Methods in Protein Sequence Analysis
Jörnvall/Höög/Gustavsson (eds.)

PATTERNS OF SEQUENCE VARIATION IN FAMILIES OF HOMOLOGOUS PROTEINS

Tom Blundell, Jon Cooper, Dan Donnelly, Huub Driessen, Yvonne Edwards, Frank Eisenmenger, Carlos Frazao, Mark Johnson, Karsten Niefind, Matthew Newman, John Overington, Andrej Sali, Christine Slingsby, V. Nalini and Zhanyang Zhu.

Imperial Cancer Research Fund Unit of Structural Molecular Biology and Laboratory of Molecular Biology, Department of Crystallography, Birkbeck College, London University,
Malet St., London, United Kingdom, WC1E7HX.

SUMMARY: X-ray structure analyses of proteins and computational approaches to the comparison of three-dimensional structures provide a basis for understanding the nature of restraints on the diversity of sequences in families of homologous proteins. Detailed examples are provided by structures defined by X-ray analysis at Birkbeck for two families of homologous proteins, the beta/gamma crystallins (five proteins) and aspartic proteinases (five enzymes). In addition all families of proteins, for which two or more well-refined high-resolution structures are available in the Brookhaven Databank, have been compared. Residue to residue substitution tables have been calculated for amino acids classified according to residue type, secondary structure, accessibility of the sidechain, and existence of hydrogen bonds from sidechain to other sidechains or peptide carbonyl or amide functions. Distinct patterns of substitution characterize most classes especially where amino acid residues are both solvent inaccessible and hydrogen-bonded through their sidechains.

INTRODUCTION

Divergent evolution gives rise to families of proteins that have similar tertiary structures but often rather different sequences.

Although random mutations occur in the DNA, the amino acid substitutions that are accepted at equivalent positions are determined by structural and functional restraints on the protein. If the pattern of amino acid substitution reflects restraints from the tertiary structure, we may ask two related questions. First, if we know the tertiary structure of a protein, can we predict the sequence variation that may occur in other family members? Second, if we observe a particular pattern of substitution in a family of homologous proteins, can we predict the structural environment - the local secondary structure, the solvent accessibility, the sidechain hydrogen bonding - of the amino acid on the assumption that it is not involved in the function? This article reviews our experimental and computational studies that address these questions.

X-RAY ANALYSES OF FAMILIES OF PROTEINS

Although much information can be gained from the study of a single three-dimensional structure of a protein in conjunction with a large number of sequences, there are many uncertainties in this procedure. These relate to the difficulties in making sequence alignments that are meaningful in structural terms especially if the sequence identities are less than 40%. Furthermore, regions that are aligned optimally by sequence may still have quite different local conformations and tertiary interactions. For these reasons it is desirable to consider families of proteins for which several three-dimensional structures are available. It is useful to have some protein families where the researchers involved in the comparisons are familiar with the details of the structure and function of each of the proteins compared. There is no better situation in this respect than a family of proteins where all or many homologues have been studied biochemically and by X-ray analysis in the laboratory. Such is the case with the beta/gamma crystallins and the aspartic proteinases at Birkbeck.

The beta/gamma crystallins comprise a family of homologous

proteins, the protomers of which are composed structurally of four Greek key motifs arranged as two globular domains. X-ray analyses at high resolution have defined the three-dimensional structures of four monomeric gamma crystallins and one oligomeric

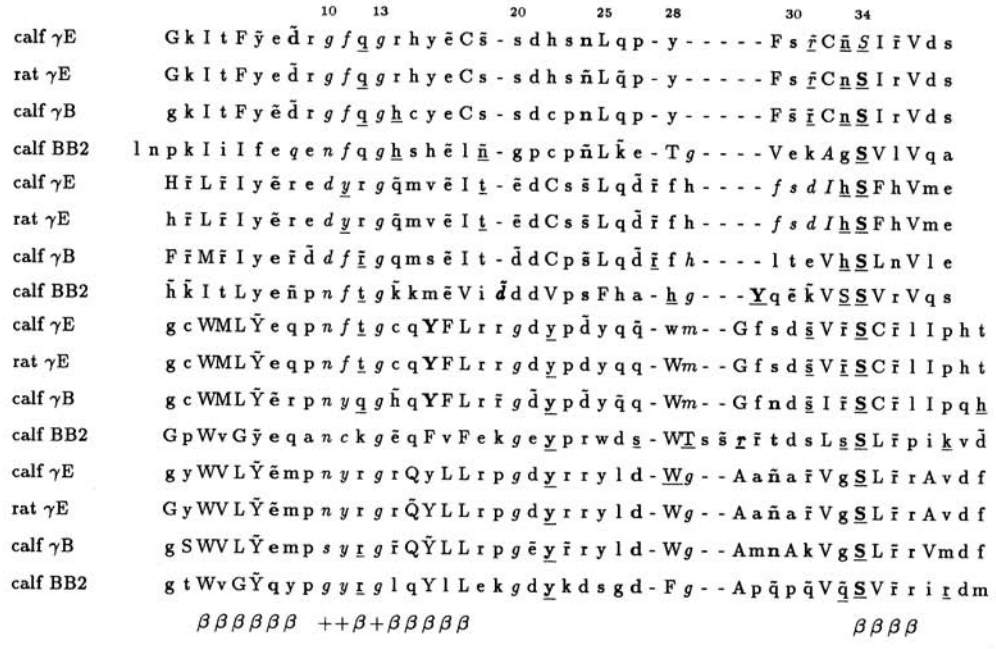

Figure 1. The alignment of sequences of the four Greek key motifs of three gamma and one beta crystallin obtained by comparing their 3-D structures using COMPARER. For each sequence there are four motifs. The numbering is that of the first motif of gamma B. The amino acid code is the standard one-letter code formatted using the following convention (Overington et al., 1990): Italic for positive phi; UPPER CASE for solvent inaccessible; lower case for solvent accessible; **bold** for hydrogen bonds to mainchain amide nitrogen; underline for hydrogen bonds to mainchain carbonyl oxygen; tilde~ for sidechain-sidechain hydrogen bonds. The secondary structure is given below where it is present in 80% or more of the proteins. α: α-helix; ß:ß-strand; +: positive PHI torsion angle.

beta crystallin (Blundell et al., 1981; Sergeev et al., 1988; White et al., 1989; Bax et al. 1990). Although the gamma crystallins have sequence identities of about 80%, ßB2 crystallin is less closely related. There are ~30% identities with the gamma crystallins. There are twenty structures for the Greek key motifs defined by X-ray analysis and more than 120 sequences of the motifs from homologous proteins available in sequence data bases. The individual motifs are not closely related. Only one residue, a glycine (Gly 13) is identical in all structures and one further residue, a serine (Ser 34), is conserved in most sequences but varied occasionally to alanine. The sequences of several motifs from beta and gamma crystallins are shown in Figure 1.

The aspartic proteinases include the pepsins, which are bilobal enzymes with a deep and extended active site cleft. The two catalytically active aspartates lie in conserved sequences (Asp-Thr-Gly) at the centre of the cleft. These catalytic aspartates (Asp 32 and Asp 215 in pepsin) occupy equivalent positions on the two lobes, which have little other sequence identity but have topologically similar structures (Tang et al. 1978). Our detailed analyses at Birkbeck are of the structures of the mammalian enzymes, chymosin (Newman, 1990; Strop et al. 1990) and pepsin (Cooper et al. 1990), and two fungal enzymes, endothiapepsin (Blundell et al. 1989) and mucorpepsin (Newman, 1990). These enzymes are between 25% and 60% identical when the sequences are considered pairwise. Other three-dimensional structures of aspartic proteinases available from the Brookhaven Databank include the fungal enzymes penicillopepsin (James and Sielecki 1983) and rhizopuspepsin (Suguna et al. 1987) and independent structures of pepsin (Andreeva et al. 1984; Sielecki et al. 1990) and chymosin (Gilliland et al. 1990). The aspartic proteinases also include the retroviral proteinases. Structures have been determined for proteinases of Rous Sarcoma Virus (RSV; Miller et al., 1989) and Human Immunodeficiency Virus (HIV; Wlodawer et al. 1989; Lapatto et al. 1989). In these dimeric enzymes each subunit corresponds to a single lobe of the pepsins and contributes one catalytic aspartate within a conserved sequence of Asp-Thr/Ser-Gly. Only three residues are identical in all lobes/subunits of

aspartic proteinases. These include the aspartate (Asp32 and equivalents) and glycine of the sequences at the active sites and a further glycine (Gly 122 and equivalents) in a strand that is closeby. Sections of the sequences of these aspartic proteinases in the conserved regions are shown in Figure 2.

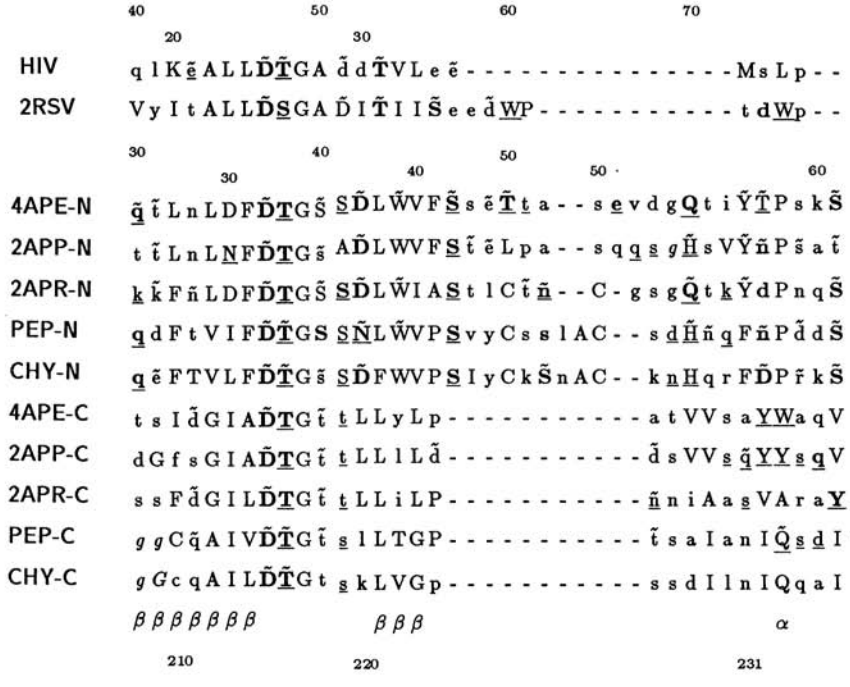

Figure 2. A section of the alignment of sequences of aspartic proteinases achieved by comparing the three-dimensional structures using COMPARER [Sali and Blundell, 1989]. APE: endothiapepsin; APP: penicillopepsin; APR: rhizopuspepsin; PEP: hexagonal porcine pepsin; CHY: calf chymosin; RSV: Rous sarcoma virus proteinase; HIV: human immunodeficiency virus proteinase. The last letter refers to the amino (N) or carboxy (C) terminal domains of the pepsins. One letter code as in Figure 1.

ENVIRONMENT SPECIFIC SUBSTITUTION TABLES

A study of these and other structures suggested that the residues

allowed at a particular position may result from structural constraints within the molecule except where the amino acids interact with the substrate or other molecules important to the function. For example, solvent-inaccessible residues, whose sidechains give a close-packed core, have a more limited set of substitutions than those on the surface [see also for example Chothia and Lesk, 1986; Hubbard and Blundell, 1987]. The requirement for an inter-residue hydrogen bond especially with peptide NH functions can also act as a constraint on the substitution of amino acids (Bajaj and Blundell, 1984; Blundell, 1986). Secondary structure also provides strong constraints on sequence variability; α-helices and ß-strands have preferred compositions.

We have attempted to characterize the structural constraints affecting the evolution of proteins (Overington et al., 1990). Our analysis depends on a systematic approach to the comparison of three-dimensional structures using COMPARER (Sali and Blundell, 1989; Zhu et al.,1990). COMPARER leads to an alignment of the sequences based upon the equivalence of the structures locally. COMPARER has been used to compare and align families of proteins such as the globins, serine and aspartic proteinases, phospholipases, cytochromes, immunoglobulins and crystallins, for which there are several high resolution X-ray analyses and coordinates in the Brookhaven Protein Databank (Bernstein et al. 1977). The alignments of the crystallins and aspartic proteinases shown in Figures 1 and 2 have been produced using this approach.

In the calculation of substitution tables (Overington et al., 1990) we first considered the structural features that appeared to be important in the families of proteins, the crystallins and aspartic proteinases. These were residue type (20 values), accessibility (2 values), side chain hydrogen bonding (8 values) and main chain conformation (4 values). However, some of these combinations do not exist. Several amino acids are unable to form hydrogen bonds through their sidechains and most polar residues are unable to act both as donors and acceptors except at extreme pH values. Furthermore, inaccessible ion pairs rarely occur except at domain or subunit interfaces which were largely omitted from the study. As a result of these factors the effective number

of classes was about three hundred.

All pairwise comparisons of structures in each alignment produced by COMPARER were considered in the analysis, and all substitutions implied by pairwise comparisons were stored in tables as a function of the features identified in the three-dimensional structures. In order to avoid very sparse tables, we considered the structural features of only one of the proteins compared. Secondly, in order to understand the role of certain structural features in constraining the mutability, we accumulated the values across various features (for example Figure 3).

Figure 3. A difference substitution table for amino acids in solvent inaccessible positions. The horizontal axis is that of an inaccessible amino acid in the 3-D structure of a protein. The vertical axis is the amino acid type in an homologous protein at a topologically equivalent position defined by COMPARER.

Figure 3 shows the difference substitution table for inaccessible residues. The values are calculated as differences between the tables for inaccessible and accessible residues. An increase in the conservation of a residue or a more favourable substitution due to the environment of the residue will be evident by a positive term. Inaccessible residues, even polar ones, are more conserved than those exposed to the solvent. Other substitution tables are shown in Overington et al. (1990).

LOCAL TERTIARY STRUCTURE CONSTRAINTS ON SEQUENCE VARIATION

The most characteristic substitution tables occur when combinations of features are considered. The difference substitution table for inaccessible residues with sidechain to mainchain carbonyl hydrogen bond shows that tryptophan is the residue whose substitution is most affected by such a hydrogen bond, followed by glutamine and tyrosine. It is surprising that, although glutamine occurs in this group, asparagine which has a similar sidechain amide function is not often found conserved forming a solvent inaccessible hydrogen bond to a carbonyl.

Figure 4 shows the substitution of buried Asp, Asn, Gln, Thr and Ser residues where there is a sidechain to mainchain nitrogen hydrogen bond. The largest value for conservation is seen for aspartic acid (Figure 4a). On the relatively infrequent occasions when substitutions are accepted at such positions, an asparagine or serine, which have similar hydrogen bonding capacity, are most likely to occur. This contrasts strongly with the substitution patterns of asparagine [Figure 4b]. Inaccessible asparagines with sidechain to mainchain NH hydrogen bonds are substituted often with aspartate or serine; leucines, alanines and many other residues are accepted. Surprisingly glutamine differs greatly from asparagine but resembles aspartate in its relatively high conservation. Its substitution profile indicates that glutamic acid and histidine are preferred substituents. Similar strong preferences for conservation are shown for solvent inaccessible serine and threonine.

There are several examples of polar inaccessible residues with their sidechains hydrogen bonded to mainchain functions in Figures 1 and 2. Some examples of the local structure and hydrogen bonding of such groups in the families of proteins with structures defined by X-ray analysis are given Figures 5 and 6. Indeed the very highly conserved threonine and serine residues of the aspartic proteinases and crystallins described above have hydrogen bonds to both mainchain NH and CO functions and are also inaccessible to solvent. Within the two domains of the pepsins Tyr 14 and Trp 189 are equivalent residues in the two lobes; each

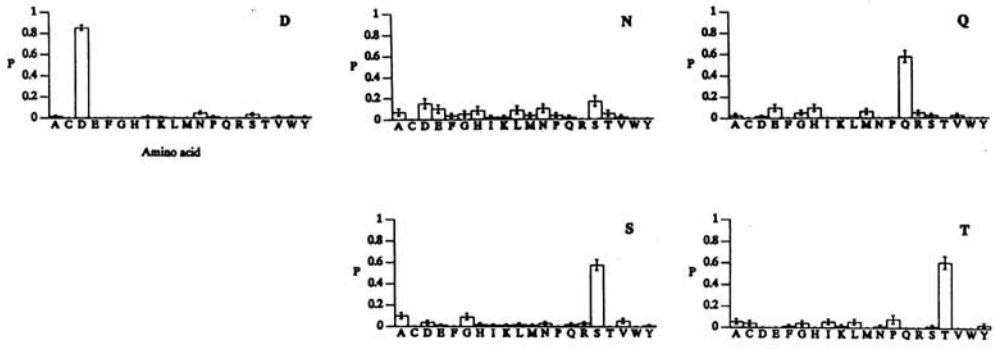

Figure 4. Patterns of substitution for amino acids that are solvent inaccessible and hydrogen-bonded to mainchain NH for (a) Asp, (b) Asn, (c) Gln, (d) Ser, (e) Thr. Probabilities (P) of a given residue being replaced by any of the 20 amino acids are given with standard errors.

is buried and hydrogen bonded to a mainchain carbonyl and is conserved within each of the domains. The substitution tables show that a buried and hydrogen-bonded tryptophan is strongly conserved and is most often substituted by a tyrosine or a hydrophobic group. In the retroviral proteinases it is replaced by a buried proline.

From these analyses it is clear that a sidechain oxygen hydrogen bond to a mainchain nitrogen is a larger factor in residue conservation than hydrogen bonds to mainchain oxygen or to another sidechain. Such effects have been noted in previous analyses of families of proteins [Bajaj and Blundell, 1985] but have not been characterized as a general factor in protein stability. The origin of the effect undoubtedly lies in the relatively greater importance of satisfying hydrogen bond donor properties of peptide NH compared to the acceptor properties of the peptide carbonyl on removal from aqueous environment. This is usually achieved with a mainchain carbonyl in regular secondary structure but in some conformations this is not possible; these conformations are characterized by the most conserved pattern of residues that occurs in protein evolution.

A positive mainchain PHI torsion angle also puts strong constraints on the variation of an amino acid. If the values PHI and PSI are in the region of +80 and -170, glycine is strongly preferred. Substitutions rarely occur if this conformation is important to the overall structure of a protein. Such an example is the uniquely conserved glycine (Gly 13; Figure 1) that facilitates the folding of a ß-hairpin onto the ß-sheet formed from Greek key motifs in the crystallins (Figure 5). Another glycine (Gly 10) also has a positive PHI angle but lies in the area of left-handed alpha helix. This is relatively well conserved but can be substituted mainly by aspartic acid or asparagine (see

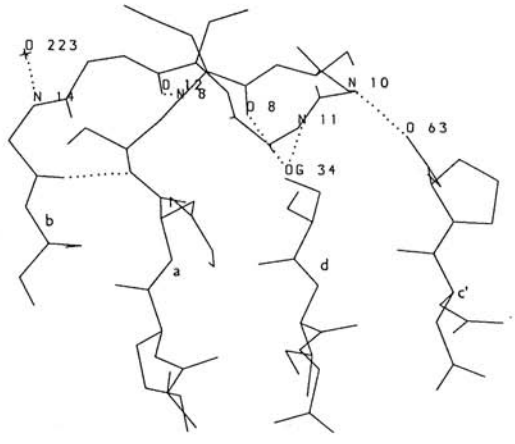

Figure 5. The hydrogen bonding arrangement of the solvent inaccessible serine (Ser 34) and two glycines with positive PHI torsion angles in motif 1 of gamma B that are strongly conserved in the repeated Greek key motifs of beta/gamma crystallins. The serine forms important hydrogen bonds to both mainchain amide and carbonyl functions. Gly 13 has values PHI and PSI in the region of +80 and -170, which give rise to steric repulsions for amino acids with sidechains. This conformation facilitates the folding of a ß-hairpin onto the ß-sheet formed from Greek key motifs in the crystallins. Another glycine (Gly 10) also has a positive PHI angle but lies in the area of left-handed alpha helix. Apart from glycine only aspartic acid, asparagine or serine are easily accommodated in this conformation (see Figure 1).

Figure 1). The restrictions on a glycine with a negative PHI are also great if the C-α is buried, e.g. the conserved glycines in aspartic proteinases at positions 34 (Figure 2) and 122.

Gly 34 and its equivalents in the sequence Asp-Thr-Gly have very low accessibility in the free enzyme and become quite inaccessible in the enzyme inhibitor, and presumably enzyme transition state, complexes. The glycine close to the active site (Gly 122 and its equivalents) is inaccessible in the free enzyme; there is no space for a side chain and it is invariant. The high conservation of such glycines is also apparent from the general substitution data for inaccessible glycines.

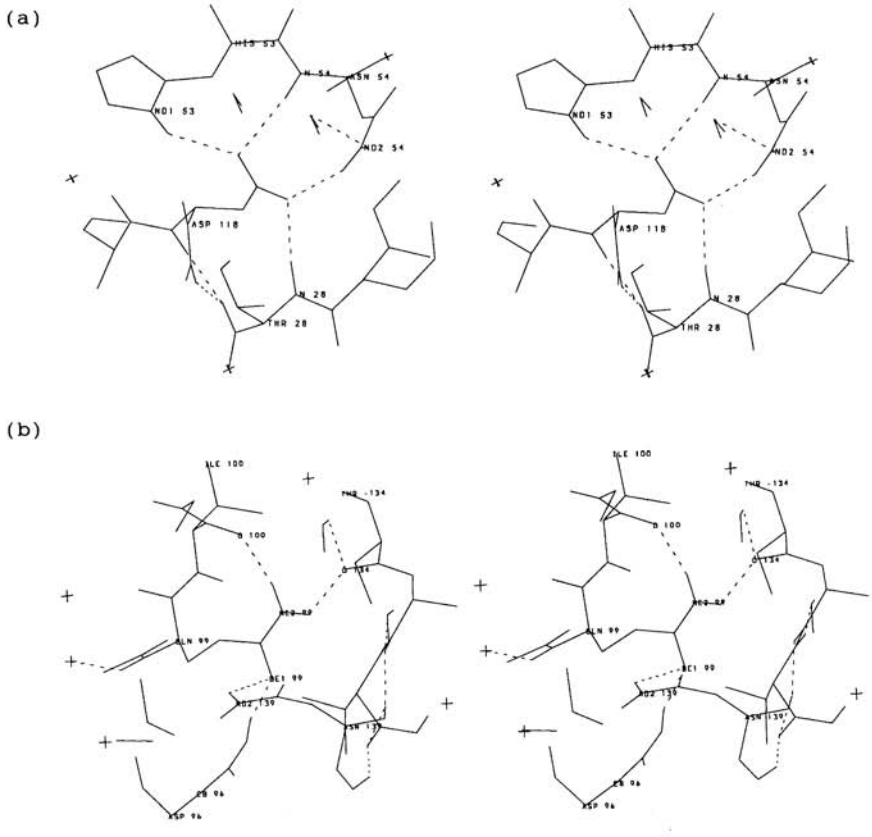

Figure 6. The hydrogen bonding arrangement of the solvent inaccessible and strongly conserved (a) aspartic acid (Asp 118) and (b) glutamine (Gln 99) in aspartic proteinases.

CONCLUSIONS

The structural data, comparisons and substitution tables provide a first quantitative analysis of amino acid diversity in homologous proteins. We have already shown that they can be used predictively to estimate the probable amino acid variation at each position in a protein of known three dimensional structure (Overington et al. 1990; M. Johnson, unpublished results). This provides a general approach to constructing templates on the basis of the tertiary structure. The method is complementary to the more theoretical approach of Ponder and Richards [1988]. Matching of the substitution patterns observed in large families of proteins for which there are many sequences but no three-dimensional structures may provide useful clues about the local secondary and tertiary structures (D.Donnelly, unpublished results). This may be achieved by scanning the sequence alignments for α-helical and ß-strand preiodicity on the basis that solvent inaccessible, solvent accessible and lipid accessible residues have distinct substitution patterns.

ACKNOWLEDGEMENTS

We thank the UK Science and Engineering Research Council, Agricultural and Food Research Council, Imperial Cancer Research Fund, American Cancer Society, Slovenian Research Council, The J. Stefan Institute, the Royal Society, the European Commision, Pfizer, and Merck Sharp and Dohme for financial support. We are grateful to our colleagues Devon Carney, Pam Thomas and Lynn Sibanda for many stimulating discussions.

REFERENCES

Andreeva, N.S., Zdanov, A.S., Gustchina, A.E. & Federov, A.A. [1984] J.Mol.Biol. 259, 11353-11365

Bajaj, M. and Blundell, T.L. (1984) Ann. Rev. Biophys. Bioeng. 13, 453-492

Bax, B., Lapatto, R., Nalini, V., Driessen, H., Lindley P.F., Mahadevan, D., Blundell, T.L. and Slingsby, C. (1990) Nature In Press

Bernstein, F.C., Koetzle, T.F., Williams, G.J.B., Mayer, E. F., Brice, M.D., Rogers, J.R., Kennard, O., Shimanouchi, T. & Tasumi, N. (1977) J. MOl. BIol. 112, 535-542.
Blundell, T,L. (1986) Chemica Scripta 26B, 213-219.
Blundell, T.L., Lindley, P., Miller,L., Moss, D.S., Slingsby, C., Tickle, I.J., Turnell, W.G. and Wistow, G. (1981) Nature 289, 771-777
Blundell, T.L., Jenkins, J.A., Sewell, B.T., Pearl. L.H., Cooper, J.B., Wood, S.P. and Veerapandian, B. [1990] J. Mol. Biol. 211, 919-941.
Chothia,C.& Lesk,A.M. (1986). EMBO J. 5, 823-826
Cooper, J.B., Kahn, G., Taylor, G., Tickle, I.J. and Blundell, T.L. (1990) J. Mol. Biol. 214, 199-222
Gilliland, G.L., Winbourne, Y.L., Nachman, J. and Wlodawer, A. (1990) Proteins 8, 82-101
Hubbard,T.J.P. & Blundell,T.L. (1987). Prot. Eng. 1, 159-171
James, M.N.G. and Sielecki, A. (1983) J. Mol. Biol. 163, 299-361
Lapatto, R., Blundell, T.L., Hemmings, A., Overington, J., Wilderspin, A., Wood, S.P., Merson, J.R., Whittle, P.E., Danley, D., Geoghegan, K., Hawrylik, S., Lee, K., Scheld, S. & Hobart, P.M. Nature, 1989, 342, 299.
Miller, M., Jaskolski, M., Rao, J.K.M., Leis, J. and Wlodawer, A. (1989) Nature 337, 576-579.
Newman, M. (1990) PhD Thesis, University of London, 1990
Overington, J., Sali, A., Johnson, M. and Blundell, T.L. (1989) Proc. Roy. Soc.Lond. B 241,132-145.
Ponder,J.W. & Richards,F.M. (1987). Proteins , 775-791
Sali,A. & Blundell,T.L. (1990). J.Mol.Biol. 212, 403-428.
Sergeev, Y.V., Chirgadze, Y.N., Mylvaganam, S.E., Driessen, H., Slingsby, C. and Blundell, T.L. (1988) Proteins 4, 137-147.
Sielecki, A., Federov, A.A., Boodhoo, A., Andreeva, N. and James, M.N.G. (1990) J. Mol. Biol. 214, 143-170
Strop, P., Newman, M., Blundell, T.L. (1990) Biochemistry ,In Press
Suguna,K, Padlan,E.A., Smith,C.W., Carlson,W.D. & Davies,D.(1987b) Proc.Natl.Acad.Sci.USA 84, 7009-7013.
Tang, J., James, M., Sielecki, A., Jenkins, J.A. and Blundell, T.L. (1978) Nature 271, 618-621
White, H.E., Driessen, H.P.C., Slingsby, C., Moss, D.S. & Lindley, P.F. (1989) J. Mol. Biol. 207, 217-235
Woldawer, A., Miller, M., Jaskolski, M., Sathyanarayana, B., Weber, I., Selk, L., Clawson, L., Schneider, J., Kent, S. (1989) Science 245: 616-621.
Zhu, Z., Sali, A. & Blundell, T.L. (1990) Unpublished Results.