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 structures against each other could reveal 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

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 pro-

(HTH) motif, but the remaining two helices are replaced by β-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

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.

[redacted]

B and C, helix C, the loop between C and D, helix D, the loop between helix D and E and the amino-terminal of helix G.

pocket of the globin. The central portion of helix E, in particular residues E7, E10, E11 and E14, makes the

DNA-binding 'recognition' helix, helix 3. One could ar-

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,

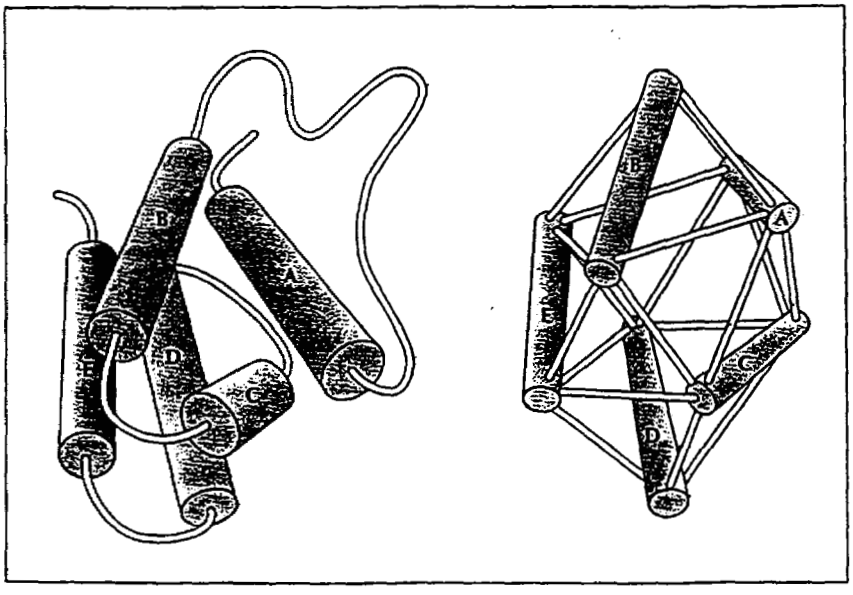

Fig. 4. The pair of cartoons, adapted from Murzin and Finkelstein [21], show how a five-helical protein (left) can be assessed in terms of mapping onto the ribs of a classical Greek sextadecahedron (right).

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 ex-

and (b) similarity of inter-Cα 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 × Number of gaps, where the summation is over all 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

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.

Acknowledgments: 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

16. BASHFORD D, CHOTHIA C, LESK AM: Determinants of a protein fold: Unique features of the globin amino acid sequence. *J Mol Biol* 1987, 196:199.
17. DICKERSON RE, GEIS I: *Hemoglobin: structure, function, evolution, and pathology.* Menlo Park, CA: Benjamin/Cummings; 1983.
18. CHOTHIA C, FINKELSTEIN AV: The classification and origins of protein folding patterns. *Annu Rev Biochem* 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. [illegible] AV: A novel super-secondary structure of pro-