# Statistical Analysis and Prediction of Protein–Protein Interfaces

**Andrew J. Bordner*** **and Ruben Abagyan**
*Molsoft LLC, San Diego, California*

***ABSTRACT*** **Predicting protein–protein interfaces from a three-dimensional structure is a key task of computational structural proteomics. In contrast to geometrically distinct small molecule binding sites, protein–protein interface are notoriously difficult to predict. We generated a large nonredundant data set of 1494 true protein–protein interfaces using biological symmetry annotation where necessary. The data set was carefully analyzed and a Support Vector Machine was trained on a combination of a new robust evolutionary conservation signal with the local surface properties to predict protein–protein interfaces. Fivefold cross validation verifies the high sensitivity and selectivity of the model. As much as 97% of the predicted patches had an overlap with the true interface patch while only 22% of the surface residues were included in an average predicted patch. The model allowed the identification of potential new interfaces and the correction of mislabeled oligomeric states. Proteins 2005;60:353–366.** © 2005 Wiley-Liss, Inc.

**Key words: protein interactions; dimerization; binding sites; protein surface annotation; statistical tests; evolutionary conservation; Support Vector Machines**

## INTRODUCTION

One fundamental goal of molecular biology is the discovery of all protein–protein interactions in an organism as well as their biochemical and biological functions. Analysis of the complete genomes that are available for many organisms provides a tentative list of participating proteins and high-throughput methods, such as yeast two-hybrid screens and mass spectrometry of coimmunoprecipitated complexes, provide evidence for specific protein interactions. However the structural details of protein interactions at the atomic level, which are essential for understanding their function and for designing drugs that modulate their interactions, are only provided by X-ray crystallographic or NMR structures of the complexes.

Predicting which residues participate in protein–protein interactions is useful since is suggests interface residues for experimental verification using mutational analysis and allows the virtual screening of ligands to alter the interaction for therapeutics discovery. The question of what properties of protein–protein interfaces differentiate them from noninterface surface regions is a necessary first step towards their computational prediction.

The physical and chemical properties of protein interfaces in known structures of complexes have been studied to determine their distinguishing features. Whereas homodimers, which often form permanent complexes, have been found to have predominantly hydrophobic interfaces, heterocomplexes were found to have interfaces whose hydrophobicity is indistinguishable from the remainder of the surface.[1–3] The frequencies of residue types, weighted by their accessible surface areas, for interfaces were compared with the whole surface in Jones and Thornton.[4] Large hydrophobic and uncharged polar residues were more prevalent in interfaces and charged residues less prevalent. Heterocomplexes were also found to have less hydrophobic and more polar residues than homodimers. An analysis of surface patches in Jones and Thornton[2] also found that interface surface patches tend to be planar, protrude from the surface, and have residues with higher solvent-accessible surface area (SASA) than other surface patches, at least for particular classes of complexes. One conclusion from that study is that no single physical property definitively distinguishes the interface for all classes of protein complexes. Geometric and electrostatic complementarity was also found to be important for some protein–protein interactions.[3,5–8] Although interfaces tended to have an area proportional to the total protein surface area,[4] and may be large, only a small fraction of interface residues are observed to make large contributions to the binding energy.[9,10]

The evolutionary conservation of residues is another property that may be utilized for predicting protein–protein interfaces. Although the residue conservation at the interface was not found to be significantly different from that in the protein interior,[11] is was observed to be slightly higher as compared with the surface residues.[11–13] Residues involved in functionally important protein–protein binding in a given protein family are expected to be conserved; however interior residues that contribute to efficient folding and stability are also generally more conserved than exposed surface residues. Therefore a strong conservation signal is only observed when the

The Supplementary Materials referred to in this article can be found at http://www.interscience.com/jpages/0887-3585/suppmat/

*Correspondence to Andrew Bordner, Computer Science and Mathematics Division, Oak Ridge National Laboratory, P.O. Box 2008, MS 6173, Oak Ridge, TN 37831. E-mail: bordner@ornl.gov

Received 1 September 2004; Revised 23 November 2004; Accepted 2 December 2004.

Published online 19 May 2005 in Wiley InterScience (www.interscience.wiley.com). DOI: 10.1002/prot.20433

residue conservation of the interface is compared with that of the surface. Although a simple definition of residue conservation that does not account for residue substitutions or evolutionary relationships may be adequate for detecting residues in an enzyme active site,[11] a method that allows for substitutions of residues with the same physicochemical properties (e.g., hydrophobic, polar, or charged) or that uses an evolutionary tree are needed to detect the weaker interface conservation signal. In order to estimate the evolutionary conservation of each site we use a novel Bayesian method, Robust EVolutionary Conservation Measure (REVCOM), that employs phylogenetic trees to calculate evolutionary rates. The rates calculation also uses misalignment probabilities to reduce errors from distantly related sequences.[14]

Although a number of methods have been described to predict protein–protein interfaces using only protein sequences, we consider only methods utilizing the structure of one protein. Of course the latter class of methods are expected to be more accurate when the structural information is available. Early methods identified hydrophobic surface regions as interaction interfaces.[1,15] Jones and Thornton[16] used a linear combination of scores for physical and geometrical properties and residue propensity to predict surface patches that maximally overlap the interface. Since the method was evaluated only by examining the overlap of the highest scoring patches it was presented as a preliminary step toward a complete interface prediction method. A recent paper describes the Optimal Desolvation Area method which identifies surface patches with minimum desolvation energy.[17] Patches with significantly low energy were mostly located in or near protein–protein binding sites. The Evolutionary Trace Method[18] uses evolutionary conservation determined from a multiple sequence alignment and associated phylogenetic tree to predict interaction interfaces. A conservation rank is assigned to each residue, based on the maximum tree depth for which the residue is absolutely conserved. Other prediction methods based on conservation emphasize the importance of evolutionary distance[19] and spatial proximity.[20] A sensitive and selective protein–protein interface prediction method should incorporate all possible discriminating factors utilized by these methods: physicochemical properties, residue type distribution, and evolutionary conservation.

Machine Learning Methods are well suited to the classification of interface and noninterface surface residues. Artificial Neural Networks trained on residue frequencies in a multiple alignment[21] or sequence profile combined with solvent accessibility[22] have been used to predict protein–protein interfaces. A paper by Koike and Takagi[23] described a prediction method employing Support Vector Machines trained on residue frequencies. We also use a Support Vector Machine for prediction but we include several improvements. First, unlike the previous study, we use information contained in the Protein Data Bank file and Swiss-Prot annotation to select a set of biologically relevant protein–protein interfaces for training and validation. Secondly, data for noninterface residues is not removed from the data set. Removing the data reduces the number of false positives in the cross validation which provides a biased measure of accuracy since the identity of noninterface residues is not known beforehand for an actual prediction. Finally, we use residue-type distribution Z-scores rather than residue frequencies as input data. This is because Z-scores are approximately identically distributed and SVM prediction works best with such data. In contrast, residue frequencies vary; for example, the mean frequency of alanine is higher than that of tryptophan.

First we analyze differences in the properties of interfaces compared with the noninterface protein surface. The hydrophobicity, solvation energy, solvent accessible surface area, residue type distribution, and evolutionary rates are compared. Next the residue composition and conservation of surface patches are used to train a Support Vector Machine (SVM) for predicting interface residues. It is shown that although the distribution of residue types in protein–protein interfaces contributes to successful predictions, most of the discrimination signal is present in the lower evolutionary rates of interface residues. The prediction method is evaluated by cross validation results on a large set of dimer interfaces as well as a smaller set of transient heterodimers. Finally, a histogram method for estimating the prediction reliability is introduced.

## MATERIALS AND METHODS

### Protein–Protein Interface Data Sets

A data set of protein intermolecular interfaces in complexes was compiled from X-ray crystal structures in the Protein Data Bank (PDB)[24] archive using the ICM scripting language.[25] Biological unit information from PDB files in mmCIF format was used to generate the structure of the complex. Complexes whose corresponding Swiss-Prot[26] entry subunit annotation was either "monomer," "homodimer," or "heterodimer" and which disagreed with the PDB information were corrected or removed after consulting the literature in order to improve the quality of the data set. Two proteins in a complex were considered interacting pairs if nonhydrogen atoms in each molecule are separated by $< 4$ Å and pairs containing short chains ($< 20$ residues) were removed. In order to remove interfaces for homologous proteins, protein interaction pairs were then clustered such that each cluster contained pairs in which both protein sequences in a pair share less $< 30\%$ sequence identity with any other pair in the cluster. Only the highest quality structure, with the least missing coordinates and the highest resolution, from each cluster was included in the data set. Alignments of the protein sequences in each cluster to a representative sequence were used to compare interface residues. Two interfaces on a protein were considered distinct if their residue sets, referred to the representative sequence, overlapped less than 20%. Interfaces including proteins that are annotated in the PDB file as having mutations and immune system proteins that are highly polymorphic or undergo somatic mutation, namely MHC, T-cell receptors, and antibodies, were excluded. Finally, only interfaces contain-

ing at least ten residues were included in the set because of the large variance of the residue-based statistics in Analysis of Protein Interfaces in Results and Discussion below as well as difficulties in validating the interface prediction for smaller interfaces. Since the protein–protein interfaces are considered as surface regions on individual proteins, both orderings of the nonidentical interacting proteins were considered distinct. The resulting set had 1494 protein–protein interfaces (1143 unordered pairs), of which 518 were homodimers, 114 were heterodimers, and the remaining 862 were multimers.

### Hydrophobicity, Solvation Energy, and Average SASA

The protein–protein interfaces were analyzed using three descriptors related to the solvation surface: the residue-based hydrophobicity, an atom-based desolvation energy, and the average residue SASA. The average hydrophobicity for a region (interface or noninterface) was calculated from the Kyte-Doolittle hydrophobicity indices,[27] $H_i$, weighted by the SASA $A_i^{res}$ of residue *i:*

$$E_{hp} = \left( \sum_{i=1}^{N_{res}} H_i A_i^{res} \right) / \sum_i A_i^{res}. \quad (1)$$

The SASA for all properties were calculated using the structure of the isolated protein, with its partners in the complex removed. The average solvation energy was calculated in a similar manner using the atomic solvation parameters $\sigma_i$ of Wesson and Eisenberg[28] and the atomic SASA $A_i^{atom}$:

$$E_{solv} = \left( \sum_{i=1}^{N_{atom}} \sigma_i A_i^{atom} \right) / \sum_i A_i^{atom}. \quad (2)$$

The average SASA for a region with $N_{res}$ residues was calculated by

$$A_{avg} = \frac{1}{N_{res}} \sum_{i=1}^{N_{res}} A_i^{res}. \quad (3)$$

### Statistical Tests

Statistical tests were used to determine whether differences in the properties between the interface and noninterface regions significantly deviated from those expected from a uniform distribution on the protein surface. The null model for residue propensity is an independent hypergeometric distribution for each residue type. This is the distribution that would result from a random arrangement of the residue types on the surface. The distribution function $F(x)$ for $x$, the number of interface residues of type *i,* is then

$$F(x) = \sum_{j=0}^{x} \frac{\binom{N_i}{j}\binom{N-N_i}{I-j}}{\binom{N}{I}} \quad (4)$$

with $N$ the total number of surface residues, $N_i$ the number of surface residues of type *i,* and $I$ the number of residues in the interface.

The confidence interval for the difference, $f_1 - f_2$, in the fractions of homodimeric and heterodimeric interfaces with propensity for a given residue type was calculated using a normal distribution with standard deviation $\sqrt{f_1(1-f_1)/n_1 + f_2(1-f_2)/n_2}$, with $n_1$ and $n_2$ the respective total numbers of interfaces.

All other statistical tests used in this study are nonparametric since the underlying distribution of the quantities is unknown and likely non-normal. The Wilcoxon signed rank test was used to compare the solvation energies, hydrophobicities, and average SASA and the Mann-Whitney-Wilcoxon rank sum test was used to compare the evolutionary rates for the interface and noninterface regions.

### Multiple Sequence Alignments

Similar sequences were collected using a BLAST search[29] of the NCBI nr database with an E-value cutoff of 0.1 and sequences with greater than 90% identity to another sequence in the set were iteratively removed. The ClustalW program[30] with default alignment parameters (Gonnet 250 scoring matrix with gap opening = 10 and gap extension = 0.1) was used to align the remaining sequences.

### Evolutionary Conservation

Evolutionary site rates were calculated using a novel Bayesian algorithm which we briefly describe.[14] First a phylogenetic tree is generated from the alignment using the neighbor-joining algorithm[31] as implemented in the Quicktree[32] program. A homogeneous Markov model of residue substitution based on the JTT matrices[33] is assumed. Pairwise PAM distances are then calculated by inverting the expression for the expected fraction of identical residues $q(t)$

$$q(t) = \sum_{i=1}^{20} (M(t))_{ii} f_i \quad (5)$$

where $M(t) = \exp[t \log M(1)]$ is the JTT matrix for PAM distance $t$ and $f_i$ are the residue occurrence probabilities. Next the branch lengths are estimated using a weighted least squares approach.[34] The $\alpha$ parameter of a gamma prior site rates distribution is estimated using the maximum likelihood method.[35,36] Finally, the site rates are calculated as the average of the posterior distribution

$$p(r \mid \vec{x}_m) = \frac{p(\vec{x}_m \mid r) f(\hat{\alpha}, r)}{\int_0^\infty p(\vec{x}_m \mid r) f(\hat{\alpha}, r) dr}, \quad (6)$$

in which $\vec{x}_m$ is the vector of residues in the multiple alignment column (gaps are ignored) and $f(\hat{\alpha}, r)$ is the gamma rates distribution for rate $r$ with the estimated parameter $\hat{\alpha}$. An alignment reliability correction based on BLAST p-values is also implemented to correct for alignment errors in distantly related sequences.

## Support Vector Machines

Support Vector Machines are a class of effective supervised learning methods that balance the number of training errors with generalization error due to overfitting.[37–39] A recent implementation of SVM in the ICM software package was used.[25] A Gaussian kernel function, $k(\vec{x}, \vec{y}) = \exp(-\gamma \|\vec{x} - \vec{y}\|^2)$, was utilized since it performed better than linear, quadratic, and cubic kernels. The training data is unbalanced since only about 15% of the surface residues are in the interface. The SVM algorithm maximizes the prediction accuracy so that training on the unbalanced set yields a prediction with a low recall, that is, the model predicts noninterface residues more accurately than interface residues. Two different regularization constants, $C_+$ and $C_-$, for interface and noninterface data, respectively, were introduced into the SVM algorithm in order to remedy this problem.[40]

### *SVM Training*

First local surface patches containing 15 residues were calculated for each surface residue. These were defined by the central residue and the closest 14 surface residues as defined by the distances between $C_\alpha$ atoms. The patch was iteratively extended with only the closest residue to the current edge being added. This was done in order to avoid a discontinuous patch that spans the interior. Next the number of residues of each type in the multiple alignment columns corresponding to the patch residues were calculated using only closely related sequences with p-value $< 10^{-4}$. The resulting residue frequencies as well as the evolutionary rates were then converted into Z-scores, $Z = (x - \mu)/\sigma$, with $\mu$ and $\sigma$ the empirical mean and standard deviation of the values for the given protein. In other words

$$\mu = \frac{1}{N_{surf}} \sum_{i=1}^{N_{surf}} x_i, \tag{7}$$

$$\sigma = \sqrt{\left( \frac{1}{N_{surf}} \sum_{i=1}^{N_{surf}} x_i^2 \right) - \mu^2} \tag{8}$$

where the sums are over all $N_{surf}$ surface residues for the protein and $x_i$ is either one of the residue frequencies or the evolutionary rate for residue *i*. The 20 residue-type Z-scores and rates Z-scores for all 15 residues in the patch were concatenated to form a 315 component vector of the training data. The residue-type Z-scores and rates Z-scores were then separately normalized to the range $[-1, 1]$.

Only the dimer interfaces from the interface data set were used for SVM prediction in order to simplify analysis. Proteins which also interact with proteins not present in the X-ray structure were removed leaving 632 dimer interfaces. The prediction accuracy was evaluated using fivefold cross validation. The complete data set is divided into five approximately equal parts with predictions made for each part in turn using an SVM trained on the remaining data. The prediction model was thereby evaluated on data not used for training the SVM. The SVM was trained to predict the class, interface or noninterface, of all surface residues in the training set.

Predictions for different values of the parameters $\gamma$, $C_-$, and $C_+/C_-$ were calculated on a grid in parameter space and the fivefold cross validation F1 parameter was optimized on approximately one quarter of the data in order to find the best parameter values. The normalization of the rates Z-scores relative to the residue Z-scores was also varied in the optimization procedure. This normalization affects the prediction results since larger magnitude components in the training vectors have a greater influence on the prediction result. The best values obtained were $\gamma = 0.2$, $C_- = 3.0$, $C_+/C_- = 5.0$ and a rates scale factor of 2.0.

### *Post-Processing to Remove Predicted Patches*

Small predicted interface patches are likely to be either small molecule binding sites or isolated false positive predictions since protein–protein binding interfaces are generally composed of large contiguous patches. Furthermore a maximum interface size criterion has been shown to effectively discriminate between interfaces that are biologically relevant and those due to crystal packing.[12,41] Although interface size information can not be included in the SVM input data, since this data relates only to residues in a 15-residue patch, it is possible to remove small interface patches from the SVM prediction.

Predicted interface residues were clustered into contiguous patches defined by all residues whose $C_\alpha$ atoms are separated by less than 6 Å and predicted interface patches with SASA $< 150$ Å$^2$ were redefined as noninterface residues.

### *Histogram Method for Estimating Prediction Reliability*

The classifier SVM used in this study provides only a binary output representing to which class the residue belongs, interface or noninterface. It would be useful to also have a continuous measure of the reliability of each prediction, i.e., the degree of certainty that the particular residue belongs to its predicted class. Then one could, for example, select only the most reliably predicted residues for experimental verification. This may be approached by assuming that predictions for data points whose normal distance to the decision hyperplane in feature space are more certain than for points near the hyperplane. In order to implement this idea, first histograms of the normal distances (200 bins) were calculated separately for interface and noninterface data. Likelihood ratios were then calculated for each bin by first normalizing the histograms to unit area under the graph so that they are estimates for the probability density functions. These normalized histograms are shown in Figure 1. The likelihood ratios calculated for each bin are then estimates of p(point with normal distance in bin $n$ is in interface)/p(point with normal distance in bin $n$ is in noninterface). A high likelihood ratio therefore indicates a reliable prediction that the residue belongs to the interface and a low ratio indicates a reliable prediction that it does not, with higher uncertainty attributed to data with intermediate values.

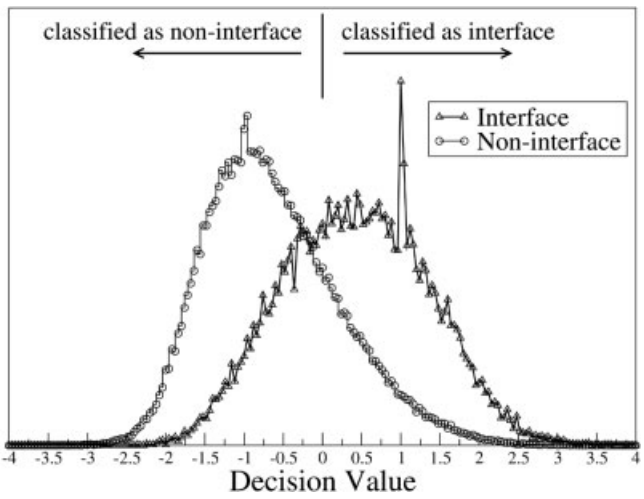


Fig. 1. Normalized histograms of the decision values, or hyperplane normal coordinates, for data corresponding to interface and noninterface residues in the training set of 438 dimers. Residues with positive decision values are predicted as interface residues and those with negative decision values are predicted as noninterface residues. The peaks at $\pm 1$ are general features resulting from the SVM optimization criterion. The distribution of decision values for interface residues is shifted toward positive values compared with the distribution for noninterface residues, as expected for a predictive model. The probability distribution functions for the two classes of residues are estimated from these histograms. Their ratio is an indicator of prediction reliability.

## RESULTS AND DISCUSSION

### Analysis of Protein Interfaces

Histograms of the number of residues in the interface and the fraction of surface area included in the interface are shown in Figure 2(a,b), respectively, for the complete set of interfaces. These show that a typical interface contains 10–30 residues which comprise 5–25% of the total surface area of the protein.

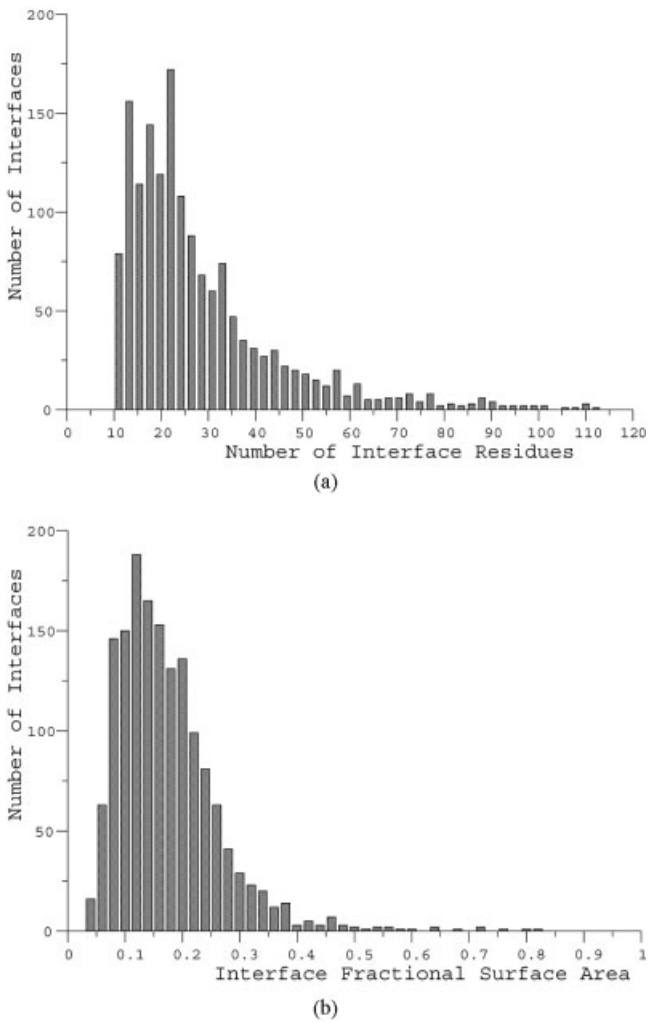


Fig. 2. Histograms of (**a**) the number of residues in the interface and (**b**) the fraction of surface area in the interface for the complete set of 1494 protein–protein interfaces. The lower cutoff of 10 interface residues for the set is evident in part (a).

#### *Physicochemical Properties*

The physicochemical properties, residue-type distribution, and residue conservation were compared for interface and noninterface surface regions of all 1494 proteins in the data set (see Protein–protein interface data sets in Materials and Methods, above). The physicochemical properties include hydrophobicity, solvation energy, and relative solvent-accessible surface area. Histograms of these three properties for interface and noninterface surface regions are shown in Figure 3(a–c). It is evident from these plots that the interface region is more hydrophobic and has a higher solvation energy, that is, is less polar. Statistical tests confirm these biases at a high significance level ($< 1 \times 10^{-15}$). The histograms also show the larger variance in interface properties due to the small number of residues in the interface compared with the number of noninterface surface residues. However, no difference is found between the average SASA for interface and noninterface regions at the 5% significance level. The histogram in Figure 3(c) is consistent with this conclusion since both distributions are fairly symmetric and peaked around 70 Å$^2$. This result disagrees with that of Jones and Thornton[4] which concluded that the SASA is significantly higher for interface surface patches. This may be due to the use of a much smaller data set (32 proteins) of different composition and the different analysis method, based on surface patches, in that paper. In any case, this property may not be useful for predicting protein–protein interfaces using the structure of an isolated protein since it is sensitive to the detailed conformations of residue side chains, which change upon forming a complex.

Because both the hydrophobicity and solvation energy are just different descriptors of the same property, with a high hydrophobicity corresponding to a high solvation energy, they are expected to be correlated. The high correlation coefficient of 0.84 between the hydrophobicity and solvation energy of interfaces in the data set confirms this hypothesis. This means that while both descriptors discriminate between the interface and the remaining surface, they provide largely redundant information.

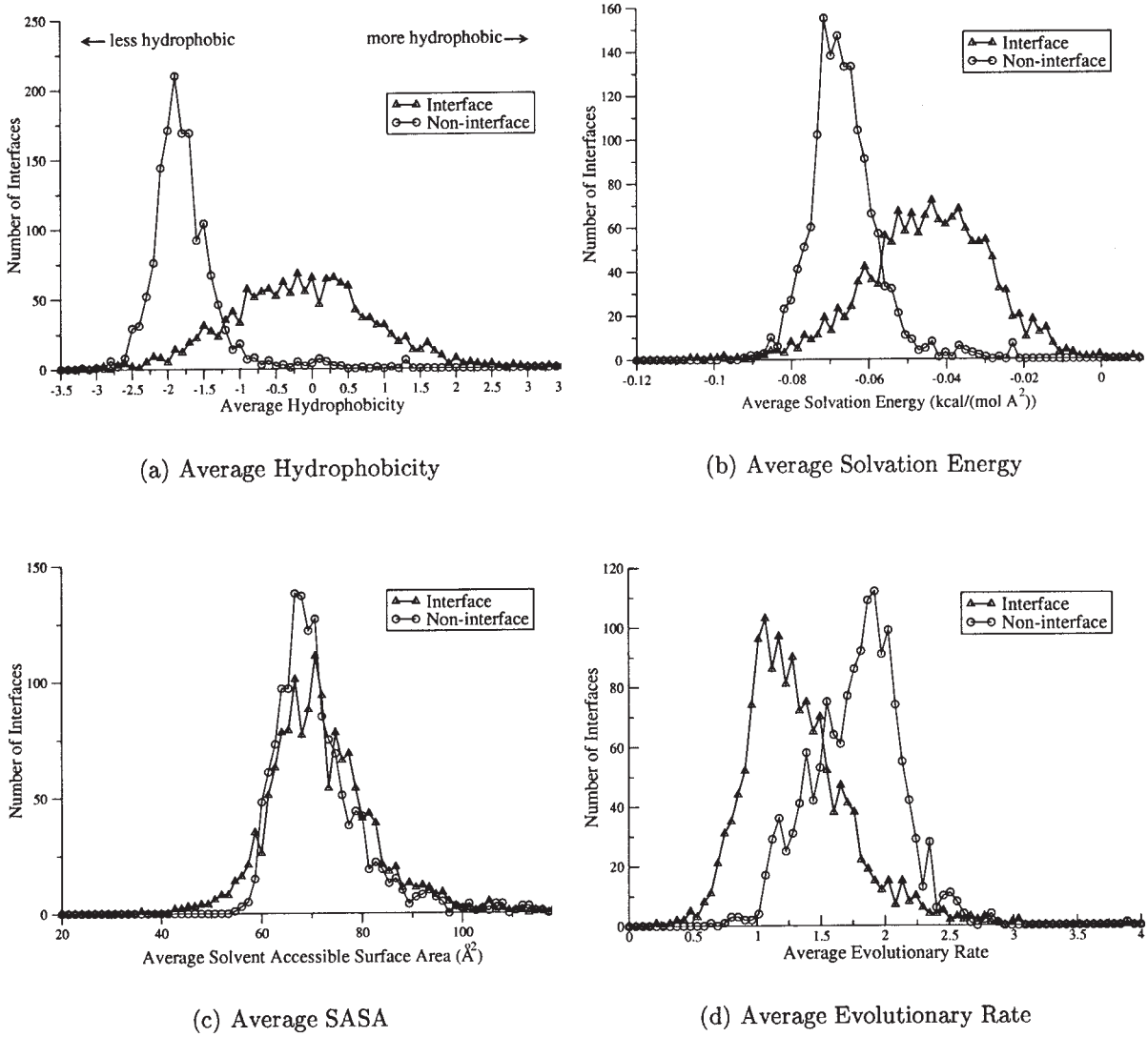


Fig. 3. Histograms of average (**a**) hydrophobicity, (**b**) solvation energy, (**c**) solvent accessible surface area, and (**d**) evolutionary rate for interface and noninterface residues in the set of 1494 protein–protein interfaces.

### *Residue Type Distribution*

The number of residues of each type in the interface was compared with a hypergeometric distribution, which describes a random distribution of residues on the protein surface. The fraction of protein–protein interfaces in the data set in which residues of a given type are more prevalent in the interface at the 5% significance level is shown in Figure 4. These may be compared with residue propensities in Jones and Thornton,[4] however only a qualitative comparison is possible since, in that study, the data set was considerably smaller and no statistical tests were performed. Both results indicate a prevalence of large hydrophobic and uncharged polar residues and sparsity of charged residues in the interface. However, whereas the propensities of certain residues, such as histidine and leucine, are considerably different for homodimers and heterodimers in the other study, the results in Figure 4 suggest that the propensities for homodimer and heterodimer interfaces differ little for most residue types. Statistical tests confirm that differences in the fractions of homodimer and heterodimer interfaces with significant propensity for a given residue type are significant (at the 5% level) only for aspartic acid and glycine, both of which are more prevalent in heterodimer interfaces. The different conclusion reached in our study may be due to the different composition of the data set in Jones and Thornton,[4] which included a larger proportion of antibody–antibody and enzyme–inhibitor complexes, which have more polar interfaces. The large fraction of interfaces with significantly higher proportions of hydrophobic and polar residue types indicates that the residue type distribution can be used to detect protein–protein interfaces on the surface.

### *Evolutionary Conservation*

Residue conservation, as expressed by the site evolutionary rates calculated using the method described in Materials and Methods, above, was also compared between the

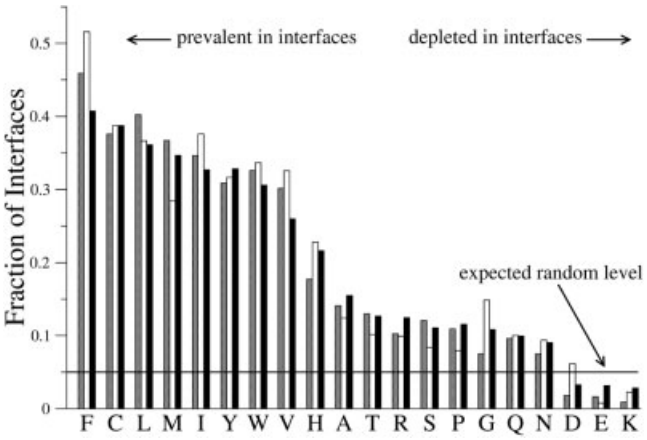


Fig. 4. Interface residue propensity measured as the fraction of protein–protein interfaces in which the occurrence of a given residue type is greater in the interface than in the non-interface regions at the 5% significance level. Values for the set of homodimers are shown as gray bars, heterodimers as white bars and multimers as black bars. Residue types are displayed in decreasing order of occurrence in the interfaces of the complete set of 1494 proteins.

interface and noninterface surface residues. Histograms of the average evolutionary rates for interface and noninterface residues are shown in Figure 3(d). A total of 934 of the 1494 interfaces in the data set, or about 63%, had higher residue conservation at the 5% significance level. This indicates that evolutionary conservation may be used to discriminate between protein–protein interfaces and the remainder of the surface for a large fraction of the data set. Its inclusion in an interface prediction method should then result in a more sensitive prediction. The fractions of homodimer interfaces (317/518) and heterodimer interfaces (79/114) with higher conservation did not appreciably differ from that of the complete set.

### Protein–Protein Interface Prediction

The previous analysis indicates that hydrophobicity, residue type distributions, and evolutionary rates may all be utilized to predict protein–protein interfaces given the structure of a query protein. Fivefold cross validation for the SVM interface prediction was performed on the complete set of 122,375 surface residues for the 632 proteins in the dimer interface set. The prediction method is described in detail in Support Vector Machines in Materials and Methods, above. Both residue-type distributions and evolutionary rates for residues in surface patches were used used as input data. Including hydrophobicity values had little effect of the prediction accuracy (data not shown), possibly because it is strongly correlated with the residue-type distribution. The cross validation statistics for the prediction are shown in Table I. It should be emphasized that these prediction statistics are for data that was not used for training. The recall, which measures the fraction of interface residues that were correctly predicted, is 35% higher than expected and the precision, which measures the fraction of predicted interface residues that were correctly predicted, is 24% higher than expected from a random assignment. In addition, a high fraction, 97%, of the predicted interface patches overlapped with the actual interface even though on average only 22% of the surface residues were included in the predicted patch. The Receiver Operating Characteristic curve for the fivefold cross validation on the complete dimer set is shown in Figure 5. This curve shows the tradeoff between sensitivity and specificity for the prediction.

### Effect of Conservation on Prediction Accuracy

Prediction cross validation statistics were also calculated for interfaces in the dimer set with high evolutionary conservation, defined by $p < 1.0 \times 10^{-4}$, and those with low evolutionary conservation, defined by $p > 0.2$, in order to assess the degree that conservation contributes to the prediction accuracy. As may be seen in Table I, highly conserved interfaces are substantially easier to predict and conversely interfaces whose evolutionary conservation is indistinguishable from other surface residues are more difficult to predict. The residue-type distribution clearly contributes to the prediction since all statistics are higher than random values for the nonconserved interfaces, however the prediction quality for these interfaces is probably too low to be useful.

### Transient Heterodimers

We also examined interface conservation and prediction accuracy for a set of 43 transient heterodimer interface from the paper of Nooren and Thornton.[42] A total of 29 of the interfaces, or about 67%, were more conserved than the remaining surface at the 5% significance level. Because this fraction is even slightly higher than for the data set used here (see Physicochemical Properties, above), which presumably contains a large proportion of permanent interfaces, it suggests that transient interfaces are no less conserved than permanent interfaces. This contrasts with the analysis of Caffrey et al.[13] which concluded that transient interface core residues are not significantly more conserved than the remaining surface. This may be due to the small size of their data set (10 interfaces). Interface conservation of the transient heterodimers was not studied in the original paper of Nooren and Thornton.

Two different SVMs, one trained on the complete set of 632 dimer interfaces and the other trained only on the subset of 114 heterodimer interfaces, were used to predict protein–protein interfaces for the transient dimers. The statistics in Table II show that prediction using the SVM trained only on heterodimer interfaces had a slightly higher precision. It also had approximately 10% lower recall, which was compensated by a correspondingly lower random recall value due to a smaller fraction of predicted interface residues. Comparison with the cross validation statistics for the complete dimer set in Table I show that while the recall for the transient heterodimer interface predictions is comparable the precision is lower, indicating a higher number of false positive results. This indicates that the predictions for this set, while reliable enough to be useful, are somewhat less accurate than the results for the dimer set. This cannot be explained by lower conservation since, as stated above, the interface conservation is as high

**TABLE I. SVM Cross Validation Statistics for the Complete Dimer Interface Set, and Subsets with High and Low Interface Evolutionary Rates[†]**

| Prediction Data Set | Interfaces | Accuracy (%) | Recall (%) | Precision (%) |
|---|---|---|---|---|
| All dimer interfaces (without post-processing) | 632 | 76 (66) | 64 (28) | 34 (15) |
| All dimer interfaces | 632 | 80 (70) | 57 (22) | 39 (15) |
| Dimer interfaces with high conservation (rates $p < 1.0 \times 10^{-4}$) | 182 | 82 (66) | 68 (24) | 51 (18) |
| Dimer interfaces with low conservation (rates $p > 0.2$) | 166 | 78 (74) | 34 (19) | 20 (11) |

[†]All results include the prediction post-processing described in Materials and Methods except those in the first row, as indicated. Expected random values are shown in parentheses. Accuracy = (true positives + true negatives)/(number of predictions), recall = (true positives)/(true positives + false negatives), and precision = (true positives)/(true positives + false positives). All $p$-values for deviation from random values are effectively zero ($< 5 \times 10^{-18}$).

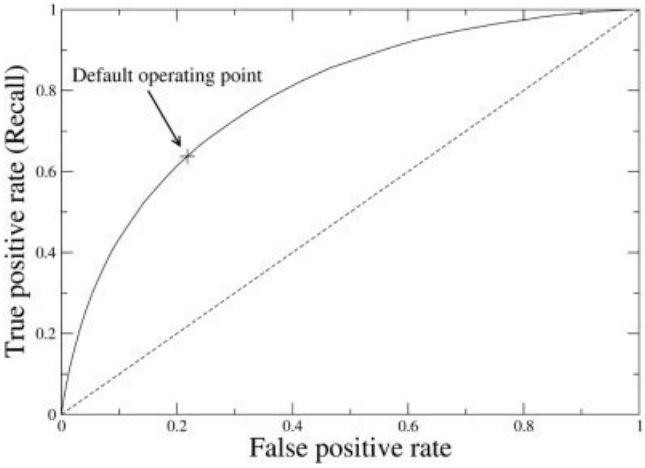


Fig. 5. Receiver Operator Characteristic (ROC) curve for fivefold cross validation on the complete set of 632 dimer interfaces. The area under the curve is 0.79. False positive rate = (false positives)/(false positives + true negatives) and true positive rate = (true positives) /(true positives + false negatives). The default SVM operating point with a decision value cutoff of zero and corresponding to the statistics in Table I is also marked.

**TABLE II. SVM Validation Statistics for the Prediction of 43 Transient Heterodimer Interfaces From Nooren and Thornton[42] Using an SVM Trained on Our Set of Dimer Interfaces (632) and the Subset of Heterodimer Interfaces (114)[†]**

| SVM training data set | Accuracy (%) | Recall (%) | Precision (%) |
|---|---|---|---|
| All dimer interfaces | 67 (60) | 67 (37) | 22 (12) |
| All heterodimer interfaces | 76 (69) | 57 (26) | 26 (12) |

[†]Expected random values are shown in parentheses (see Table 1 for definition).

as for the dimer set. An examination of the interface hydrophobicity shows that, like those in the dimer set, the interfaces are more hydrophobic than the non-interface regions ($p = 1.7 \times 10^{-7}$). The residue composition of the interfaces also has a similar bias to that of the complete set (shown in Fig. 4). Investigation of the proteins with the worst prediction statistics reveals that the apparent lower prediction accuracy is rather due to the larger number of alternative binding partners for transient heterodimers, most of which are involved in intracellular signalling. Since these other binding partners are not present in the X-ray crystal structure used for the prediction, the predicted interface residues are labelled as a false positives. For example, the transient heterodimer data set contains complexes of the GTP-binding nuclear protein Ran with four different binding partners: regulator of chromosome condensation (RCC1), nuclear transport factor 2 (NTF2), importin β, and the Ran binding domain of RanBP2. Four interface patches, labelled as P1–P4, are predicted for Ran using the complex with RCC1 (PDB entry 1I2M[43]). Structural alignment with Ran in the other complexes shows that all four predicted interfaces largely overlap the actual protein–protein interfaces inferred from the X-ray structures. RCC1 and NTF2 bind at patch P1, importin β binds at patches P1 and P2, and RanBP2 binds at patches P3 and P4. It is interesting that the largest predicted binding patch, P1, contains the binding interfaces of three different proteins, all of which have little overlap with each other.

### Detailed Analysis of a Representative Subset

Since the complete set of dimer interfaces is too large for a detailed analysis of all prediction results a representative sample of interfaces was selected in an unbiased manner. The subset of 21 protein–protein interfaces shown in Table III were randomly selected from the set of all interfaces for proteins of an intermediate sequence length, $250 < L < 300$. The prediction results for these proteins are also displayed on the protein surfaces in Figure 6. It is apparent from this figure that many residues classified as false positives are adjacent to the interface. This may indicate their functional role in stabilizing the protein–protein interactions. No prediction method can predict the interface with perfect accuracy because of the arbitrariness in the definition of which residues belong to the interface, particularly around the periphery. The choice of whether interface residues are defined by the relative SASA lost upon forming a complex or by $C_\alpha$ distances as well as the choice of cutoff values leads to different sets of interface residues and hence different assignments of erroneously predicted residues. Figure 6 also shows that residues far from the interface are usually correctly classified as noninterface residues.

Table III shows that many of the false positives are actually residues in active sites or ligand binding sites in which the ligand is missing from the structure. These sites are known to be highly conserved and so difficult to distinguish from predicted protein–protein interfaces, ex-

**TABLE III. Representative Sample of 21 Protein–Protein Interfaces in Dimer Set (Randomly Selected Proteins With Sequence Length 250 < *L* < 300)†**

| PDB chains | Reference | Protein(s) | Prediction comments |
|---|---|---|---|
| 1ABR B-A | 48 | Abrin-a | B-chain lectin |
| 1AD4 A-B | 49 | Dihydropteroate synthetase | |
| 1ALN M-M | 50 | Cytidine deaminase | |
| 1F5Q A-B | 51 | CDK2, γ-Herpesvirus Cyclin | Not natural protein binding partners (human cyclins). Ligands missing (ATP and peptide) False positives in active site. |
| 1F89 A-B | 52 | Hypothetical protein YLR351C | Unknown function. False positives in putative catalytic cleft. |
| 1FK8 A-B | 53 | 3α-Hydroxysteroid dehydrogenase | Substrate (steroid) missing. False positives near NAD cofactor and in putative substrate binding site. |
| 1G60 A-B | 54 | Methyltransferase MboIIA | Equilibrium between monomers and dimers. Ligand (DNA) missing. False positives in putative DNA binding loop. |
| 1GS5 A-A | 55 | Acetylglutamate kinase | False positives near catalytic site. |
| 1GZ0 A-F | 56 | rRNA Methyltransferase homolog (Yjfh) | Ligands (AdoMet and RNA) missing. False positives in putative cofactor (AdoMet) and rRNA binding sites. |
| 1J5P A-A | | Hypothetical protein Tm1643 | Unknown function. False positives in cleft containing NAD. |
| 1JXI A-B | 57 | Phosphomethylpyrimidine kinase | ATP and $Mg^{2+}$ missing. False positives near substrate (HMP) binding site. |
| 1KC3 A-A | 58 | Dtdp-glucose oxidoreductase | False positive near active site. |
| 1KZQ A-B | 59 | Major surface antigen p30 | Natural ligand unknown (may be sulfated proteoglycans). False positive patch around C-terminal region (GPI-anchored to cell membrane). Also false positives in deep groove between D1 and D2 domains. |
| 1L5X A-B | 60 | Survival protein E homolog | Unknown function. Substrate missing. |
| 1MEE A-I | 61 | Subtilisin, eglin-C | Not natural ligand (peptide). |
| 1MG5 A-B | | Alcohol dehydrogenase | Substrate missing. |
| 1N57 A-A | 62 | Chaperone Hsp31 | Substrate (peptide) and ATP missing. False positives in putative hydrophobic substrate binding pocket. |
| 1NPD A-B | 63 | Shikimate 5-dehydrogenase homolog (Ydib) | Substrate missing. False positives in putative substrate binding site and near cofactor (NAD) binding site. |
| 1O0W A-B | | Ribonuclease III (Tm1102) | Substrate missing. |
| 1O0Y A-B | | Deoxyribose-phosphate aldolase (Tm1559) | Substrate missing. |
| 2HHM A-B | 64 | Inositol monophosphatase | Substrate missing. False positives in putative binding pocket. |

†Comments are for the prediction cross validation results shown graphically in Figure 6.

cept by their smaller size. Post-processing of prediction results, as described in Materials and Methods, above, only removed predicted interface patches smaller than most of these ligand binding sites. The surface area cutoff was not set higher than 150 Å$^2$ since this caused many correctly predicted interface patches to be removed (data not shown). One interesting case is the *Toxoplasma gondii* Major Surface Antigen p30 which had a large predicted interface patch surrounding the C-terminus. This is probably a correct prediction of the interaction of this region with the membrane since the C-terminus is glycosylphosphatidylinositol (GPI) anchored to the membrane.

### Predicted Tetramer Interface for a Glycyl-tRNA Synthetase

Examination of the prediction results revealed a large predicted interface with no binding partner in the structure for PDB entry 1J5W, a *Thermotoga maritima* glycl-tRNA synthetase. The structure is annotated as a *homodimer* in the PDB file. Figure 7(a, b) shows the clearly predicted homodimeric interface as well as another interface on the opposite side. Although eukaryotic and archeal glycyl-tRNA synthetases are homodimeric, eubacterial glycyl-tRNA synthetases have been observed to usually have an $\alpha_2\beta_2$ structure. This prediction suggests that the *T. maritima* glycl-tRNA synthetase also has an $\alpha_2\beta_2$ structure and indicates residues for mutagenesis studies to verify this hypothesis. There are no available structures for homologous aminoacyl-tRNA synthetases with the same quaternary structure. The prediction result for *Thermus thermophilus* seryl-tRNA synthetase, a homodimer, shown in Figure 7(c, d) for comparison. This result shows only the well-predicted dimeric interface, thus confirming that this enzyme is indeed a homodimer. These two proteins have no significant sequence homology, as is the case for all proteins in the data set, and so are independent predictions.

### Higher Prediction Reliability for Central Interface Residues

The likelihood ratio described in Materials and Methods above, may be used to rank predicted interface residues by the reliability of the prediction. Inspection of likelihood ratio values for large interfaces showed that often the central interface residues have higher values than peripheral interface residues. This is illustrated for *Escherichia coli* cytidine deaminase in Figure 8. In order to verify this

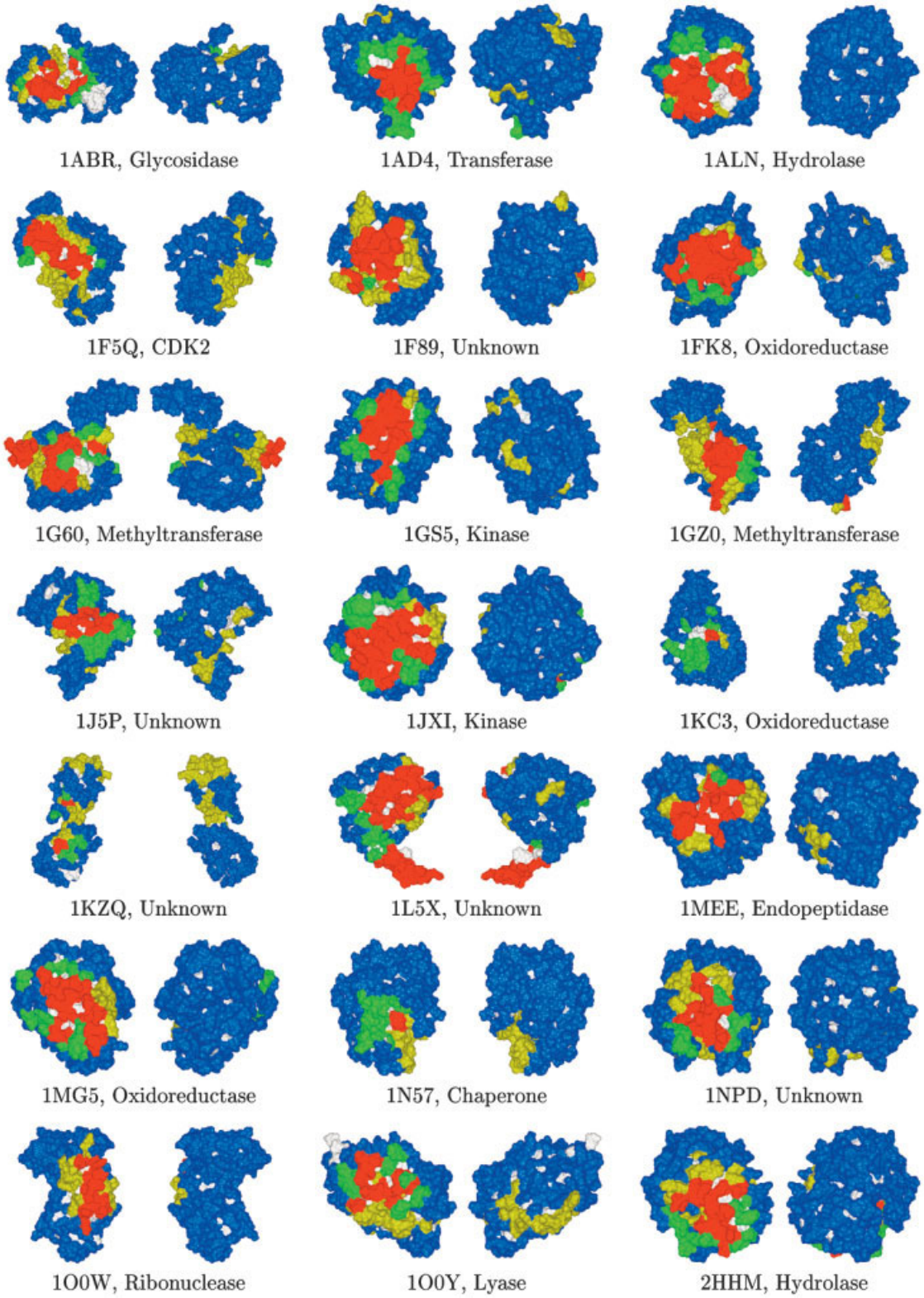


Fig. 6. Protein–protein interface predictions for a representative sample of proteins in the test set (randomly selected proteins with sequence length $250 < L < 300$). The solvent-excluded surface for each residue is colored as follows: red, true positive; blue, true negative; yellow, false positive; and green, false negative. White indicates residues that were not included in the prediction either because they bind to small molecules or have zero SASA. More information is given in Table III.

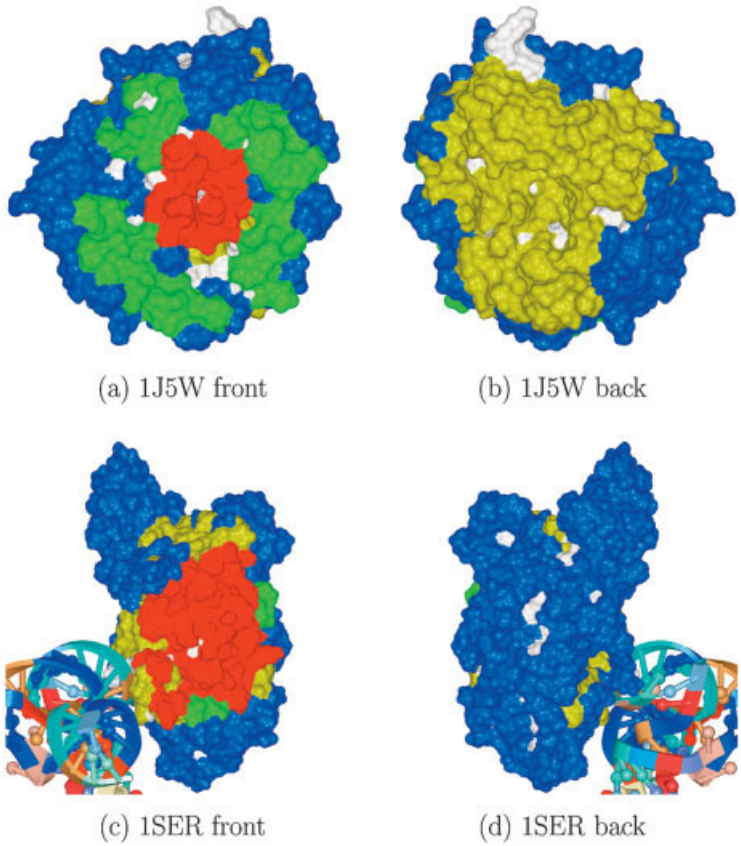


Fig. 7. Interface prediction results for PDB entry 1J5W, *T. maritima* glycl-tRNA synthetase, and 1SER,[65] *T. thermophilus* seryl-tRNA synthetase (refer to Fig. 6 for an explanation of the surface colors). A portion of the bound tRNA is shown in the 1SER figures. The presence of a large predicted interface opposite the dimeric interface for the glycl-tRNA synthetase but not the seryl-tRNA synthetase suggests that their quaternary structures are an $\alpha_2\beta_2$ tetramer and a homodimer, respectively.

hypothesis, a set of core residues was defined as those with no $C_\alpha$ atoms within 8 Å of a noninterface residue $C_\alpha$ atom. The remainder of the interface residues were defined as peripheral residues. A total of 109 of the 145 dimer interfaces with greater than six central residues had higher likelihood ratios for the central interface residues than the peripheral interface residues at the 5% significance level. This implies that central residues are, in fact, more strongly predicted than peripheral ones. A statistical test for evolutionary rates showed that 65 of the 145 interfaces had more conserved central interface residues than peripheral interface residues at the same significance level. Since this is a considerably higher number of interfaces than expected randomly it may be concluded that the higher conservation of central residues is a major contribution to their more confident prediction. Another factor that contributes is the fact that, compared with central residues, patches for peripheral residues contain a larger proportion of noninterface residues, thus diluting any discriminating signal.

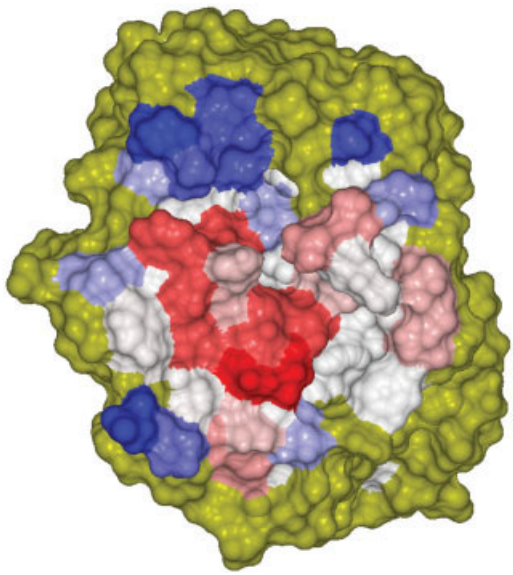

Fig. 8. Higher likelihood ratio for central residues in the 1ALN interface. The protein's surface is colored by the corresponding residue's likelihood ratio: red, high; white, intermediate; and blue, low. The surface that is not part of the dimer interface is shown in yellow. The molecule has the same orientation as in Figure 6.

## Future Directions

There are several possible extensions of the prediction method presented here. Studies of alanine-scanning mu-

tagenesis experimental results show that a small number of interface residues contribute a disproportionately large fraction of the binding energy.[9,10] These hot-spot residues are also observed to be structurally conserved[44] as well as maintain a similar conformation in the unbound structure and bury the largest side-chain SASA in the complex.[45] This implies that the addition of *structural* conservation to our method may improve the prediction accuracy when a sufficient number of similar X-ray crystal structures are available. Also it may be possible to discriminate hot spot residues from other interface residues using an SVM trained on residue propensities and evolutionary and structural conservation. The use of local surface patches in the prediction, as are used for the protein–protein interface prediction, may be important because of evidence that hot spot residue interactions are protected from solvent by surrounding occluding residues.[10]

A study by Ofran and Rost[46] found that six different types of protein–protein interfaces significantly differed in their residue composition and residue–residue interaction preferences. This suggests that it may be possible to train an SVM to classify predicted interfaces by their type, for example, permanent or transient, using the same input data as for the protein–protein interface prediction. A related application is differentiating biologically relevant interfaces from nonspecific interfaces in X-ray crystal structures, as demonstrated by Janin and coworkers.[47]

Binding sites for ligands, such as small molecules, DNA/RNA, and peptides, were frequently predicted as interfaces, even when they were not present in the structure. This is likely due to higher conservation at these sites. Because these sites are usually in large surface pockets, unlike protein–protein binding sites, limiting the prediction to surface pockets may enable the prediction of ligand binding sites while filtering out larger protein–protein interfaces.

Many protein–protein docking algorithms begin by sampling all possible relative conformations of two rigid protein structures. Interface predictions for two potential binding partners may be used to limit the number of conformations sampled and thus speed up the docking calculation.

## CONCLUSION

Statistical analysis of protein–protein interfaces in the large data set demonstrated that while interfaces are generally more hydrophobic and have a higher solvation energy than the remaining surface the average residue SASA does not significantly differ. A test of whether the number of residues in the interface of a specific type is significantly larger than expected from a random distribution gave results qualitatively in agreement with a previous study using a smaller data set,[4] namely the interface residue composition is enriched in large hydrophobic and uncharged polar residues but depleted of charged residues. However, unlike the previous study, the composition was not appreciably different for homodimers and heterodimers, except for aspartic acid and glycine. Overall, the fact that about 30–50% of the interfaces had detectably higher fractions of eight different residue types suggests that this is an important discriminating signal. Finally, about 63% of the interfaces were shown to have significantly higher evolutionary conservation indicating that it is also a strong discriminating property.

Fivefold cross validation using the complete set of 632 dimer interfaces was used to test the SVM prediction performance. Z-scores of residue frequencies in multiple alignment columns and evolutionary rates were used as input. Statistics for the test set confirm that the prediction was both considerably more sensitive (recall difference = 35%) and more selective (precision difference = 24%) than expected randomly. The high fraction, 97%, of the predicted interfaces overlapping with the actual interface, even though an average of only 22% of the surface residues were predicted as belonging to the interface, reflects the high accuracy of the prediction model. Detailed examination of an arbitrary subset of predictions revealed that many wrongly classified residues were near the interface boundary and that false positives were often near the binding sites of ligands absent in the structure. Comparing the prediction results for subsets of interfaces that have high conservation and average conservation demonstrated that evolutionary conservation is a major discriminating signal for interface prediction. Likelihood ratios calculated using a histogram method showed that prediction results for central interface residues had a higher confidence level than for peripheral interface residues. This may be partially attributed to the higher evolutionary conservation observed for central residues. In any case, likelihood ratios are useful for prioritizing putative interface residues for experimental verification.

A smaller set of transient heterodimers was also studied in order to discover any differences from permanent complexes which presumably comprise the majority of the larger interface set. The fraction of interfaces that had significantly higher conservation than the remaining surface was comparable to that of the larger set, in contrast with a previous paper.[13] Although statistics showed that the interface prediction for transient heterodimers was somewhat less accurate than for the large set, an examination of the least accurate predictions revealed large predicted interfaces that actually localize to the interfaces for alternative binding partners not present in the structures. This is not surprising as most proteins in the set are involved in intracellular signalling and have multiple protein–protein interactions.

## SUPPLEMENTARY MATERIAL

Tables with information on the protein–protein interface data set and detailed cross validation prediction results are available online as Supplementary Material.

## ACKNOWLEDGMENTS

This work was funded by a grant from the Department of Energy (No. DE-FG03-01ER83282). We would like to thank L. Budagyan for assistance with the SVM software that he developed.

## REFERENCES


1. Korn AP, Burnett RM. Distribution and complementarity of hydropathy in multi-subunit proteins. Proteins 1991;9:37–55.
2. Jones S, Thornton JM. Analysis of protein–protein interaction sites using surface patches. J Mol Biol 1997;272:121–132.
3. Lo Conte L, Chothia C, Janin J. The atomic structure of protein–protein recognition sites. J Mol Biol 1999;285:2177–2198.
4. Jones S, Thornton JM. Principles of protein–protein interactions. Proc Natl Acad Sci USA 1996;93:13–20.
5. Lawrence MC, Colman PM. Shape complementarity at protein/protein interfaces. J Mol Biol 1993;234:946–950.
6. McCoy AJ, Epa V. Chandana, Colman PM. Electrostatic complementarity at protein/protein interfaces. J Mol Biol 1997;268:570.
7. Xu D, Tsai CJ, Nussinov R. Hydrogen bonds and salt bridges across protein-protein interfaces. Protein Eng 1997;10:999–1012.
8. Sheinerman FB, Norel R, Honig B. Electrostatic aspects of protein-protein interactions. Curr Opin Struct Biol 2000;10:153–159.
9. Clackson T, Wells JA. A hot spot of binding energy in a hormone-receptor interface. Science 1995;267:383–386.
10. Bogan AA, Thorn KS. Anatomy of hot spots in protein interfaces. J Mol Biol 1998; 280:1–9.
11. Grishin NV, Phillips MA. The subunit interfaces of oligomeric enzymes are conserved to a similar extent to the overall protein sequences. Protein Sci 1994;3:2455–2458.
12. Valdar WS, Thornton JM. Protein-protein interfaces: analysis of amino acid conservation in homodimers. Proteins 2001;42:108–124.
13. Caffrey DR, Somaroo S, Hughes JD, Mintseris J, Huang ES. Are protein-protein interfaces more conserved in sequence than the rest of the protein surface? Protein Sci 2004;13:190–202.
14. Bordner AJ, Abagyan RA. REVCOM: a robust Bayesian method for evolutionary rate estimation. Bioinformatics. Advanced Access published on March 4, 2005; doi: 10.1043/bioinformatics/b.
15. Young L, Jernigan RL, Covell DG. A role for surface hydrophobicity in protein-protein recognition. Protein Sci 1994;3:717–729.
16. Jones S, Thornton JM. Prediction of protein-protein interaction sites using patch analysis. J Mol Biol 1997;272:133–143.
17. Fernandez-Recio J, Totrov M, Skorodumov C, Abagyan R. ODA (optimal desolvation area): new predictor for protein-protein interaction sites. Proteins 2005;58:134–143.
18. Lichtarge O, Bourne HR, Cohen FE. An evolutionary trace method defines binding surfaces common to protein families. J Mol Biol 1996;257:342–358.
19. 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:S71–77.
20. Landgraf R, Xenarios I, Eisenberg D. Three-dimensional cluster analysis identifies interfaces and functional residue clusters in proteins. J Mol Biol 2001;307:1487–1502.
21. Fariselli P, Pazos F, Valencia A, Casadio R. Prediction of protein–protein interaction sites in heterocomplexes with neural networks. Eur J Biochem 2002;269:1356–1361.
22. Zhou HX, Shan Y. Prediction of protein interaction sites from sequence profile and residue neighbor list. Proteins 2001;44:336–343.
23. Koike A, Takagi T. Prediction of protein-protein interaction sites using support vector machines. Protein Eng Des Sel 2004;17:165–173.
24. Berman HM, Westbrook J, Feng Z, Gilliland G, Bhat TN, Weissig H, Shindyalov IN, Bourne PE. The Protein Data Bank. Nucleic Acids Res 2000;28:235–242.
25. Molsoft LLC. ICM software manual. Version 3.0. 2004.
26. Apweiler R, Bairoch A, Wu CH, Barker WC, Boeckmann B, Ferro S, Gasteiger E, Huang H, Lopez R, Magrane M, and others. UniProt: the universal protein knowledgebase. Nucleic Acids Res 2004;32: D115–D119.
27. Kyte J, Doolittle RF. A simple method for displaying the hydropathic character of a protein. J Mol Biol 1982;157:105–132.
28. Wesson L, Eisenberg D. Atomic solvation parameters applied to molecular dynamics of proteins in solution. Protein Sci 1992;1:227–235.
29. Altschul SF, Gish W, Miller W, Myers EW, Lipman DJ. Basic local alignment search tool. J Mol Biol 1990;215:403–410.
30. Thompson JD, Higgins DG, Gibson TJ, Clustal W: improving the sensitivity of progressive multiple sequence alignment through sequence weighting, positions-specific gap penalties and weight matrix choice. Nucleic Acids Res 1994;22:4673–4680.
31. Saitou N, Nei M. The neighbor-joining method: a new method for reconstructing phylogenetic trees. Mol Biol Evol 1987;4:406–425.
32. Howe K, Bateman A, Durbin R. QuickTree: building huge Neighbour-Joining trees of protein sequences. Bioinformatics 2002;18: 1546–1547.
33. Jones DT, Taylor WR, Thornton JM. The rapid generation of mutation data matrices from protein sequences. Comput Appl Biosci 1992;8:275–282.
34. Bulmer M. Use of the method of generalized least-squares in reconstructing phylogenies from sequence data. Mol Biol Evol 1991;8:868–883.
35. Neyman J. Molecular studies of evolution: a source of novel statistical problems. In: Gupta SS, Yackel J, editors. Statistical decision theory and related topics. New York: Academic Press; 1971. p 1–27.
36. Felsenstein J. Evolutionary trees from DNA sequences: a maximum likelihood approach. J Mol Evol 1981;17:368–376.
37. Boser BE, Guyon IM, Vapnik VN. A training algorithm for optimal margin classifiers. In: Haussler D, editor. Proceedings of the 5th Annual ACM Workshop on Computational Learning Theory. New York: Association for Computational Machinery; 1992. p 144–152.
38. Schölkopf B. Support vector learning. Munich: Oldenbourg Verlag. 1997.
39. Vapnik VN. Statistical learning theory. New York: Wiley. 1998.
40. Chang CC, Lin CJ, LIBSVM: a library for Support Vector Machines. Technical report. Department of Computer Science and Information Engineering, National Taiwan University. 2003.
41. Ponstingl H, Henrick K, Thornton JM. Discriminating between homodimeric and monomeric proteins in the crystalline state. Proteins 2000;41:47–57.
42. Nooren IM, Thornton JM. Structural characterisation and functional significance of transient protein-protein interactions. J Mol Biol 2003;325:991–1018.
43. Renault L, Kuhlmann J, Henkel A, Wittinghofer A. Structural basis for guanine nucleotide exchange on Ran by the regulator of chromosome condensation (RCC1). Cell 2001;105:245–255.
44. 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;100:5772–5777.
45. Rajamani D, Thiel S, Vajda S, Camacho CJ. Anchor residues in protein-protein interactions. Proc Natl Acad Sci USA 2004;101: 11287–11292.
46. Ofran Y, Rost B. Analysing six types of protein-protein interfaces. J Mol Biol 2003; 325:377–387.
47. Bahadur RP, Chakrabarti P, Rodier F, Janin J. A dissection of specific and non-specific protein-protein interfaces. J Mol Biol 2004;336:943–955.
48. Tahirov TH, Lu TH, Liaw YC, Chen YL, Lin JY. Crystal structure of abrin-a at 2.14 Å. J Mol Biol 1995;250:354–367.
49. Hampele IC, D'Arcy A, Dale GE, Kostrewa D, Nielsen J, Oefner C, Page MG, Schonfeld HJ, Stuber D, Then RL. Structure and function of the dihydropteroate synthase from staphylococcus aureus. J Mol Biol 1997;268:21–30.
50. Xiang S, Short SA, Wolfenden RJ, Carter CW. Cytidine deaminase complexed to 3-deazacytidine: a "valence buffer" in zinc enzyme catalysis. Biochemistry 1996;35:1335–1341.
51. Card GL, Knowles P, Laman H, Jones N, McDonald NQ. Crystal structure of a γ-herpesvirus cyclin-cdk complex. EMBO J 2000;19: 2877–2888.
52. Kumaran D, Eswaramoorthy S, Gerchman SE, Kycia H, Studier FW, Swaminathan S. Crystal structure of a putative CN hydrolase from yeast. Proteins 2003;52:283–291.
53. Grimm C, Maser E, Mobus E, Klebe G, Reuter K, Ficner R. The crystal structure of 3α-hydroxysteroid dehydrogenase/carbonyl reductase from *Comamonas testosteroni* shows a novel oligomerization pattern within the short chain dehydrogenase/reductase family. J Biol Chem 2000;275:41333–41339.
54. Osipiuk J, Walsh MA, Joachimiak A. Crystal structure of MboIIA methyltransferase. Nucleic Acids Res 2003;31:5440–5448.
55. Ramon-Maiques S, Marina A, Gil-Ortiz F, Fita I, Rubio V. Structure of acetylglutamate kinase, a key enzyme for arginine biosynthesis and a prototype for the amino acid kinase enzyme family, during catalysis. Structure 2002;10:329–342.
56. Michel G, Sauve V, Larocque R, Li Y, Matte A, Cygler M. The

structure of the RlmB 23S rRNA methyltransferase reveals a new methyltransferase fold with a unique knot. Structure 2002;10: 1303–1315.
57. Cheng G, Bennett EM, Begley TP, Ealick SE, Crystal structure of 4-amino-5-hydroxymethyl-2-methylpyrimidine phosphate kinase from Salmonella typhimurium at 2.3 Å resolution. Structure 2002;10:225–235.
58. Blankenfeldt W, Kerr ID, Giraud MF, McMiken HJ, Leonard G, Whitfield C, Messner P, Graninger M, Naismith JH. Variation on a theme of SDR. dTDP-6-deoxy-L-lyxo-4-hexulose reductase (RmlD) shows a new Mg2+-dependent dimerization mode. Structure 2002;10:773–786.
59. He XL, Grigg ME, Boothroyd JC, Garcia KC. Structure of the immunodominant surface antigen from the Toxoplasma gondii SRS superfamily. Nat Struct Biol 2002; 9:606–611.
60. Mura C, Katz JE, Clarke SG, Eisenberg D. Structure and function of an archaeal homolog of survival protein E (SurEα): an acid phosphatase with purine nucleotide specificity. J Mol Biol 2003;326: 1559–1575.
61. Dauter Z, Betzel C, Genov N, Pipon N, Wilson KS. Complex between the subtilisin from a mesophilic bacterium and the leech inhibitor eglin-C. Acta Crystallogr B 1991;47:707–730.
62. Quigley PM, Korotkov K, Baneyx F, Hol WG. The 1.6 Å crystal structure of the class of chaperones represented by *Escherichia coli* Hsp31 reveals a putative catalytic triad. Proc Natl Acad Sci USA 2003;100:3137–3142.
63. Benach J, Lee I, Edstrom W, Kuzin AP, Chiang Y, Acton TB, Montelione GT, Hunt JF. The 2.3 Å crystal structure of the shikimate 5-dehydrogenase orthologue YdiB from *Escherichia coli* suggests a novel catalytic environment for an NAD-dependent dehydrogenase. J Biol Chem 2003;278:19176–19182.
64. Bone R, Springer JP, Atack JR. Structure of inositol monophosphatase, the putative target of lithium therapy. Proc Natl Acad Sci USA 1992;89:10031–10005.
65. Biou V, Yaremchuk A, Tukalo M, Cusack S. The 2.9 Å crystal structure of T. thermophilus seryl-tRNA synthetase complexed with tRNA(Ser). Science 1994;263:1404–1410.