Novel protein folds and their nonsequential structural analogs

AYSAM GUERLER AND ERNST-WALTER KNAPP
Department of Chemistry and Biochemistry, Freie Universität Berlin, 14195, Berlin, Germany
(RECEIVED March 20, 2008; FINAL REVISION May 1, 2008; ACCEPTED May 2, 2008)

Abstract

Newly determined protein structures are classified to belong to a new fold, if the structures are sufficiently dissimilar from all other so far known protein structures. To analyze structural similarities of proteins, structure alignment tools are used. We demonstrate that the usage of nonsequential structure alignment tools, which neglect the polypeptide chain connectivity, can yield structure alignments with significant similarities between proteins of known three-dimensional structure and newly determined protein structures that possess a new fold. The recently introduced protein structure alignment tool, GANGSTA, is specialized to perform nonsequential alignments with proper assignment of the secondary structure types by focusing on helices and strands only. In the new version, GANGSTA+, the underlying algorithms were completely redesigned, yielding enhanced quality of structure alignments, offering alignment against a larger database of protein structures, and being more efficient. We applied DaliLite, TM-align, and GANGSTA+ on three protein crystal structures considered to be novel folds. Applying GANGSTA+ to these novel folds, we find proteins in the ASTRAL40 database, which possess significant structural similarities, albeit the alignments are nonsequential and in some cases involve secondary structure elements aligned in reverse orientation. A web server is available at http://agknapp.chemie.fu-berlin.de/gplus for pairwise alignment, visualization, and database comparison.

Keywords: nonsequential protein structure alignment; novel protein fold; protein fold space; protein structure/folding

Supplemental material: see www.proteinscience.org

The specific biochemical abilities of a protein result from its three-dimensional (3D) native structure. For enzymes, the structure optimizes the geometric arrangement of catalytically active amino acid side chains and cofactors, and simultaneously allows efficient access and removal of educts and products, respectively. For proteins, where one of the functions is to form specific complexes with other proteins, the shape of the contact surface and the residue pair interactions in the contact surface are also relevant (Shulman-Peleg et al. 2007). Consequently, proteins from different species that perform the same function often possess the same structures and the same key residues. However, there are exceptions where nature uses alternatively designed protein 3D structures with equivalent or different key residues and cofactors to perform the same function in different species.

Under physiological conditions, the native 3D structure of a protein is determined solely by the primary sequence (Anfinsen et al. 1961). On the other hand, the native 3D structure does not belong to a unique primary sequence. Mutational studies demonstrated that often, only a small fraction of amino acids is crucial to define and stabilize the 3D structures of proteins (Guo et al. 2004; Russ et al. 2006). Consequently, only structurally and functionally relevant residues of a protein are conserved among

Reprint requests to: Ernst-Walter Knapp, Department of Chemistry and Biochemistry, Freie Universität Berlin, Fabeckstrasse 36A, 14195 Berlin, Germany; e-mail: knapp@chemie.fu-berlin.de; fax: 49-30-838-56921.

Article published online ahead of print. Article and publication date are at http://www.proteinscience.org/cgi/doi/10.1110/ps.035469.108.

different species. This sequence conservation is used to assess the unknown function of proteins by sequence comparison, which may fail if the sequence homology is too low. In case the protein 3D structure is available, structure comparison can be more useful to assess the protein function, since the universe of protein structures is much smaller than the universe of protein sequences. The number of different protein folds is estimated to be about 1000 only (Wang 1998; Leonov et al. 2003). But there are also less optimistic views expecting a much larger number of distinct protein folds of 4000 or even 8000 (Govindarajan et al. 1999; Grant et al. 2004; Liu et al. 2004).

Sequence comparison as a tool to identify protein function is now well established and is sufficiently reliable if the sequence similarity is higher than 40%, but the results of sequence comparisons become increasingly uncertain with lower sequence identity (Whisstock and Lesk 2003; Lee et al. 2007). Since not only sequence but also structure similarity of proteins correlates with their function, structure comparison of proteins is most useful to characterize a protein of yet unknown function, if its 3D structure is available (Lee et al. 2007). This approach can be particularly successful, since at present the Protein Data Bank (PDB) (Berman et al. 2000) already contains a considerable fraction of the universe of protein folds to predict the structures of soluble proteins (Kolodny et al. 2005; Zhang and Skolnick 2005a).

Protein 3D structure comparison is still a challenging task and depends critically on the alignment algorithm, the similarity measure, and the fractions of the protein structures considered for the pairwise structure alignment (Kolodny et al. 2005). An actual but still incomplete listing of available methods for protein structure alignment can be found on the web page http://en.wikipedia.org/wiki/Structural_alignment_software, already containing more than 40 different programs, for example, DaliLite (Holm and Park 2000) and TM-align (Zhang and Skolnick 2005b).

To identify a potential new fold, the considered protein structure must be aligned to all representative protein structures of the PDB. Suitable databases of representative protein structures are the ASTRAL databases provided by SCOP (Chandonia et al. 2004). These databases contain subsets of the PDB with domain structures whose sequence similarities are below a threshold value of, say, 40% or 70% sequence identity. In the past, the structure alignment methods used to identify the same fold in a database were often restricted or biased considering protein structures possessing the same connectivity of secondary structure elements (SSEs) (i.e., α-helices and β-strands) as defined by the polypeptide chain. Only a few methods are available that allow for nonsequential protein structure alignments, for example, MASS (Dror et al. 2003), TOPOFIT (Ilyin et al. 2004), SCALI (Yuan and Bystroff 2005), and others (Szustakowski and Weng 2000, 2002; Shatsky et al. 2002; Shih and Hwang 2004; Chen et al. 2006). Recently, the program GANGSTA (Kolbeck et al. 2006) appeared, which ignores the loops connecting different SSEs, like, for example, MASS (Dror et al. 2003), PRISM (Yang and Honig 1999), and SARF (Alexandrov and Fischer 1996), and allows nonsequential protein structure alignment. In addition, it offers alignment of yet unknown protein structures against a database of more than 3000 domains of protein structures with <40% sequence identity. Since GANGSTA is relatively slow, we have redesigned it completely. GANGSTA+ is more than a factor of 10 faster, yields alignments of higher quality, and offers alignments of arbitrary protein structures against the complete ASTRAL40 (1.71) database containing about 7500 domains of protein structures with <40% sequence identity of the corresponding polypeptides (Chandonia et al. 2004).

Here, we like to demonstrate that protein folds, which are considered to be new, appear to be known folds if one considers only the topological arrangement of the SSEs and disregards the connectivity of the polypeptide chain defined by the loops connecting the SSEs. GANGSTA+ is also capable of performing structure alignments where the SSEs can be aligned in reverse orientation; that is, aligned SSE pairs are of the same type, but oriented such that the C-terminal end of one SSE is superimposed on the N-terminal end of the other SSE. This can be used to enhance the likeliness of finding similar structures for a given protein structure. We like to point out that this study can also be performed with several other protein structure alignment programs mentioned above. We use GANGSTA+ in this application since with our own method we have the procedures better under control. To contrast nonsequential with sequential alignment results, we applied TM-align (Zhang and Skolnick 2005b), which works exclusively sequentially.

Results

Large-scale database comparison

We applied TM-align (Zhang and Skolnick 2005b) and GANGSTA+ on all protein pairs from the ASTRAL40 (SCOP 1.71) database of protein structure alignment (DPSA) (see Materials and Methods) and evaluated the TM-score, Equation 3, of all protein pairs, obtained with TM-align and with GANGSTA+. Figure 1 shows the distribution of the highest TM-scores obtained by structure alignment of each protein of the DPSA with respect to the whole DPSA. The results illustrate that both methods yield comparable protein structure alignments. For high TM-scores, the score distribution generated with GANGSTA+ is slightly below that of TM-align, which is

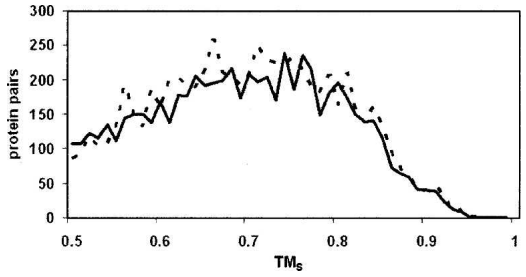

Figure 1. TM-score, Equation 3, distribution of the highest-scoring protein structure alignments generated by (dashed line) TM-align (Zhang and Skolnick 2005b) and (solid line) GANGSTA+ for each DPSA entry with respect to whole DPSA.

understandable, since GANGSTA+ optimizes the SAS but not the TM-score. Moreover, GANGSTA+ is a nonsequential protein structure alignment tool, solving a problem of higher complexity. Both methods operate in the time range of less than a second CPU time per protein pair on an AMD/OPTERON with 1600 MHz.

Nonsequential structure similarities of novel folds with the ASTRAL40 protein structure database

In 2007, Koo et al. published the crystal structure of the hypothetical protein TA0956 (**2JMK**) from *Thermoplasma acidophilum*, which was stated to possess a new fold (Koo et al. 2007). Database searches with TM-align (Zhang and Skolnick 2005b) and DaliLite (Holm and Park 2000) did not yield significant structural similarities to other proteins in the ASTRAL40 database (SCOP 1.71) (DPSA). In contrast to DaliLite and TM-align, GANGSTA+ was able to detect several nonsequential structure alignments with complete assignment of all seven SSEs of **2JMK** by scanning the whole DPSA in ~72 min (i.e., about 0.6 s per protein pair) on an AMD/OPTERON with 1600 MHz CPU. The structure alignments of all three considered programs were evaluated by accounting the number of aligned residues and the RMSD. However, GANGSTA+ also considers proper assignment of SSE types (α-helix, β-strand) and completeness of assigned SSEs. Regarding RMSD and the number of aligned residues, the most similar structure to **2JMK** is **1GO4** (Fig. 2C; Sironi et al. 2002). **1GO4** is a protein involved in cell cycle regulation. GANGSTA+ succeeded to align all seven SSEs of **2JMK** to the structure of **1GO4** at RMSD = 1.8 Å with 61 residues in total (see Fig. 2A,B), while the sequence identity between the two proteins is only about 19%. However, in this structure alignment, five out of the seven SSEs from **2JMK** have been aligned in reverse orientation albeit the SSE types, where properly assigned (Fig. 2B). Figure 2C displays the overall result of the applied database search correlating the

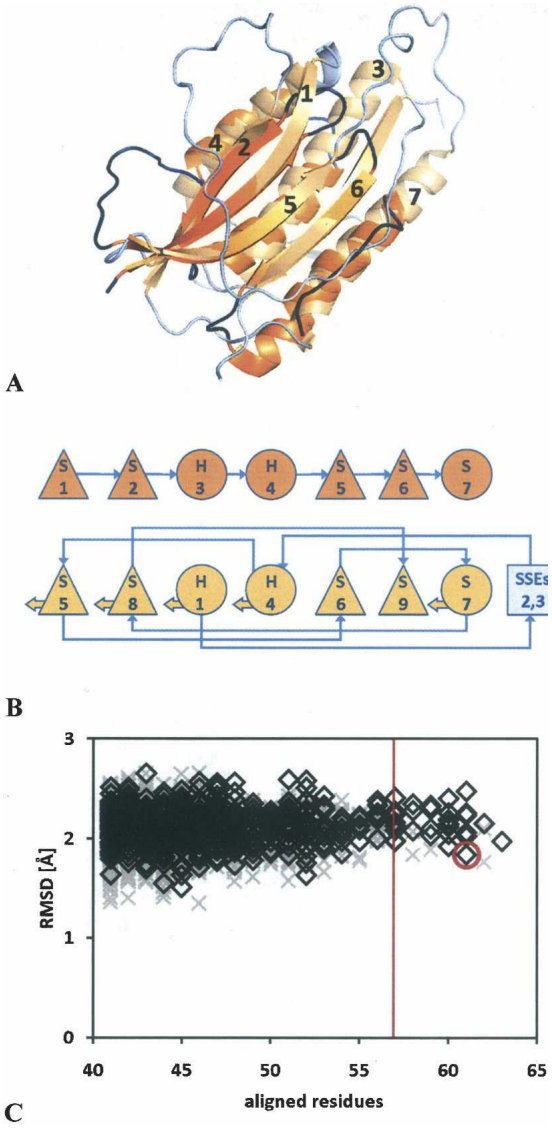

Figure 2. (*A*) Protein structure alignment with GANGSTA+. New fold **2JMK** (reference protein) (Koo et al. 2007) (dark colors, blue for loops and orange for SSEs) aligned with GANGSTA+ on **1GO4** (Sironi et al. 2002) (detected protein) (light colors, blue and orange) yielding the RMSD = 1.8 Å with 61 aligned residues and seven aligned SSEs. The aligned SSEs of **2JMK** (**1GO4**) are represented in dark (light) orange; non-aligned parts (SSEs and loops) are in dark (light) blue. The SSEs are numbered sequentially for the reference protein **2JMK** (dark orange). (*B*) Connectivity graph of protein structure alignment. Aligned SSE pairs are on *top* of each other (color code same as in *A*). SSEs are numbered in sequential order for both proteins. The SSE numbering in *A* refers to protein **2JMK**, *top* part. Letters denote the following: H, α-helices (circles); S, β-strands (triangles); blue arrows, connecting loops. (*Top* part) **2JMK** (all SSEs in dark orange); (*bottom* part) **1GO4** (aligned SSEs in light orange, unaligned SSEs in light blue). SSE pairs, assigned in reverse orientation, are marked by arrows pointing to the *left*. (*C*) Diagram correlating the number of aligned residues with the RMSD for the structure alignment results of GANGSTA+ with respect to **2JMK** and the ASTRAL40 data set (diamonds mark alignments involving all seven SSEs **2JMK**; "X" marks incomplete alignments). All results with more than 40 aligned residues are displayed. The structure of **2JMK** consists of seven SSEs, which comprise a total of 57 residues, marked by the red line. The red circle marks the structure alignment with **1GO4** shown in *A* and *B*.

number of aligned residues with the RMSD for all aligned protein pairs. In total, 1534 protein structure alignments of **2JMK** were found, involving more than 40 aligned residues. From these alignments, 469 provide a complete type-consistent SSE assignment of **2JMK**. But these alignments are all nonsequential in SSE connectivity, explaining that it is difficult to find these similar protein folds. Given the protein pair **2JMK** and **1GO4**, DaliLite (Holm and Park 2000) aligned 75 residues at RMSD = 11.0 Å (Z-Score = 1.7 < 2.0). But none of the SSEs was aligned in a type-consistent way. Also with TM-align (Zhang and Skolnick 2005b), the detected protein structure was less similar to the reference protein **2JMK**. The best alignment result considered 67 residues at RMSD = 4.0 Å (TM-score = 0.24 < 0.50) (see also Table 1) with a single type-consistent SSE pair only (α-helix no. 3) (see Fig. 2A).

In 2006, Sue et al. published the structure **2AJE**, stating that this DNA-binding protein appears to be a new fold, where a particular argument was the additional C-terminal helix (SSE no. 4 in Fig. 3A) (Sue et al. 2006). In a structural database search of the DPSA using DaliLite and TM-align, no protein structures with significant similarities were found. However, application of GANGSTA+ with respect to the ASTRAL40 data set revealed significant nonsequential similarities to **1J7N**, a domain of the anthrax lethal factor published in 2001 (Pannifer et al. 2001). GANGSTA+ aligned all four α-helices, including the additional C-terminal helix (see Fig. 3A) with a total of 53 aligned residues at a RMSD = 2.1 Å and a sequence identity of ~15%. The four α-helices of **2AJE** have been aligned to **1J7N** without the need of reverse SSE orientations (see Fig. 3B). The database scan took ~51 min (~0.4 s per protein pair). The overall results depicted in Figure 3C display only 31 protein

Table 1. *Database search of similar protein structures with GANGSTA+ and comparison of pairwise structure alignments made with DaliLite and TM-align*

New fold	Detected structural analog	DaliLite	TM-align	GANGSTA+
2JMK/7/57[a]	1GO4:A_[b]	11.0/0/75[c]	4.0/1/67[c]	1.8/7/61[c]
2AJE/4/44	1J7N:A2	3.9/3/45	3.4/3/45	2.1/4/53
2ES9/5/58	1SXJ:E1	2.5/4/57	4.0/5/65	1.8/5/69

Scanning the ASTRAL40 database, only GANGSTA+ was able to detect protein structures that are structurally similar to the listed three new folds. These detected similar protein structures correspond to alignments that are nonsequential in the SSE connectivity. For the structurally similar protein pairs found with GANGSTA+, it is possible to use also DaliLite and TM-align for pairwise structure alignments with results listed below.

[a] PDB ID of new fold/number of SSEs of the new fold/number of residues in these SSEs.

[b] PDB ID: domain ID according to SCOP of aligned protein structure.

[c] RMSD/number of type consistently aligned SSEs/number of aligned residues.

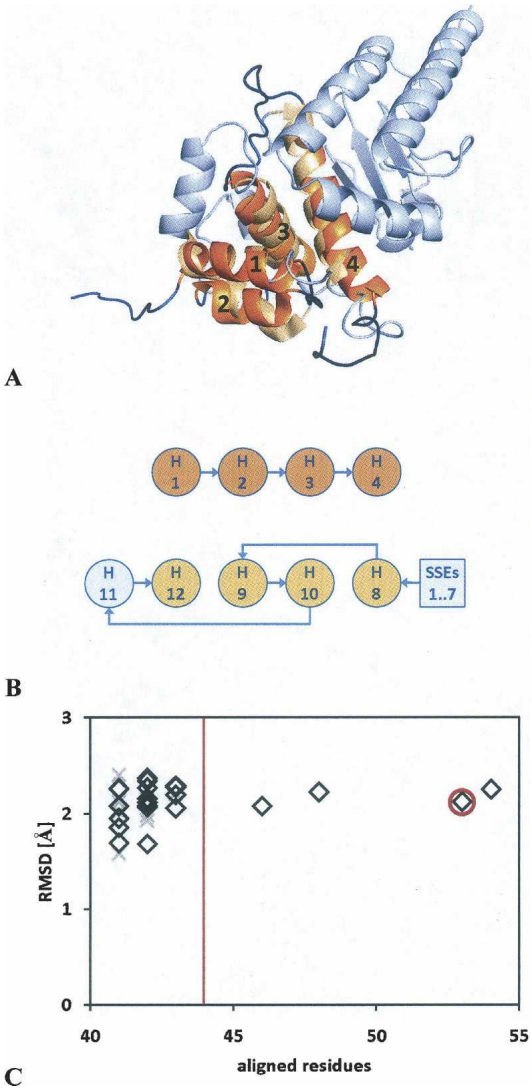

Figure 3. (*A*) Protein structure alignment with GANGSTA+. New fold **2AJE** (Sue et al. 2006) (reference protein) (dark colors, blue for loops and orange for SSEs) aligned with GANGSTA+ on **1J7N** (Pannifer et al. 2001) (detected protein) (light colors, blue and orange) yielding the RMSD = 2.1 Å with 53 residues and four aligned SSEs. The aligned SSEs of **2AJE** (**1J7N**) are represented in dark (light) orange; unaligned parts (SSEs and loops) are in dark (light) blue. The SSEs are numbered sequentially for the reference protein **2AJE** (dark orange). (*B*) Connectivity graph of protein structure alignment. Aligned SSE pairs are on *top* of each other (color code same as in *A*). SSEs are numbered in sequential order for both proteins. The SSE numbering in *A* refers to protein **2AJE**, *top* part. Letters denote the following: H, α-helices (circles); S, β-strands (triangles); connecting loops, blue arrows. (*Top* part) **2AJE** (all SSEs in dark orange); (*bottom* part) **1J7N** (aligned SSEs in light orange, unaligned SSEs in light blue). For this alignment, all SSE pairs are in the same orientation. (*C*) Diagram correlating the number of aligned residues with the RMSD for the structure alignment results of GANGSTA+ with respect to **2AJE** and the ASTRAL40 data set; (diamonds) alignments involving all four SSEs of **2AJE**; (X) incomplete alignments. All results with more than 40 aligned residues are displayed. (Red line) The structure of **2AJE** consists of four SSEs, which comprise a total of 44 residues. (Red circle) The structure alignment with **1J7N** shown in *A* and *B*.

structure alignments that involve more than 40 residues. Twenty-two of these structure alignments are complete, involving all four α-helices of **2AJE**. But all these structure alignments are nonsequential in SSE connectivity, and therefore more difficult to find than sequential alignments. Given the protein pair **2AJE** and **1J7N**, DaliLite aligned 45 residues at RMSD = 3.9 Å (Z-Score = 0.9 < 2.0) with three type-consistent SSE pairs. Here, TM-align also aligns 45 residues with three type-consistent SSE pairs at RMSD = 3.4 Å (TM-score = 0.16 < 0.50) (see also Table 1).

A third example is **2ES9**, deposited in the PDB in the year 2005 (Benach et al. 2005). It was claimed to possess a new fold, according to the SCOP classification library (Murzin et al. 1995). We searched with DaliLite for structures similar to **2ES9**, which yielded a sequential structure alignment of significant similarity (Z-Score = 5.2 > 2.0) to the structure of **1SZA** published in 2004 (Meinhart and Cramer 2004). DaliLite aligns 67 residues at RMSD = 2.5 Å for this protein pair. However, the generated alignment is insufficient to describe the fold of **2ES9** in sufficient detail, since only the four-helix bundle, which is a common motive (Mehl et al. 2003; Eckenhoffa et al. 2005), was aligned, while the fifth lateral α-helix was skipped.

The structure alignment with GANGSTA+ for **2ES9** with respect to the ASTRAL40 data set took 40 min (~0.3 s per protein pair) and revealed a nonsequential alignment with **1SXJ,** involving all five SSEs of **2ES9** (see Fig. 4A,B). Figure 4C depicts the overall result from database search. In this case, GANGSTA+ found 485 protein structure alignments, involving more than 40 residues. From these structure alignments, 140 involve all five SSEs of **2ES9**. But also here, all structure alignments found with GANGSTA+ are nonsequential in SSE connectivity. In the case of **1SXJ**, 69 residues were aligned at RMSD = 1.8 Å with a sequence identity of only ~12%. Four out of the five SSEs of **2ES9** were aligned in reverse orientation on the equivalent SSEs in **1SXJ** (see Fig. 4B). The structure of **1SXJ** was published in 2004 (Bowman et al. 2004). Both proteins **1SXJ** found with GANGSTA+ and **1SZA** found with DaliLite are involved in DNA or RNA polymerization, respectively. Hence, the connection between **2ES9** and **1SZA** found with DaliLite is relevant.

Given the protein pair **2ES9** and **1SXJ**, DaliLite aligned 57 residues at RMSD = 2.5 Å with a Z-Score of 3.2 > 2.0. It succeeded in aligning the four-helix bundle, but again skipped the lateral α-helix (see Fig. 5A,B). Although TM-align aligned all five SSEs of **2ES9** and **1SXJ** considering 65 residues, the detected structural similarity was low yielding a RMSD of 4.0 Å with a TM-score of 0.41 < 0.50 (see Table 1). Given the protein pair **2ES9** and **1SZA**, GANGSTA+ revealed comparable

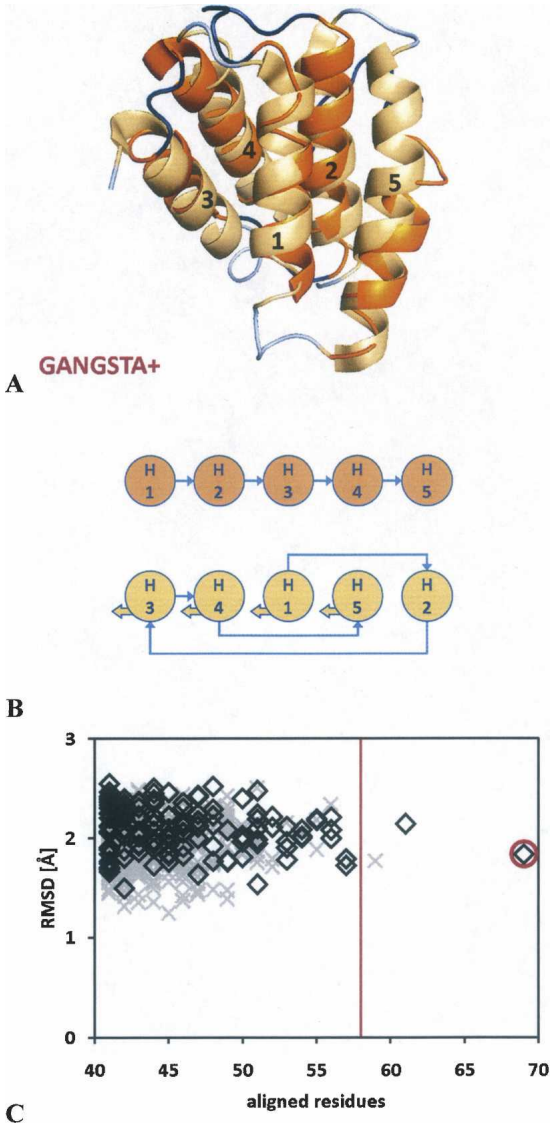

Figure 4. (*A*) Protein structure alignment with GANGSTA+. New fold **2ES9** (Benach et al. 2005) (reference protein) (dark colors, blue for loops and orange for SSEs) aligned with GANGSTA+ on **1SXJ** (Bowman et al. 2004) (detected protein) (light colors, blue and orange) yielding the RMSD = 1.8 Å with 69 aligned residues and five aligned SSEs. The aligned SSEs of **2ES9** (**1SXJ**) are represented in dark (light) orange; unaligned parts (SSEs and loops) are in dark (light) blue. The SSEs are numbered sequentially for the reference protein **2ES9** (dark orange). (*B*) Connectivity graph of protein structure alignment. Aligned SSE pairs are on *top* of each other (color code same as in *A*). SSEs are numbered in sequential order for both proteins. The SSE numbering in *A* refers to protein **2ES9**, *top* part. Letters denote the following: H, α-helices (circles); S, β-strands (triangles); connecting loops, blue arrows. (*Top* part) **2ES9** (all SSEs in dark orange); (*bottom* part) **1SXJ** (aligned SSEs in light orange). SSE pairs assigned in reverse orientation are marked by arrows pointing to the *left*. (*C*) Diagram correlating the number of aligned residues with the RMSD for the structure alignment results of GANGSTA+ with respect to **2ES9** and the ASTRAL40 data set; (diamonds) alignments involving all SSEs of **2ES9**; (X) incomplete alignments. All results with more than 40 aligned residues are displayed. (Red line) The structure of **2ES9** consists of five α-helices, which comprise a total of 58 residues. The red circle marks the structure alignment with **1SXJ** shown in *A* and *B*.

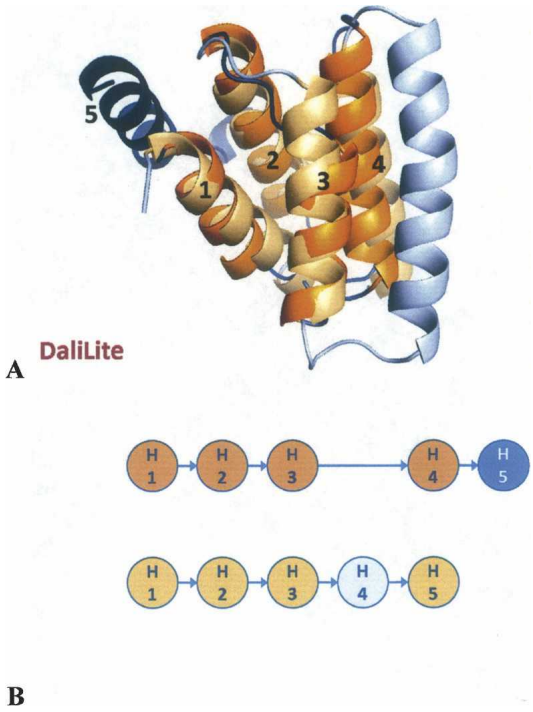

Figure 5. (*A*) Protein structure alignment with DaliLite. New fold **2ES9** (Benach et al. 2005) (reference protein) (dark colors, blue for loops and orange for SSEs) aligned with GANGSTA+ on **1SXJ** (Bowman et al. 2004) (detected protein) (light colors, blue and orange) yielding the RMSD = 2.5 Å with 57 aligned residues and five aligned SSEs. The aligned SSEs of **2ES9** (**1SXJ**) are represented in dark (light) orange; unaligned parts (SSEs and loops) are in dark (light) blue. The SSEs are numbered sequentially for the reference protein **2ES9** (dark orange). The fifth lateral α-helix of **2ES9** has not been aligned. (*B*) Connectivity graph of protein structure alignment. Aligned SSE pairs are on *top* of each other (color code same as in *A*). SSEs are numbered in sequential order for both proteins. The SSE numbering in *A* refers to protein **2ES9**, *top* part. Letters denote the following: H, α-helices (circles); S, β-strands (triangles); connecting loops, blue arrows. (*Top* part) **2ES9** (aligned SSEs in dark orange, unaligned SSEs in dark blue); (*bottom* part) **1SXJ** (aligned SSEs in light orange, unaligned SSEs in light blue). All SSE pairs are assigned in the same orientation.

similarities as detected by DaliLite. It aligned 49 residues sequentially at 2.1 Å RMSD, and similar to DaliLite it could not find an α-helix in **1SZA** that corresponds to the lateral helix in **2ES9**. The results of this example illustrate that the detection of a structure motive that resembles a four-helix bundle with respect to the enormous number of four-helix bundle related motives requires a sensitive and robust nonsequential structure alignment approach.

GANGSTA+ allows the detection of structural similarities between proteins also, if the assignment of SSE pairs is restricted to SSEs having the same orientation. With these constraints, we performed additional database searches for the two protein structures, **2JMK** and **2ES9**, with the DPSA, since without such constraints the best structure alignments were obtained with reverse orientations of some of the aligned SSEs. Initially, we restricted the alignments to avoid only inversion of α-helix orientations, but allowed inversion of β-strand orientations. This partial restriction is of particular interest, since inversion of a helix axis goes along with inversion of the helix dipole moment, while inversion of a strand has fewer consequences for function and energetics of a protein. Under these conditions, GANGSTA+ found for **2JMK** nonsequential structure similarities to the DPSA entry **1Q6Z** (Bera et al., in press), aligning 60 residues at RMSD = 2.0 Å involving inversion of two β-strand orientations. Since it is mandatory for applications with GANGSTA+ to rank structure alignments highest, which align all SSEs of the reference protein, all seven SSEs of **2JMK** were aligned SSE-type consistent with **1Q6Z**. Avoiding also inversion of β-strand orientations, GANGSTA+ detected nonsequential similarities between **2JMK** and **1VJU** (SGPP), aligning only 45 residues at RMSD = 2.4 Å. Note that for this more conservative alignment result, the number of aligned residue pairs was below the total number of residues of 58 belonging to the SSEs of the reference protein **2JMK**. Here, still, all seven SSEs of **2JMK** were aligned in the same orientation as the equivalent SSEs in **1VJU** (see Supplemental Figs. S1A–C).

Finally, we performed a search with GANGSTA+ to detect similarities for **2ES9** with the DPSA with SSE pairs that possess the same orientation. This yielded the nonsequential structure alignment of **2ES9** with **1H6K** (Mazza et al. 2001) with complete SSE assignment, aligning 48 residues at RMSD = 2.2 Å (see Supplemental Figs. S2A–C). **1H6K** was published by Mazza et al. in 2001 and is a nuclear protein containing an RNA-binding motif. Thus, likewise, **1SXJ** and **1SZA**, the function of **1H6K** also relates to nucleic acids (Berman et al. 2000). These results illustrate that GANGSTA+ is capable of generating high-quality nonsequential structure alignments, considering SSE pairs with the same or reverse orientations, depending on the needs of the application.

Discussion

The protein structure alignment tool GANGSTA+ solves alignment problems in a three-stage hierarchical approach starting with an alignment on the secondary structure level, where only α-helices and β-strands are considered. In the second stage, the residue pair assignment is performed on the basis of the results from the first stage. In a subsequent last stage, a refinement of the residue pair assignment is performed to complete the SSE assignment from the first stage, and to find possible reassignments of SSEs, and to extend the residue pair assignment beyond the SSE boundaries. The latter procedure leads often to a larger number of aligned residue

pairs as the number of residues in the aligned SSEs of the smaller protein. This is clearly visible in the diagrams correlating RMSD with the number of aligned residues for the proteins whose structures were found to be similar to the reference protein structure (Figs. 2C, 3C, and 4C). In all three cases of proteins considered to be new folds, (**2JMK)** (Koo et al. 2007), (**2AJE)** (Sue et al. 2006), and (**2ES9)** (Benach et al. 2005), GANGSTA+ found structure alignments where the number of aligned residue pairs was larger than the total number of residues in all SSEs considered for the alignment. See the alignment results in the correlation diagrams (Figs. 2C, 3C, and 4C), where the total number of residues in all considered SSEs are marked by the red vertical line.

The four key features of GANGSTA+ are: (1) It can perform sequential, but alternatively, also nonsequential structure alignments, disregarding the polypeptide connectivity in the latter case; (2) it assigns SSE pairs only if they are of the same type (α-helix or β-strand); (3) it is capable of aligning SSEs having the same or reverse mutual orientations; and (4) it maximizes the number of assignable SSEs. GANGSTA+ manages to find nonsequential structure alignments, since it ignores the loops connecting the SSEs in the first SSE alignment stage. Considering the loops in the initial stage of structure alignment would favor sequential SSE alignments. However, after the SSE assignment is terminated in the final (third) stage of structure alignment, the residue pair assignment also involves residues belonging to the loop regime of the protein structures. While GANGSTA+ assigns SSEs only type consistently, that is, helix on helix and strand on strand, it has the option to align SSEs in the same or opposite orientations.

GANGSTA+ provides nonsequential protein structure alignments in the same time range as the fastest commonly used sequential structure alignment methods with less than a second per protein pair, on average, on an AMD/OPTERON with 1600 MHz. Furthermore, a comparison with TM-align and the TM-score illustrates that GANGSTA+ is able to solve sequential protein structure alignments according to the TM-score with comparable quality.

We demonstrated that GANGSTA+ is able to detect nonsequential similarities for protein chains stated to possess new folds considering three examples. Although the question whether a new protein structure contains a new fold or not remains difficult to judge, the results illustrate that the application of nonsequential structure alignment tools can yield additional insight to understand protein structures and fold characteristics, presenting new starting points for protein function analysis and protein structure comparison. GANGSTA+ is not bound to a sequential connectivity of SSEs in the polypeptide chains of proteins. Thus, it can detect structure similarities of different proteins that have common ancestors, but whose SSE connectivity was reshuffled by genetic operations during evolution (Cooper et al. 1997). Although DaliLite and TM-align proved to be very accurate structural alignment methods on representative data sets (Hou et al. 2002; Day et al. 2003; Pandit et al. 2006; Barthel et al. 2007), they are restricted or biased toward sequential structure alignments. This can lead to failures in detection of structural relations between protein chains that possess nonsequential similarity.

In future investigations, we aim to unravel functional relationships of proteins with yet unknown functions using GANGSTA+ to detect nonsequential structure similarity. Furthermore, we aim to improve protein structure prediction approaches on the basis of nonsequential structural relations. In this context, multiple structure alignments with GANGSTA+ that can be used to define new sequence similarity measures for sequence alignment methods could be a promising direction (Schwartz and Dayhoff 1978; Pearson and Lipman 1988; Henikoff and Henikoff 1992; Altschul et al. 1997; Pearson and Sierk 2005).

Further investigations will focus on structurally similar proteins with SSEs pairs aligned in reverse orientation. In contrast to inversion of β-strand orientation, the inversion of an α-helix axis goes along with the inversion of the large helix dipole (Chakrabarti 1994). The helix dipole can have a strong influence on protein stability, its intrinsic function (Chou et al. 1988; Fairman et al. 1989; Aqvist et al. 1991; Ben-Tal and Honig 1996; Sengupta 2005), and ability of complex formation with other proteins (Miura et al. 1999). β-Strands in proteins can be organized in alternative orientations (parallel or antiparallel), which indicates functional robustness of proteins toward inverse orientations of β-strands. Therefore, we would expect to observe significantly more β-strand inversions than α-helix inversions. GANGSTA+ enables us to analyze the functional relevance of the α-helices dipole orientation, and its impact on SSE arrangements in common structural motifs for large databases. We are able to discriminate between single α-helix inversions or arbitrary many inversions of each SSE type, for example, to count parallel and antiparallel β-strands within a certain protein family. These possibilities underline the wide range of GANGSTA+ applicability to analyze the protein fold spaces and its properties.

Materials and Methods

Structure alignment

Here we provide a short overview of the protein structure alignment method used in the present study. A more detailed technical description is given in the ESM. GANGSTA+ as well

as GANGSTA (Kolbeck et al. 2006) have in common the ability to align protein structures hierarchically starting with an alignment of secondary structure elements (SSE; first stage). Only α-helices and β-strands are considered as SSEs. Nonsequential structure alignment is facilitated, since loops and coils connecting the SSEs are ignored. GANGSTA used a genetic algorithm to explore similarities between two protein structures based on contact maps of SSE, while GANGSTA+ uses a combinatorial approach, which is more efficient and reliable. For the highest ranked SSE assignments, preliminary alignments on the residue level are performed (second stage), using rigid body energy minimization with attractive soft interactions between C_α atoms belonging to different proteins. This minimizes the spatial distances of the assigned C_α atom pairs. In stage three, the preliminary structural overlay is used to assign the C_α atoms of both proteins to points on the same rectangular grid. C_α atom pairs assigned to the same grid points are used for a more accurate and complete SSE assignment. Finally, the assignment on the residue level (second stage) is repeated, also aligning residues belonging to loops and coils.

Scoring results of structure alignments

Protein structure alignments are evaluated with the structure alignment score (SAS) (Kolodny et al. 2005):

$$SAS = (RMSD * 100)/N_{aligned} \quad (1)$$

that weights the RMSD of C_α atoms relative to the number of aligned residues $N_{aligned}$. In GANGSTA, we used the more complex GANGSTA score:

$$G_s = \frac{RMSD + 2 * N_{gap}}{N_{aligned} * q^{st}_{res} * (1 - \Delta SSE) + \varepsilon}, \quad (2)$$

where N_{gap} is the number of unassigned SSEs of the reference protein; q^{st}_{res}, varying between 0 and 1, is the residue contact map overlap; and ΔSSE measures the similarity of the SSE pair distance map between the two aligned protein structures as defined in Kolbeck et al. (2006). To prevent division by zero, $\varepsilon = 10^{-5}$ is used. G_s is particularly useful to detect more distant similarities between protein structures. In the present application, we focus on protein structure pairs possessing a high degree of similarity, where the more simple similarity measure of SAS, Equation 1, is sufficient. No different alignments were found using G_s instead. Both scores; SAS and G_s, are available in GANGSTA+.

The TM-score (Zhang and Skolnick 2005b), TM_S, is also used for a large-scale database comparison of sequential protein structure alignments from GANGSTA+ and TM-align (see section on large-scale database comparison). It is defined by

$$TM_S = \max\left[\frac{1}{L_{Ref}} \sum_{i=1}^{N_{aligned}} \frac{(d_0 * L_{Ref})^2}{(d_0 * L_{Ref})^2 + d_i^2}\right], \text{ where}$$
$$d_0 = 1.24\sqrt[3]{L_{Ref} - 15} - 1.8\ , \quad (3)$$

where L_{Ref} is the total number of residues of the generally smaller reference protein to which other protein structures from a database are aligned, $N_{aligned}$ is the number of aligned residues, d_i is the C_α–C_α distance between the *i*th pair of aligned residues, and d_0 is a distance parameter normalizing the distances to make the average TM-score independent of the protein size for random structure pairs (see details in Zhang and Skolnick 2005b).

Database for structure alignments

The database of proteins used for structure alignment (DPSA) by GANGSTA+ involves all protein domains from ASTRAL40 (SCOP 1.71) (Murzin et al. 1995) with at least three SSEs (7347 protein chains in total), yielding about 27 10^6 possible pairs for protein structure alignment. Application of GANGSTA+ to the whole DPSA yielded about 8.4 10^6 pairs of successfully aligned protein structures with an SAS <10, where >50% of the SSEs from the smaller protein are involved in the structure alignment.

Acknowledgments

The authors thank Björn Kolbeck, Tobias Schmidt-Goenner, and Gernot Kieseritzky for useful discussions. This project was funded by the International Research Training Group (IRTG) on ''Genomics and Systems Biology of Molecular Networks'' (GRK1360, Deutsche Forschungsgemeinschaft [DFG]).

References

Alexandrov, N. and Fischer, D. 1996. Analysis of topological and nontopological structural similarities in the PDB: New examples with old structures. *Proteins* **25:** 354–365.

Altschul, S.F., Madden, T.L., Schaffer, A.A., Zhang, J., Zhang, Z., Miller, W., and Lipman, D.J. 1997. Gapped BLAST and PSI-BLAST: A new generation of protein database search programs. *Nucleic Acids Res.* **25:** 3389–3402.

Anfinsen, C.B., Haber, E., Sela, M., and White, F.H. 1961. The kinetics of formation of native ribonuclease during oxidation of the reduced polypeptide chain. *Proc. Natl. Acad. Sci.* **47:** 1309–1314.

Aqvist, J., Luecke, H., Quiocho, F.A., and Warshel, A. 1991. Dipoles localized at helix termini of proteins stabilize charges. *Proc. Natl. Acad. Sci.* **88:** 2026–2030.

Barthel, D., Hirst, J.D., Blazewicz, J., Burke, E.K., and Krasnogor, N. 2007. ProCKSI: A decision support system for protein (structure) comparison, knowledge, similarity and information. *BMC Bioinformatics* **8:** 416. doi: 10.1186/1471-2105-8-416.

Benach, J., Abashidz, E.M., Jayaraman, S., Rong, X., Acton, T.B., Montelione, G.T., and Tong, L. 2005. *Crystal structure of Q8ZRJ2 from Salmonella typhimurium NESG TARGET STR65.* RCSB Protein Data Bank, Piscataway, NJ(in press). doi: 10.2210/pdb2es9/pdb.

Ben-Tal, N. and Honig, B. 1996. Helix–helix interactions in lipid bilayers. *Biophys. J.* **71:** 3046–3050.

Bera, A.K., Anderson, N.L., and Hasson, M.S. High-resolution structure of E28A mutant benzoylformate decarboxylase from *Pseudomonas putida* complexed with thiamin thiazolone diphosphate (in press).

Berman, H., Westbrook, J., Feng, Z., Gilliland, G., Bhat, T., Weissig, H., Shindya-lov, I., and Bourne, P. 2000. The Protein Data Bank. *Nucleic Acids Res.* **28:** 235–242.

Bowman, G.D., O'Donnell, M., and Kuriyan, J. 2004. Structural analysis of a eukaryotic sliding DNA clamp–clamp loader complex. *Nature* **429:** 724–730.

Chakrabarti, P. 1994. An assessment of the effect of the helix dipole in protein structures. *Protein Eng.* **7:** 471–474.

Chandonia, J., Hon, G., Walker, N., Lo, C., Koehl, P., Levitt, M., and Brenner, S. 2004. The ASTRAL compendium in 2004. *Nucleic Acids Res.* **32:** 189–192.

Chen, L., Wu, L.Y., Wang, Y., Zhang, S., and Zhang, X.S. 2006. Revealing divergent evolution, identifying circular permutations and detecting active sites by protein structure comparison. *BMC Struct. Biol.* **6:** 18. doi: 10.1186/1472-6807-6-18.

Chou, K.C., Maggiora, G.M., Némethy, G., and Scheraga, H.A. 1988. Energetics of the structure of the four-α-helix bundle in proteins. *Proc. Natl. Acad. Sci.* **85:** 4295–4299.

Cooper, D.N., Ball, E.V., and Krawczak, M. 1997. The human gene mutation database. *Nucleic Acids Res.* **26:** 285–287.

Day, R., Beck, D.A.C., Armen, R.S., and Daggett, V. 2003. A consensus view of fold space: Combining SCOP, CATH, and the Dali Domain Dictionary. *Protein Sci.* **12:** 2150–2160.

Dror, O., Benyamini, H., Nussinov, R., and Wolfson, H.J. 2003. MASS: Multiple structural alignment by secondary structures. *Bioinformatics* (Suppl 1) **19:** i95–i104. doi: 10.1093/bioinformatics/ftg1012.

Eckenhoffa, R.G., Liua, R., Johanssona, J.S., and Lollb, P.J. 2005. The four-helix bundle: An attractive fold. *Int. Congr. Ser.* **1283:** 15–20.

Fairman, R., Shoemaker, K.R., York, E.J., Stewart, J.M., and Baldwin, R.L. 1989. Further studies of the helix dipole model: Effects of a free α-NH^{3+} or α-COO^{-} group on helix stability. *Proteins* **5:** 1–7.

Govindarajan, S., Recabarren, R., and Goldstein, R.A. 1999. Estimating the total number of protein folds. *Proteins* **35:** 408–414.

Grant, A., Lee, D., and Orengo, C. 2004. Progress towards mapping the universe of protein folds. *Genome Biol.* **5:** 107.

Guo, H.H., Choe, J., and Lawrence, L.A. 2004. Protein tolerance to random amino acid change. *Proc. Natl. Acad. Sci.* **101:** 9205–9210.

Henikoff, S. and Henikoff, J.G. 1992. Amino acid substitution matrices from protein blocks. *Proc. Natl. Acad. Sci.* **89:** 10915–10919.

Holm, L. and Park, J. 2000. DaliLite workbench for protein structure comparison. *Bioinformatics* **6:** 566–567.

Hou, J., Sims, G.E., Zhang, C., and Kim, S.H. 2002. A global representation of the protein fold space. *Proc. Natl. Acad. Sci.* **100:** 2386–2390.

Ilyin, V., Abyzov, A., and Leslin, C. 2004. Structural alignment of proteins by a novel TOPOFIT method, as a superimposition of common volumes at a topomax point. *Protein Sci.* **13:** 1865–1874.

Kolbeck, B., May, P., Schmidt-Goenner, T., Steinke, T., and Knapp, E.W. 2006. Connectivity independent protein-structure alignment. *BMC Bioinformatics* **7:** 510. doi: 10.1186/1471-2105-7-510.

Kolodny, R., Koehl, P., and Levitt, M. 2005. Comprehensive evaluation of protein structure alignment methods. *J. Mol. Biol.* **346:** 1173–1188.

Koo, B.K., Jung, J., Jung, H., Nam, H.W., Kim, Y.S., Yee, A., and Lee, W. 2007. Solution structure of the hypothetical novel-fold protein TA0956 from *Thermoplasma acidophilum*. *Proteins* **69:** 444–447.

Lee, D., Redfern, O., and Orengo, C. 2007. Predicting protein function from sequence and structure. *Nat. Rev. Mol. Cell Biol.* **8:** 995–1005.

Leonov, H., Mitchell, J.S.B., and Arkin, I.T. 2003. Monte Carlo estimation of the number of possible protein folds: Effects of sampling bias and folds distributions. *Proteins* **51:** 352–359.

Liu, X., Fan, K., and Wang, W. 2004. The number of protein folds and the distribution over families in nature. *Proteins* **54:** 491–499.

Mazza, C., Ohno, M., Segref, A., Mattaj, I.W., and Cusack, S. 2001. Crystal structure of the human nuclear cap-binding complex. *Mol. Cell* **8:** 383–396.

Mehl, A.F., Heskett, L.D., Jain, S.S., and Demeler, B. 2003. Insights into dimerization and four-helix bundle formation found by dissection of the dimer interface of the GrpE protein from *Escherichia coli*. *Protein Sci.* **12:** 1205–1215.

Meinhart, A. and Cramer, P. 2004. Recognition of RNA polymerase II carboxy-terminal domain by 3′-RNA-processing factors. *Nature* **430:** 223–226.

Miura, Y., Kimura, S., Kobayashi, S., Iwamoto, M., Imanishi, Y., and Umemura, J. 1999. Negative surface potential produced by self-assembled monolayers of helix peptides oriented vertically to a surface. *Chem. Phys. Lett.* **315:** 1–6.

Murzin, A.G., Brenner, S.E., Hubbard, T., and Chothia, C. 1995. SCOP. *J. Mol. Biol.* **247:** 536–540.

Pandit, S.B., Zhang, Y., and Skolnick, J. 2006. TASSER-Lite: An automated tool for protein comparative modeling. *Biophys. J.* **91:** 4180–4190.

Pannifer, A.D., Wong, T.Y., Schwarzenbacher, R., Renatus, M., Petosa, C., Bienkowska, J., Lacy, D.B., Collier, R.J., Park, S., Leppla, S.H., et al. 2001. Crystal structure of the anthrax lethal factor. *Nature* **414:** 229–233.

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

Pearson, W.R. and Sierk, M.L. 2005. The limits of protein sequence comparison? *Curr. Opin. Struct. Biol.* **15:** 254–260.

Russ, W., Lowery, D., Mishra, D., Yaffe, M., and Ranganathan, R. 2006. Natural-like function in artificial WW domains. *Nature* **437:** 579–583.

Schwartz, R.M. and Dayhoff, M.O. 1978. Detection of distant relationships based on point mutation data. In *Evolution of protein molecules* (eds. H. Matsubara and T. Yamanaka), pp. 1–16. Academic Japan, Tokyo, Japan.

Sengupta, D., Behera, R.N., Smith, J.C., and Ullmann, G.M. 2005. The α-helix dipole: Screened out? *Structure* **13:** 849–855.

Shatsky, M., Nussinov, R., and Wolfson, H.J. 2002. MultiProt—a multiple protein structural alignment algorithm. In *Proceedings of the Second International Workshop on Algorithms in Bioinformatics,* Lecture Notes in Computer Science, Vol. 2452, pp. 235–250. Springer-Verlag, London, UK.

Shih, E.S.C. and Hwang, M.J. 2004. Alternative alignments from comparison of protein structures. *Proteins* **56:** 519–527.

Shulman-Peleg, A., Shatsky, M., Nussinov, R., and Wolfson, H. 2007. Spatial chemical conservation of hot spot interactions in protein–protein complexes. *BMC Biol.* **5:** 43. doi: 10.1186/1741-7007-5-43.

Sironi, L., Mapelli, M., Knapp, S., Antoni, A.D., Jeang, K.T., and Musacchio, A. 2002. Crystal structure of the tetrameric Mad1–Mad2 core complex: Implications of a ''safety belt'' binding mechanism for the spindle checkpoint. *EMBO J.* **21:** 2496–2506.

Sue, S.C., Hsiao, H., Chung, B.C.P., Cheng, Y.H., Hsueh, K.L., Chen, C.M., Ho, C.H., and Huang, T. 2006. Solution structure of the *Arabidopsis thaliana* telomeric repeat-binding protein DNA binding domain: A new fold with an additional C-terminal helix. *J. Mol. Biol.* **356:** 72–85.

Szustakowski, J.D. and Weng, Z. 2000. Protein structure alignment using a genetic algorithm. *Proteins* **38:** 428–440.

Szustakowski, J.D. and Weng, Z. 2002. Protein structure alignment using evolutionary computing. In *Evolutionary computation in bioinformatics* (eds. G. Fogel et al.) pp. 59–86. Morgan Kaufman, San Francisco, CA.

Wang, Z.X. 1998. A re-estimation for the total numbers of protein folds and superfamilies. *Protein Eng.* **11:** 621–626.

Whisstock, J.C. and Lesk, A.M. 2003. Prediction of protein function from protein sequence and structure. *Q. Rev. Biophys.* **36:** 307–340.

Yang, A.S. and Honig, B. 1999. Sequence to structure alignment in comparative modeling using PrISM. *Proteins* **3:** 66–72.

Yuan, X. and Bystroff, Y. 2005. Non-sequential structure-based alignments reveal topology-independent core packing arrangements in proteins. *Bioinformatics* **7:** 1010–1019.

Zhang, Y. and Skolnick, J. 2005a. The protein structure prediction problem could be solved using the current PDB library. *Proc. Natl. Acad. Sci.* **102:** 1029–1034.

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