Biochemistry (Moscow), Vol. 70, No. 8, 2005, pp. 835-840. Translated from Biokhimiya, Vol. 70, No. 8, 2005, pp. 1013-1020.
Original Russian Text Copyright © 2005 by Chelliah, Blundell.
Originally published in Biochemistry (Moscow) On-Line Papers in Press, as Manuscript BM05-063, May 29, 2005.

Quantifying Structural and Functional Restraints on Amino Acid Substitutions in Evolution of Proteins

V. Chelliah and T. L. Blundell*

Department of Biochemistry, University of Cambridge, 80 Tennis Court Road,
CB2 1GA, Cambridge, England; E-mail: tom@cryst.bioc.cam.ac.uk

Received March 3, 2005

An essay in memory of Oleg Ptitsyn who was a major influence on our thinking about protein structure, folding, and function

Abstract—One of Oleg Ptitsyn's most important papers (Shakhnovich, E., Abkevich, V., and Ptitsyn, O. (1996) *Nature*, **379**, 96-98) describes how knowledge of structure and function can be used to understand better the nature of amino acid substitutions in families and superfamilies of proteins. The selective advantages of retaining structure and function during evolution can be expressed as restraints on the amino acid substitutions that are accepted.

Key words: substitution tables, local environments, functional restraints, structural restraints

SELECTIVE PRESSURES ON EVOLUTION

Although selective pressures in evolution operate at the level of the whole organism, they are mediated largely through the functions of proteins, expressed by the arrangement of the polypeptide chain in space. Thus, during evolution the overall architecture of the protein and the "active site" amino acids are selected for optimal function. Once optimized, structure and interactions become restraints on further evolutionary change, as long as the function remains advantageous to the whole organism.

Classical analyses of amino acid substitutions in protein evolution such as those of Dayhoff [1, 2] tend to ignore these subtleties and treat all amino acids equivalently. Nevertheless, the Dayhoff amino acid substitution tables can be used in the search for homologs that are not too divergently evolved. They have been used extensively to construct phylogenetic trees, but the distances between proteins on different branches are meaningful only if the amino acid substitutions are selectively neutral [3]. Changes in function or even the way that function is manifested, i.e., through different oligomeric states or allosteric controls, might be selectively advantageous and therefore increase the rate of acceptance of amino acid substitutions [4].

For true orthologs, the general architecture is conserved. The local environment of each amino acid will determine the rate of acceptance and the nature of amino acids that are substituted. Strictly this will be the environment of the amino acid in the productive state of the protein, for example in the transition state of an enzyme or in a multicomponent complex of a regulatory system. This is reflected in the conservative variation of those amino acids required for stable three-dimensional structure and maintenance of function.

Over great evolutionary distances when functional restraints might vary, the amino acid substitutions can appear to be random when considered at the level of the amino acid sequence. Because many differing amino acid sequences are able to attain the same three-dimensional arrangement of the polypeptide main-chain, there may be little restraint on amino acid sequences over long periods of time. As a consequence the invariant amino acids may comprise a very small percentage of the total and the sequence relationships between protein superfamily members fall into the twilight zone or become statistically insignificant.

* To whom correspondence should be addressed.

0006-2979/05/7008-0835 ©2005 Pleiades Publishing, Inc.

INTRODUCING PROTEIN ARCHITECTURE INTO EVOLUTIONARY ANALYSES

Because function is dependent on protein architecture, protein structure tends to be more conserved in evolution than sequence. Thus, it is in tertiary structures that the most distant relations between proteins may be evident [5]. We therefore need to find the relationship between amino acid substitutions and protein three-dimensional structure.

Over the years many analyses have been made of the amino acid substitutions occurring in specific protein families and superfamilies where three-dimensional structures of one or more members had been determined, starting with myoglobin and hemoglobin, but including lysozyme and α-lactalbumin, the serine proteinases, and the insulins [6, 7]. An early attempt to define quantitatively amino acid substitution patterns in terms of local amino acid environments in protein families [8, 9] considered a protein as a string of structural subsites, each of which could accommodate a selection of amino acids and still maintain its structural/functional properties. The local environment was defined by the secondary structure, solvent accessibility, and side-chain hydrogen bonding of the amino acid, all features that were known to impose restraints on amino acid substitutions.

Overington et al. [8, 9] defined secondary structure simply in terms of α, β, coil, and amino acids with a positive φ conformation. The first two reflected the differing propensities for the α-helix and β-strand, and the third the conformational variability found in loop regions where local structures frequently change during the evolution [6]. The fourth acknowledged the infrequency of conformers with a positive φ conformation due to unfavorable interactions between the side-chain of L-amino acids and the peptide groups in this conformation. The principal exception is for glycine that lacks a side-chain; less frequent are asparagine and aspartic acid that can exhibit favorable dipole–dipole interactions (which may not include hydrogen bonding) between the side-chains and the peptide carbonyls [10].

Solvent accessibility of the side-chain is also an important determinant of evolutionary variation. This is most usefully defined as inaccessible if the side-chains were less than 7% accessible to a 1.4 Å probe, when compared to an amino acid in an extended and isolated polypeptide chain. Residues that are completely buried are less easily substituted than those that are on the surface and accessible. However, substitutions are made of completely inaccessible residues, partly because compensating changes are made in adjacent amino acid side-chains in the core, but more usually because the relative positions of the elements of secondary structure change in evolutionary time as first described by Chothia and Lesk [11].

Finally, side-chain hydrogen bonding is a critical determinant of amino acid substitutions. Hydrogen bonds from a side-chain to a main-chain NH function are most important. This is because during protein folding many hydrogen bonds to water in the unfolded chain must be replaced by others within the protein, most usually NH to CO groups in the standard α and β secondary structures and in β turns. However, where this is not possible, side-chain functions substitute for main chain functions, becoming part of the critical hydrogen bond network and the amino acids involved are consequently less often substituted. This is also true for side-chain hydrogen bonds to main-chain carbonyl functions, although these are more often found bonds buried in a folded protein without hydrogen. Thus, polar side-chains such as glutamine or aspartic acid that are inaccessible and hydrogen bonded to one or more main-chain functions are the most highly conserved amino acids in evolution.

Overington et al. [8, 9] defined the three kinds of side-chain hydrogen bonds by donor–acceptor distances of not more than 3.5 Å, giving 2^3 classes of involvement in hydrogen bonding. Combined with two classes of solvent accessibility and four classes of secondary structure this leads to 64 classes. The variation of sequence under each set of constraints was studied using environment-specific amino acid substitution tables (ESSTs) [8, 9]. Such amino acid substitution tables can be quantitatively evaluated and shown to be useful by examining their affects on performance when they are used in sequence-structure homology recognition programs like Fugue [12]. The effect of structural features on alignment accuracy can be examined systematically by eliminating the structural features (e.g. positive φ angle, accessibility, hydrogen bonding) one by one [12]. The elimination of any of these structural features causes reduction in the alignment accuracy.

DIRECT FUNCTIONAL RESTRAINTS ON EVOLUTIONARY CHANGE

Although the local environment constrained substitution tables provide more accurate predictions of amino acid substitutions than those devised by Dayhoff, they are still unsatisfactory in many ways. This is partly because the structure used in the application of the substitution tables should really comprise the biologically active system: the productive enzyme–cofactor–substrate complex, the active multiprotein system, the nucleic acid protein complex, and so on.

Indeed the most evolutionary conserved part of a protein is the identity and arrangement of the amino acid residues in the active site (Fig. 1). Over great evolutionary distances, this conservation is achieved by maintaining the core and topological equivalence of structural elements, even though helices and sheets may translate and rotate relative to each other. Residues in loops or turns undergo substantial insertions, deletions, or conforma-

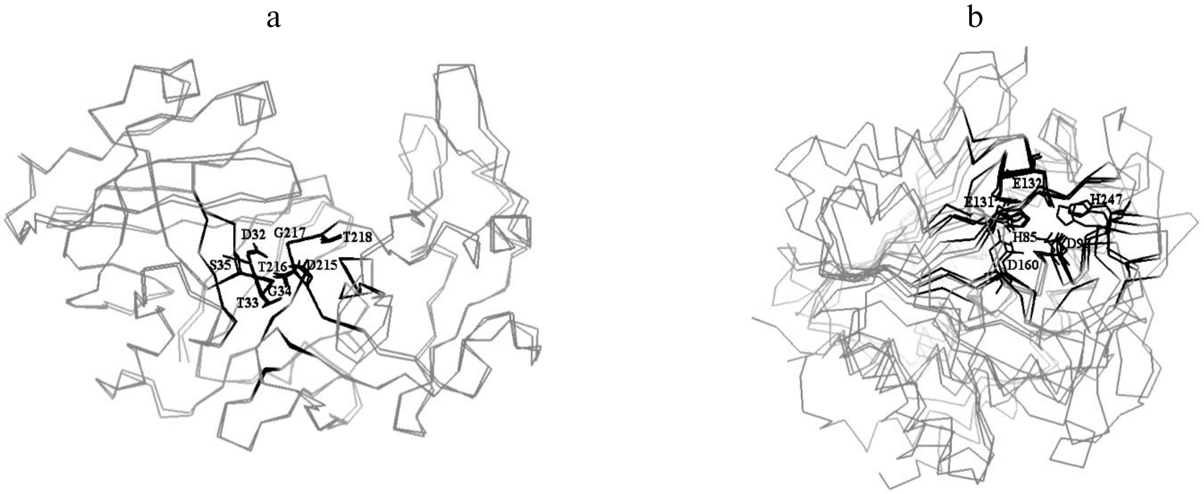

Fig. 1. The identity and arrangement of amino acid residues in the active site. The residues that are near the active site (within 9 Å distance from the active site center) are colored dark gray, the important identical residues in the active site are colored black and labeled. a) Superposition of aspartic proteinase family (2apr, 5pep, 1psn). Active site residues are numbered according to the PDB numbering of 5pep. b) Superposition of bacterial exopeptidase family (1xjo, 1amp, 1cg2). Active site residues are numbered according to the PDB numbering of 1xjo.

tional changes that accommodate relative movements of helices and sheets while leaving essential residues in the same arrangement. This has been described in [13], where it was shown that the relative positions of functional residues are highly conserved.

Mirny and Shakhnovich [14] have made an analysis on the molecular evolution of five of the most populated protein folds: immunoglobulin fold, oligonucleotide-binding fold, Rossman fold, alpha/beta plait, and TIM barrels, in order to distinguish between functional and structural reasons for amino acid conservation. We have developed an alternative method to identify regions mediating functional interactions with other molecules using the observations of the conservation of the nature and relative positions of active site residues [21]. Positions where environment-specific substitution tables make poor predictions of the overall amino acid substitution pattern are identified using Jensen and Shannon divergent score [15-17]. These are mapped onto the protein 3D structures and contoured, allowing identification of clusters of residues with strong evolutionary restraints. The clusters of residues apparently subjected to these additional restraints in evolution correlate well with the functional sites in proteins defined by experimental methods.

PREDICTING AMINO ACID SUBSTITUTIONS AS A FUNCTION OF THE RELATIONSHIP TO THE FUNCTIONAL SITE

Since the residues involved in protein interactions have strong evolutionary pressure to remain unchanged, these have different substitution patterns compared to those in non-interacting sites. However, the residues close to these functional residues will also tend to be under evolutionary restraints greater than those elsewhere, as they often provide the necessary infrastructure that supports the retention of the arrangement of the residues directly involved in functional interactions. This is most evident in enzymes because the catalytic activities are usually performed by a small highly conserved constellation of residues and therefore the region critical to the function can be most easily identified and located.

Thus, we have selected a set of enzymes of known structure with a well-defined active site. By associating the HOMSTRAD [18] database (a database of protein structure alignments for homologous families) with the Enzyme Structure Database (http://www.biochem.ucl.ac.uk/bsm/enzymes), 560 enzyme families were retrieved from HOMSTRAD. The Enzyme Structure Database contains all known enzyme structures deposited in the Protein Data Bank (PDB), classified by their Enzyme Commission (EC) [19] numbers. HOMSTRAD families with only one representative structure, those where domains are found in different entries in HOMSTRAD, and those that have insufficient information about their active sites were not included. The final data set contained 241 families with well-defined active site residues.

Active site information was retrieved from primary literature or from ACT_SITE annotations of the Swiss-Prot database [20] for at least one member of each family. For enzymes with known catalytic residues, only the residues directly involved in catalysis were included. For enzymes that do not have their catalytic residues precisely defined, the conserved residues that are likely to be involved in substrate or substrate–analog binding as reported in the literature were taken. The active site center was calculated using the average coordinates of the functional atoms (side-chain atoms that are likely to be involved in interactions, as defined in Chelliah et al. [21]) of the active site residues.

CALCULATION OF FUNCTION-DEPENDENT SUBSTITUTION TABLES

The local environments were described by the usual four classes of secondary structure (α-helix, β-strand, irregular structure (coil), and residue with a positive φ main-chain torsion angle), two classes of solvent accessibility (residues with side-chain relative accessibilities greater than 7% were defined as accessible, otherwise inaccessible), and three types of hydrogen bonds (side-chain to side-chain/heteroatom, side-chain to main-chain NH, and side-chain to main-chain CO) [8, 9, 12, 22]. In addition, the amino acid residues were defined according to the distance (less than or equal to 9 Å or greater than 9 Å from the active site center) between the active site center and the average coordinates of the functional atoms of each residue. The value of 9 Å was chosen based on the functional site prediction method described in Chelliah et al. [21]. When this method was applied to the 241 enzyme families, the distance between the observed and predicted active site centers was less than 9 Å for 82% of the families, a high enough success rate without compromising the definition of active site center.

Different types of amino acid substitution matrices (see the web site http://www-cryst.bioc.cam.ac.uk/~viji/sub_table) were derived based on these environments as described in the table. For each set of environments, tables of substitution frequencies were derived. The substitution tables were derived using the program SUBST (http://www-cryst.bioc.cam.ac.uk/~kenji/subst) with the same clustering and smoothing parameters as in Shi et al. [12]. The function-dependent environment-specific substitution tables were derived from 140 families with a total of 610 structures. Cystine and cysteine were considered to be different residues as they have distinct preferences for both local environment and differences in the patterns of accepted mutations. The assignment as a cystine was defined on the basis of a 2.5 Å sulfur-to-sulfur atom distance cutoff.

COMPARISON OF DIFFERENT AMINO ACID SUBSTITUTION TABLES

The amino acid substitution patterns observed in FD9-Non-ESSTs were compared with those observed in Non-ESST. The difference in substitution patterns between two matrices was quantified by calculating the Euclidean distance (D_{XY}) between the two matrices X and Y as:

$$D_{X,Y} = [\Sigma_i \Sigma_j (X_{i->j} - Y_{i->j})^2]^{1/2} ,$$

i, j = 1 to 21 (21, to distinguish cystine from cysteine);

where $X_{i\rightarrow j}$ and $Y_{i\rightarrow j}$ are the substitution frequency (multiplied by 100), of the amino acid type i substituted with amino acid type j in the matrices X and Y, respectively. The Euclidian distance between the Non-ESST and FD9-Non-ESSTs was 145 for the residues near the active site and 34 for the others. The larger distance (145) between the Non-ESST and FD9-Non-ESSTs near the active site implies that FD9-Non-ESSTs have distinct substitution patterns near the active site.

Classification of local environments

Type of substitution table	Structural/functional features	Number of local environments
ESSTs	Secondary structure, solvent accessibility and hydrogen bonding [12]	64 (4 × 2 × 2 × 2 × 2)
Non-ESST	None (analogous to Dayhoff)	1
Acc-only-ESSTs	Accessibility	2
FD9-ESSTs	As in ESSTs with two distance categories (those that are at the distance less than or equal to 9 Å and that are greater than 9 Å from the active site center)	128 (4 × 2 × 2 × 2 × 2 × 2)
FD9-non-ESSTs	Distance from the active site center (less than or equal to 9 Å and greater than 9 Å from the active site center)	2
Acc-only-FD9-ESSTs	Accessibility and the distance from the active site center feature	4 (2 × 2)

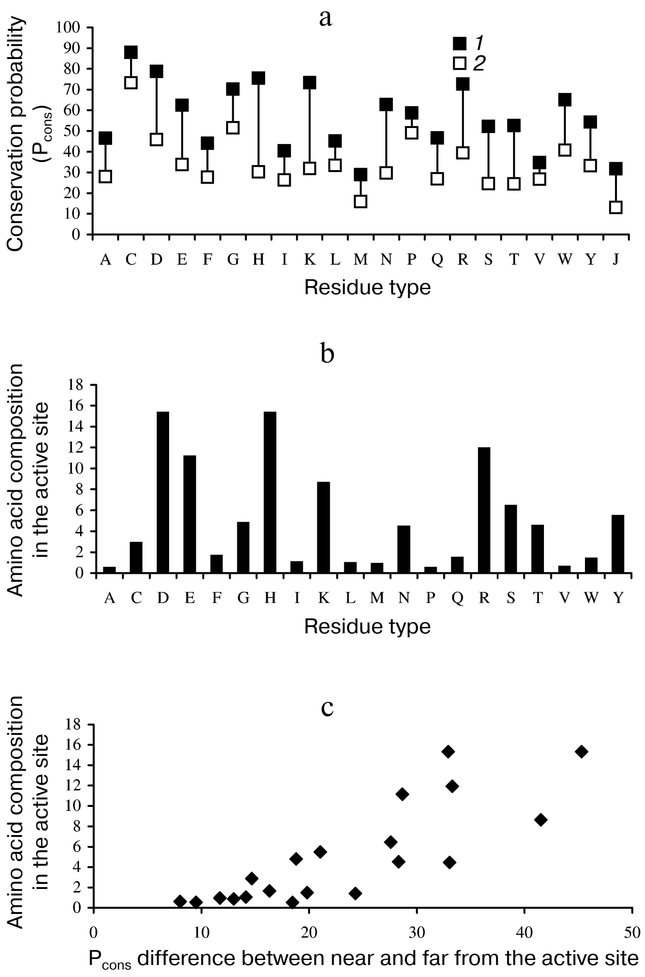

Fig. 2. a) Difference between conservation probability (P_{cons}) of the residues when they are near (*1*) and far from the active site (*2*). The standard one-letter amino acid code is used with the exception of C for cystine and J for cysteine. b) Amino acid composition in the active site of the 241 enzyme families (a total 1148 residues) compiled from the literature and Swiss-prot annotation. The standard one-letter amino acid code is used. c) Relationship between the difference in P_{cons} of the residues that are near and far from the active site and the amino acid composition in the active site (i.e. relationship between the information in Figs. 2a and 2b).

The conservation probability (P_{cons}) was defined as the probability that a residue will not be substituted by any other residue type (the entries along the diagonals in the matrix). The P_{cons} values clearly show that the residues near the active site are more highly conserved than those that are far from the active site (Fig. 2a). The differences between the P_{cons} values near the active site and far from the active site are greater for the residues that have higher likelihood to be directly involved in catalysis. The largest differences in the conservation probability (P_{cons}) are observed for His, Asp, Glu, Arg, and Lys. Figure 2b shows the amino acid composition in the active site for a total of 1148 residues compiled from the literature and the SWISS-PROT annotation for the 241 enzyme families analyzed. The frequencies of His, Asp, Glu, Arg, and Lys in the active site are the highest among all the amino acids. The differences in P_{cons} near the active site and far from the active site correlate well with the amino acid composition in the active sites, the correlation coefficient being 0.82 (Fig. 2c). This suggests that the substitution patterns are well defined by the residue proximity to the active site.

AMINO ACID TYPE AND ACCESSIBILITY

The largest differences in patterns of accepted mutations occur between accessible and inaccessible classes; for all residues a buried position is better conserved than a surface position [8]. The residues that show the greatest differences in conservation as shown by Acc-only-ESSTs are generally polar residues, for example, Asp, His, Arg, and Lys as shown in Fig. 3a.

When the distance from the active site center is included along with the accessibility (compare Acc-only-

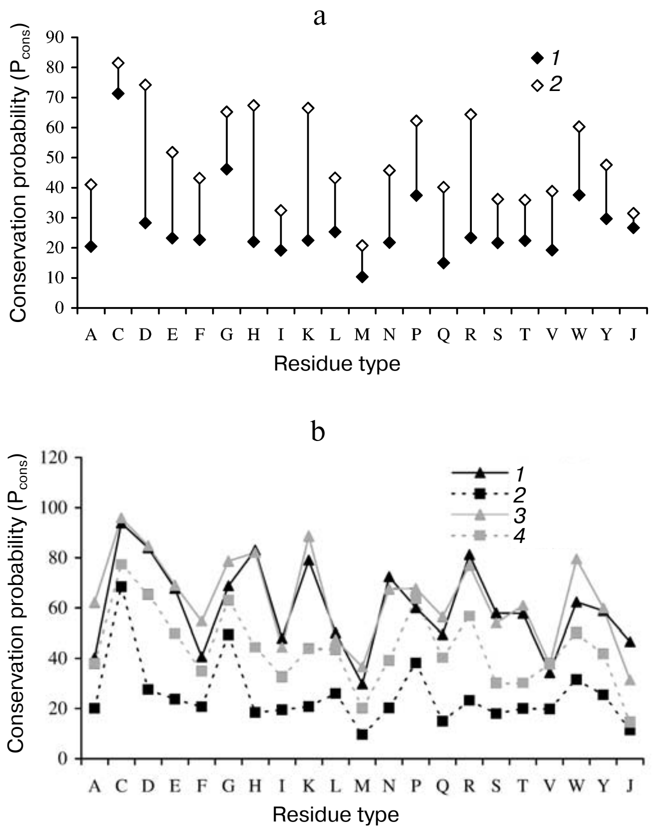

Fig. 3. a) Difference in P_{cons} of the residues when they are accessible (*1*) and buried (*2*) as observed in Acc-only-ESST. b) Comparison of conservation pattern (P_{cons}) of the residues when they are accessible near the active site (*1*), accessible far from the active site (*2*), buried near the active site (*3*), and buried far from the active site (*4*) as observed in Acc-only-FD9-ESST. The standard one-letter amino acid code is used with the exception of C for cystine and J for cysteine (see text for more details).

FD9-ESSTs with Acc-only-ESSTs), it can be seen that the differences in P_{cons} for accessible and inaccessible residues near the active site (less than 9 Å) are smaller than those of residues far from the active site (Fig. 3b).

The substitution patterns of the accessible residues in the active site are similar to those buried near the active site. This is not surprising because they are buried in enzyme substrate complexes and hydrogen bonded. In addition to the observation of [8, 9] that buried polar residues are more conserved and have more distinct conservation patterns than those that are accessible, we have found that the buried polar residues are even more conserved near the active site than in other places. Accessible aromatic residues near the active site are less conserved than buried residues close to the active site, presumably because they are usually not directly involved in catalysis but more often in substrate binding and specificity (for example, Phe and Trp).

SEQUENCE–STRUCTURE HOMOLOGY RECOGNITION USING FD9-ESSTs

The FD9-ESSTs were tested in the sequence-structure homology recognition program FUGUE [12] and compared with the recognition performance obtained using the standard ESSTs and other sequence alignment programs [23]. Both the recognition performance and the alignment accuracy are improved using the FD9-ESSTs. The alignments near the active site are greatly improved using the FD9-ESSTs and the improvements are statistically significant (p value is 0.02, according to Wilcoxon [24, 25] signed rank test).

In this essay, we have described how knowledge of structure and function can be used to understand better the nature of amino acid substitutions in families and superfamilies of proteins. The selective advantages of retaining structure and function during evolution can be expressed as restraints on the amino acid substitutions that are accepted. Oleg Ptitsyn would have immediately identified another restraint that should be included, that due to protein folding, and indeed we have been trying to address this (P. Fanon, C. M. Dodson, and T. L. Blundell, unpublished results). A particular influence on this work has been one of Oleg Ptitsyn's most important papers on the effects of protein folding on conservation [26]. A task for the future is to successfully include these in the substitution tables.

REFERENCES

1. Dayhoff, M. O., Schwartz, R. M., and Orcutt, B. C. (1978) *A Model for Evolutionary Change*, National Biomedical Research Foundation, Washington, D.C.
2. Jones, D. T., Taylor, W. R., and Thornton, J. M. (1992) *Comput. Appl. Biosci.*, **8**, 275-282.
3. Kimura, M. (1983) *The Neutral Theory of Molecular Evolution*, Cambridge University Press, Cambridge.
4. Blundell, T. L., and Wood, S. P. (1975) *Nature*, **257**, 197-203.
5. Ptitsyn, O. B., and Finkelstein, A. V. (1980) *Q. Rev. Biophys.*, **13**, 339-386.
6. Bajaj, M., and Blundell, T. (1984) *Annu. Rev. Biophys. Bioeng.*, **13**, 453-492.
7. Ptitsyn, O. B. (1974) *J. Mol. Biol.*, **88**, 287-300.
8. Overington, J., Donnelly, D., Johnson, M. S., Sali, A., and Blundell, T. L. (1992) *Protein Sci.*, **1**, 216-226.
9. Overington, J., Johnson, M. S., Sali, A., and Blundell, T. L. (1990) *Proc. Roy. Soc. Lond. B Biol. Sci.*, **241**, 132-145.
10. Deane, C. M., Allen, F. H., Taylor, R., and Blundell, T. L. (1999) *Protein Eng.*, **12**, 1025-1028.
11. Chothia, C., and Lesk, A. M. (1986) *EMBO J.*, **5**, 823-826.
12. Shi, J., Blundell, T. L., and Mizuguchi, K. (2001) *J. Mol. Biol.*, **310**, 243-257.
13. Irving, J. A., Whisstock, J. C., and Lesk, A. M. (2001) *Proteins: Struct. Func. Genet.*, **42**, 378-382.
14. Mirny, L. A., and Shakhnovich, E. I. (1999) *J. Mol. Biol.*, **291**, 177-196.
15. Kullback, S. (1959) *Information Theory and Statistics*, John Wiley and Sons, New York.
16. Lin, J. (1991) *IEEE Trans Info. Theory*, **37**, 145-151.
17. Yona, G., and Levitt, M. (2002) *J. Mol. Biol.*, **315**, 1257-1275.
18. Mizuguchi, K., Deane, C. M., Blundell, T. L., and Overington, J. P. (1998) *Protein Sci.*, **7**, 2469-2471.
19. Webb, E. C. (ed.) (1992) *Enzyme Nomenclature 1992. Recommendations of the Nomenclature Committee of the International Union of Biochemistry and Molecular Biology on the Nomenclature and Classification of Enzymes*, Academic Press, New York.
20. Boeckmann, B., Bairoch, A., Apweiler, R., Blatter, M. C., Estreicher, A., Gasteiger, E., Martin, M. J., Michoud, K., O'Donovan, C., Phan, I., Pilbout, S., and Scheider, M. (2003) *Nucleic Acids Res.*, **31**, 365-370.
21. Chelliah, V., Chen, L., Blundell, T. L., and Lovell, S. C. (2004) *J. Mol. Biol.*, **342**, 1487-1504.
22. Johnson, M. S., Overington, J. P., and Blundell, T. L. (1993) *J. Mol. Biol.*, **231**, 735-752.
23. Chelliah, V., Mizuguchi, K., and Blundell, T. L. (2005) *Proteins: Struct. Func. Bioinf.*, in press.
24. Siegel, S., and Castellan, N. J. (1988) *Nonparametric Statistics for the Behavioral Sciences*, McGraw-Hill International Editions, New York.
25. Wilcoxon, F. (1947) *Biometrics*, **3**.
26. Shakhnovich, E., Abkevich, V., and Ptitsyn, O. (1996) *Nature*, **379**, 96-98.