J Mol Model (2007) 13:1157–1167
DOI 10.1007/s00894-007-0237-0

ORIGINAL PAPER

Statistical analysis of physical-chemical properties and prediction of protein-protein interfaces

Surendra S. Negi · Werner Braun

Received: 11 May 2007 / Accepted: 30 July 2007 / Published online: 9 September 2007

Abstract We have developed a fully automated method, InterProSurf, to predict interacting amino acid residues on protein surfaces of monomeric 3D structures. Potential interacting residues are predicted based on solvent accessible surface areas, a new scale for interface propensities, and a cluster algorithm to locate surface exposed areas with high interface propensities. Previous studies have shown the importance of hydrophobic residues and specific charge distribution as characteristics for interfaces. Here we show differences in interface and surface regions of all physical chemical properties of residues as represented by five quantitative descriptors. In the current study a set of 72 protein complexes with known 3D structures were analyzed to obtain interface propensities of residues, and to find differences in the distribution of five quantitative descriptors for amino acid residues. We also investigated spatial pair correlations of solvent accessible residues in interface and surface areas, and compared log-odds ratios for interface and surface areas. A new scoring method to predict potential functional sites on the protein surface was developed and tested for a new dataset of 21 protein complexes, which were not included in the original training dataset. Empirically we found that the algorithm achieves a good balance in the accuracy of precision and sensitivity by selecting the top eight highest scoring clusters as interface regions. The performance of the method is illustrated for a dimeric ATPase of the hyperthermophile, *Methanococcus jannaschii*, and the capsid protein of Human Hepatitis B virus. An automated version of the method can be accessed from our web server at http://curie.utmb.edu/prosurf.html.

Keywords Hot spots · Molecular recognition · Physical chemical properties of interface residues · Protein-protein interface

Electronic supplementary material The online version of this article (doi:10.1007/s00894-007-0237-0) contains supplementary material, which is available to authorized users.

S. S. Negi · W. Braun (✉)
Department of Biochemistry and Molecular Biology,
Sealy Center for Structural Biology and Molecular Biophysics,
University of Texas Medical Branch,
301 University Blvd,
Galveston, TX 77555-0857, USA
e-mail: webraun@utmb.edu

Introduction

Protein-protein interactions play an important role in many biological processes in the cell, e.g., formation of active sites of oligomeric enzymes and maintenance of their effective conformation, regulatory processes including signal transduction, electron transport systems, DNA synthesis, antibody antigen interaction, formation of inter cellular structures [1, 2]. In order to understand how proteins recognize their partner or how these interactions build molecular complexes, it is important to examine the role of amino acid residues present in the protein interface. A protein interface consists of 6–30% of the monomer surface area that vary from 500–5000 Å^2 and the average value of the contact surface area in a monomer is about 800 Å^2 [3, 4]. Earlier studies have shown that the hydrophobic interaction between amino acid residues plays a major role in binding of the protein interfaces [5]. These residues were found to form small patches on the protein surface that includes both polar and charges residues [6, 7]. The hydrogen bonding and the pairing of the polar residues also plays a significant role in binding the protein interfaces but this complementary nature of electrostatic interaction varies from protein to protein interfaces [8].

A number of databases are available today to study the structural basis of protein-protein interactions [9–12], however, the basic problem to characterize intrinsic properties of interfaces which distinguish them from other surface

areas has not yet been solved. Alanine scanning of protein-protein interfaces has shown that the free energy is not uniformly distributed in protein interfaces and only few residues contribute to the bulk of the binding energy called hot spot residues [13–16]. ASE-db [9] is a database for alanine mutation that provides information about hot-spots of amino acid residues on the protein surface. These residues are found to be clustered at the center of the interface and surrounded by a small subset of residues. It has been found that a single residue can contribute a large fraction of binding free energy despite the large binding interface [17]. We show how propensities for hot spots are correlated to a large extent with our propensities for interface residues.

Early statistical studies [7, 18–20] showed that interfaces have few statistically significant differential characteristics, yet some trends were observed. More recent methods tried to implement prediction methods for interaction surfaces including additional information from evolutionary information, such as profile methods or correlated mutations. These methods include patch analysis [21–23], clustering methods [24, 25], computational alanine scanning [26–28], prediction from sequence profile [29], prediction based on machine learning algorithm [22, 30–32], hydrophobic moment [33], structure based method [34, 35] and the phylogenetic information [36, 37]. In addition, shape and size of the protein also play an important role in deciding the functional sites on the protein surface [35, 38–40] and few prediction methods are publicly accessible as web servers [23, 41–44].

Based on a statistical analysis of 72 known protein complexes we have developed two new methods to predict potential interface regions on the surface of a monomeric protein. The two methods, a patch analysis and a cluster method, locate regions on the surface of monomers with a high proportion of residues with high interface propensities, but differ in the computational techniques to decompose the monomeric protein surface. In the patch analysis, a patch of radius R was drawn around the central surface exposed residue. In the cluster method, the entire protein surface was partitioned into *n* clusters. Score functions were developed to rank clusters or patches according to their preferences to be in an interface. The number of high ranking clusters or patches were empirically determined to obtain a good balance between sensitivity and precision. Our InterProSurf method was tested for a training set of 72 protein complexes as well as a test data set of 21 protein complexes. This accuracy of the method is in the same range as recently published methods [22, 23, 41, 42].

In addition we identified differences in physical chemical properties of residues present in interfaces as compared to other surface residues subunits. The discriminating properties of interface and surface regions are quantified by five physical-chemical descriptors developed previously by our group [45]. We investigated spatial pair correlations of these descriptors in interface and surface areas, and compared log-odds ratios for interface and surface areas. These observations might help to understand the physical-chemical nature of interfaces and to improve prediction algorithms for hot-spots and for potential binding sites on the protein surface. Our prediction method InterProSurf was already successfully used to design entry sensitive mutants of the E1 envelope protein of the Venezuelan Equine Encephalitis Virus [46], and an automated version of the method can be accessed from our web server at http://curie.utmb.edu/prosurf.html.

Material and methods

Propensity scale

A final set of 72 protein complexes was selected from the protein data bank to derive propensity values for residues being in interfaces or on the surfaces of complexes. A different set of 21 different complexes was used to assess the accuracy of our prediction method (Table 1a,b). These sets do not contain redundant protein complexes as protein complexes with similar sequences (>37% sequence identity) were discarded using psiblast [47] and clustalw [48] se-

Table 1 List of protein complexes

a) List of PDB id's used deriving the propensity of amino acid residues at the protein interface and on the surface

1a2d	1a2y	1a3r	1a4y	1a6t	1a6v	1a8j	1ad9	1agr
1ahw	1bab	1bdj	1bj3	1brs	1bxi	1c8o	1cbw	1cdk
1cee	1cho	1cjt	1cly	1cn3	1cse	1cz7	1dan	1d3b
1dfj	1dgr	1ds8	1e6t	1ee4	1efu	1epb	1fc2	1fdl
1gg2	1hhj	1hvi	1jck	2jel	1jhl	1lya	1mmo	1mnm
1msb	1nca	1pyt	1reg	1rhi	1scu	1seb	1smp	1stf
1tab	1tgs	1udi	1xso	2bbk	2gst	2mev	2mta	2rcs
2scp	2sic	2vis	3hhr	4aah	4mdh	4ts1	7fab	8ruc

b) In addition to above 72 protein complexes, performance of the prediction algorithm was also tested for following set of 21 PDB id's

1abr	1bun	1fcd	1htt	1zbd	1aoh	1cmx	1fin	1jtd
1apy	1eg9	1frv	1pdk	1bmq	1emv	1fug	1pvd	1bpl
1f60	1g7k	1rrp						

quence alignment. Interface residues in these complexes were identified by the change in the solvent accessibility area (δASA) of the residues in the complex and in the monomeric form. The surface areas of the amino acid residues in the protein complex were compared to those of the monomeric form using the GetArea program [49] with a probe radius of 1.4 Å. We considered residues as being buried in a structure, complex or monomeric form, as residues where the ratio of the side-chain surface area to the "random coil" value per residue is less than 20%. The remaining residues in the complexes were then separated into surface and interface residues depending on the absolute value of the change in the solvent accessibility area, δASA. If δASA is larger than 15 $Å^2$ then the residue was classified as an interface residue, otherwise the residue was considered as a surface residue in the complex. The random coil value of a residue *X* in the tripeptide *Gly-X-Gly* was calculated as the average value in an ensemble of thirty random conformations [49].

The propensity of an amino acid residue being in a protein interface ($\boldsymbol{P_{interface}}$) and on the protein surface ($\boldsymbol{P_{surface}}$) was calculated by using the following two equations:

$$P_{interface}=\frac{\frac{n_i}{\sum n_i}}{\frac{N_i}{N}} \tag{1}$$

$$P_{surface}=\frac{\frac{s_i}{\sum s_i}}{\frac{N_i}{N}} \tag{2}$$

Here n_i , i=1, 2 ...20, is the number of residues of type i at the interface and s_i is the number of residues of type i in the surface. Σn_i is the total number of residues at the interface and Σs_i is the total number of residues in the surface. N_i is the total number of residue of type i at the protein interface and on the protein surface; N is the total number of the residue at the protein interface and on the surface, respectively. Interface propensity greater than one indicates that the residue is more frequent at the protein interface while surface propensity greater than one indicates that residue is more frequent on the protein surface.

The interface residues were identified by calculating the distance between each atom of amino acid residues in different chains of the protein complex. Two atoms across the protein interface were assumed to be interacting with each other, if the cartesian distance between them was less then the sum of their van der Waals radii plus a constant of 1 Å, *i.e.*,

$$dist\left(r_i, r_j\right) \le r(vwd)_i + r(vwd)_j + 1 \tag{3}$$

Where $dist(r_i, r_j)$ is the distance between atom r_i and r_j and $r(vwd)_i$ and $r(vwd)_j$ are van der Waals radii of the atom r_i and r_j, respectively. Once the interface residues on the protein surface were identified, then the environment around the interface and surface residues can be determined by defining a sphere of radius 5 Å around the interface and surface residues. Any residue inside this sphere was assumed to be interacting with its neighbors and a frequency table was developed for amino acid residues interacting with each other with in this distance range.

Prediction algorithm

(A) Clustering method

To identify regions on the protein surface with many residues of high interface propensities, we first decompose the set of surface residues in clusters of spatially related residues [25, 50, 51]. All amino acid residues of the protein surface are represented by their C_β atom ($C\alpha$ atom in the case of the *Gly* residue). These residues represent a set of points in the three dimensional space. This new three dimensional space is then partitioned into n clusters (Ω_n) by a clustering method that is frequently used in data compression techniques and known as Linde, Buzo, and Gray (LBG) algorithm [52–57]. Each cluster contains a certain number of surface residues near in space and is represented by the centroid c_n of its cluster.

The LBG algorithm partitions the input space of k m-dimensional vectors into n non-overlapping regions (Ω_n), such that each vector of a region Ω_n is nearest to its centroid (c_n) and the average squared distance of all input vectors to their centroid is minimal. Each vector belongs to a particular region Ω_n, which is represented by its centroid c_n. In this way the whole protein surface is partitioned into n clusters where each cluster of the protein surface is represented by the centroid of the cluster. The boundary of each region Ω_n is defined such that each vector in Ω_n is nearest to its own centroid c_n:

$$\Omega_n = \left\{x : d(x, c_n) \le d\left(x, c_{n'}\right) \forall n' \ne n\right\} \tag{4}$$

If $|\Omega_n|$ is the total number of elements in the encoding region (Ω_n), then the centroid position is given by Eq. (5) [53, 54]

$$c_n = \frac{\sum\limits_{x_k \in \Omega_n} x_k}{|\Omega_n|} \tag{5}$$

The LBG algorithm build the codebook vector in an iterative procedure and guarantees that the distortion $d(x,c_n)$ from one-iteration to next will not increase [53]. Empirically we found that a fixed number of clusters (n=32) gives a satisfactory clustering of the protein surface.

(B) Patch analysis

In patch analysis, the interface residues are predicted by defining a spherical patch around each surface exposed residue in the protein. The amino acid residues are represented by their C_β atom ($C\alpha$ atom for *Gly* residue). A surface patch is defined as the central surface solvent exposed residue surrounded by n-nearest neighbors within a sphere of radius R in the unbound protein. Therefore for a protein having *n* surface exposed amino acid residues, we have *n* surface patches. Patch sizes were varied from 8 to15 Å and score of each patch was calculated by Eq. (6). A patch is predicted to be part of the protein interface if its score is greater than χ% of the maximum patch scored on the protein surface. This hypothesis was tested for χ=88, 89...94. Once a high scoring surface patch is predicted, the residues in the patch are counted and the predicted residues in each patch are compared with the actual amino acid residues present in the protein interface.

Scoring function

Each cluster or patch of surface residues is evaluated by a scoring function to find surface regions of a protein with many residues of high interface propensities. We used the average propensity of a cluster or a patch as a scoring function. The average values, *Score*, are calculated with weighting factors proportional to the solvent accessible surface area (ASA_i), as both, propensity and solvent accessible surface area of an amino acid residue at the protein interface are influencing the protein interaction:

$$Score = \frac{\sum_{i\in\Omega_n} p_i ASA_i}{\sum_{i\in\Omega_n} ASA_i} \quad (6)$$

Here p_i is the interface propensity ($P_{interface}$) or surface propensity ($P_{surface}$) of an amino acid residue in a cluster or in a patch. The final scores were sorted in increasing order and the highest scoring patches or clusters for interface propensities were predicted to be part of an interface. We varied systematically the number of high scoring clusters in the clustering method and the patch size in the patch analysis to find empirically the optimal parameters for prediction in each method. Our analysis showed that the optimal range is eight to ten high ranking clusters in the clustering method, and that the optimal patch size is 11 to 12 Å with a score cutoff greater than 92% of the maximum patch score. This scoring scheme was tested for our training set of 72 protein complexes that have been used to derive the propensity values, and for 21 new independent test proteins.

Assessment of the prediction accuracy

The two prediction methods were assessed by counting the number of true positives, TP, i.e., interface residues which were correctly predicted as interfaces residues; false positives, FP, surface residues wrongly predicted as interface residues (overpredicted); true negatives, TN, surface residues correctly predicted as surface residues; and false negatives, FN; not predicted interface residues (underpredicted). The overall accuracy (Q_{Total}), sensitivity ($Q_{Sensitivity}$) and precision ($P_{precision}$) of the prediction methods were assessed with standard measures, reviewed by Baldi et al. [58] for bioinformatics studies:

$$Q_{Total} = 100\frac{TP + TN}{TP + TN + FP + FN} \quad (7a)$$

$$Q_{Sensitivity} = 100\frac{TP}{TP + FN} \quad (7b)$$

$$Q_{\mathrm{Pr}ecision} = 100\frac{TP}{TP + FP} \quad (7c)$$

Results and discussions

Propensity of the amino acid residues and correlation with experimental data

In the current work 72 different protein complexes derived from the protein data bank (shown in Table 1a) based on their solvent accessible surface area, distribution of the amino acid residues, their physical chemical properties and distribution of amino acid across the protein interface and on the protein surface were analyzed. Our results show that the propensities of most amino acid residues at protein interfaces are significantly different from the values on the protein surface (Fig. 1). As expected high values are found for hydrophobic residues, such as Phe, Trp, Ile, Leu, Met, Pro and Val, however, Ala does not have a particular preference for interfaces. In addition, some residues with polar or charged functional groups, such as Tyr, Cys, His and Arg are also found more frequently at protein interfaces as expected by chance.

High propensities of Leu, Trp, Tyr, and Phe at interface regions were also found in other studies [7, 59]. However, some major differences are found to the recently published values of Ma et al., e.g., the values of Ile (1.41 versus 0.23) and Asp (0.72 versus 1.55) reverse the preferences for interfaces in our and their conservation propensity scales. Our values are qualitatively similar to the study by the

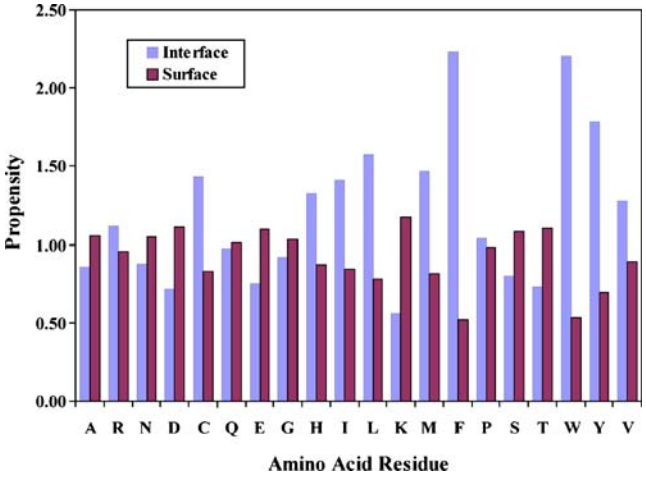

Fig. 1 Propensity of amino acid residues at protein interface (▪) and on the protein surface (▪). Propensity greater than 1 at the interface or on the surface indicates that the amino acid residues are more frequent at the interface or surface respectively

Thornton group [19]. Interestingly the residues Ser and Thr which are capable of forming hydrogen bonds to side chains, are less frequently found in interfaces which is also consistent with the low enrichment factors of these residues as hot spot residues [16]. Overall, our values correlate with the hot spot enrichment factors to a similar extent as the conservation propensity values [59] (r=0.35 versus r=0.36) if we include all amino acid residues. The r factor between our values and the enrichment factor increases to 0.50 as compared to 0.43 between the conservation propensity values and the enrichment factors, if we exclude the three residues Cys, Leu and Val with zero enrichment factors.

We believe that the low values of these three residues in the Ala scanning experiments [16] underestimate the importance of these residues in forming interfaces as the hydrophobic residue Ala is used as a reference point, and not for example a polar residue. For cysteine residues, not forming disulfide bridges, the sulfhydryl group (-SH) is inactive toward the water molecule and does not form hydrogen bond [60, 61], thus its physical chemical properties are similar to hydrophobic residues. The high propensity of Cys found in our study is partially due to its ability to form a disulfide bond across the protein interface. This is confirmed by a statistical analysis of short distances across interfaces. If we analyzed all short distances of atoms of the same type, restricted to distances less than 2.5 Å, the probability of SG atom was the highest (data not shown). If we increase the distance range to 2.5–3 Å, the pairing frequency of the oxygen and nitrogen atoms was prominent due to the formation of hydrogen bonds across the interface [15]. Beyond the range greater than 3 Å, the Cα and C_β interactions were found to be increasing, and finally the overall distributions of same atoms interacting with each other showed that the $C_\beta - C_\beta$ interaction has the highest frequency across the protein interface. Thus we chose C_β interactions as one of the criteria to define interacting residues (Appendix 1.1).

Distribution of amino acid residues from their physical chemical properties

In our study we would like to quantitatively characterize physical-chemical characteristics of interfaces which make them unique as compared to surfaces not involved in protein-protein interactions. Thus we analyzed the distribution of general physical chemical properties of interface residues using physical-chemical property scales derived in our previous work [45, 62]. In that work we have constructed five vectors, E1, E2, E3, E4 and E5, each representing a property with specific values for each amino acid. The vectors were generated by multidimensional scaling of a large number of 237 physical chemical properties, and we demonstrated that the distribution of the 20 amino acids in the 5-dimensional vector space is similar to the distribution in the original high dimensional property space. The new five properties have a clear physical interpretation and correlate well with the hydrophobicity (E1), size (E2), frequency of amino acid in α-helix (E3), number of degenerate codons (E4) and frequency of residue in β-strands (E5). Each vector gives a different decomposition of amino acid residues into five groups of amino acid residues namely, E1(α_{11}(VLIMFW), α_{12}(CY), α_{13}(AH), α_{14}(PTQR), α_{15}(GSNKDE)), E2(α_{21}(KRE), α_{22}(MFWQYHD), α_{23}(AVLITCN), α_{24}(PS), α_{25}(G)), E3(α_{31}(A), α_{32}(VLME), α_{33}(GISTQKHD), α_{34}(FCNR), α_{35}(WPY)), E4(α_{41}(VLIPKR), α_{42}(AGFSTY), α_{43}(WNQE), α_{44}(MHD), α_{45}(C)) and E5(α_{51}(VTCR), α_{52}(IYQS), α_{53}(NKH), α_{54}(AGLMFD), α_{55}(WPE)) [45]. In this quantitative scheme of property analysis the hydrophobic/hydrophilic separation of residues, for example, is shown in vector E1, with the hydrophobic residues in bin E1(α_{11}) and the polar and charged residues in bin E1(α_{15}).

Figure 2 shows the distribution of amino acid residues in each of the property vectors E1 and E2 at the protein interface and on the protein surface. In this figure the distribution of the E1 vector clearly shows the high propensity of hydrophobic residues across the interface relative to the protein surface (bin E1(α_{11})), and the high propensity of hydrophilic residues on the protein surface (bin E1(α_{15})). These bins represent distinct quantitative finger prints for distinguishing surface from interface regions. Further inspections of bins of other vectors reveal additional markers, for example, bin E2(α_{21}) and bin E2(α_{22}) in the vector E2 show also remarkable differences. The group of amino acid residues in bin E2(α_{21}) of vector E2 is highly populated with surface residues due to the large residues K, R and E, and bin E2(α_{22}) of E2 is highly populated by interface residues and reflects the high propensities of Phe,

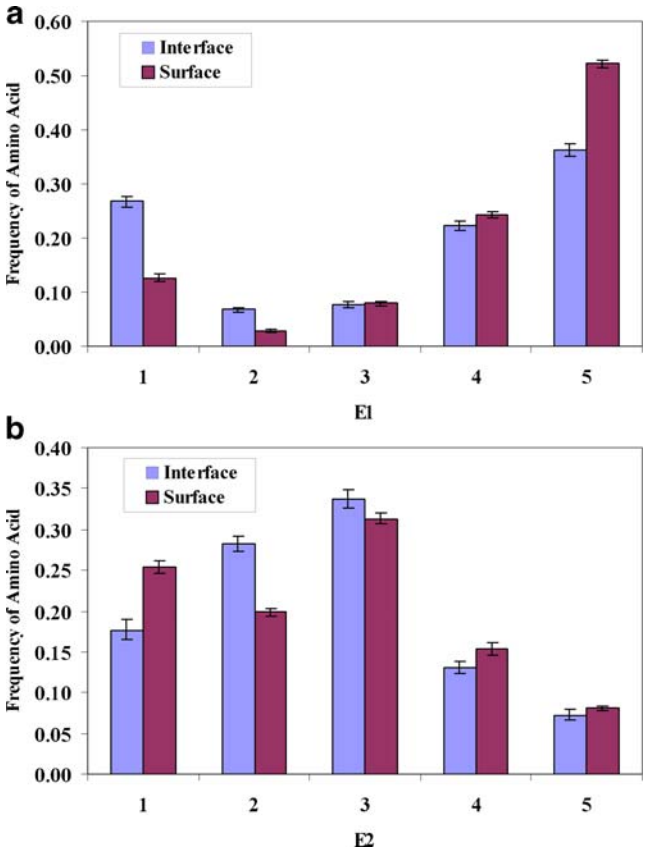

Fig. 2 Distribution of amino acid residues at the protein interface (blue) and on the protein surface (red) based on their physical chemical properties. The error bar in each bin represents the standard error calculated by $\sigma/\sqrt{n}$, where, n=72 and σ is the standard deviation calculated from the average value of **a**) E1 and **b**) E2 vector in all protein complexes

Met, Trp, Tyr and His residues for protein interfaces. Other bins with large differences are bin E3(α_{33}) of vector E3 and bin E5(α_{53}) of vector E5 (data not shown, see Appendix 1.2). Not all vectors contribute to a distinction between surface and interface areas to the same extent, as seen in the similar distributions of surface and interface residues in the bins of vector E4. Our analysis is different from a simple correlation analysis of physical chemical properties and interface propensities, as can be seen for the distributions of vector E3. We find no particular preferences for helical residues in interface regions, as the distributions of the bins with helix formers (bin E3(α_{31}), bin E3(α_{32})) and helix breakers are fairly similar, however a distinct difference is found for bin E3(α_{33}).

Our analysis also allows to quantitatively describing the nature of protein interfaces by calculating the pair frequencies of the groups of amino acid residues. The amino acid residues in interface and the surface regions are grouped into bins for each of the five vectors mentioned above, and we determined how often a residue of each group is near a residue of the same or another group in the some interface or on the surface. This statistics is different from evaluating interactions across the surface. A patch of radius 5 Å around each residue as described in the method section is used for this statistics. We calculated the distribution of pair-frequencies of residues around the interface residues $PI\left[Ei\left(\alpha_{i,j},\alpha_{i,j'}\right)\right]_{Interface}$ and around the surface residues $PS\left[Ei\left(\alpha_{i,j},\alpha_{i,j'}\right)\right]_{Surface}$. The pair distributions of interface and surface residues are shown in Fig. 3 for the property vector E1, and the pair-frequencies of interface and surface regions are compared by the log-odds ratio ξ, given in Table 2.

$$\xi\left(\alpha_{i,j},\alpha_{i,j'}\right)=\log_2\frac{PI\left[Ei\left(\alpha_{i,j},\alpha_{i,j'}\right)\right]_{Interface}}{PS\left[Ei\left(\alpha_{i,j},\alpha_{i,j'}\right)\right]_{Surface}} \quad (8)$$

Figure 3 clearly shows that hydrophobic/hydrophobic pairs dominate in interfaces as compared to surface regions. The pair-frequency of bin1 residues interacting with residues of bin1 in interfaces is more than twice as compared to

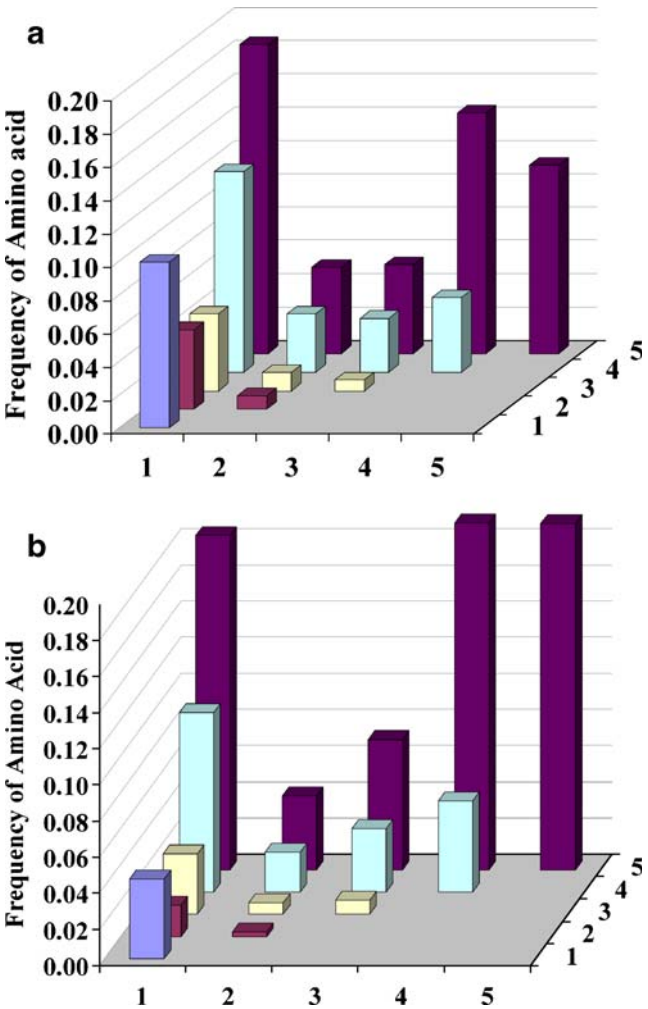

Fig. 3 **a**) Distribution of amino acid residues around the interface residues in E1 vector, and **b**) on the protein surface in E1 vector

Table 2 Log odd ratio table for pairing of amino acid residues(ξ) between different groups of property vector E1 to E5

	Bin1	Bin2	Bin3	Bin4	Bin5
a)					
Bin1	1.1811	1.4687	0.4624	0.2791	0.0044
Bin2	1.4687	1.5930	0.7272	0.6628	0.3399
Bin3	0.4624	0.7272	−0.2222	−0.1169	−0.4365
Bin4	0.2791	0.6628	−0.1169	−0.1772	−0.4110
Bin5	0.0044	0.3399	−0.4365	−0.4110	−0.7612
b)					
Bin1	−0.6128	−0.1766	−0.4211	−0.6665	−0.5070
Bin2	−0.1766	0.8736	0.3883	0.3050	0.3068
Bin3	−0.4211	0.3883	0.1033	−0.1517	−0.1045
Bin4	−0.6665	0.3050	−0.1517	−0.3935	−0.4552
Bin5	−0.5070	0.3068	−0.1045	−0.4552	0.2032
c)					
Bin1	−0.5932	−0.0942	−0.4914	0.1562	0.3373
Bin2	−0.0942	0.2957	−0.2121	0.4508	0.6812
Bin3	−0.4914	−0.2121	−0.5400	−0.0158	0.2503
Bin4	0.1562	0.4508	−0.0158	0.5974	0.8592
Bin5	0.3373	0.6812	0.2503	0.8592	1.1217
d)					
Bin1	0.1051	0.0278	−0.1508	−0.2061	0.4765
Bin2	0.0278	0.0319	0.0563	−0.0063	0.4967
Bin3	−0.1508	0.0563	−0.0553	−0.1995	0.4866
Bin4	−0.2061	−0.0063	−0.1995	−0.0468	0.2322
Bin5	0.4765	0.4967	0.4866	0.2322	0.0603
e)					
Bin1	−0.1448	0.0694	−0.3971	0.1251	−0.0768
Bin2	0.0694	0.3019	−0.2482	0.3091	0.1714
Bin3	−0.3971	−0.2482	−0.8124	−0.3691	−0.5266
Bin4	0.1251	0.3091	−0.3691	0.2568	0.0932
Bin5	−0.0768	0.1714	−0.5266	0.0932	−0.0951

a) Log odd ratio table for pairing between groups in residues in different bin of E1, **b)** between different bin of E2, **c)** of E3, **d)** of E4 and **e)** of E5

surface regions. Hydrophilic/hydrophilic pairs dominate in surfaces (bin5-bin5, Fig. 3b). This can also be seen in Table 2, where large and positive values of ξ indicate that pairing among these groups of residues is more often in protein interface while a negative value of ξ indicates that the pairing among residue group is more in the protein surface.

A large value of ξ shows that groups of amino acid residues having high propensity at protein interface (e.g., F, W, Y, C, M), often interact with each other. This can be seen from Table 2 that the value of ξ for a group of residues in bin1 interacting with bin1 (=1.1811), bin1 with bin2 (=1.4687) and bin2 with bin2 (=1.593) of vector E1 is greater than its value for group of residues in bin5-bin5 (=−0.7612). Bin1 and bin2 in E1 is dominated by the hydrophobic, including Cys and large residues while bin5 is dominated by the charge and hydrophilic residues. Similar results are also found for ξ in bin2-bin2 of E2 vector. A large value of ξ for Cys and Tyr in bin2 of E1 indicates that these residues are often found in protein interface and confirms high propensity of Cys residue at protein interface, see also Fig. 1. On other hand, low value of ξ=−0.7612 in bin5-bin5 of vector E1, $\xi_{i,j}$=−0.612 for bin1-bin1 of E2, ξ=−0.5932 for bin1-bin1, ξ=−0.54 for bin3-bin3 and ξ=−0.4914 for bin1-bin3 of E3 vectors shows that protein surface is mainly dominated by charged and hydrophilic residues. Also, the low value of ξ=0.0044 for interaction between the residues in bin1-bin5 of E1 vector clearly shows that the interaction between hydrophobic and hydrophilic residues at protein interface is smaller than protein surface.

Prediction of the functional sites and accessing the performance

We have calculated the performance of our scoring method using patch and cluster analysis as discussed in the previous sections. The performance of the method is based on the size of the patch or cluster used in the scoring function to predict

the interface residues on the protein surface (Fig. 4a,b). The overall accuracy of the method was found to be about ~70% when tested against the training and test data set. For smaller protein interfaces it has been shown that hydrophobic nature of amino acid residues plays an important role in binding protein interfaces together while for large interfaces both polar and hydrophobic residues contribute largely [3, 40]. Each patch or cluster on the protein surface contains approximately two to four amino acid residues if the size of the protein is small and four to eight residues if the size of protein is large. Empirically we find that in practice we achieve a good balance between precision and sensitivity by choosing the number of clusters to be eight and the patch size to be 12 Å, Fig. 4a,b. Patches or clusters predicted on the protein surface may include the residues present in close vicinity of the interface residues. In some cases we also observed that a high ranking cluster contained residues close to the interface. These residues can play an important role in stabilizing the protein interfaces.

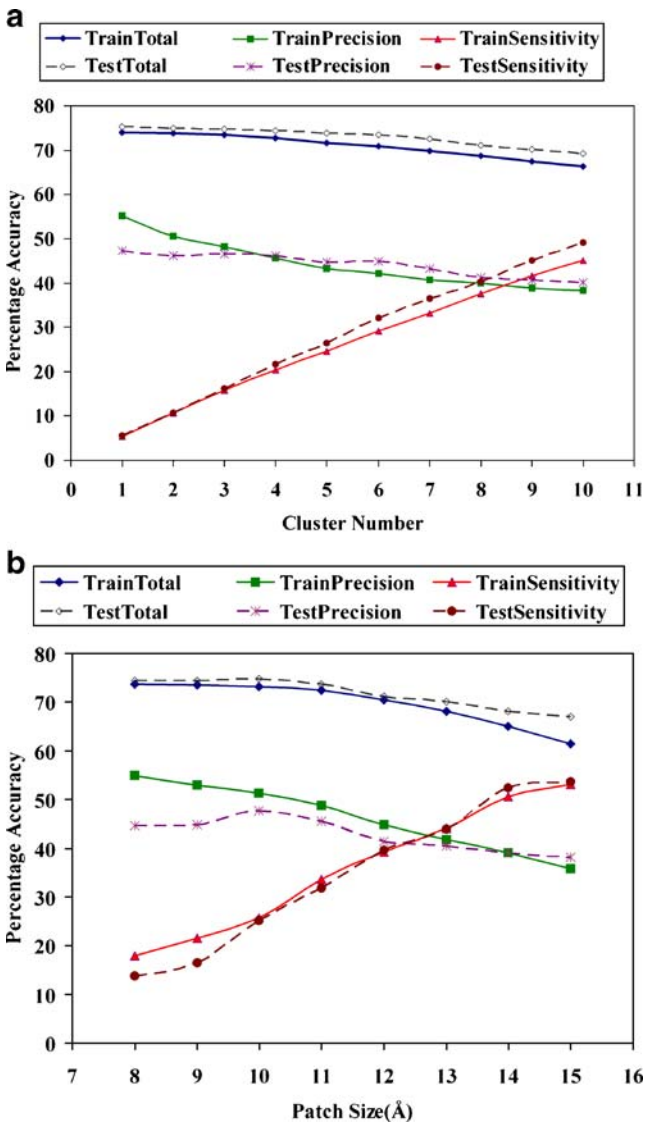

Fig. 4 Prediction accuracy for training and test data set obtained from **a**) cluster analysis and **b**) patch analysis

In our study, we have found that most of the predicted residues were either present in the actual interface or close to the interface residues as shown in the following examples. Our analysis shows that the characteristic of the interacting sites also depend on the nature of interacting residues as well on the geometry of the protein surface because the geometry of the protein surface carries more information about its function at the molecular level. The performance of the prediction method can be increased by including the environment of the residue at the interface which may varies from protein to protein [31]. Since a limited number of methods are available online [23, 41, 42, 44], we hope the method described here will provide users a new interface to predict the functional sites on the protein surface based on the structural information of the protein only. In our future work we are also planning to introduce the role of five amino acid descriptors in the scoring function and compare the performance of our prediction method against other available methods.

Comparison of the predicted and observed interface for a dimeric ATPase

The crystal structure of the dimeric ATPase MJ0577 of the hyperthermophile Methanococcus jannaschii (PDB id: 1MJH) consists of a five stranded parallel beta sheet and two helices on each side of the beta sheet [63]. The crystal structure of the protein shows different ATP binding motifs that are shared among many homologous protein of this family. The amino acid residues present at the actual protein interface are calculated by a change in ASA and shown by blue color in the protein surface, Fig. 5a. The amino acid residues predicted by clustering and patch analysis are shown by red and yellow color in protein surface, Fig. 5b,c. In case of 1MJH, the extra prediction comes from the cluster number 9 (contain residue V66, E67, E70, N71, L73, having interface score 1.04 and surface score 0.99 calculated by Eq. (7)) and cluster number 10 (contain residue no: T30, L31, K32, A33, Interface score=1.01 and surface score=1.00). Based on the scoring scheme, one can predict these two clusters as part of protein interface or the part of protein surface. Therefore, if we exclude these two clusters as part of the predicted protein interface, than we have found that the predicted residues are located in the same region of the actual dimeric interface, Fig. 5. However, the amino acid residues in these two clusters may involved in interactions with some other proteins.

We tested the performance of InterProSurf method against two previously published popular methods: ProMate [42]

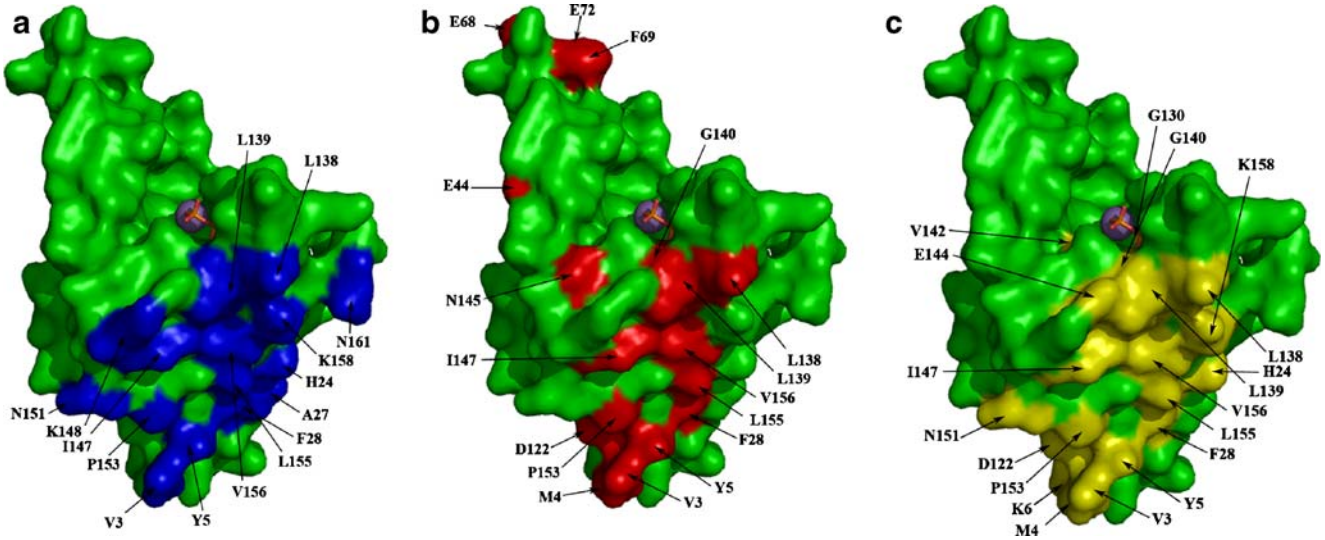

Fig. 5 Comparison of the actual and predicted residues in ATPase using InterProSurf. **a**) Actual interface residues present in crystal interface (blue). **b**) Predicted residues using cluster analysis (red). **c**) Predicted residues by patch analysis (yellow)

and ConSurf [41] by analyzing the interface residues in the dimeric structure of ATPase. First, we determined the experimentally observed interface residues by calculating the change in the solvent accessible surface area of amino acid residues upon complexation as described in the methods section. These interface residues were then compared with the residues predicted by the InterProSurf, ProMate and ConSurf webservers using default parameters (Fig. 6). All methods predict the experimentally observed interface to some extent, however, the predictions are far from perfect. Most high scoring residues of InterProSurf are within the observed interface regions. In practice it might be useful to use a consensus prediction to improve the prediction, as the basis for the three prediction methods implemented in InterProSurf, ProMate and ConSurf are different.

Illustration for the capsid protein of the human hepatitis B virus

The crystal structure of the human hepatitis B virus capsid protein [64] (PDB id: 1QGT) shows that two alpha helical hairpins form the dimer interface and the spikes on the capsid surface. The capsid contains a highly conserved C-terminal having amino acid residues from R112 to E127 followed by an irregular proline rich loop (residue no T128 to N136). These residues were found to be highly conserved in their sequence alignment and play an important role in inter-subunit interaction of the virus capsid. The Cys61 residue in each of the monomer protein forms a disulphide bridge at the dimer interface and is predicted correctly by our method, Fig. 7. The amino acid residues contributing to the antigenic

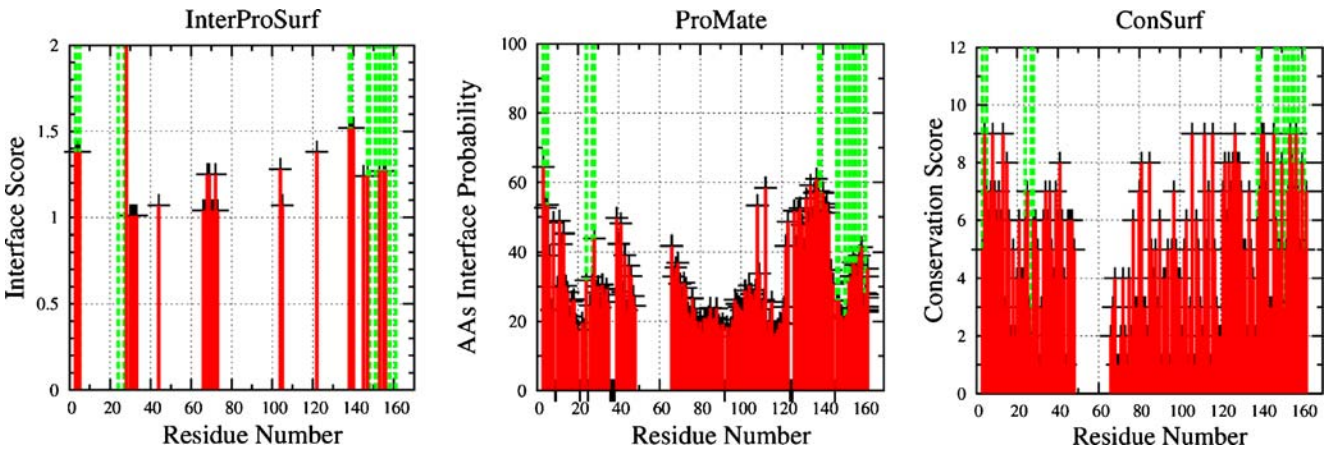

Fig. 6 Comparison of interface residues predicted for ATPase using the web servers InterProSurf, ProMate and ConSurf. The experimentally observed interface residues are shown in dashed lines while the predicted residues are shown as solid lines. High scoring residues of InterProSurf coincide to a large extent with the experimentally observed interface residues

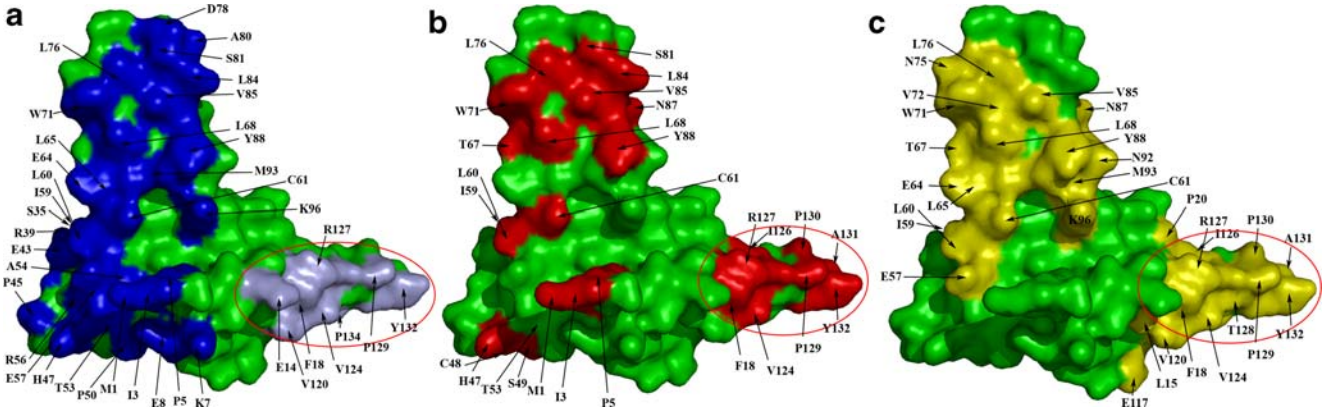

Fig. 7 Comparison of actual and predicted residues in the capsid protein of hepatitis B virus. **a**) The actual interface residues (blue). **b**) Predicted residues using cluster analysis (red). **c**) Predicted residues using patch analysis (yellow). Residues shown inside the circle predicted by our method are found to play an important role in inter subunit packing

site around the residue A80 form the dimer spikes were also predicted. Tyr132 was found to be fully exposed in the isolated dimer and buried in the protein complex also predicted. In addition, the predicted residues P129, P130 and I139 play an important role in the stability and inter-subunit packing [64]. The amino acid residues predicted by our method are also found to be highly conserved in their sequence alignment. As an illustration for the quality of our prediction in comparison to other publicly available webservers, we show the predictions for the human hepatitis B virus capsid protein of InterProSurf, ProMate and Consurf in Fig. 8. The quality of the prediction of InterProSurf is similar or slightly better than those methods.

Conclusions

For many protein complexes three-dimensional structures of the monomeric units are available, however the 3D structures and molecular details of the complexes are not known and require a large experimental effort. Our statistical analysis and the prediction tools we provide on our website can help to elucidate these unknowns. Our method is different from other studies using evolutionary information or correlated mutations across interfaces, however can be combined with these methods in practice to achieve a higher reliability. Our results also give some insights of hot spot residues without any prior knowledge of thermodynamic analysis. The protein interfaces are not found to be uniform in terms of amino acid residue properties. We found a characteristic difference in the pair frequencies of residues in interface and surface regions in terms of their physical chemical properties which can be used to further characterize hot spots in quantitative terms from sequence information.

The new method for predicting the interacting residues has been implemented in a completely automated procedure and is publicly available through our web site at http://curie.utmb.edu/prosurf.html. The prediction method described

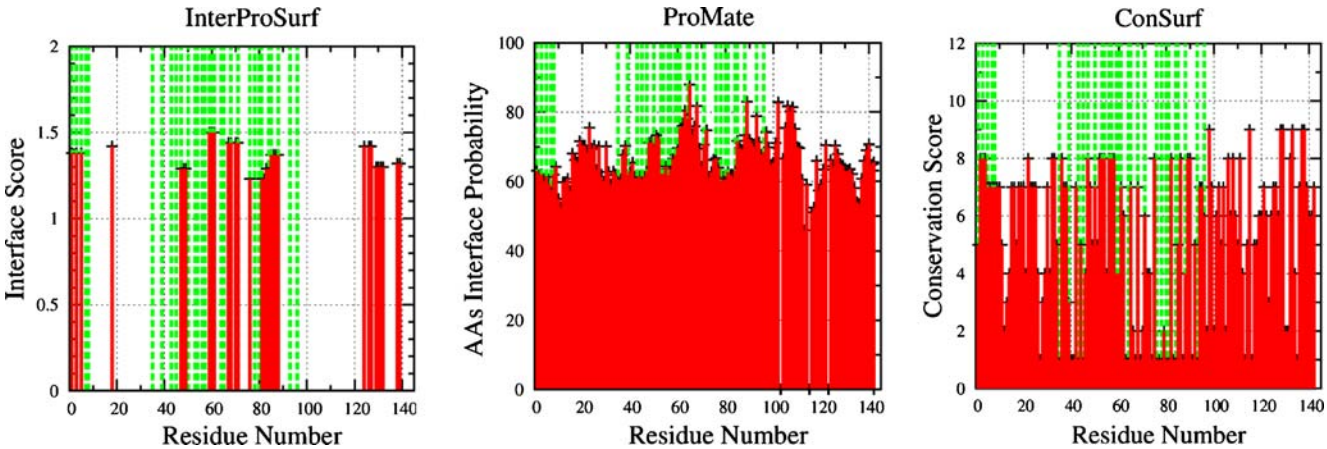

Fig. 8 Comparison of interface residues predicted for the capsid protein of hepatitis B using InterProSurf, ProMate and ConSurf web servers. In addition to actual interface residues the InterProSurf method also predicts one highly conserved cluster of amino acid residues from 137-143 which can be seen with ConSurf analysis

here does not include any information about the partner protein and is solely based on the solvent accessible surface area of the monomer protein and propensity of amino acid residues [46]. In defining the scoring function, we assumed that the contribution of each amino acid residue in binding protein interfaces is independent and therefore their contribution can be summed. The performance of the method was tested for a dataset of 21 protein complexes independent from the complexes used in deriving the propensity values of the algorithm. We showed that the overall accuracy obtained from both patch analysis and cluster analysis is about 70%. The accuracy of the method can be further increased by combining our method with evolutionary information or choosing more detailed scoring functions using the statistical analysis of pair frequencies we present. We already used the methods in practice to guide experimental mutations of the envelope protein E1 of the Venezuelan Equine Encephalitis Virus to help in designing of new drugs [46].

Acknowledgements This work was supported by National Institutes of Health Grants R21 AI055746 and R01 AI064913.

References

1. Kinoshita K, Nakamura H (2003) Curr Opin Struct Biol 13:396–400
2. Archakov AI, Govorun VM, Dubanov AV, Ivanov YD, Veselovsky AV, Lewi P, Janssen P (2003) Proteomics 3:380–391
3. Lo Conte L, Chothia C, Janin J (1998) J Mol Biol 285:2177–2198
4. Janin J, Chothia C (1990) J Bio Chem 265:16027–16030
5. Tsai CJ, Lin SL, Wolfson HJ, Nussinov R (1997) Protein Sci 6:53–64
6. Hu ZJ, Ma BY, Wolfson H, Nussinov R (2000) Proteins-Structure Function Genetics 39:331–342
7. Jones S, Thornton JM (1996) Proc Natl Acad Sci USA 93:13–20
8. McCoy AJ, Epa VC, Colman PM (1997) J Mol Biol 268:570–584
9. Thorn KS, Bogan AA (2001) Bioinformatics 17:284–285
10. Bader GD, Betel D, Hogue CWV (2003) Nucleic Acids Res 31:248–250
11. Xenarios I, Salwinski L, Duan XQJ, Higney P, Kim SM, Eisenberg D (2002) Nucleic Acids Res 30:303–305
12. Xenarios L, Eisenberg D (2001) Curr Opin Biotechnol 12:334–339
13. DeLano WL (2002) Curr Opin Struct Biol 12:14–20
14. DeLano WL, Ultsch MH, de Vos AM, Wells JA (2000) Science 287:1279–1283
15. Kortemme T, Baker D (2002) Proc Natl Acad Sci USA 99:14116–14121
16. Bogan AA, Thorn KS (1998) J Mol Biol 280:1–9
17. Clackson T, Wells JA (1995) Science 267:383–386
18. Jones S, Thornton JM (1997) J Mol Biol 272:133–143
19. Jones S, Thornton JM (1997) J Mol Biol 272:121–132
20. Jones S, Thornton JM (1995) Prog Biophys Mol Biol 63:31–35
21. Jones S, Marin A, Thornton JM (2000) Protein Eng 13:77–82
22. Bordner AJ, Abagyan R (2005) Proteins: Structure, Function, Bioinformatics 60:353–366
23. Murakami Y, Jones S (2006) Bioinformatics 22:1794–1795
24. Brinda KV, Kannan N, Vishveshwara S (2002) Protein Eng 15:265–277
25. Landgraf R, Xenarios I, Eisenberg D (2001) J Mol Biol 307:1487–1502
26. Morrison KL, Weiss GA (2001) Curr Opin Chem Biol 5:302–307
27. Kortemme T, Kim DE, Baker D (2004) Sci STKE 219:12
28. Massova I, Kollman PA (1999) J Am Chem Soc 121:8133–8143
29. Zhou HX, Shan Y (2001) Proteins 44:336–343
30. Fariselli P, Pazos F, Valencia A, Casadio R (2002) Eur J Biochem 269:1356–1361
31. Koike A, Takagi T (2004) Protein Engineering Design & Selection 17:165–173
32. Bradford JR, Westhead DR (2004) Bioinformatics 21:1487–1494
33. Gallet X, Charloteaux B, Thomas A, Brasseur R (2000) J Mol Biol 302:917–926
34. Bock JR, Gough DA (2001) Bioinformatics 17:455–460
35. Gao Y, Wang RX, Lai LH (2004) J Mol Model 10:44–54
36. Yao H, Kristensen DM, Mihalek I, Sowa ME, Shaw C, Kimmel M, Kavraki L, Lichtarge O (2003) J Mol Biol 326:255–261
37. Glaser F, Pupko T, Paz I, Bell RE, Bechor-Shental D, Martz E, Ben-Tal N (2003) Bioinformatics 19:163–164
38. Chakrabarti P, Janin J (2002) Proteins 47:334–343
39. Bahadur RP, Chakrabarti P, Rodier F, Janin J (2004) J Mol Biol 336:943–955
40. Glaser F, Steinberg DM, Vakser IA, Ben-Tal N (2001) Proteins-Structure Function Genetics 43:89–102
41. Landau M, Mayrose I, Rosenberg Y, Glaser F, Martz E, Pupko T, Ben-Tal N (2005) Nucleic Acids Res 33:W299–W302
42. Neuvirth H, Raz R, Schreiber G (2004) J Mol Biol 338:181–199
43. Ogmen U, Keskin O, Aytuna AS, Nussinov R, Gursoy A (2005) Nucleic Acids Res 33:W331–W336
44. Mihalek I, Res I, Lichtarge O (2006) Bioinformatics 22:1656–1657
45. Venkatarajan MS, Braun W (2001) J Mol Model 7:445–453
46. Negi SS, Kolokoltsov AA, Schein CH, Davey RA, Braun W (2006) J Mol Model 12:921–929
47. Altschul S, Madden T, Schaffer A, Zhang JH, Zhang Z, Miller W, Lipman D (1998) Faseb J 12:A1326–A1326
48. Thompson JD, Higgins DG, Gibson TJ (1994) Nucleic Acids Res 22:4673–4680
49. Fraczkiewicz R, Braun W (1998) J Comp Chem 19:319
50. Liang S, Liu Z, Li W, Ni L, Lai L (2000) Biopolymers 54:515–523
51. Singh RK, Tropsha A, Vaisman II (1996) J Comp Biol 3:213–221
52. Linde Y, Buzo A, Gray RM (1980) IEEE Trans Commun 28:84–95
53. Sayood K (2000) Introduction to data compression, 2nd edn. Morgan Kaufmann Publishers Inc
54. Patane G, Russo M (2001) Neural Netw 14:1219–1237
55. Equitz HW (1989) IEEE Trans Acoustics Speech Signal Process 37:1568–1575
56. Cosman PC, Oehler KL, Riskin EA, Gray RM (1993) Proc I E E E 81:1326–1341
57. Lin CL, Tai SC (1998) IEEE Trans Circuits Syst, II Analog Digit Signal Process 45:432–435
58. Baldi P, Brunak S, Chauvin Y, Andersen CAF, Nielsen H (2000) Bioinformatics 16:412–424
59. Ma B, Elkayam T, Wolfson H, Nussinov R (2003) Proc Natl Acad Sci USA 100:5772–5777
60. Nagano N, Ota M, Nishikawa K (1999) FEBS Lett 458:69–71
61. Karlin S, Zhu ZY, Baud F (1999) Proc Natl Acad Sci USA 96:12500–12505
62. Mathura VS, Schein CH, Braun W (2003) Bioinformatics 19:1381–1390
63. Zarembinski TI, Hung LW, Mueller-Dieckmann HJ, Kim KK, Yokota H, Kim R, Kim SH (1998) Proc Natl Acad Sci USA 95:15189–15193
64. Wynne SA, Crowther RA, Leslie AGW (1999) Mol Cell 3:771–780