Volume 315, number 3, 301–306 FEBS 11991 January 1993

Structural alignment of globins, phycocyanins and colicin A

Liisa Holm and Chris Sander

Protein Design Group, European Molecular Biology Laboratory, D-6900 Heidelberg, Germany

Received 28 October 1992; revised version received 30 November 1992

A database search employing a novel algorithm for protein structure comparison by alignment of distance matrices has revealed a striking resemblance between the tertiary structures of the bacterial toxin colicin A and globins. The globin-like domain in colicin A contains all elements essential for the toxin's lethal ionophoric activity. The structural similarity between colicin A and globins is comparable to that between globins and phycocyanins. This suggests that these three protein families, which have unrelated sequences and different functional contexts, are an example of physical convergence to a stable folding motif, the three-on-three helical sandwich.

Globin fold; Phycocyanin; Colicin A; Structural similarity; Folding motif; Convergent evolution

1. INTRODUCTION

A substantial fraction of presently known protein structures can be described in terms of a relatively small set of folding motifs, such as helical bundles, β-barrels, and the like. The delineation of common structural cores is useful for the analysis of protein evolution, for the elucidation of folding principles, for model building by homology and for modular protein design. We have developed a novel algorithm for protein structure comparison by alignment of distance matrices [1] and performed an all-against-all comparison of more than 150 different proteins. An exemplary result of this search is the discovery that the membrane-insertion domain of the bacterial toxin colicin A closely resembles the globin fold: six α-helices (and one turn of 3_{10}-helix) are arranged similarly in space, the packing between the helices is similar as well as the sequential order in which the chain threads through the helices [2]. The colicin A domain thus joins the structural class that so far has consisted of the globin and phycocyanin families. In this paper, we present and discuss the structural alignment of these three protein families that contain the globin fold.

2. MATERIALS AND METHODS

2.1. *Structure alignment*

The method for structure alignment [1] is based on the exploitation of distance matrices, a rotation and translation invariant representation of three-dimensional structure. The alignment algorithm optimizes a structural similarity score that measures the agreement of all equivalent intramolecular distances in two proteins. More precisely, consider two proteins, labeled A and B. An alignment consisting of L equivalenced residue pairs is evaluated by an additive similarity score of the form

$$S = \sum_{i=1}^{L} \sum_{j=1}^{L} \phi\,(i,j), \tag{1}$$

where i and j label the equivalenced pairs (i^A,i^B) and (j^A,j^B), and ϕ is a measure of the similarity of the C^α–C^α distances d_{ij}^A in protein A and d_{ij}^A in protein B. Residues with no equivalent in the other protein do not contribute to the score. ϕ is defined so that smaller distance deviations correspond to higher similarity. Tolerance to spatially extended geometrical distortions is achieved by using relative rather than absolute distance deviations, preventing dominance of long intramolecular distances:

$$\phi\,(i,j) = \begin{cases} (\theta - \dfrac{|\,d_{ij}^A - d_{ij}^B\,|}{d_{ij}^*})\; w(d_{ij}^*),\ i \neq j \\ \theta,\ i = j \end{cases} \tag{2}$$

where d_{ij}^* is the arithmetic average of d_{ij}^A and d_{ij}^B, θ is a constant that determines the threshold of similarity (set to $\theta = 0.20$, i.e. 20% deviation), and the damping factor $w(d)=\exp(-d^2/\alpha^2)$ limits the range of the score to the radius of a typical domain (α = 20 Å).

The optimal structural alignment is defined as the set of residue equivalences that maximizes the similarity score. Because of the complex pairwise dependencies, a Monte Carlo procedure is used for optimization. The algorithm is not guaranteed to reach the global optimum, but in practice the method identifies structural similarities in excellent agreement with intuitive notions of structural equivalence [1]. The method allows sequence gaps of any length and free topological connectivity of aligned segments, including matches in reversed chain direction.

2.2. *Database scan*

Atomic coordinates except those for phycocyanins (R. Huber and T. Schirmer, pers. commun.) were retrieved from the Protein Data Bank (PDB) [3]. A representative set of 154 protein chains was selected from the Protein Data Bank such that no pair had higher than 30% sequence identity. A database containing all-against-all structure comparisons within this set is available [4]. Structural similarity scores were measured in units of standard deviations (σ) above average for the comparison of a given structure against all other structures in the database. Starting from the complete pool of similarity scores, the

Correspondence address: L. Holm, Protein Design Group, European Molecular Biology Laboratory, D-6900 Heidelberg, Germany.

calculation of average score and standard deviation was repeated excluding outliers with scores higher than 3 σ until no outliers remained in the modified pool.

2.3. *Contact maps*

Structural similarity can be evaluated using a more physical measure than merely C^{α}–C^{α} distances, namely two-dimensional plots of residue–residue contacts [5]. Interaction energies were calculated as the sum of an interatomic linear-square well potential (equal to 1.0 from 0.0 Å to 3.6 Å, then decreasing linearly to zero at 6.4 Å), where the sum was taken over all atoms belonging to a given residue pair. The cutoff energy for counting a residue–residue contact was 0.01.

3. RESULTS

Ribbon diagrams (Fig. 1) highlight the similar overall fold of the common core of globins, phycocyanins and colicin A in spite of small shifts in the relative orientations of some of the helix pairs. The structural alignment method is able to pick up the similarity because it optimizes the agreement of pairwise relationships rather than of absolute atomic positions. Optimal rigid-body superimposition according to the structural alignment results in a root mean square deviation (rmsd) of C^{α} positions of 3.2 Å for 112 structurally equivalent residues between colicin A and myoglobin, and 3.5 Å for 115 structurally equivalent residues between colicin A and the α-chain of phycocyanin (Figs. 2 and 3). For a pairwise comparison of globins and phycocyanins, our fully automated alignment method generates results essentially identical to the manual alignment of Pastore and Lesk [6].

In a representative database of more than 150 protein chains, globins, phycocyanins and colicin A form one distinct structural class comprising three families. The mutual distance matrix similarity scores (Fig. 4) between the globin, phycocyanin and colicin A families are 5–8 standard deviations (σ) above background (more than 12 σ within each family). All other protein structures score less than 3.5 σ, with at most a few helices matching. For example, one of the top matches is cytochrome *P450* (PDB dataset 6CPP) with one α-helical layer matching globin helices B-G-H and a turn of 3_{10}-helix matching the 3_{10}-helix C. Two toxin structures not deposited in the Protein Data Bank are said to have helical domains with similarity to colicin A [7–9]. The seven-helix bundle of δ-endotoxin [8] appears to be unique. Diphtheria toxin [9], however, does have a domain that contains a compact helical sandwich which topologically resembles the globin fold: globin helices B-C-E-G-H can be superimposed with diphtheria toxin helices TH3-4-5-8-9, but the equivalents of globin helices A and F are less clear. Diphtheria toxin domain has two helices, TH6-7, roughly corresponding to helix F in globins. In sequence, diphtheria toxin helix TH1 would correspond to globin helix A, but structurally these two helices are located on opposite sides of the molecule.

The most notable differences in helix packing between the three families that contain the globin fold are associated with the binding of cofactors. For example, to accommodate the haem group, one end of helix E in myoglobin moves about 5 Å outwards compared to the equivalent helix 6 in colicin A (Fig. 3). The agreement of all-atom residue-residue contact patterns (Fig. 5) indicates general similarities in size, orientation and intercalation of side chains in the core in spite of very few actual residue identities. Quantitatively, nearly 50% of inter-helical contacts between structurally equivalent residues are common between each pair of the three families (between 61 and 69 of 140). The similarity of contacts is stronger between distantly related globins,

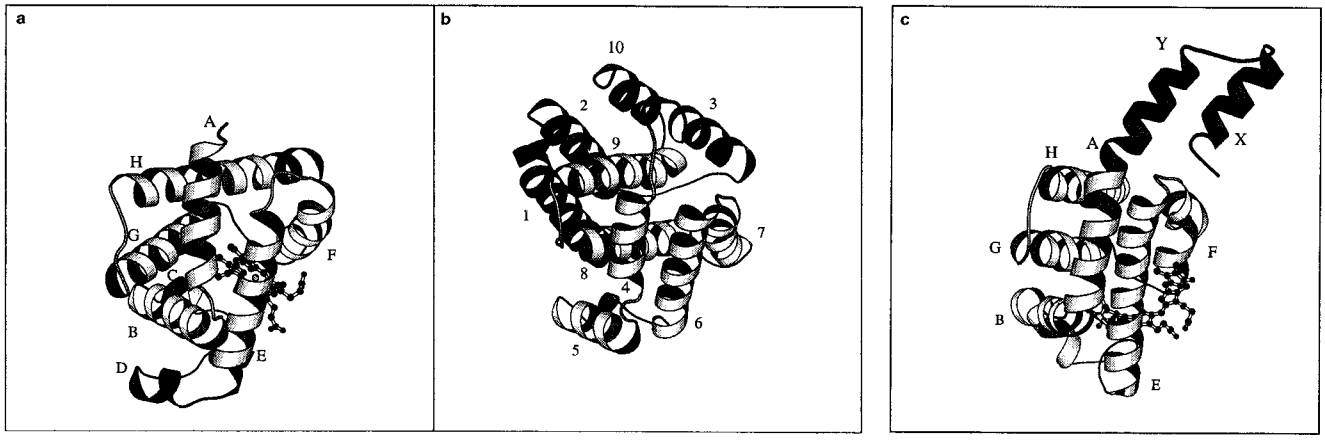

Fig. 1. Ribbon diagrams of (a) myoglobin [23], (b) the carboxyterminal fragment of colicin A [18], and (c) the phycocyanin α-chain [24]. The ribbon for the common core has lighter shading. Helices are labelled according to the original crystallographic reports. The globin-like domain in the crystal structure of colicin A consists of helices 4–9, covered on one side by helices 1–3 and 10. The equivalent α-helices in colicin A and globins, respectively, are helix 4 = A, 5 = B, 6 = E, 7 = F, 8 = G, and 9 = H. One turn of 3_{10}-helix between helices 5 and 6 in colicin A matches the 3_{10}-helix C in globins and phycocyanins. The haem group is included with the globin (a), and the bilin group with the phycocyanin (c) structure. Drawn with MolScript [25].

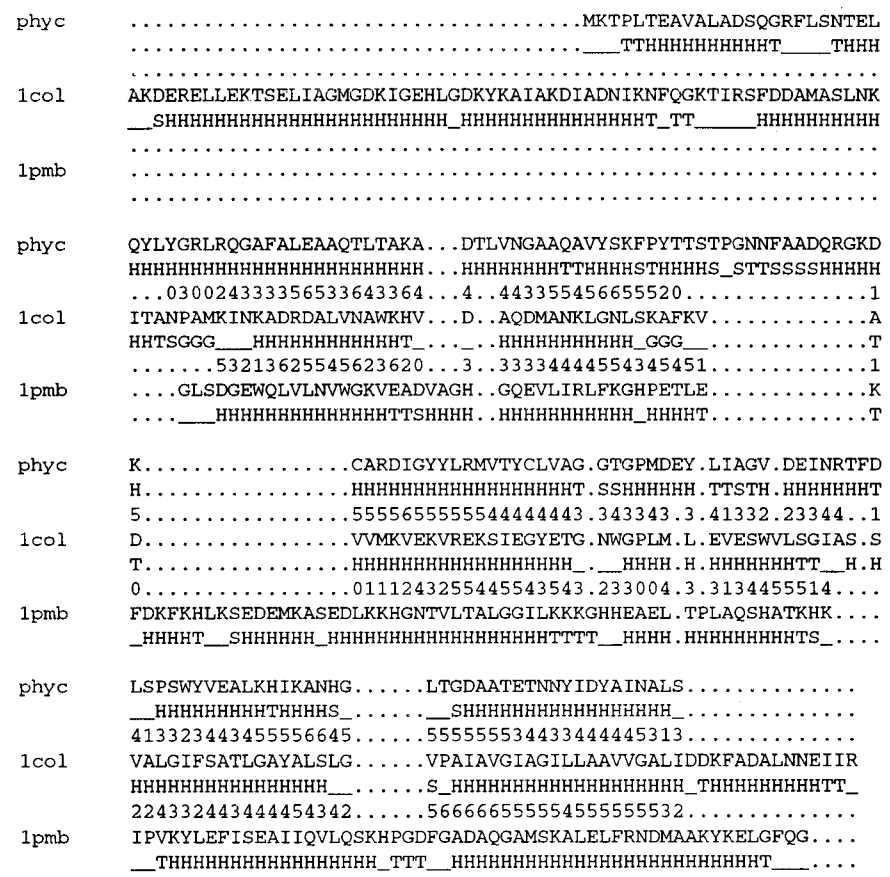

Fig. 2. Pairwise structural alignment of porcine myoglobin (1pmb) and the α-chain of phycocyanin (phyc) with the carboxyterminal fragment of colicin A (1col). The crystal structure of colicin A consists of residues 393–592 of the entire polypeptide chain. The secondary structure [26] is shown below the sequences. The numbers between the sequences indicate the similarity of residue environments in the two aligned proteins (see section 2), on a scale from 0 to 10. Dots mark gaps or trailing ends. If the structures of myoglobin and phycocyanin are aligned, there is a shift of one residue in the alignment of the A helices and a shift of 4–5 residues in the alignment of the F helices (the F helix is broken in the middle in phycocyanin). The alignment of myoglobin and the α-chain of phycocyanin contains 121 structurally equivalent residues and gives an rmsd of 3.7 Å.

e.g. 77% for myoglobin and leghemoglobin. A rare feature shared by globins and phycocyanins, crossed-ridge packing of the B and E helices [6], can also be seen in colicin A.

Residue patterns conserved between the three families are very difficult to find. Among known globin and phycocyanin sequences, only two positions have been reported to have a highly specific pattern of residue conservation: Leu, Phe, or Tyr at position B14 and Pro at position C2 [6]. However, these two residues cannot be essential determinants of the globin fold, because the same backbone conformation in the BC corner is achieved with Gln or Ser at C2 in the hemoglobin of ark clam (PDB dataset 1SDH) and relatives [10] and by completely different sequence combinations in colicins [11].

4. DISCUSSION

In spite of similar folding patterns, globins, phycocyanins and colicins lack significant sequence similarity, bind different cofactors (or none), and are involved in very different biological functions. Phycocyanins are part of the light-harvesting antennae in cyanobacteria and red algae. Globins are oxygen transporters found in bacteria, insects, vertebrates, and plants. Colicins are antibacterial toxins produced by *Escherichia coli* and closely related bacteria. Colicin molecules consist of several domains, each associated with one step of the toxin's lethal activity (receptor binding, translocation across the outer membrane, killing activity) [12]. The major group of colicins, including colicin A, kill sensitive cells by forming pores in the cytoplasmic membrane. The part of colicin A that has been solved crystallographically is a proteolytically cleaved carboxyterminal fragment (residues 389–592), which is water-soluble and has ionophoric properties similar to those of the entire protein.

The globin-like domain of colicin A (helices 4–9) is essentially identical with the membrane-pore-forming unit. Helices 1, 2 and 10 are excluded from membrane insertion by proteolytic data and spectroscopic meas-

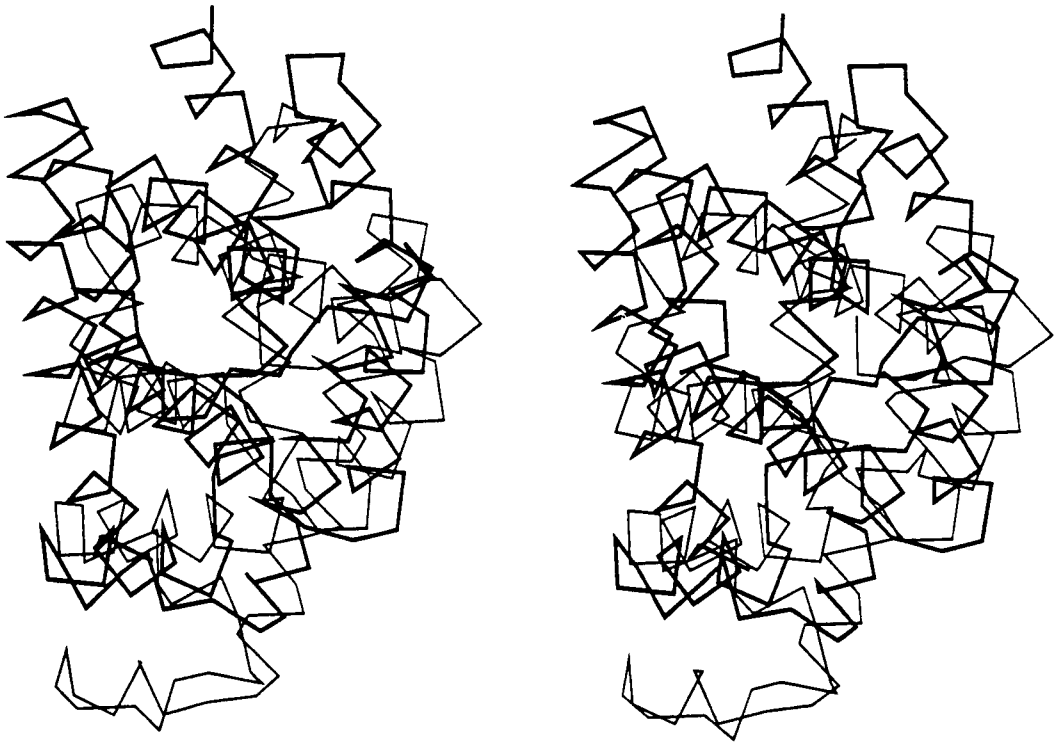

Fig. 3. Optimal rigid-body superimposition [27] of myoglobin (thin line) and the carboxyterminal fragment of colicin A (thick line) corresponding to the structure-derived alignment in Fig. 2. The aminoterminus of the globin fold is at the upper right. Plotted with WHAT IF [28].

urements [13–15]. Deletion mutants of the homologous colicin E1 [16] show that helices 5–9 are sufficient for ionophoric activity. As a monomer would be too short to form an ion channel, a trimeric model has been proposed [17]. The secondary structure of the soluble form, as seen in crystals, appears to be preserved in the membrane-bound conformation [13,18]. In addition, recent spectroscopic results by Lakey et al. [7] show that helices 5–9 remain closely packed on membrane insertion rather than opening up like an 'umbrella' as initially suggested [11,17]. These facts indicate that the globin-like domain is a structural and functional component of

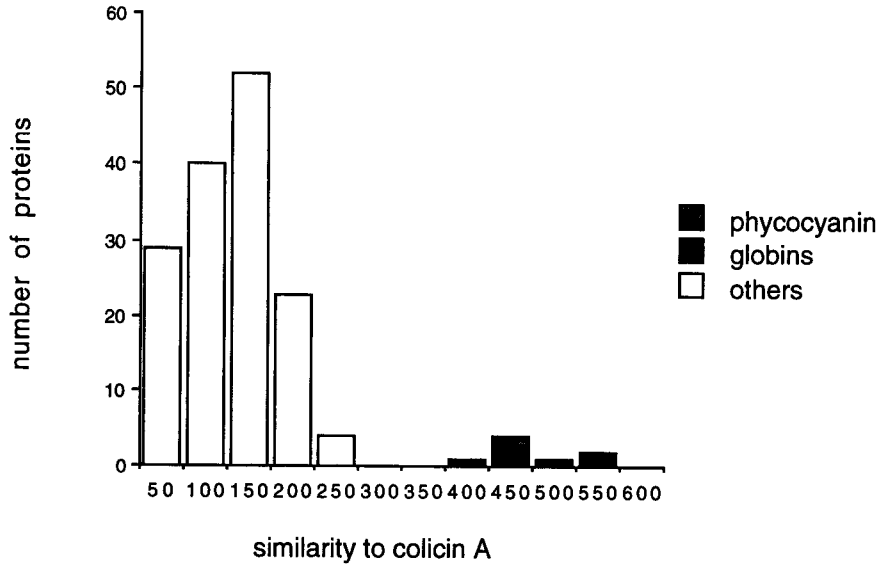

Fig. 4. Histogram of structural similarity scores in a scan of colicin A (1COL-A) against a representative set of 154 protein chains selected from the Protein Data Bank and the α- and β-chains of phycocyanin. The representative set contains six distantly related globins (1ECA, 1HDS-B, 1PMB-A, 1SDH-A, 2LH4, 2LHB). Globins and phycocyanins are clearly separated from the rest of the database, and are about equally similar to colicin A.

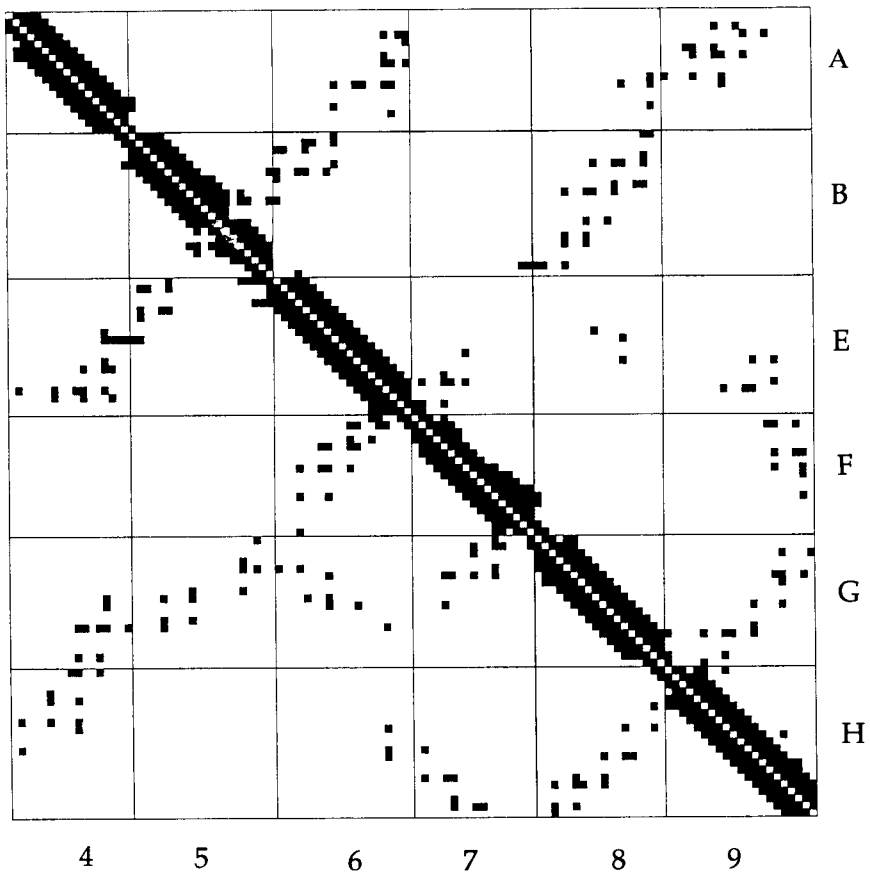

Fig. 5. Aligned contact maps of colicin A (bottom) and myoglobin (top). Residues in contact (see section 2) are indicated by black squares. Only the structurally equivalent parts (see Fig. 2) are shown. The lines separate six segments corresponding to globin helices A, B–C, E, F, G, and H. The absence of direct contacts between the F and G helices in myoglobin is due to the heme group. Plotted with CONAN [5].

colicin A. Bacteria may have acquired this component by adapting a gene fragment of some, as yet unidentified, protein with a globin-like fold.

Colicin A is as similar to either globins or phycocyanins as they are to each other both in terms of overall topography of the fold and in terms of side chain packing. It has been proposed that architectural features shared by globins and phycocyanins are the trace of a distant evolutionary relationship [6,19]. The observation of the same fold in *three* families that inhabit different regions of sequence space and have radically different functional contexts, indicates particular stability of their common structural scaffold, but leaves open the question of divergent vs. convergent evolutionary origin. The three families could have similar folds simply because of physical constraints that govern secondary structure packing and chain topology. These principles are partly understood. Basic folds, including helical globules, can be described by simple geometrical models [20,21]. Considering that antiparallel packing of pieces of secondary structure adjacent in sequence is favored entropically [21], the globin fold represents a rather uncomplicated antiparallel helical meander (or 'Greek key' [22]) motif. We expect that the three-on-three helical sandwich framework of the globin fold will turn out to be as widely spread among proteins as are, e.g. parallel $(\alpha\beta)_8$-barrels, antiparallel meander β-barrels, parallel α/β domains or four-helical bundles [22].

Acknowledgments: We thank the crystallographers for making available their coordinates, P. Kraulis for the program MolScript, M. Scharf for the program CONAN, G. Vriend for the program WHATIF, and F. Pattus for informative discussions.

REFERENCES

[1] Holm, L. and Sander, C., J. Mol. Biol., in press.
[2] Holm, L. and Sander, C. Nature, in press.
[3] Bernstein, F.C., Koetzle, T.F., Williams, G.J.B., Meyer, E.F., Brice, M.D., Rodgers, J.R., Kennard, O., Shimanouchi, T. and Tasumi, M. (1977) J. Mol. Biol. 112, 535–542.
[4] Holm, L., Ouzounis, C., Sander, C., Tuparev, G. and Vriend, G. Protein Sci. 1, 1691–1698.
[5] Scharf, M. (1989) Diploma thesis, University of Heidelberg.
[6] Pastore, A. and Lesk, A.M. (1990) Proteins 8, 133–155.

[7] Lakey, J.H., Duche, D., Gonzales-Manas, J.-M., Baty, D. and Pattus, F., submitted.
[8] Li, J., Carroll, J and Ellar, D.J. (1991) Nature 353, 815–821.
[9] Choe, S., Bennett, M.J., Fujii, G., Curmi, P.M.G., Kantardjieff, K.A., Collier, R.J. and Eisenberg, D. (1992) Nature 357, 216–222.
[10] Sander, C. and Schneider, R. (1991) Proteins 9, 56–68.
[11] Parker, M.W., Pattus, F., Tucker, A.D. and Tsernoglou, D. (1989) Nature 337, 93–96.
[12] Baty, D., Frenette, M., Lloubes, R., Geli, V., Howard, S.P., Pattus, F. and Lazdunski, C. (1988) Mol. Microbiol. 2, 807–811.
[13] Lakey, J.H., Massotte, D., Heitx, F., Dasseux, J.-L., Faucon, J.-F., Parker, M.W. and Pattus, F. (1991) Eur. J. Biochem. 196, 599–607.
[14] Lakey, J.H., Baty, D. and Pattus, F. (1991) J. Mol. Biol. 218, 639–653.
[15] Xu, S., Cramer, W.A., Peterson, A.A., Hermondson, M. and Montecucco, C. (1988) Proc. Natl. Acad. Sci. USA 85, 7531–7535.
[16] Liu, Q.R., Crozel, V., Levinthal, F., Slatin, S., Finkelstein, A. and Levinthal, C. (1986) Proteins 1, 218–229.
[17] Parker, M.W., Postma, J.P.M., Pattus, F., Tucker, A.D. and Tsernoglou, D. (1992) J. Mol. Biol. 224, 639–657.
[18] van der Goot, F.G., Gonzales-Manas, J.M., Lakey, J.H. and Pattus, F. (1991) Nature 354, 408–410.
[19] Schirmer, T., Bode, W., Huber, R., Sidler, W. and Zuber, H. (1985) J. Mol. Biol. 184, 257–277.
[20] Murzin, A.G. and Finkelstein, A.V. (1988) J. Mol. Biol. 204, 749–769.
[21] Chothia, C. and Finkelstein, A.V. (1990) Annu. Rev. Biochem. 59, 1007–1039.
[22] Richardson, J.S. (1981) Adv. Protein Chem. 34, 167–339.
[23] Dodson, G., Hubbard, R.E., Oldfield, T.J., Smerdon, S.J. and Wilkinson, A.J. (1988) Prot. Eng. 2, 233–237.
[24] Schirmer, T., Bode, W. and Huber, R. (1987) J. Mol. Biol. 196, 677–695.
[25] Kraulis, P. J. (1991) Appl. Cryst. 24, 946–950.
[26] Kabsch, W. and Sander, C. (1983) Biopolymers 22, 2577–2637.
[27] Kabsch, W. (1978) Acta Cryst. A34, 827–828.
[28] Vriend, G. (1990) J. Mol. Graph. 8, 52–56.