Volume 293, number 1,2, 149–152 FEBS 10360 November 1991

Relative-residue surface-accessibility patterns reveal myoglobin and catalase similarity

V.B. Cockcroft and D.J. Osguthorpe

Molecular Graphics Unit, Bath University, Bath BA2 7AY, UK

Received 27 August 1991

A novel sliding-window search method using relative-residue surface-accessibility patterns identified extensive, but unsuspected, structural similarity over a 3-helix region in the C-terminus of the evolutionarily unrelated proteins sperm-whale myoglobin and beef liver catalase. This clear example of structural similarity between non-homologous proteins highlights the importance of relative-residue surface-accessibility patterns in understanding the local folded structure in proteins.

Relative-residue surface-accessibility pattern; Non-homologous similarity; Myoglobin; Catalase

1. INTRODUCTION

We reasoned that patterns in the relative-residue surface-accessibilities of proteins may be used to search through the Brookhaven databank of protein structures for similarities in proteins considered evolutionarily unrelated (i.e. non-homologous) by classical sequence analysis methods. This was based on the early observation that regular secondary-structure elements (e.g. α-helices, β-strands) in proteins are often rationalizable in physical-chemical terms by the periodic sequence patterns in the hydrophobicities of amino acid side-chains [1–4]. Here we present the findings of a model study in which a sliding-window search method was implemented and that highlighted previously unsuspected structural similarity between sperm-whale myoglobin and beef liver catalase.

2. EXPERIMENTAL

2.1. *Establishment of Database*

Relative-residue surface-accessibilities to solvent (RCSCQ), calculated using the algorithm of Richmond and Richard, were obtained from the BIPED database (Daresbury Laboratories, UK) [5]. The accessibility database consisted of 103 protein structures with resolutions <2.6 Å from the Brookhaven databank, excluding myoglobin structures. Each accessibility file contained the residue number identifier (UNIQID), amino acid type (IUPAC one-letter code), assigned secondary-structure (STRB), and the accessibility (RCSCQ) for each residue of the protein (UNIQID, STRB and RCSCQ refer to the attributed name in BIPED).

2.2. *Database Scanning*

The probe comprised the RCSCQ values of sperm-whale myoglobin. The window-search method used the comparison function:

$$\text{score} = \sum_{i}^{n_w} (\text{probe\_RCSCQ}_{p+i} - \text{database\_RCSCQ}_{q+i})^2$$

where n_w is the window length in residues (20); p is the offset in the residue sequence to the start residue of the window for the probe; and q is the offset in the residue sequence to the start residue of the window for the current accessibility database protein. All probe-database window matches were sorted in ascending order of score.

3. RESULTS AND DISCUSSION

A myoblobin probe was constructed using the relative-residue surface-accessibility values that were calculated from the X-ray structure of sperm-whale myoglobin (i.e. 1MBD of the Brookhaven database [6]) and was used to scan a relative-residue surface-accessibility database (derived also from known protein structures but with myoglobins excluded). The use of relative rather than absolute values provides a normalization of the differences in size of the amino acid side-chains, with the final indices being scaled between 0 and 100%. Using a 20-residue window, out of 1.6×10^6 matches of the probe with the database, the top-match (i.e. lowest score) was of a C-terminal segment in sperm-whale myoglobin and a C-terminal segment of beef liver catalase (8CAT) [7] (see Table I). The Cα RMS-fit is relatively high compared to other matches, but this is owing to a structural difference in the 5 residues at the N-terminus of the matched segments; over the remaining 15 residues the Cα RMS-fit was 0.4 Å. The second lowest match was with *E. coli* arabinose binding protein (1ABP) [8] where the similarity was with α-helical segments which had their C-termini packed more tightly onto their protein cores than their N-termini. The first match to a member of the globin family was with eryth-

Correspondence address: D.J. Osguthorpe, Molecular Graphics Unit, Bath University, Bath BA2 7AY, UK.

rocruorin (1ECD) [9], a haemoglobin of midge-larvae. This involved topologically identical segments (i.e. segments that would be classified as equivalent by a sequence alignment), which form an unusual surface loop in close contact with the bound haeme-centre. The fourth and fifth matches were human α-haemoglobin (4HHB) [10] and erythrocruorin, respectively. In these cases the α-helical segments are structurally similar (according to their RMS-fit values), but would be identified as non-homologous segments by conventional sequence alignment methods [11]. The above demonstrates that relative-residue surface-accessibility patterns can be used to search for structural similarity, irrespective of homology.

Notably, visual analysis of the top-matched segments in the structures of sperm-whale myoglobin and beef liver catalase revealed extensive similarity not only in the superimposed matched segments, but in the surrounding protein structure (see Fig. 1). This involved a 3-helix region present in a similar context, with the last of the 3 consecutive helices occurring at the C-terminus in both proteins and the first helix packing onto the third helix (i.e. the matched segment) with a ≈45° cross-over angle. Here the helices are labelled X, Y and Z in amino-acid sequence order (Helices X, Y and Z correspond to helices F, G and H in myoglobin and helices 11A, 12A and 13A in catalase). For the 34 positions that were assigned as being topologically equivalent (see Fig. 2) in the 2 proteins a Cα RMS-fit of 2.1 Å was obtained. The polypeptide leading into helix X also displays similarity, being an α-helical region that is set by a sharp turn to almost a right-angle to helix X. The major difference in the 2 regions is the length of the loop linking the overlap of helices Y and Z. This loop is longer in sperm-whale myoglobin, mainly because of a 3-turn extension in both helices. It is this structural detail which leads to the large RMS-fit value for the matched segments (see discussion above).

A further level of similarity of the 3-helix region involves a quartet of core residues that interdigitate helix X and Z, namely the residue positions Leu^{89}, Ala^{90} (helix X), and Phe^{138}, Ile^{142} (helix Z) of sperm-whale myoglobin and the corresponding positions Ile^{462}, Ala^{463}, Tyr^{488} and Ile^{492} of beef liver catalase (see Fig. 1). It is notable that the amino acids at the corresponding positions are identical or very similar. In addition, their side-chain atoms are similarly placed, although the phenylalanine ring at position 138 in sperm-whale myoglobin is not so well aligned structurally with the tyrosine ring at position 488 in beef liver catalase (χ_1 side-chain torsion-angles are 166° and −158°, respectively). Interestingly, superpositioning and structural comparison of these interdigitating positions gave a better Cα RMS-fit for sperm-whale myoglobin and beef liver catalase, 0.6 Å than the homologous globins, 0.8–1.2 Å (Table II). This greater structural similarity in the non-homologous case than the homologous cases is explained by the greater amino acid similarity of the interdigitating positions (see Table II).

No significant sequence similarity over the 3-helix region could be detected by inspection of a multiple alignment, which included related sequences for both proteins (Fig. 2). This supports the existing view that the catalases and globins evolved independently. If diver-

Table I

The top 5 lowest surface accessibility matches

	BRK Codes and residue positions	Aligned sequences	SCORE	RMS-fit (Å)
1	1MBD:130-149:	AMNKALELFRKDIAAKYKEL	2560	4.0
	8CAT:462-467:	NFSDVHPEYGSRIQALLDKY		
2	1MBD:57-76:	ASEDLKKHGVTVLTALGAIL	2717	1.8
	1ABP:42-61:	DGEKTLNAIDSLAASGAKGF		
3	1MBD:28-47:	ILIRLFKSHPETLEKFDRFK	2755	0.8
	1ECD:23-42:	ILYAVFKADPSIMAKFTQFA		
4	1MBD:55-74:	MKASEDLKKHGVTVLTALGA	2831	3.0
	4HHB:70-89:	VAHVDDMPNALSALSDLHAH		
5	1MBD:128-147:	QGAMNKALELFRKDIAAKYK	3189	1.1
	1ECD:94-113:	HDQLNNFRAGFVSYMKAHTD		

Sequence listing of the top 5 lowest scores of relative-residue surface-accessibility matches. For each match, the Brookhaven codes of the protein, the residue numbering of the matched segments, the aligned amino-acid sequences (asterisk: identity; dot: conservation), and the RMS-fit of a superposition of the Cα atoms are listed. NB. As a check, a sample of pseudorandom Cα RMS-fits generated for 20-residue length segments gave a mean RMS-fit of 7.4 Å and standard deviation of 1.5 Å (data not shown).

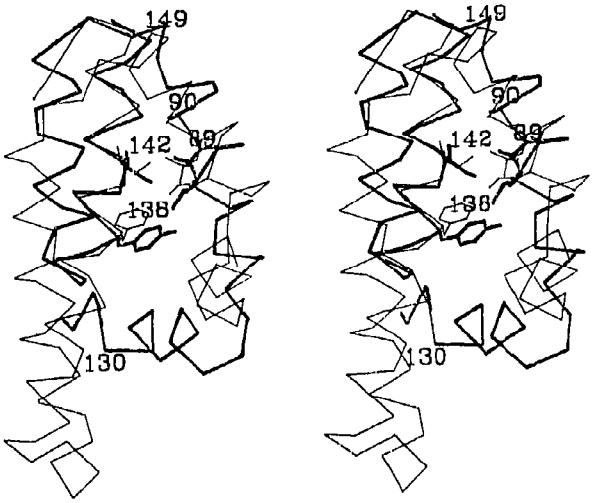

Fig. 1. Stereoview of superposition of the 3-helix packing region of sperm-whale myoglobin and beef liver catalase. Myoglobin and catalase are coloured blue and yellow, respectively, except over the matched segments in which case they are coloured green and red, respectively. The side-chains of the core quartet of residues (see text) are shown and the residue numbering (89, 90, 138 and 142) of these residues refers to sperm-whale myoglobin.

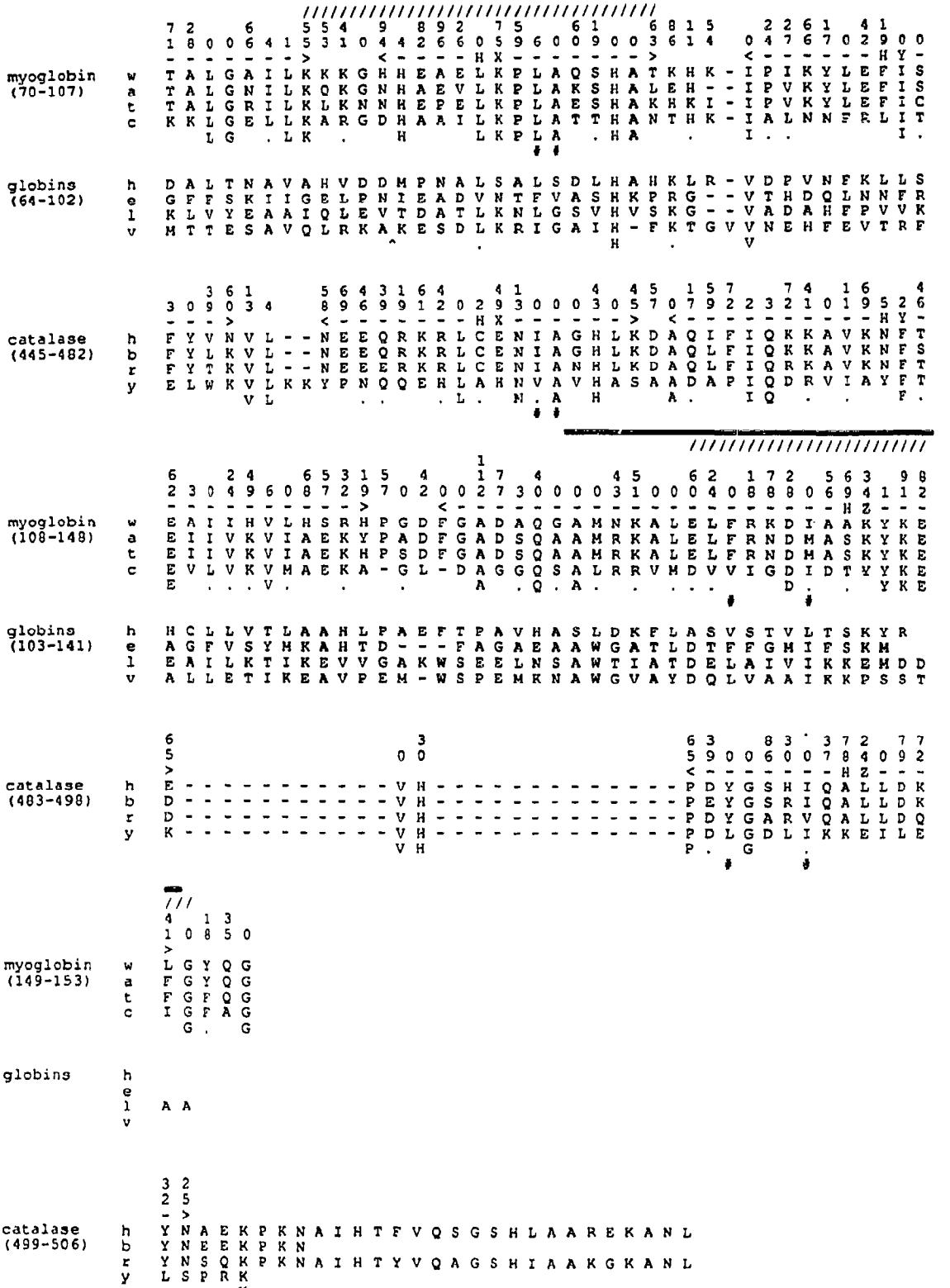

Fig. 2. Alignment of myoglobin and catalase amino acid sequences over the 3-helix packing region. Myoglobin sequences are: (w) sperm-whale, (a) alligator, (t) map turtle and (c) carp. Globin sequences are: (h) human α-haemoglobin, (e) midge-larvae erythrocruorin, (l) leghaemoglobin and (v) vitreoscilla haemoglobin [12]. Catalase sequences are: (h) human, (b) beef liver, (r) rat and (y) yeast peroxisomal. The consensus sequences show invariant and conserved (dot; equivalence groupings: DNEQ, RKH, YFW, MILV, STPAGC) residue positions. Assignments of helix structure are from the Brookhaven structures 1MBD and 8CAT. Relative residue surface-accessibility values from these structures are given above their corresponding sequence sets. Circumflex indicates the position of insertion sequence in vitreoscilla heamoglobin not included in the alignment. Slashes above the alignment indicate assigned topologically equivalent residue positions. Solid bar shows the position of the matched segments of the database search. Hashes indicate the packing-core quartet residue positions.

Table II

Conservation of interdigitating core quartet residues

	Superpositioned structures	Interdigitating positions				No. IDs	RMS fit (Å)
		1	2	3	4		
1	8CAT-1MBD	I-L	A-A	Y-F	I-I	2	0.6
2	8CAT-1ECD	L-F	A-V	Y-F	I-I	1	0.6
3	1MBD-1ECD	L-F	A-V	F-F	I-I	2	0.8
4	8CAT-2LH6	L-L	A-G	Y-L	I-I	2	0.8
5	1ECD-2LH6	F-L	V-G	F-L	I-I	1	0.8
6	1MBD-4HHB	L-L	A-S	F-V	I-L	1	0.9
7	2LH6-4HHB	L-L	G-S	L-V	I-L	1	1.0
8	1MBD-2LH6	L-L	A-G	F-L	I-I	2	1.0
9	8CAT-4HHB	L-L	A-S	Y-V	I-L	1	1.1
10	1ECD-4HHB	F-L	V-S	F-V	I-L	0	1.2

IDs = number of identical amino acids at core quartet positions. The superimposed Brookhaven database structures are: 8CAT = beef liver catalase; 1MBD = sperm-whale myoglobin; 1ECD = erythrocruorin; 2LH6 = leghaemoglobin; 4HHB = human haemoglobin.

gent evolution were the case, the similarity in their structures must have been maintained over a long period of time, as both globin [12] and catalase [13] are present in bacteria. In the absence of an evolutionary link, the observed structural similarity of the 3-helix region as revealed by the sliding-window method presented here is clearly surprising; many more examples may come to light as our analysis is applied to other groupings of proteins. We propose the term non-homologous similarity to describe such situations.

It is suggested that the relative-residue surface-accessibilities reflect the tendencies of residue positions to partition between the aqueous solvent and the hydrophobic core in proteins, and that the above, by way of a non-homologous example, indicates what a crucial determinant of the local folded structure such tendencies through the polypeptide can be. Consideration of such information may be of use in developing sequence-to-structure algorithms [14].

Acknowledgement: We thank Professor G. Lunt for useful suggestions and comments on the manuscript.

REFERENCES

[1] Perutz, M.F., Kendrew, J.C. and Watson, H.C. (1965) J. Mol. Biol. 13, 669–678.
[2] Hubbard, T.J.P. and Blundell, T.L. (1987) Protein Eng. 1, 159–171.
[3] Gaboriaud, C., Bissery, V., Benchetrit, T. and Mornon, J.P. (1987) FEBS Lett. 224, 149–155.
[4] Bowie, J.U., Clarke, N.D., Pabo, C.O. and Sauer, R.T. (1990) Proteins 7, 257–264.
[5] Islam, S.A. and Sternberg, M.J.E. (1989) Protein Eng. 2, 431–442.
[6] Phillips, S.E.V. (1980) J. Mol. Biol. 142, 531.
[7] Melik-Adamyan, W.R., Barynin, V.V., Vagin, A.A., Borisov, V.V., Vainshtein, B.K., Eita, I., Marthy, M.R.N. and Rossman, M.G. (1986) J. Mol. Biol. 188, 63.
[8] Newcomer, M.E., Gilliard, G.L. and Quicho, F.A. (1981) J. Biol. Chem. 256, 13213.
[9] Steigeman, W. and Webster, E. (1979) J. Mol. Biol. 127, 309.
[10] Fermi, G., Perutz, M.F., Shaanan, B. and Fourme, R. (1984) J. Mol. Biol. 75, 159.
[11] Bashford, D., Chothia, C. and Lesk, A.M. (1987) J. Mol. Biol. 196, 199–241.
[12] Wakabayashi, S., Matsubara, H. and Webster, D.A. (1986) Nature 322, 481–483.
[13] Triggs-Raine, B.L., Doble, B.W., Mulvey, M.R., Sorby, P.A. and Loewen, P.C. (1988) J. Bacteriol. 170, 4425–4429.
[14] Bowie, J.U., Reidhaar-Olson, J.F., Lim, W.A. and Sauer, R.T. (1990) Science 247, 1306.