Structural similarity of DNA-binding domains of bacteriophage repressors and the globin core

S. Subbiah, D.V. Laurents and M. Levitt

Beckman Laboratories for Structural Biology, Departments of Cell Biology and Biochemistry, Stanford University school of Medicine, Stanford, California 94305-5307, USA.

Background: In recent years, the determination of large numbers of protein structures has created a need for automatic and objective methods for the comparison of structures or conformations. Many protein structures show similarities of conformation that are undetectable by comparing their sequences. Comparison of structures can reveal similarities between proteins thought to be unrelated, providing new insight into the interrelationships of sequence, structure and function.

Results: Using a new tool that we have developed to perform rapid structural alignment, we present the highlights of an exhaustive comparison of all pairs of protein structures in the Brookhaven protein database. Notably, we find that the DNA-binding domain of the bacteriophage repressor family is almost completely embedded in the larger eight-helix fold of the globin family of proteins. The significant match of specific residues is correlated with functional, structural and evolutionary information.

Conclusion: Our method can help to identify structurally similar folds rapidly and with high-sensitivity, providing a powerful tool for analyzing the ever-increasing number of protein structures being elucidated.

Current Biology 1993, **3**:141–148

Background

Although the number of protein structures deposited in the Brookhaven protein database (PDB) has grown rapidly in recent years [1], the subset of new protein folds has grown at a significantly slower rate [2]. This rate difference still persists after allowing for the many structural determinations of homologous, mutant and drug-complexed versions in the same basic protein family. Therefore, assuming there is no systematic bias in the selection criteria in deciding which particular protein structure is to be determined, it has been suggested that we are 'closing-in' on the complete repertoire of folds that are allowable from the multitude that constitute all possible protein structures [3]. The limited number of these folds may be due to evolution: once there are enough folds to create all possible protein functions there is then no pressure to evolve new folds. On the other hand, the limit to the number of folds may be due to the existence of basic structural limitations that dictate, and thus relate, the three-dimensional structures of proteins. Finding and understanding such principles of protein construction will help in the design of new and variant proteins.

Assuming that the reservoir of unobserved folds is depleting rapidly, any structural constraints should be detectable in the structural database presently available to us. Suitable and exhaustive comparisons of these unexpected similarities that could help catalogue and, perhaps, define structural principles. In this context, it is worth noting that analogous studies of the one-dimensional DNA and protein sequence databases, made possible by the development of elegant computer algorithms, have borne much fruit in identifying and cataloguing many novel sequence motifs of functional interest [4,5]. With regard to the problem of comparing two different three-dimensional protein structures considered here, despite early (and more recently plentiful) work in the development of suitable computer algorithms, systematic studies have been limited [6–11]. Many of the available methods have been hampered by limitations in accuracy, speed and sensitivity.

Here we present a new method for protein structure comparison that is accurate, fast and sensitive. Using this improved tool, we present the highlights of an exhaustive comparison of all pairs of protein structures in the PDB. The discovery of a significant structural similarity between two well-studied protein families, the bacteriophage repressors and the globins, emphasizes the power of our method. With its speed and sensitivity, it can aid the crystallographer and NMR spectroscopist in rapid identification of the relatedness of a newly determined structure to all previously reported ones. Such discoveries will in turn help to identify the rules that govern higher order structural motifs.

Results

Aligning structures

Our method aligns two protein structures by starting with an arbitrary equivalence of residues that are superimposed in three-dimensions. A structural alignment matrix, which is calculated from distances between pairs of residues that are not in the same protein, is searched to achieve the optimal alignment. This gives

Correspondence to: M. Levitt

Fig. 1. A stereo view of the best structural alignment produced by our method between the Cα backbone of azurin (1AZU) and plastocyanin (2PCY). The matched residues of 1AZU are coloured green and other residues yellow; the matched residues of 2PCY are coloured red and other residues blue. For this alignment, the cRMS value is 2.80 Å, for 89 matching Cα atoms.

a new set of equivalent residues, which are again superimposed, and the procedure is iterated to convergence (see Materials and methods).

Known cases

To test the accuracy of this method, we used the technique to align some familiar cases that are known to be at the extreme limit of detectability. Our alignment of the two copper-binding metalloproteins, azurin (PDB entry 1AZU) and plastocyanin (2PCY) (Fig. 1), agrees with the findings of Taylor and Orengo [8], but differed from the earlier alignment results of Chothia and Lesk [12] for 2PCY residues 45–65. In addiiton, our list of the twelve PDB entries that are most related to hen egg-white Lysozyme (2LYM), as sorted by our structural alignment score (SAS), illustrates a sensible rank ordering of related structures in the database (Table 1). Significantly, the seventh entry in our SAS list correctly identifies the structural similarity of T4 lysozyme (2LZM) and hen egg-white lysozyme. This similarity, which is commonly believed to be at the limit of detection, can be considered as defining the boundary between convergent and divergent evolution [13]. Using standard methods, the root mean square (RMS) of equivalent Cα atoms (cRMS) for 2LZM and hen egg-white lysozyme is found to be **4.82** Å over 89

Table 1. Best matches to hen egg-white lysozyme (2LYM 129).

Protein	N	cRMS	SAS	I%	N_{brk}	n	Biological name	Source
2HFL	126	0.42	**0.34**	**99.2**	0	554	Fab HyHEL-5/lysozyme	Mouse/chicken
2LZ2	119	0.43	0.36	94.9	0	129	Lysozyme	Turkey
3HFM	123	0.46	0.37	100.0	0	558	Fab HyHEL-10/lysozyme	Mouse/chicken
1LYM	122	0.47	0.38	100.0	0	258	Lysozyme	Hen egg-white, monoclin
1LZ1	123	0.54	0.44	61.7	1	130	Lysozyme	Human
1ALC	115	0.97	0.84	36.5	4	122	α-Lactalbumin	Baboon
2LZM	89	4.82	5.41	6.7	8	164	Lysozyme	Bacteriophage T4
2PRK	98	6.06	6.19	10.2	10	279	Proteinase K	Tritirach. alb. limb
8ADH	81	5.47	6.75	2.4	8	374	Alcohol dehydrogenase (apo)	Horse liver
1GP1	81	5.50	6.79	4.9	11	368	Glutathione peroxidase	Bovine
2RUB	85	5.86	6.89	9.4	9	882	Rubisco	*Rhodospirillum rubrum*
1PPD	78	5.45	6.99	6.4	7	212	Papain D	Papaya

The top 12 matches found by our method to the 129 residue hen egg-white lysozyme structure (2LYM), using the 295 most representative coordinates sets from the entire July 1991 release of the Brookhaven protein structure database. The list is in order of decreasing similarity as measured by structural alignment scores (SAS). The PDB entry identifier of the structure is given on the left; N is the number of equivalenced residues between the matched protein and 2LYM; I% is the percentage sequence identity for the equivalenced residues; N_{brk} denotes the number of gaps in the alignment; n is the total number of residues in the matched structure. The first five matched structures are all lysozyme or Fab-complexes of lysozymes, the sixth is α-lactalbumin, a member of the lysozyme family, and the seventh is T4 lysozme, which is clearly similar to hen egg-white lysozyme in the active site, but the structures into which the active sities are embedded are dissimilar.

equivalent Cα atoms. By our method we obtained an SAS value of 5.41 Å — this lies in the middle of the SAS range of 5–7 Å that our exhaustive study of the database implicates as indicating the cutoff for probable structural relatedness. For example, the recently discovered similarity between ubiquitin (1UBQ) and ferrodoxin (3FXC) has a cRMS of 2.1 Å for 47 equivalenced Cα's[10]. Our method found an SAS of 4.14 Å with a cRMS of 2.62 Å for **64** equivalenced Cα atoms, which, at least technically, is slightly better than the originally reported value of 4.47 Å (100 x (2.1/47); see Materials and methods).

Searching a database of 295 structures, we also found that ferredoxin is the structure most closely related to ubiquitin — the entire search took only 20 minutes of cpu time using a single processor of a 25 Mhz Silicon Graphics Iris 4D/240 workstation. This demonstrates the method's utility in rapidly identifying any similarity between a newly determined structure and previously reported ones.

The globin and repressor folds are well characterized

We now present an example of unexpected and significant similarities between structures in the database, which were found after conducting a study of all possible pairwise alignments. Sperm whale myoglobin is a member of the globin family of folds that includes myoglobins, hemoglobins, erythocrourins, leghemoglobins, and plant phycocyanins; its X-ray structure was determined 35 years ago [14]. The structure of this heme-binding heme-binding, all-helical fold is the most extensively studied of all protein folds: more than 400 sequences and 12 X-ray structures of globin folds from different species have been determined, and there have also been extensive theoretical studies of globin fold architecture [15–17]. Globins vary in size from 132–157 residues, with 145 being typical of a monomeric protein [18]. Usually, each monomer is composed of eight α-helices labeled A through to H; the short helix D is not present in some globins, such as the hemoglobin a-chain. To a first approximation, helices C, E, F, and G provide the residues that line the heme-binding pocket, whereas helices A, B, and D, when present, do not contact the heme. In many globins, four monomers tetramerize using two portions of the monomer surface: the loop between helices F and G together with the nearby carboxy-terminal end of the last helix H; and the carboxy-terminal half of helix G, the amino-terminal third of helix H and the intervening loop between these two segments.

Similarly, the bacteriophage repressor family of proteins, which includes λ cro, λ repressor, 434 cro, 434 repressor, P22 cro and P22 repressor, have been structurally well-characterized [19]. Of the four X-ray structures available, λ cro, λ repressor, 434 cro and the DNA binding amino-terminal domain of 434 repressor (1R69), the latter three are composed of five helices that fold into a compact ball. λ cro has the first three helices, including the DNA-binding helix-turn-helix (HTH) motif, but the remaining two helices are replaced by b-strands. Structurally, 434 cro and 434 1R69 are almost identical, having no relative insertions or deletions, 52 % sequence identity, and a cRMS deviation of 0.79 Å. Structurally, λ repressor differs from the two 434 proteins in that it has a relative insertion of three residues in the loop between helices 1 and 2, an additional residue in the loop between helices 3 and 4 and an additional dimerization helix (number **6);** the Cα RMS deviation over the structurally equivalent residues in all five helices between 1R69 and λ repressor is 1.79 Å (residues 9–76 of λ repressor overlap all of 1R69, with the dimerization helix not included). The three helices of λ cro are almost identical to the corresponding helices of λ repressor. Over 70% of the residues in P22 repressor and cro proteins, and 434 repressor and cro proteins, are similar, and there are no significant insertions or deletions.

Thus, this HTH bacteriophage repressor family is closely knit. The crystallographically best-studied protein, 1R69, can be taken as the archetypal structure [20]. Both cro and repressor proteins bind as dimers to the same 2-fold symmetric DNA operator site. Although the carboxy-terminal domain of the intact 434 repressor is primarily responsible for the dimerization involved in the cooperativity of binding to the operator DNA, adjacent amino-terminal 1R69 domains make significant dimer contacts in the X-ray structure of the protein-DNA complex. To a first approximation, this symmetry-related dimer interaction involves the carboxy-terminal half of the loop preceding helix 4 and the adjacent small fifth helix.

The globin and repressor folds are similar

To our surprise, our method found that a striking structural similarity exists between these two families of proteins. Searching with 1R69 against our database of 295 structures, alignments between protein strucure were made and a list of the top matches are shown in Table 2. Apart from the obvious matches to three other repressor proteins, significant matches (with SAS values of 6-7) are found to seven proteins in the globin family. The best match is to a hemoglobin (3HHB) with 54 equivalent residues, a cRMS of 3.26 Å, and an SAS of 6.03 Å; the worst match is found to erythrocruorin with 57 equivalent residues, a cRMS deviation of 3.87 Å and an SAS value of 6.93 Å. The actual superpositions are identical for all these globins and are illustrated in stereo for 3HHB (Fig. 2) and in cartoon form in Figure **3.** Both views show that the five helices of 1R69, helices l-5, can be superimposed remarkably well onto the longer helices A, B, E, G and H of the globin (accordingly, helices D, C and F have no counterpart in 1R69). Helices D, C and F, as well as the unmatched portions of the five superimposed helices, can be viewed simplistically as large insertions into the smaller 1R69 fold (Fig. 3). In such a view, the turn between helices 2 and **3** of 1R69 — the classic turn of the HTH DNA-binding motif — 'receives' an insertion corresponding to the

Table 2. Best matches to to phage 434 repressor.

Protein	N	cRMS	SAS	I%	N_{brk}	n	Biological name	Source
2OR1	63	0.49	0.78	100.0	0	126	Repressor	434, 1-69/ORI
2CRO	63	0.79	1.26	52.3	0	65	Cro	Phage 434.
1LRP	60	1.79	2.98	28.3	2	89	λ Repressor	Phage
3HHB	54	3.26	6.03	12.9	5	295	Hemoglobin (deoxy)	Human
2DHB	54	3.34	6.18	12.9	5	295	Hemoglobin (deoxy)	Horse
1LH4	57	3.66	6.42	14.0	4	157	Leghemoglobin (deoxy)	Yellow lupin
2LHB	58	3.74	6.45	3.4	2	154	Hemoglobin V (cyn., met.)	Lamprey
1HDS	55	3.55	6.46	20.0	4	588	Hemoglobin (sickle cell)	Deer
2LZM	52	3.46	6.66	9.6	6	164	Lysozyme	Bacteriophage T4
5MBN	57	3.87	6.78	10.5	4	157	Myoglobin (deoxy)	S.whale
1ECD	57	3.95	6.93	8.7	4	140	Erythrocruorin (deoxy)	*Chironomous thummi*
1CTF	52	4.00	7.69	1.9	5	68	L7/L12 50S Ribosomal Protein (C)	*E.coli*
9PAP	51	4.20	8.24	5.8	6	212	Papain (oxidized cys25)	Papaya
4FD1	55	4.55	8.28	3.6	4	106	Ferredoxin	*Axobacter vinelandii*
2CHA	53	4.45	8.39	5.6	8	236	α-Chymotrypsin (tosyl)	cow
2MHR	49	4.12	8.41	2.0	3	118	MyoHemerythrin	Sipuncular worm
1HRB	47	3.98	8.46	4.2	3	113	Hemerythrin B	Marine worm
2LYM	53	4.53	8.55	0.0	5	129	Lysozyme (1 atm)	Hen egg-white

The 18 structures from the PDB that best match the repressor structure, 1R69, are listed in order of decreasing similarity, as in Table 1. The best match is to 1R69 complexed to DNA, the next two are closely related repressor structures, and the several next best matches include a series from the goblin family: hemogoblin, leghemoglobin, myogoblin and erythrocruorin.

carboxy-terminal half of helix B plus the loop between B and C, helix C, the loop between C and D, helix D, the loop between helix D and E and the amino-terminal half of helix E. Similarly, the loop between helices 3 and 4 of 1R69 receives the insertion corresponding to the loop between helix E and helix F plus helix F, the loop between helices F and G and the amino-terminal half of helix G.

Interestingly, with the exception of portions of helix E, these additions include all the structural elements that are necessary and sufficient to line the heme-binding

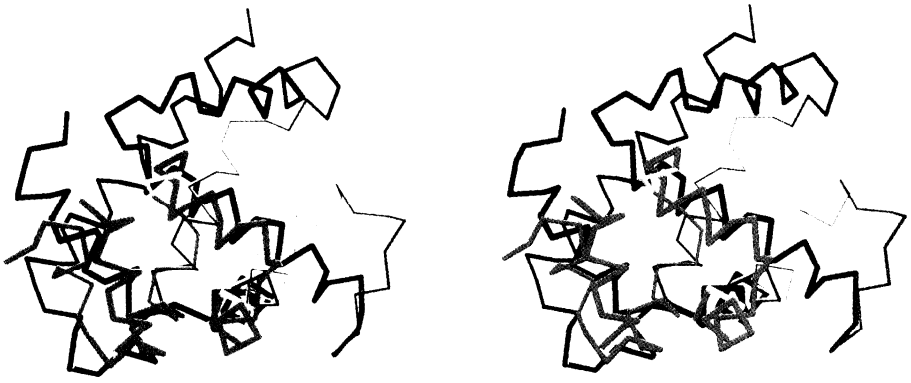

Fig. 2. A stereo view of the best structural alignment produced by our method between the Cα backbone of hemoglobin (3HHB) and the amino-terminal DNA-binding **domain of** 434 **repressor** (1R69). The **matched residues of** 1R69 **are coloured green and other residues** are yellow; **the matched residues of** 3HHB are coloured red and other residues blue. Note that the short helix D is missing in 3HHB. For **this alignment, the** cRMS **value is** 3.26 Å, **for** 54 **matching** Cα **atoms.**

pocket of the globin. The central portion of helix E, in particular residues E7, E10, E11 and E14, makes the only helix-E contacts to the heme in at least eight of the nine globin structures studied by Lesk and Chothia [15]. In our superimposition, these helix-E residues map onto 1R69 in the vicinity of the turn before helix 3, and helix 3 residues 29, 30 and 33 (Fig. 3). It is intriguing to note that this portion of helix 3 (residues 28–30 and 33) provides the primary mode of sequence-specific recognition of DNA by repressor [20]. Therefore, the heme and DNA-binding sides of both proteins are located in the same region of space after the superposition.

Curiously, there also appears to be similarity in the portions of the structures used in quartenary interactions between monomers (Fig. 3). Most of the quartenary interactions in the globins come from the carboxy-terminal half of helix G, helix H and the intervening loop (the remaining globin interactions involve portions of the globin structure that can be viewed as insertions in 1R69). These three elements of globins correspond in 1R69 repressor to all of helix 4, all of helix 5 and the intervening loop in 1R69. Significantly, all the 1R69 dimer contacts are in either helix 4 or helix 5. The globin structural elements that are involved in quartenary interactions are distant from the heme-binding pocket. Similarly, in 1R69 the dimer contacts are made by structural elements distant from the principal DNA-binding 'recognition' helix, helix 3. One could argue that the oligomerization interface would have to be away from the binding site so that it does not sterically impede binding; it is noteworthy that in our alignment, the heme-binding portion of globin consists of two regions inserted before and after this 'recognition' helix.

In a study of helix packing, Efimov [21] noted that the arrangement of helices in globins is similar to the arrangement in λ repressor. A careful inspection of his results shows that λ repressor helices 1–4 and 6 correspond to hemoglobin helices C and E-H. This is quite different from that found here: repressor helices 1–5 correspond to hemoglobin helices A, B, E, G and H. The best structural alignment of λ repressor and hemoglobin (3HHB) helices C-H gives an RMS deviation of 5.0 Å for 53 matches, which is not significant. Our alignment of λ repressor and 3HHB gives a highly significant RMS deviation of 3.1 Å for 58 matches. Although Efimov noted the general similarity of repressors and globins, he matched helices incorrectly. It is only with the present precise correspondence of repressor and globin residues that one can infer a significant similarity beyond that expected from the packing of helices.

Given that the globin monomer has roughly twice as many residues as the 1R69 monomer, it is reasonable to suggest that the 1R69 portion of the globin forms some kind of stable structural core with a capacity to

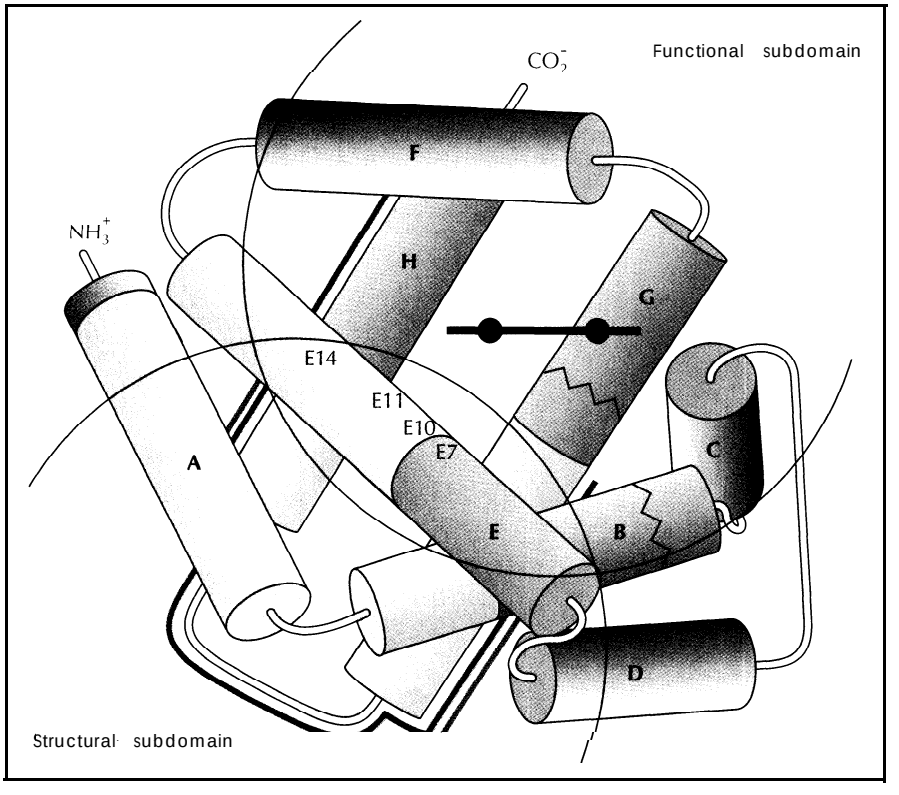

Fig. 3. The 8 helices of the globin fold, labeled A-H, are shown in cartoon form and in a view similar to that in Figure 2. The portions of these helices that correspond to the 5 helices of the IR69 repressor in our superimposition are coloured orange. The heme molecule can be seen edge-on as the solid red horizontal bar centered on the partial circle that encompasses the functional subdomain of globin. Note that the overlap between the structural subdomain and the functional subdomain of globin involves helix E and its principal contacts to the heme, residues E7, E10, E11 and E14 only. The corresponding repressor helix is the DNA-recognition helix, helix 3 and specifically residues 29, 30 and 33. The globin secondary structural elements that are primarily involved in quartenary interaction between neighboring monomers in hemoglobin are marked by the purple line. The portions of IR69 involved in dimer interaction on binding to the operator are marked by the blue line. In both cases, the surfaces of the secondary structure elements are used for quaternary contacts. The two jagged lines that divide helix B (residues B12–B13) and helix G (residues G6–G7) mark the beginning and end, respectively, of the second of the three exons that comprise the hemoglobin gene. This exon encodes all of the hemoglobin structure that makes significant direct contacts to the heme. This figure is adapted from [15,16].

interact with other monomers, while the other half of globin constitutes a heme-binding pocket that has been grafted on to the 1R69 five-helix core framework. However, any arguments for common ancestry between the two proteins based on the sequence similarity are not convincing, because the sequence similarity over equivalenced residues is insignificant (between **3** and 20%, see Table 2), particularly after allowing for the general dominance of hydrophobic residues in protein cores. This leaves open the question of whether the five-helix 1R69 motif is the structural core of the eight-helix globin fold.

Discussion

The possibility of a simple and general theory of folding for stable, all-helical and ball-like structural cores has been addressed by Murzin and Finkelstein [22]. They proposed that well-packed globular bundles of idealized helices of similar lengths can be described by ideal regular Greek polyhedra (Fig. 4). Based on notions of good packing, they argued that structural cores were limited to between three and six helices and can be represented by a series of polyhedra: octahedron, dodecahedron, sextadecahedron and icosahedron. For instance, allowing for the different loop connections between helices, suitable ribs selected from the edges of a sextadecahedron should be able to represent the axes of the five helices in an ideal five-helix core. Additional helices, as in the globins, would be accommodated as additional layers about the central helical core. In 1988, Murzin and Finkelstein [23] considered the 43 then-known cases of helical cores from the protein structure database and systematically assayed their fit to an idealized helical core inscribed in an appropriate polyhedron. Except for two proteins (calcium-binding parvalbumin, 3CPV, and the **6** major helices of the globin fold, 2MHB), the overall deviation of the real helix axes from those in the model polyhedra were all under the theoretically expected error of 3 Å.

The cases that did not fit the theory both involved six helices; Murzin and Finkelstein were able to delete the single offending helix and obtain a much better fit between the remaining five-helix core and a sextadecahedron. In particular, deletion of helix F in globin, decreased the overall error from 4.3 Å to 2.6 Å. This led them to suggest that the five helices A, B, E, G and H of globin form the structural core of the globin fold. Our independent alignment of protein structures superimposes these same five helices onto 1R69.

If the repressor-like half is indeed the structural core of globin, the remaining half would be expected to contribute to the heme-binding function. The 'genes-in-pieces' arguments that propose that secondary structure is encoded at the exon-intron level appear to lend some support to this division of the globin fold [24–26]. As the repressor gene is from a prokaryote, it has no exonic structure. However, globin chains come in 3 exons with the middle one splicing at residues B12–B13 and G6–G7 (Fig. 3) [24,25]. Thus it appears that, to within a couple of residues, the middle exon corresponds almost exactly to a replacement of helix 3 of repressor with helix E and all the other heme-binding structural elements that are present in globin but not in repressor, In other words, in going from repressor to hemoglobin, the recognition helix of repressor can be viewed as being replaced by a single exon that encodes the heme-binding functionality of globin. Incidentally, it has already been shown spectroscopically that the proteolytic fragment corresponding largely to this middle exon can independently bind heme when expressed by itself [27]. Given this evidence for necessary and sufficient functionality of the non-repressor-like half of globin, the argument for the 1R69 fold being the structural core of the globin fold is strengthened. Further support is offered by recent NMR evidence that upon the removal of heme, myoglobin retains the A,

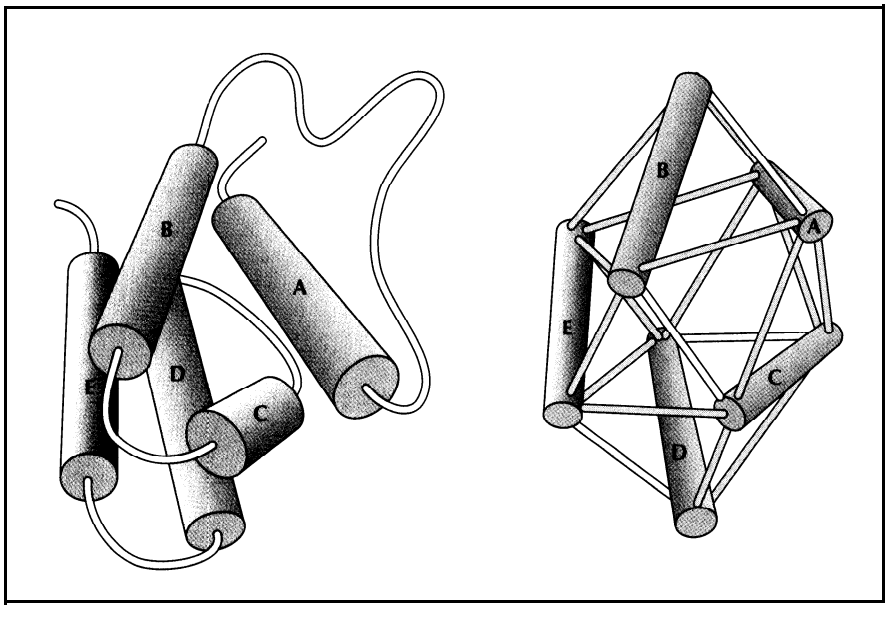

Fig. 4. The pair of cartoons, adapted from Murzin and Finkelstein [21], show how a five-helical protein (left) can be assessed in terms of mapping onto the ribs of a classical Greek sextadecahedron (right).

B, E, G and H helices and their common hydrophobic core while the C, D and F helices are disrupted [28]. A final proof that repressor is the structural core of globin would lie in deleting appropriate portions of the globin sequence as suggested by our superimposition, and obtaining a stable fold. Similar studies, using 1R69 as a prototype, can also attempt to determine the structural details that define such all-helical cores.

In summary, we have a versatile tool that an X-ray crystallographer or an NMR spectroscopist with a newly determined structure can use to ascertain rapidly the existence of structural cousins in the exponentially expanding database. Additionally, this tool can be used to study classes of structural patterns and to catalogue structural cores in a quantitative manner. The ability to do so, at least in the case of the all-helical core, allows the consideration of large-scale modifications to wild-type proteins beyond the present successes in experimental 'tinkering' with protein design.

Materials and methods

The method used here to align two protein three-dimensional structures optimally starts with an arbitrary equivalence of the residues of the two proteins (referred to as proteins A and B). This provides a list of equivalent residue pairs, denoted as i(A),j(B) to indicate that residue i of protein B is equivalenced to residue j of protein **B. For** example, if the initial alignment matches the start of protein A to that of B, then the equivalent pairs are (1,1), (2,2), (3,3)... (n,n), where n is the number of residues in the shorter protein. This list of equivalent residues is used to make a list of equivalent Cα atoms that are then superimposed in three-dimensions using standard, well-studied methods [29–32].

Once the two 3-dimensional structures are superimposed, it is possible to calculate the structural alignment (SA) matrix that is at the heart of our method. The element $SA_{i(A)j(B)}$ of this matrix measures the structural similarity of the i-th residue of protein A with the j-th residue of protein B. Here $SA_{i(A)j(B)}$ is a function of the Euclidean distance between $C\alpha_{i(A)}$ and $C\alpha_{j(B)}$ after coordinate superimposition. These distances give rise to a cross-distance matrix $d_{i(A)j(B)}$, which is different from the conventional distance matrix in that positions i and j are in different proteins. A new alignment is then determined by searching the structural alignment matrix for the alignment that has the best score. This follows exactly the methods used to align a pair of sequences, where a sequence alignment matrix is used; dynamic programming rapidly ($O(n^2)$ operations) finds the optimum for the given alignment matrix and a deletion/insertion penalty value [5]. The new alignment leads, in turn, to a new set of residue equivalencies and the Cα coordinates of these equivalent residues are then used to re-superimpose the two proteins in three-dimensions. This gives a new structural alignment matrix and the whole process is repeated until the structural alignment remains unchanged.

In normal sequence alignment, dynamic programming gives a globally optimum alignment without any iteration — the initial residue equivalences are unimportant. Structural alignment using dynamic programming must be iterated, and the alignment obtained may depend on the initial residue equivalences. We use five different initial residue equivalences to start the iterative procedure. Three are particularly simple and involve aligning the chain beginnings, the chain ends and the chain mid-points, respectively, without allowing any gaps. The two other initial alignments are obtained by maximizing (a) sequence identity and (b) similarity of inter-Ca torsion angles. For each initial set of residue equivalences, we superimpose coordinates, calculate the structural alignment matrix, and then use the standard Needleman-Wunsch dynamic programming method to find the best structural alignment for the current SA matrix. This finds the alignment with the highest score (given by $\Sigma SA_{i(A)j(B)}$ − Penalty x Number of gaps, where the summation is over ad residue pairs that are equivalenced) [5]. The same value of Penalty = 10 is used for all structural comparisons; this corresponds to half the best score for a single aligned pair of residues. After repeating the scheme for each of the five initial set of equivalent residues, the optimal alignment is taken as that with the highest score. Extensive studies have shown that no one of the five schemes for initial residue equivalences works better than another; no information on features in either protein is needed to obtain the optimal structural alignment.

Structural alignment equivalences a set of N residues and it is easy to calculate the RMS deviation of the equivalent Cα atoms, denoted here as cRMS. The values of N and cRMS are not independent; if fewer residues are matched, the cRMS value will be less for the same general quality of superimposition. To allow for this, we define a structural alignment score (SAS) as 100 x cRMS / N. We also define I % as the percentage amino acid identity for the N matched residues.

This method is so simple that one must ask why it succeeds and how it can be original in a field of intense previous study [6–11]. The present work succeeds for two reasons; first, the form of the function that relates the inter-Cα distance $d_{i(A)j(B)}$ to the matrix element $SA_{i(A)j(B)}$; and second, the large number of different initial residue equivalencies tried. The function used to calculate the structural alignment matrix is simple, with $SA_{i(A)j(B)} = 20 / (1 + 5(d_{i(A)j(B)})^2)$, but has the important properties of being positive, decreasing monotonically with increasing $d_{i(A)j(B)}$, and changing most rapidly for $d_{i(A)j(B)} = \sqrt{5} \simeq 2$ A. This was the first function tried; subsequent tests showed that other functions with similar properties work as well.

A number of other methods use an iterative dynamic programming approach (see [33] for a recent review). The method closest to ours, and indeed the method on which our work is based, is the program ALIGN written by Cohen and is used in the study of antibody structures [34]. His method worked well for proteins that were closely similar, but lacks the sensitivity and range of convergence we have. We believe this is because we use a different definition of the SA matrix and start from five different initial residue equivalences. Several methods use dynamic programming on a linear array of residue properties, which makes the procedure exactly equivalent to sequence alignment. One of the first studies aligned the (ϕ,ψ) torsion angles [35] and this has been generalized to many more residue properties [36]. These methods will not be able to detect similarity of global folding that is not reflected at the local residue level. The method of Taylor and Orengo [s] also uses dynamic programming but avoids superimposing coordinates, Instead the similarity of a pair of residues is calculated by comparing the set of distances of each residue with its neighbors; two residues are closely similar if the distances to their respective neighbors are the same. In contrast our method superimposes the two sets of coordinates and judges similarity by the direct distance between the pair of residues, which is much faster. In an improved method, Taylor and Orengo [38] sped their search by first doing an initial alignment using residue properties, such as solvent accessible area and main-chain torsion angles. Their running times appear comparable to ours (10 weeks to compare all known structures compared to approximately 10 days using our technique on a more modern computer). Relying on initial matching of residue properties will not work for comparisons of protein that are very different and will make the method less sensitive. The speed of

our method arises from its extreme simplicity. Use of local features to provide a better initial equivalencing of residues will speed our method even further.

Acknowledgment& We acknowledge G. Cohen for invaluable discussion of the method and P. David for helpful suggestions. S. S. is supported by a Damon Runyon-Walter Winchell Cancer Research Fellowship and D. V. L. by a NSF pre-doctoral fellowship. This work was supported by the Office of Energy Research, Office of Basis Energy Science, Divisions of Materials Sciences and also Energy Biosciences of the US Department of Energy.

References

1. **Bernstein** FC, ET *AL.*: The Protein data bank: a computer-based archival file for macromolecular structures. *J Mol Biol* **1977, 112**:535.
2. HILL CP, **Anderson DH, Wesson** L, DEGRADO **WF, Eisenberg** D: Crystal structure of a 1: implications for protein design. *Science* **1990, 249**:543-546.
3. **Chothia** C: Proteins. One thousand families for the molecular biologist [news]. *Nature* **1992, 357**:543-544.
4. DOOLITTLE RF: *Of Urfs and Orfs: a primer on how to analyze derived amino acid sequences*. Mill Valley, CA: University Science Books; 1986.
5. **Needleman** SB, **Wunsch** CD: A general method applicable to the search for similarities in the amino acid sequence of two proteins. *J Mol Biol* **1970, 48**:443.
6. **REmington** SJ, **Matthews** BW: A systematic approach to the comparison of protein structures. *J Mol Biol 1980*, **140**:77.
7. ROSSMANN MG, MORAS D, OLSEN KW: Chemical and biological evolution of a nucleotide-binding protein. *Nature* 1974, **250: 195.**
8. TAYLOR WR, ORENGO CA: Protein structure alignment. *J Mol Biol* **1989, 208(1)**:1–22.
9. SALI A, **Blundell** TL: 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* **1990, 212(2)**:403–28.
10. **Vriend** *G*, **Sander** C: Detection of common three-dimensional substructures in proteins. *Proteins* **1991,** 1 1(**1**):52-58.
11. ZUKER M, **Somorjai** RL: The alignment of protein structures in three dimensions. *Bull Math Biol* **1989, 51(1)**:55-78.
12. **Chothia** C, LESK AM: Evolution of proteins formed by β-sheets. *J Mol Biol* **1982, 160**:309.
13. **Weaver** LH, **ET** *AL.*: Comparison of goose-type, chicken-type, and phage-type lysozymes illustrates the changes that occur in both amino acid sequence and three-dimensional structure during evolution. *J Mol Evol* **1985, 21**:97.
14. KENDREW JC, BODO G, DINTZIS **HM, Parrish** RG, **Wyckoff** H: A three-dimensional model of the myoglobin molecule obtained by X-ray analysis. *Nature* **1958, 181**:662.
15. LESK AM, CHOTHIA C: How different amino acid sequences determine similar protein structures: the structure and evolutionary dynamics of the globins. *J Mol Biol* **1980, 136**:225.
16. BASHFORD D, **Chothia** C, LESK AM: Determinants of a protein fold: Unique features of the globin amino acid sequence. *J Mol Biol* **1987, 196199.**
17. DICKERSON RE, **Geis I:** *Hemoglobin: structure, function, evo-Zution, and pathology*. Menlo Park, CA: Benjamin/Cummings; 1983.
18. CHOTHIA **C,** FINKELSTEIN AV: The classification and origins of protein folding patterns. *Annu Rev Biocbem* **1990, 59**:1007–1039.
19. **Harrison** SC, **Aggarwal** AK: DNA recognition by proteins with the helix-turn-helix motif. *Annu Rev Biochem* **1990, 59**:933–969.
20. MONDRAGON **A,** SUBBIAH **S,** ALMO **SC,** DROTTAR **M, Harrison** SC: Structure of the amino-terminal domain of phage 434 repressor at 2.0 Å resolution. *J Mol Biol* **1989, 205(1)**: 189–200.
21. **Efimov,** AV: A novel super-secondary structure of proteins and the relation between structure and amino acid sequence. *FEBS Letters* **1984, 166**:33.
22. MURZIN AG, **Finkeistein** AV: Polyhedra describing the packing of helixes in a protein globule. *Biofizika* **1983, 28**:905.
23. **Murzin** AG, FINKELSTEIN AV: General architecture of the α-helical globule. *J Mol Biol* **1988, 204(3)**:749–769.
24. *Go* M: Correlation of DNA exonic regions with protein structural units in haemoglobin. *Nature* **1981, 291**:90.
25. **Eaton** WA: The relationship between coding sequences and function in haemoglobin. *Nature* **1980, 284**:183.
26. **Gilbert** W: Genes In Pieces Revisited. *Science* **1985, 228**:823.
27. CRAIK CS, **Buchman** SR, **Beychok** S: Characterization of globin domains: Heme binding to the central exon product. *Proc Natl Acad Sci USA* **1980,** 77:1384.
28. COCCO **MJ, Lecomte** JT: Characterization of hydrophobic cores in apomyoglobin: a proton NMR spectroscopy study. *Biochemistry* **1990, 29(50)**:11067–11072.
29. **Diamond** R: On the comparison of conformations using linear and quadratic transformations. *Acta Cys* 1976, **A32**:1.
30. MCLACHIAN AD: A mathematical procedure for superimposing atomic coordinates of proteins. *Acta Cys* **1972, A28**:656.
31. KABSCH W: A solution for the best rotation to relate 2 sets of vectors. *Acta Cys* 1976, **A32**:922.
32. RAO ST, ROSSMANN MG: Comparison of super-secondary structures in proteins. *J Mol Biol* **1973, 76**:241.
33. **Johnson** MS: Comparison of protein structure. *Curr Opin Strut Biol* **1991, 1**:334.
34. **Satow Y, Cohen** GH, PADLAN EA, **Davies** DR: Phosphocholine binding immunoglobulin Fab McPC603. *J Mol Biol* **1986, 190**:593–604.
35. LEVINE **M, Stuart** D, **Williams** J: A method for the systematic comparison of the three-dimensional structures of proteins and some results. *Acta Cyst* **1984, A40**:600.
36. **Johnson** MS, SALI A, **Blundell** TL: Phylogenetic relationships from three-dimensional protein structures. *Methods Enzymol* **1990, 183**:670.
37. ORENGO **CA, Taylor** WR: A rapid method of protein structure alignment. *J Theor Biol* **1990, 147**:517.

Received:25 **November** 1993; revised: 27 **January** 1993.
Accepted: 27 **January** 1993.