Exploiting Sequence and Structure Homologs to Identify Protein–Protein Binding Sites

Jo-Lan Chung,[1,3] **Wei Wang,**[1] **and Philip E. Bourne**[2,3*]
[1]*Department of Chemistry and Biochemistry, University of California, San Diego, La Jolla, California*
[2]*Department of Pharmacology, University of California, San Diego, La Jolla, California*
[3]*San Diego Supercomputer Center, University of California, San Diego, La Jolla, California*

ABSTRACT **A rapid increase in the number of experimentally derived three-dimensional structures provides an opportunity to better understand and subsequently predict protein–protein interactions. In this study, structurally conserved residues were derived from multiple structure alignments of the individual components of known complexes and the assigned conservation score was weighted based on the crystallographic B factor to account for the structural flexibility that will result in a poor alignment. Sequence profile and accessible surface area information was then combined with the conservation score to predict protein–protein binding sites using a Support Vector Machine (SVM). The incorporation of the conservation score significantly improved the performance of the SVM. About 52% of the binding sites were precisely predicted (greater than 70% of the residues in the site were identified); 77% of the binding sites were correctly predicted (greater than 50% of the residues in the site were identified), and 21% of the binding sites were partially covered by the predicted residues (some residues were identified). The results support the hypothesis that in many cases protein interfaces require some residues to provide rigidity to minimize the entropic cost upon complex formation. Proteins 2006;62:630–640.** **© 2005 Wiley-Liss, Inc.**

Key words: structurally conserved surface residues; protein interfaces; support vector machines

INTRODUCTION

Structural genomics projects promise to yield a large increase in the number of experimentally determined protein structures and hence a more comprehensive view of protein fold space.[1,2] Lacking a functionally driven impetus, many of these newly determined structures are functionally uncharacterized. Exploiting this raw structural information and at the same time identifying functional sites, such as protein–protein binding sites, has become an important question.[3] Moreover, because proteins function through their interactions with other molecules, identification of these binding sites is important for mutant and drug design, functional annotation, and understanding the mechanism of the molecular recognition processes. How then can we further our understanding of protein–protein interactions with the advent of additional structure information, most often consisting of single uncomplexed components?

A number of studies on the characteristics of protein interfaces have provided clues for binding site prediction. In terms of physical chemistry, protein interfaces are generally observed to be more hydrophobic than the remainder of the protein surface.[4] Moreover, the interfaces of permanent complexes tend to be more hydrophobic when compared to those of transient complexes.[4,5] Different amino acid compositions have also been observed among the interaction sites of homo-permanent complexes, homo-transient complexes, hetero-permanent complexes, and hetero-transient complexes.[6] In addition, through alanine-scanning mutagenesis, it has been found that the binding free energy is not distributed equally across these protein interfaces.[7,8] A small subset of residues at the interfaces form energy hot spots, enriched in tyrosine, tryptophan, and arginine.[8,9] These hot spots are surrounded by hydrophobic rings that potentially occlude bulk solvent.[10] From a geometric perspective, most protein interfaces, especially for large protein–protein complexes, are relatively flat when compared with the remainder of the surface, whereas most enzyme-binding sites form the largest surface cavities.[4,11,12] Hu et al.[13] and Ma et al.[14] have analyzed representative families of protein–protein interfaces. They found that protein-binding sites are studded with structurally conserved residues, especially polar residues and provide a signature for energy hot spots. Further, Ma et al.[14] showed that these structurally conserved residues distinguish between binding sites and exposed protein surfaces. They suggested that this finding might be used for predicting binding sites or hot spots.

A number of further *in silico* approaches for predicting protein–protein binding sites have been proposed.[15] Docking procedures identify the binding mode of two interacting proteins with known structures and build an atomic model of the complex by using geometric and electrostatic

Grant sponsor: National Institutes of Health; Grant numbers: GM63208, GM08326

*Correspondence to: Philip E. Bourne, Department of Pharmacology, San Diego Supercomputer Center, University of California, San Diego, Gilman Drive, La Jolla, CA 92093-0537. E-mail: bourne@sdsc.edu

Received 12 January 2005; Revised 14 June 2005; Accepted 14 July 2005

Published online 2 December 2005 in Wiley InterScience (www.interscience.wiley.com). DOI: 10.1002/prot.20741

complementary.[16,17] Homology modeling[18] and multimeric threading[19] approaches, which apply the knowledge learned from a known complex structure to the component (apo) homologs and examine the predicted contact residues using a specific potential function, are able to predict interaction partners of a protein and model the complex using sequence alignments. The success of these two approaches usually require at least 30% amino acid identity between homologs.[15,20]

Several additional computational methods can also predict protein-binding sites without any knowledge of binding partners. Evolutionary trace and related methods[21–25] (e.g., ConSurf[25]) map the conserved residues of a protein, derived from sequence alignment and associated phylogenetic trees, onto its structure and then searched for 3D clusters. Pazos et al.[26] used correlated mutations from sequence alignments to predict binding partners and also simultaneously define binding sites. Other methods identify the binding sites on the basis of the residue hydrophobicity,[27] the properties of surface patches,[28,29] and the charge distributions at protein interaction interfaces.[30] Recently, neural networks and support vector machines (SVMs) were trained to predict binding sites using sequence profiles and accessible surface areas.[31–35]

All these methods use the evolutionary information from the sequence alignment, and/or residue properties, and/or structural properties. It is difficult to compare the predictive power of these various methods because of the difference in training data and definition of binding sites as well as the lack of common test sets.[15] To the best of our knowledge, none of the above methods exploits the information of spatially conserved residues in similar structures to predict the protein-binding sites without the prior knowledge of their binding partners, although recently structurally conserved residues were reported to correspond to hot spots on protein interfaces and can be derived from multiple structure alignments.[14] Integrated structure-based approaches will likely become more important as more protein structures are determined by structural genomics.

The key to using structure-based information is to determine structurally conserved residues beforehand. In this study, we identified all structurally conserved residues in the nonredundant chains of hetero-complexes from the Protein Data Bank (PDB),[36] where each of these chains was structurally aligned to at least three other nonredundant members. These conserved residues were determined using a structural conservation score derived from the multiple structure alignments and weighted by the normalized B factors to consider the impact of the flexibility on the quality of the alignment. To combine evolutionary information derived from sequence alignments and physical chemistry information derived from structures, a support vector machine (SVM) was trained to identify binding site residues using structural conservation, accessible surface area, and sequence profile of structurally neighboring residues. We show that structural conservation can significantly improve the predictive power of the SVM.

MATERIALS AND METHODS

Data Set

All non-NMR protein structures that have multiple chains and a resolution better than 3.5 Å were collected from the PDB (March, 2004).[36] For each structure, the reduction of solvent accessible surface area (ASA) upon binding between any two chains was calculated using the DSSP program.[37] A pair of chains was selected if the reduction of ASA was $\geq$ 450 Å^2. The pairs containing chains belonging to SCOP class $\geq$ 8[38,39] or less than 80 amino acids long were discarded to filter out small fragments and focus our attention on relatively large protein–protein interfaces.

The nonredundant chains of hetero-complexes were selected by the method of Zhou et al.[35] with the following modifications. All the chains of selected pairs were compared using BLAST.[40] Two chains were assigned to the same cluster if the residue identity was $\geq$ 30% and $>$ 90% of the amino acids were aligned. All the interacting pairs were clustered this way. For proteins that interact with multiple partners, the one with the most binding partners was chosen as a representative. Collectively, the representative pairs comprise the nonredundant set of hetero-complexes.

The structural homologs corresponding to these nonredundant chains were retrieved at the SCOP family level from the Astral database[41,42] at the 40% sequence purge level. Each chain was then aligned with its homologs using CE-MC,[43,44] a multiple structure alignment algorithm. If a structure was composed of more than one SCOP domain, different parts of this structure were aligned separately. A chain was selected if at least four members were aligned over 60% of their lengths and with a Z score[45] above 4.0. The selected chains were further filtered by discarding chains with less than 20 interfacial contacts or with no B-factors provided. The final data set consisted of 274 nonredundant chains of hetero-complexes, containing 319 binding sites. Each of these chains was accompanied by a structure alignment of at least four nonredundant members. The data are available upon request.

Definition of Surface Residues and Interface Residues

A residue was defined as a surface residue if its ASA was at least 15% of its nominal maximum area.[46] The ASA for each chain was calculated separately using the DSSP program.[37] The coordinates of a single chain were obtained from the corresponding complex structure.

A residue was considered to form an interfacial contact if the distance between any of its heavy atoms and any heavy atoms of its interacting proteins were $<$ 5 Å. A surface residue was defined to be an interface residue if it formed an interfacial contact. According to this definition, about 28% (10,615) of all surface residues in the data set represented protein–protein binding sites.

Normalization of B-factor

B-factors determined by X-ray crystallographic experiments provide an indication of degree of mobility and disorder of an atom in a protein.[47] Several studies have used B-factors to estimate conformational flexibilities.[48,49] Because of the difference among structures in the data set, the B-factors must be normalized before comparison. In this study, the C_α B-factors were normalized using the method described by Smith et al.[48,49] First, a median-based method was used to detect B-factor outliers in a chain; outliers correspond to residues located within segments of unusually high mobility. These outliers were removed before normalization. A residue with $M_i > 3.5$ was said to be an outlier. A M_i value was calculated as follows:

$$M_i = 0.6745 \times |x_i - \tilde{x}|/MAD$$

where x_i is the B-factor of the ith residue, $\tilde{x}$ is the median of the B-factors, and MAD is the median of absolute displacements from the median. The B-factor was then normalized as follows:

$$B_{norm,i} = (B_i - \mu_{noout})/\sigma_{noout}$$

where μ_{noout} and σ_{noout} are the mean and standard deviation of the C_α B-factors for a chain calculated after removal of outliers.

The Structural Conservation Score

The structural conservation score was derived from the structure alignment generated by CE-MC algorithm. The raw conservation score at each aligned position x was defined as follows:

$$C(x) = \frac{2}{N(N-1)} \times \sum_{i}^{N}\sum_{j>i}^{N} L(s_i(x),s_j(x))$$

where N is the total number of aligned structures, $s_i(x)$ is the amino acid at position x in the ith structure in the alignment. L is defined as follows and weighted by the residue similarity (M):

$$L(s_i(x),s_j(x)) = \exp(-d(s_i(x),s_j(x))) \times M(s_i(x),s_j(x))$$

where d is the distance between the C_α atom of $s_i(x)$ and $s_j(x)$. If a gap is present in either structure, a value of 6 is assigned. M is defined by Valdar et al.[50] as follows:

$$M(s_i(x),s_j(x)) = \begin{cases} \dfrac{m(a,b) - min(m)}{max(m) - min(m)} & a \neq gap \cap b \neq gap \\ 0 & else \end{cases}$$

where m is a modified PET substitution matrix.[50]

The raw structural conservation score was then weighted by the normalized B-factor ($B_{norm,i}$) from the query structure to account for the structural flexibility that will lead to a poor alignment. Flexible regions are usually poorly aligned hence the conservation scores derived form the structural alignments of those regions are less reliable than the scores derived from other regions. Therefore, for two residues with the same raw structural conservation scores, the one located in the flexible regions (higher B-factor) should be given a lower conservation score than the one in the rigid regions (lower B-factor). Thus:

$$_{weighted}C(x) = C(x)^{r(x)}$$

where

$$r(x) = \exp(B_{norm,i})$$

Since $C(x)$ is between 0 and 1, a residue with a smaller $B_{norm,i}$ score (less flexible) would have a higher $_{weighted}C(x)$. The $_{weighted}C(x)$ is equivalent to $C(x)$ when $B_{norm,i}$ is 0.

The Implementation of Support Vector Machines

SVMs[51] are powerful supervised learning algorithms for making binary decisions. The data samples are mapped to a high-dimensional feature space using a kernel function. An optimal separating hyper-plane is then constructed to separate two classes, maximizing the margin between them.[51] In this study, the SVMs were trained to predict whether or not a surface residue was located at the interface. It was implemented using SVM^{light} [52] with the radial basis function as a kernel. The value of γ and the regularization parameter C were tuned to 0.01 and 10, respectively (see References 51 and 52 for details of tuning these parameters).

Four SVM predictors were trained with different inputs. Each predictor was input a window containing a surface residue and its 12 spatially nearest surface residues (not to be confused with contiguous residues). An interface residue was defined to belong to the positive class, and a noninterface residue was defined to belong to the negative class. For predictor 1, the inputs were sequence profile and accessible surface area of residues in the window. The sequence profile was obtained from three iterations of a PSI-BLAST search against the NCBI nonredundant database (NR) with $e = 0.001$ and $h = 0.001$.[53] Accordingly, each residue was encoded as a feature vector with 13×21 dimensions: (the surface residue to be predicted + 12 nearest neighbors) × (20 amino acids + accessible surface area). For predictor 2, the sequence profile, accessible surface area, and structural conservation score (weighted by the normalized B-factor) were used as inputs (13×22 dimensions). The inputs for predictor 3 and predictor 4 were the same as those for predictor 2 except the weighted structural conservation score was replaced by the raw conservation score (without weighting) and the normalized B-factor, respectively. All the input values were scaled between -1 and 1 before being fed into the SVM.

It is known that a SVM makes more correct decisions on a large characterized data set. In this data set, only 28% of the surface residues were interface residues. If all surface residues were used in the training, the machine would be biased to predict a residue as a noninterface residue. To address this issue, a set of noninterface surface residues was randomly selected to make the ratio of positive and negative data 1:1. Threefold cross-validation was then used to train the SVM. The whole data set was randomly divided into three subgroups with an approximately equal number of chains. Each SVM was run three times with three different training and test sets. For each run, two of

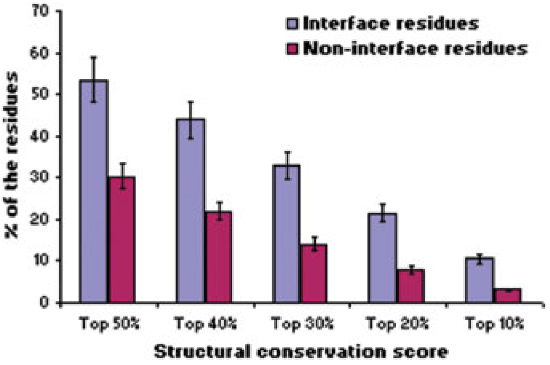

Fig. 1. Discrimination between interface and noninterface residues for different structural conservation scores. The error bars present ±10% deviation of the measurements.

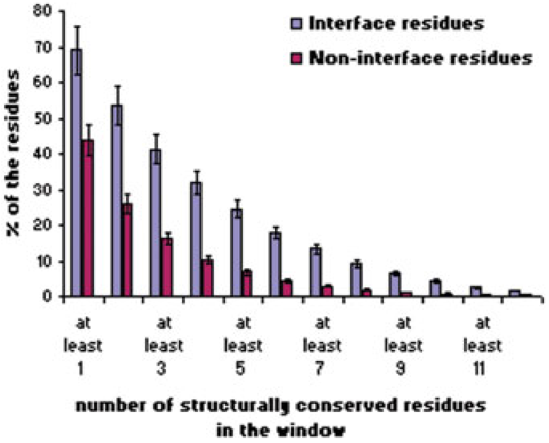

Fig. 2. Proportion of the interface and noninterface surface residues surrounded by at least a certain number of structurally conserved surface residues in a window sided by 12 spatially nearest surface residues. A residue scoring within the top 20% was said to be structurally conserved. The error bars present ±10% deviation of the measurements.

the subgroups were used in training with the remaining subgroups used as the test set. The performance of the SVM was measured by the precision (the proportion of the correctly predicted interface residues with respect to the total positively identified residues) and recall (the proportion of the correctly predicted interface residues with respect to the total number of interface residues).

The Clustering of Predicted Residues

Clustering was used to remove the isolated interface residues predicted by SVMs and include the noninterface residues surrounded by several predicted interface residues. Any two predicted interface residues were clustered together using the hierarchical clustering method if the distance between their C_β atoms was < 6.5 Å. Clusters with only one or two residues were then removed. A predicted noninterface residue was converted to an interface residue if its C_β atom and at least three predicted interface residues were located within 6 Å.

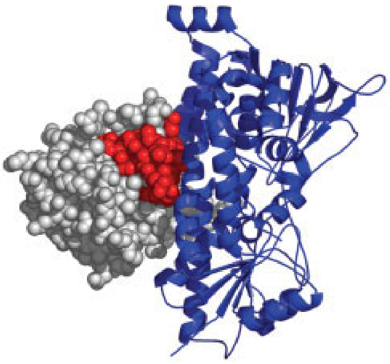

Fig. 3. Residues with the top 20% of structural conservation scores (red) mapped to adrendoxin (Adx, PDB code 1E6E:B) and known to bind adrenodoxin reductase (AR, blue). The secondary structures of Adx are shown in Figure 6(d).

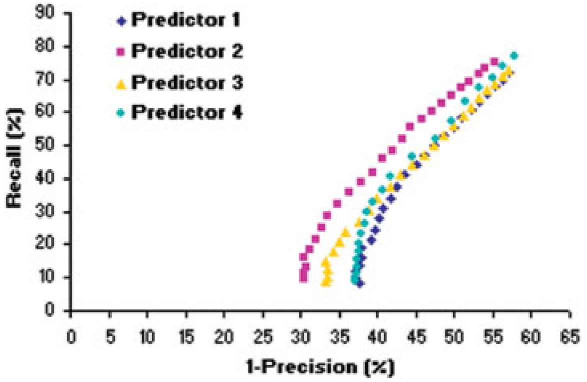

Fig. 4. Comparative performances of the predictors. Predictor 1: sequence profile + ASA. Predictor 2: sequence profile + ASA + structural conservation score. Predictor 3: sequence profile + ASA + raw structural conservation score unweighted by the normalized B factor. Predictor 4: sequence profile + ASA + normalized B factor.

RESULTS AND DISCUSSION

Structurally Conserved Surface Residues

The structural conservation scores of all 60,834 surface residues in the 274 nonredundant chains of the heterocomplexes were calculated. The proportion of residues scoring above a given threshold score at the interface and noninterface was compared (Fig. 1). It is clear that structurally conserved residues are predominant at the binding interfaces compared to the rest of the protein surface. For example, the proportion of interface residues with the top 40% of scores was more than twice that of the noninterface residues. Figure 2 illustrates the comparison between interface and noninterface residues surrounded by a specified number of structurally conserved surface residues for a window sided by the 12 spatially nearest surface residues (x-axis). A residue scoring in the top 20% was defined as structurally conserved. Note that the residues within the window were spatially close on the surfaces but not necessarily contiguous in sequence. Compared to noninterface residues, interface residues have significantly more structurally conserved residues in their spatial neighborhoods. Figures 1 and 2 illustrate that structurally conserved residues were more clustered at interfaces and distinguished the interfaces from the remainder of the surface. Figure 3 illustrates these structurally conserved residues mapped onto the protein surface. Most of the structurally conserved residues on the surface of adrendoxin (Adx) were located at the interface that binds to adrenodoxin reductase (AR).

Ma et al.[14] derived structurally conserved residues and showed that these residues distinguish binding sites from the remainder of the surface. In their study, 10 representative interfaces and their homologous members were aligned using MUSTA,[54,55] a sequence-independent structure comparison algorithm. A residue was defined as structurally conserved if >80% of the members in the alignment had an identical residue aligned within 1.0 Å. This approach is more precise because similar interfaces were directly aligned. However, structurally conserved residues cannot be derived when the structures of complexes or their homologous members are not available. In the present study, the individual components of the complex, rather than the whole complex, were aligned to their nonredundant structural homologs belonging to the same SCOP family retrieved from the ASTRAL database. According to SCOP, the proteins are in the same SCOP family if their residue identity is >30%, or they have lower residue identity but with very similar functions and structures. The structurally conserved residues were then derived from the alignment with B-factors as weighting factors (discussed later). This method allows large-scale analyses because there are more structure homologs of unbound structures than complexes. Figures 1 and 2 show that the structurally conserved residues derived from single structures by our approach successfully discriminate interface and noninterface surface residues. It is anticipated that the majority of structures provided by the structural genomics projects are not complexes and without information on their binding partners. Our approach can exploit these newly determined structures to identify structurally conserved residues, which can be further used for binding site predictions.

The Predictor Using the Structural Conservation Score

SVM predictors were trained using threefold cross-validation with different inputs. Because all the chains in the data set have below 30% residue identity, the interaction mode cannot be directly inferred from sequence homology alone.[20] The performances of predictors 1 and 2 were compared in Figure 4. The sequence profile and accessible surface area of the surface residue to be predicted and its 12 spatially close surface residues were input to both of the predictors. Predictor 2 included the structural conservation score as an additional input feature. The curves in Figure 4 were made by raising the discrimination value of the SVMs from 0 with a 0.25 increment each time until the average number of predicted interface residues of a protein was around 5. The performance of predictor 2 was better than that of predictor 1, especially at high precision. It should be noted that this improvement at high precision is especially useful for the selection of target residues for site-specific mutagenesis. The predicted interface residues of predictors 1 and 2 were then clustered to locate the binding sites on the protein surface. The discrimination value of each predictor was adjusted to give 50% precision after clustering, where predicted interfaces residues were roughly 35% of the total surface residues in the data set. Compared to predictor 1, improvements in correctly or precisely predicted binding sites were achieved with predictor 2 (the definitions for different levels of prediction are described in Table I). The number of correctly predicted binding sites increased 17%, and the number of precisely predicted binding sites increased 13% from predictor 1 to 2. Both predictors predicted significantly better than the random predictions (shown in parentheses in Table I). Finally, 52% of the binding sites were precisely predicted, 77% of the binding sites were correctly predicted, and 21% of the binding sites were partially predicted. These results indicate that structural conservation score is able to be effectively exploited in the identification of protein-binding sites.

TABLE I. Prediction Performances

(a) The precision and recall before and after clustering

	Before Clustering		After Clustering	
Predictor[a]	Precision (%)	Recall (%)	Precision (%)	Recall (%)
Predictor 1	49.8	56.8	50.0	59.7
	(27.7)[c]	(31.6)	(30.6)	(23.0)
Predictor 2	51.0	63.7	50.0	67.3
	(27.6)	(34.5)	(29.3)	(27.1)

(b) Percentage of predicted binding sites for the two predictors at 50% precision after clustering

	% of All Binding Sites After Clustering[b]			
Predictor	Precise Prediction	Correct Prediction	Partial Prediction	Wrong Prediction
Predictor 1	39.5	59.9	31.7	8.5
	(0.9)	(6.6)	(81.5)	(11.9)
Predictor 2	52.0	76.5	21.0	2.5
	(0.6)	(9.4)	(82.1)	(8.5)

[a] Predictor 1: sequence profile + ASA. Predictor 2: sequence profile + ASA + structural conservation score.

[b] Precise prediction: at least 70% of the interface residues were identified. Correct prediction: at least 50% of the interface residues were identified. Partial prediction: some but less than 50% of the interface residues were identified. Wrong prediction: no interface residues were identified.

[c] Random predictions were obtained by random shuffling of the predicted interface and noninterface residues.

Sequence profile and accessible surface area used in predictor 1 have been used in previous studies using SVMs or neural networks as predictors[31,34,35] to define interface residues. Although similar inputs were used in these studies, the predictions differ slightly. This may be due to differences in the data sets and in the definitions of interface residues.[31] Koike et al.[31] also included hydrophobicity, sequence conservation score, patch flatness, amino acid ratio in the patch, and interaction site ratio to train the SVMs on a data set composed of homo- and hetero-complexes. Only the predicted interaction site ratio slightly improved the predictions when the actual interaction site ratio was unknown. According to the ROC curves in their article, the recall increased about 8% when the precision was 60%, but no improvement was observed when the precision dropped to 50%. In the present study, by incorporating the structural conservation score into the SVM, the recall increased about 18% at precision 60% and increased 10% at 50% precision.

The B-factor and the Structural Conservation Score

Flexible regions in a protein tend to be aligned poorly in structure alignment compared to other regions and therefore have less reliable structural conservation scores. To take this into consideration, the raw structural conservation score was weighted by the normalized B-factor, reducing the conservation scores of the residues in the flexible regions (higher B-factors) and magnifying those in the rigid regions (lower B-factors). To test if the weighting was advantageous, two additional predictors were tested. Predictor 3 took predictor 1 and included the raw conservation score without the weighting. Predictor 4 included only the normalized B-factor and not the structural conservation score. The impact of the four predictors is given in Figure 4, with the best results achieved at the top left of the graph. Clearly the SVM using the weighted structural conservation score (predictor 2) out performed the other predictors. Improvement was only achieved at higher precisions when using the raw structural conservation score (predictor 3). On the other hand, the little improvement of predictor 4 also implied that the enhanced performance achieved by the weighted structural conservation score was not directly caused by the B-factor alone.

In order to validate our prediction, the data set used in this study was composed of individual components of the complexes. One may argue that the impact of the weighting came from the fact that generally the interface residues in a crystallized complex have lower B-factors than the rest of the surface. However, as observed by Neuvirth et al.,[28] the interface residues already have lower B-factors and have more bound water molecules in the unbound state. In addition, they observed the same biophysical properties for the interfaces in the bound and unbound states. This implies our data are suitable in evaluating the method for predicting the binding sites from a structure without knowing its binding partners.

The Predictions

Incorporating information on structurally conserved residues enhanced predictions for 58% of the 319 binding sites. As expected, many of these binding sites were located in or very close to the surfaces of relatively rigid regions, such as β strands. Examples of the binding sites identified by our approach (predictor 2) are presented in Figures 5 to 8. The predictions were also compared with those predicted by predictor 1 and by the ConSurf server.[25] As previously mentioned, predictor 1 used only sequence profile and accessible surface area, which are common input features used by previous studies.[31–35] ConSurf is a tool identifying the functionally important regions on the surface of a known 3D structure from the phylogenetic information derived from multiple sequence alignments. We used the default parameters provided by the sever but changed the maximum number of the homologues to 100, number of psi-blast interaction to 3, and chose maximum likelihood as the method of calculating conserved scores.

Figure 5 illustrates the predicted binding site on domain 1 of human coxsackie and adenovirus receptor (CAR D1) with the knob domain of the adenoviruses serotype 12 (Ad12) as the binding partner. CAR is an integral membrane protein expressed in a broad range of human and murine cell types. Domain 1 of CAR (CAR D1) is one of two extracellular domains mediating Ad and coxsackie virus B infections.[56] It adopts an immunoglobulin-like β-sandwich fold [Fig. 5(d)] and belongs to the V set domains (antibody variable domain-like) SCOP family. As shown in Figure 5(a), only about 10% of the interface residues on CAR D1 were identified by predictor 1. After incorporating the structural conservation score into the SVM, the prediction improved significantly, covering about 71% of the actual binding site [Fig. 5(b)]. The predicted residues located in the red circle were not bound to the partner and seemed to be false positives. However, Bewley et al.[56] pointed out that each CAR D1 domain actually binds at the interface between two adjacent Ad12 knob domains in the trimer consisting of the virus and virus receptor [Fig. 5(e)]. This is consistent given that most neutralizing antibodies to knob are directed against the trimer, rather than the monomer.[56,57] When the predicted residues in red circle in Figure 5(b) were mapped to the biologically active molecule of the complex, they were actually bound to another Ad12 knob domain [Fig. 5(e)], indicating a correct prediction. In this case, ConSurf [Fig. 5(c)] did not identify clustered conserved residues in any of the two binding regions.

A second example is the adrendoxin (Adx)-adrenodoxin reductase (AR) complex (Fig. 6). In the mitochondria of cells of the adrenal cortex the steroid hydroxylating system requires the transfer of electrons from the membrane-attached flavoprotein AR via the soluble Adx to the membrane-integrated cytochrome P450 of the CYP 11 family.[58,59] The predictor 1 has already predicted 56% of the total interface residues in the middle of the binding site [Fig. 6(a)]. However, those predicted residues did not cover the primary interaction region determined by the electrostatic analysis[59] (within the red circle). This region was identified and the total coverage reached 72% after incorporating the structurally conserved residues using predictor 2 [Fig. 6(b)]. Consurf also correctly predicted the binding residues on this protein but our predictions were more clustered toward the binding partner and have fewer false positives [Fig. 6(c)].

Apert syndrome (AS) is caused by substitution of one of two adjacent residues, Ser252Trp or Pro253Arg, in the fibroblast growth factor receptor 2 (FGFR 2).[60,61] These two mutations augment the affinity to the fibroblast growth factor 2 (FGF 2) by introducing additional interac-

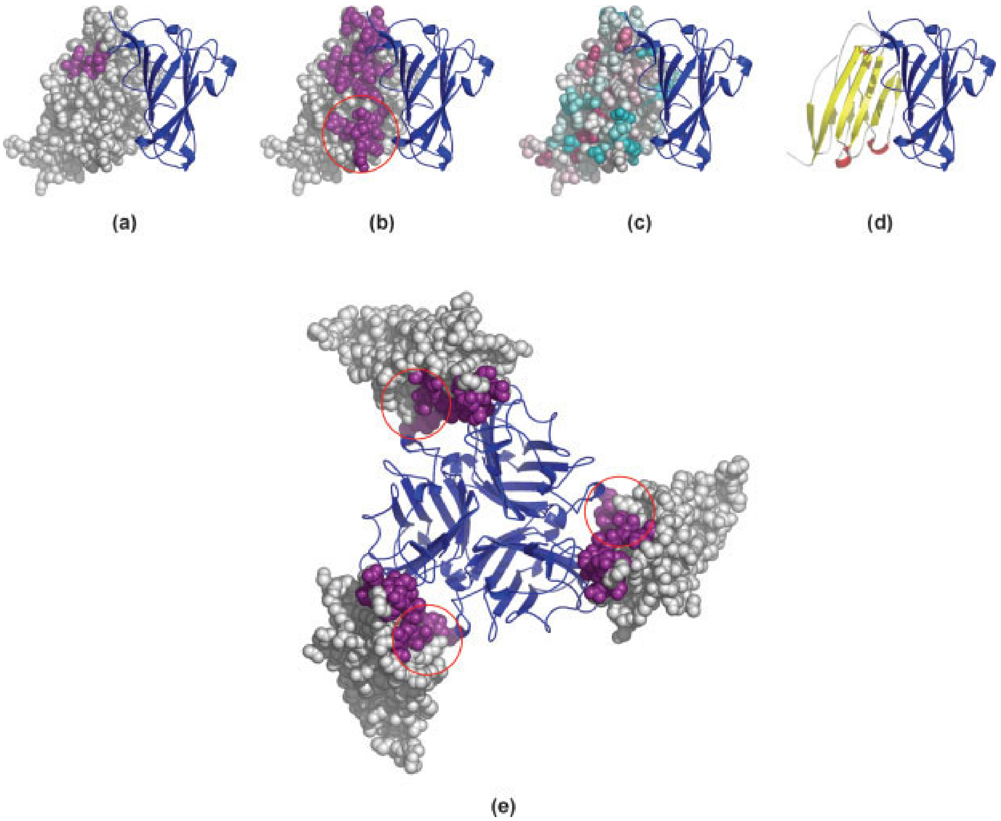

Fig. 5. Predicted interface residues (purple) for domain 1 of the human coxsackie and adenovirus receptor (CAR D1; PDB code 1KAC:B) identified by (a) predictor 1 and (b) predictor 2. The binding partner is the knob domain of the adenoviruses serotype 12 (Ad12; blue). (c) The conserved residues predicted by the ConSurf server are colored as a gradient between red (conserved) and dark cyan (variable). (d) The secondary structures of CAR D1 are colored red for helices and yellow for strands. (e) The interface residues identified by predictor 2 are mapped to the Ad12 knob–CAR D1 complex. The residues within the red circles in (b) and (e) are bound to other virus molecules in the complex.

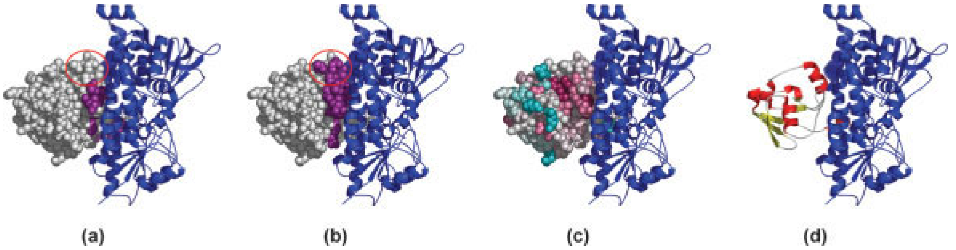

Fig. 6. Predicted interface residues (purple) for adrendoxin (Adx, PDB code 1E6E:B) identified by (a) predictor 1 and (b) predictor 2. The binding partner is adrenodoxin reductase (AR; blue). The interface residues defined by predictor 1 are in the middle of the binding site; however, they failed to cover the primary interaction region determined by electrostatic analysis (red circle). This region is identified when incorporating structurally conserved residues using predictor 2. (c) The conserved residues predicted by the ConSurf server are colored as a gradient between red (conserved) and dark cyan (variable). (d) The secondary structures of Adx are colored red for helices and yellow for strands.

tions but do not affect the relative positions of their two domains.[62] Figure 7 shows the Ser251Trp mutant of FGFR2. Predictor 2 successfully identified 74% of the binding residues, including the two positions that mutations occur [Fig. 7(b)]. ConSurf identified many potential binding residues as well. However, it is hard to define where the actual binding site is [Fig. 7(c)].

Improved predictions were also observed in proteins with multiple binding partners. Nitrogenase molybdenum-iron protein from *Clostridium pasteurianum* consists of tetramers with two α and two β subunits [Fig. 8(d)].[63] The interfaces on the β subunit chain B contacting the α subunit chain A was correctly predicted by both predictors (66% for predictor 1 and 70% for predictor 2). However, the interface contacting the α subunit chain D was only partially predicted (29%) by predictor 1 [Fig. 8(a)], whereas predictor 2 correctly (54%) identified this interface [Fig. 8(b)]. For this protein, ConSurf correctly identified the

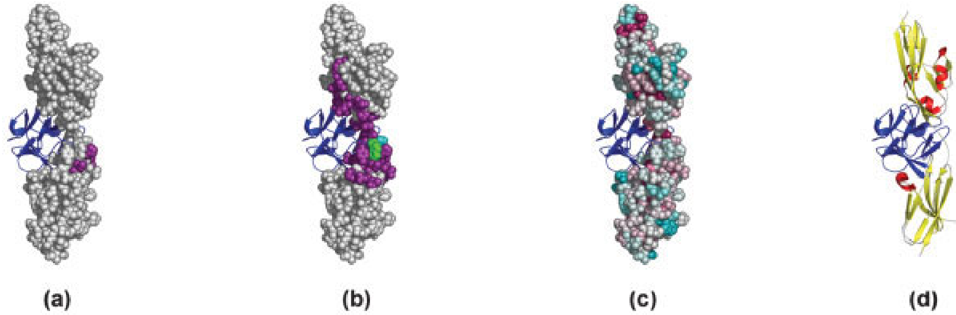

Fig. 7. Predicted interface residues (purple) for the Ser252Trp mutant of fibroblast growth factor receptor 2 (FGFR 2; PDB code 1II4:E) identified by (a) predictor 1 and (b) predictor 2. The binding partner is fibroblast growth factor 2 (FGF 2; blue). Substitution of one of the two adjacent residues, Ser252Trp (green) or Pro253Arg (cyan), in FGFR 2 causes Apert syndrome. (c) The conserved residues predicted by the ConSurf server are colored as a gradient between red (conserved) and dark cyan (variable). (d) The secondary structures of FGFR 2 are colored red for helices and yellow for strands.

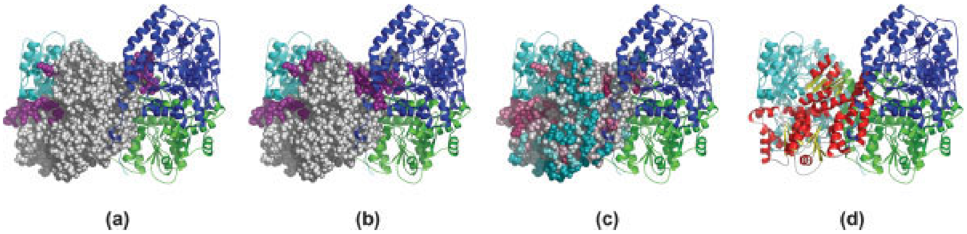

Fig. 8. Predicted interface residues (purple) for the nitrogenase molybdenum-iron protein of *Clostridium pasteurianum* β subunit (PDB code 1MIO:B) identified by (a) predictor 1 and (b) predictor 2. The two α subunits and the other β subunit are colored blue, cyan, and green, respectively. (c) The conserved residues predicted by the ConSurf server are colored as a gradient between red (conserved) and dark cyan (variable). (d) The secondary structures of the β subunit are colored red for helices and yellow for strands.

sites contacting the α subunit chain A but, like predictor 1, only partially identified a few residues contacting the α subunit chain D [Fig. 8(c)].

Twenty-five of the 319 binding sites were poorly predicted by predictor 2. For each of these sites, predictor 1 identified at least five more interface residues than predictor 2. Visual inspection revealed that many of these binding sites involved flexible loops. One example is the coat protein VP 1 in p1/mahoney poliovirus (1AL2: 1). The deteriorated prediction of the interfaces on VP 1 may be caused by loops that are in contact with the two other coat proteins, VP 2 and VP 3. In other case scenarios, poorly predicted binding sites resided in regions of poor structure alignment. For example, >30% of the residues of the methane monooxygenase hydroxylase α subunit (1FYZ: A, C) were in the gapped positions of its structure alignment. This led to an artificially low value for the structural conservation score. This begs the question, what is the overall impact of structure misalignments? To answer this question we undertook a simple empirical test. Three representative proteins (all alpha, all beta, alpha and beta) were chosen from the data set to test the effects of structure alignment. The manually curated structure alignments retrieved from the Homstrad database[64] were mapped to these proteins. The structural conservation score was derived from these alignments and sent to the predictor with the rest of the input features of predictor 2. It is hard to test the whole data set in this way because many Homstrad families do not have enough structures. Figures 9 to 11 compare the interface residues identified by the predictor using the CE-MC alignment versus that using the Homstrad curated alignment. Although the structure entries in the two alignments were not identical, the predicted results only showed slight variation. This indicates that our approach is not over sensitive to the details of the structure alignment and is able to produce consistent results.

In the present study, the precision is adjusted to 50% to locate binding sites. At this precision, the number of predicted interface residues was approximately 35% of the total surface residues in the data set. Increasing the cutoff will increase the coverage of the prediction (recall) but with a decreasing precision (more false positives) as a trade-off. As described above, incorporating the structural conservation score did improve the prediction of the binding sites (Table I and Fig. 4). Larger improvements are expected at higher precision (Fig. 4). It is not possible to distinguish false positives from sites that could be bound to an unidentified binding partner. For example, some false positives in the CAR D1 [Fig. 5(b)] were actually bound to the other knob domain and should be counted as true positives.

The structural conservation score, which measures the residue conservation in 3D space, is a major factor in improving the predictions. A better scoring function or weighting function could be developed in the future.

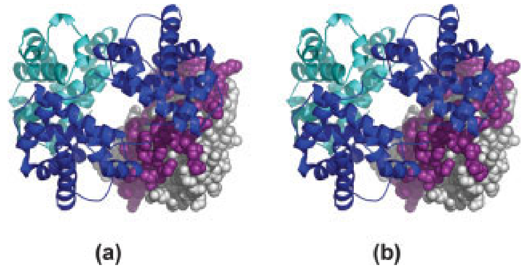

Fig. 9. Predicted interface residues (purple) for the hemoglobin β chain (PDB code 1ABW:B) identified by predictor 2 with structure alignments provided by (a) CE-MC, (b) Homstrad.

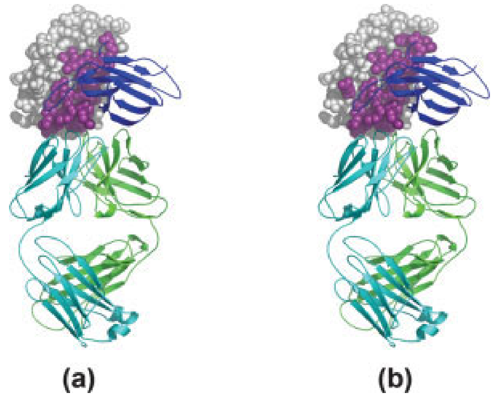

Fig. 10. Predicted interface residues (purple) for the murine T-cell receptor variable domain (PDB code 1KB5:A) identified by predictor 2 with structure alignments provided by (a) CE-MC, (b) Homstrad.

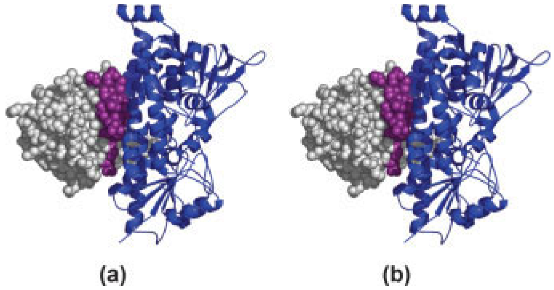

Fig. 11. Predicted interface residues (purple) for adrendoxin (Adx, PDB code 1E6E:B) identified by predictor 2 with structure alignments provided by (a) CE-MC, (b) Homstrad.

However, the main focus of our study is to test if the information provided by structurally conserved residues can help the prediction. As such, this study represents an initial trial that exploits multiple structure alignments on a large scale for the prediction of functional regions.

A limitation of our method is requiring structure homologs be available. To assess this limitation we surveyed the 1824 nonredundant structures with at least 80 amino acids obtained form the PDB-select database[65] (25% list, Oct, 2004). Forty-five percent of these structures are X-ray structures with <3.5 Å resolution and have at least three nonredundant structure homologs with SCOP class <8 and <3.5 Å resolution. This number is expected to increase as more and more protein structures are determined by structural genomics and traditional structure determination projects. Thus this approach already has the potential to have significant impact since of the 45% identified many have no or incomplete experimental binding sites information available.

CONCLUSION

We have surveyed the structurally conserved residues in 274 nonredundant chains of hetero-complexes. These conserved residues were defined using a structural conservation score derived from multiple structure alignments followed by a weighting process using normalized B-factor. We have found that, although derived from the alignments of structures without knowing their binding partners in advance, these structurally conserved residues did distinguish the protein interface from the rest of the surface. An SVM was then trained to predict the interface residues by incorporating the structural conservation score with the accessible surface area and sequence profile of neighboring residues as inputs. At a precision of 50%, the number of correctly predicted binding sites increased 17%, and the number of precisely predicted binding sites increased 13% when incorporating the information from the structural conservation. In total, 52% of the binding sites were precisely predicted, 77% of the binding sites were correctly predicted, and 21% of the binding sites were partially predicted. A manual investigation revealed that many of these improvements occurred at the binding sites located on or very close to the surfaces of relatively rigid regions, somewhat surprising perhaps, because it could be argued that flexibility is required for recognition. It appears that, on balance, rigidity is preferred and is assumed to minimize the entropic cost upon complex formation. Details of the impact of rigidity require further study. This study indicates the feasibility of using information derived from multiple structure alignments on a large scale. Moreover, the method proposed here could be used to guide site-specific mutagenesis experiments or combined with docking procedures to limit the search space.[15]

ACKNOWLEDGMENTS

The authors would like to thank Dr. Chittibabu Guda for his help on using the CE-MC program. They would also like to thank Dr. Boojala V. B. Reddy, Dr. C. Y. Kao, and the entire Bourne group for helpful discussions. This work was supported by the National Institutes of Health grant GM63208 and GM08326.

REFERENCES

1. Burley SK, Almo SC, Bonanno JB, Capel M, Chance MR, Gaasterland T, Lin D, Sali A, Studier FW, Swaminathan S. Structural genomics: beyond the human genome project. Nat Genet 1999;23(2): 151–157.
2. Grant A, Lee D, Orengo C. Progress towards mapping the universe of protein folds. Genome Biol 2004;5(5):107.
3. Shapiro L, Harris T. Finding function through structural genomics. Curr Opin Biotechnol 2000;11(1):31–35.
4. Jones S, Thornton JM. Analysis of protein–protein interaction sites using surface patches. J Mol Biol 1997;272(1):121–132.
5. Lo Conte L, Chothia C, Janin J. The atomic structure of protein–protein recognition sites. J Mol Biol 1999;285(5):2177–2198.
6. Ofran Y, Rost B. Analysing six types of protein–protein interfaces. J Mol Biol 2003;325(2):377–387.

7. Wells JA. Systematic mutational analyses of protein–protein interfaces. Methods Enzymol 1991;202:390–411.
8. DeLano WL. Unraveling hot spots in binding interfaces: progress and challenges. Curr Opin Struct Biol 2002;12(1):14–20.
9. Clackson T, Wells JA. A hot spot of binding energy in a hormone-receptor interface. Science 1995;267(5196):383–386.
10. Bogan AA, Thorn KS. Anatomy of hot spots in protein interfaces. J Mol Biol 1998;280(1):1–9.
11. Jones S, Thornton JM. Principles of protein–protein interactions. Proc Natl Acad Sci USA 1996;93(1):13–20.
12. Laskowski RA, Luscombe NM, Swindells MB, Thornton JM. Protein clefts in molecular recognition and function. Protein Sci 1996;5(12):2438–2452.
13. Hu Z, Ma B, Wolfson H, Nussinov R. Conservation of polar residues as hot spots at protein interfaces. Proteins 2000;39(4):331–342.
14. Ma B, Elkayam T, Wolfson H, Nussinov R. Protein–protein interactions: structurally conserved residues distinguish between binding sites and exposed protein surfaces. Proc Natl Acad Sci USA 2003;100(10):5772–5777.
15. Wodak SJ, Mendez R. Prediction of protein–protein interactions: the CAPRI experiment, its evaluation and implications. Curr Opin Struct Biol 2004;14(2):242–249.
16. Halperin I, Ma B, Wolfson H, Nussinov R. Principles of docking: an overview of search algorithms and a guide to scoring functions. Proteins 2002;47(4):409–443.
17. Smith GR, Sternberg MJ. Prediction of protein–protein interactions by docking methods. Curr Opin Struct Biol 2002;12(1):28–35.
18. Aloy P, Russell RB. Interrogating protein interaction networks through structural biology. Proc Natl Acad Sci USA 2002;99(9):5896–5901.
19. Lu L, Lu H, Skolnick J. MULTIPROSPECTOR: an algorithm for the prediction of protein–protein interactions by multimeric threading. Proteins 2002;49(3):350–364.
20. Aloy P, Ceulemans H, Stark A, Russell RB. The relationship between sequence and interaction divergence in proteins. J Mol Biol 2003;332(5):989–998.
21. Lichtarge O, Bourne HR, Cohen FE. An evolutionary trace method defines binding surfaces common to protein families. J Mol Biol 1996;257(2):342–358.
22. Landgraf R, Xenarios I, Eisenberg D. Three-dimensional cluster analysis identifies interfaces and functional residue clusters in proteins. J Mol Biol 2001;307(5):1487–1502.
23. Lichtarge O, Sowa ME. Evolutionary predictions of binding surfaces and interactions. Curr Opin Struct Biol 2002;12(1):21–27.
24. Pupko T, Bell RE, Mayrose I, Glaser F, Ben-Tal N. Rate4Site: an algorithmic tool for the identification of functional regions in proteins by surface mapping of evolutionary determinants within their homologues. Bioinformatics 2002;18 Suppl 1:S71–S77.
25. Glaser F, Pupko T, Paz I, Bell RE, Bechor-Shental D, Martz E, Ben-Tal N. ConSurf: identification of functional regions in proteins by surface-mapping of phylogenetic information. Bioinformatics 2003;19(1):163–164.
26. Pazos F, Valencia A. In silico two-hybrid system for the selection of physically interacting protein pairs. Proteins 2002;47(2):219–227.
27. Gallet X, Charloteaux B, Thomas A, Brasseur R. A fast method to predict protein interaction sites from sequences. J Mol Biol 2000;302(4):917–926.
28. Neuvirth H, Raz R, Schreiber G. ProMate: a structure based prediction program to identify the location of protein–protein binding sites. J Mol Biol 2004;338(1):181–199.
29. Jones S, Thornton JM. Prediction of protein–protein interaction sites using patch analysis. J Mol Biol 1997;272(1):133–143.
30. Sheinerman FB, Honig B. On the role of electrostatic interactions in the design of protein–protein interfaces. J Mol Biol 2002;318(1):161–177.
31. Koike A, Takagi T. Prediction of protein–protein interaction sites using support vector machines. Protein Eng Des Sel 2004;17(2):165–173.
32. Yan C, Dobbs D, Honavar V. A two-stage classifier for identification of protein–protein interface residues. Bioinformatics 2004;20 (Suppl 1):I371–I378.
33. Ofran Y, Rost B. Predicted protein–protein interaction sites from local sequence information. FEBS Lett 2003;544(1–3):236–239.
34. Fariselli P, Pazos F, Valencia A, Casadio R. Prediction of protein–protein interaction sites in heterocomplexes with neural networks. Eur J Biochem 2002;269(5):1356–1361.
35. Zhou HX, Shan Y. Prediction of protein interaction sites from sequence profile and residue neighbor list. Proteins 2001;44(3):336–343.
36. Berman HM, Westbrook J, Feng Z, Gilliland G, Bhat TN, Weissig H, Shindyalov IN, Bourne PE. The Protein Data Bank. Nucleic Acids Res 2000;28(1):235–242.
37. Kabsch W, Sander C. Dictionary of protein secondary structure: pattern recognition of hydrogen-bonded and geometrical features. Biopolymers 1983;22(12):2577–2637.
38. Murzin AG, Brenner SE, Hubbard T, Chothia C. SCOP: a structural classification of proteins database for the investigation of sequences and structures. J Mol Biol 1995;247(4):536–540.
39. Lo Conte L, Brenner SE, Hubbard TJ, Chothia C, Murzin AG. SCOP database in 2002: refinements accommodate structural genomics. Nucleic Acids Res 2002;30(1):264–267.
40. McGinnis S, Madden TL. BLAST: at the core of a powerful and diverse set of sequence analysis tools. Nucleic Acids Res 2004; 32(Web Server issue):W20–W25.
41. Brenner SE, Koehl P, Levitt M. The ASTRAL compendium for protein structure and sequence analysis. Nucleic Acids Res 2000; 28(1):254–256.
42. Chandonia JM, Hon G, Walker NS, Lo Conte L, Koehl P, Levitt M, Brenner SE. The ASTRAL Compendium in 2004. Nucleic Acids Res 2004;32 Database issue:D189–D192.
43. Guda C, Scheeff ED, Bourne PE, Shindyalov IN. A new algorithm for the alignment of multiple protein structures using Monte Carlo optimization. Pac Symp Biocomput 2001:275–286.
44. Guda C, Lu S, Scheeff ED, Bourne PE, Shindyalov IN. CE-MC: a multiple protein structure alignment server. Nucleic Acids Res 2004;32(Web Server issue):W100–103.
45. Shindyalov IN, Bourne PE. Protein structure alignment by incremental combinatorial extension (CE) of the optimal path. Protein Eng 1998;11(9):739–747.
46. Rost B, Sander C. Conservation and prediction of solvent accessibility in protein families. Proteins 1994;20(3):216–226.
47. Drenth J. Principles of protein X-ray crystallography. New York: Springer-Verlag; 1994. p xiii, 305.
48. Vihinen M, Torkkila E, Riikonen P. Accuracy of protein flexibility predictions. Proteins 1994;19(2):141–149.
49. Smith DK, Radivojac P, Obradovic Z, Dunker AK, Zhu G. Improved amino acid flexibility parameters. Protein Sci 2003;12(5):1060–1072.
50. Valdar WS. Scoring residue conservation. Proteins 2002;48(2):227–241.
51. Vapnik VN. The nature of statistical learning theory. New York: Springer; 1995. p xv, 188.
52. Schölkopf B, Burges CJC, Smola AJ. Advances in kernel methods: support vector learning. Cambridge, MA: MIT Press; 1999. p vii, 376.
53. Altschul SF, Madden TL, Schaffer AA, Zhang J, Zhang Z, Miller W, Lipman DJ. Gapped BLAST and PSI-BLAST: a new generation of protein database search programs. Nucleic Acids Res 1997; 25(17):3389–3402.
54. Leibowitz N, Nussinov R, Wolfson HJ. MUSTA—a general, efficient, automated method for multiple structure alignment and detection of common motifs: application to proteins. J Comput Biol 2001;8(2):93–121.
55. Leibowitz N, Fligelman ZY, Nussinov R, Wolfson HJ. Automated multiple structure alignment and detection of a common substructural motif. Proteins 2001;43(3):235–245.
56. Bewley MC, Springer K, Zhang YB, Freimuth P, Flanagan JM. Structural analysis of the mechanism of adenovirus binding to its human cellular receptor, CAR. Science 1999;286(5444):1579–1583.
57. Fender P, Kidd AH, Brebant R, Oberg M, Drouet E, Chroboczek J. Antigenic sites on the receptor-binding domain of human adenovirus type 2 fiber. Virology 1995;214(1):110–117.
58. Grinberg AV, Hannemann F, Schiffler B, Muller J, Heinemann U, Bernhardt R. Adrenodoxin: structure, stability, and electron transfer properties. Proteins 2000;40(4):590–612.
59. Muller JJ, Lapko A, Bourenkov G, Ruckpaul K, Heinemann U. Adrenodoxin reductase-adrenodoxin complex structure suggests

electron transfer path in steroid biosynthesis. J Biol Chem 2001;276(4):2786–2789.

60. Lajeunie E, Cameron R, El Ghouzzi V, de Parseval N, Journeau P, Gonzales M, Delezoide AL, Bonaventure J, Le Merrer M, Renier D. Clinical variability in patients with Apert's syndrome. J Neurosurg 1999;90(3):443–447.

61. Slaney SF, Oldridge M, Hurst JA, Moriss-Kay GM, Hall CM, Poole MD, Wilkie AO. Differential effects of FGFR2 mutations on syndactyly and cleft palate in Apert syndrome. Am J Hum Genet 1996;58(5):923–932.

62. Ibrahimi OA, Eliseenkova AV, Plotnikov AN, Yu K, Ornitz DM, Mohammadi M. Structural basis for fibroblast growth factor receptor 2 activation in Apert syndrome. Proc Natl Acad Sci USA 2001;98(13):7182–7187.

63. Kim J, Woo D, Rees DC. X-ray crystal structure of the nitrogenase molybdenum-iron protein from Clostridium pasteurianum at 3.0-A resolution. Biochemistry 1993;32(28):7104–7115.

64. Mizuguchi K, Deane CM, Blundell TL, Overington JP. HOMSTRAD: a database of protein structure alignments for homologous families. Protein Sci 1998;7(11):2469–2471.

65. Hobohm U, Sander C. Enlarged representative set of protein structures. Protein Sci 1994;3(3):522–524.