BIOINFORMATICS ORIGINAL PAPER Vol. 28 no. 3 2012, pages 324–330
doi:10.1093/bioinformatics/btr654

Structural bioinformatics Advance Access publication November 29, 2011

SCPC: a method to structurally compare protein complexes

Ryotaro Koike* and Motonori Ota
Department of Complex Systems Science, Graduate School of Information Science, Nagoya University, Furo-cho, Chikusa-ku, Nagoya 464-8601, Japan

Associate Editor: Alfonso Valencia

ABSTRACT

Motivation: Protein–protein interactions play vital functional roles in various biological phenomena. Physical contacts between proteins have been revealed using experimental approaches that have solved the structures of protein complexes at atomic resolution. To examine the huge number of protein complexes available in the Protein Data Bank, an efficient automated method that compares protein complexes is required.

Results: We have developed Structural Comparison of Protein Complexes (SCPC), a novel method to structurally compare protein complexes. SCPC compares the spatial arrangements of subunits in a complex with those in another complex using secondary structure elements. Similar substructures are detected in two protein complexes and the similarity is scored. SCPC was applied to dimers, homo-oligomers and haemoglobins. SCPC properly estimated structural similarities between the dimers examined as well as an existing method, MM-align. Conserved substructures were detected in a homo-tetramer and a homo-hexamer composed of homologous proteins. Classification of quaternary structures of haemoglobins using SCPC was consistent with the conventional classification. The results demonstrate that SCPC is a valuable tool to investigate the structures of protein complexes.

Availability: SCPC is available at http://idp1.force.cs.is.nagoya-u.ac.jp/scpc/.

Contact: rkoike@is.nagoya-u.ac.jp

Supplementary information: Supplementary data are available at *Bioinformatics* online.

Received on June 23, 2011; revised on October 28, 2011; accepted on November 24, 2011

1 INTRODUCTION

Proteins are crucial molecules in biological phenomena that function via specific interactions with target molecules, including small ligand molecules, other proteins and nucleic acids. Such molecular interactions are involved in biological processes including various kinds of metabolism, signal transduction and transcriptional regulation. Numerous protein–protein interactions have been experimentally identified using high-throughput techniques, such as the yeast two-hybrid system (Fields and Song, 1989; Ito *et al.*, 2001; Uetz *et al.*, 2000), and the results are available in open public databases (Alfarano *et al.*, 2005; Aranda *et al.*, 2010; Ceol *et al.*, 2010; Keshava Prasad *et al.*, 2009; Salwinski *et al.*, 2004). In addition, structures of protein complexes have been determined at atomic resolution by X-ray crystallography or nuclear magnetic resonance, and this information has been deposited in the Protein Data Bank (PDB) (Berman *et al.*, 2007). Currently, the number of structural complexes in the PDB is rapidly increasing (Supplementary Fig. S1).

To process such large numbers of structural complexes and derive valuable biological insights from them, an automated method to compare the structures of protein complexes is required. Data processing of the huge amount of sequences and structures of biomacromolecules has been facilitated by automated comparison methods; for example, the Needleman–Wunsch algorithm (Needleman and Wunsch, 1970), FASTA (Pearson and Lipman, 1988), BLAST (Altschul *et al.*, 1997) and the S-Search (Smith and Waterman, 1981) for protein and nucleic acid sequences, and DALI (Holm and Sander, 1993), SSAP (Orengo and Taylor, 1996), VAST (Gibrat *et al.*, 1996), CE (Shindyalov and Bourne, 1998), MATRAS (Kawabata and Nishikawa, 2000) and TM-align (Zhang and Skolnick, 2005) for protein structures. These methods have also made significant contributions to various studies on the structure, function and evolution of biomacromolecules.

An improved understanding of many aspects of structural complexes and protein–protein interactions is obtained by employing automated comparison methods. One of these is the comparative analysis of structural complexes composed of evolutionary-related proteins. When proteins constitute a protein complex, their orthologous proteins also frequently form complexes. By applying an automated comparison method to these protein complexes, the evolutionary relationships between structure variation and sequence diversity can be revealed. Such a method also enables us to compare a query structural complex with each structure in databases and retrieve the structures that are similar to the query structure. This approach can be further extended to an exhaustive comparison among all existing structures of protein complexes. Similar structures are grouped into a cluster and a variety of clusters each consisting of similar complexes are subsequently obtained from the analysis. The resulting data present a comprehensive classification of protein complexes. Therefore, the development of an automated method to compare the structures of protein complexes would facilitate future investigations into protein complexes and protein–protein interactions.

In general, in the comparison of two structural complexes, each subunit of a complex exclusively corresponds to one subunit of the other complex. Exploring subunit correspondence is a unique process of comparing protein complexes that is not required for comparing protein structures. Currently, only a few methods have been proposed. Aloy *et al.* (2003) investigated structures of dimeric proteins, in which the subunit correspondence was provided in

*To whom correspondence should be addressed.

advance. To compare the relative orientation of two subunits of a dimer with that of another dimer, interaction root mean square deviation (iRMSD) was invented. By plotting iRMSD against sequence identity between subunits, they showed that interactions of closely related proteins were almost invariable, whereas those for structurally analogous proteins were not. Levy *et al.* (2006) simplified the structure of a protein complex to the form of a graph, in which a node and an edge indicated a subunit and the direct interaction of two subunits, respectively. Graphs were a suitable representation that could be compared and classified, and thus the subunit correspondence was explored easily. The results were summarized in the 3D Complex database (Levy *et al.*, 2006), which exhibits the classification of structural complexes. Mukherjee and Zhang (2009) developed a method, MultiMer-align (MM-align), as an extension of TM-align (Zhang and Skolnick, 2005). In MM-align, multiple subunits in each complex are connected in a head-to-tail manner, and two virtual monomeric chains are aligned so that one subunit in a complex corresponds to one in the other. All possible orders of subunits in virtual monomeric chains are evaluated. Note that similar ordering issues were addressed in protein structure comparisons if the structurally related parts in two proteins were located at sequentially different positions (Guerler and Knapp, 2008; Krissinel and Henrick, 2004; Yuan and Bystroff, 2005), as seen in circularly permuted proteins (Lindqvist and Schneider, 1997; Schmidt-Goenner *et al.*, 2010).

Our study aimed to develop a novel automated method, Structural Comparison of Protein Complexes (SCPC), to compare the structure of two protein complexes employing secondary structure elements (SSEs). SCPC divides complexes into subunits and detects structurally similar subunits from each of the complexes. Subunit similarity was explored using a heuristic approach that has previously been employed for the comparison of protein structures (Mizuguchi and Go, 1995), in which sequential orders of SSEs were considered. The structurally similar subunits are used to construct two similar substructures in complexes that show the subunit correspondence. The construction is independent of the orders of the subunits in the PDB files. SCPC was applied to three types of structural complexes: (i) homo- and heterodimers; (ii) homo-oligomers consisting of homologous subunits exhibiting different oligomeric states; and (iii) oxyhaemoglobins and deoxyhaemoglobins, to evaluate the performance of this method.

2 METHODS

2.1 Overview

Let us consider two structures of protein complexes, *A* and *B*, composed of *N* and *M* subunits, respectively. The subunits of *A* (or *B*) are labelled and represented as $a_1, a_2, \cdots, a_N$ (or $b_1, b_2, \cdots, b_M$). SCPC detects similar substructures within *A* and *B*, which are indicated by two sets of subunits. These sets are composed of the same number (at most the smaller of *N* and *M*) of subunits. Each subunit of one set exclusively corresponds to one subunit of the other set. A subunit and its corresponding subunit adopt similar structures and the two subunits are paired as a structurally-similar subunit pair (SSP). The two subunits are also located at the same position when two similar substructures are superimposed. Thus, the spatial arrangements of the subunits are similar. The one-to-one correspondence of the subunits is represented by an SSP set. For example, $\{(a_1, b_i), (a_2, b_j), \cdots, (a_M, b_k)\}$, where two subunits enclosed by a parenthesis is an SSP, and SSPs are arranged in ascending order of the subscript of a. SCPC produces a final set of SSPs and calculates the similarity score for the substructures.

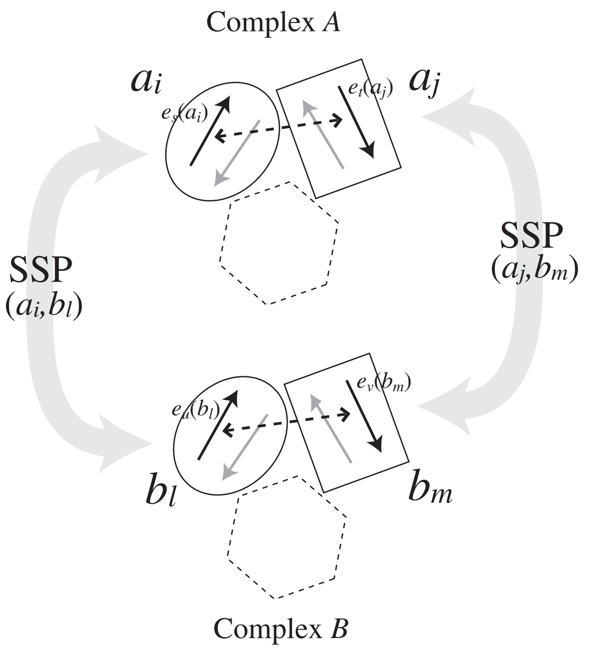

Fig. 1. Subunits and SSEs examined in the subunit arrangement comparison process. An ellipse, rectangle or hexagon indicates a subunit, and the same shape illustrates structurally similar subunits. Therefore, a_i and b_l (ellipses), or a_j and b_m (rectangles), are each SSPs. SSEs (arrows) in a_i and b_l, or a_j and b_m, are aligned structurally. For example, two SSEs, $e_s(a_i)$ and $e_u(b_l)$, or $e_t(a_j)$ and $e_v(b_m)$, are aligned. In comparing the spatial arrangement of a_i and a_j with that of b_l and b_m, we focused on two SSEs in the distinct subunits of a complex and their counterparts in another complex. For example, the spatial arrangement of $e_s(a_i)$ and $e_t(a_j)$ (linked with a dashed bidirectional arrow) was compared with that of $e_u(b_l)$ and $e_v(b_m)$. Spatial arrangement of every pair of SSEs in the distinct subunits was also considered.

SCPC is divided into two parts: the subunit comparison process and the subunit arrangement comparison process. In the subunit comparison process, a subunit in *A* is structurally compared with a subunit in *B* based on SSEs (Mizuguchi and Go, 1995). If SSEs of the subunits are structurally aligned, then the subunits are similar. Structurally similar subunits are stored as an SSP with the SSEs alignment. A pairwise comparison is then performed on all pairs of subunits from *A* and *B*. In the subunit arrangement comparison process, very large numbers of SSP sets are generated in a stepwise manner. During each step, SCPC adds an SSP to the set generated previously, evaluates the spatial arrangements of the subunits and retains the significant SSP sets. Finally, the best set of SSPs is obtained. Details of the subunit arrangement comparison process are presented in the following section. The subunit comparison process is described in the Supplementary Material because an existing method was employed (Mizuguchi and Go, 1995).

2.2 Subunit arrangement comparison process

2.2.1 Examination of a pair of SSPs The subunit arrangement comparison process involves the examination of a pair of SSPs, (a_i, b_l) and (a_j, b_m). SCPC compares the spatial arrangement of a_i and a_j in *A* with that of b_l and b_m in *B*, and scores the similarity between the two arrangements as $S_{SU}(a_i, a_j, b_l, b_m)$. To calculate $S_{SU}(a_i, a_j, b_l, b_m)$, we focused on two SSEs, the *s*-th SSE of a_i and the *t*-th SSE of a_j, denoted as $e_s(a_i)$ and $e_t(a_j)$, respectively. Since SSEs were aligned and stored in the subunit comparison process (Supplementary Material), the counterparts of $e_s(a_i)$ and $e_t(a_j)$ were already determined [$e_u(b_l)$ and $e_v(b_m)$, respectively; Fig. 1]. The spatial arrangement of $e_s(a_i)$ and $e_t(a_j)$ is compared with that of $e_u(b_l)$ and $e_v(b_m)$, and the similarity is scored as $S_{SSE}(e_s(a_i),e_t(a_j),e_u(b_l),e_v(b_m))$ (We modified the existing score to evaluate the structural similarity of SSEs and defined the score S_{SSE}. See Supplementary Material for details.) The similarity score $S_{SU}(a_i,a_j,b_l,b_m)$ was defined as,

$$S_{SU}(a_i,a_j,b_l,b_m)=\sum_{(s,u)\in I(a_i,b_l)}\sum_{(t,v)\in I(a_j,b_m)} S_{SSE}(e_s(a_i),e_t(a_j),e_u(b_l),e_v(b_m)), \tag{1}$$

where $I(a_i,b_l)$ is the set of indices that specifies aligned SSEs of a_i and b_l, and the SSEs not observed in the alignments are discarded.

2.2.2 Total score of a set of SSPs To evaluate a set of SSPs, the total score S_{total} was introduced in SCPC. A set of N SSPs is depicted as $\Gamma_N=\{(a_{x(1)},b_{y(1)}),(a_{x(2)},b_{y(2)}),\cdots,(a_{x(i)},b_{y(i)}),\cdots,(a_{x(N)},b_{y(N)})\}$, where $x(i)$ and $y(i)$ indicate subunit labels of the i-th SSP. We defined the total score of Γ_N as,

$$\begin{aligned}S^N_{total}&=\sum_{i=1}^{N-1}\sum_{j>i}^{N}S_{SU}(a_{x(i)},a_{x(j)},b_{y(i)},b_{y(j)})\\&=\sum_{SSPpair}S_{SU}(a_{x(i)},a_{x(j)},b_{y(i)},b_{y(j)})\end{aligned}, \tag{2}$$

where $\Sigma_{\mathrm{SSP_{pair}}}$ represents the summation of all combinations of two SSPs in Γ_N. As clearly shown by Equation (2), the similarity scores of two subunit arrangements, S_{SU} s, constitute the total score, S_{total}.

2.2.3 Exploring the best set of SSPs When SCPC explores the best set of SSPs that maximizes S_{total}, SSP sets are produced in a stepwise manner and their total scores are evaluated. Initially, all sets of two SSPs are generated, their total scores are calculated using Equation (2) and the best 500 sets are retained. At the next step, possible SSPs are added to each of the retained sets. A number of sets of three SSPs are generated and their total scores are calculated. Sets of three SSPs are subsequently discarded if the addition of an SSP does not improve the total score. SCPC assembles the new sets (sets of three SSPs) and previously retained sets (sets of two SSPs), and selects the best 500 sets. This step is repeated until the addition of a new SSP is not possible. Finally, SCPC decides the best set of SSPs among all the calculated sets. This set presents the two most similar substructures in the protein complexes.

2.3 *R*-score

The total score, S_{total}, is additive, and thus largely depends on the protein complex size. To reduce the size dependency of the similarity score, we introduced the scaled score, *R*-score (Kawabata, 2003). As a result of the structure comparison between complexes *A* and *B*, the total score, $S_{total}(A,B)$, is obtained. The *R*-score is defined as,

$$R(A,B)=100\times\frac{S_{total}(A,B)-S_{\min}}{S_{\max}-S_{\min}}, \tag{3}$$

where $S_{\max}$ and $S_{\min}$ are,

$$S_{\max}=\frac{S_{total}(A,A)+S_{total}(B,B)}{2}$$

$$S_{\min}=0$$

$S_{total}(A,A)$ and $S_{total}(B,B)$ are the total scores obtained by self-comparisons.

3 RESULTS

3.1 Comparison of homo- and heterodimers

First, we compared Ulp1/Smt3 (PDB ID: 1euv) (Mossessova and Lima, 2000) and Den1/Nedd8 (1xt9) (Reverter *et al.*, 2005) as examples of heterodimers. SCPC produced a set of SSPs, {(1euvA, 1xt9A), (1euvB, 1xt9B)}, where A and B were the chain identifiers. Ulp1 (1euvA) and Smt3 (1euvB) corresponded to Den1 (1xt9A) and Nedd8 (1xt9B), respectively. The SSEs of the corresponding subunits were structurally aligned with SCPC (Fig. 2a). Aligned SSEs were represented by the centres of their Cα atoms and the structures were superimposed (Fig. 2b). The total score S_{total} was also obtained and converted to the *R*-score (50.4).

To examine the performance of SCPC, we prepared a benchmark set composed of homo- and heterodimers. This set is a subset of Mukherjee and Zhang's set that was constructed to evaluate MM-align (Mukherjee and Zhang, 2009). The set consists of two groups of dimers; the first group is the set of queries (105 dimers) and the second group (3662 dimers) is the database used for the similarity search in structures. The detailed process of dataset construction is described in the Supplementary Material. The examination of Mukherjee and Zhang provides a list of dimer pairs, a query and its closest structural neighbour (CSN) in the second group identified by MM-align based on the similarity score (TM-score). We applied SCPC to the 105 pairs. The *R*-score and the TM-score of the pairs are plotted in Figure 2c. The two values correlate well with a correlation coefficient of 0.78, indicating that SCPC can also evaluate the structural similarity between the queries and the CSNs, as calculated by MM-align.

In the similarity search of the database by SCPC, the CSNs are not always the same as that of MM-align. The CSNs were identical for 62 queries (red crosses in Fig. 2c); however, SCPC and MM-align proposed different CSNs for 43 queries (black crosses). An additional measure was introduced to examine which structure was better. We divided the dimers of a query and its CSN into subunits, and aligned the corresponding subunits independently using MATRAS (Kawabata and Nishikawa, 2000). Two residue-wise structural alignments were subsequently obtained for the smaller and larger subunits. According to the alignment of the larger subunits, the two complexes were superimposed, and we calculated the RMSD of the aligned Cα atoms of the smaller subunits (referred to as the ligand RMSD). The ligand RMSD values are plotted in Figure 2d for 41 queries (the two omitted plots are described in the Supplementary Material). This plot shows that both CSNs from SCPC and MM-align exhibit almost the same similarities to the queries. The difference between ligand RMSDs is only 0.05 Å on average and SCPC proposed better CSNs for 15 queries. All the data on the 105 queries and their CSNs are summarized in Supplementary Table S1. Consequently, SCPC and MM-align identified the same or comparable structural neighbours for the 103 query dimers. This demonstrates that SCPC performs well as MM-align in the structural comparison of dimers.

3.2 Comparison of homo-oligomers composed of homologous proteins

While many homologous proteins adopt similar structures, molecular functions and binding modes, some also assemble into quite different protein complexes using distinct binding modes (Akiva *et al.*, 2008; Hashimoto and Panchenko, 2010; Nishi *et al.*, 2011; Nishi and Ota, 2010). Structural comparisons of such protein complexes are important in understanding protein–protein interactions and protein oligomerization processes. In this section, we focus on N-carbamoyl-D-amino acid hydrolase (2ggk) (Chiu *et al.*, 2006) and aliphatic amidase (2uxy) (Andrade *et al.*, 2007). These proteins are homologous, because their sequences are

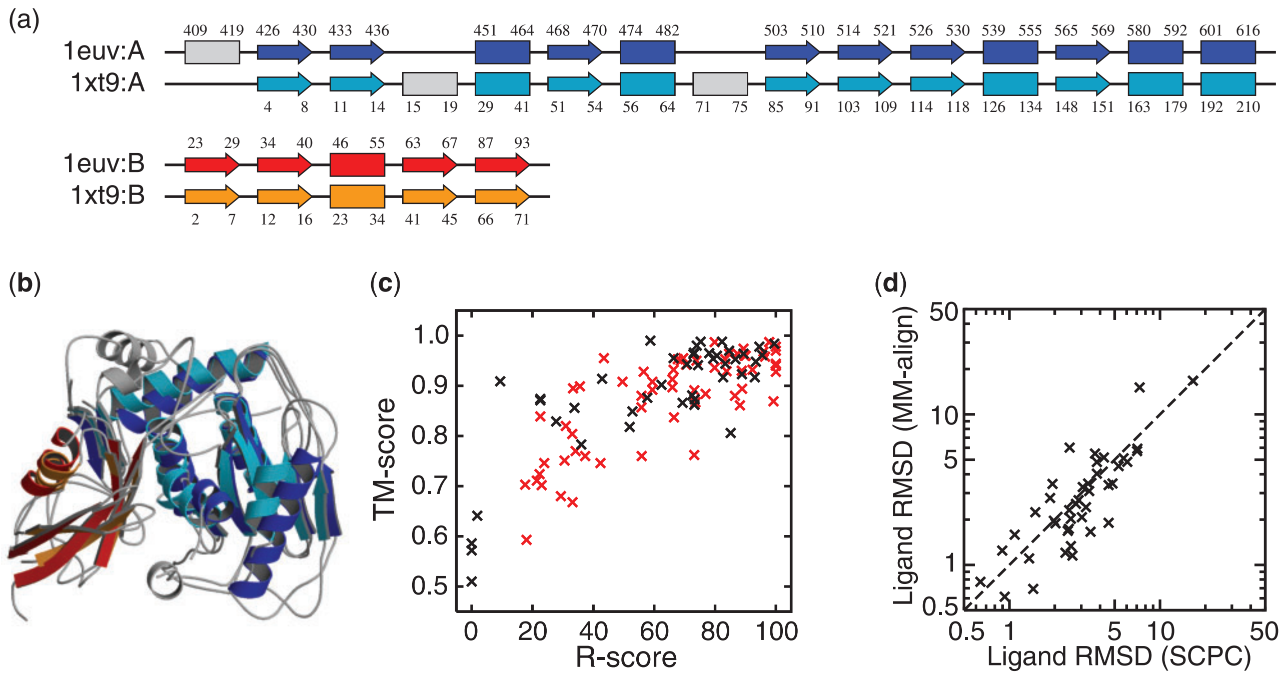

Fig. 2. Structural comparison of dimers by SCPC. (**a**) SSE alignment of Ulp1/Smt3 and Den1/Nedd8. A helix and a strand are illustrated as a rectangle and an arrow, respectively. N- and C-terminal residue numbers of an SSE are presented. Ulp1, Den1, Smt3 and Nedd8 are indicated in blue, cyan, red and orange, respectively. Unaligned SSEs are indicated in grey. (**b**) Superposition of Ulp1/Smt3 and Den1/Nedd8. Structural complexes were drawn with Molscript (Kraulis, 1991). SSEs are coloured in the same manner as (a). (**c**) The *R*-score and the TM score for 105 queries and their CSNs of MM-align. (**d**) Structural similarities of the queries and their CSNs proposed by SCPC and MM-align. The ligand RMSD value is plotted on a logarithmic axis.

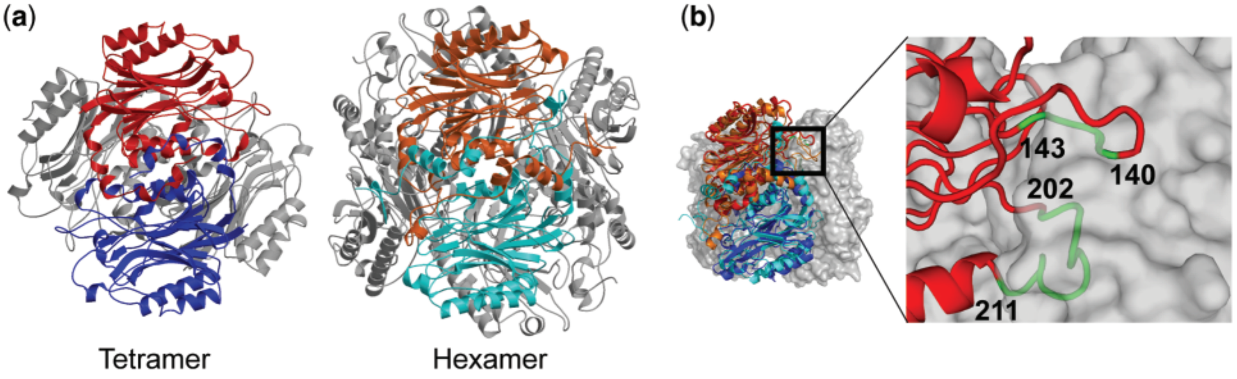

Fig. 3. Structural similarity of *N*-carbamoyl-D-amino acid hydrolase and aliphatic amidase. (**a**) Structures of the complexes of *N*-carbamoyl-D-amino acid hydrolase and aliphatic amidase. Only non-grey subunits were found to be similar. *N*-carbamoyl-D-amino acid hydrolase subunits are shown in red (2ggkA) and blue (2ggkB), and those of aliphatic amidase are shown in orange (2uxyA) and cyan (2uxyE). (**b**) Superposition of similar substructures of *N*-carbamoyl-D-amino acid hydrolase and aliphatic amidase. Other subunits of aliphatic amidase are shown as grey surfaces (2uxyBCDF). The two regions, residues 140-143 and 202-211, of the red subunit (2ggkA) are shown in light green. The detailed view of these regions is shown in the right panel, in which the blue, orange and cyan subunits are omitted. Similar steric clashes between the loops and the subunit were also observed in 2ggkB and 2uxyD.

similar [BLAST *E*-value: 10^{-10} (Altschul *et al.*, 1997)]. However, *N*-carbamoyl-D-amino acid hydrolase and aliphatic amidase exist in homo-tetrameric and homo-hexameric forms, respectively. When they were compared, SCPC detected similar substructures composed of two subunits, depicted by {(2ggkA, 2uxyA), (2ggkB, 2uxyE)} (Fig. 3a). We further observed that a subunit of the homo-tetramer interacts with the other three subunits via different interfaces and that each interaction represents a binding mode.

The two subunits in the detected set of SSPs (2ggkAB) represented the binding mode that gave the largest interface area (2475 Å^2) among all three modes. Similarly, the two detected subunits in the homo-hexamer (2uxyAE) also represented the binding mode that gave the largest interface area (4581 Å^2). In this case, the homologous proteins maintain the binding mode that gives the largest interface area, as previously reported (Levy *et al.*, 2008). Since the structures of homo-oligomers are essentially symmetrical, we can observe almost the same binding modes in different dimers of the homo-oligomers (e.g. 2ggkCD and 2uxyBD). These dimers were identified by suboptimal solutions, whereas the other dimers that presented distinct binding modes (e.g. 2ggkAC and 2uxyAB) were not detected.

Based on superimposing similar substructures that were detected (Fig. 3b), we investigated why the oligomeric states of the two proteins were different. We found that four inserted loop regions, residues 140-143 and 202-211 in two subunits of the homo-tetramer (2ggkAB), were protruding and abolished the interaction with the subunits of the homo-hexamer (2uxyCB, see the right panel). These steric hindrances potentially inhibit the *N*-carbamoyl-D-amino acid hydrolase from forming a homo-hexamer (Akiva *et al.*, 2008; Hashimoto and Panchenko, 2010). Accordingly, SCPC successfully identified similar substructures from different protein complexes, thereby offering valuable insight into protein interactions.

3.3 Classification of haemoglobin quaternary structures

Haemoglobin (Hb) is a well-known oligomeric protein that forms a tetramer composed of two α-subunits and two β-subunits. Each subunit contains a haem group, to which a ligand molecule (oxygen or carbon monoxide) binds. Hb exists in two states that exhibit different ligand affinities; the R state with a high affinity and the T state with a low affinity. These two states differ in their quaternary structures (Baldwin and Chothia, 1979; Perutz, 1972). We collected 18 Hb structures from the PDB and protein quaternary structure file server (PQS) (Henrick and Thornton, 1998) (Supplementary Material). Eight of these structures contain bound carbon monoxide (CO-Hb), whereas three structures contain bound oxygen (oxy-Hb). The other seven structures do not include any ligands (deoxy-Hb). Ligand-bound Hb (CO-Hb and oxy-Hb) and deoxy-Hb are usually in the R and T states, respectively. We applied SCPC to the 18 Hb quaternary structures. Every pair of Hb structures was examined and a set of four SSPs was obtained for each pair. This result revealed that all subunits of a given Hb correspond to all subunits of any Hb. Structural differences between Hbs were not large, because all *R*-scores were >36.0 (see Figs. 2c and 2d for the significance of the *R*-score). Average-linkage clustering (equivalent to the UPGMA method) was also performed according to the *R*-scores (Fig. 4a). The dendrogram revealed that the Hb structures were divided into two groups, Hbs with bound ligands and deoxy-Hbs, with only two exceptions that are described below.

To clarify the structural difference between Hb with bound ligands and deoxy-Hb, we compared human oxy-Hb (2dn1) and deoxy-Hb (2dn2) (Park *et al.*, 2006). We obtained the set {(2dn1A, 2dn2A), (2dn1B, 2dn2B), (2dn1C, 2dn2C), (2dn1D, 2dn2D)} using these proteins with an *R*-score of 67.6. We further examined the spatial arrangements of the two subunits to analyse the structural differences in more detail. Two SSPs were chosen from the set to calculate the S_{SU} [Equation (1)]. All S_{SU} scores are summarized in Figure 4b. The spatial arrangement of an α-subunit and a β-subunit of the oxy-Hb (2dn1CD) was closest to that of the deoxy-Hb (2dn2CD), and the S_{SU} was found to be the largest (see α_2–β_2 in Fig. 4b).

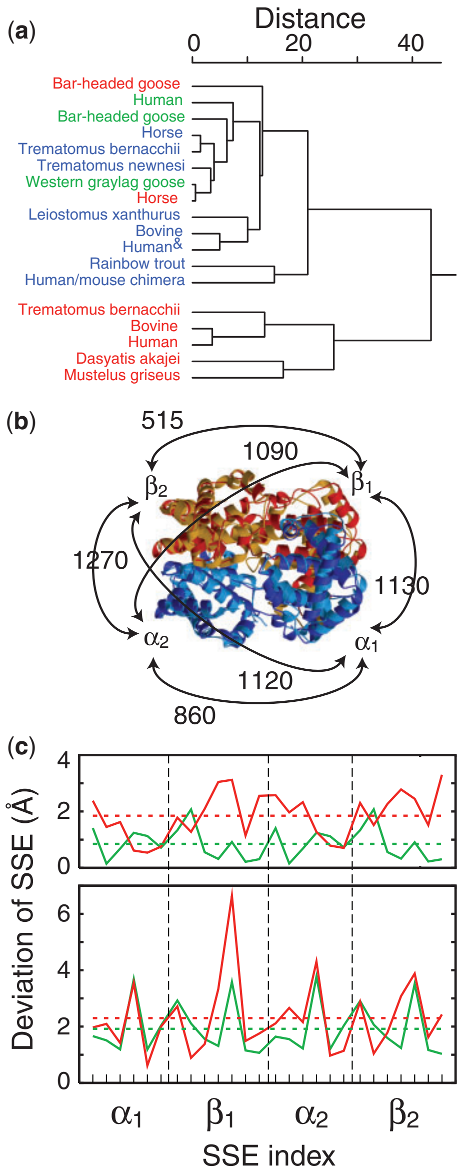

Fig. 4. Classification of Hb quaternary structures. (**a**) A dendrogram showing the similarity of 18 Hb structures. Using the *R*-score (*R*), the distance of the dendrogram, *D*, is given as $D = 100 - R$. Each Hb is indicated by its species name. Human embryonic Hb is indicated by &. CO-Hb, oxy-Hb and deoxy-Hb are blue, green and red, respectively. (**b**) Superposition of the structures of the human oxy-Hb (2dn1) and deoxy-Hb (2dn2) forms. The α-subunits and the β-subunits of the oxy-Hb are blue and red, and those of the deoxy-Hb are cyan and orange, respectively. Structures are superimposed using the structures of α_2 and β_2 (2dn1CD and 2dn2CD). S_{SU} scores are indicated at the middle of the arrows connecting the two subunits. (**c**) Deviation of two atypical Hbs, horse Hb (upper panel) and bar-headed goose Hb (lower panel), from human oxy-Hb and deoxy-Hb. Each atypical structure was superimposed with both structures of human Hbs. Green and red lines indicate the deviations of the centres of the corresponding SSEs from the human oxy-Hb structure and the human deoxy-Hb structure, respectively. The average deviations are shown as dashed lines. SSE indices at the horizontal axis specify the positions of helices from the N-terminus in a subunit.

The arrangements of the other α-subunits and β-subunits (2dn1AB and 2dn2AB, α_1-β_1) were also similar, and the S_{SU} was the second largest. This result indicates that the spatial arrangements of an α-subunit and its counterpart β-subunit, i.e. α_1 and β_1, or α_2 and β_2, are quite similar in structure for oxy-Hb and deoxy-Hb, as indicated by the careful analysis of Baldwin and Chothia (1979). It also suggests that the heterodimer, α_1 and β_1, or α_2 and β_2, is rigid. Therefore, the difference in the quaternary structures of Hbs with a bound ligand and deoxy-Hbs can be described based on the movement of these two rigid dimers (Fig. 4b).

In the dendrogram (Fig. 4a), the two deoxy-Hbs from horse (1ibe) (Wilson *et al.*, 1996) and bar-headed goose (1hv4) (Liang *et al.*, 2001) were classified into the Hb cluster with bound ligands. The horse Hb has an oxidized iron ion associated with its haem group. This kind of Hb is known to adopt the quaternary structure of the high-affinity (R) state (Wilson *et al.*, 1996) and it is therefore reasonable that this Hb structure is classified into the Hb cluster with bound ligands. The bar-headed goose Hb is known to have high oxygen affinity. The difference in quaternary structure between the oxy-Hb and the deoxy-Hb forms isolated from the bar-headed goose was reported to be much smaller than that observed in the corresponding human structures (Liang *et al.*, 2001), supporting the classification of the bar-headed goose deoxy-Hb. The structures of the two exceptions were superimposed with the structures of the human oxy-Hb and deoxy-Hb, and the deviations from the centres of the corresponding SSEs are shown in Figure 4c. This result shows that these two structures are closest to the structure of human oxy-Hb, rather than the structure of human deoxy-Hb. Consequently, SCPC is capable of accurately detecting the structural differences between Hbs due to ligand binding.

4 CONCLUSIONS

We have developed SCPC, a novel method to compare the structures of protein complexes. SCPC was initially applied to a dataset of homo- and heterodimers that was used to evaluate an existing method, MM-align. The results were examined and the performances of MM-align and SCPC were compared. SCPC was further applied to homo-oligomers, which do not have unique oligomeric states, despite the apparent homology of their subunits. Additional examples have been previously reported (Akiva *et al.*, 2008; Hashimoto and Panchenko, 2010; Nishi *et al.*, 2011; Nishi and Ota, 2010) and their structural comparison would be potentially useful for clarifying the evolution of protein complexes, as supported by the evolutionarily conserved binding mode identified for these homo-oligomers. Finally, SCPC was applied to the well-studied Hb and the results were compared with the conventional classification. Since Hbs are tetrameric proteins, the application of our method to higher order protein complexes was also considered. Satisfactory results were obtained for all tested cases. In addition, SCPC is fast. For example, in the structural comparison of bacterial ribosomes from *Escherichia coli* (2qbd) and *Thermus thermophilus* (1fjg), each of which represents one of the largest complexes in the PDB, the calculation time was only 6.3 s on a 2.5 GHz AMD Opteron processor. These results indicate that SCPC is an effective and efficient tool for characterizing structures of protein complexes.

As described in Section 1, the subunit correspondence is crucial in the structural comparison of protein complexes. We have observed that SCPC is suitable for the analysis of domain fusions or fissions in multi-domain proteins if the correspondence between domains should be examined rather than the subunits. Such enhancements will make SCPC more powerful and useful.

Funding: Institute for Bioinformatics Research and Development, Japan Science and Technology Agency; Ministry of Education, Culture, Sports Science and Technology, Japan.

Conflict of Interest: none declared.

REFERENCES

Akiva,E. *et al.* (2008) Built-in loops allow versatility in domain-domain interactions: lessons from self-interacting domains. *Proc. Natl Acad. Sci. USA*, **105**, 13292–13297.

Alfarano,C. *et al.* (2005) The Biomolecular Interaction Network Database and related tools 2005 update. *Nucleic Acids Res.*, **33**, D418–D424.

Aloy,P. *et al.* (2003) The relationship between sequence and interaction divergence in proteins. *J. Mol. Biol.*, **332**, 989–998.

Altschul,S.F. *et al.* (1997) Gapped BLAST and PSI-BLAST: a new generation of protein database search programs. *Nucleic Acids Res.*, **25**, 3389–3402.

Andrade,J. *et al.* (2007) Structure of amidase from Pseudomonas aeruginosa showing a trapped acyl transfer reaction intermediate state. *J. Biol. Chem.*, **282**, 19598–19605.

Aranda,B. *et al.* (2010) The IntAct molecular interaction database in 2010. *Nucleic Acids Res.*, **38**, D525–D531.

Baldwin,J. and Chothia,C. (1979) Haemoglobin: the structural changes related to ligand binding and its allosteric mechanism. *J. Mol. Biol.*, **129**, 175–220.

Berman,H. *et al.* (2007) The worldwide Protein Data Bank (wwPDB): ensuring a single, uniform archive of PDB data. *Nucleic Acids Res.*, **35**, D301–D303.

Ceol,A. *et al.* (2010) MINT, the molecular interaction database: 2009 update. *Nucleic Acids Res.*, **38**, D532–D539.

Chiu,W.C. *et al.* (2006) Structure-stability-activity relationship in covalently cross-linked N-carbamoyl D-amino acid amidohydrolase and N-acylamino acid racemase. *J. Mol. Biol.*, **359**, 741–753.

Fields,S. and Song,O. (1989) A novel genetic system to detect protein-protein interactions. *Nature*, **340**, 245–246.

Gibrat,J.F. *et al.* (1996) Surprising similarities in structure comparison. *Curr. Opin. Struct. Biol.*, **6**, 377–385.

Guerler,A. and Knapp,E.W. (2008) Novel protein folds and their nonsequential structural analogs. *Protein Sci.*, **17**, 1374–1382.

Hashimoto,K. and Panchenko,A.R. (2010) Mechanisms of protein oligomerization, the critical role of insertions and deletions in maintaining different oligomeric states. *Proc. Natl Acad. Sci. USA*, **107**, 20352–20357.

Henrick,K. and Thornton,J.M. (1998) PQS: a protein quaternary structure file server. *Trends Biochem. Sci.*, **23**, 358–361.

Holm,L. and Sander,C. (1993) Protein structure comparison by alignment of distance matrices. *J. Mol. Biol.*, **233**, 123–138.

Ito,T. *et al.* (2001) A comprehensive two-hybrid analysis to explore the yeast protein interactome. *Proc. Natl Acad. Sci. USA*, **98**, 4569–4574.

Kawabata,T. (2003) MATRAS: a program for protein 3D structure comparison. *Nucleic Acids Res.*, **31**, 3367–3369.

Kawabata,T. and Nishikawa,K. (2000) Protein structure comparison using the markov transition model of evolution. *Proteins*, **41**, 108–122.

Keshava Prasad,T.S. *et al.* (2009) Human Protein Reference Database–2009 update. *Nucleic Acids Res.*, **37**, D767–D772.

Kraulis,P.J. (1991) MOLSCRIPT: a program to produce both detailed and schematic plots of protein structures. *J. Appl. Crystallogr.*, **24**, 946–950.

Krissinel,E. and Henrick,K. (2004) Secondary-structure matching (SSM), a new tool for fast protein structure alignment in three dimensions. *Acta Crystallogr. D Biol. Crystallogr.*, **60**, 2256–2268.

Levy,E.D. *et al.* (2006) 3D complex: a structural classification of protein complexes. *PLoS Comput. Biol.*, **2**, e155.

Levy,E.D. *et al.* (2008) Assembly reflects evolution of protein complexes. *Nature*, **453**, 1262–1265.

Liang,Y. *et al.* (2001) The crystal structure of bar-headed goose hemoglobin in deoxy form: the allosteric mechanism of a hemoglobin species with high oxygen affinity. *J. Mol. Biol.*, **313**, 123–137.

Lindqvist,Y. and Schneider,G. (1997) Circular permutations of natural protein sequences: structural evidence. *Curr. Opin. Struct. Biol.*, **7**, 422–427.

Mizuguchi,K. and Go,N. (1995) Comparison of spatial arrangements of secondary structural elements in proteins. *Protein Eng.*, **8**, 353–362.

Mossessova,E. and Lima,C.D. (2000) Ulp1-SUMO crystal structure and genetic analysis reveal conserved interactions and a regulatory element essential for cell growth in yeast. *Mol. Cell*, **5**, 865–876.

Mukherjee,S. and Zhang,Y. (2009) MM-align: a quick algorithm for aligning multiple-chain protein complex structures using iterative dynamic programming. *Nucleic Acids Res.*, **37**, e83.

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

Nishi,H. and Ota,M. (2010) Amino acid substitutions at protein-protein interfaces that modulate the oligomeric state. *Proteins*, **78**, 1563–1574.

Nishi,H. *et al.* (2011) Cover and spacer insertions: small nonhydrophobic accessories that assist protein oligomerization. *Proteins*, **79**, 2372–2379.

Orengo,C.A. and Taylor,W.R. (1996) SSAP: sequential structure alignment program for protein structure comparison. *Methods Enzymol.*, **266**, 617–635.

Park,S.Y. *et al.* (2006) 1.25 Å resolution crystal structures of human haemoglobin in the oxy, deoxy and carbonmonoxy forms. *J. Mol. Biol.*, **360**, 690–701.

Pearson,W.R. and Lipman,D.J. (1988) Improved tools for biological sequence comparison. *Proc. Natl Acad. Sci. USA*, **85**, 2444–2448.

Perutz,M.F. (1972) Nature of haem-haem interaction. *Nature*, **237**, 495–499.

Reverter,D. *et al.* (2005) Structure of a complex between Nedd8 and the Ulp/Senp protease family member Den1. *J. Mol. Biol.*, **345**, 141–151.

Salwinski,L. *et al.* (2004) The Database of Interacting Proteins: 2004 update. *Nucleic Acids Res.*, **32**, D449–D451.

Schmidt-Goenner,T. *et al.* (2010) Circular permuted proteins in the universe of protein folds. *Proteins*, **78**, 1618–1630.

Shindyalov,I.N. and Bourne,P.E. (1998) Protein structure alignment by incremental combinatorial extension (CE) of the optimal path. *Protein Eng.*, **11**, 739–747.

Smith,T.F. and Waterman,M.S. (1981) Identification of common molecular subsequences. *J. Mol. Biol.*, **147**, 195–197.

Uetz,P. *et al.* (2000) A comprehensive analysis of protein-protein interactions in Saccharomyces cerevisiae. *Nature*, **403**, 623–627.

Wilson,J. *et al.* (1996) The crystal structure of horse deoxyhaemoglobin trapped in the high-affinity (R) state. *J. Mol. Biol.*, **264**, 743–756.

Yuan,X. and Bystroff,C. (2005) Non-sequential structure-based alignments reveal topology-independent core packing arrangements in proteins. *Bioinformatics*, **21**, 1010–1019.

Zhang,Y. and Skolnick,J. (2005) TM-align: a protein structure alignment algorithm based on the TM-score. *Nucleic Acids Res.*, **33**, 2302–2309.