is given as follows:

$$\varepsilon^2 = (1/n)\{\|YK\|^2 + \|XK\|^2 - 2\,\mathrm{tr}(DS)\}, \qquad (5)$$

where

$$K = I - w_0^{-1} w w^T. \qquad (6)$$

Here, I is the identity matrix, $w_0 = \sum_k w_k$, D is derived by letting a singular value decomposition of YKK^TX^T be UDV^T [$D = \mathrm{diag}(d_i)$, $d_1 \geq d_2 \geq d_m \geq 0$] and

$$S = \begin{cases} I & \text{if } \det(YKK^TX^T) \geq 0 \\ \mathrm{diag}(1, 1, \ldots, 1, -1) & \text{if } \det(YKK^TX^T) < 0. \end{cases} \qquad (7)$$

When $\mathrm{rank}(YKK^TX^T) \geq m - 1$, the optimum transformation parameters are determined uniquely as follows:

$$R = USV^T, \qquad (8)$$

$$t = (1/w_0)(Yw - RXw), \qquad (9)$$

where S in (8) must be chosen as

$$S = \begin{cases} I & \text{if } \det(U)\det(V) = 1 \\ \mathrm{diag}(1, 1, \ldots, 1, -1) & \text{if } \det(U)\det(V) = -1 \end{cases} \qquad (10)$$

when $\mathrm{rank}(YKK^TX^T) = m - 1$.

The above theorem is an extended version of Umeyama's (1991) theorem. Also, it can be easily proved according to the proof of Umeyama's theorem noting that in the proof process c and h in Umeyama's formulae are replaced by 1 and w, respectively; accordingly, n is replaced by w_0 in the computation of K. Here c is the scaling transformation parameter and $h = (1, 1, \ldots, 1)^T$.

From Theorem 1, we can see that it gives not only the rotation matrix and translation vector of transformation for optimal superimposition but also the minimum value of the mean squared deviation directly from the given two vector sets and weights assigned to data points.

4. Examples and contrast

In order to elucidate the advancement of the proposed method in contrast with the existing methods, which are based on the precondition that the centroids of the two vector sets coincide, we show some numerical results of the RMSD (root mean squared deviation) for some examples in Table 1. These results are calculated using two different methods, called the centroid-coincidence method (CCM) and the non-centroid-coincidence method (NCCM, *i.e.* the method proposed in this paper), which differ by forcing the centroids of the two vector sets to coincide or not, respectively. Here we select Kabsch's method (Kabsch, 1976, 1978) as the representative of the CCM. Generally, according to the atoms of protein molecules used in the superimposition process, specialists and scholars adopt two kinds of superimposition. One is when only the α-carbon atom and the other is where four atoms, nitrogen α-carbon, carbon and oxygen, of the peptide backbone are used per residue. The weights assigned to these four kinds of atoms in this paper are 0.6, 1, 0.8 and 0.9, which may not have any practical significance to biology but demonstrates the

numerical differences of RMSDs derived from the CCM and the NCCM defined above.

From Table 1, we can see that the results of the RMSD of the NCCM are obviously smaller than those of the CCM. Thus we can say that the precondition that the centroids of the two vector sets coincide is not strictly fulfilled, and the superimposition obtained based on it is in fact not optimal.

5. Conclusions

In this paper, we have presented a strict solution in the mathematical sense for the optimal superimposition between two vector sets of protein atom coordinates, which actually belongs to the problem of the weighted optimal rigid superimposition between two m-dimensional (m is a discretionary natural number) vector sets.

This work was partially funded by the National High Technology Development 863 Program of China under Grant No. 2001AA231071. The authors thank DaHua He for discussions about the theory of matrices used in this paper.

References

Diamond, R. (1976). *Acta Cryst.* A**32**, 1–10.
Diamond, R. (1988). *Acta Cryst.* A**44**, 211–216.
Ferro, D. R. & Hermans, J. (1977). *Acta Cryst.* A**33**, 345–347.
Kabsch, W. (1976). *Acta Cryst.* A**32**, 922–923.
Kabsch, W. (1978). *Acta Cryst.* A**34**, 827–828.
Kaindl, K. & Steipe, B. (1997). *Acta Cryst.* A**53**, 809.
Lesk, A. M. (1986). *Acta Cryst.* A**42**, 110–113.
McLachlan, A. D. (1972). *Acta Cryst.* A**28**, 656–657.
McLachlan, A. D. (1979). *J. Mol. Biol.* **128**, 49–79.
McLachlan, A. D. (1982). *Acta Cryst.* A**38**, 871–873.
Steipe, B. (2002). *Acta Cryst.* A**58**, 506.
Umeyama, S. (1991). *IEEE Trans. Pattern Anal. Mach. Intell.* **13**, 376–386.

Acta Crystallographica Section A
Foundations of Crystallography
ISSN 0108-7673

addenda and errata

A strict solution for the optimal superimposition of protein structures. Retraction

Chuanbo Chen and Qishen Li*

College of Computer Science and Technology, Huazhong University of Science and Technology, Wuhan, People's Republic of China. Correspondence e-mail: li_qishen@yahoo.com

In the paper 'A strict solution for the optimal superimposition of protein structures' by Chuanbo Chen & Qishen Li [*Acta Cryst.* (2004), A**60**, 201–203], we claimed that existing methods for the optimal superimposition of two point sets, requiring the precondition of coincident centroids, are mathematically not strict. It has been brought to our attention that this claim is erroneous. We therefore retract the publication.

  doi: 10.1107/S0108767304024730