The structural alignment between two proteins: Is there a unique answer?

ADAM GODZIK

Department of Molecular Biology MB-1, The Scripps Research Institute, La Jolla, California 92037

(RECEIVED April 10, 1996; ACCEPTED April 23, 1996)

Abstract

Structurally similar but sequentially unrelated proteins have been discovered and rediscovered by many researchers, using a variety of structure comparison tools. For several pairs of such proteins, existing structural alignments obtained from the literature, as well as alignments prepared using several different similarity criteria, are compared with each other. It is shown that, in general, they differ from each other, with differences increasing with diminishing sequence similarity. Differences are particularly strong between alignments optimizing global similarity measures, such as RMS deviation between C α atoms, and alignments focusing on more local features, such as packing or interaction pattern similarity. Simply speaking, by putting emphasis on different aspects of structure, different structural alignments show the unquestionable similarity in a different way. With differences between various alignments extending to a point where they can differ at all positions, analysis of structural similarities leads to contradictory results reported by groups using different alignment techniques.

The problem of uniqueness and stability of structural alignments is further studied with the help of visualization of the suboptimal alignments. It is shown that alignments are often degenerate and whole families of alignments can be generated with almost the same score as the “optimal alignment.” However, for some similarity criteria, specially those based on side-chain positions, rather than C α positions, alignments in some areas of the protein are unique. This opens the question of how and if the structural alignments can be used as “standards of truth” for protein comparison.

Keywords: protein similarity; protein structural alignment

The existence of whole groups of sequentially unrelated but structurally similar proteins (Pascarella & Argos, 1992; Holm & Sander, 1993; Orengo et al., 1993) recently received a lot of attention and allowed us to ask many fundamental questions about protein structures and rules of folding (Finkelstein & Pritsyn, 1987; Chothia, 1992; Orengo et al., 1993). However, the apparently simple question about similarity between two protein structures is, in fact, quite difficult. On the one hand, the considerable success of threading methods (Bowie et al., 1991; Godzik et al., 1992; Jones et al., 1992; Lemer et al., 1995) seems to suggest that there is a significant similarity in interactions stabilizing structurally similar but sequentially unrelated proteins. On the other hand, other groups claimed that interactions stabilizing different proteins from the same structural family are, in fact, quite different (Russell & Barton, 1994).

These inconsistencies hamper progress in other fields, which use structural alignments for calibration or evaluation of their results. For instance, structural alignments are often used as a “standard of truth” when evaluating sequence based alignment

methods and choosing the best scoring matrices and alignment strategies (Vogt et al., 1995). Although most agree that some pairs and groups of proteins are similar, there is no agreement on how are they similar. In this contribution, I address this problem by careful analysis of several different structural alignments for a few pairs of structurally similar but sequentially unrelated proteins.

There are two main issues in comparing two protein structures. One is the problem of a measure used to quantify the similarity. The RMS deviation (RMSD) for the optimal superposition of C α atoms of aligned positions is the most popular similarity measure (Maiorov & Crippen, 1994). Other groups use the difference between distance maps of two proteins (Holm & Sander, 1993), contact map overlap (Godzik et al., 1993), or more complicated scoring systems (Lesk & Chothia, 1980; Sali & Blundell, 1990; Fischer et al., 1992; Jones et al., 1992; Russell & Barton, 1992), including, among others, such additional structural features as local secondary structure, hydrogen bonding pattern, burial status, or interaction environment. More detailed discussion of various similarity measures is presented as a part of Materials and methods.

The second problem arises from the combination of the non-locality of the scoring function (see detailed discussion in Materials and methods) and the existence of insertions/deletions in the alignment. It can be proved that, for these reasons, finding an

Reprint requests to: Adam Godzik, Department of Molecular Biology MB-1, The Scripps Research Institute, 10666 N. Torrey Pines Road, La Jolla, California 92037; e-mail: adam@scripps.edu; www: <http://www.scripps.edu/adam/mosaic/adam.html>.

optimal structural alignment between two protein structures is an NP-hard problem (Lathrop, 1994) and thus there are no fast structural alignment algorithms that are guaranteed to be optimal within any given similarity measure. Therefore, all of the existing structural alignment algorithms use some simplifications, either of the scoring function or of the search procedure, to arrive at the reasonable, if not the optimal, alignment. Breaking down the dimensionality of the problem by performing at least part of the search at the level of secondary structure elements is an often used approach (Mitchell et al., 1990; Sali & Blundell, 1990; Holm & Sander, 1993). Another possibility is to represent a three-dimensional structure as a one-dimensional pseudo-sequence (Jones et al., 1992). Different optimization methods, such as the dynamic programming, two level dynamic programming (Jones et al., 1992), and Monte Carlo minimization (Sali & Blundell, 1990; Godzik et al., 1993; Holm & Sander, 1993) are used by different groups. Algorithms are described in the formalism of sequence homology analysis, graph theory (Mitchell et al., 1990), or computer vision technique (Fischer et al., 1992). As a result, protein structure comparison is a vibrant, very active field where different techniques are used to answer the same question of how different/similar are various protein structures. These efforts lead to several independent classifications of existing protein structures into structural families, with many of these classifications available publicly, either as part of original publications (Pascarella & Argos, 1992; Fischer et al., 1995) or as Internet resources (CATH, 1995; DALI, 1995; SCOP, 1995).

Although there was no systematic effort to compare and cross-validate the results of classifications and comparisons performed by different groups, the results are clearly similar. All methods recognize the existence of structurally similar proteins, and all well-known examples, such as globin-like topology of colicin A and phycocyanin or immunoglobulin-like topology of fasciulin, are reproduced in all existing classifications. However, by putting emphasis on different aspects of the protein structure, the unquestionable similarity is seen in a different way by various methods. There are usually few or no ambiguities on the level of whole structural elements, such as helices or strands. In fact, some of the structure alignment programs (Mitchell et al., 1990; Fischer et al., 1992) stop at this level and do not produce the residue-residue alignment. It is at the residue level where differences between various methods become clearly visible. At the same time, even the most recent papers talk about the structural alignment between two proteins, as if the existence of the only one and true alignment was obvious and did not require any explanation.

In fact, the opposite is true, and the implicit assumption about the existence of the unique structural alignment "borrowed" from the field of homology searches and sequence comparison may not be warranted here. There is a fundamental difference between these two fields. When comparing sequences of two related proteins, we are certain that there is a unique, correct solution, even if we are not able to find it. Both proteins have evolved by a series of mutations-deletions-insertions from the common ancestor, and there is an actual molecular process that is linking one sequence to another. Therefore, there is a unique, one to one correspondence between positions in each protein and positions in the common ancestor. This correspondence, if it could only be known, could be used to create the true alignment. Of course, we do not know this correspondence, so the sequence alignments are approximate, subject to our approximate knowledge of the evolution process. On the other hand, the process of transform-

ing one protein structure into another, or deriving both from a common ancestor, does not exist. The two proteins may fold due to a different balance of forces and there is not necessarily a one to one correspondence between positions in both proteins. In fact, it is entirely possible that structural alignment between two distantly homologous proteins could be different from the correct evolutionary alignment, if, for instance, some fragments of the sequence changed their function during evolution.

To keep clear the differences between various alignments and purposes of making and using them, we would use a convention where term "homologous" would be used only to denote an implied evolutionary relationship between proteins. In contrast, the term "similarity" would be used to describe just that, the similarity between two sequences or structures, without implying any relationship between proteins. In this sense, this paper deals only with similarities between proteins, and in particular, between protein structures. Questions of protein homology lay outside its scope.

The paper is organized as follows: results of the comparison between several existing structural alignments for few chosen protein families will be presented in the first part of the Results. It will be followed by the analysis of the suboptimal structural alignments for the same groups of proteins. Different similarity measures and algorithms for finding the optimal structural alignment, as well as various visualization techniques for comparing different alignments for the same protein pair and studying the distribution of suboptimal alignments will be reviewed in Materials and methods.

Results

Among the approximately 200 structural families known so far (CATH, 1995; DALI, 1995; SCOP, 1995), close to 100 of them display a very large sequence diversity, touching or crossing the boundary of nonhomologous proteins with similar structure. Only a few examples were chosen for detailed analysis here, mostly well studied structural families that were analyzed before by several different groups, because the availability of several types of structural alignments was necessary for the analysis. In particular, the following structural pairs were analyzed in detail:

copper binding proteins – azurin (1azcA) and plastocyanin (1plc)
flavodoxin fold – flavodoxin (4fxn) and chemotaxis protein (3chy)
globin – myoglobin (1mbc) and phycocyanin (1cpcA)

A more comprehensive database of various types of structural alignments is under construction at the author's laboratory and a continuously updated current version is accessible via the World Wide Web at <http://www.scripps.edu/adam/mosaic/structure.html>. Its current size is not sufficient to allow statistical analysis and comparison of structural alignments of different type, but as would be shown in the remainder of this paper, it is definitely large enough to notice systematic differences and similarities between structural alignments of different type.

Comparison of various alignments

First, for each protein group different available, structural alignments were identified in the literature or in on-line databases. Alignments were downloaded from the Internet or copied from literature. In the following part of this paper, various protein families and their alignments will be discussed case by case.

Copper binding proteins

Plastocyanin and azurin are homologous proteins with the topology of a Greek key beta sandwich, and a similar function of binding blue copper in a homologous binding site and participating in the electron transfer chain in plants and bacteria, respectively. The sequence similarity between plastocyanin and azurin is on the edge of the "twilight zone" of proteins barely recognizable by sequence alignment tools. Both structures have been known for quite some time and thus become a favorite case study for various structural alignment tools. The first structural alignment based on the analysis of both structures was built by hand on the basis of the sequence similarity, but include information about residue function and published by the team that crystallized azurin (Adman, 1985). It was followed by the detailed analysis of the similarity between this pair of proteins, where packing in the core was used as a criterion for an alignment (Chothia & Lesk, 1982). Structural alignments between plastocyanin and azurin were included in two alignment databases, FSSP and ali_3D, both maintained at EMBL. The first of those was obtained with the DALI program, optimizing the difference of two C α -C α distance maps by a MC minimization (Holm & Sander, 1993). The second was obtained by iteratively extending the small group of initially aligned positions, identified by sequence alignment and visual inspection (Pascarella & Argos, 1992). Finally, plastocyanin-azurin case was used as an example in several papers presenting new approaches to optimize the structural alignment between two proteins (Taylor & Orengo, 1989; May & Johnson, 1994). Contact map-based alignment was prepared following a procedure described previously (Godzik et al., 1993; Godzik & Skolnick, 1994; see also Materials and methods). All these alignments are compared with each other in Figure 1 and Table 1. The first part of Figure 1 shows various alignments presented in a dot plot format, with points on the plane identifying aligned positions between two proteins according to different alignments. Contact-map overlap-based alignment is identified by a solid line, other alignments as dashed lines. In the second part of the Figure 1, all alignments are shown in a format similar to a multiple alignment, but here different alignments of the same protein are presented with the table in-

dexed by positions in the template protein being aligned. Alignments are also compared by calculating different measures of protein similarity according to different alignments (see Table 1). For comparison, a sequence alignment is presented, as obtained by the GAP program from GCG program suite (Genetics Computer Group, 1991), using the Gonnet substitution matrix, which was recently shown to produce consistently better results than other similarity matrices (Vogt et al., 1995).

The most interesting observation is that all these alignments are really very different, with RMSD ranging from 1.5 to 6.7 Å and number of equivalenced interactions between 13 and 108. At the same time, as seen in Figures 1A and 1B, with the exception of the central part of two molecules, where structures differ the most, all alignments are quite close to each other, differing at most by 2–3 positions. In three regions, the second, fourth, and two last beta strands encompassing the copper ligands, all structural alignments actually coincide. Another quite obvious observation is that alignments equivalencing longer fragments of both proteins tend to have large RMSDs. This results from the strong similarity as measured by RMSD being confined to a relatively small section of the protein (the core). For this reason, extending the alignment beyond the core inevitably brings unwanted positions to the alignment.

The last statement can be justified on the basis of the RMSD similarity measure, but the contact overlap gives a slightly different picture. The contact overlap alignment is by 25% longer than the otherwise very similar core packing alignment of Chothia and Lesk (1982) (see Fig. 1B), but it aligns 44% more interactions. Clearly residues from outside the core, despite their diverging C α positions, manage to maintain very similar interaction pattern. Another interesting feature of the RMSD-based alignment, which would be discussed in more details later, is illustrated by comparing GAFIT and MNYFIT alignments (May & Johnson, 1994). With a very small, 0.1-Å difference in their RMSD score, both alignments differ at almost all positions. As would be illustrated later using visualization of suboptimal alignments, RMSD-based alignments are very unstable and can change dramatically with very small change of the similarity score.

It is also very informative to compare alignments obtained using various similarity measures. Both alignments based on side

Table 1. Comparison of several published structural alignments of azurin (1azcA) and plastocyanin (1plc) structures^a

Alignment source	RMSD	No. contacts	No. identities	Alignment length	Structural information
Contact overlap	5.1	108	15	89	Interactions
Adman (1985)	6.7	76	18	94	Visual
Taylor and Orengo (1989)	3.9	96	13	86	Environment
Chothia and Lesk (1982)	2.9	72	11	70	Core packing
DALI (1995)	4.5	68	13	93	C α distances
C α distance	2.4	72	12	80	C α distances
Gerstein et al. (1994)	5.7	75	14	96	Core packing
Pascarella and Argos (1992)	2.5	75	16	73	Hybrid
May and Johnson (1994)	4.7	73	12	89	C α RMSD
GAFIT	1.5	65	11	60	C α RMSD
MNYFIT	1.6	13	4	51	C α RMSD
Sequence	8.4	52	24	98	Gonnet

^a The entire length of azurin (plastocyanin) is 129 (99) amino acids and there are 239 (173) pair interactions in the 1azcA (1plc) structure.

chain positions, as well as the alignment based on structural environments, give remarkably similar results, with the contact based potential simply extending the core packing alignment beyond the core. All these similarity measures involve only residues from directly interacting regions, and a global similarity measure can be factorized into the contributions coming from various regions of the protein. This is to be contrasted with similarity measures that cannot be factorized, such as RMSD and, to some extent, distance matrix difference, where shifts in the alignment in one part of the structure might be necessary to compensate for some structural differences somewhere else. As a result, some fragments can be aligned with a total disregard for important local features. For instance, the first beta strand in azurin has a strong hydrophobic-hydrophilic pattern of HPHPHP (C-E-A-T-I-E). In both "side-chain" alignments, this is aligned to a similar, HPHHHP pattern in plastocyanin (I-D-V-L-L-D), with only one error. At the same time, DALI, based on distance matrix difference aligns this pattern to a null-HPHHH (.-I-D-V-L-L) pattern with four errors, which is clearly out of phase and RMSD-based alignments shift this fragment even further into the (.-.-I-D-V-L). This example also illustrates the danger of blindly using structural alignments to analyze patterns of residue conservation. Statistical analysis would conclude that hydrophobic/hydrophilic character of the residues is preserved exactly if based on "side-chain" alignments, is not conserved at all if based on the $C\alpha$ distance alignment, or is very weakly conserved if based on RMSD alignments.

This effect is illustrated in Figure 2, where a short, 13-residue fragment of two plastocyanin-azurin alignments is shown in detail. The contact map overlap alignment, shown on the left, has (for this fragment) 90% accuracy for contact map overlap, but more than 1.49 Å RMSD between backbone atoms RMSD based alignment has RMSD of 0.52 Å, but decreasing contact overlap to about 50%.

At the same time, alignments optimizing side chain-based similarity measures are prone to other problems. For instance, they are likely to distort backbone similarity by shifting fragments so as to preserve side-chain interactions. Both these tendencies are illustrated in Figure 2, where interaction overlaps and backbone superpositions are shown according to different alignments.

The sequence alignment, despite an obvious, if weak, homology between both proteins, is markedly worse than the structural alignments in both RMSD and number of predicted contacts. It is clear that at this homology level the gain of additional 6–10% in sequence similarity comes from spurious similarities and can actually be misleading for a global alignment. Still, the sequence alignment is quite close to structural ones, involving shifts of no more than 2–3 residues in the protein core.

Of course, this analysis is based on just one example of the alignment between two homologous proteins, so the observations might not be general and applicable to cases of more distant structural similarity, especially not accompanied by protein homology.

Flavodoxin-CheY chemotaxis protein

These two proteins share the same topology of an $\alpha/\beta/\alpha$ layered structure, with the 21 345 order of strands in the 5-strand central beta sheet. Flavodoxin is a low potential flavoprotein, substituting for ferredoxin as the native electron acceptor in Photosystem I in cyanobacteria under the low-iron conditions. FNM molecule is an integral part of flavodoxin molecule. Chemotaxis CheY protein is a component of a signal transduction system in bacteria. Phosphorylation of CheY changes its con-

Fig. 2. Differences between various structural alignments. A 13-residue fragment of a contact map based alignment between plastocyanin and azurin is shown on the right. On a lower part of the figure, a superposition of the contact patterns of both proteins is shown, with black squares showing contacts present in both molecules. Above it, a backbone overlap according to the same alignment is shown. On the right hand of the figure, analogous figures are shown for a $C\alpha$ RMSD based alignment. For fragments shown in the figure, the backbone RMSD is 1.49 Å for the contact based alignment and 0.52 for the RMSD one.

formation and modulates its interactions with its target within the flagellar motor. CheY does not have any prosthetic groups and it has no nucleotide-binding activity. Sequence similarity between these two proteins is about 15% of identical residues, which is practically random. There is no evidence for their homology and these two proteins probably represent an example of structurally convergent evolution, where the same structural solution was independently "discovered" by two protein families. Therefore, analysis and comparison of their alignments takes us into another region of protein similarity than the plastocyanin-azurin case discussed earlier.

As before, all the available alignments are compared directly and as a plot (Fig. 3A,B) and their differences are summarized in a separate table (Table 2). It is clear that structural differences between both proteins are somewhat larger than in the previous case; on the other hand, we do not have purely $C\alpha$ RMSD based alignments on our list. Despite the general agreement between various alignments as to equivalencing of the secondary structure elements, the diversity between various alignments is large, with not a single pair aligned in the same way in all structural alignments. At the same time, both structures display quite strong structural similarity. For instance, the number of equivalenced pair interactions in a contact based alignment is still significant, at 45% (see Table 2), which should be compared with 62% in the case of plastocyanin-azurin pair and with about 20–30% for alignment between dissimilar structures. The best RMSD (in the straight $C\alpha$ distances based alignment) is 2.93 Å

Table 2. Comparison of several published structural alignments of flavodoxin (4fxn) and CheY protein (3chy) structures^a

Alignment source	RMSD	No. contacts	No. identities	Alignment length	Structural information
Contact	4.3	107	14	108	Interactions
DALI (1995)	3.8	44	11	112	C α distances
Pascarella and Argos (1992)	4.6	48	9	82	Hybrid
Taylor and Orengo (1989)	4.6	63	13	97	Environment
C α distance	2.9	72	12	100	C α distances
C α distance	2.6	66	9	80	C α distances
Sequence	9.0	30	19	127	Gonnet

^a The entire length of flavodoxin (CheY) is 139 (128) amino acids and there are 257 (239) pair interactions in the 4fxn (3chy) structure.

for 80% of the shorter protein length (100 positions), which could be compared with 1.5 Å RMSD for plastocyanin-azurin case on one hand, and about 7–8 Å RMSD between dissimilar compact structures of this size.

As before, alignments based on structural features other than the C α position (contact and SSAP) are quite close, with only a middle section misaligned. But again, this apparently minor difference accounts for more than 40% of interactions common to both proteins. Also, as before, alignments based on the positions of C α alone differ most from all the others, and tend not to preserve the pattern of physicochemical characteristic along the sequence. Also of interest is the weighted C α distance difference used in the program DALI, which in the previous case gave results almost identical to that obtained with a simple C α distance difference, in this case, resulted in a significantly different alignment. Apparently, implicitly assumed length of the alignment, disguised as the tolerance level (see Equation 2B and accompanying discussion), in this case forces the alignment to extend beyond the common core of the two proteins, resulting in a significantly worse RMSD.

For alignments minimizing C α distance difference or RMSD, the target length of alignment must be specified in advance, as discussed in Materials and methods (unless one uses weighted C α distance difference, as in the DALI algorithm). The last two alignments in Figure 3 and Table 2 were obtained with the same target function, but with different target lengths. It is interesting to note that the difference between these two alignments is not simply adding or deleting some equivalenced pairs. Instead, forcing the alignment to include additional pairs results in shifts in the alignment in other areas.

Myoglobin-phycocyanin

The last past pair of structures that will be analyzed in detail is myoglobin and phycocyanin. Both are helical proteins, with a similar topology of a two-layered helical sandwich of eight helices. The former is an example of globins, a wide class of proteins that incorporate heme as a prosthetic group that binds oxygen reversibly. Proteins from this group are involved in oxygen transport in a variety of organisms, ranging from bacteria, insects, and plants to vertebrates. Some classes of globins (such as hemoglobins) form tetramers or dimers. Myoglobin has a distinction of being the first protein for which a complete three-dimensional structure was determined. With almost 200 solved structures and close to 2,000 known sequences, this is probably the best known protein group. Phycocyanins are ac-

cessory light-harvesting pigments, part of light-harvesting complex of cyanobacteria and red algae. Members of this group are covalently bound to its prosthetic group, linear tetrapyrrole chromophores, and in natural conditions usually form stacks of hexamers, containing two related alpha and beta chains.

Differences in the function of both proteins, together with the lack of recognizable sequence similarity between them suggest that these two protein groups are not related or that their relation is extremely distant (Pastore & Lesk, 1990). As in the previous cases, contact and side-chain packing alignments are quite close, with C α distance difference alignments differing substantially (see Figs. 4A and 4B).

As seen on the three sets of examples presented here, there is little quantitative difference between the level of disparity of different structural alignments between three groups of proteins, each coming from a different structural class (all alpha, alpha/beta, and all beta) and ranging from weakly homologous to nonhomologous.

Distribution of suboptimal alignments

For any similarity measure and alignment method used to obtain the optimal alignment, knowing just the best alignment and its score does not really tell much about the similarity between the two proteins. It can be compared with knowing just the numerical value of the "best" in a series of measurements for some experiment, without knowing the overall accuracy of the experiment, distribution of measurements, and errors of each particular measurement. Another analogy would be knowing the best path between two points, without knowing if it is going across a flat area, with many different but almost as good paths nearby, or if its going along a mountain edge, with a uniquely defined path. Interpretation of the structural alignment could be dramatically different if it would be just one among many similarly scoring alignments or if some parts of the alignment would be much better defined than others.

Distribution of suboptimal alignments provides more information about the similarity "landscape." An example presented in Figure 5 shows a planar plot, where a gray level of a box located at the position $[i, j]$ is used to show a score of the best possible alignment of position i from the first proteins with position j from the second. Alignments with scores within 10% of the best alignment are shown. Alignments with energies equal to that of the best alignment can be seen as a series of black squares. In regions where the black squares form a narrow line,

A

B

Fig. 4. Comparison of several published structural alignments of phycocyanin (1cpcA) and sperm whale myoglobin (1mbc) structures. **A:** "Dot plot" format. Different alignments are identified by points on the alignment plane. **B:** "Multiple alignment" format. The phycocyanin sequence is given in the first row, subsequent rows display the myoglobin sequence, aligned to the phycocyanin from the first row according to various structural alignments, identified by appropriate references.

Fig. 5. Density of suboptimal structural alignments for the contact overlap based structural alignments. Alignments within top 90% of the maximal score are shown. **A:** Plastocyanin-azurin. **B:** Flavodoxin-CheY protein. **C:** Phycocyanin-myoglobin. (*Continues on following pages.*)

the best alignment is unique; in regions where they merge into a wide large black area, many different alignments have the same score as the optimal one.

There are many areas where the optimal alignment is basically unique and such areas often coincide with well-defined secondary structure elements. At the same time, areas where the best alignment is not unique often coincide with loops and areas of structure without a strong local regularity. It is clear that the area between residues 45 and 60 in plastocyanin and residues 50 and 80 in azurin (Fig. 5A) could not be uniquely aligned and, to some

extent, the same is true for the large loop between residues 62 and 70 in plastocyanin and residues 90 and 100 in azurin. The structural alignments presented in Figure 1 were most divergent in the same areas.

The most interesting features are the double single lines visible in regions 1–10 and 30–40 of Figure 5A. In these areas, there are two possible alignments with almost the same score. In both cases, a beta strand from one structure can be aligned to its equivalent strand in the second structure in two different ways, shifted by two residues. Some criteria, such as hydrogen bond

pattern and RMSD, favor one alignment, whereas contact overlap favors (albeit slightly) the second alignment. As mentioned in the introduction, on examples like this, it is most visible that the question of what we mean by structural similarity might actually lead to different answers in terms of structural alignment between proteins.

Distribution of suboptimal alignments in other cases studied here gives very much the same answer—there are some areas where alignment is unique, but in others, most notably loops,

there are many possibilities of aligning two structures. The existence of alternative alignments, visible as double or even triple lines, is common, with one or two alternative alignments for each case shown here. Again there is no significant, qualitative difference between weakly homologous proteins such as plastocyanin-azurin and sequentially and functionally unrelated proteins such as flavodoxin-cheY or myoglobin-phycoerythrin, nor between proteins coming from very different structural classes.

Fig. 5. Continued.

An algorithm used to obtain the distribution of suboptimal alignment and described in Materials and methods could, in principle, be used with any similarity measure. In examples presented so far, the contact map overlap, as defined by Equation 1, was used. Using other similarity measures gives qualitatively different answer (Godzik et al., 1993).

Discussion

Following the tradition from sequence analysis, the results of comparing structures of two proteins are eventually expressed in the form of an alignment: a set of one to one equivalencies between positions in both proteins. As shown in several examples in this paper, different similarity measures used for structural comparisons may lead to different alignments. As could be ex-

pected, the structural alignment optimizing one similarity measure can be close, but by no means identical, to the alignment optimizing a different similarity measure. All structural alignments are virtually identical on the level of secondary structure elements, but start to differ by 2–4 positions on the residue level. Therefore, in principle, for the same pairs of proteins we can have the $C\alpha$ -RMSD, $C\alpha$ distance difference, contact map overlap, or some other structural based alignment. Even within any given measure, alignments are often degenerate and whole families of alignments can be generated with almost the same score.

Therefore, it is clear that there is not such a thing as the structural alignment that could be used as a standard of truth to judge and validate other alignment methods, such as threading or sequence alignments. Instead, there are different structural alignments, emphasizing structural similarity as seen by this or

other similarity measure. For closely homologous proteins, the differences between various structural alignments are minor and confined to residues outside the hydrophobic core, where sequence variations are biggest. But the picture changes dramatically for more distant homologues and unrelated proteins with similar structure. Structural alignments start to differ from each other, not only in loops and irregular fragments of structure, but also for well-defined structural elements within the core. Different similarity measures, such as C α -based RMSD or distance difference contradict other similarity measures such as H-bond networks or residue interactions. This problem influences very much the analysis of structural similarities, leading to contradictory results reported by groups using different types of alignments.

Of course, all these efforts serve as a preliminary step in answering a much more important question – why proteins fold the way they do? It is not unreasonable to expect that once we understand how protein structure changes with small and large sequence changes, it would be possible to predict such changes for point mutations, homologous proteins, and eventually for new protein structures. The existence of pairs and groups of non-homologous proteins with similar structure provides us with an important hint to point us to what in protein sequence really defines its structure. However, in order to use this hint, we have to understand first what is really similar between structurally similar proteins. Every measure of structural similarity is, in fact, based on singling out a particular structural feature (C α positions, side-chain packing, H-bonding network) as more important than others. Therefore, the ambiguity present in structural analysis manifested in differences between structural alignments obtained with different similarity measures illustrate very well the problems in folding predictions, where we do not know what features of a structure are most important for how the protein actually folded.

Materials and methods

Measures of protein structural similarity

For quantitative information about the similarity between protein structures as described by a particular alignment between the two proteins, the most widely used measure is the RMSD after optimal superposition of equivalenced positions of two structures:

$$RMSD(AB) = \sqrt{\frac{1}{N_{aligned}} \sum_{i=1}^{N_{aligned}} \left(\bar{r}^A(i) - \bar{r}^B(AB(i)) \right)^2}, \quad (1)$$

where $\bar{r}^A(i)$ is the position of the i th C α atom in the protein A, $\bar{r}^B(AB(i))$ is the Cartesian vector of the i th C α position of protein B after optimal superposition, numbered by the position in the protein A according to a particular alignment AB, with only aligned positions used in the calculations. The set of equivalenced positions $[i, AB(i)]$ is defined beforehand in the form of the alignment between two proteins, and for each set the optimal superposition can be found by the deterministic algorithm (Kabsch, 1978). Because the optimal superposition depends on the whole alignment, it is not possible to define contributions to the final score coming from individual residues.

Another measure is the difference of the distance matrices:

$$D(AB) = \frac{1}{N_{pairs}} \sum_{i>j} |d_{ij}^A - d_{AB(i)AB(j)}^B|, \quad (2)$$

where distances between C α atoms i and j in protein A is compared with the distance between C α atoms of equivalenced atoms in protein B, i.e., $AB(i)$ and $AB(j)$. The problem with both similarity measures defined in Equations 1 and 2 is that they have a trivial minimum at zero equivalenced residue pairs. For every alignment, its score can be improved by deleting positions with the largest deviation. Therefore, for both measures, the target length of the alignment must be specified in advance. Making this number too large may result in a significantly shifted alignment, where to accommodate very distant pairs, some other fragments of the alignment would be shifted. On the other hand, using a small target length of the alignment leads to omitting potentially important fragments from the alignment. Some authors (Pascarella & Argos, 1992) use a buildup procedure, to extend the alignment until a target RMSD value is reached. Others (Holm & Sander, 1993) modified the simple distance difference, as used in Equation 2, to compare relative, rather than absolute, distortions and to exclude very distant pairs from calculations. For instance, in program DALI (Holm & Sander, 1993), the following formula was used:

$$D(AB) = \frac{1}{N_{pairs}} \sum_{i>j} \left(0.2 - \frac{|d_{ij}^A - d_{AB(i)AB(j)}^B|}{d_{ij}^*} \right) \exp\left(-\frac{d_{ij}^{*2}}{\alpha^2}\right), \quad (2A)$$

where d_{ij}^* is the geometric mean between the two distances, and by using α equal to 20 Å, contribution from more distant pairs was weighted down. The tolerance level, set to 0.2, is an indirect target for the expected difference between equivalent distances in both proteins.

Another approach is to focus directly on interactions between amino acid side chains; one can employ the contact map description of the protein structure. The overlap of two contact maps, as defined by the number of contacts between equivalenced residues in two proteins that are simultaneously present in both structures (Godzik et al., 1993) can be used as a measure of similarity between two protein structures. It is calculated by the following equation, where the summation is taken over all aligned pairs.

$$S_C(AB) = \frac{1}{N_c} \sum_{i>j} C_{ij}^A C_{AB(i)AB(j)}^B, \quad (3)$$

where the C_{kl}^A is the binary contact map of proteins A and B, respectively, $C_{AB(i)AB(j)}^B$ is a contact map of a protein B, but numbered according to the alignment AB. Finally, the normalization factor N_c is the maximal possible overlap between the two maps. Figure 2 shows the overlap between contact maps of structurally similar proteins. This similarity measure does not have to have a predefined length, and it is possible to factorize a total score to contributions coming from different fragments of the protein.

It is also possible to calculate alignment score based on still different measures of structure similarity, such as continuous C α -C α distances, H-bond networks, and many others.

Finding an optimal alignment

Any of the equations given above can be used to quantify the similarity between two protein structures using an already existing alignment. However, the problem of finding the optimal alignment is not trivial. Because in each of the scoring functions

1–4 the gain or penalty for equivalencing amino acids at positions i and $AB(i)$ depends on the alignment at other positions, the problem of finding the best alignment in NP complete (Lathrop, 1994).

As yet, there is no fast and deterministic solution to this problem. There are, however, a number of algorithms that can produce a reasonable answer (Taylor & Orengo, 1989; Holm & Sander, 1993; Godzik & Skolnick, 1994).

Groups of structurally similar but sequentially unrelated proteins

Using one such approximate algorithm, one can prepare a structural alignment that would be reasonably close to the optimal alignment with a chosen similarity measure. Comparing a score of this alignment to a distribution of scores between unrelated proteins in the database, the structural similarity between our two proteins can be rated as significant or marginal. For instance, using the contact map overlap defined by Equation 3, contact maps of topologically related proteins can be superimposed on the level of 50–70% contact overlap. Figure 2A shows the superposition between the contacts maps of plastocyanin-azurin, two copper-binding Greek key beta proteins involved in electron transfer in plants. The sequence similarity expressed as a percentage of identical residues in the sequence based alignment is close to random: 22.5% for the poplar plastocyanin-algae azurin comparison (Genetics Computer Group, 1991). The contact map overlap, however, is substantial, with 65% overlap. This should be compared with about 80% overlap between independently solved structures of highly homologous proteins and about 20% overlap between proteins with different topology. Therefore, contact map overlap between sequentially unrelated proteins is clearly distinguishable from the random overlap and quite close to the overlap between close sequential homologues. This makes this measure of structural similarity different from the RMSD between core residues, which was shown to increase rapidly with changes in sequence (Lesk & Chothia, 1980).

Properly scaled similarity scores can be used to cluster the protein structure database into groups of structurally similar proteins, or even to build similarity trees dividing protein structures into similar, close, or distant structural cousins. This strategy was used by several groups to identify groups of proteins with similar structures (Blundell & Johnson, 1993; Holm & Sander, 1993; CATH, 1995; Fischer et al., 1995).

Another very extensive classification was created using a “human classification” system (SCOP, 1995). Several other, less complete lists of structurally similar proteins were published earlier (Pascarella & Argos, 1992). Finally, many structurally similar proteins that were discovered by individual researchers are subjects of separate publications.

All this data, available in a published form and/or over the Internet, was used here to identify potentially interesting protein pairs and groups used in this analysis.

Visualization of protein alignments

As mentioned earlier, an alignment is a set of equivalencies between positions in one protein and positions in the second protein. Such a set can be represented as a set of points $[x, y]$ on the plane, where x and y position of each point is equivalent to its position in the first or second protein, respectively. A path linking a set of such points is equivalent to the linear alignment,

as usually presented in the literature (Godzik & Skolnick, 1994). Such a path can be visualized on the plane by linking consecutive alignment points. There is an obvious analogy between depicting protein alignments in this way and an alignment technique known as a “dot-plot” (Gribskov & Dereveux, 1991). Several other alignment programs used a similar visualization technique (Vingron & Argos, 1990).

Depicting alignments on the plane makes it much easier to compare different alignments of the same pair of proteins to each other. Paths representing different alignments would overlap in the areas where both alignments coincide and would diverge in areas where they differ. Distance between paths represents the amount of shift between alignments.

Structural alignments of proteins

Once pairs or groups of proteins with similar structure are identified, one can prepare pairwise or multiple structural alignments of all proteins within the group. In at least one case, a constantly updated database of protein structural alignments (DALI, 1995) was set up on the Internet and made available to the public. This particular database and alignments created by the program DALI were used extensively as a reference in this paper.

Several other groups also made some alignments publicly available. For instance, Sali and Overington created a series of structural alignments for 36 structural families (Sali & Blundell, 1990). In a series of elegant papers Lesk, Chothia, and their co-workers published alignments for several better known examples of structurally similar but sequentially unrelated proteins, based on the analysis of packing in the common core of the proteins. Many other alignments, obtained by methods ranging from some less known structural alignment programs to alignments by hand, were published in papers discussing some specific proteins. Finally, Pascarella and Argos put together a database of such structural alignments, adding some alignments prepared by a hybrid method, including sequence information and iterative extending an existing “seed” alignment using $C\alpha$ RMSD criteria. Unfortunately, some very well known structure comparison programs, including the SSAP program by Taylor and Orengo (1989), and the STAMP program developed by Russell and Barton (1994), could not be systematically included in the current analysis, because most of their alignments are not available publicly. To complement this list, additional alignments were created using a MC alignment optimization program (Godzik & Skolnick, 1994), which uses the analogy between a lattice chain conformation and an alignment between two or more proteins. Two similarity measures, contact map overlap as defined by Equation 3 and simple $C\alpha$ distance difference as defined in Equation 2, were used.

Analysis of sub-optimal structural alignments

Most alignment algorithms, both sequence and structure based, aim at finding the best possible score and the optimal alignment within a given scoring system. But knowing the optimal alignment itself does not tell everything about the similarity between both proteins. The equally important question concerns the uniqueness and stability of the best alignment in respect to the particular scoring system used, and, in the case of structural alignments, also to small changes in structure.

For instance, if there are many quite different alignments with the same or very close score, one can argue that the alignment is not stable, and any observations or statistics obtained

from this particular alignment might be rather arbitrary. Of course, an alignment can be uniquely defined (i.e., stable) in some areas (such as the hydrophobic core) and ubiquitous in others. Identifying regions where alignments are strong and unique, and where they are not well-defined and vague, would be logically equivalent to knowing error bars for a set of experimental points. Similar questions in the sequence alignments are addressed by the analysis of suboptimal alignments (Vingron & Argos, 1990; Saqi & Sternberg, 1991; Zuker, 1991).

In this method, each pair $[x, y]$, where x and y represents a position in the first or second protein, respectively, is assigned a value equal to a score of the best alignment that incorporates this particular pair. Now the path equivalent to the best alignment is visible on the "alignment plane" as the edge of a "highest mountain range." Areas where the best alignment is unique are visible as narrow strips; areas where there are many degenerate solutions appear as wide "plateaus." Representing alignments with scores within some fixed distance from the highest score in shades of gray or with different symbols, it is possible to identify regions where the best alignment is stable by visual inspection. Such representation of suboptimal alignments was used previously for sequence alignments (Vingron & Argos, 1990; Saqi & Sternberg, 1991; Zuker, 1991), structural alignments (Godzik et al., 1993), and threading alignments (MatchMaker, 1994; A. Godzik, in prep.).

In sequence and threading alignments, which use dynamic programming, it is possible to identify suboptimal alignments by summing up instantaneous alignment scores for the regular and reverse alignments (Vingron & Argos, 1990), a technical trick that was used in all other implementations (Saqi & Sternberg, 1991; Zuker, 1991; MatchMaker, 1994; A. Godzik, in prep.). In structural alignments, at least as implemented here, it is not possible; therefore, a different strategy was used. After a stable solution was obtained in an MC-based simulated annealing procedure, a long, high-temperature isothermal MC run was performed. During this run, the system explored the alignment basin in the vicinity of the best alignment. During such a run, statistics about the highest score of the alignment incorporating a given $[x, y]$ pair was gathered and subsequently displayed in a modified version of a dot plot, as seen, for example, in Figure 5. It is important to note that while the information about the suboptimal alignments obtained from MC simulation in alignment space is formally equivalent to that obtained from dynamic programming (Vingron & Argos, 1990; Zuker, 1991), it was obtained in the nondeterministic procedure, and therefore it represents only an approximation to the real distribution of suboptimal alignments.

Acknowledgments

This research was supported by NSF grant MCB-9506278.

References

- Adman ET. 1985. Structure and function of small blue copper proteins. In: Harrison P, ed. *Metalloproteins, part I*. London: Macmillan. pp 1–42.
- Blundell TL, Johnson MS. 1993. Catching a common fold. *Protein Sci* 2: 877–883.
- Bowie JU, Luethy R, Eisenberg D. 1991. A method to identify protein sequences that fold into a known three dimensional structure. *Science* 253:164–170.
- CATH. 1995. Protein structure classification. London: UC BSM. <http://www.biochem.ucl.ac.uk/bsm/cath/>.
- Chothia C. 1992. Proteins. One thousand families for the molecular biologist. *Nature* 357:543–544.
- Chothia C, Lesk AM. 1982. Evolution of proteins formed by β -sheets. I. Plastocyanin and azurin. *J Mol Biol* 160:309–324.
- DALI. 1995. Protein structure comparison by alignment of distance matrices. Heidelberg. <http://www.embl-heidelberg.de/dali/>.
- Finkelstein AV, Ptitsyn OB. 1987. Why do globular proteins fit the limited set of folding patterns? *Prog Biophys Mol Biol* 50:171–190.
- Fischer D, Bachar O, Nussinov R, Wolfson H. 1992. An efficient automated computer vision based technique for detection of three dimensional structural motifs in proteins. *J Biomol Struct Dyn* 9:769–789.
- Fischer D, Tsai CJ, Nussinov R, Wolfson H. 1995. A 3D sequence independent representation of the Protein Data Bank. *Protein Eng* 8:981–997.
- Genetics Computer Group. 1991. *Program manual for the GCG package, version 7*, Madison, WI.
- Gerstein M, Sonnhammer E, Chothia C. 1994. Volume changes on protein evolution. *J Mol Biol* 236:1067–1078.
- Godzik A, Skolnick J. 1994. Flexible algorithm for direct multiple alignment of protein structures and sequences. *CABIOS* 10:587–596.
- Godzik A, Skolnick J, Kolinski A. 1992. A topology fingerprint approach to the inverse folding problem. *J Mol Biol* 227:227–238.
- Godzik A, Skolnick J, Kolinski A. 1993. Regularities in interaction patterns of globular proteins. *Protein Eng* 6:801–810.
- Gribskov M, Dereux J. 1991. *Sequence analysis primer*. New York: Stocton Press.
- Holm L, Sander C. 1993. Protein structure comparison by alignment of distance matrices. *J Mol Biol* 233:123–138.
- Jones DT, Taylor WR, Thornton JM. 1992. A new approach to protein fold recognition. *Nature* 358:86–89.
- Kabsch W. 1978. A discussion of the solution of the best rotation to relate two sets of vectors. *Acta Crystallogr A* 34:827–828.
- Lathrop RH. 1994. The protein threading problem with sequence amino acid interaction preferences is NP-complete. *Protein Eng* 7:1059–1068.
- Lemer C, Rooman M, Wodak S. 1995. Protein structure prediction by threading methods: Evaluation of current techniques. *Proteins Struct Funct Genet* 23:337–355.
- Lesk AM, Chothia C. 1980. How different amino acid sequences determine similar protein structures. *J Mol Biol* 136:225–270.
- Mayorov VN, Crippen GM. 1994. Significance of root-mean square deviation in comparing three-dimensional structures of globular proteins. *J Mol Biol* 235:625–634.
- MatchMaker. 1994. *MatchMaker 1.0*. St. Louis, Missouri: Tripos Associates.
- May AC, Johnson MS. 1994. Protein structure comparison using a combination of a genetic algorithm, dynamic programming and least-square minimization. *Protein Eng* 7:475–485.
- Mitchell EM, Arytmiuk PJ, Rice DW, Willet P. 1990. Use of techniques derived from graph theory to compare secondary structure motifs in proteins. *J Mol Biol* 212:151–166.
- Orengo CA, Flores TP, Jones DT, Taylor WR, Thornton JM. 1993. Recurring structural motifs in proteins with different functions. *Curr Biol* 3:131–139.
- Pascarella S, Argos P. 1992. A data bank merging related protein structures and sequences. *Protein Eng* 5:121–137.
- Pastore A, Lesk AM. 1990. Comparison of the structure of globins and phycocyanins: Evidence for evolutionary relationship. *Proteins Struct Funct Genet* 8:133–155.
- Russell R, Barton G. 1994. Structural features can be unconserved in proteins with similar folds. *J Mol Biol* 244:332–350.
- Russell RB, Barton GJ. 1992. Multiple protein sequence alignment from tertiary structure comparison: Assignment of global and residue confidence levels. *Proteins Struct Funct Genet* 14:309–323.
- Sali A, Blundell TL. 1990. Definition of general topological equivalence in protein structures. A procedure involving comparison of properties and relationships through simulated annealing and dynamic programming. *J Mol Biol* 212:403–428.
- Saqi M, Sternberg M. 1991. A simple method to generate non-trivial alternate alignments of protein sequences. *J Mol Biol* 219:727–732.
- SCOP. 1995. Structural classification of proteins. <http://www.bio.cam.ac.uk/scop>.
- Taylor WR, Orengo CA. 1989. Protein structure alignment. *J Mol Biol* 208:1–22.
- Vingron M, Argos P. 1990. Determination of reliable regions in protein sequence alignment. *Protein Eng* 3:565–569.
- Vogt G, Etzold T, Argos P. 1995. An assessment of amino acid exchange matrices in aligning protein sequences: The twilight zone revisited. *J Mol Biol* 249:816–831.
- Zuker M. 1991. Suboptimal sequence alignment in molecular biology. Alignment with error analysis. *J Mol Biol* 221:403–420.