doi:10.1016/j.jmb.2006.07.028 J. Mol. Biol. (2006) 362, 365–386

JMB

Available online at www.sciencedirect.com
ScienceDirect

ELSEVIER

Insights into Protein–Protein Interfaces using a Bayesian Network Prediction Method

James R. Bradford[1], Chris J. Needham[2], Andrew J. Bulpitt[2] and David R. Westhead[1]*

[1]*Institute of Molecular and Cellular Biology, University of Leeds, Leeds, LS2 9JT, UK*

[2]*School of Computing University of Leeds Leeds, LS2 9JT, UK*

**Corresponding author*

Identifying the interface between two interacting proteins provides important clues to the function of a protein, and is becoming increasing relevant to drug discovery. Here, surface patch analysis was combined with a Bayesian network to predict protein–protein binding sites with a success rate of 82% on a benchmark dataset of 180 proteins, improving by 6% on previous work and well above the 36% that would be achieved by a random method. A comparable success rate was achieved even when evolutionary information was missing, a further improvement on our previous method which was unable to handle incomplete data automatically. In a case study of the Mog1p family, we showed that our Bayesian network method can aid the prediction of previously uncharacterised binding sites and provide important clues to protein function. On Mog1p itself a putative binding site involved in the SLN1-SKN7 signal transduction pathway was detected, as was a Ran binding site, previously characterized solely by conservation studies, even though our automated method operated without using homologous proteins. On the remaining members of the family (two structural genomics targets, and a protein involved in the photosystem II complex in higher plants) we identified novel binding sites with little correspondence to those on Mog1p. These results suggest that members of the Mog1p family bind to different proteins and probably have different functions despite sharing the same overall fold. We also demonstrated the applicability of our method to drug discovery efforts by successfully locating a number of binding sites involved in the protein–protein interaction network of papilloma virus infection. In a separate study, we attempted to distinguish between the two types of binding site, obligate and non-obligate, within our dataset using a second Bayesian network. This proved difficult although some separation was achieved on the basis of patch size, electrostatic potential and conservation. Such was the similarity between the two interacting patch types, we were able to use obligate binding site properties to predict the location of non-obligate binding sites and *vice versa*.

Keywords: Bayesian network; protein–protein binding sites; interaction types; drug targets; Mog1p family

Introduction

Structural genomics projects are beginning to produce protein structures with unknown function, and therefore accurate, automated predictors of protein function are required if all these structures are to be annotated in reasonable time. Identifying the interface between two interacting proteins provides important clues to the function of a protein and can reduce the search space required by docking algorithms to predict the structures of complexes. Detecting novel protein–protein binding sites is also

Abbreviations used: SVM, support vector machine; ASA, accessible surface area; ROC, receiver operating characteristic; AUC, area under a ROC curve; MCC, Matthews correlation coefficient; MSA, multiple sequence alignment.

E-mail address of the corresponding author: D.R.Westhead@leeds.ac.uk

becoming increasing important to the drug discovery process given recent evidence that protein–protein interactions make "drugable" targets.[1,2]

Nooren & Thornton[3] describe two ways of classifying protein–protein interactions based on the components and lifetime of the complex. The term "obligate" describes complexes in which individual components cannot exist as stable structures independently *in vivo*. By contrast, if each component can exist as a stable structure under physiological conditions then the complex is described as non-obligate. Interactions can be further classified as permanent, weak transient or strong transient according to the lifetime of the complex. Permanent interactions are very stable and mostly, but not always occur in obligate complexes. Weak transient interactions occur between two proteins that need to associate and dissociate continuously *in vivo*. A transient interaction may become permanent under certain cellular conditions but usually the type of interaction is inferred by the function of the protein.

In general, binding sites share common properties that distinguish them from the rest of the protein.[4–7] For example, in their bound conformation they are often the most planar and accessible of all the surface patches regardless of interaction type.[6] Hydrophobic residues also cluster at some interfaces,[8–10] especially large interfaces of obligate or permanent complexes,[5,9,11] whilst other smaller, transient interfaces are less hydrophobic and have a significant number of polar residues.[5,11–13] Hydrophobic residues tend to be scattered over these interfaces in order to accommodate electrostatic interactions,[12] hydrogen bonding and salt bridges.[11,14] Charged side-chains are often excluded from protein–protein interfaces with the exception of arginine. Arginine is one of the most abundant interface residues regardless of interaction type[15,16] mainly due to its hydrogen bonding capacity and role in cation-π interactions.[17] Patches of low desolvation energy or "optimal docking areas" (ODAs) often correspond to protein–protein interfaces.[18] However, secondary structure composition appears to be of little discriminatory value, since neither α-helices nor β-sheets dominate at transient binding sites.[13] Alanine-scanning has shown that binding free energy is not equally distributed at a protein–protein interface[19,20] with the majority of the binding affinity provided by a small number of conserved, polar "hot-spot" residues[21,22] often at the centre of the binding site.[23]

Evolutionary conservation has some discriminatory power for obligate and more permanent interactions,[24,25] although protein–protein interfaces in general are often not conserved to the extent where they can be distinguished from other surface patches.[26–28] Nevertheless conservation scoring systems such as Evolutionary Trace have been used with some success to locate protein–protein binding sites.[29,30] Interestingly, the interface core tends to be more conserved than the periphery in both obligate and non-obligate cases.[31]

No single property absolutely differentiates protein–protein interfaces from other surface patches[6] therefore most binding site prediction methods combine more than one physical–chemical property. Jones & Thornton[32] defined roughly circular patches on the protein surface, then scored and ranked each patch according to its chemical and physical properties. Similarly, Neuvirth *et al.*[33] applied a probabilistic approach to assess the likelihood of surface patches being part of a binding site using a dataset of unbound proteins involved in transient interactions. Several groups have used machine learning methods such as neural networks[34–37] and support vector machines[25,38–41] (SVMs), although the most effective of these make extensive use of structural information.[38,39] In particular, we used an SVM in combination with surface patch analysis to predict binding sites with a success rate of 76% on a benchmark dataset of 180 proteins containing both obligate and non-obligate binding sites.[38] However, the SVM was unable to handle incomplete data automatically, such as instances where evolutionary information was unavailable. Furthermore, given that a random method achieves a success rate of 36% on the same data set there is still a need to improve prediction accuracy.

A number of attempts have been made to differentiate the interface types assigned by Nooren & Thornton.[3] Per-residue surface and interfaces areas of non-obligate interactions tend to be smaller than those of obligate interactions,[42] with obligate interactions involving more non-polar contacts.[43] Mintseris & Weng[44] found that obligate interfaces evolve more slowly than transient interfaces. This allows them to co-evolve with their interaction partners and so correlated mutations are rare between transient interfaces.[44] In earlier work, the same authors used atomic contact vectors to discriminate obligate from non-obligate interactions with a success rate of 91% although this required knowledge of the binding partner.[45] Recently De *et al.*[43] found that involvement of defined secondary structure elements such as β-sheets and helices is much more common across subunits at an obligate interface than a non-obligate interface. Despite these differences, there remains a need for an accurate classifier of interaction type that combines structural and sequence information and requires no knowledge of the binding partner.

In this work, we have devised a highly accurate protein–protein binding site prediction method using a Bayesian network in combination with surface patch analysis. We also attempt to distinguish obligate from non-obligate binding sites using a second Bayesian network. Bayesian networks are probabilistic graphical models which provide compact representations for expressing joint probability distributions and for inference.[46] This representation and use of probability theory makes Bayesian networks suitable for learning from incomplete datasets, expressing causal relationships, combining domain knowledge and data, and avoid over-fitting a model to data. Consequently, a host of applications in computational biology have used Bayesian networks and Bayesian learning methodologies:[47,48] analysis of gene expression data,[49–58] prediction of

transcription factor binding sites and other functional DNA regions,[59–62] prediction of sub-cellular location,[63] discovering structural correlations in α-helices,[64] protein–protein interaction prediction,[65] and gene function prediction.[66] To our knowledge, Bayesian networks have yet to be applied to protein–protein binding site prediction.

Overview

This work is motivated in two ways: to predict both protein–protein binding site location and type (whether obligate and non-obligate), and in doing so provide insights into the properties that characterise a binding site and drive complex formation.

The first part of this work, binding site location prediction, consists of two separate phases. In the training phase, we train two Bayesian networks (one analogous to a naïve Bayes classifier and another designed using expert knowledge) to distinguish between interacting and non-interacting surface patches taken from a benchmark dataset of 180 proteins.[38] To do this we exploit several surface properties previously implicated in distinguishing protein–protein binding sites from the rest of the protein surface: hydrophobicity, residue interface propensity, shape, sequence conservation, electrostatic potential, and solvent accessible surface area (ASA). The best performing classifier is then carried forward to the prediction phase where we perform two cross-validation tests: one using all available data, the other without access to sequence conservation scores. In addition, we carry out a study on four proteins in the Mog1p family that share the same fold but little sequence similarity. The family represents an ideal test case for our method, since it includes two structural genomics targets, one of which has little or no detectable sequence homology to any known protein, and two other proteins involved in protein–protein interactions but with binding sites yet to be located experimentally. Finally in the prediction phase, we demonstrate our method's applicability to the drug discovery process by predicting known binding sites involved in the protein–protein interaction network of papilloma virus infection.

In the second part of this work, we train a second Bayesian network to distinguish obligate from non-obligate binding sites using similar properties to those used in binding site location prediction but with the addition of patch size and secondary structure nodes. Based on findings from this study we carry out a heterogeneous cross-validation test where we train our binding site location Bayesian network above on obligate data in order to predict non-obligate binding sites, and *vice versa*.

Results and Discussion

Training phase

Two Bayesian network structures

We compared binding site prediction performance of two Bayesian network structures: a structure analogous to a naïve Bayes classifier (Figure 1(a)), and an "expert" Bayesian network (Figure 1(b)), both with 14 nodes representing the mean and standard deviation of seven surface properties across a patch, and a class node (binding site patch?).

A naïve structure contains only edges from the class node to the other observations (thus assuming that all the variables are independent) and is called a naïve Bayes classifier. We derived an expert Bayesian network structure with edges between the residue interface propensity and hydrophobicity nodes (Figure 1(b)). Our rationale was that the hydrophobic nature of a patch of protein surface is strongly correlated with the residues found within that region; the correlation coefficients were 0.93 and 0.73 between the means and standard deviations, respectively, of patch hydrophobicity and residue interface propensity.

Training procedure

For each protein within our benchmark, non-redundant training set of 180 proteins[38] (see also Materials and Methods), we generated one protein surface patch involved in interactions (interacting patch) and one patch taken from the non-interacting parts of the surface (non-interacting patch) of equivalent size to the interacting patch. We then trained both the naïve and expert Bayesian networks to distinguish between the two patch types and com-

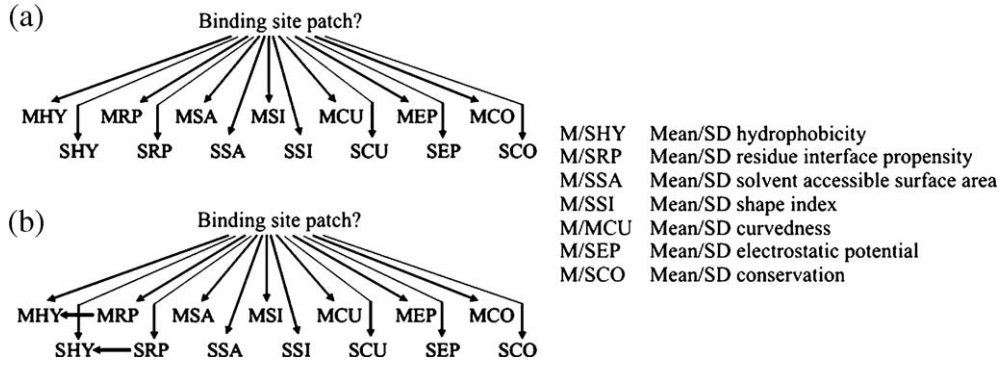

Figure 1. The two Bayesian network structures tested in this work for their ability to distinguish between interacting and non-interacting patches. (a) Naïve and (b) expert.

pared performance. Note that no sequence homologues were found in Swissprot (release 46) for 12 of the 180 proteins. In these cases, the expectation maximisation (EM) algorithm was used to estimate the expected values of the conservation scores (see Materials and Methods).

The Bayesian network assigns a probability score to each patch under test. Consequently, classification performance depends on the probability threshold below which a patch is classed as "non-interacting" and above which a patch is classified as "interacting" and so calculating sensitivity (TP/TP + FN, where TP and FN are the numbers of true positives and false negatives) or specificity (TN/TN + FP, where TN and FP are the numbers of true negatives and false positives) values at a single threshold is potentially misleading. Therefore, we plotted receiver operating characteristic (ROC) curves to evaluate performance. A ROC curve is a plot of sensitivity (true positive rate) *versus* (1-specificity) (false positive rate) across a range of probability thresholds (in our case from 0.0 to 1.0). The area under a ROC curve (AUC) gives a measure of classifier performance; an AUC of 1.0 is indicative of a perfect classifier whereas the AUC of a classifier no better than random is 0.50. In order to derive a probability threshold, p, to define our "predicted patches" in the prediction phase, we chose a point on the ROC curve of the best classifier at which the gradient is equal to one, and is closest to the point (0, 1). This point represents an equal cost of a false positive and a false negative.

Training performance

ROC curves for each of the Bayesian network structures are shown in Figure 2. The AUC for the naïve classifier was 0.89±0.01, compared to 0.90± 0.01 for the expert Bayesian network, suggesting little gain in associating hydrophobicity with residue interface propensity. The high AUC for the naïve classifier was nevertheless indicative of a very good classifier. In both cases, the equal costs probability p was approximately 0.50, where mean sensitivity and specificity values were both 0.81 for the naïve classifier, and 0.83 for the expert Bayesian network.

The small size of the data set was a potential source of overfitting so it was important to measure the significance of our AUC values obtained above. Randomisation testing has previously been found to be very effective at assessing over-fitting.[67,68] Here, the original training set is copied and class labels are replaced with random class labels. Then the Bayesian network is trained on these data using the same methodology that is used with the original data. Any estimate of accuracy greater than random for the randomly labelled data reflects the bias in the methodology, and this reference distribution can then be used to adjust the estimates on the real data. In this work, randomization testing was implemented by training both Bayesian networks on five datasets each containing 360 patches randomly classified into equal numbers of interacting and non-interacting

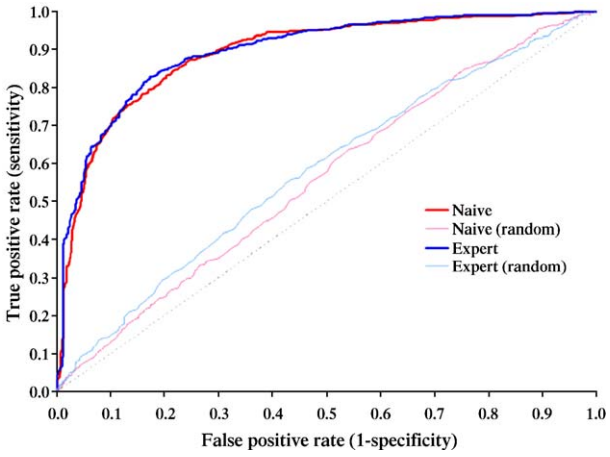

Figure 2. ROC curves for naive (AUC = 0.89±0.01) and expert (AUC = 0.90±0.01) Bayesian network structures.

patches. From these random data, ROC curves were plotted that gave "baseline" AUC values to which our AUC values above could be compared. The extent to which these baseline AUC values exceeded 0.50 (the performance of a random classifier) indicated the level of potential over-fitting of the model.

Baseline AUC values of 0.57±0.01 and 0.59±0.01 were obtained for the naïve and expert structures, respectively, suggesting that the slight increase in training performance with the expert structure could be related to the larger number of free parameters in this model. It is also worth noting that the Bayesian network structures learnt automatically from the training data using maximum weight spanning tree and greedy search algorithms (see Materials and Methods) displayed no edges other than those between the class node and the other variables. These learned structures equated to the naïve structure in Figure 1(a).

We concluded that there were no useful connections between nodes other than those between the class node and the other variable nodes. Therefore, we used a naïve Bayesian network in all subsequent analyses.

Comparison with previous work

In previous work,[38] a support vector machine (SVM) was trained to distinguish interacting patches from non-interacting patches from the same benchmark dataset with a mean Matthews Correlation Coefficient[69] (MCC; equation (1)) of 0.63±0.03. This was used as the basis for comparison here, because AUC values were not reported in the earlier study. This amounts to comparing the performance of the classifiers at a single, optimal value of the classification threshold (0.0 for the SVM score, and $p = 0.5$ for the Bayesian network).

$$\mathrm{MCC} = \frac{(\mathrm{TP} \times \mathrm{TN}) - (\mathrm{FP} \times \mathrm{FN})}{\sqrt{(\mathrm{TP} + \mathrm{FP})(\mathrm{TP} + \mathrm{FN})(\mathrm{TN} + \mathrm{FP})(\mathrm{TN} + \mathrm{FN})}} \quad (1)$$

An MCC of +1 represents perfect training classification (no false positives or negatives) whereas –1 represents a complete failure (all positives classified as negative and *vice versa*).

For the SVM, the baseline MCC, calculated by training on random data (analogous to the random data used for calculating baseline AUC), was 0.13± 0.03. The Bayesian network achieved a mean MCC of 0.62±0.03 and baseline MCC of 0.09±0.05. Therefore, even though gross MCC scores were similar (0.62 *versus* 0.63), the baseline MCC of the naïve Bayesian network was marginally less than that of the SVM indicating that the SVM model had a tendency to over-fit the training data to a slightly greater extent than the Bayesian network.

Prediction phase

Prediction strategy

For each protein subject to binding site prediction, we generated enough patches for complete coverage of the protein surface (one patch per surface atom). Each patch was then assigned a probability value by the naïve Bayesian network (trained as above) according to the likelihood that the patch was part of a protein–protein binding site. These probability values were used to produce a ranked set of non-overlapping predicted patches with probabilities above $p = 0.50$.

Success criteria

We used two measures to define the success of our predicted patches: patch precision, indicating the proportion of the predicted patch residues that were interface residues (equation (2)), and interface coverage, indicating the proportion of interface residues that were included in the predicted patch (equation (3)).

$$\text{Patch precision} = \frac{\text{No. of interface residues in patch}}{\text{No. of patch residues}} \quad (2)$$

$$\text{Interface coverage} = \frac{\text{No. of interface residues in patch}}{\text{No. of interface residues}} \quad (3)$$

The patch precision measure is equivalent to reliability used by Neuvirth *et al.*,[33] with interface coverage equivalent to the percentage overlap measure by Jones & Thornton.[32] Neither group considered both measures for their success criteria. In terms of assessing prediction performance, our priority was high patch precision with a reasonable level of interface coverage. Thus, a prediction was deemed a success if a binding site patch with over 50% precision (equation (2)) and 20% interface coverage (equation (3)) was ranked in the top three predicted patches.

As a measure of the significance of our predictions we also calculated the number of successes one would expect to achieve across our data set if each protein surface was sampled at random (equation (4)) in Materials and Methods).

Predictive performance

Leave-one-out cross-validation. Leave-one-out cross-validation involved removing one protein from the training set (and the interface residue propensity calculation to avoid bias), training the Bayesian network on the remaining proteins, and then predicting the position of the binding site on the selected protein. This process was repeated until all proteins had been left out. Because non-interacting patches were chosen at random for the training step, results varied slightly between separate cross-validation runs. Therefore, we repeated the entire cross-validation procedure five times and evaluated average performance.

The naïve Bayesian network performed considerably better than the SVM of earlier work,[38] achieving a success rate of 82% (148/180) compared to 76% (136/180) with the SVM (Table 1C). Both methods performed significantly above the 36% (65/180) success rate expected across the whole data set by random chance. An example set of results taken from one of the five cross-validation runs providing details for each individual test case is given in Supplementary Data, Table 3.

The percentage of top ranked binding site patches also increased from 45% (81/180) with the SVM to 52% (94/180) with the Bayesian network. The overall increase in performance with the naïve Bayesian network can, for the most part, be attributed to its higher success rate of 79% (52/66) on non-obligate interfaces, compared to 65% (43/66) with the SVM (Table 1A). By contrast, performance on the obligomers increased by less than 1% (Table 1B). It is interesting to note that whilst the SVM compared well with the naïve Bayesian network in training (see above), the Bayesian network appears to generalise better to unseen data. This supports our initial theory based

Table 1. Comparison of leave-one-out cross-validation success rates between a naïve Bayesian network and an SVM used in previous work[38]

Predictor	No. of examples	No. of successes[a]	No. of patches ranked 1st
A. *Non-obligates*			
Naïve BN	66	52	35
SVM	66	43	28
B. *Obligates*			
Naïve BN	114	96	60
SVM	114	93	53
C. *All*			
Naïve BN	180	148	94
SVM	180	136	81

Standard deviations were 1–2% of the mean value in all cases.
[a] No. of proteins with a patch of over 50% precision and 20% interface coverage ranked in the top 3.

on the random training results that the SVM model was overfitting the data to a greater extent than the Bayesian network.

The upper limits of patch precision and interface coverage and specificity are constrained by the size and shape of the patch relative to the interface. Our calculation of patch size (see Materials and Methods) was such that the majority of the interacting patches (67%) during training were smaller than their corresponding binding sites, with only 4% matching their binding site exactly. However, small patch sizes that cover only the core of the binding site are potentially an advantage, since "hot-spot" residues contributing most of the binding affinity tend to cluster at the centre of interfaces.[23] This suggests that patches consisting of mainly the central binding site residues provided a strong binding site "signal" for the Bayesian network to learn. Furthermore, in the majority of cases the side-chains of these hot spot residues are in their bound conformation even before any physical interaction occurs.[70] There is also evidence from molecular dynamics simulations to suggest that the core of the interface is generally less mobile than the periphery.[71] Therefore, our smaller patches may be more tolerant to conformational changes during complex formation than patches that cover the entire interface.

Discrimination between protein binding sites and other functional regions. Most functional site predictors, particularly those that use evolutionary information alone, are indiscriminate in terms of the type of site they predict, whether a protein, ligand or DNA binding site. This is because most important sites functional sites on a protein surface tend to be highly conserved. With the inclusion of the six properties in addition to sequence conservation, and the training of the Bayesian network on protein binding sites alone, our method should be specific to protein binding site prediction. However, we have shown previously that there is a possibility that in cases where a successful patch is not found, one of the top three patches could be located at another important functional region on the protein surface, particularly a ligand binding site.[38] This is not surprising, since protein–protein binding sites often share common properties with ligand binding sites such as high conservation and presence of clefts. There are even cases where both protein inhibitors and ligands can bind at enzyme active sites, and it is now thought possible that small molecule drugs can bind at protein–protein interaction sites.[1,2] Therefore, in cases where we do predict a known ligand binding site, this ligand binding site may also be the location for an as yet unknown protein binding site. Given this, it is difficult to quantify how well the Bayesian network can distinguish between protein binding sites and ligand binding sites. Nevertheless, recent evidence has shown that ligand and protein binding site clefts can display properties distinctive from one another[72] and so our method should be biased towards predicting protein–protein binding sites over ligand binding sites.

The ability of our method to discriminate a DNA from a protein binding site is more critical since, unlike in the case of a ligand binding site, a protein is unlikely to bind to a DNA binding site. To test whether we were indeed predicting protein binding sites in preference to DNA binding sites, we carried out a small study on 33 proteins that are known to bind both proteins and DNA. These test cases were selected from two recent datasets of DNA binding proteins[73,74] using similar criteria to our derivation of the 180 protein training set (see Materials and Methods). The important filtering steps included the removal of proteins sharing over 20% sequence identity with another DNA binding protein or a protein within our 180 protein training set. NMR structures and structures whose resolution was worse than 3.0 Å were also disallowed as were complexes whose interfaces were made up of more than one separate chain or proteins containing more than one known protein binding site.

Training a Bayesian network on our 180 protein training set and testing on each DNA binding protein in turn, we achieved a success rate of 88% (29/33) in protein binding site prediction, whilst the DNA binding site was predicted amongst the top three patches in only 33% (11/33) of cases (Table 2). Twenty-two top ranked patches formed part of a protein binding site whereas only two covered a DNA binding site. In only three cases (grey highlighted in Table 2) a DNA binding site patch was ranked higher than a protein binding site patch within the top three.

Most of the dataset comprises of obligate protein interactions since DNA binding proteins with non-obligate protein binding sites are poorly represented in the PDB. This could explain our high success rate since obligate binding sites are generally easier to predict than non-obligates (see above). Nevertheless, the two non-obligate binding sites in our dataset were both predicted successfully and more accurately than their respective DNA binding sites. These results suggest that our method is heavily biased towards predicting protein binding sites over DNA binding sites, and that the properties of a DNA binding site are sufficiently different from a protein binding site to make this possible.

ROC curve. To assess prediction performance further, we generated a ROC curve (Figure 3) based on probability scores calculated by the Bayesian network for all 133,600 patches used in leave-one-out cross-validation. The data set therefore contained 8249 positive and 125,351 negative patches where positive patches were those with over 50% patch precision and 20% interface coverage. The area under the curve (AUC) was 0.86 ± 0.02, which indicated that predictive performance was comparable with training performance and the model had generalized well to unseen data even though the training set of 180 proteins was relatively small.

Table 2. Ability of Bayesian network to predict protein binding sites in preference to DNA binding sites

Protein		Highest ranked binding site patch[a]	
Query chain	Binding partner	Protein	DNA
A. *Homo-obligates*			
1a74_A	1a74_B	1	–
1b3t_A	1b3t_B	1	2
1c0w_A	1c0w_B	1	–
1cma_A	1cma_B	1	–
1d02_A	1d02_B	1	–
1ddn_A	1ddn_B	1	–
1f4k_A	1f4k_B	1	2
1g9z_A	1g9z_B	1	–
1gdt_A	1gdt_B	1	–
1gu4_A	1gu4_B	1	–
1jt0_A	1jt0_C	1	3
1kc6_A	1kc6_B	1	–
1llm_C	1llm_D	1	–
1lmb_3	1lmb_4	1	–
1lq1_A	1lq1_B	1	3
1pvi_A	1pvi_B	1	3
1srs_A	1srs_B	1	–
1zme_C	1zme_D	1	
1dh3_A	1dh3_C	2	–
1j59_A	1j59_B	2	–
1je8_A	1je8_B	2	–
1ku7_A	1ku7_D	2	–
3cro_R	3cro_L	2	1
1bhm_A	1bhm_B	–	–
1gxp_A	1gxp_B	–	–
1l3l_A	1l3l_C	–	–
1n6q_A	1n6q_B	–	2
B. *Hetero-obligates*			
1awc_A	1awc_B	1	3
1h9d_A	1h9d_B	1	–
1nkp_A	1nkp_B	1	3
1yrn_A	1yrn_B	2	1
C. *Non-obligates*			
1cf7_A	1cf7_B	1	2
1t7p_A	1t7p_B	2	–

Cases where a DNA binding site patch is ranked higher than the best ranked protein binding site patch.

[a] Within the top three patches only, the highest ranked binding site patch with greater than 50% precision and 20% interface coverage to either a protein or DNA binding site. "–" Denotes no binding site patch found in top three.

Probability score versus patch precision/interface coverage. To study the relationship between probability score and patch precision/interface coverage further we generated all 133,600 patches possible across the 180 protein dataset according to our patch definition (see Materials and Methods), and then calculated patch precision and interface coverage for each patch. The Pearson correlation coefficient, R, derived by linear regression, between probability score output by the Bayesian network and patch precision was 0.48, and between probability score and interface coverage was 0.42. Over this large dataset these correlations are highly significant with p-values for zero correlation equal to 0.0 (reported by Matlab). These results suggest a strong correlation between probability score and the degree of interface in the patch.

Handling missing data

Effect of hiding the conservation nodes. One of the strengths of a Bayesian network is its ability to handle incomplete or missing data. A common source of missing data for our patches is evolutionary information if no homologues of the query protein are found (this occurs in 12/180 proteins in our data set). To demonstrate that the Bayesian network was robust to the removal of evolutionary information, we performed leave-one-out cross-validation with conservation scores on the query protein missing. That is, for the naïve structure trained with evolutionary information (where available), the probability of a patch from the query protein being part of a binding site was calculated when evolutionary information (the mean and standard deviation of the conservation scores across a patch) was missing and the two nodes treated as hidden. An overall success rate of 82% (147/180) was achieved despite the missing data, which was comparable with that achieved using all available data. Therefore, the Bayesian network was successfully able to infer the binding site patch class node by marginalising over all possible values of the hidden variables. It is important to remember that this kind of study would have been very difficult with our SVM without resorting to assigning dummy conservation scores on the test protein, since it is not possible to train an SVM on a full set of attributes and test with one or more attributes hidden.

The use of conservation does have some effect on prediction as illustrated by results on non-obligate and obligate interfaces separately. Performance on obligate interfaces actually increased from 84% to 89% (96/114 to 101/114) upon removal of evolutionary information, but decreased from 79% to 70% (52/66 to 46/66) on non-obligates, suggesting that evolutionary properties make a significant contribution to the detection of some non-obligate interfaces, whilst hindering the detection of a number of obligate interfaces.

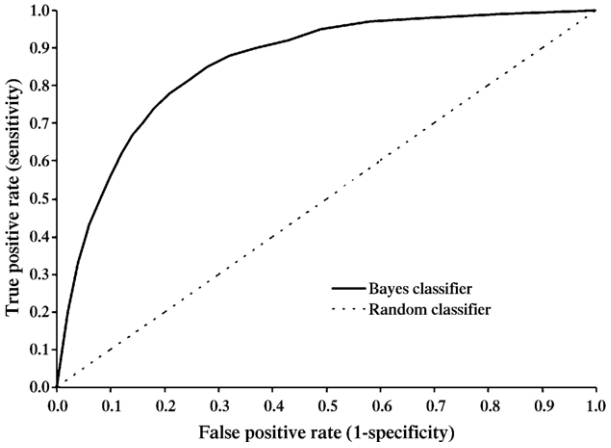

Figure 3. ROC curve (AUC = 0.86 ± 0.02) to assess prediction performance. Ten probability thresholds were considered between 0.0 and 1.0.

Two case studies. In an attempt to explain the results above further, we studied two proteins where hiding the conservation nodes made a difference to binding site prediction: Zn, Cu superoxide dismutase[75] (an obligate homo-dimer; PDB code: 1xso, chain A), and a cytotoxic T-lymphocyte protein that makes a non-obligate interface with T lymphocyte activation antigen CD80[76] (PDB code: 1i8l, chain C). In the case of superoxide dismutase, hiding the conservation nodes improved binding site prediction, whereas the opposite was true for the T-lymphocyte protein. Note that here we assume the Bayesian network has learnt that binding site patches tend to be hydrophobic and conserved with high residue interface propensity. Whilst this may be true, the Bayesian network uses other nodes and could have learnt more complex, non-linear ways to separate interacting from non-interacting patches.

The majority of the site of interaction between the superoxide dismutase subunits is hydrophobic with high interface propensity, although it is not particularly well conserved with respect to the rest of the protein surface with the majority of conservation scores around 0.5. Adjacent to this binding site is a patch consisting of a convex protrusion of high hydrophobicity and interface propensity surrounded by a ring of highly conserved, hydrophilic residues of low interface propensity. Utilising all available data, including evolutionary information, the naïve Bayesian network was unable to predict a binding site patch in the top three ranked patches on superoxide dismutase (Figure 4). Instead, the adjacent patch was predicted as the most likely binding site with a probability of about 91%. Conversely, with the two conservation nodes hidden, a top ranked patch with 75% interface coverage and 67% patch precision on the surface of this protein was predicted with a probability of approximately 93%. Interestingly, the adjacent patch is predicted as the third ranked patch but with probability reduced to 75%. So why is prediction of this obligate binding site hindered by using evolutionary information? The key perhaps lies in the differences between the binding site itself and the patch adjacent to it. Mean hydrophobicity and residue interface propensity values at binding site are slightly higher than those at the adjacent patch, although the adjacent patch is much more conserved. Hiding the conservation nodes effectively reduces the influence of evolutionary information (although some knowledge of how to use conservation comes from the training data), and so the remaining nodes, particularly hydrophobicity and interface residue propensity, both of which are favourable at the binding site, become more critical. Consequently, without evolutionary information, the probability of the adjacent patch being a binding site patch reduces to 75% whilst high hydrophobicity and interface propensity mean that a binding site patch is now ranked top despite the lack of conservation. Of course, the failure to locate the obligate binding site on superoxide dismutase using all available data is not necessarily a negative result, since the patch adjacent to the binding site may itself be part of another, maybe non-obligate, binding site that is essential for the enzyme's function. Indeed, further study revealed that the patch included a significant proportion of the active site cavity, in particular the invariant His61 residue that forms a bridge between the copper and zinc ions. This may be a common problem in the specific prediction of obligate binding sites on proteins with conserved functional sites (of any type) elsewhere on the surface, and may explain why evolutionary infor-mation tends to hinder the detection of obligate rather than non-obligate binding sites. Whereas the majority of the

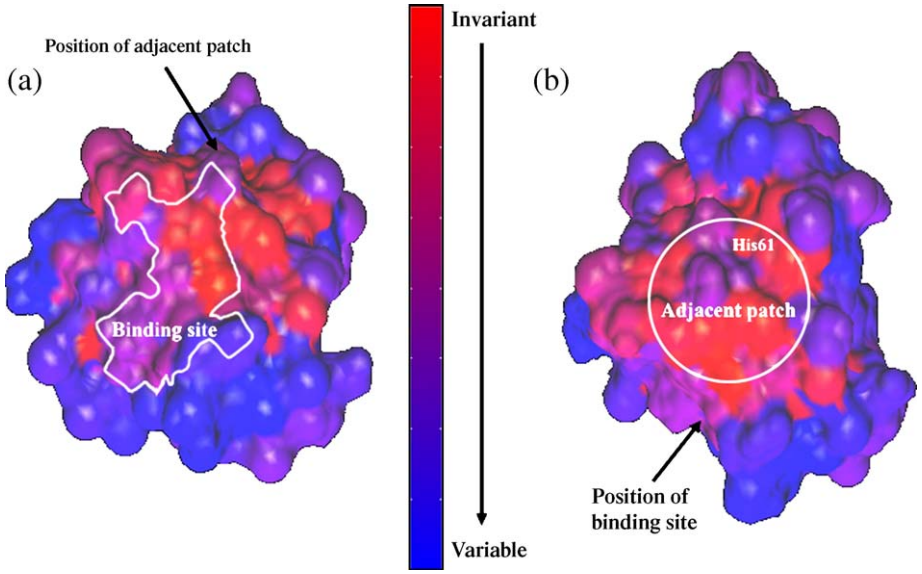

Figure 4. Effect of hiding the conservation nodes on prediction of an obligate binding site. (a) The homo-dimeric binding site on superoxide dismutase that is predicted by the Bayesian network when the conserved nodes are hidden. (b) A more conserved patch predicted as the most likely binding site location when all data is made available. Included in this patch is the His61 residue, which forms an integral part of the enzyme active site.

non-obligate proteins in our dataset are dimers with one protein–protein binding site, some of the proteins defined as obligate may also have other conserved functional sites that would be detected by algorithms based on conservation information.

With regard to the non-obligate example (cytotoxic T-lymphocyte protein), a top ranked patch with 82% interface coverage and 60% patch precision at 98% probability is predicted by the naïve Bayesian network when all available data are utilised. However, with the two conservation nodes hidden, no successful patches in the top three ranked patches are predicted. Unlike the obligate binding site on superoxide dismutase, the non-obligate binding site on T-lymphocyte protein is part of the most conserved region of the protein surface. It is also of high interface residue propensity but, unusually since interface propensity and hydrophobicity are often well correlated, only average hydrophobicity. This could be the reason why hiding the conservation nodes has a negative affect on prediction, since without evolutionary information a patch with average conservation but both high hydrophobicity and residue interface propensity is predicted as the top ranked patch.

In the non-obligate case therefore, conservation provides an essential guide for the Bayesian network to locate the binding site. By contrast, a highly conserved, non-binding site patch on superoxide dismutase misdirects predictions away from the obligate binding site, perhaps due to the presence of the enzyme active site in this patch. In this case, hiding evolutionary information is an advantage, since other nodes become more influential and the higher hydrophobicity and residue interface propensity at the obligate binding site ensure that it predicted with greater probability than the competing patch. However, these results suggest that while inclusion or exclusion of conservation information has different effects on different individual cases, average performance over the whole test set is only marginally changed. Interfaces can therefore be successfully predicted without using homology.

Studies on the Mog1p structural family

We applied our prediction method to the four proteins comprising the Mog1p structural family: (1) Mog1p, a regulatory protein for the nuclear transport of Ran GTPase in yeast; (2) TM1622 from *Thermotoga maritime*, solved by the Joint Centre for Structural Genomics and one of the seven structural genomics targets chosen for the inaugural Automated Protein Function Prediction Assessment (APFPA) exercise 2005; (3) hypothetical protein Pa94 from *Pseudomonas aeruginosa* (termed APC22056 by Midwest Centre for Structural Genomics); and (4) PsbP protein, a regulator of the photosystem II complex from higher plants. There is little sequence similarity between these four proteins but they do show strong structural similarity (Figure 5) that indicates divergent evolution. Mog1p, Pa94 and PsbP are classified in SCOP[77] as having a Mog1p/PsbP-like fold. TM1622 has yet to be classified formally. In terms of function, Mog1p is known to bind GTP and GTPases, including Gsp1p, a homolog of mammalian Ran, the Ras family GTPase.[78,79] Mog1p may also play a role in yeast SLN1-SKN7 signal transduction, regulating the Skn7p transcription factor in response to osmotic stress through binding of three proteins, Sln1p, Skn1p and Ypd1p.[80] It is not known whether these three proteins can bind simultaneously or not so up to four binding sites may exist on the Mog1p surface. PsbP is unlikely to bind Ran, but does bind to other GTPases.[81] The exact functions of TM1622 and Pa94 remain unknown but since they are likely to bind other proteins in the form of GTPases they provide ideal targets for protein–protein binding site prediction. Therefore, one aspect of this work was to first predict binding sites on the best characterised protein, Mog1p, and then use these predictions as a reference for predictions on the other members of the family. In this way, any functional similarities between members of the family could be inferred.

Throughout this work, we restricted ourselves to using the fully annotated Swissprot release 46 database as a source of sequence homologues. However, insufficient numbers of homologues were detected in this database for Mog1p, TM1622, or Pa94 in order for us to calculate useful conservations scores, and so predictions on these three proteins were made without reference to evolutionary information. As such, these made good test cases for the ability of our method to predict on the basis of other information. If the larger Uniprot[82] database, containing a number of un-annotated sequences, is considered then homologues can be found for Mog1p and Pa94 (but not TM1622). This enabled us to compare our predictions on these two proteins (made without access to evolutionary information) with those of Consurf,[83] a state-of-the-art method for identifying functionally important residues from multiple sequence alignments with the option of both a Swissprot and, more importantly for our purposes, a Uniprot[82] sequence search. In the case of Mog1p, we also compared our predictions with conserved residues found manually by Stewart & Baker.[79]

Mog1p

Stewart & Baker[79] observed a concave cluster of conserved residues between residues 30–70 and identified this as a putative Ran binding site on Mog1p. By multiple sequence alignment of Mog1p with a *Schizosaccharomyces pombe* homologue and three expressed sequence tag (EST) sequences from human, mouse and *Caenorhabditis elegans*, seven invariant and seven conserved surface residues were found in this region. Later studies revealed that mutations to either one of two conserved acidic residues Asp62 or Glu65 prevent Ran binding.[84] The putative Ran binding site is flanked on one side by a conserved ridge that separates the Ran binding site from another concave surface patch containing a

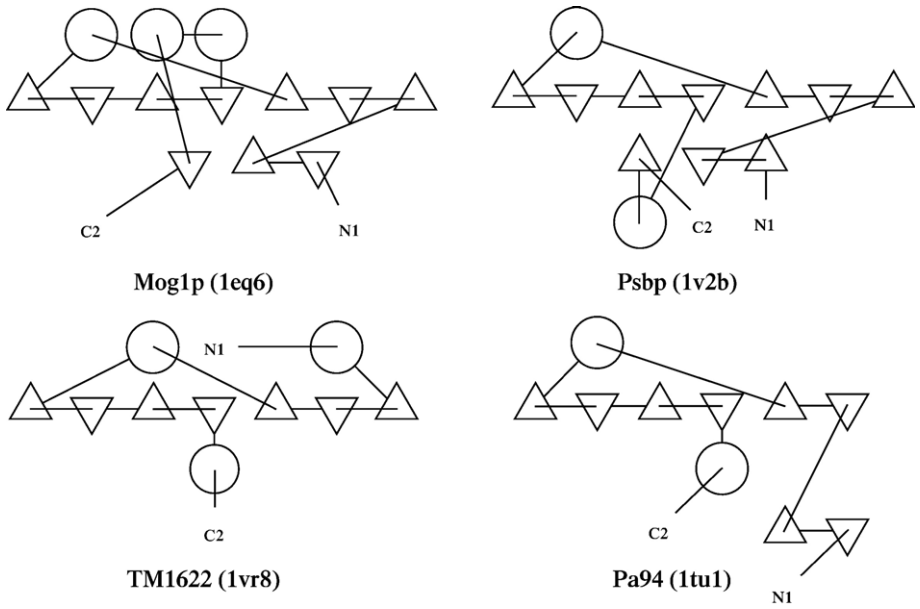

Figure 5. TOPS diagrams[85] of the four members of the Mog1p family. Note that any small secondary structures such as single turn helices or isolated β-bridges were removed for clarity.

second cluster of conserved residues. In addition to these, a further two clusters of conserved residues can be found on the surface of Mog1p.

We predicted the top three most probable binding site patches on the surface of Mog1p (PDB code: 1eq6[79]) using our trained naïve Bayesian network. From each patch we only considered the core residues (residues in which over 50% of the corresponding surface vertices were patch vertices) and compared them to the conserved and invariant residues identified by Stewart & Baker[79] (note that the *S. pombe* homologue used by Stewart & Baker,[79] even though detected in Swissprot, was seen by our automated method as providing insufficient information to calculate useful conservation scores). Predicted patches were located on three of the four conserved clusters of residues on the surface of Mog1p. The third ranked patch included the proposed Ran binding site capturing three of the seven conserved residues and two invariant residues including Glu65. Interestingly, probability score for this patch decreased from 68% to 65% in the E65A mutant[84] unable to bind Ran (PDB code: 1jhs), suggesting that predictions can be sensitive even to single mutations at an interface. The top ranked patch (probability score: 93%) included three conserved residues and one invariant residue and covered the concave patch separated from the Ran binding site by the invariant loop.

Little is known about the binding sites involved in SLN1-SKN7 signal transduction except that they are located beyond residue 78.[80] Some of the residues contributing to the top ranked patch were within the 30–70 residue range; however, the second ranked patch (probability score: 83%) was located at a significant cluster of conserved residues beyond residue 78. Indeed, seven of the 13 residues contributing to the core of the second ranked patch were conserved of which one was invariant. Considering that the Bayesian network made this prediction without use of this evolutionary information, we propose that the second ranked patch represents a potential binding site for one of the three SLN1-SKN7 signal transduction proteins.

Overall we captured 12 of the 46 conserved, and four of the seven invariant surface residues observed by Stewart & Baker.[79] If we include peripheral as well as core predicted residues these numbers increase to 18 conserved and all seven invariant residues. Therefore, even though our predicted patches together only covered 34% of the surface, we retained all the invariant residues and three of the four conserved clusters of residues. It should be noted that a number of unpredicted conserved residues are part of the "ridge" between the top and third ranked patches suggesting that these have a role in maintaining the integrity of one or both of these possible binding sites.

We also compared our predicted patches to the residues considered by the Consurf[83] web server to be functionally important. Using a multiple sequence alignment as input, Consurf implements the Rate4Site algorithm,[87] an extension of the evolutionary trace method devised by Lichtarge *et al.*,[29] to build an evolutionary tree and calculate a conservation score for each residue position. Each score is normalised so that the average score for all residues is zero, and the standard deviation is one. From these scores, nine levels of conservation are derived. Here we considered residues undergoing the slowest rates of evolution in levels 7–9 as "conserved". In order to acquire sufficient numbers of sequence homologues for the multiple sequence alignment (MSA) we used Uniprot[82] as the source sequence database. With Consurf default parameters, 39 unique sequence homologues were found. From the MSA,

52 of the 160 (33%) surface residues on Mog1p were calculated to be conserved. There was considerable agreement between the study by Stewart & Baker[79] and Consurf despite the larger numbers of sequences acquired by the latter, indeed all seven invariant surface residues found by Stewart & Baker[79] were calculated as conserved by Consurf. All three of the patches predicted to be binding sites by our method, operating without the use of evolutionary information, were enriched with Consurf conserved residues. Considering both the core and periphery of the patch, 80%, 52% and 63% of residues in the top, second and third ranked patches, respectively, were considered functionally important by Consurf. Overall, 73% (38/52) of Consurf conserved surface residues were captured by the three patches and 92% (22/24) of the most conserved (level 9) residues. These results further support our hypotheses that the second ranked patch is indeed a binding site for one of the three SLN1-SKN7 signal transduction proteins, and that the third ranked patch includes the Ran binding site. The strong correlation between the top ranked patch and conserved residues also suggests this patch is of functional significance.

TM1622, Pa94 and PsbP

Despite the striking structural similarities between each member of the Mog1p family (Figure 5), there was little evidence from our binding site predictions that they shared interaction partners.

Structural alignment with SSM[86] revealed that the majority of the binding sites were predicted at non-equivalent positions between each protein. The exceptions were approximately 25% surface area overlap between the top ranked patch of Mog1p and that of TM1622 (PDB code: 1vr8), and 75% overlap between the same patch on Mog1p and the second ranked patch on Pa94 (PDB code: 1tu1).

The predictions on TM1622 and Pa94 were carried out without evolutionary information, since no sequence homologues were detectable in the Swissprot database. However, for Pa94 19 unique sequence homologues were found in Uniprot[82] so, as with Mog1p, we were able to compare our method with Consurf in this instance. Despite Consurf only finding 25% (35/138) of surface residues on Pa94 to be functionally important, 91% (32/35) of these residues were included in our top three predicted patches which together covered only 56% (77/138) of the total number of surface residues. Given this strong agreement between our method and Consurf, we suggest that our patches represent functionally important regions on the surface of Pa94 and possible protein-protein binding sites.

Predictions on PsbP (PDB code: 1v2b[81]) were carried out using evolutionary information, since we identified numerous homologues for this protein in Swissprot. All the top three ranked patches included a hydrophobic pocket. The pocket of the top ranked patch (probability score: 96%) was less well conserved than its surrounding residues whilst the pockets on the other two patches (probability scores: 89% and 85%) were conserved. It is possible that one of these patches represents an interface to another protein in the photo-system II complex, although the complex form of PsbP has yet to be elucidated experimentally.

No patches equivalent to the Ran binding site patch on Mog1p were found on TM1622, Pa94 or PsbP. Furthermore, all predicted patches lacked the acidic residues essential to Ran binding on Mog1p. This suggests either the other family members do not bind Ran, or Ran has evolved different modes of binding for each of the proteins. The latter may be possible because Ran is known to bind NFT2 and importin β at different places on its surface.[88,89]

These results demonstrate the potential for a range of functions and interactions even between proteins of similar structure. In cases such as these, binding site prediction methods such as ours are essential, since function cannot be inferred from either sequence or structural homology. Details of the residues involved in all the predicted patches on Mog1p, Pa94, TM1622 and PsbP are given in Supplementary Data, Table 4.

Locating potential drug targets

Many human diseases result from abnormal protein–protein interactions so finding drugs that inhibit these interactions is of critical importance. Despite the large and often hydrophobic areas involved, it is now thought possible that small molecules can inhibit protein–protein interactions,[1,2] particular in light of the recent advances in screening techniques.[90,91] As an aid to the screening process, one could also locate potential sites of interaction on a protein surface *in silico* using a prediction program such as ours and then design inhibitors that bind at these sites. Use of a reliable computational method in this way to complement "wet-lab" studies would be more cost effective than experimental work alone. A recent review on protein–protein interactions in human disease[92] provides some ideal test cases to assess the ability of our method to predict possible drug targets on protein surfaces. Here, we concentrate on the protein–protein interaction network involved in papilloma virus infection.

Papilloma viruses are double-stranded DNA viruses that invade the basal layer of epithelial cells to cause benign lesions and cancer in higher eukaryotes.[93,94] Once within the basal layer, their genome is maintained and replicated as an extra-chromosomal plasmid. The virus persists by replicating and then segregating this plasmid into the nuclei of the daughter cells of the epithelium using a network of protein–protein interactions involving viral E1 and E2 proteins, and the human protein Brd4. The E2 protein consists of a C-terminal domain that binds DNA as a dimer,[95] and a regulatory N-terminal *trans*-activation domain that binds viral E1 protein.[96] E2 also targets the cellular protein Brd4, which anchors the viral plasmid to chromosomes of the host cell.[97,98] Inhibition of any one of these protein–protein interactions should

perturb viral replication and thus curb the infection. We therefore used our Bayes prediction strategy to predict potential binding sites on the surface of both the E2 *trans*-activation (PDB code: 1r6k[99]) and DNA binding (PDB code: 1jj4[95]) domains. Note that the E2 *trans*-activation domain was in unbound form.

Results were encouraging (Figure 6). On the E2 *trans*-activation domain (Figure 6(a)), the top ranked patch covered the E1 helicase/E2 *trans*-activation domain interaction site, including residues Tyr19 and Glu39 that are critical for binding.[99] More specifically, the patch covered the interaction site of the peptide inandione, an inhibitor of the E1/E2 interaction,[99,100] with 40% precision and 92% binding site coverage. Precision was low because the calculation was based on a small molecule binding site covered by a large protein–protein interaction patch; precision on the whole E1 binding site would have been higher. Figure 6(c) shows two possible positions of the inhibitor, A and B, found in the crystal structure. Inhibitor B is most likely the *in vivo* position of the inhibitor as inhibitor A binds only weakly with a secondary binding pocket.[99] The binding pocket of inhibitor B can only be found on the bound form of E2, which makes our predictions on the unbound form even more impressive.

The third ranked patch was located at the interaction site between the DNA binding domain and the E2 *trans*-activation domain. Thus two of the three highest ranked patches were located at a site of functional importance on the surface of the E2 *trans*-activation domain. On the E2 DNA binding domain (Figure 6(b)), the top ranked patch was located at the site of dimerisation, a process required for E2 binding to DNA.[95] These results strongly suggest that our protein–protein interface method can help locate critical regions on a protein surface that can be targeted by drugs, especially in light of the recent evidence to suggest that small molecule inhibitors can make effective protein–protein interaction blockers.[1,2] We compared our results with Promate,[33] a web server dedicated to protein–protein binding site prediction. Promate predicted one patch of four residues close to the E1 helicase binding site; however, only one of these residues actually formed part of the peptide inandione interaction site.

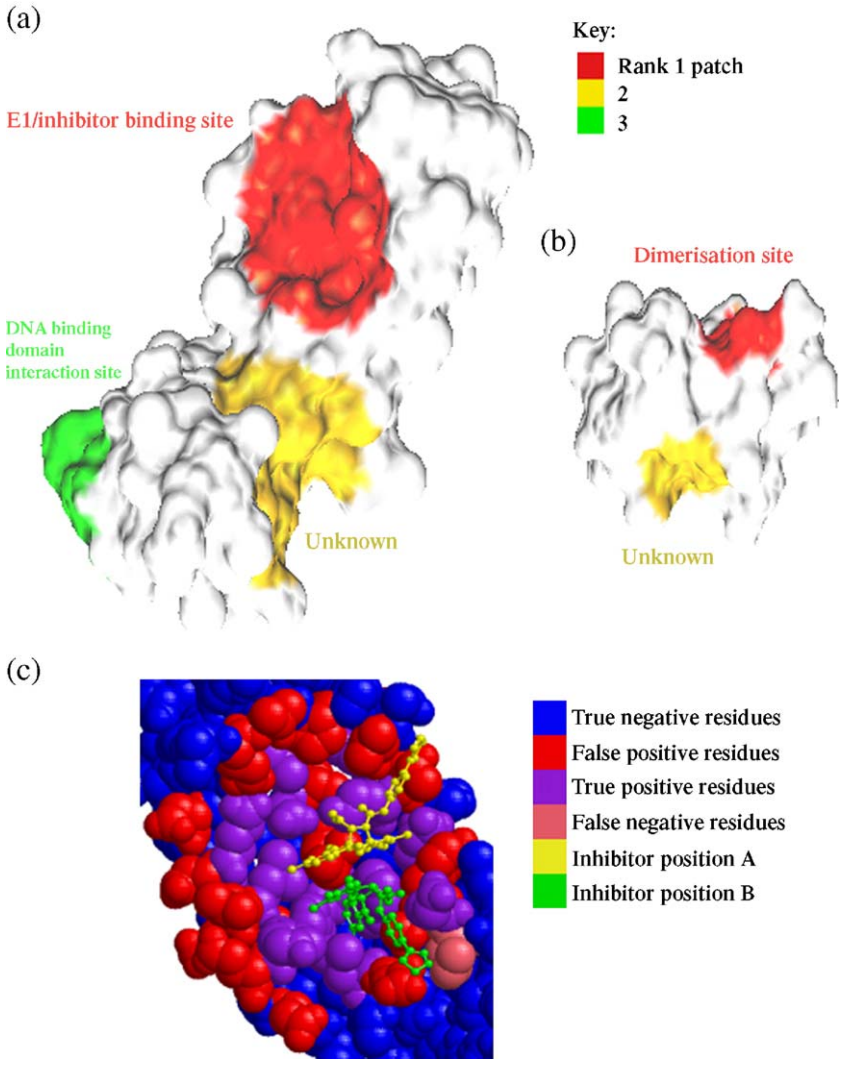

Figure 6. Predicted patches on papilloma virus E2 protein. (a) *Trans*-activation domain (PDB code: 1r6k); (b) DNA binding domain (PDB code: 1jj4); (c) detail of the inhibitor peptide inandione interaction site.

Distinguishing non-obligate interfaces from obligate interfaces: a second Bayesian network

Pre-selection of discriminatory properties

We tested a total of eight properties for their ability to distinguish obligate from non-obligate interacting patches: hydrophobicity, conservation, electrostatic potential, ASA, shape index, curvedness, secondary structure and patch size. As in the binding site patch prediction Bayesian network, two nodes (mean and standard deviation) were calculated from hydrophobicity, conservation, electrostatic potential, ASA, shape index and curvedness. Single nodes were calculated from secondary structure and patch size. Therefore, excluding the class node, a maximum 14 node Bayesian network was possible.

Secondary structure and interacting patch size nodes were added on the basis of previous work by Mintseris & Weng,[44] and De *et al*.[43] Secondary structure was assigned by Stride[101] and used as a discrete node that could take one of four values dependent on the majority of secondary structure comprising the patch: α−helix (discrete value = 1), β-sheet (2), other (3), or mixed (4) if no single secondary structure element constituted over 50% of the patch. Patch sizes were equivalent to 6% of the size of the surface area of the protein under test therefore no knowledge other than the size of the query protein was required. Size was expressed as the radius of the patch and was treated as a continuous node. We did not normalise this node as it is difficult to assign a maximum possible limit to the radius of a protein–protein interface.

Two further changes were made from the Bayesian network we used to predict interacting patches. First, we discarded residue interface propensity, which was used specifically to identify interface regions (it would have been counter-intuitive to use it on a data set consisting of only interacting patches). Second, we used un-normalised conservation scores prompted by the work of Mintseris & Weng[44] who noted that obligate interfaces evolve more slowly than non-obligate interfaces. The normalised conservation scores we used in the binding site patch prediction Bayesian network were relative to the other conservation scores within the same protein, meaning that the highest scoring residue was assigned a score of 1.0 regardless of its actual conservation score. Therefore, they were useful in distinguishing the more conserved interacting patches from non-interacting patches on the same protein. Here we attempt to distinguish obligate from non-obligate interfaces that occur on separate proteins so any evolutionary differences between these two types of patches would have been obscured by normalisation.

Assessing individual property contributions

Eight naïve Bayesian networks based on one property were trained on our "interaction type" data set consisting of 132 patches taken from the 180 interacting patches described above: 66 non-obligate patches and five subsets of 66 patches chosen at random from the 114 obligate patches. This ensured a balanced training set of equal numbers of each interaction type. We therefore carried out five training runs per property and plotted ROC curves and "gross AUCs" from the mean data. Note that specificity and sensitivity measures for the ROC curves were derived as if non-obligates were labelled as positive and obligate interfaces negative.

As with the binding site patch prediction Bayesian networks, with such a small data set there was always a danger of overfitting, therefore it was important to measure the significance of our gross AUC values. Each Bayesian network was trained on the same five training sets but with half of the obligate interacting patches randomly assigned as non-obligate, and *vice versa*. The extent to which baseline AUC values obtained on this data exceeded 0.50 indicated the level of potential overfitting of the single property model. The net AUC value was calculated by subtracting baseline AUC from gross AUC. Results for each individual property are shown in Table 3.

Patch size was the best discriminator achieving a net AUC of 0.24 and gross AUC of 0.76±0.03. These results appear to agree with previous studies that have shown that larger interfaces are usually obligate.[42,43] What is surprising is that our patch sizes were estimates of interface size based on the size of the query protein and so our results suggest that obligate and non-obligate interactions can be distinguished at this accuracy level solely on the basis of protein size. Electrostatic potential was the next best performing property achieving an AUC of 0.70±0.03. This may have been due to the prevalence of enzyme–inhibitor interfaces in our non-obligate set where electrostatic interactions are critical. Conservation also performed well appearing to support the findings by Mintseris & Weng[44] that obligate interfaces evolve more slowly than non-obligate interfaces. Encouragingly, net AUC achieved with un-normalised conservation score was 0.11 in contrast to 0.02 with normalised score so justifying our decision to use un-normalised conservation in the interface type Bayesian network.

Table 3. Assessing the ability of different properties/property combinations to classify interaction type

BN variable	Gross AUC	+/−	Random AUC	+/−	Net AUC
Patch size	0.76	0.03	0.52	0.02	0.24
Electrostatic potential	0.70	0.03	0.55	0.02	0.15
Conservation	0.69	0.02	0.57	0.02	0.11
Secondary structure	0.67	0.02	0.58	0.03	0.09
Shape index	0.62	0.03	0.55	0.01	0.07
Curvedness	0.63	0.02	0.56	0.03	0.06
ASA	0.64	0.02	0.59	0.03	0.05
Hydrophobicity	0.61	0.02	0.59	0.03	0.02

AUC, area under ROC curve.

ROC curves illustrating the discriminatory powers of patch size, electrostatic potential and conservation are shown in Figure 7. Also plotted is the probability threshold as a function of false positive rate. In order to derive the probability threshold (p) at which false positives and false negatives have an equal cost (not necessarily $p=0.50$; see earlier), we identified the point on the ROC curve where the gradient equals one (and is closest to (0, 1)). This gave the following: $p=0.43$ for patch size, $p=0.53$ for electrostatic potential and $p=0.58$ for conservation. In particular, the electrostatic potential and conservation probability curves involved extensive plateaus from around 0.6 to 0.5 therefore any slight change to the threshold in this region would cause a large change in both specificity and sensitivity. This illustrates the danger of evaluating performance at just one threshold.

There was some useful information provided by secondary structure although net AUC was 0.09. De *et al.*[43] found that involvement of defined secondary structure elements such as β-sheets and helices is much more common across subunits at an obligate interface than a non-obligate interface. Our data support these claims: 70% of the obligate interfaces were predominantly α-helix, β-strand or "mixed" compared to only 52% of the non-obligate interface (Table 4). Interestingly, only 11% of the obligate interfaces were predominantly β-strand whereas 45% were α-helix. In contrast, 48% of non-obligate interfaces were predominantly loops or other less well-defined secondary structure. Despite these differences, it appeared that the distinguishing power of secondary structure was limited overall.

Low net AUC values achieved by shape index, curvedness and ASA suggested little significant difference in topography between obligate and non-obligate interfaces. Perhaps surprisingly hydrophobicity was the poorest discriminator, since permanent interfaces, making up the majority of the obligate set, are characterised by a larger, hydrophobic interfaces than transient interfaces constituting the majority of the non-obligate set.[7,11] One of the reasons for this may be our use of patches at the interface centre that are frequently smaller than the actual interface in order to achieve high precision in binding site patch predictions. Therefore, our results imply that the cores of the two interface types have similar hydrophobicity levels and the difference between obligate and non-obligate interfaces is the number of non-polar contacts between binding

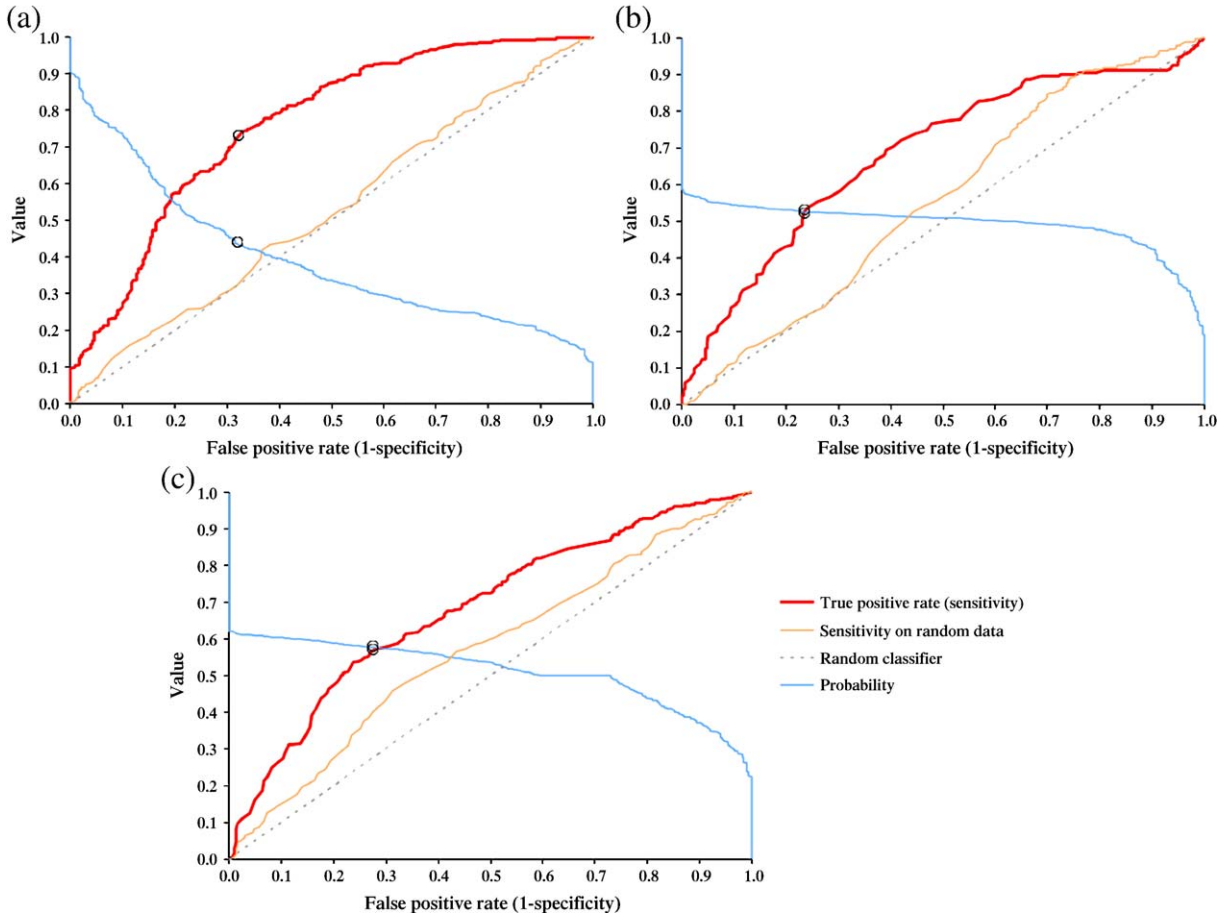

Figure 7. ROC curves to assess training performance of three interface type Bayesian networks consisting of (a) patch size, (b) electrostatic potential or (c) conservation nodes. Also shown are changes in probability threshold. Circles indicate gradient equal to one on ROC curve and the classification threshold giving equal costs on the probability curve.

Table 4. Secondary structure composition of an obligate versus a non-obligate interface

Secondary structure	Interface type	
	Obligate	Non-obligate
α Helix	0.45	0.27
β Strand	0.11	0.21
Other	0.30	0.48
Mixed	0.15	0.03

sites[43] rather than mean hydrophobicity over a given surface area. This is consistent with groups who have found that obligate interfaces are characterised by large hydrophobic patches.[5,9,11]

The Bayesian network consisting of all 14 nodes achieved a gross AUC of 0.84±0.02 but baseline AUC increased to 0.70. Therefore any discriminatory power gained by combining properties was probably a result of overfitting. Even combining electrostatic potential and patch size reduced net AUC to 0.20.

We chose to differentiate obligates from non-obligate interfaces because they are more descriptive of protein function whereas the terms "permanent" and "transient" describe the timescale of association. However, our results reflect the difficulties of assigning interface type even manually and there was a strong possibility that we were not discriminating obligates from non-obligates but rather transient and permanent interfaces as it is these that made up the majority of the non-obligate and obligate sets, respectively. For example, the good performance of the patch size node may have been due to permanent interfaces being generally larger than transient interfaces. Furthermore, the apparent difference in electrostatic potential properties between obligate and non-obligate interfaces may be solely due to the number of enzyme-inhibitor interfaces in the non-obligate set and not a general difference between the interface types. A larger data set is required before stronger conclusions can be made.

Further studies with the binding site patch prediction Bayesian network

Heterogeneous cross-validation

Given the weak performance of the interface type Bayesian network we investigated the possibility of using obligate interface information to predict non-obligate interfaces, and *vice versa* using the original set of 14 nodes (not secondary structure or patch size). To do this we carried out heterogeneous cross-validation in which we trained the binding site patch prediction Bayesian network shown in Figure 1(a) to distinguish obligate interacting patches from non-interacting patches and then used the model to predict binding site patches on the non-obligate complex types, and *vice versa*. Whichever interaction type was removed from the training set was also left out of the interface residue propensity calculation as well. As in previous work,[38] we found it possible to use non-obligate interface information to predict obligate interfaces with a high success rate of 83% (95/114), and *vice versa* with a success rate of 80% (53/66). The latter success rate was slightly better than the rate of 74% (49/66) obtained when non-obligate interface properties were used to predict non-obligate interfaces by leave-one-out cross-validation within the non-obligate training set. Normalisation of the nodes may have had the effect of suppressing the differences, if any, between non-obligate and obligate interface types although we believe that this was only significant with regard to conservation scores (see above).

Our results suggest the physical–chemical properties at non-obligate and obligate interfaces, particularly in small patches that include the "core" of the interface, are similar to such an extent that one can use information from one type of interface to predict the other. This explains our lack of success at distinguishing non-obligate interfaces from obligate interfaces. Perhaps the characteristics of an obligate interface provide stronger interface detection signals than non-obligate interface properties, hence the improved prediction of non-obligate interfaces with obligate information. In other words, the difference between an obligate interface and the rest of the protein surface is more marked than on a non-obligate complex even if the nature of the differences is similar. The strength of this "binding site signal" could govern the stability of the contacts between the two binding sites involved at the interface, which in turn govern the nature of the interaction, with obligate interactions more stable than non-obligate contacts.

Binding affinity versus prediction success

So far we have treated obligate and non-obligate interaction types as two discrete subsets in our benchmark of protein complexes.[38] However, a continuum probably exists between the two interaction types.[3] We therefore considered our predictions in relation to theoretical binding affinity of each protein complex as calculated by Dcomplex.[102] As expected, all the non-obligate complexes exhibited lower binding affinities of less than 30.0 kcal/mol (binding free energy of −30.0 kcal/mol) suggesting that most non-obligate interactions are transient. However, obligate complexes exhibited a range of binding affinities between 7.9 to 132.0 kcal/mol, although the majority of these were greater than 25.0 kcal/mol. This was probably because our obligate data set consisted of both permanent and transient interactions.

In order to study the relationship between prediction success and binding affinity, we identified only those proteins upon which prediction success was achieved in all five leave one-out cross-validation rounds and called these "hits", a protein with a top ranked binding site patch in all five rounds was called a "top hit". A total of 75% of proteins in complexes with less than 30.0 kcal/mol binding affinity were hits and 48% were top hits in contrast to 88% and 56%, respectively, of proteins in

complexes with binding affinity above 30.0 kcal/mol (all obligate complexes). Therefore, as binding affinity increases, the probability of achieving prediction success increases. There are probably a number of reasons for this. Binding affinity is strongly correlated ($R = 0.94$) with interface size so in most cases large interfaces give high binding affinities. Intuitively, a larger binding site means that more patches with over 50% precision and 20% interface coverage will be available for the Bayesian network to recognize as interacting patches, especially when one considers that patch sizes are usually smaller than the actual interface in these cases. More interesting perhaps is that binding sites in high affinity complexes may provide a stronger binding site signal, hence higher probability interacting patches. Indeed, there is a correlation between binding affinity and probability values of patches at binding sites (Figure 8), with low probability patches much less likely to occur at high affinity binding sites.

Conclusion

In this work, we have devised a method to predict both protein–protein binding site location and interface type (obligate or non-obligate) using a Bayesian network in combination with surface patch analysis. We trained two Bayesian network structures to distinguish between interacting and non-interacting surface patches taken from a benchmark dataset of 180 proteins and found no significant performance advantage in adding extra connections to a simple naïve Bayes classifier. We therefore carried forward the naïve Bayes classifier to the prediction phase. This simple classifier achieved a success rate of 82% in homogenous leave- one-out cross-validation which was significantly better than 76% achieved with an SVM in previous work.[38] Unlike a general functional site predictor such as those that use only evolutionary information, our method is specific to protein–protein binding sites as demonstrated by a small study on a test set of 33 proteins having both a DNA and protein binding site on their surfaces. Predicted patches overlapping the DNA binding site were higher ranked than the protein binding site in only three of the 33 cases, and protein binding sites were predicted with an 89% success rate.

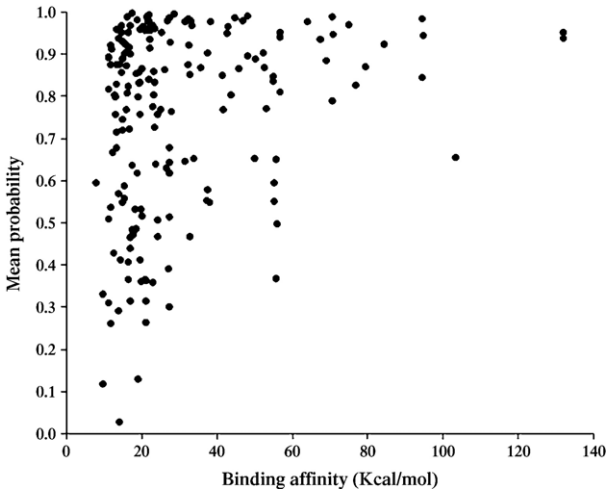

Figure 8. Mean probability is calculated from the ten patches with highest precision and interface coverage in each protein of the training set. Note that only positive examples are represented.

We also tested the ability of the Bayesian network to handle missing data. Overall performance of the Bayesian network without to access evolutionary information on the test protein was comparable to that of the original classifier, indeed some predictions benefited from having access to less information. In a case study of four proteins in the Mog1p family sharing the same fold, two of which have been structural genomics targets, we demonstrated that our Bayesian network method can provide important clues as to the functionally important areas on a protein surface even when the test protein has no detectable sequence homology to any known protein. On Mog1p we detected a putative binding site involved in the SLN1-SKN7 signal transduction pathway, and predicted a Ran binding site, previously characterized solely by conservation studies, even without reference to evolutionary information. There was little correspondence between predicted binding sites on other members of the Mog1p family and those on Mog1p itself suggesting that each of the four proteins is involved in different protein interactions or functions despite the overall structural similarities. Finally in the prediction phase, we demonstrated our method's applicability to the drug discovery process by successfully locating a number of binding sites involved in the protein–protein interaction network of papilloma virus infection. These results are particularly significant in light of recent evidence that protein-protein interactions make drugable targets.[1,2]

Distinguishing obligate from non-obligate binding sites using a second Bayesian network proved more difficult, although some separation was achieved on the basis of patch size, electrostatic potential and conservation. This result may have been due to our use of interacting patches that were frequently smaller than the actual interface. For example, interfaces of obligate complexes are characterised by large hydrophobic patches[5,9,11] but whereas the number of non-polar contacts across the two interface types may differ significantly,[43] the mean hydrophobicity values across the central region of the binding site may still be comparable. Such was the similarity in the two interacting patch types, we were able to use obligate binding site properties to predict the location of non-obligate binding sites and *vice versa*. Indeed, results on non-obligate binding sites actually improved if only obligate information was used as oppose to a mixture of non-obligate and obligate information. This led us to believe that the difference between an obligate interface and the rest of the protein surface is more marked than on a non-obligate complex even if the nature of the differences is similar.

Overall we have demonstrated that our method can not only predict protein–protein binding sites with high success rate but can also be used to provide valuable insights into the properties that characterise them.

Materials and Methods

Training set

The details of our benchmark training set of 180 proteins have been described.[38] A comprehensive set of complexes was chosen from the Protein Data Bank[103] (PDB) and then subjected to a number of stringent filtering steps. Proteins sharing over 20% sequence identity with a higher resolution structure (or the most recently determined structure if resolutions were equal) of the same complex type were removed. Evidence in the literature had to exist that the complex occurred naturally and was stable as a dimer so we eliminated interfaces only present as a result of crystal packing. NMR structures were not used, nor were mutant complexes or structures whose resolution was worse than 3.0 Å. Fragments were allowed unless the interface was severely truncated, but dimers containing a protein of less than 20 residues were discarded. Complexes whose interfaces were made up of more than one separate chain or complexes containing more than one binding site of the same type were also removed, as well as complexes containing broken interfaces where one protein contacts the other at two points. As far as possible we aimed to include only proteins involved in dimeric interactions to reduce the possibility of a non-interacting patch (see below) containing residues involved in an interaction different to the one given in the PDB file. A total of 180 proteins taken from 149 complexes survived the filtering process of which 36 were involved in enzyme–inhibitor interactions, 30 in "other" non-obligate interactions, 27 in hetero-obligate interactions and 87 homo-obligate interactions. Homo-obligate complexes were classed as such if their subunits shared over 80% sequence identity; only the subunit with the largest binding site was retained. A complete list of proteins in the training set is given in Supplementary Data, Table 1.

Molecular surface generation and interface definition

We computed the solvent excluded surface (SES) of each individual protein and/or complex with the MSMS surface generation program[104] using a probe radius of 1.5 Å. An atom was defined as part of the interface if it lost more than 99% of its solvent accessible surface area upon complex formation. Any atom not allocated to the interface was deemed part of the non-interacting surface. A list of interface residues for each protein in the training set is listed in Supplementary Data, Table 2.

Bayesian networks

Bayesian networks are probabilistic graphical models, which provide a compact representation for expressing joint probability distributions and for inference. A set of variables $x=\{X_1,..,X_n\}$ can be represented as nodes of the Bayesian network, with relationships between the variables represented as directed edges connecting related nodes. This defines a graph structure. To allow for efficient inference and learning, a directed acyclic graph (DAG) must be formed. The graph structure is chosen to exploit the conditional independence between the variables; this provides a concise representation in terms of simple component distributions (factors), reducing the number of parameters to be estimated. A naïve Bayes classifier has a structure whereby the class variable C is a parent to each variable X_i, which are each treated as being independent. Model parameters in the form of conditional probability distributions (CPDs) between a variable and those it depends on are learned from the data, that is, for the naïve Bayes classifier we learn the CPDs $p(X_i \mid C)$.

Implementation

The Bayesian networks were implemented using the Bayesian Network Toolbox for Matlab[105] (BNT). All variable distributions with regard to mean and standard deviation of each property and patch radius were approximately Gaussian (data not shown) allowing us to treat their corresponding nodes as continuous. The class nodes and the secondary structure node were discrete.

Bayesian network structures were learnt using the BNT structure learning package[106] (SLP) with the maximum weight spanning tree and greedy search algorithms.

Learning

One advantage of using a Bayesian network is that it is possible to learn the model parameters from data when the training data sets are incomplete. For example, if a protein has no known homologues, the two conservation score nodes in Figure 1(a) will have no data associated with them. To learn from incomplete data, we use the Expectation-Maximisation (EM) algorithm, which estimates the missing values by computing the expected values and updating parameters using these expected values as if they were observed values.

Inference

The joint probability distribution of the naïve Bayesian network shown in Figure 1(a) can be expressed as:

$$p(C, X_1, \ldots, X_n) = p(C) \prod_i p(X_i|C)$$

Where C represents the "binding site patch?" node (more generally C represents the classes that we are trying to classify into), and X_i represent the other nodes in the network. From the definition of conditional probability:

$$p(C|X_1, \ldots, X_n) = \frac{p(C, X_1, \ldots, X_n)}{p(X_1, \ldots, X_n)}$$

Since we are assuming all of the variables (X_i) are independent (i.e. $p(X_1,..,X_n)=p(X_1)...p(X_n)$), the predictive distribution of C given the variables X_i can be written:

$$p(C|X_1, \ldots, X_n) = p(C) \prod_i \frac{p(X_i|C)}{p(X_i)}$$

The learned model enables predictions to be made about the expected states of variables even if evidence for one or more variables is missing. For example, if the test protein has no known homologues we must marginalise over the conservation score nodes in order to infer the state of the variable we are interested in. To marginalise

over the unknown variable X_j, (in the discrete case), the probability of belonging to class C incorporates a contribution for the probability of X_j being each of its possible values (x_j) given that particular classification C:

$$p(C|\text{all the } X_i \text{ except } X_j) = p(C)\prod_{i\neq j}\frac{p(X_i|C)}{p(X_i)}\left(\sum_{x_j}\frac{p(X_j=x_j|C)}{p(X_j=x_j)}\right)$$

A similar methodology is used for continuous variables except that an integral is taken over all possible values of the probability distribution $p(X_j \mid C)$ in place of the summation.

Surface patches

Patch properties

Approximately circular regions of protein surface of defined size known as patches formed the basis for much of this work. Each surface vertex, generated by MSMS,[104] within a patch was labelled with seven physical–chemical surface properties as described:[38] hydrophobicity calculated using the Fauchère & Pliska scale,[107] residue interface propensity, shape index and curvedness,[108,109] electrostatic potential calculated by Delphi,[110,111] and solvent accessible surface area (ASA) generated by MSMS.[104] Sequence conservation score was calculated by Scorecons,[112] which was locally installable and robust enabling high throughput processing of results. These properties were subsequently normalised between zero and one and then the mean and standard deviation of each property was calculated across the patch to produce the 14 Bayesian network variables listed in Figure 1.

Patch definitions

Interacting patch. We defined an interacting patch as an approximately circular region of protein surface directly involved in a protein–protein interaction. The centre of the interacting patch was the centre of geometry of the actual interface. The size of each patch was equivalent to 8% of the size of the smallest of the two proteins known to be involved in the interaction (see predicted patches definition, below, as to how we derived this figure).

Non-interacting patch. A non-interacting patch, of equivalent size to an interacting patch, was taken from the non-interacting parts of the surface. The centre of the non-interacting patch was chosen at random from the set of non-interacting surface vertices.

A more detailed explanation of the generation of interacting and non-interacting patches can be found in the work by Bradford & Westhead.[38]

Predicted patches. Predicted patches were generated as follows: for each protein subject to interface prediction, we generated enough patches for complete coverage of the protein surface (one patch per surface atom). The centre of each patch was the surface vertex closest to the centre of geometry of each surface atom. Patch sizes were estimated from a short study of the relationship between the size of the interface and the sizes of the two proteins within the complex.[38] For each test case, the sizes of the proteins and their interface were calculated in terms of number of surface vertices. Using linear regression, it was found that the interface size was equivalent to about 13% of the size of the smallest protein in the complex, and about 12% of the size of its parent protein. To avoid an excess of non-interacting vertices in the interacting patch, and because of the non-circularity of most interfaces, we favoured a conservative patch size that was less than the average values found above. Therefore, we set our patch size to 8% of the size of the smallest of the two proteins known to be involved in the interaction unless the binding partner was unknown, in which case we used a patch size equivalent to 6% of the surface area of the query protein.

Each patch was then assigned a probability value by the trained Bayesian network according to the likelihood that the patch was part of a protein–protein binding site. At this point we discarded any patch with a probability value below 0.50 and ranked the remaining patches according to probability, with patches most likely to be located at a binding site (and therefore having the highest probability values) ranked highest. We then removed overlapping patches by discarding any patch that shared more than 10% of its residues with a patch above it in the ranked list. The outcome was a set of non-overlapping patches ranked according to probability of being part of a binding site. These ranked patches were defined as our "predicted patches".

Expected number of successful predictions by chance

First, we calculated the probability, p, of finding a binding site patch (a patch with at least 50% patch precision and 20% interface coverage) at random amongst the set of patches generated for each test case:

$$p = \frac{\text{No. of binding site patches}}{\text{Total no. of patches}}$$

When considering the top three patches, we were, in effect, making three attempts at finding a binding site patch so, given p, we calculated the probability of succeeding at least once in these three attempts:

$$P(\text{at least one binding site patch in top three}) = 1-(1-p)^3$$

These P values allowed us to calculate the number of successful predictions (a patch with at least 50% patch precision and 20% interface coverage in the top three) one would expect (E) to achieve by chance across our data set (equation (4)).

$$E[\text{number of successful predictions}] = \bar{P}N \qquad (4)$$

where $\bar{P}$ is the mean of P across the dataset and N is the number of proteins in the dataset.

Acknowledgements

The project was supported by the BBSRC E-Science Initiative, grant number: BBS/B/16585. We thank Nick Burgoyne for useful discussions and would like to acknowledge the constructive criticisms and inputs of two anonymous reviewers who made this a better paper.

Supplementary Data

Supplementary data associated with this article can be found, in the online version, at doi:10.1016/j.jmb.2006.07.028

References

1. Arkin, M. R. & Wells, J. A. (2004). Small-molecule inhibitors of protein-protein interactions: progressing towards the dream. *Nature Rev. Drug Discov.* **3**, 301–317.
2. Arkin, M. R., Randal, M., DeLano, W. L., Hyde, J., Luong, T. N., Oslob, J. D. *et al.* (2003). Binding of small molecules to an adaptive protein-protein interface. *Proc. Natl Acad. Sci. USA*, **100**, 1603–1608.
3. Nooren, I. M. & Thornton, J. M. (2003). Diversity of protein-protein interactions. *EMBO J.* **22**, 3486–3492.
4. Chothia, C. & Janin, J. (1975). Principles of protein-protein recognition. *Nature*, **256**, 705–708.
5. Jones, S. & Thornton, J. M. (1996). Principles of protein-protein interactions. *Proc. Natl Acad. Sci. USA*, **93**, 13–20.
6. Jones, S. & Thornton, J. M. (1997). Analysis of protein-protein interaction sites using surface patches. *J. Mol. Biol.* **272**, 121–132.
7. Nooren, I. M. & Thornton, J. M. (2003). Structural characterisation and functional significance of transient protein-protein interactions. *J. Mol. Biol.* **325**, 991–1018.
8. Young, L., Jernigan, R. L. & Covell, D. G. (1994). A role for surface hydrophobicity in protein-protein recognition. *Protein Sci.* **3**, 717–729.
9. Glaser, F., Steinberg, D. M., Vakser, I. A. & Ben-Tal, N. (2001). Residue frequencies and pairing preferences at protein-protein interfaces. *Proteins: Struct. Funct. Genet.* **43**, 89–102.
10. Tsai, C.-J., Lin, S. L., Wolfson, H. J. & Nussinov, R. (1997). Studies of protein-protein interfaces: a statistical analysis of the hydrophobic effect. *Protein Sci.* **6**, 53–64.
11. Lo Conte, L., Chothia, C. & Janin, J. (1999). The atomic structure of protein–protein recognition sites. *J. Mol. Biol.* **285**, 2177–2198.
12. Larsen, T. A., Olson, A. J. & Goodsell, D. S. (1998). Morphology of protein-protein interfaces. *Structure*, **6**, 421–427.
13. Ansari, S. & Helms, V. (2005). Statistical analysis of predominantly transient protein-protein interfaces. *Proteins: Struct. Funct. Genet.* **61**, 344–355.
14. Xu, D., Tsai, C. J. & Nussinov, R. (1997). Hydrogen bonds and salt bridges across protein-protein interfaces. *Protein Eng.* **10**, 999–1012.
15. Bahadur, R. P., Chakrabarti, P., Rodier, F. & Janin, J. (2004). A dissection of specific and non-specific protein–protein interfaces. *J. Mol. Biol.* **336**, 943–955.
16. Chakrabarti, P. & Janin, J. (2002). Dissecting protein-protein recognition sites. *Proteins: Struct. Funct. Genet.* **47**, 334–343.
17. Crowley, P. B. & Golovin, A. (2005). Cation-π interactions in protein-protein interfaces. *Proteins: Struct. Funct. Genet.* **59**, 231–239.
18. Fernandez-Recio, J., Totrov, M., Skorodumov, C. & Abagyan, R. (2004). Optimal docking area: A new method for predicting protein-protein interaction sites. *Proteins: Struct. Funct. Genet.* **58**, 134–143.
19. Clackson, T. & Wells, J. A. (1995). A hot spot of binding energy in a hormone-receptor interface. *Science*, **267**, 383–386.
20. Bogan, A. A. & Thorn, K. S. (1998). Anatomy of hot spots in protein interfaces. *J. Mol. Biol.* **280**, 1–9.
21. Hu, Z., Ma, B., Wolfson, H. & Nussinov, R. (2000). Conservation of polar residues as hot spots at protein interfaces. *Proteins: Struct. Funct. Genet.* **39**, 331–342.
22. Ma, B., Elkayam, T., Wolfson, H. & Nussinov, R. (2003). Protein-protein interactions: structurally conserved residues distinguish between binding sites and exposed protein surfaces. *Proc. Natl Acad. Sci. USA*, **100**, 5772–5777.
23. Keskin, O., Ma, B. & Nussinov, R. (2005). Hot regions in protein–protein interactions: the organisation and contribution of structurally conserved hot spot residues. *J. Mol. Biol.* **345**, 1281–1294.
24. Valdar, W. S. J. & Thornton, J. M. (2001). Protein-protein interfaces: analysis of amino acid conservation in homodimers. *Proteins: Struct. Funct. Genet.* **42**, 108–124.
25. Bordner, A. J. & Abagyan, R. (2005). Statistical analysis and prediction of protein–protein interfaces. *Proteins: Struct. Funct. Genet.* **60**, 353–366.
26. Caffrey, D. R., Somaroo, S., Hughes, J. D., Mintseris, J. & Huang, E. S. (2004). Are protein-protein interfaces more conserved in sequence than the rest of the protein surface? *Protein Sci.* **13**, 190–202.
27. Grishin, N. V. & Phillips, M. A. (1994). The subunit interfaces of oligomeric enzymes are conserved to a similar extent to the overall protein sequences. *Protein Science*, **3**, 2455–2458.
28. Bradford, J. R. & Westhead, D. R. (2003). Asymmetric mutation rates at enzyme-inhibitor interfaces: implications for the protein-protein docking problem. *Protein Sci.* **12**, 2099–2103.
29. Lichtarge, O., Bourne, H. R. & Cohen, F. E. (1996). An evolutionary trace method defines binding surfaces common to protein families. *J. Mol. Biol.* **257**, 342–358.
30. Aloy, P., Querol, E., Aviles, F. X. & Sternberg, M. J. E. (2001). Automated structure-based prediction of functional sites in proteins: applications to assessing the validity of inheriting protein function from homology in genome annotation and to protein folding. *J. Mol. Biol.* **311**, 395–408.
31. Guharoy, M. & Chakrabarti, P. (2005). Conservation and relative importance of residues across protein–protein interfaces. *Proc. Natl Acad. Sci. USA*, **102**, 15447–15452.
32. Jones, S. & Thornton, J. M. (1997). Prediction of protein–protein interaction sites using patch analysis. *J. Mol. Biol.* **272**, 133–143.
33. Neuvirth, H., Raz, R. & Schreiber, G. (2004). ProMate: A structure based prediction program to identify the location of protein-protein binding sites. *J. Mol. Biol.* **338**, 181–199.
34. Zhou, H.-X. & Shan, Y. (2001). Prediction of protein interaction sites from sequence profile and residue neighbour list. *Proteins: Struct. Funct. Genet.* **44**, 336–343.
35. Zhou, H.-X. & Shan, Y. (2005). Prediction of interface residues in protein–protein complexes by a consensus neural network method: test against NMR data. *Proteins: Struct. Funct. Genet.* **61**, 21–35.
36. Fariselli, P., Pazos, F., Valencia, A. & Casadio, R. (2002). Prediction of protein-protein interaction sites in heterocomplexes with neural networks. *Eur. J. Biochem.* **269**, 1356–1361.
37. Ofran, Y. & Rost, B. (2003). Predicted protein-protein

interaction sites from local sequence information. *FEBS Letters*, **544**, 236–239.
38. Bradford, J. R. & Westhead, D. R. (2005). Improved prediction of protein-protein binding sites using a support vector machines approach. *Bioinformatics*, **21**, 1487–1494.
39. Chung, J.-L., Wang, W. & Bourne, P. E. (2006). Exploiting sequence and structure homologs to identify protein-protein binding sites. *Proteins: Struct. Funct. Genet.* **62**, 630–640.
40. Yan, C., Dobbs, D. & Honavar, V. (2004). A two stage classifier for identification of protein-protein interface residues. *Bioinformatics*, **20**, I371–I378.
41. Reš, I., Mihalek, I. & Lichtarge, O. (2005). An evolution based classifier for prediction of protein interfaces without using protein structures. *Bioinformatics*, **21**, 2496–2501.
42. Gunasekaran, K., Tsai, C.-J. & Nussinov, R. (2004). Analysis of ordered and disordered protein complexes reveals structural features discriminating between stable and unstable monomers. *J. Mol. Biol.* **341**, 1327–1341.
43. De, S., Krishnadev, O., Srinivasan, N. & Rekha, N. (2005). Interaction preferences across protein-protein interfaces of obligatory and non-obligatory components are different. *BMC Struct. Biol.* **5**, 15.
44. Mintseris, J. & Weng, Z. (2005). Structure, function, and evolution of transient and obligate protein-protein interactions. *Proc. Natl Acad. Sci. USA*, **102**, 10930–10935.
45. Mintseris, J. & Weng, Z. (2003). Atomic contact vectors in protein-protein recognition. *Proteins: Struct. Funct. Genet.* **53**, 629–639.
46. Needham, C. J., Bradford, J. R., Bulpitt, A. J. & Westhead, D. R. (2006). Inference in Bayesian networks. *Nature Biotech.* **24**, 51–53.
47. Husmeier, D., Dybowski, R. & Roberts, S. (2005). Editors of *Probabilistic Modelling in Bioinformatics and Medical Informatics*. Springer, London.
48. Beaumont, M. A. & Rannala, B. (2004). The Bayesian revolution in genetics. *Nature Rev. Genet.* **5**, 251–261.
49. Friedman, N., Linial, M., Nachman, I. & Pe'er, D. (2000). Using Bayesian networks to analyze expression data. *J. Comput. Biol.* **7**, 601–620.
50. Hartemink, A. J., Gifford, D., Jaakkola, T. & Young, R. (2001). Using graphical models and genomic expression data to statistically validate models of genetic regulatory networks. *Pac. Symp. Biocomput.* **6**, 422–433.
51. Hartemink, A. J., Gifford, D., Jaakkola, T. & Young, R. (2002). Combining location and expression data for principled discovery of genetic regulatory network models. *Pac. Symp. Biocomput.* **7**, 437–449.
52. Pe'er, D., Regev, A., Elidan, G. & Friedman, N. (2001). Inferring subnetworks from perturbed expression profiles. *Bioinformatics*, **17**, S215–S224.
53. Yoo, C., Thorsson, V. & Cooper, G. (2002). Discovery of causal relationships in a gene regulation pathway from a mixture of experimental and observational DNA microarray data. In *Procedings of the Pacific Symposium on Biocomputing*, pp. 498–509, World Scientific Publishing Co. Pte. Ltd., Singapore.
54. Tamada, Y., Kim, S., Bannai, H., Imoto, S., Tashiro, K., Kuhara, S. & Miyano, S. (2003). Estimating gene networks from gene expression data by combining Bayesian network model with promoter element detection. *Bioinformatics*, **19**, ii227–ii236.
55. Husmeier, D. (2003). Sensitivity and specificity of inferring genetic regulatory interactions from microarray experiments with dynamic Bayesian networks. *Bioinformatics*, **19**, 2271–2282.
56. Friedman, N. (2004). Inferring cellular networks using probabilistic graphical models. *Science*, **303**, 799–805.
57. Lee, P. H. & Lee, D. (2005). Modularized learning of genetic interaction networks from biological annotations and mRNA expression data. *Bioinformatics*, **21**, 2739–2747.
58. Nariai, N., Tamada, Y., Imoto, S. & Miyano, S. (2005). Estimating gene regulatory networks and protein-protein interactions of *Saccharomyces cerevisiae* from multiple genome-wide data. *Bioinformatics*, **21**, ii206–ii212.
59. Cai, D., Delcher, A., Kao, B. & Kasif, S. (2000). Modelling splice sites with Bayes networks. *Bioinformatics*, **16**, 152–158.
60. Castelo, R. & Guigo, R. (2004). Splice site identification by idlBNs. *Bioinformatics*, **20**, i69–i76.
61. Ben-Gal, I., Shani, A., Gohr, A., Grau, J., Arviv, S., Shmilovici, A. *et al.* (2005). Identification of transcription factor binding sites with variable-order Bayesian networks. *Bioinformatics*, **21**, 2657–2666.
62. Pudimat, R., Schukat-Talamazzini, E.-G. & Backofen, R. (2005). A multiple-feature framework for modelling and predicting transcription factor binding sites. *Bioinformatics*, **21**, 3082–3088.
63. Drawid, A. & Gerstein, M. (2000). A Bayesian system integrating expression data with sequence patterns for localizing proteins: comprehensive application to the yeast genome. *J. Mol. Biol.* **301**, 1059–1075.
64. Klinger, T. M. & Brutlag, D. L. (1994). Discovering structural correlations in a-helices. *Protein Sci.* **3**, 1847–1857.
65. Jansen, R., Yu, H., Greenbaum, D., Kluger, Y., Krogan, N. J., Chung, S. *et al.* (2003). A Bayesian networks approach for predicting protein-protein interactions from genomic data. *Science*, **302**, 449–453.
66. Troyanskaya, O. G., Dolinski, K., Owen, A. B., Altman, R. B. & Botstein, D. (2003). A Bayesian framework for combining heterogeneous data sources for gene function prediction (in *Saccharomyces cerevisiae*). *Proc. Natl Acad. Sci. USA*, **100**, 8348–8353.
67. Cohen, P. R. & Jensen, D. (1997). Overfitting explained. In *Preliminary Papers: Sixth Intl. Workshop on Artificial Intelligence and Statistics*, pp. 115–122.
68. Salzberg, S. L. (1997). On comparing classifiers: pitfalls to avoid and a recommended approach. *Data Mining Knowledge Discov.* **1**, 317–327.
69. Matthews, B. W. (1975). Comparison of the predicted and observed secondary structure of T4 phage lysozyme. *Biochim. Biophys. Acta*, **405**, 442–451.
70. Kimura, S. R., Brower, R. C., Vajda, S. & Camacho, C. J. (2001). Dynamical view of the positions of key side chains in protein-protein recognition. *Biophys. J.* **80**, 635–642.
71. Smith, G. R., Sternberg, M. J. E. & Bates, P. A. (2005). The relationship between the flexibility of proteins and their conformational states on forming protein–protein complexes with an application to protein–protein docking. *J. Mol. Biol.* **347**, 1077–1101.
72. Burgoyne, N. J. & Jackson, R. M. (2006). Predicting protein interaction sites: Binding hot-spots in protein-protein and protein-ligand interfaces. *Bioinformatics*, **22**, 1335–1342.
73. Lejeune, D., Delsaux, N., Charloteaux, B., Thomas, A. & Brasseur, R. (2005). Protein-nucleic acid recognition: statistical analysis of atomic interactions and

influence of DNA structure. *Proteins: Struct. Funct. Genet.* **61**, 258–271.
74. Ahmad, S., Gromiha, M. M. & Sarai, A. (2004). Analysis and prediction of DNA-binding proteins and their binding residues based on composition, sequence and structural information. *Bioinformatics*, **20**, 477–486.
75. Djinovic Carugo, K., Battistoni, A., Carr, M. T., Polticelli, F., Desideri, A., Rotilio, G. *et al.* (1996). Three-dimensional structure of *Xenopus laevis* Cu, Zn superoxide dismutase B determined by X-ray crystallography at 1.5 angstroms resolution. *Acta Crystallog. sect. D*, **52**, 176–188.
76. Stamper, C. C., Zhang, Y., Tobin, J. F., Erbe, D. V., Ikemizu, S., Davis, S. J. *et al.* (2001). Crystal structure of the B7-1/CTLA-4 complex that inhibits human immune responses. *Nature*, **410**, 608–611.
77. Murzin, A. G., Brenner, S. E., Hubbard, T. & Chothia, C. (1995). SCOP: a structural classification of proteins database for the investigation of sequences and structures. *J. Mol. Biol.* **247**, 536–540.
78. Oki, M. & Nishimoto, T. (1998). A protein required for nuclear-protein import, Mog1p, directly interacts with GTP–Gsp1p, the *Saccharomyces cerevisiae* Ran homologue. *Proc. Natl Acad. Sci. USA*, **95**, 15388–15393.
79. Stewart, M. & Baker, R. P. (2000). 1.9 Å resolution crystal structure of the *Saccharomyces cerevisiae* Ran binding protein Mog1p. *J. Mol. Biol.* **299**, 213–223.
80. Lu, J., M.-Y., Deschenes, R. J. & Fassler, J. S. (2004). Role for the Ran binding protein, Mog1p, in *Saccharomyces cerevisiae* SLN1-SKN7 signal transduction. *Eukaryotic Cell*, **3**, 1544–1556.
81. Ifuku1, K., Nakatsu, T., Kato, H. & Sato, F. (2004). Crystal structure of the PsbP protein of photosystem II from *Nicotiana tabacum*. *EMBO Rep.* **5**, 362–367.
82. Wu, C. H., Apweiler, R., Bairoch, A., Natale, D. A., Barker, W. C., Boeckmann, B. *et al.* (2006). The Universal Protein Resource (UniProt): an expanding universe of protein information. *Nucl. Acids Res.* **34**, D187–D191.
83. Glaser, F., Pupko, T., Paz, I., Bell, R. E., Bechor-Shental, D., Martz, E. & Ben-Tal, N. (2003). ConSurf: identification of functional regions in proteins by surface-mapping of phylogenetic information. *Bioinformatics*, **19**, 163–164.
84. Baker, R. P., Harreman, M. T., Eccleston, J. F., Corbett, A. H. & Stewart, M. (2001). Interaction between Ran and Mog1 is required for efficient nuclear protein import. *J. Biol. Chem.* **276**, 41255–41262.
85. Michalopoulos, I., Torrance, G. M., Gilbert, D. R. & Westhead, D. R. (2004). TOPS: An enhanced database of protein structural topology. *Nucl. Acids Res.* **32**, D251–D254.
86. Krissinel, E. & Henrick, K. (2004). Secondary-structure matching (SSM), a new tool for fast protein structure alignment in three dimensions. *Acta Crystallor. sect. D*, **60**, 2256–2268.
87. Pupko, T., Bell, R. E., Mayrose, I., Glaser, F. & Ben-Tal, N. (2002). Rate4Site: an algorithmic tool for the identification of functional regions in proteins by surface mapping of evolutionary determinants within their homologues. *Bioinformatics*, **18**, S71–S77.
88. Stewart, M., Kent, H. M. & McCoy, A. J. (1998). Structural basis for molecular recognition between nuclear transport factor 2 (NTF2) and the GDP-bound form of the Ras-family GTPase Ran. *J. Mol. Biol.* **277**, 635–646.
89. Vetter, I. R., Arndt, A., Kutay, U., Gorlich, D. & Wittinghofer, A. (1999). Structural view of the Ran-Importin beta interaction at 2.3 Å resolution. *Cell*, **97**, 635–646.
90. Zhao, L. & Chmielewski, J. (2005). Inhibiting protein-protein interactions using designed molecules. *Curr. Opin. Struct. Biol.* **15**, 31–34.
91. Pagliaro, L., Felding, J., Audouze, K., Nielsen, S. J., Terry, R. B., Krog-Jensen, C. & Butcher, S. (2004). Emerging classes of protein-protein interaction inhibitors and new tools for their development. *Curr. Opin. Chem. Biol.* **8**, 442–449.
92. Ryan, D. P. & Matthews, J. M. (2005). Protein-protein interactions in human disease. *Curr. Opin. Struct. Biol.* **15**, 441–446.
93. zur Hausen, H. (2002). Papillomaviruses and cancer: from basic studies to clinical application. *Nature Rev. Cancer*, **2**, 342–350.
94. Baseman, J. G. & Koutsky, L. A. (2005). The epidemiology of human papillomavirus infections. *J. Clin. Virol.* **32**, S16–S24.
95. Kim, S. S., Tam, J. K., Wang, A. F. & Hegde, R. S. (2000). The structural basis of DNA target discrimination by papillomavirus E2 proteins. *J. Biol. Chem.* **275**, 31245–31254.
96. Abbate, E. A., Berger, J. M. & Botchan, M. R. (2004). The X-ray structure of the papillomavirus helicase in complex with its molecular matchmaker E2. *Genes Dev.* **18**, 1981–1996.
97. You, J., Croyle, J. L., Nishimura, A., Ozato, K. & Howley, P. M. (2004). Interaction of the bovine papillomavirus E2 protein with Brd4 tethers the viral DNA to host mitotic chromosomes. *Cell*, **117**, 349–360.
98. Brannon, A. R., Maresca, J. A., Boeke, J. D., Basrai, M. A., McBride, A. A., McPhillips, M. G. & Oliveira, J. G. (2005). Reconstitution of papillomavirus E2-mediated plasmid maintenance in Saccharomyces cerevisiae by the Brd4 bromodomain protein Brd4: tethering, segregation and beyond. *Proc. Natl Acad. Sci. USA*, **102**, 2998–3003.
99. Wang, Y., Coulombe, R., Cameron, D. R., Thauvette, L., Massariol, M.-J., Amon, L. M. *et al.* (2004). Crystal structure of the E2 transactivation domain of human papillomavirus type 11 bound to a protein interaction inhibitor. *J. Biol. Chem.* **279**, 6976–6985.
100. White, P. W., Titolo, S., Brault, K., Thauvette, L., Pelletier, A., Welchner, E. *et al.* (2003). Inhibition of human papillomavirus DNA replication by small molecule antagonists of the E1-E2 protein interaction. *J. Biol. Chem.* **278**, 26765–26772.
101. Frishman, D. & Argos, P. (1995). Knowledge-based protein secondary structure assignment. *Proteins: Struct. Funct. Genet.* **23**, 566–579.
102. Liu, S., Zhang, C., Zhou, H. & Zhou, Y. (2004). A physical reference state unifies the structure-derived potential of mean force for protein folding and binding. *Proteins: Struct. Funct. Genet.* **56**, 93–101.
103. Berman, H. M., Westbrook, J., Feng, Z., Gilliland, G., Bhat, T. N., Weissig, H. *et al.* (2000). The Protein Data Bank. *Nucl. Acids Res.* **28**, 235–242.
104. Sanner, M. F. & Olson, A. J. (1996). Reduced surface: an efficient way to compute molecular surfaces. *Biopolymers*, **38**, 305–320.
105. Murphy, K. P. (2001). The Bayes Net Toolbox for Matlab. *Comput. Sci. Stat.* **33**.
106. Leray, P. & Francois, O. (2004). *BNT structure learning package: documentation and experiments*. Technical report, Laboratoire PSI, Université et INSA de Rouen.

107. Fauchère, J. L. & Pliska, V. (1983). Hydrophobic parameters of amino-acid side chains from the partitioning of *N*-acetyl-amino-acid amides. *Eur. J. Med. Chem.* **18**, 369–375.
108. Duncan, B. S. & Olson, A. J. (1993). Shape analysis of molecular surfaces. *Biopolymers*, **33**, 231–238.
109. Koenderink, J. J. (1991). *Solid Shape.* Cambridge, MA, MIT Press.
110. Rocchia, W., Alexov, E. & Honig, B. (2001). Extending the applicability of the non-linear Poisson-Boltzmann equation: multiple dielectric constants and multivalent ions. *J. Phys. Chem. ser. B*, **105**, 6507–6514.
111. Rocchia, W., Sridharan, S., Nicholls, A., Alexov, E., Chiabrera, A. & Honig, B. (2002). Rapid grid-based construction of the molecular surface for both molecules and geometric objects: applications to the finite difference Poisson-Boltzmann method. *J. Comp. Chem.* **23**, 128–137.
112. Valdar, W. S. (2002). Scoring residue conservation. *Proteins: Struct. Funct. Genet.* **48**, 227–241.

Edited by M. J. E. Sternberg

(Received 9 February 2006; received in revised form 15 June 2006; accepted 13 July 2006)
Available online 21 July 2006