ELSEVIER

Biophysical Chemistry 114 (2005) 63–69

Biophysical Chemistry

http://www.elsevier.com/locate/biophyschem

Solvent accessibility in native and isolated domain environments: general features and implications to interface predictability

Mohd Firdaus Raih[a], Shandar Ahmad[b,*], Rong Zheng[c], Rahmah Mohamed[a]

[a]*National Institute for Genomics and Molecular Biology (Interim Laboratory), and School of Biosciences and Biotechnology, Universiti Kebangsaan Malaysia, 43600 UKM Bangi, Malaysia*

[b]*Department of Biochemical Science and Engineering, Kyushu Institute of Technology, Iizuka 820 8502, Fukuoka, Japan*

[c]*Weill Medical College of Cornell University, New York, NY 10021, USA*

Received 20 August 2004; received in revised form 20 October 2004; accepted 26 October 2004
Available online 23 November 2004

Abstract

A non-redundant database of 4536 structural domains, comprising more than 790,000 residues, has been used for the calculation of their solvent accessibility in the native protein environment and then in the isolated domain environment. Nearly 140,000 (18%) residues showed a change in accessible surface area in the above two conditions. General features of this change under these two circumstances have been pointed out. Propensities of these interfacing amino acid residues have been calculated and their variation for different secondary structure types has been analyzed. Actual amount of surface area lost by different secondary structures is higher in the case of helix and strands compared to coil and other conformations. Overall change in surface area in hydrophobic and uncharged residues is higher than that in charged residues. An attempt has been made to know the predictability of interface residues from sequence environments. This analysis and prediction results have significant implications towards determining interacting residues in proteins and for the prediction of protein–protein, protein–ligand, protein–DNA and similar interactions.

Keywords: Domain–domain interactions; Interface area; Solvent accessibility; Neural network; Structure classification

1. Introduction

Importance of protein–protein and domain–domain interactions has been widely recognized and several efforts to their understanding have been reported [1–3]. Pairing preference of amino acid residues at protein interfaces [4] and features of the domain–domain interactions [5] are particularly interesting studies aimed at characterization of protein–protein interfaces. Attempts have also been made at predicting the interface residues [6,7] and classification of interfaces [8]. It has also been reported that the intra-chain domain interfaces have very similar characteristics to oligomeric interfaces [5]. Although some of these studies did compare domain–domain interactions and chain–chain interactions, most of the treatment was made separately. Furthermore, the role of water and other heterogeneous atoms has not been included in these studies. Here we try to extract general features of solvent accessibility of amino acid residues in the interfacing regions of protein domains. A broad definition of interface, inclusive of interfaces with water, ligands, heterogeneous atoms, intra- and inter-chain domain interfaces, has been used. The idea is that any region in the protein structure may be considered to have "unsaturated bonds", which will be satisfied by either the long-range contacts with intra-chain domain or by the atoms coming from inter-chain domain, ligands, other protein or DNA chains. Thus, we characterize interfacing of residues by calculating their accessible surface area (ASA) after isolating their domains from the rest of the protein and comparing them with the ASA in the native protein (original PDB coordinates). Statistics of residue propensities to be on the interface have been collected before, in some of the above-mentioned references. However, our present study on interface ASA differs in two ways. First, here the definition of

* Corresponding author. Tel./fax: +81 948 297841.
E-mail address: shandar@bse.kyutech.ac.jp (S. Ahmad).

0301-4622/$ - see front matter © 2004 Elsevier B.V. All rights reserved.
doi:10.1016/j.bpc.2004.10.005

the interface is much more general and hence the statistics represents general nature of "unsaturated" bonds in the residues rather than restricting it to protein–protein or domain–domain interfaces. Secondly, previous works have simply characterized residues to be on the interface or otherwise if the ASA change was anything more than zero. The actual overlap or the net change in the ASA due to interfacing was not studied. Here we obtain the pre- and post-interface ASAs of residues and compare the actual loss in the ASA for these residues. This feature gives additional information about the degree of overlap and is shown not necessarily to reflect propensity. Further, we characterize ASA change in the post- and pre-interfacing conditions separately for different secondary structure types, trying to find if there is any significant difference in residue propensities to be on the interface by virtue of their secondary structure. A large data set of more than 790,000 residues (including 140,000 in the interface) has been used to extract these features of interfacing patterns. If the information about these "unsaturated bonds" is present in the local sequence of the protein, it should be possible to predict such interfacing residues without looking at the complementary parts. An attempt has therefore been made to predict these interfacing residues from sequence. This approach results in moderate success but opens the way for further exploration of predictability of interfacing residues from other sets of descriptors, perhaps coming from the structural properties of the residue environment.

2. Materials and methods

2.1. Data sets

The domain information about proteins has been taken from SCOP [9]. PDB style coordinates for isolated domains have been taken from ASTRAL database [10]. Version 1.63 of SCOP lists consists of 5226 domains from 3332 collected proteins such that no two domain sequences have more than 40% sequence identity [9]. Redundancy was further removed by clustering this sequence data at 25% identity threshold using blastclust [11]. This resulted in 4536 domains, which were used in this work. It may be noted that some of these domains are identical to their native environments. However, the isolated domain environments are free from the heterogeneous and other non-amino acid atoms, and therefore may have implications to the final interpretation of results.

2.2. Calculation of ASA and secondary structure

Solvent accessibility or accessible surface area (ASA) and secondary structure were calculated using DSSP program [12]. For a comparison between the native and domain ASA values, ratio of the ASA in the native PDB coordinate file to that in the ASTRAL database (with only isolated domains) in percentage terms were calculated. This value of the native ASA relative to the isolated domain ASA is termed as post-interface ASA of the residue. Residues in the ASTRAL data have been mapped to their corresponding residues in the PDB file by first calculating the ASA, as such and then tagging them for the residue number and chain names. Tagged residues were then collected in columnar data of ASA values, residue number, chain name and secondary structure. Relative post-interface Area (PIA) has been defined as the ratio of the ASA of the same residue in the native state to that in the isolated domain state (in percentage terms). Isolated domain here refers to the coordinate data in the ASTRAL data sets.

$$\mathrm{PIA} = 100 \times (\text{Relative native ASA of a domain}) \,/(\text{Relative ASA in isolated domain})$$

Difference in the native and isolated domain states has been demonstrated for an example domain in Fig. 1(a) and (b).

2.3. Interfacing or unsaturated residue

In this work, an interfacing residue has been defined as any residue whose accessible surface area in the native

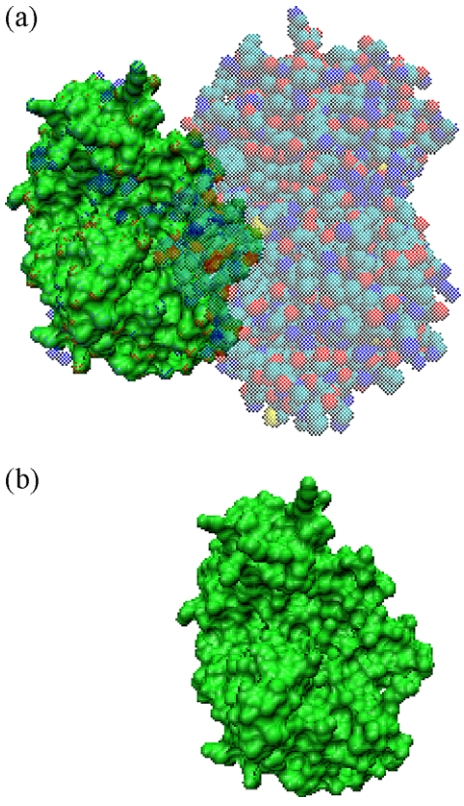

Fig. 1. A protein domain in its (a) native environment (name: GTP-specific succinyl-COA synthetase, PDB code: 1EUC) and (b) isolated state (second domain of chain B, SCOP domain code: d1eucb2). Domain atoms from residues with unsaturated bonds interact with several heterogenous atoms, water and other domains to saturate them, resulting in a reduced accessible surface area.

protein is at least 1 Å less than in its isolated domain environment. This interfacing therefore includes interaction with other intra- or inter-chain domains and other heterogeneous atoms, including water molecules. This may be noted that interface definition in some other works is somewhat different from this one. For example, Ofran and Rost [8] used a definition of contact between residues such that any atom from the first residue should be within 6 Å distance from any atom of the other residue. However, definitions such as ours, in terms of change in the solvent accessibility, are also widely used [5]. Since we are dealing here with the change in solvent accessibility in particular, a domain interface defined exclusively in these terms is more useful than other methods.

2.4. Propensity

Propensity of a residue type (e.g. alanine) is defined as the ratio of its relative frequency/composition in the interface relative to that in the surface, rescaled by the overall propensity. Thus, propensity (P_i) of a residue type i is:

$$\text{Propensity} = p_i/p_{\text{average}} \tag{2}$$

where

$$p_i = \text{FI}_i/\text{FS}_i \tag{3}$$

FI_i=relative composition of residue i in the interface=(the frequency of residue i in interface)/(total number of

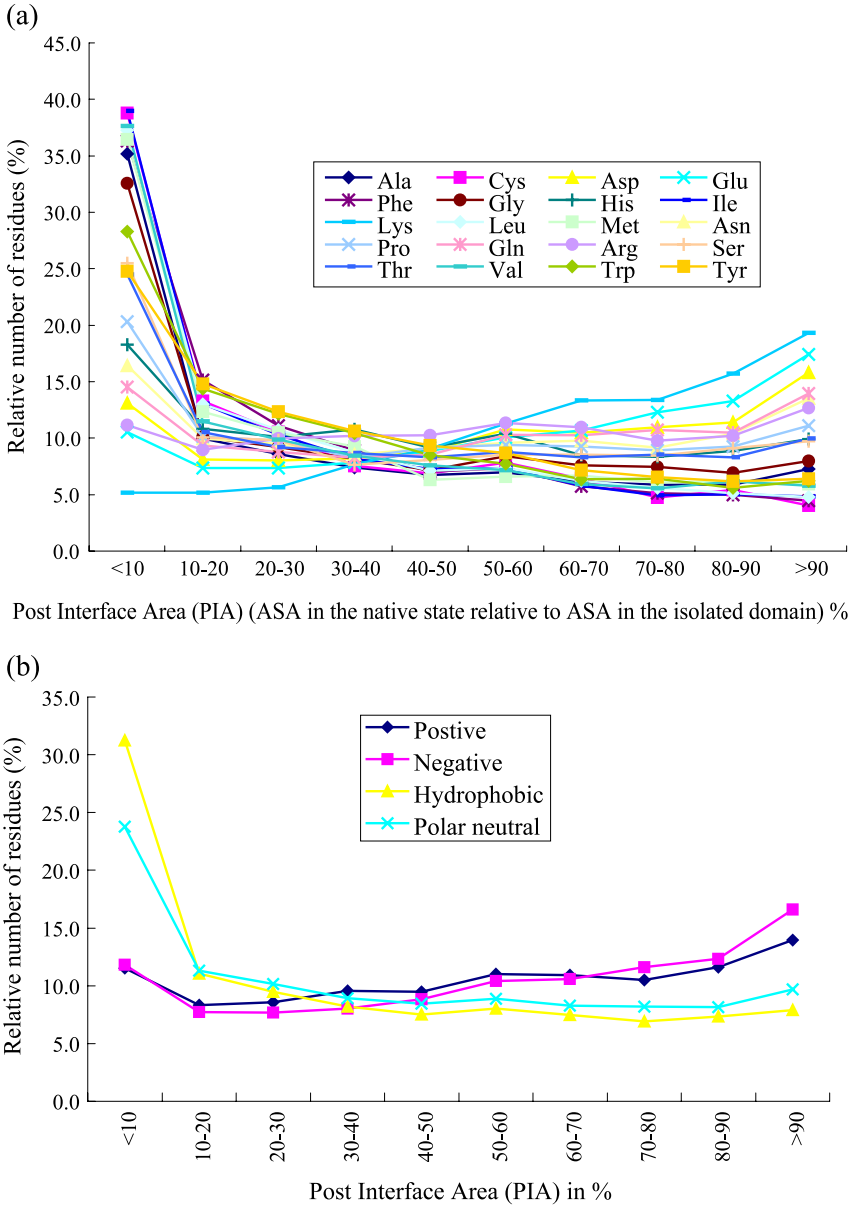

Fig. 2. Frequency of amino acid residues in different ranges of post-interface accessible surface area (PIA). PIA is defined as the ASA of the residue in the native state (post-interface area) relative to the ASA in the isolated domain environment (pre-interface area). (a) PIA data for the 20 amino acid residue types. (b) PIA values with residues grouped by their side chain properties.

interface residues); FS_i=relative composition of residue i in the surface=(frequency of residue i on surface)/(total number ofsurface residues); $p_{average}$ is the mean of 20 values of p_i's corresponding to each residue type.

For these calculations, a residue is regarded to be on surface if it has a non-zero accessible surface area. All propensity values are calculated separately for each domain and averaged at the end to obtain the final values. Propensities for secondary structures are also calculated in the same way except that the residue type is replaced by secondary structure type.

3. Results and discussion

3.1. Post-interface accessible area of amino acids

Fig. 2(a) shows the results of post-interface area (PIA) of amino acid residues relative to the area in an isolated domain. Higher value of PIA means smaller change in the interface and thereby smaller overlap. Fig. 2(b) shows the same results by grouping the 20 amino acid types into their well-known categories. It is very clearly observed that negative and positive charged residues show a similar pattern of retaining more of their ASAs compared with the hydrophobic and neutral counterparts. For example, just about 12% charged residues lose as high as 90% of their ASAs, compared with 32% of the hydrophobic and 24% of the polar residues undergoing such a large change. This highlights the fact that fewer atoms from a charged residue will participate in the interaction with the residues on a different domain.

Greater loss of ASA in hydrophobic residues, at first sight leads to the impression that the interfaces are being formed to avoid contact with water and that greater overlap will help in avoiding more molecules of water. However, a similar behavior by the polar uncharged residues stands against such a conclusion. More likely, it is the charged residues which stand out of the lot and show more targeted interactions. We attribute this to the fact that charged residues may form strong bonds with their complementary domain residues, by disposing fewer of their atoms and thereby losing less ASA. Requirement of retaining most of the solvent accessibility by the charged residues may also be imposed by biological functions that need to be performed by these residues, in addition to forming the interfaces. Hydrophobic residues on the other hand occur less frequently in the active sites and hence seem to be more important for the stability of the domain–domain complexes—a role similar to the role of these residues in protein stability by way of core formation [15].

3.2. Mean post-interface area and propensity

Propensities may be defined in many ways differing, for example, in the way they are normalized by number of surface residues and the fact that there is no unique cutoff of ASA which may be used for defining a surface residue. Different definitions of propensity lead to different statistics. Our definition of propensity is simply the ratio of amino-acid compositions in the interface to that in the surface, where surface residue is any residue with non-zero ASA. Fig. 3 shows mean post-interface area (PIA) of the 20 amino acid residues and compares them with the propensity of these residues to be in the interface. As described in Materials and methods, residues with no preference to be on the interface will have 100% propensities according to the definition used here. Interesting aspect of this analysis is that the mean overlap or PIA of residues does not necessarily correspond exactly to their propensity. Results of propensity reported here are in broad agreement with those by other investigators [5]. However, results of PIA have never been reported and here we try to investigate for the first time how the propensity values of residues may or may not correspond to their PIAs. Notable among these results is the fact that only one of the positively charged residues Arg has a high propensity whereas the other such residue, Lys, does not show a similar high propensity.

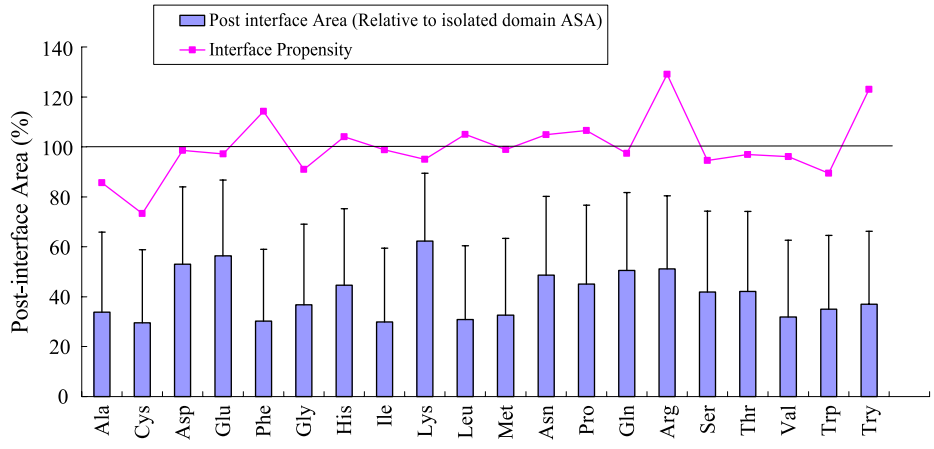

Fig. 3. Mean post-interface area (PIA) for all 20 amino acid residue types. Vertical bars represent standard deviations. Interface propensity of residues is also plotted together.

Higher propensity of Arg signifies their role in providing domain–domain, domain–ligand and chain–chain interactions. Lys, although similarly charged, has been implicated in cation–pi interactions which are required to stabilize the structure of protein itself [13]. In fact, Lys, together with other charged residues, is also among those whose most of the ASA is preserved upon interfacing.

3.3. Interface propensity and secondary structure

We tried to see if certain secondary structures have a preference to be on the interface. We had previously found that DNA binding does not significantly favor any particular secondary structure [14]. However, the overall domain–domain interactions here do seem to be biased by their secondary structure. Fig. 4(a) shows the interface propensity of the secondary structure types, defined by DSSP [12]. As seen from the figure, coil seems to be the most favored conformation for the interfacing residues, whereas the isolated bridge conformation has a negative preference to be on the interface. Bend conformation has a significantly higher presence in the interface than other regular secondary structures. This observation may be attributed to the fact that bend conformation may actually be the effect of domain

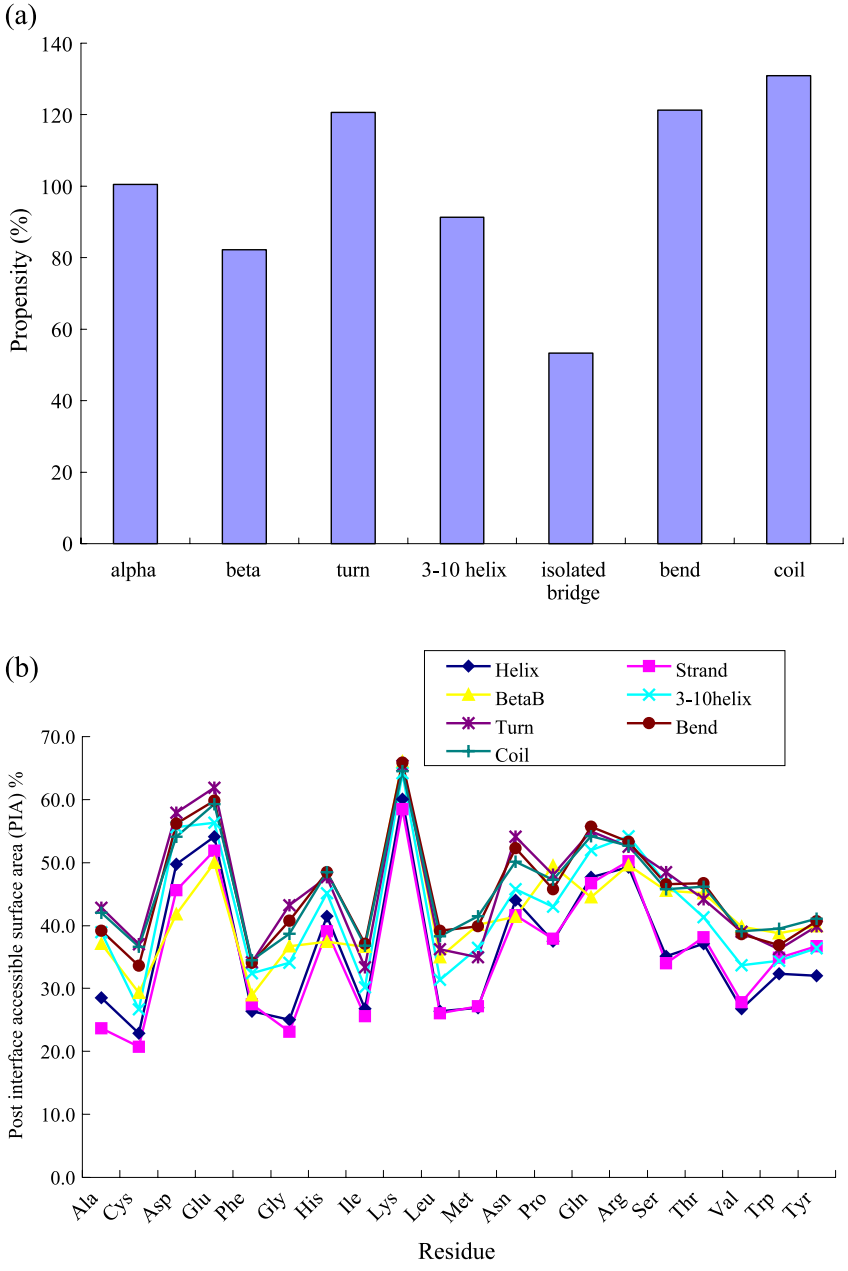

Fig. 4. Interface propensity of different secondary structures types. (a) Coil has a much higher propensity than other secondary structure types. (b) Secondary structure propensity values do not seem to be sensitive to residue type, as all secondary structures show similar pattern in their residue-wise propensity. Pi conformation has too few residues and hence not plotted.

interactions with the external environment rather than being the cause of it or because propensity values are exaggerated by under-representation of these structures in the surface. Water molecules, for example, are known to cause an increase in the bend conformations and the current statistics probably just represents that fact. Similarly, there may be a loss of secondary structure by way of coiling in order to provide necessary conformational changes for interactions.

Looking at Fig. 4(b), results of interfacing seem to be slightly different from Fig. 4(a). It seems more evident that the helical and strand conformations lose more ASA upon interfacing compared to their less frequent counterparts such as 3–10 helix, bend and even coil conformations.

3.4. Unsaturated bonds and prediction of interface residues

In view of the above observations, it is hoped that positive and negative propensities of residues should make it possible to predict an unsaturated bond on a given domain and hence candidates of interface residues. This predictability will improve significantly, if the domain-interacting residues are located close to each other on the sequence, in which case, we can simply give the local sequence environments as the inputs to a prediction model. However, it was observed that the cases of residues forming interfaces in successive positions and being conserved at these positions are not overwhelmingly frequent. Forty percent of all such interfacing residues are isolated singlets, meaning that their C- and N-terminal neighbors are not on the interface. Twenty-six percent of all such environments occur as doublets, 10% as triplets and less than 1% each has more than nine successive residues on the interface. We tried to build a prediction model to assess the predictability of these interfacing residues or the residues with unsaturated bonds, ready to interact with external atoms (data not shown). It was noted that the information about immediate neighbor on the C- and N-terminal sides significantly improved the prediction, but the sequence neighbors beyond the first one do not significant improve results (see next section). It was also observed that predictions were better if the interacting residues occurred successively close on a sequence. Predictability of such interfacing residues was higher than what is expected from a simple propensity analysis, suggesting that the residue neighbors also play a role in determining their interfacing capacity. The prediction quality, although not being extraordinarily high, anyway opens a way for the prediction of unsaturated residues, likely to be improved by including spatial information, when available.

3.5. Residue environments likely forming unsaturated bonds

As mentioned above, the first neighbor information on either side of the residue was found to contribute significantly in the prediction of interfacing residues and subsequent residues seem to be not so significant. This suggests that the residues have an unsaturated bond—making them prone for interfacing—because the C- and N-terminal neighbors of this residue do not entirely satisfy its bonds. Since second and subsequent neighbors have little interaction with the residue anyway, they do not determine the degree of saturation and hence make little difference in prediction. A natural role would be expected for the residues taking part in the formation of secondary structure, e.g. fourth neighbor in alpha helix. However, interface prediction was not improved if the information of distant neighbors was included. This is obviously due to the fact that interfacing due to unsaturated bonds on protein sequences does not proceed by way of hydrogen bonds-at least in a large proportion of interfacing residues. That is to say that the interactions of domain atoms with those on other domains or ligands are not caused by hydrogen bond. Van der Waals, dipole–dipole, partial ionic and covalent bonds therefore should account for the degree of saturation on the main chain residues. To examine which residue neighbors leave them with unsaturated bonds, we calculated the over-representation of C- and N-terminal environments (immediate neighbors) for the 20 amino acid residues. Table 1 gives some of the residue environments whose frequency of occurrence in the interfacing and non-interfacing regions is significantly different from each other. It is noticed that residue environments such as Trp negatively affect the propensity of residues to be in the interface, except in some cases, e.g. WVF, when both neighbors of Val are aromatic. Thus, a large number of clear preferences seem to have been made by nature in terms of sequence arrangement in order to leave some sites unsaturated for interactions with other residues and atoms.

Table 1
Most significant residue environments which enhance interfacing probability of residues: the three residue fragments in the first column are the residue environments of the central residue in the C- and N-terminals for which the frequency in the interfacing regions is significantly higher (at least five times) that in the non-interfacing region

Environments (of the central residue) favoring interface	Environments (of the central residue) avoiding interface
TCE, KCI, NCV, WPD, DDH, WVI, PCK, PCR, QAK, FGC, AKC, WVF, RFW, KQC, DQW, DCK	CWA, WMN, NCW, CYH, NWC, WTM, CCY, CWE, GCW, HYW, MFC, MYM, QMC, WHW, WQM, CHF

In the second column, some of the patterns whose frequency in the interfacing region is significantly lower than in the interfacing regions (at least one fifth) are collected.

4. Conclusion

Solvent accessibility of amino acid residues in their isolated domains and native environment provides useful information about unsaturated bonds in them. Propensity of residues to form interfaces with external atoms does not necessarily correlate with the mean post-interface exposed

area. Charged residues have a much smaller overlap than their hydrophobic and polar counterparts, a property not necessarily shared by propensity values. Certain residue environments enhance propensity of residues to be in the interface whereas some others reduce it. In general, results of this study have useful applications to the prediction of protein–protein, protein–ligand and protein–DNA (including water-mediated) interactions.

References

[1] I.M. Nooren, J.M. Thornton, Structural characterization and functional significance of transient protein–protein interactions, J. Mol. Biol. 325 (5) (2003) 991–1018.
[2] A.I. Archakov, V.M. Govorun, A.V Dubanov, Y.D. Ivanov, A.V. Veselovsky, P. Lewi, P. Janssen, Protein–protein interactions as a target for drugs in proteomics, Proteomics 3 (4) (2003) 380–391.
[3] Y. Gao, R. Wang, L. Lai, Structure-based method for analyzing protein–protein interfaces, J. Mol. Model (2003 (Nov.)) 22 (online).
[4] F. Glaser, D.M. Steinberg, I.A. Vakser, N. Ben-Tal, Residue frequencies and pairing preferences at protein–protein interfaces, Proteins 43 (2) (2001) 89–102.
[5] S. Jones, A. Marin, J.M. Thornton, Protein–domain interfaces: characterization and comparison with oligomeric protein interfaces, Protein Eng. 13 (2) (2000) 77–82.
[6] S. Jones, J.M. Thornton, Prediction of protein–protein interaction sites using patch analysis, J. Mol. Biol. 272 (1997) 133–143.
[7] H.X. Zhou, Y. Shan, Prediction of protein interaction sites from sequence profile and residue neighbour list, Proteins 44 (3) (2001) 336–343.
[8] Y. Ofran, B. Rost, Analysing six types of protein–protein interfaces, J. Mol. Biol. 325 (2003) 377–387.
[9] A.G. Murzin, S.E. Brenner, T. Hubbard, C. Chothia, SCOP: a structural classification of proteins database for the investigation of sequences and structures, J. Mol. Biol. 247 (1995) 536–540.
[10] J.M. Chandonia, G. Hon, N.S. Walker, L. Lo Conte, P. Koehl, M. Levitt, S.E. Brenner, The ASTRAL compendium in 2004, Nucleic Acids Res. 32 (2004) D189–D192.
[11] S.F. Altschul, W. Gish, W. Miller, E.W. Myers, D.J. Lipman, Basic local alignment search tool, J. Mol. Biol. 215 (1990) 403–410.
[12] W. Kabsch, C. Sander, Dictionary of protein secondary structure: pattern recognition of hydrogen-bond and geometrical features, Biopolymers 22 (1983) 2577–2637.
[13] M.M. Gromiha, C. Santhosh, S. Ahmad, Structural analysis of cation-interactions in DNA binding proteins, Int. J. Biol. Macromol. 34 (3) (2004) 203–211.
[14] S. Ahmad, M.M. Gromiha, A. Sarai, Analysis and prediction of DNA-binding proteins and their binding residues based on composition, sequence and structural information, Bioinformatics 20 (2004) 477–486.
[15] N. Kannan, S. Vishveshawara, Aromatic clusters: a determinant of thermal stability of thermphilic proteins, Protein Eng. 13 (2000) 753–761.