Structural and functional restraints in the evolution of protein families and superfamilies

Sungsam Gong*, Catherine L. Worth*†, G. Richard J. Bickerton*‡, Semin Lee*, Duangrudee Tanramluk* and Tom L. Blundell*[1]

*Department of Biochemistry, University of Cambridge, 80 Tennis Court Road, Cambridge CB2 1GA, U.K., †Structural Bioinformatics Group, Leibniz-Institut für Molekulare Pharmakologie, Campus Berlin-Buch, Robert-Rössle-Strasse 10, 13125 Berlin, Germany, and ‡Medicinal Informatics, Division of Biological Chemistry and Drug Discovery, College of Life Sciences, University of Dundee, Dundee DD1 5EH, U.K.

Abstract

Divergent evolution of proteins reflects both selectively advantageous and neutral amino acid substitutions. In the present article, we examine restraints on sequence, which arise from selectively advantageous roles for structure and function and which lead to the conservation of local sequences and structures in families and superfamilies. We analyse structurally aligned members of protein families and superfamilies in order to investigate the importance of the local structural environment of amino acid residues in the acceptance of amino acid substitutions during protein evolution. We show that solvent accessibility is the most important determinant, followed by the existence of hydrogen bonds from the side-chain to main-chain functions and the nature of the element of secondary structure to which the amino acid contributes. Polar side chains whose hydrogen-bonding potential is satisfied tend to be more conserved than their unsatisfied or non-hydrogen-bonded counterparts, and buried and satisfied polar residues tend to be significantly more conserved than buried hydrophobic residues. Finally, we discuss the importance of functional restraints in the form of interactions of proteins with other macromolecules in assemblies or with substrates, ligands or allosteric regulators. We show that residues involved in such functional interactions are significantly more conserved and have differing amino acid substitution patterns.

Introduction

An understanding of protein evolution requires not only knowledge of genomes, protein sequences, structures and functions, but also an understanding of selective pressures at the level of the whole organism and the role of the protein in cells and whole-organism systems [1,2].

Insights into the relationship of protein structure, function and evolution began to emerge nearly 50 years ago as protein structures were determined for which there were multiple sequences. For example, insulin sequences from Fred Sanger in the 1950s ([3,4] and see [5] for a review), together with the three-dimensional structure from Dorothy Hodgkin a decade later [6,7], provided clues about the impacts of amino acid substitutions on tertiary structure and precursor activation, on quaternary interactions at dimer and hexamer interfaces, and on the putative receptor-binding region [8]. They demonstrated that amino acid substitutions were accepted during evolution in a way that satisfied restraints arising from structure and function [9]. Thus the core of the protein tended to be relatively conserved (Figure 1), and residues in helices and strands were substituted in ways that maintained the overall stabilities of these secondary structures. Most interestingly, a glycine residue with a positive φ main-chain torsion angle that allowed the chain to change direction sharply was conserved in all insulins. Substitutions of amino acids at positions involved in dimer formation retained their hydrophobic character in all species except the hystricomorpha. The conservation of B10 histidine in most mammalian, fish and bird insulins was evidence of restraints arising from the existence of a hexamer.

The insulin structure also provided evidence of restraints from functional interactions. Residues in a patch mainly on the surface of the monomer appeared to have greater restraints on their substitution than could be explained by retention of the structure of insulin throughout evolution; this observation provided the clues about restraints in evolution arising from function, in this case the binding of insulin with its receptor.

Much of the sequence variation appeared to be selectively neutral; the accepted amino acids were able to fulfil the same structural and functional roles. Some amino acid changes, such as those occurring in the hystricomorph insulins, had been proposed to be the result of selectively neutral substitutions [10]. However, these substitutions proved to be consistent with loss of ability to dimerize and stabilization of the monomeric form. This presumably resulted from change of storage form, possibly related to zinc availability, and was therefore probably also selectively advantageous.

Thus the analyses of insulins, along with parallel work on globins, lysozymes and serine proteinases, provided strong evidence for the conservation of tertiary structure during

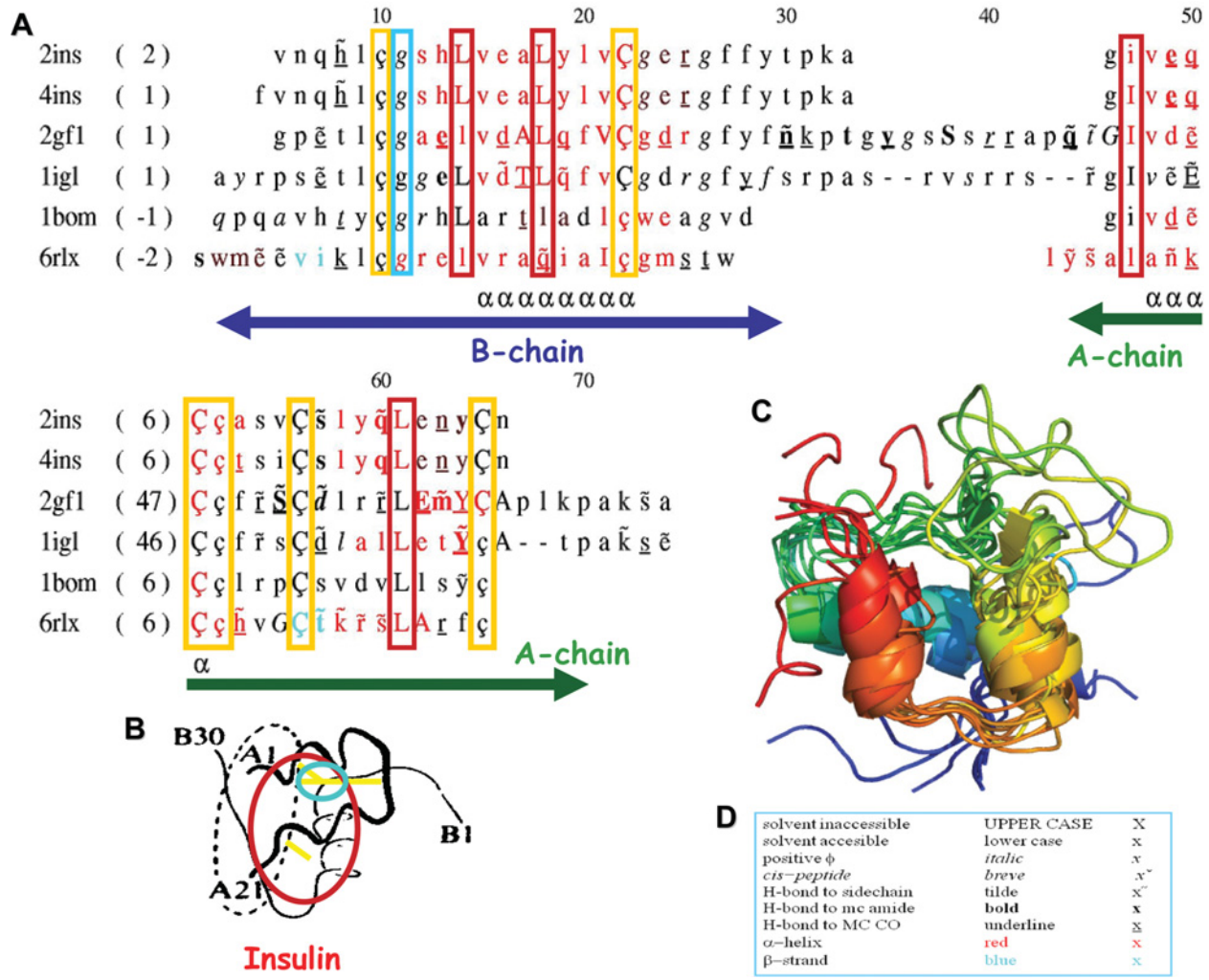
Key words: buried residue, hydrogen bond, local structural environment, protein evolution, protein structure.

Abbreviations used: ESST, environment-specific substitution table; PCA, principal component analysis.

[1]To whom correspondence should be addressed (email tom@cryst.bioc.cam.ac.uk).

Figure 1 | Evolution of insulins as demonstrated by sequence alignment and structural superimposition

(**A**) Six members of the insulin superfamily, structurally aligned and represented in JOY format [30]: (i) core residues that are largely inaccessible in all members in red box; (ii) half-cystines responsible for disulfide bonds in yellow box; and (iii) a glycine with a positive φ main-chain torsion angle in blue box. (**B**) The key residues represented within the three-dimensional structure of insulin. (**C**) Six superimposed structures of insulin proteins. (**D**) Legend of JOY format.



evolution, and emphasized the importance of considering restraints from protein interactions, in this case in terms of oligomers and receptor activation. They underlined the importance of local environment in the acceptance of amino acid substitutions during protein evolution.

Protein superfamilies

Subsequent analyses of more divergent members of superfamilies showed that these general features could still be preserved while allowing considerable sequence divergence so that sequence similarities were in the twilight zone (20–35% identity) and below. Indeed, analyses of existing protein structures and sequences provided an approach not only to understand the structure and function of proteins, but also to solving the inverse folding problem. It enabled sequence profiles for each protein family and its more distant relatives, the protein superfamily, to be generated. For insulin, this led to the identification of somatomedin C (insulin-like growth factor 1), relaxin and later many other molecules as members of the insulin superfamily [11].

Similar analyses led to the recognition of other distant superfamilies and at the same time to identification of further restraints on amino acid substitutions arising from local structural features. Thus, in the family of $\beta\gamma$-crystallins which have four copies of a Greek-key motif in each protomer, a buried serine residue, which is hydrogen-bonded to a main-chain amide function, stabilizes the folding of the loop-joining strands *a* and *b* over the globular domain in each motif and leads to the most conserved sequence pattern [12]. This motif was later identified in *Myxococcus xanthus* spore coat protein S [13], both proteins probably being selected for as a consequence of their stability.

A more celebrated example was the evolution of the aspartic proteinases where two gene duplication and fusion events were predicted in the evolution and a symmetrical ancestral dimer was proposed [14,15]. The buried threonine residue next to the catalytic aspartate residue in each Asp-Thr-Gly motif, which forms an important buried hydrogen-bond to a buried main-chain amide function, appears to be a critical structural restraint and an easily recognizable sequence motif. A similar symmetrical dimer was later predicted and identified in retroviral proteases from rous sarcoma and HIV [16], where the Asp-Thr-Gly sequence forms a similar role and retention of the dimeric form as a regulatory requirement in its activation appears to have provided an evolutionary selective advantage to the viruses.

Taken together, the aspartic proteinases and crystallins implied that, apart from solvent accessibility and secondary structure, side-chain hydrogen-bonding, particularly to main-chain functions in buried positions, might also be strong restraints on sequence variation.

Tertiary structural restraints

Such analyses of superfamilies led to the idea that propensities for amino acids [17] and their substitution patterns [18,19] might be systematically defined in terms of local structural environments. Solvent accessibility of the side chain and occurrence in regular secondary structures were local environments used by most groups [17,20,21]. Two further classes of local environment were added to these by Overington et al. [19]: (i) amino acids with a positive φ main-chain torsion angle (learnt from the B8 glycine residue of insulin); and (ii) amino acids with side chains that formed hydrogen bonds to main-chain or other side-chain functions (inspired by the conserved serine and threonine residues of the crystallins and aspartic proteinases).

Although buried main-chain functions achieve hydrogen-bond satisfaction through the formation of secondary structures in most positions in the core, buried polar residues often carry out this role. Interestingly, those amino acid residues with polar side chains whose hydrogen-bonding potential is satisfied tend to be more conserved than their unsatisfied or non-hydrogen-bonded counterparts, particularly when buried [22]. Indeed, such buried and satisfied polar residues are significantly more conserved in sequence identity than individual hydrophobic residues such as leucine in similar solvent-inaccessible local environments.

An ESST (environment-specific substitution table; http://www-cryst.bioc.cam.ac.uk/esst) describes the substitution of amino acids as a function of structural environments which restrict the possible and allowable substitutions [18]. The combination of environmental descriptors for solvent accessibility, secondary structure and side-chain hydrogen-bonding gives 64 matrices for each amino acid in this model, and each is associated with a distinct pattern of amino acid substitution. Figures 2(A)–2(D) demonstrate that amino acid substitution patterns are influenced by local structural environment. In particular, a solvent-inaccessible environment (Figure 2B) restricts the possible substitution of amino acids most strongly, enhancing the diagonal of the substitution matrix, but secondary structure and the existence of side-chain hydrogen bonds also lead to different substitution patterns. The relative importance of these has been demonstrated by an analysis of distances among the 64 tables, each characterized by a different set of restraints, followed by PCA (principal component analysis) [23] based on a matrix of substitution profiles for all 64 environments over 441 (21×21) possible substitutions (Figures 2E and 2F). In Figure 2(E), PCA divides the 64 environments by solvent accessibility, which corresponds to the primary principal component (PC1). As expected, we observed that, for all 21 amino acids, the degree of residue conservation in the solvent-inaccessible regions is much higher than that of solvent-accessible regions (see Supplementary Figure S1 at http://www.biochemsoctrans.org/bst/037/bst0370727add.htm). In Figure 2(G) and 2(H), it is very evident that the relative importance of the three types of hydrogen bond is very hierarchical; eight environments are divided by the existence of a hydrogen bond from a side-chain to main-chain amide (N/n) followed by a main-chain carbonyl (O/o), which correspond to the first and second component of PCA respectively.

Restraints from functional interactions with other proteins

As shown in the case of the insulins, functional interactions with other macromolecules can also provide restraints on the acceptance of amino acid substitutions. For protein–protein interactions, two structural environments can be defined for interfacial residues: (i) interface core for residues with relative accessibilities less than 7% [24]; and (ii) interface periphery for those with relative accessibilities greater than 7% (Figure 3A). Residue propensities can be calculated as the relative proportion of each residue type in each of the interfacial accessibility structural environments and for buried and exposed non-interface residue environments (Figure 3B). Propensities for the majority of residues at the interface core and periphery are intermediate between those for the protein core and exposed surface, with the interface periphery most similar to the exposed protein surface and the interface core most similar to the protein core. The exceptions to this are methionine, glycine, alanine, histidine, tryptophan, tyrosine and arginine. Of these, alanine and glycine, the two smallest residues, are disfavoured at the interface periphery. Histidine and arginine, two positively charged residues, are favoured at the periphery; in fact, this is the structural environment in which these residues are most enriched. Arginine is capable of multiple types of favourable interaction: it can simultaneously form up to five hydrogen bonds and an ionic salt bridge with the positive charge carried on its guanidinium motif. Tryptophan, tyrosine and methionine, the three largest hydrophobic residues that can engage in a range of interactions, are all favoured at the interface core, corresponding with the observations of Ofran and Rost [25]. The enrichment

Figure 2 | Different amino acid substitution patterns and the relative importance of structural restraints

(**A**–**D**) Four examples of ESST matrices. (**A**) Substitutions of residues occurring in β-strands, that are solvent accessible and which do not form any side-chain hydrogen bonds. (**B**, **C** and **D**) differ from (**A**) by one parameter defining the structural environment; (**B**) differs in the solvent accessibility (second letter), (**C**) in the element of secondary structure (first letter), (**D**) in the fulfilment of the hydrogen-bond from a side-chain to a main-chain amide group (fifth letter). Environments are shown using five-letter code representation: (i) first letter for secondary structures (H, α-helix; E, β-strand; P, positive φ main-chain torsion angle; C, coil); (ii) second letter for solvent accessibility (A, accessible; a, inaccessible); and (iii) remaining three letters for the existence (upper case) or absence (lower case) of hydrogen-bonds from a side-chain to other side-chain (S/s, third) and main-chain carbonyl (O/o, fourth) and main-chain amide (N/n, fifth). Each entry, within a column of an ESST, is coloured by the degree of a residue substitution for a given amino acid. The colour scale ranges from white (less conserved) to red (more conserved). The order of amino acids in a row (or column) is ACDEFGHIKLMNPQRSTVWYJ. [Note that there are 21 amino acids where C represents half-cystine (disulfide-bonded) and J represents cysteine (non-disulfide-bonded).] (**E** and **F**) A total of 64 environments plotted in the axis of first (PC1), second (PC2) and third (PC3) component of PCA. PC1, PC2 and PC3 are responsible for 31, 13 and 8% of the total variance respectively. PCA was performed from the substitution profile (64 × 21 × 21, see the text for details). A total of 64 environments are coloured using different schemes: (i) by the solvent accessibility (red: inaccessible, blue: accessible) in (**E**); (ii) by the element of secondary structure (red, α-helix; blue, β-strand; black, positive φ main-chain torsion angle; green, coil) in (**F**). (**G** and **H**) Eight hydrogen-bond types projected on to PC1, PC2 and PC3. PCA was performed on the 8 × 8 distance matrix by averaging the effect of the solvent accessibility (A/a) and the elements of secondary structure (H/E/P/C). Hence, the distance matrix reflects only the effect of hydrogen bonds from side chains. Eight environments are coloured by the existence (blue) or absence (black) of hydrogen bond from side-chain to main-chain amide and by the existence (red) or absence (black) of hydrogen bond from side-chain to main-chain carbonyls. PC1, PC2 and PC3 are responsible for 55, 18.7 and 11% of the total variance respectively. PCA projection was drawn using the RGL package of R software (http://www.r-project.org/).

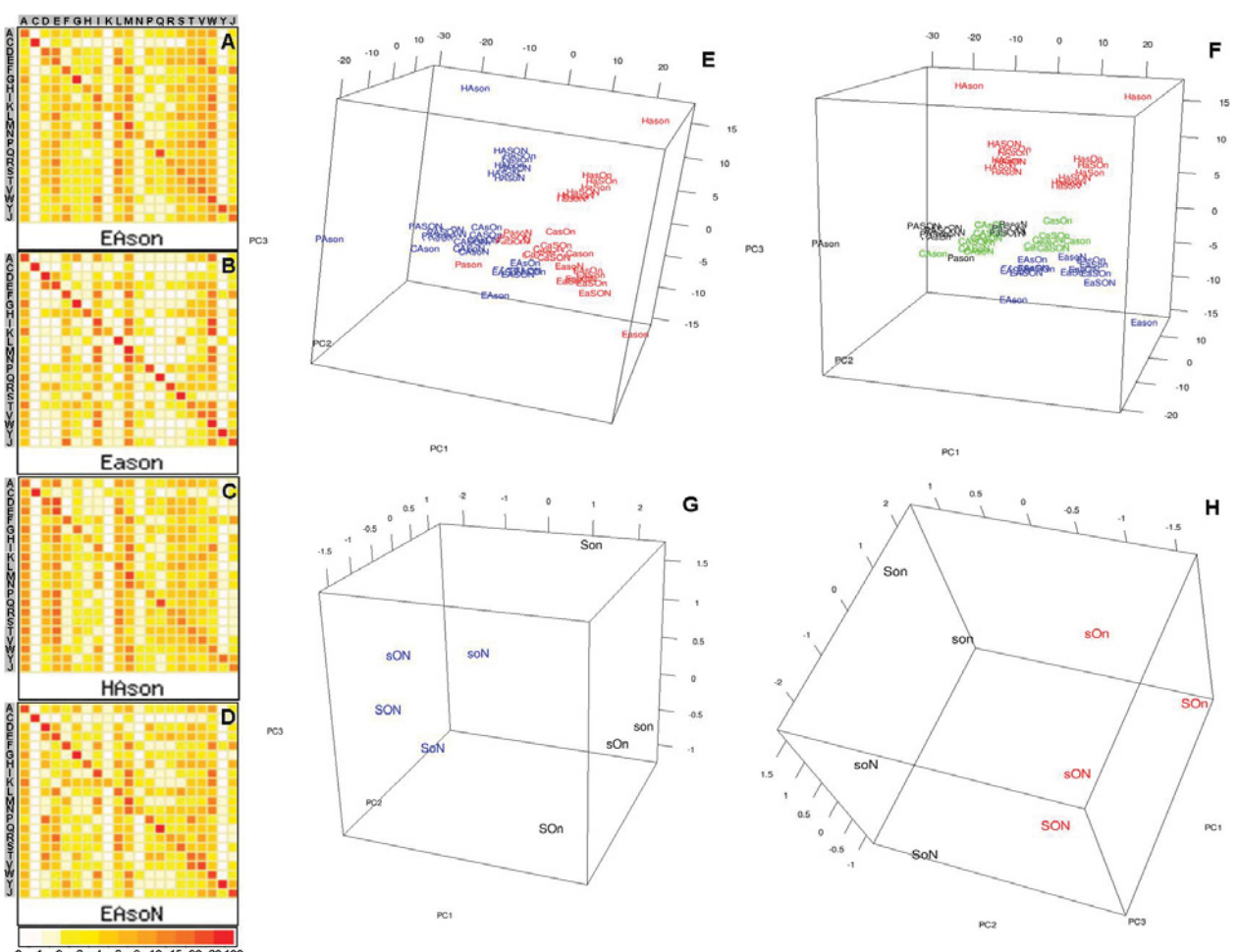
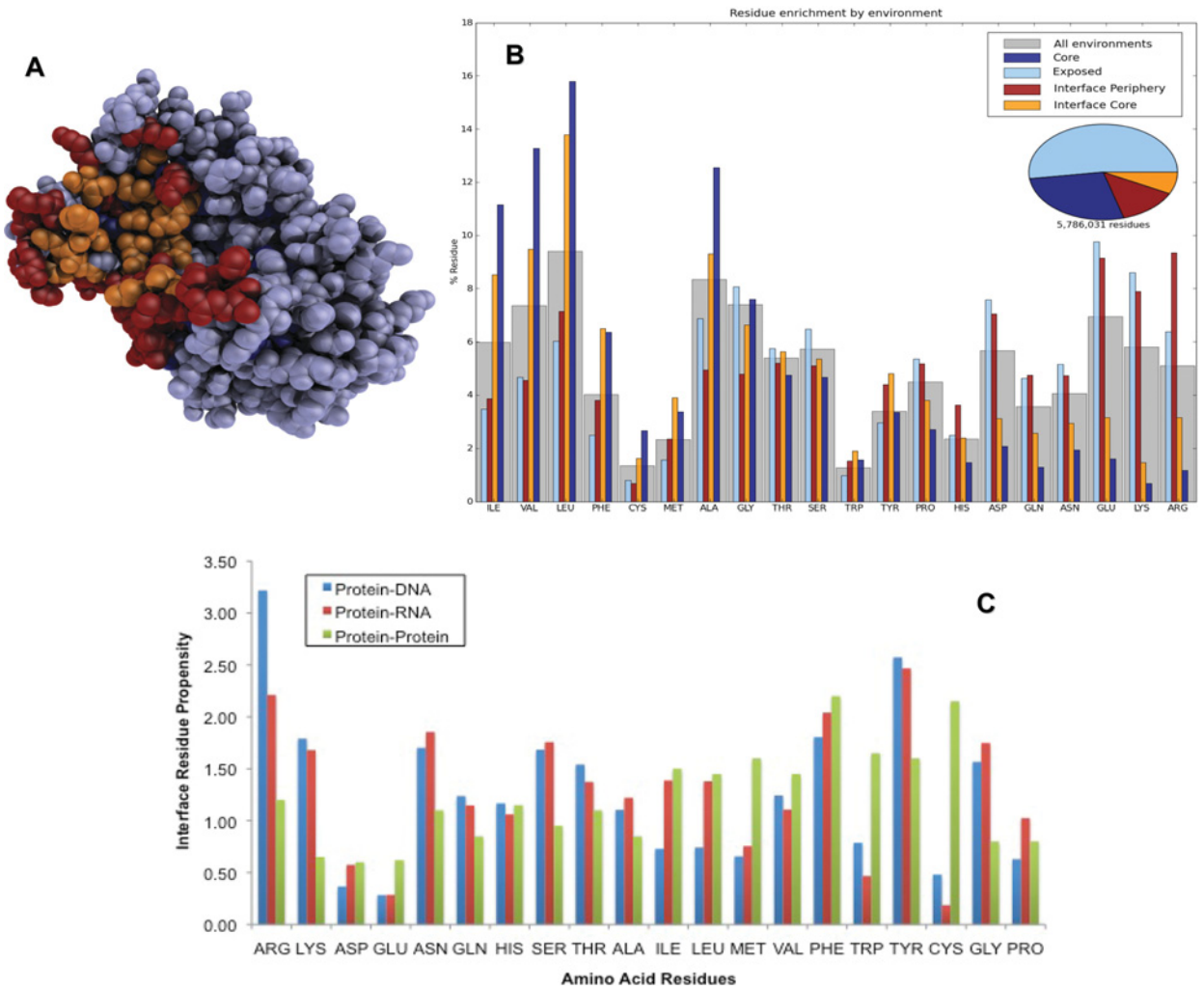


Figure 3 | Definition of interface residues and their propensities

(**A**) Four structural environments, defined for all protein–protein interfaces. Interfacial residues with relative accessibilities less than 7% are defined as interface core (orange), interfacial residues with relative accessibilities greater than 7% are defined as interface periphery (dark red), non-interface buried residues (dark blue) and exposed residues (light blue). (**B**) Residue propensities for each of the four structural environments (G.R.J. Bickerton, unpublished work). Colour scheme is consistent with (**A**). (**C**) Residue propensities at protein–DNA/RNA interfaces [31]. A propensity of >1 indicates that a residue occurs more frequently on the interface than on the protein surface.



of aromatic tyrosine may be explained by its contribution to the hydrophobic effect without a large entropic penalty owing to the side chain having few rotatable bonds as well as the hydrogen-bonding capacity of its 4-hydroxy group. Tryptophan has a very large aromatic side chain that can mediate aromatic π-interactions, act as hydrogen-bond donor and form extensive hydrophobic contacts.

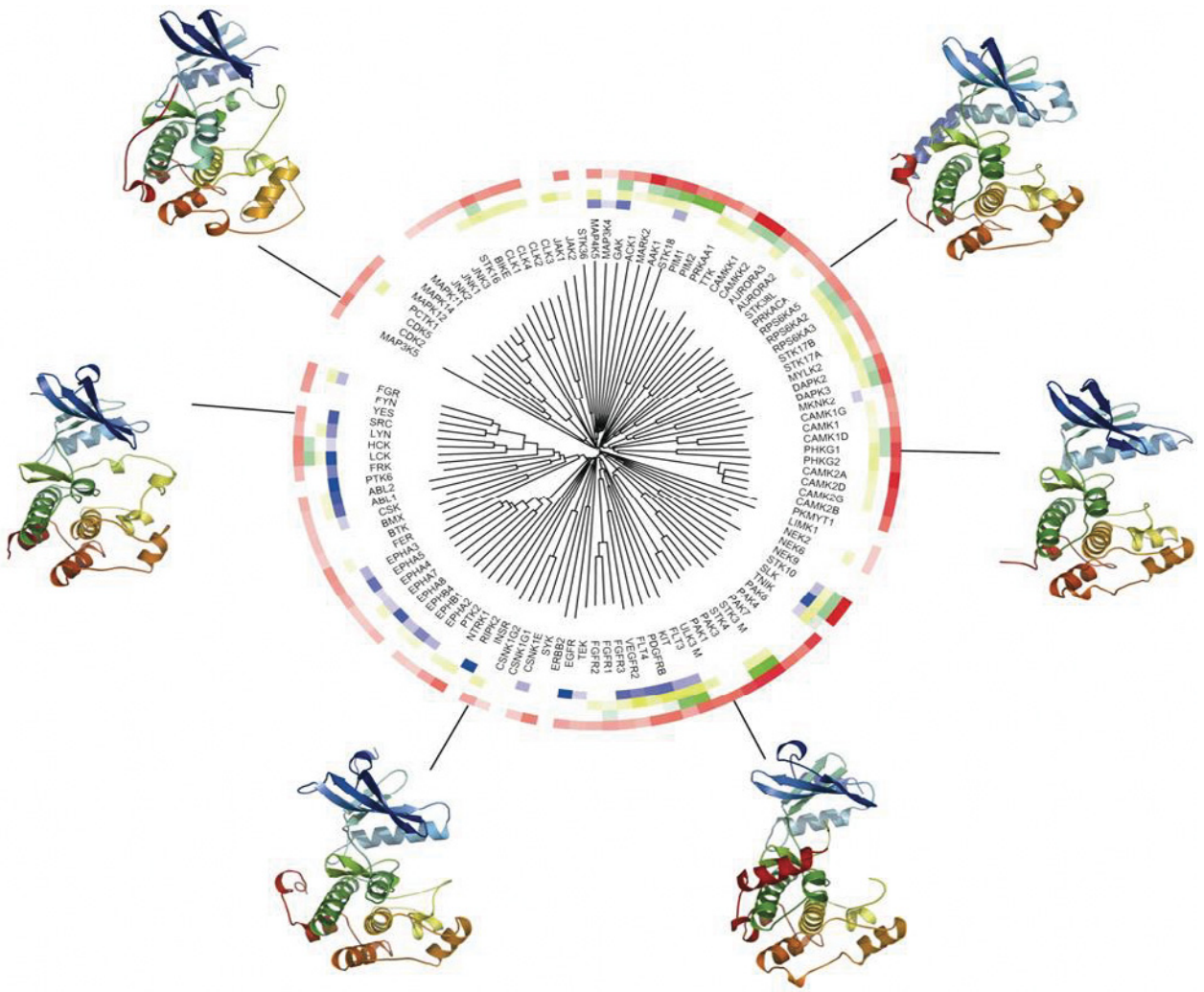
Increasing the number of descriptors captures the residue environments more accurately. However, the combinatorics of environment definitions can result in a large number of environments such that the available alignment data may be partitioned into individual environments that are only sparsely populated. A total of 48 protein–protein interaction-specific ESSTs can be derived using a combination of the four categories of interface accessibility environment, four categories of main-chain conformation and secondary structure (helix, strand, coil and positive φ main-chain torsion angle), two categories of intermolecular hydrogen-bonding (bonded and unbonded) and two categories of intramolecular hydrogen-bonding (bonded and unbonded). The main determinant appears to be the interface accessibility environment (results not shown). As in tertiary interactions, intramolecular hydrogen-bonding status is a further strong determinant (results not shown).

Functional restraints from protein–nucleic acid interactions

The mechanism and nature of protein–nucleic acid recognition is believed to be different from those of protein–protein

Figure 4 | The kinase fold and drug discovery

The binding affinities to promiscuous inhibitors, staurosporine (red), LY-333531 (green), SU11248 (yellow) and ZD-6474 (blue), demonstrates that the similarity in sequence cannot predict the ability to bind inhibitor. The dendrogram is generated by ClustalX [32] multiple alignment of the catalytic domain sequence of kinases, and then the tree is mapped to binding constants of the four inhibitors reported by Fabian et al. [33] using program iTOL [34].



interaction. Hence, the restraints of protein–nucleic acid interaction should be also considered differently. Protein–nucleic acid interfaces, which arise as a consequence of functional restraints, are often conserved and show distinctive amino acid propensities owing to the polar nature of DNA/RNA (Figure 3C).

To describe the differences between amino acid substitution patterns arising from structural restraints and those under functional restraints of nucleic acid binding, new sets of restraints have been incorporated into ESSTs to represent the nature of protein–nucleic acid interactions [26]. Hence, residues involved in intermolecular interactions with nucleic acids are classified further into three types: (i) hydrogen bond; (ii) water-mediated hydrogen bond; and (iii) van der Waals contact. A total of 128 ESSTs were created using four categories of secondary structure (as for tertiary interactions), and two categories of each of solvent accessibility, hydrogen-bonding to nucleic acid, water-mediated hydrogen-bonding to nucleic acid and van der Waals contact to nucleic acid.

By measuring distances among the new sets of ESSTs and constructing phylogenetic trees using the distance matrices, the residues interacting with nucleic acids have shown distinct substitution patterns when compared with the other sites (S. Lee and T.L. Blundell, unpublished work). The new ESSTs were also tested using the sequence–structure homology recognition program, FUGUE [27], to compare the recognition performance with the conventional substitution tables. Significant improvements were achieved in both recognition performance and alignment accuracy using the new substitution tables.

Conclusions

Residues involved in functional interactions with substrates, ligands and multi-component assemblies are clearly more

conserved and have differing amino acid substitution patterns. It is clear therefore that they need to be removed when calculating substitution tables for amino acid residues that are under restraints of tertiary interactions [28]. However, amino acids some distance away can also be under restraint, particularly from the need to maintain precise arrangements of catalytic residues and cofactor-binding interactions. Conversely, amino acid substitutions may affect substrate- and drug-binding specificity, a factor evident in the analysis of kinase drugs (Figure 4) [29].

Acknowledgement

S.L. thanks Juok Cho for statistical analysis.

Funding

T.L.B. thanks the Wellcome Trust for support of our structural studies of protein assemblies. C.L.W. and G.R.J.B. thank the Biotechnology and Biological Sciences Research Council for a studentship. D.T. thanks the Royal Thai Government for funding the study of inhibitor selectivity in protein kinase. S.L. thanks Mogam Science Scholarship Foundation for partial funding for the study of protein–nucleic acid interactions.

References

1 Bajaj, M. and Blundell, T. (1984) Evolution and the tertiary structure of proteins. Annu. Rev. Biophys. Bioeng. **13**, 453–492
2 Orengo, C.A. and Thornton, J.M. (2005) Protein families and their evolution: a structural perspective. Annu. Rev. Biochem. **74**, 867–900
3 Sanger, F. and Tuppy, H. (1951) The amino-acid sequence in the phenylalanyl chain of insulin. 2. The investigation of peptides from enzymic hydrolysates. Biochem. J. **49**, 481–490
4 Sanger, F. and Tuppy, H. (1951) The amino-acid sequence in the phenylalanyl chain of insulin. I. The identification of lower peptides from partial hydrolysates. Biochem. J. **49**, 463–481
5 Sanger, F. (1988) Sequences, sequences, and sequences. Annu. Rev. Biochem. **57**, 1–29
6 Adams, M.J., Blundell, T.L., Dodson, E.J., Dodson, G.G., Vijayan, M., Baker, E.N., Harding, M.M., Hodgkin, D.C., Rimmer, B. and Sheat, S. (1969) Structure of rhombohedral 2 zinc insulin crystals. Nature **224**, 491–495
7 Blundell, T.L., Cutfield, J.F., Cutfield, S.M., Dodson, E.J., Dodson, G.G., Hodgkin, D.C., Mercola, D.A. and Vijayan, M. (1971) Atomic positions in rhombohedral 2-zinc insulin crystals. Nature **231**, 506–511
8 Blundell, T.L., Cutfield, J.F., Cutfield, S.M., Dodson, E.J., Dodson, G.G., Hodgkin, D.C. and Mercola, D.A. (1972) Three-dimensional atomic structure of insulin and its relationship to activity. Diabetes **21**, 492–505
9 Blundell, T.L. and Wood, S.P. (1975) Is the evolution of insulin Darwinian or due to selectively neutral mutation? Nature **257**, 197–203
10 Kimura, M. (1968) Evolutionary rate at the molecular level. Nature **217**, 624–626
11 Blundell, T.L. and Humbel, R.E. (1980) Hormone families: pancreatic hormones and homologous growth factors. Nature **287**, 781–787
12 Slingsby, C., Driessen, H.P., Mahadevan, D., Bax, B. and Blundell, T.L. (1988) Evolutionary and functional relationships between the basic and acidic β-crystallins. Exp. Eye Res. **46**, 375–403
13 Wistow, G., Summers, L. and Blundell, T. (1985) *Myxococcus xanthus* spore coat protein S may have a similar structure to vertebrate lens $\beta\gamma$-crystallins. Nature **315**, 771–773
14 Blundell, T.L., Jenkins, J.A., Sewell, B.T., Pearl, L.H., Cooper, J.B., Tickle, I.J., Veerapandian, B. and Wood, S.P. (1990) X-ray analyses of aspartic proteinases: the three-dimensional structure at 2.1 Å resolution of endothiapepsin. J. Mol. Biol. **211**, 919–941
15 Tang, J., James, M.N., Hsu, I.N., Jenkins, J.A. and Blundell, T.L. (1978) Structural evidence for gene duplication in the evolution of the acid proteases. Nature **271**, 618–621
16 Blundell, T.L., Lapatto, R., Wilderspin, A.F., Hemmings, A.M., Hobart, P.M., Danley, D.E. and Whittle, P.J. (1990) The 3-D structure of HIV-1 proteinase and the design of antiviral agents for the treatment of AIDS. Trends Biochem. Sci. **15**, 425–430
17 Luthy, R., McLachlan, A.D. and Eisenberg, D. (1991) Secondary structure-based profiles: use of structure-conserving scoring tables in searching protein sequence databases for structural similarities. Proteins **10**, 229–239
18 Overington, J., Donnelly, D., Johnson, M.S., Sali, A. and Blundell, T.L. (1992) Environment-specific amino acid substitution tables: tertiary templates and prediction of protein folds. Protein Sci. **1**, 216–226
19 Overington, J., Johnson, M.S., Sali, A. and Blundell, T.L. (1990) Tertiary structural constraints on protein evolutionary diversity: templates, key residues and structure prediction. Proc. Biol. Sci. **241**, 132–145
20 Koshi, J.M. and Goldstein, R.A. (1995) Context-dependent optimal substitution matrices. Protein Eng. **8**, 641–645
21 Levitt, M. (1978) Conformational preferences of amino acids in globular proteins. Biochemistry **17**, 4277–4285
22 Worth, C.L. and Blundell, T.L. (2009) Satisfaction of hydrogen-bonding potential influences the conservation of polar sidechains. Proteins **75**, 413–429
23 Hotelling, H. (1933) Analysis of complex statistical variables into principal components. J. Educ. Psychol. **24**, 417–441
24 Hubbard, T.J. and Blundell, T.L. (1987) Comparison of solvent-inaccessible cores of homologous proteins: definitions useful for protein modelling. Protein Eng. **1**, 159–171
25 Ofran, Y. and Rost, B. (2003) Analysing six types of protein–protein interfaces. J. Mol. Biol. **325**, 377–387
26 Lee, S. and Blundell, T.L. (2009) Ulla: a program for calculating environment-specific amino acid substitution tables. Bioinformatics, doi:10.1093/bioinformatics/btp300
27 Shi, J., Blundell, T.L. and Mizuguchi, K. (2001) FUGUE: sequence-structure homology recognition using environment-specific substitution tables and structure-dependent gap penalties. J. Mol. Biol. **310**, 243–257
28 Gong, S. and Blundell, T.L. (2008) Discarding functional residues from the substitution table improves predictions of active sites within three-dimensional structures. PLoS Comput. Biol. **4**, e1000179
29 Tanramluk, D., Schreyer, A., Pitt, W.R. and Blundell, T.L. (2009) On the origins of enzyme inhibitor selectivity and promiscuity: a case study of protein kinase binding to staurosporine. Chem. Biol. Drug Des. **74**, 16–24
30 Mizuguchi, K., Deane, C.M., Blundell, T.L., Johnson, M.S. and Overington, J.P. (1998) JOY: protein sequence-structure representation and analysis. Bioinformatics **14**, 617–623
31 Lee, S. and Blundell, T.L. (2009) BIPA: a database for protein–nucleic acid interaction in 3D structures. Bioinformatics **25**, 1559–1560
32 Thompson, J.D., Gibson, T.J., Plewniak, F., Jeanmougin, F. and Higgins, D.G. (1997) The CLUSTAL_X windows interface: flexible strategies for multiple sequence alignment aided by quality analysis tools. Nucleic Acids Res. **25**, 4876–4882
33 Fabian, M.A., Biggs, 3rd, W.H., Treiber, D.K., Atteridge, C.E., Azimioara, M.D., Benedetti, M.G., Carter, T.A., Ciceri, P., Edeen, P.T., Floyd, M. et al. (2005) A small molecule-kinase interaction map for clinical kinase inhibitors. Nat. Biotechnol. **23**, 329–336
34 Letunic, I. and Bork, P. (2007) Interactive Tree Of Life (iTOL): an online tool for phylogenetic tree display and annotation. Bioinformatics **23**, 127–128

Received 17 February 2009
doi:10.1042/BST0370727

SUPPLEMENTARY ONLINE DATA

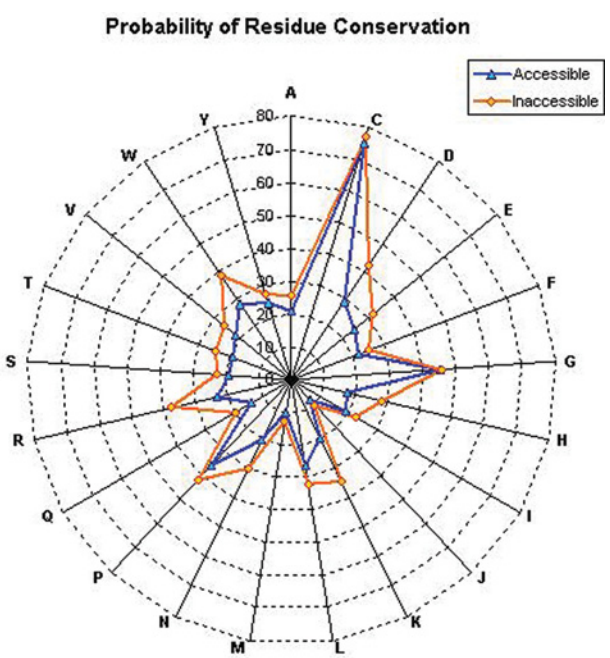
Structural and functional restraints in the evolution of protein families and superfamilies

Sungsam Gong*, Catherine L. Worth*†, Richard Bickerton*‡, Semin Lee*, Duangrudee Tanramluk* and Tom L. Blundell*[1]

*Department of Biochemistry, University of Cambridge, 80 Tennis Court Road, Cambridge CB2 1GA, U.K., †Structural Bioinformatics Group, Leibniz-Institut für Molekulare Pharmakologie, Campus Berlin-Buch, Robert-Rössle-Strasse 10, 13125 Berlin, Germany, and ‡Medicinal Informatics, Division of Biological Chemistry and Drug Discovery, College of Life Sciences, University of Dundee, Dundee DD1 5EH, U.K.

Figure S1 | **Probabilities of residue conservation by solvent accessibility**

The probabilities of residue conservation in the solvent accessible area (blue) are compared with those in the solvent inaccessible region for 21 amino acids. From the 64 ESSTs, the probabilities on the diagonal axis were averaged for each of the two groups: solvent-accessible and -inaccessible. Note that we distinguish cysteine and half-cystine using one-letter codes J and C respectively.



Received 17 February 2009
doi:10.1042/BST0370727

[1]To whom correspondence should be addressed (email tom@cryst.bioc.cam.ac.uk).