Supplementary information

MICAN-SQ: A sequential protein structure alignment program that is applicable to monomers and all types of oligomers

Shintaro Minami\textsuperscript{1,\#}, Kengo Sawada\textsuperscript{2,\#}, Motonori Ota\textsuperscript{1}, and George Chikenji\textsuperscript{2,*}

\textsuperscript{1}Department of Complex Systems Science, Graduate School of Informatics, Nagoya University, Nagoya, Aichi, Japan,
\textsuperscript{2}Department of Applied Physics, Graduate School of Engineering, Nagoya University, Nagoya, Aichi, Japan

\# Equally contributed.
\* To whom correspondence should be addressed. E-mail: chikenji@tbp.cse.nagoya-u.ac.jp

Supplementary Note S1

This note provides a brief description of the MICAN program, and highlights similarities and differences between MICAN-SQ, RW, and RR. In all the three alignment modes, the alignment algorithm comprises the following three basic steps: (i) alignments of Secondary Structure Elements (SSEs), (ii) alignments of residues, and (iii) ranking of alignments, as described below.

The 1st step: alignments of SSEs. In the first step, for given two protein structures, a query (\(Q\)) and a template (\(T\)), the program performs SSE alignments in all the three modes. Initially, all SSEs of the two proteins are represented by multiple vectors. Secondly, the geometric hashing algorithm (Nussinov and Wolfson, 1991) identifies superpositions of the two structures that show many correspondences of SSE vectors between the two structures. Thirdly, the algorithm reports the top scoring 50 superpositions, and these are used to set up an initial superimposition of the two structures in the residue level alignment step. This procedure of SSE alignment is essentially the same in all the three modes, except that SSE matches with reverse N- to C-terminal directions of protein chains are only allowed in RR. Because SQ and RW utilize an identical algorithm in this step, these alignments may violate the sequential rule in SQ at that moment. A complete description of this step is provided in our previous report (Minami et al., 2013).

The 2nd step: alignment of residues. One-to-one mapping of C\(_\alpha\) atoms is performed for each of the 50 superimpositions obtained in the SSE alignment step. Based on the initial superimposition obtained by SSE alignments, an initial one-to-one map of C\(_\alpha\) atoms is generated using the greedy algorithm in all three modes. This algorithm aligns residue pairs from those superimposed with smaller to larger distances. Subsequently, iterative refinements of the alignment are performed until the structural similarity score converges. This procedure is essentially the same for SQ, RW, and RR, although restrictions are imposed on each alignment scheme. A detailed description of this algorithm for RW and RR is provided in a previous paper (Minami et al., 2013). The SQ algorithm for constructing sequential alignments from the initial superimposition is reported in the main text.

The 3rd step: selecting the best alignment. The best scoring alignment is reported based on SSE match-weighted TMscores (sTMscore) (Minami et al., 2013) from the 50 alignments generated by the residue level alignment, and the sTMscore between the query (\(Q\)) and the template (\(T\)) for a given superimposition is defined as follows:

\[
s\text{TMscore}(Q \rightarrow T) = \frac{1}{N_Q} \sum_{(i,j)} \left[ \frac{1}{1 + d_{ij}^2/d_0^2} \times \frac{\delta_{\sigma_i,\sigma_j} + 1}{2} \right],
\]

(1)
where $N_Q$ is the number of residues of the query protein. The first factor in the brackets of Equation (1), $1/(1 + d_{ij}^2/d_0^2)$, measures spatial proximities of two $C_\alpha$ atoms, where $d_{ij}$ is the distance between the $C_\alpha$ atoms of the $i$-th residue of the query and that of the $j$-th residue of the template. $d_0$ is used as a scaling factor to normalize match differences and is defined as $d_0 = 1.24 \sqrt[3]{N_Q - 15} - 1.8$. This formula was originally introduced by Zhang & Skolnick to eliminate protein size dependences of score functions (Zhang and Skolnick, 2004). The second factor $(\delta_{\sigma_i,\sigma_j} + 1)/2$ was introduced so that aligned residues belong to the same SSE type (Minami et al., 2013), where $\sigma_i$ represents the three states of the secondary structure (helix, strand, or coil) of residue $i$, and $\delta_{\sigma_i,\sigma_j}$ is Kronecker $\delta$ delta. Sub-optimal alignments can also be reported if a user specifies the rank of the sub-optimal alignment. Note that completely the same scoring functions, stTMscore, is used for all three modes. Accordingly, differences in structure alignments arise purely from differences in the restrictions imposed on each alignment mode.

Supplementary Note S2

In this note, we introduce several quantities of the MICAN program. In addition to stTMscores, MICAN returns TMscores (Zhang and Skolnick, 2004), Dali scores (Holm and Sander, 1993), SP-scores (Yang et al., 2012), numbers of aligned residues, root-mean-square deviations (RMSD) of aligned residues, and sequence identities of aligned residues. These quantities are generated for all alignments constructed by MICAN and offer helpful criteria on which to evaluate global alignments.

In addition, estimates of local structure similarities are available and are required when protein pairs have locally similar fragments, such as functional motifs, but adopts globally different structures. MICAN calculates stTMscores that are normalized by the number of aligned residues (stTMscore($N_{\text{ali}}$)), and these are defined by the following equation:

$$\text{stTMscore}(N_{\text{ali}}) = \frac{1}{N_{\text{ali}}} \sum_{(i,j)} \left[ \frac{1}{1 + d_{ij}^2/d_0^2} \times \frac{\delta_{\sigma_i,\sigma_j} + 1}{2} \right], \quad (2)$$

in which $d_0$ is almost the same as $d_0$ from Equation (1), but $N_Q$ is replaced by the number of aligned residues ($N_{\text{ali}}$). stTMscores ($N_{\text{ali}}$) range from 0 to 1, and higher values indicate stronger similarity in aligned regions. The quantity in the brackets of the Equation (2) indicates closeness scores for each of aligned residue pairs and is denoted as $C(i,j)$. This parameter indicates spatial proximities between residue $i$ of the query and the residue $j$ of the template for superimposed structures as follows:

$$C(i,j) = \frac{1}{1 + d_{ij}^2/d_0^2} \times \frac{\delta_{\sigma_i,\sigma_j} + 1}{2}.$$  

$C(i,j)$ values of aligned residue pairs can be obtained by specifying the command line option “-a”.

Supplementary Note S3

This note defines the matrix that is used in residue level alignments. To identify residue pairs that are superimposed with smaller distances and similar local-backbone geometries for a given superimposition of the query ($Q$) and the template ($T$) structure, an $N_Q \times N_T$ matrix $M_{i,j}$ is constructed, where $N_Q$ and $N_T$ are numbers of residues of the query protein and template, respectively. This matrix differs slightly from that used in the previous paper (Minami et al., 2013), and modifications of the matrix were designed to accommodate geometries of local-backbone conformations. The matrix element $M_{i,j}$ for a given superimposition is defined as follows:

$$M_{i,j} = \frac{1}{1 + d_{ij}^2/d_0^2} \times \frac{\delta_{\sigma_i,\sigma_j} + 1}{2} \times e^{-\theta_i^2/\theta_0^2} \times e^{-\omega_i^2/\omega_0^2}. \quad (3)$$
In this equation, the first and second factors are the same as those in Equation (1) in the Supplemental Note S1. The third factor \( e^{-\theta_0^2/\theta_0^2} \) was introduced so that aligned residues have the same \( C_\alpha \) to \( C_\beta \) direction. Because MICAN uses only \( C_\alpha \) atoms, pseudo \( C_\beta \) atoms, which are defined by three consecutive \( C_\alpha \) atoms, are used to calculate vectors instead of actual \( C_\beta \) atoms. Accordingly, the vector \( \mathbf{v}_i \) from \( C_\alpha \) to pseudo \( C_\alpha \) atom of residue \( i \) is defined as follows:

\[
\mathbf{v}_i = \frac{(r_i - r_{i-1}) + (r_i - r_{i+1})}{|(r_i - r_{i-1}) + (r_i - r_{i+1})|},
\]

where \( r_i \) denotes the coordinate of the \( C_\alpha \) atom of residue \( i \), and \( \theta_0 \) is set at \( \pi/2 \). The fourth factor \( e^{-\omega_0^2/\omega_0^2} \) was introduced so that aligned residues run in the same backbone direction. We define the backbone direction \( \mathbf{b}_i \) of residue \( i \) as

\[
\mathbf{b}_i = \frac{(r_{i+1} - r_{i-1})}{|(r_{i+1} - r_{i-1})|}.
\]

The angle \( \omega_0 \) is calculated as \( \omega_0 = \cos^{-1}(\mathbf{b}_i \cdot \mathbf{b}_j) \), and \( \theta_0 \) is set to \( \pi/2 \). Although essentially the same matrix is used for SQ, RW, and RR in the current version of the MICAN program, in RR, \( \omega_0 \) is calculated as \( \omega_0 = |\cos^{-1}(\mathbf{b}_i \cdot \mathbf{b}_j)| \) to allow matching with the reverse orientation of protein chains.

**Supplementary Note S4**

In this note, we describe the procedure for analyses of the glyoxalase/bleomycin resistance protein/dihydroxybiphenyl dioxygenase family.

**The dataset**

In these analyses, we employed the Evolutionary Classification of protein Domains (ECOD) database (version develop203) (Cheng et al., 2014), and domains of this protein family were obtained from the F40 non-redundant domain list (< 40% sequence identity). For each domain, a biological assembly provided by the PDB and the corresponding quaternary structure of the PISA server (Proteins, Interfaces, Structures and Assemblies) (Krissinel and Henrick, 2007) were compared. When two assembly states were the same, we assumed that the corresponding oligomeric structure was biologically true. Furthermore, domains of >350 amino acids were eliminated, because the typical size of this family is approximately 250, and 350 would be uncharacteristically long. Subsequently, we identified 64 domains and included them in clustering analyses.

**The score function for discriminating the assembly states**

To estimate similarities of chain traces over multiple chains, we introduced the Chain-to-Chain Correspondence score (\( C^3 \)score), which is defined as follows:

\[
C^3 \text{score} = \frac{N_{\text{ali}}'}{N_{\text{ali}}},
\]

where \( N_{\text{ali}} \) is the number of aligned residues and \( N_{\text{ali}}' \) is the maximum number of aligned residues calculated assuming that one chain of a protein corresponds to only one chain of another. To define the meaning of \( C^3 \)scores, let us consider the three alignments of protein pairs in Figure S12 (A)-(C). TMscores were calculated by MICAN-SQ for the three pairs and are essentially the same, whereas chain-to-chain correspondences differ considerably. The structural alignment in Figure S12(A) shows that most residues of chain A of a protein correspond to those of A in another protein,
and the same correspondences are relevant to B chains, indicating perfect one(chain)-to-one(chain) correspondences. In contrast, Figures S12(B) and (C) show one-to-two chain correspondences. For example, in Figure S12(B), chain A of 1kll corresponds to the N terminal half of chain A and the C terminal half of chain B of 1kmz, and the chain-traces over multiple chains differ. Thus, C3 scores discriminate alignments of Figure S12(A) from those of (B) and (C). Note that Nali′ is always smaller than or equal to Nali, because Nali is calculated by subtracting lengths that are aligned with other chains from Nali. The C3 score of the protein pair in Figure S12(A) is 1.0, whereas those of (B) and (C) are around 0.5. C3 scores are designed to evaluate only chain-to-chain correspondences and are not subject to structural similarities. Therefore, C3 scores are complementary to structural similarity scores. For example, C3 scores are always one if the compared two proteins are monomers, regardless of structure similarities.

The clustering analysis

For the 64 structures of the glyoxalase/bleomycin resistance protein/dihydroxybiphenyl dioxygenase family, we performed all-vs.-all structure alignments using MICAN-SQ and computed TMscore and C3 score. In this analysis, distances between the two structure i and j are defined as

\[ d_{i,j} = 1 - \frac{1}{2} (\text{C3 score}(i, j) + \text{TMscore}(i, j)), \]

where TMscore(i, j) is the average of TMscore(i → j) and TMscore(j → i). Thus, we constructed a dendrogram using the average linkage scheme (Figure 5 in the main text) with R statistical software.

References


Figure S1: An example of a structure alignment of a domain-swapped dimer and its structurally similar monomer; (A) structure of the domain-swapped dimer (the platelet activator convulxin, PDB ID: 1umr); chains B and D are red and blue, respectively. (B) The structure of the non-swapping monomer of a similar protein; mannose-binding protein, PDB ID: 1msb; (C) Superimposed structure of 1umr and 1msb; (D) alignment plot of the two proteins.
Input protein Q and M

SQ

SSE level alignment
prohibiting the reverse orientation

generates 50 superpositions

RW

SSE level alignment
allowing the reverse orientation
generates 50 superpositions

RR

residue level alignment with sequential constraint

1 2 50

residue level alignment allowing only rewiring

1 2 50

residue level alignment allowing both rewiring & reverse orientation

1 2 50

ranking the alignments based on sTM-score from the 50 candidates

output results

Figure S2: Flowchart of the MICAN algorithm; Black-colored regions represent the same sub step of the previously described algorithm. Red-colored regions represent newly introduced parts.
Figure S3: Procedure for residue level alignments of MICAN-SQ; for simplicity, the query protein comprises one chain and the template comprises two chains. The vertical axis represents residue numbers of the query protein, and the horizontal axis represents the residue number of the template. (A) All segments for a given superposition of the query and the template are shown as black diagonal lines. (B) The segment with the highest score among the segments $A_1$ is indicated as a red diagonal line. (C) The matrix elements interfering with $A_1$ and those violating the sequential rule are indicated as cyan- and magenta-colored regions, respectively. (D) Modified matrix; matrix elements interfering with $A_1$ and those violating the sequential rule are set to zero. The first member of the set of segments of the alignment is indicated as a blue diagonal line. (E)-(J) The sequential alignment that approximately maximizes the total score $S_{tot}$ is obtained by repeating the procedure (B)-(D).
Figure S4: Box plots of the distributions of Q-scores obtained using all of the methods compared in the test sets for MALIDUP (upper) and MALISAM (lower).
Figure S5: Scatter matrix of the Q-scores for all pairs of compared methods in the MALIDUP set; Upper right panels with diagonal lines are scatter plots and P values are shown in the lower left panels; (* P < 0.05; ** P < 0.01; *** P < 0.001).
Figure S6: Matrix of Q-score Scatterplots for all pairs of compared methods in the MALISAM set. Upper right panels with diagonal lines are scatter plots and lower left panels show P-values; (* P < 0.05; ** P < 0.01; *** P < 0.001).
Figure S7: An example of structure alignments using SQ(top), RW(middle), and RR(bottom) schemes for protein pairs of aldehyde reductase (ALDH; PDB ID: 1o02) and hypothetical protein MT938 (PDB ID: 1ihn). The representations on the left are superpositions of 1o02 and 1ihn, and the corresponding alignment plots are shown in the right panels, in which aligned residue pairs are represented as colored circles. Horizontal axes represent aldehyde reductase residue numbers, and the vertical axis represents the residue numbers of the hypothetical protein MT938. Reference alignments from the MALISAM database are shown as gray circles.
Figure S8: Box plots of SID-score distributions from all methods for comparisons in the 3DSwap-PS test set.
Figure S9: Scatter matrix of SID scores for all pairs of methods for the 3DSwap-PS set; the upper right panels with diagonal lines are scatter plots, and P values are shown in the lower left panels; (* P < 0.05; ** P < 0.01; *** P < 0.001).
Figure S10: An example of structure alignment of the trimeric single-domain lectin (PDB ID: 2bt9) and its corresponding monomeric single-domain lectin (PDB ID: 1ofz); (A) structure of 2bt9; chains A, B, and C are shown in blue, green, and red, respectively; (B) structure of 1ofz (yellow); (C) superposition of 2bt9 and 1ofz by MICAN-SQ; (D) alignment plot of the protein pair obtained by MICAN-SQ.
Figure S11: Structure comparison of the bleomycin resistance protein (BRP; PDB ID: 1ewj) and the uncharacterized protein Atu1953 (PDB ID: 2pjs); the middle illustrations are structures from the symmetrical axis of dimers (top view), and the left illustrations are their schematics. The right illustrations show the same structures rotated by 90° (side view). (A) Complex structures of BRP and bleomycin; chains A and B are shown in blue and pink, respectively. The two bound bleomycins are shown in gray; (B) structure of the uncharacterized protein Atu1953; chains A and B are shown in yellow and green, respectively. The loops connecting the 4th and 5th strands in 2pjs (residue 55-63) are shown in red. (C) Structure superimposition of the two proteins by MICAN-SQ; strong similarities of protein structures were detected with a TMscore of 0.71, but the clear structural overlap between bleomycin and the red loops indicates that Atu1953 does not bind bleomycin with the same binding mode as BRP.
Figure S12: Examples of structural pairs of the glyoxalase/bleomycin resistance protein/dihydroxy-biphenyl dioxygenase family; schematic of protein structure pairs (left and middle panels) and their alignment plot (right); (A) 1kll and 2a4w, (B) 1kll and 1kmz, and (C) 1kll and 5d7z; Note that 1kll, 2a4w, and 1kmz are dimers, and 5d7z is a monomer.