WILEY InterScience® DISCOVER SOMETHING GREAT

PROTEINS STRUCTURE ■ FUNCTION ■ BIOINFORMATICS

An automated decision-tree approach to predicting protein interaction hot spots

Steven J. Darnell,[1] David Page,[2] and Julie C. Mitchell[1,3*]

[1] Department of Biochemistry, University of Wisconsin-Madison, Madison, Wisconsin 53706
[2] Department of Biostatistics and Medical Informatics, University of Wisconsin-Madison, Madison, Wisconsin 53706
[3] Department of Mathematics, University of Wisconsin-Madison, Madison, Wisconsin 53706

ABSTRACT

Protein–protein interactions can be altered by mutating one or more "hot spots," the subset of residues that account for most of the interface's binding free energy. The identification of hot spots requires a significant experimental effort, highlighting the practical value of hot spot predictions. We present two knowledge-based models that improve the ability to predict hot spots: K-FADE uses shape specificity features calculated by the Fast Atomic Density Evaluation (FADE) program, and K-CON uses biochemical contact features. The combined K-FADE/CON (KFC) model displays better overall predictive accuracy than computational alanine scanning (Robetta–Ala). In addition, because these methods predict different subsets of known hot spots, a large and significant increase in accuracy is achieved by combining KFC and Robetta–Ala. The KFC analysis is applied to the calmodulin (CaM)/smooth muscle myosin light chain kinase (smMLCK) interface, and to the bone morphogenetic protein-2 (BMP-2)/BMP receptor-type I (BMPR-IA) interface. The results indicate a strong correlation between KFC hot spot predictions and mutations that significantly reduce the binding affinity of the interface.

Proteins 2007; 68:813–823.

Key words: atomic density; complemented pocket; computational alanine scanning; decision tree; FADE; protein–protein interface; shape complementarity; shape specificity; site-directed mutagenesis.

INTRODUCTION

An increased desire to modify and redesign protein interfaces highlights the importance of predicting mutagenesis hot spots. The principles governing protein interactions are not fully understood,[1] but it is known that a small subset of "hot spot" residues account for most of a protein interface's free energy of binding.[2,3] In practice, the site-directed mutation of hot spots is an effective means of disrupting a protein–protein interaction,[4] but the systematic identification of hot spots requires a significant experimental effort. Predictive models, in contrast, can be applied quickly to help make experimental design more efficient.

The stability of protein complexes is mediated by a collection of biophysical properties, including hydrophobicity, van der Waals forces, shape specificity, hydrogen bonds, salt bridges, solvent accessibility, and so on.[1,5–10] Current methods for predicting hot spots include energetic-based and structure-based approaches. Some energetic approaches, such as computational alanine scanning, are virtual mutagenesis techniques that use potential energy functions and thermodynamic cycles to predict the change in the free energy of binding ($\Delta\Delta G$) for a theoretical mutation within a protein interface.[11–13] These techniques account for atomic packing interactions, hydrogen bonds, electrostatics, and solvation effects. In contrast, structural approaches focus on a single or small set of chemical and physical features. If the sequence or structure is known for a set of related protein complexes, hot spots can be predicted by identifying sequentially-conserved or structurally-conserved residues.[14,15] Hot spot predictions can also be made from individual structures by identifying chemically-complemented polar residues surrounded by a ring of hydrophobic residues,[3] or interface regions with a well-matched protrusion and crevice.[1,16] Although structural approaches often have a narrower predictive range, they can identify important binding interactions that are not captured by energetics.

In this study, we analyze the features of experimentally determined hot spot residues from protein complexes having known structures. Knowledge-based models are created from this data to predict hot spot residues in new protein complexes. Our approach is related broadly to knowledge-based automated neural networks[17] and knowledge-based support vector machines[18] in the

The Supplementary Material referred to in this article can be found at http://www.interscience.wiley.com/jpages/0887-3585/suppmat/
Grant sponsor: National Library of Medicine Training Grant to the Computation and Informatics in Biology and Medicine Training Program; Grant number: 5T15LM007359
*Correspondence to: Julie C. Mitchell, 433 Babcock Drive, Madison, Wisconsin 53706.
E-mail: mitchell@math.wisc.edu
Received 10 August 2006; Revised 2 December 2006; Accepted 12 February 2007
Published online 6 June 2007 in Wiley InterScience (www.interscience.wiley.com). DOI: 10.1002/prot.21474

sense that we use machine learning to modify an initial classifier constructed by domain experts. Our approach is also knowledge-based in that the feature types are carefully constructed from background knowledge about the general nature of protein interactions.

Two predictive models are created: Knowledge-based FADE, or *K-FADE*, based on shape specificity features calculated by the Fast Atomic Density Evaluation (FADE) program,[19] and Knowledge-based Biochemical Contact analysis, or *K-CON*, based on biochemical contacts. A combination of K-FADE and K-CON (KFC) modestly exceeds the predictive accuracy of Robetta's computational alanine scanning (Robetta–Ala),[20] and it uses considerably less computing time. Unlike Robetta–Ala, KFC cannot be trained on protein stability data due to differences in methodology; KFC is trained on a limited set of interface mutation data. Nevertheless, KFC predicts different hot spots than alanine scanning, and the combination of KFC and Robetta–Ala produces a statistically significant increase in predictive accuracy over Robetta–Ala.

To create the knowledge-based models, the atomic contacts, hydrogen bonds, salt bridges, and surrounding shape specificity of known hot spots are quantified as features. Next, a machine learning algorithm analyzes these features and produces a model for predicting new hot spots. We test the resulting models using a cross-validation statistical analysis, and the predictive performance of the knowledge-based models is compared against Robetta–Ala. Finally, we use these new methods to analyze two protein complexes, the calmodulin/myosin kinase complex and the BMP-2/BMP receptor complex, and comment on the models description of these binding interactions. K-FADE and K-CON predict several hot spots that are missed by Robetta–Ala, further illustrating that our knowledge-based methods improve the scope and accuracy of hot spot predictions within protein interfaces.

MATERIALS AND METHODS

The process of making our knowledge-based decision tree models requires four steps. First, a data set of protein–protein interfaces with experimentally characterized interface mutations must be assembled. Then, a vector of features describing the local environment of each residue is calculated. Next, a machine learning algorithm selects the most important features and creates a predictive model. Finally, the predictive performance of the model is analyzed. The following section describes the details of this implementation.

Hot spot data sets

Data set for cross-validation

Each protein interface has a crystal structure with a resolution <3.0 Å, and a set of interface residues experimentally mutated to alanine. Structures were obtained from the Protein Data Bank.[21] Each interface residue was required to have one atom within 4 Å of the binding partner and to have a reported $\Delta\Delta G$ for mutation to alanine. Alanine mutation data was obtained from the Alanine Scanning Energetics database,[22] the Binding Interface database (BID),[23] and a data set from Kortemme and Baker.[13] Additionally, the sequence identity of at least one protein in each interface was required to be <35% from all other proteins in the data set. The sequence identity between proteins was calculated using the PISCES sequence culling server.[24] The resulting data set contained 16 nonredundant protein complexes with 249 mutated interface residues, of which 60 residues were hot spots ($\Delta\Delta G$ greater than 2 kcal/mol[1,3]). Table I lists the protein structures in the data set. The on-line Supplementary Material catalogs the individual mutations with their respective $\Delta\Delta G$ (Table S1).

Table I
Data Set of Protein Structures

PDB	Molecule 1 Name	Molecule 2 Name
1A4Y	Angiogenin	Ribonuclease Inhibitor
1AHW	Immunoglobulin Fab 5G9	Tissue Factor
1BRS	Barnase	Barstar
1BSR	Bovine Seminal Ribonuclease	Bovine Seminal Ribonuclease
1BXI	Colicin E9 Immunity Protein	Colicin E9 DNase Domain
1CBW	Chymotrypsin	BPTI Trypsin Inhibitor
1DAN	Blood Coagulation Factor VIIa	Tissue Factor
1DN2	Immunoglobulin Lambda	Peptide
1DVF	Idiotopic Antibody D1.3	Anti-idiotopic Antibody D1.3
1DX5	Thrombin	Thrombomodulin
1GC1	Envelope Protein	CD4
1JTG	β-lactamase	β-lactamase Inhibitory Protein
1NMB	FAB NC10	N9 Neuraminidase
1VFB	Mouse Monoclonal Antibody D1.3	Lysozyme
3HFM	Ig FAB Fragment HY HEL-10	Lysozyme
3HHR	Human Growth Hormone	Human Growth Hormone Receptor

Independent test set

A second data set—independent of the previous set—was collected from the BID to further validate our decision tree models. Each protein structure has a set of interface residues experimentally mutated to alanine; however, the mutations are characterized by experiments

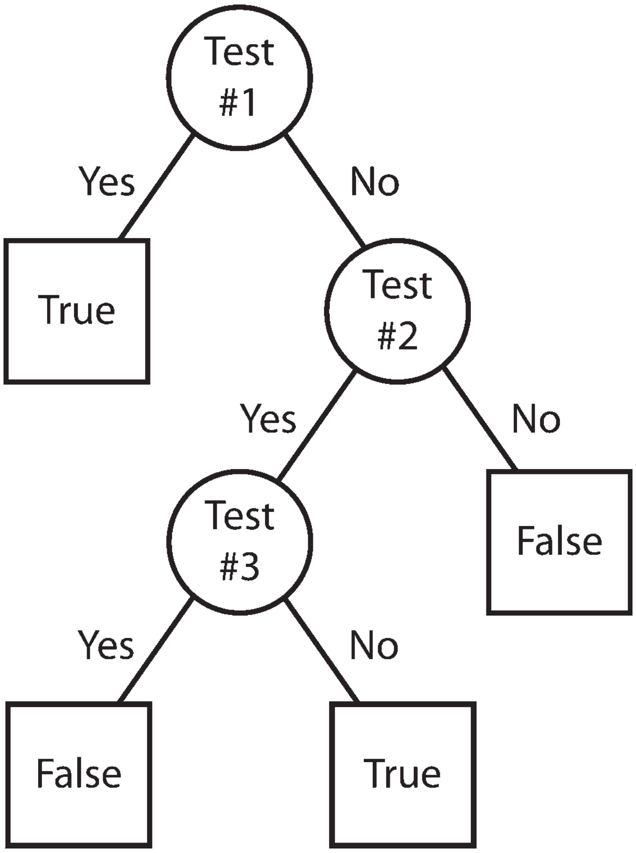

Figure 1
A decision tree is a predictive model used to predict the classification of a record. The tree defines a set of Boolean tests (circles), where the result of each test determines the next test to apply. The process continues until the path terminates, where the model predicts the record's class (squares).

that do not report $\Delta\Delta G$ values. Instead, the relative disruptive effect of the mutation is listed in the BID as either strong, intermediate, weak, or insignificant. Given the subjective nature of the classification, we only considered "strong" mutations to identify binding hot spots. The secondary test set contained 19 protein complexes with 112 mutated interface residues, of which 50 residues were hot spots. The Supplementary Material catalogs the protein complexes (Table S2), and the individual mutations with their respective disruptive strength (Table S3).

Learned decision tree models

The machine learning algorithm C5.0[25] was used to search for patterns within our training data, and to generate a learned decision tree that predicts the classification of residues within new protein complexes. A decision tree is a series of Boolean tests, where each test determines the next test to apply (Figure 1). Every path through the tree terminates with a prediction as to whether a given residue is a hot spot. C5.0 also specifies a confidence level for each path. It classifies the training data using the learned decision tree, and it defines a path's confidence level as the percentage of correctly classified samples for that path.

Two decision tree models were created: the K-FADE model and the K-CON model. K-FADE was trained using shape-related features (shape specificity, FADE points, residue size), and K-CON was trained using all available contact features (shape-related features, atomic contacts, hydrogen bonds, salt bridges, chemical type). For each model, an exhaustive search identified the combination of features that best described a hot spot environment. In some respects, the feature construction mirrors other methods,[26,27] but our approach is algorithmically distinct and answers different questions.

Comparison of machine learning algorithms

Implementations of other "standard" machine learning algorithms are available, thus it is prudent to consider these alternatives in addition to learned decision trees. We performed an empirical comparison between learned decision trees and two rival machine learning algorithms: support vector machines (SVM) as implemented in SVMlight[28] and Bayesian Networks (Bayes Net) as implemented in Weka.[29] The rivals create predictors in very different manners. SVMlight calculates a hyperplane using the described features to separate hot spots from other residues, whereas Weka creates a directed graph using the same features to describe the conditional probability of whether a residue is a hot spot. A comparison of the three algorithms will illustrate that learned decision trees perform best for this classification task.

Physical and chemical features

Initially, we trained models using large sets of features (e.g. combinations of a putative hot spot's amino acid identity, its relative size, chemical nature, noncovalent interactions, intermolecular contacts, surrounding shape specificity, and 80 other radially distributed properties described by Bagley and Altman[30]); however, these models clearly overfitted the training data. Subsequently, we searched for smaller sets of features that generated robust and general predictive models.

The features we used to train our final models combine knowledge of the biochemical properties of amino acids with information about the local degree of shape match within the protein interface. The following sections describe the collection of features used by K-FADE and K-CON, along with the details of how they were computed.

Size and chemical nature

The residues in the data set were characterized by their physical and chemical properties. The relative size and chemical nature of each residue were classified as:

small	Ala, Gly, Ser
medium	Asn, Asp, Cys, Pro, Thr, Val
large	Arg, Gln, Glu, His, Ile, Leu, Lys, Met, Phe, Trp, Tyr
nonpolar	Ala, Gly, Ile, Leu, Met, Phe, Pro, Trp, Val
polar	Asn, Cys, Gln, Ser, Thr, Tyr
charged	Arg, Asp, Glu, His, Lys

Shape specificity

Our K-FADE model uses shape specificity features to predict interface hot spots. Local shape specificity can be derived from atomic density within the interface.[19] In a crevice surrounded by nearby atoms, the local atomic density will be high. The density is characterized by an atomic density exponent (λ), which is higher in a crevice than a protrusion. Atomic density exponents relative to both proteins are calculated over a regular grid containing the binding interface. Each grid point that is within 3 Å of both molecules is identified as a *FADE point*. The *shape specificity potential* is evaluated for all FADE points using the expression

$$S_p = (\lambda_1 - \lambda_0) \cdot (\lambda_2 - \lambda_0) \qquad (1)$$

where λ_1 and λ_2 are atomic density exponents relative to molecules 1 and 2, and λ_0 is a median exponent describing flat edge geometry.[19] S_p is negative where a crevice and protrusion are well matched, and it's positive near a geometric mismatch.

To calculate a residue's shape microenvironment, a spherical volume with a radius of 10 Å centered at the residue's center of mass is partitioned into 10 nested spherical shells each with a thickness of 1 Å. The local shape specificity score describes the degree of geometric complementarity around the interface residue. The distribution of FADE points suggests a residue's local packing density and its relative position within the interface; numerous FADE points in distant shells of the microenvironment suggests a residue lies in the interior of an interface rather than at its edge. The number of FADE points in each shell and the sum of their S_p values were recorded as features for the K-FADE model.

Biochemical contacts

To create features for the K-CON model, the molecular modeling package WHAT IF[31] was used to identify noncovalent interactions within protein complexes and to score hydrogen bonds. There were three types of noncovalent interactions recorded: atomic contacts, hydrogen bonds, and salt bridges.

An *atomic contact* was identified when the distance between the van der Waals surfaces of two atoms was < 0.25 Å. If an atom had multiple contacts, its contact was assigned to the atom pair with the shortest distance. Atomic contacts were then divided into three categories: polar, nonpolar, and generic. Polar contacts occur between a nitrogen atom and an oxygen or sulfur atom. Nonpolar contacts occur between two carbon atoms that are not directly bonded to nitrogen, oxygen, or sulfur. The residue's polar contact feature was defined as the number of polar contacts its side chain made with the binding partner. The nonpolar and generic contact features were defined similarly.

A *hydrogen bond* was identified from an optimized hydrogen bond network model.[32] Each hydrogen bond was scored on a scale from 0 to 1, where 0 indicates no hydrogen bond, and 1 indicates an ideal hydrogen bond. The value of a residue's hydrogen bond feature was defined by the sum of the hydrogen bond scores associated with its atoms; therefore, residues that participate in more favorable or multiple hydrogen bonds are weighted more heavily.

A *salt bridge* was identified when the distance between the centers of a basic nitrogen and an acidic oxygen was <4.5 Å. In this study, histidine was considered basic, and only one salt bridge was acknowledged between any unique pair of acidic and basic residues. The number of salt bridges a residue made with its binding partner was the residue's salt bridge feature.

Precision, recall and *F*1 score

Our training set contains an uneven distribution of hot spots; as such, it is unwise to search for models that maximize *statistical accuracy*, or the fraction of correctly classified residues. For this data set, a model that predicts zero hot spots achieves an arbitrarily high accuracy of 76%. We evaluate model performance in terms of the widely-used *F*1 score to avoid this situation.

The *F1 score* (*F*1)—a robust metric of a model's overall accuracy—is defined as a function of precision and recall (Eq. 2). *Precision* (P) measures the accuracy of predicting the hot spot class, that is, the fraction of positive hot spot predictions that are correct. *Recall* (R) indicates how many total hot spots are correctly predicted, in other words, the fraction of known hot spots that are predicted. These measures are defined in terms of true positives (TP: residue is a predicted hot spot and an actual

Table II

The Confusion Matrices List the Number of True Positive (TP), False Positive (FP), True Negative (TN), and False Negative (FN) Hot Spot Predictions for Each Model

		Predicted	
		Hot Spot	Other
K-FADE			
Actual	Hot Spot	22 (TP)	38 (FN)
	Other	27 (FP)	162 (TN)
K-CON			
Actual	Hot Spot	27 (TP)	33 (FN)
	Other	26 (FP)	163 (TN)
Robetta–Ala			
Actual	Hot Spot	28 (TP)	32 (FN)
	Other	27 (FP)	162 (TN)

Each column represents the class predicted by the model, and each row represents the class as determined by experiment.

hot spot), false positives (FP: residue is a predicted hot spot and is not an actual hot spot), true negatives (TN: residue is not a predicted hot spot and not an actual hot spot), and false negatives (FN: residue is not a predicted hot spot and is an actual hot spot).

$$F1 = \frac{2PR}{P + R} \quad (2)$$

F1: A measure to balance precision and recall rates

$$P = \frac{TP}{TP + FP} \quad (3)$$

Precision: Fraction of predicted hot spots that are true hot spots

$$R = \frac{TP}{TP + FN} \quad (4)$$

Recall: Fraction of true hot spots that are predicted hot spots

Each measure's worst value is 0, and its best value is 1. For practical purposes, any useful method must have an *F*1 score greater than 0.24, since 24% of the residues in our data set were hot spots. The *F*1 score is used in the next section to compare the performance of K-FADE, K-CON, and Robetta–Ala.

Cross-validation

To guarantee that our models were not overtrained on a single data set, the *F*1 scores for K-FADE and K-CON were estimated by cross-validation—a resampling technique that averages the analysis over a series of training and testing events on different subsets of the same data set.[33,34] This approach guards against calculating overly optimistic *F*1 scores for models suggested from the data.

A 16-fold "leave one protein complex out" cross-validation was performed to estimate the accuracy of K-FADE and K-CON when presented with a new interface. For each of the 16 protein complexes, we withheld the mutation data for that complex, trained a model using the remaining 15 complexes, and tested the model using the omitted interface to gauge predictive accuracy. This method was more logical than a typical 10-fold cross-validation (data is subdivided into 10 equal partitions),[35] given the inherent dependencies between residues in the same binding interface.

The difference in *F*1 score between methods, relative to Robetta–Ala, was assessed using a two-tailed, paired *t*-test. Simplistically, the test indicates whether the performance between two models is different. The statistics package R[36] was used to calculate a *P-value* for the *F*1 score differences. The lower the *P*-value, the lower the probability that the observed difference occurred by chance. A *P*-value < 0.05 is commonly regarded as a statistically significant threshold, and was used in this study.

RESULTS

The results in this section are presented in the following order. First, we compare the predictive performance of decision trees, SVMs, and Bayes Nets—and find that learned decision trees are the most accurate predictors of binding hot spots. Then, after calculating the performances of our tree-based models and Robetta–Ala (true and false predictions listed in Table II), we illustrate that KFC exceeds the predictive accuracy of Robetta–Ala. Finally, we combine the KFC and Robetta–Ala models, demonstrate that their union significantly improves predicted hot spot accuracy when compared with Robetta–Ala alone (Table III), and validate this result using an independent test set (Table IV). The Supplementary Material catalogs the individual interface residues for each data set with their respective hot spot classification predictions from each model (Tables S1, S3).

Table III

Hot Spot Prediction Statistics for Individual and Combined Models

Model	Precision	Recall	*F*1 score	Δ*F*1	*P*-value
Robetta–Ala	0.51	0.47	0.49	**	**
K-FADE	0.47	0.37	0.41	−0.08	0.88
K-CON	0.52	0.45	0.48	−0.01	0.54
KFC	0.49	0.58	0.53	+0.04	0.22
KFCA	0.44	0.72	0.55	+0.06	0.02

Δ*F*1 is the change in *F*1 score for a model when compared with Robetta–Ala. The reported *P*-value is the probability that the observed Δ*F*1 occurred by chance (*i.e.*, the probability that KFCA outperformed Robetta–Ala by chance is 2%). Either of K-FADE and K-CON is less predictive than Robetta–Ala. The combination of K-FADE and K-CON (KFC) performs better, and the combination of all three models (KFCA) offers the best overall performance.

Table IV
Hot Spot Prediction Statistics for the Independent Test Set

Model	Precision	Recall	$F1$ score	$\Delta F1$	P-value
Robetta–Ala	0.64	0.28	0.39	**	**
KFC	0.51	0.36	0.42	+0.04	0.81
KFCA	0.53	0.48	0.51	+0.12	0.0071

Again, KFC and KFCA perform better than Robetta–Ala, further validating the methodology. Additionally, the significance of KFCA's improved performance is stronger than previously observed.

It is important to emphasize that the $F1$ score for a random model represents the practical baseline for the $F1$ score. For the random model, precision will remain constant, and recall will change as the expected frequency of hot spots changes. If the frequency of hot spots is 24% (observed in our data set), then the expected $F1$ score for a random model is $F1 = 0.24$. It is not important to have a 50-50 balance between true positives and negatives within the data set. Rather, the essential goal is to perform considerably better than a random model. The following results illustrate that K-FADE, K-CON, and Robetta–Ala each individually perform much better than a random model, and that combined models have superior hot spot prediction capabilities overall.

Decision trees vs. SVMs and Bayes nets

To identify the best machine learning algorithm for predicting binding hot spots, we determined the performance of learned decision trees, SVMs, and Bayes Nets using our training data set. For each algorithm, the $F1$ score was estimated by cross-validation for models trained using the optimal feature configurations identified by C5.0, namely shape specificity features (shape specificity, FADE points, and residue size) and biochemical contact features (polar atomic contacts, generic atomic contacts, hydrogen bonds, FADE points, and chemical type).

Compared with SVMs, decision trees were superior for predicting hot spots. A boosted decision tree trained on biochemical contact features classified hot spots with $F1 = 0.48$, whereas a linear kernel SVM trained on the same features classified hot spots with $F1 = 0.33$. Optimizing the SVM cost ratio, a parameter describing the trade-off between training error and separation margin, increased the SVM $F1$ score by +4 percentage points. Regardless, a paired t-test illustrated the difference in $F1$ score between the decision tree and the optimized SVM was significant at the 0.05 level ($P = 0.012$). For the shape specificity feature set, the decision tree classified hot spots with $F1 = 0.48$ while the SVM classified hot spots with $F1 = 0.06$. This difference in performance was substantial and very significant ($P = 0.0016$). Additionally, changing the cost ratio actually decreased SVM performance. Thus, applying SVM to K-FADE features produced a model that was considerably worse than random, while for K-CON features the model was better than random but still did not match the accuracy achieved by decision trees.

Unlike SVM, Bayes Net performed comparably to decision trees. A repeated hill climber Bayes Net (with Bayesian scoring metric) trained on biochemical contact features classified hot spots with $F1 = 0.46$. Decision trees outperformed the Bayes Net for these features, but a paired t-test indicated the difference was not significant ($P = 0.34$). For the shape specificity feature set, the Bayes Net classified hot spots with $F1 = 0.47$. Although Bayes Net outperformed decision trees for shape features, the difference was not significant ($P = 0.80$). Finally, the $F1$ score obtained by combining the two Bayes Net models was 0.47, which was not enough to surpass the decision tree approach ($F1 = 0.53$). The P-value for this difference was 0.23. An optimal combination of decision trees and Bayes Nets may exist, and we intend to investigate this possibility in the future. However, for this study, learned decision trees were found to be the best machine learning algorithm for predicting binding hot spots.

Robetta–Ala

The Robetta server's Computational Interface Alanine Scanning service[20] was used as a standard to analyze the predictive performance of the knowledge-based models. Since the method predicts the $\Delta\Delta G$ for an interface mutation to alanine, a residue was deemed a predicted hot spot when its $\Delta\Delta G$ for mutation to alanine was greater than 2 kcal/mol.[*]

Robetta–Ala correctly predicted hot spots with $P = 0.51$, $R = 0.47$, and $F1 = 0.49$. This means that the method correctly predicted 47% of alanine scanning hot spots for this data set, and 51% of its predicted hot spots were correctly classified. The $F1$ score indicates the tradeoff between precision and recall. We will see that while neither the K-FADE or K-CON predictive models individually exceeds the $F1$ score for Robetta–Ala, their combined predictions do. In addition, a model combining all three predictive methods achieves a statistically large improvement in $F1$ score over any individual model.

Knowledge-based decision trees

The following three models are based on the use of physical/biochemical feature vectors. The two individual models are based on the best geometric feature vector and the best biochemical feature vector we found for predicting hot spots. The third model, a combination

[*]This implementation does not predict the $\Delta\Delta G$ for mutations of glycine or proline; therefore, these residues were not included in the data set.

DOI 10.1002/prot

of the first two, predicts a significant fraction of interface hot spots.

The K-FADE shape specificity model

K-FADE was trained using the following features: specificity, FADE points, and residue size. K-FADE correctly predicted 22 (true positive) hot spots and 162 (true negative) residues that are not hot spots ($P = 0.47$, $R = 0.37$, and $F1 = 0.41$). In comparison, Robetta–Ala correctly predicted hot spots with $P = 0.51$, $R = 0.47$, and $F1 = 0.49$. The recall of K-FADE was 10 percentage points less than that of Robetta–Ala, because hot spots need not be located in regions with well-matched shape specificity. However, K-FADE correctly predicted 8 hot spots that Robetta–Ala did not predict, representing 13% of the hot spots in the data set. This suggests that energetic approaches to represent the dispersion forces between binding partners underestimate the importance of shape specificity (i.e. well-matched crevices and protrusions), but shape specificity alone does not predict hot spots comprehensively.

The K-CON biochemical contact model

K-CON was trained using the following features: polar atomic contacts, generic atomic contacts, hydrogen bonds, FADE points, and chemical type. These properties are consistent with the observation that alanine-scanning experiments inherently weight polar residue mutations greater than nonpolar residue mutations when identifying hot spots.[9] K-CON correctly predicted 27 (true positive) hot spots and 163 (true negative) residues that are not hot spots ($P = 0.52$, $R = 0.45$, $F1 = 0.48$). The scores for precision, recall, and F1 score of K-CON are comparable to those of Robetta–Ala, even though K-CON was trained on a small set of features. K-CON correctly predicted 11 hot spots that Robetta–Ala did not predict, representing 18% of the hot spots in the data set. Seven of these hot spots were also overlooked by the K-FADE model.

The K-FADE and K-CON (KFC) combined model

In this study, models were combined by taking the union of the individual results, where a positive prediction from any model resulted in a hot spot prediction. When the K-FADE and K-CON models are combined, they correctly predict hot spots from the data set with $P = 0.49$, $R = 0.58$, and $F1 = 0.53$. That is, 58% of the hot spots were recalled, and 49% of the predicted hot spots were experimentally identified. The $\Delta F1$ for K-FADE/CON (or "KFC") was +4 percentage points versus Robetta–Ala, which suggests the combined model modestly exceeds the predictive accuracy of alanine scanning. Reviewing Table III, we see that KFC predicts 58% of hot spots, as compared with 47% for Robetta–Ala, and KFC achieves a higher $F1$ score than Robetta–Ala. In addition, KFC requires 1–2 orders of magnitude less computation time than is used by the Robetta–Ala server (Figure 2).*

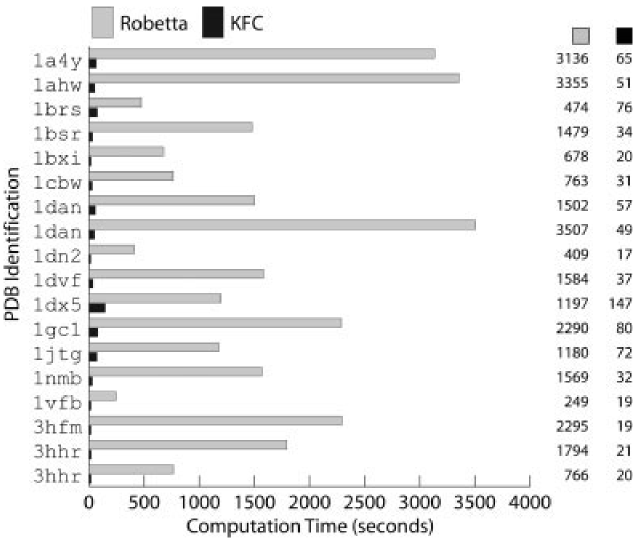

Figure 2

The computation time required by the combined K-FADE and K-CON model (KFC) is less than that required by Robetta–Ala. The columns on the right list the elapsed time (in seconds) for the analysis of each protein interface by Robetta–Ala (gray) and KFC (black). Two protein complexes had more than one analyzed interface: 1DAN (chains L/TU and LTU/H) and 3HHR (chains A/B and A/C).

Combining knowledge-based and Robetta–Ala models

Another increase in predictive accuracy is achieved by combining all three models ($P = 0.44$, $R = 0.72$, and $F1 = 0.55$). An impressive 72% of alanine scanning hot spots are predicted by KFC+Robetta–Ala (KFCA), and its F1 score is the highest observed. The ΔF1 for the three-way combination relative to Robetta–Ala alone was +6 percentage points. This difference is quite significant, as it has a P-value of 0.02. The KFCA model loses some precision relative to the individual models, due to a slightly higher false positive rate. However, the increase in recall more than offsets this effect, leading to an improved $F1$ score (Table III). Each of the models correctly predicts a different subset of hot spots, implying that the models are not comprehensive individually, but

*A Linux desktop computer was used to run K-FADE and K-CON, and timings for Robetta–Ala were taken from the runtimes reported by the server queue.

Table V
Hot Spot Predictions for Interface Residues of smMLCK

Residue	Mutation	Activity (%)	Strength	K-FADE	Conf.	K-CON	Conf.	Robetta–Ala	$\Delta\Delta G$
K799	A	98	N	—	0.73	—	0.53	—	0.61
W800	A	5	S	X	0.70	X	0.61	X	5.18
K802	A	40	I	—	0.70	—	0.81	—	1.53
R808	A	40	I	—	0.54	—	0.82	X	2.70
I810	A	15	S	—	0.52	X	0.52	—	1.90
R812	I	5	S	X	0.66	—	0.90	—	1.24
L813	A	15	S	—	0.71	—	0.82	X	2.01

The kinase activity of each mutant (relative to MK-40K) is described in Ref. 33, and the strength of each mutation is described in the BID as strong (S), intermediate (I), or insignificant (N). Hot spot predictions for each model are marked with an X. The score for K-FADE and K-CON indicates the confidence of the prediction, where its worst value is 0 and its best value is 1. However, the score does not reflect the true accuracy of the prediction since the boosted models artificially weight the training data.[25] The score for Robetta–Ala is the predicted $\Delta\Delta G$ (kcal/mol) when the interface residue is mutated to alanine.

together they predict a considerable subset of alanine scanning hot spots with a reasonable false positive rate.

Independent KFC model validation

Cross-validation indicates that our methodology will predict hot spots with greater accuracy than Robetta–Ala, but the analysis produces several decision trees to choose from. Consequently, final versions of the K-FADE and K-CON models are trained using the entire training set. These final models are validated using the independent test set from the BID. Again, the KFC model outperforms Robetta–Ala (Table IV).

Robetta correctly predicts hot spots from the data set with $P = 0.64$, $R = 0.28$, and $F1 = 0.39$, while KFC predicts hot spots with $P = 0.51$, $R = 0.36$, and $F1 = 0.42$. The $\Delta F1$ for KFC relative to Robetta–Ala is +4 percentage points. Once more, an increase in accuracy is achieved by combining all three models ($P = 0.53$, $R = 0.48$, $F1 = 0.51$). This +12 percentage point increase in $F1$ score is associated with a P-value of 0.0071—the probability of this increase over Robetta–Ala occurring by chance is 0.7%. Given the persistent improvement in predictive performance, these results suggest that KFC is a robust and general model for predicting binding hot spots.

DISCUSSION

In this study, two new knowledge-based hot spot models, K-FADE and K-CON, are introduced. These models analyze the intermolecular geometric contacts within a known protein complex structure and predict binding hot spots within the protein–protein interface—they do not analyze remote regions of the complex, nor do they model site-directed mutations. Together, K-FADE and K-CON (KFC) exceed the predictive accuracy of Robetta–Ala. In addition, our KFC method is considerably faster than Robetta–Ala. By joining forces, KFC and Robetta–Ala (KFCA) together predict nearly three-quarters of experimentally observed hot spot residues.

We now demonstrate the practical value of our approach using two examples selected to illustrate key differences between KFC and Robetta–Ala. The KFCA analysis of two protein complexes from the BID demonstrates the power of using K-FADE and K-CON to predict hot spots. In these cases, the interface residue mutations are experimentally characterized by methods other than measuring the $\Delta\Delta G$ of the perturbation. That is, experimental evidence suggests the presence of a hot spot, but it does not describe the residue's mutation in terms of a change in binding energy. K-FADE and K-CON predictions correlate well with hot spot mutations in this broader context.

Case studies

The calmodulin/myosin kinase complex

Calmodulin (CaM) is a Ca^{+2} binding protein and regulates cellular function by interacting with other target proteins. For example, CaM regulates the activity of smooth muscle myosin light chain kinase (smMLCK), which regulates the phosphorylation of the myosin P-light chain, and subsequently muscle contraction. The CaM/smMLCK CaM-binding domain structure (PDB: 1CDL) has been used to study the interaction between CaM and a 40 kDa fragment of smMLCK with kinase activity (MK-40K),[37] thus we use this structure to characterize the binding interface. The mutations listed in Table V span the length of the CaM-binding domain, and are noted in the BID to disrupt the complex to varying degrees.

KFCA correctly predicts 4 "strong" hot spots: Ile810 (K-CON), Arg812 (K-FADE), Leu813 (Robetta–Ala), and Trp800 (all three). Ile810, Arg812, and Leu813 form a cluster of hot spots near the C-terminus of the smMLCK CaM-binding domain. Experiments indicate that I810A, R812I, and L813A greatly reduce or prevent the binding of CaM and the MK-40K mutant, and similarly reduce smMLCK's kinase activity in the presence of CaM and myosin light chain.[38] R812I is the most disruptive

DOI 10.1002/prot

MK-40K mutation out of the three. The fact that it was predicted only by K-FADE suggests that its role is in determining shape specificity of binding. The extra width introduced by a mutation to isoleucine may hinder its ability to fit into the narrow crevice occupied by the native arginine. Ile810 and Leu813 are also in regions with high shape specificity; however, they are not predicted by K-FADE, but rather by their biochemical environment. K-CON predicts Ile810 is a hot spot because it forms a total of three contacts between CaM Met36 and Leu39. It is not clear what energetic features predict Leu813, since Robetta–Ala only reports a single score.

Trp800, located on the N-terminal side of the smMLCK CaM-binding domain, is far removed from the previous hot spot cluster. Its side chain protrudes into an extremely well-matched, and very hydrophobic crevice. The mutation W800A prevents MK-40K from binding to CaM, and results in a total loss of smMLCK kinase activity. K-FADE, K-CON, and Robetta–Ala each predict that Trp800 is a binding hot spot. The level of confidence in this prediction is greatly enhanced by a consensus of the three models.

The BMP-2/BMP receptor complex

Bone morphogenetic protein 2 (BMP-2) is an extracellular signal molecule that induces bone formation and regeneration in adult vertebrates. When BMP-2 binding to its cell surface receptors (BMPR), the kinase activities of the receptors are activated and the signal is transferred to a family of transcription factors. The BMP-2/BMPR-IA ectodomain complex (PDB: 1ES7) represents a key, high-affinity protein interaction in the signaling event.

The mutations F49A, P50A, H54E, S69R are located on the "wrist epitope" binding surface of BMP-2,[39] and are noted in the BID to disrupt the complex in a strong or intermediate manner. KFC correctly predicts 3 of these 4 residues to be hot spots (Phe49, Pro50, and Ser69), while Robetta–Ala only predicts one (Phe49). His54 is not predicted to be a hot spot by any model, suggesting that the opposite charge introduced by the H54E mutation is the major contributor of the destabilization.

S69R is the most disruptive mutation observed in this interface. K-CON predicts Ser69 is a hot spot because it participates in hydrogen bonds with the Arg297 residue of BMPR-IA. The strong effect of the S69R mutation likely originates from the disruption of the hydrogen bonds, and the steric conflicts introduced into the binding interface. The adjacent pair of residues Phe49 and Pro50 also play an important role in this binding interaction. The mutations F49A and P50A each cause a moderate (10-fold) decrease in association rate when compared with wild type.[39] All three models predict Phe49 as a hot spot, but only K-FADE and K-CON predict Pro50 (Robetta–Ala will not mutate prolines). Phe49 and Pro50 form a protrusion complemented by a cavity on the BMPR-IA surface. These residues establish substantial atomic contacts with four BMPR-IA residues (Figure 3). Interestingly, the F49A/P50A double mutation decreases the association rate by 50-fold and eliminates biological activity,[40] suggesting the protrusion is an essential binding element that is only eliminated by mutating *both* hot spots.

Advantages of knowledge-based decision tree models

K-FADE and K-CON are learned decision trees that together exceed the predictive accuracy of Robetta–Ala. In the cross-validation analysis, the KFC model identifies 58% of alanine scanning hot spots while maintaining a low rate of false positive predictions. This model not only exceeds the predictive accuracy of Robetta–Ala, but its computations are *orders of magnitude faster.*

Although KFC exceeds the predictive accuracy of Robetta–Ala, the advantages of our knowledge-based decision tree models are maximized when we use them in conjunction with Robetta–Ala. The best *P*-value ($P = 0.02$) in the cross-validation analysis was achieved by the combination of KFC+Robetta–Ala (KFCA). The second lowest *P*-value was achieved by our KFC model. Both models show improved predictive accuracy over Robetta–Ala, with that of KFCA being statistically quite significant (Table III). The KFCA combined model also predicts 72% of observed alanine mutation hot spots, an impressive fraction of the total.

Similar trends are observed when KFC and Robetta–Ala are applied to an independent test set. Again, both KFC and KFCA show improved predictive accuracy over Robetta–Ala (Table IV). This time, stronger evidence suggests KFCA performs significantly better than Robetta–Ala; the probability of KFCA's +12 percentage point improvement over Robetta–Ala occurring by chance is 0.7%. KFC alone consistently outperforms Robetta–Ala over two independent data sets, illustrating that validity of using machine learning to predict binding hot spots within protein–protein interactions.

Limitations of alanine scanning

It would be beneficial to re-examine predicted hot spots that were experimentally characterized as weak or insignificant, since alanine mutagenesis data defines a limited range of mutations. Matson and Nilles report that alanine mutations introduced into the interface of the low calcium response (Lcr) protein complex LcrG/LcrV do not disrupt the interaction; however, non-alanine mutations that alter the interface's chemical complementarity successfully disrupt the complex.[41]

Although experimental alanine-scanning is a useful technique for understanding the principles of protein interactions, a single type of mutation is not always sufficient to determine a residue's importance. The creation of comprehensive mutagenesis datasets for even a small number of protein interfaces would be immensely valua-

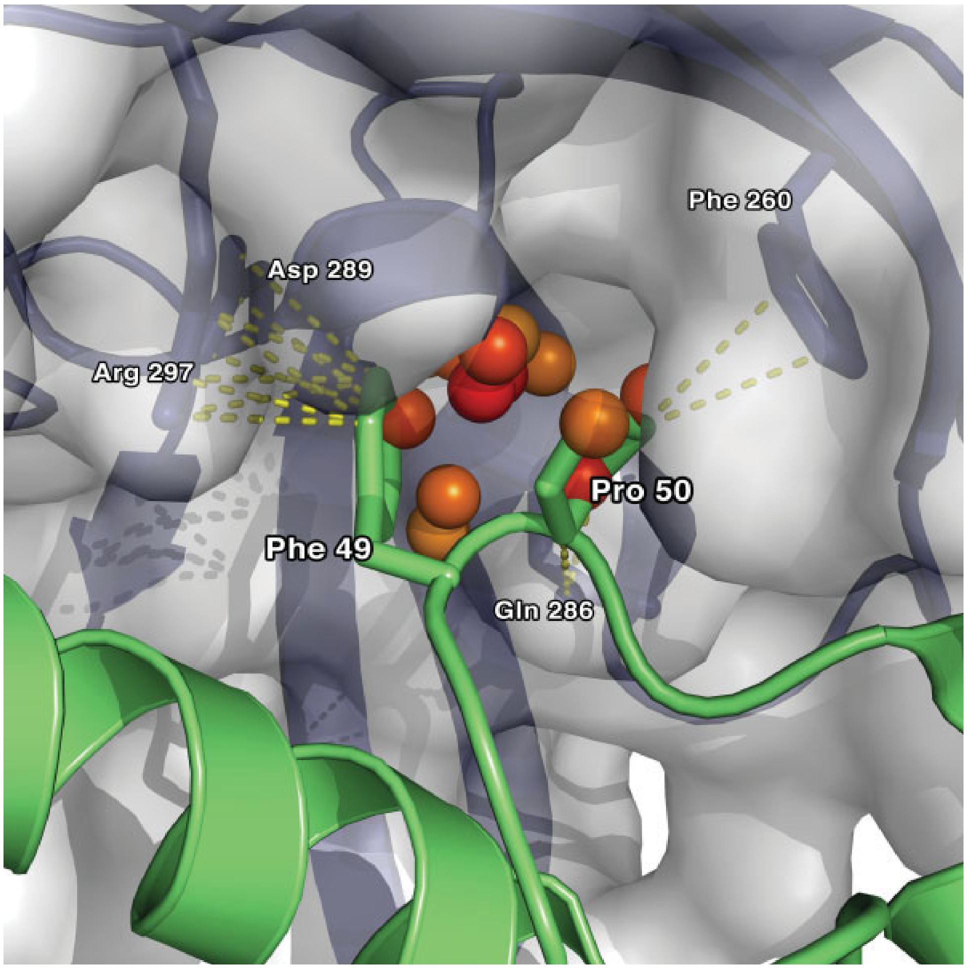

Figure 3

Phe49 and Pro50 from BMP-2 form an essential binding element. They protrude into a crevice on the surface of BMPR-IA, making contact with Phe260, Gln286, Asp289, and Arg297. The spheres within the pocket highlight regions with well-matched shape specificity, where deeper shades of red indicate a more negative S_p value.

ble toward the creation of more detailed knowledge-based decision tree models for hot spot predictions and interface redesign.

CONCLUSIONS

This study demonstrates our success in using machine-learning to select features that best predict the identity of binding hot spots, thus limiting the complexity of the problem to a small set of features. The properties governing hot spot residues are diverse, and are not completely described by any one model. We have illustrated the importance of using both structural and energetic models to describe different aspects of protein interfaces. The KFC model, which combines our shape specificity and biochemical contact models, is shown to exceed the predictive accuracy of Robetta–Ala. KFC's automation and rapid computational speed allow it to quickly generate a list of putative binding hot spots, making it a valuable tool capable of facilitating the interactive exploration of protein interfaces.

In addition, KFCA, which combines KFC and Robetta–Ala, gives rise to a statistically significant increase in the accuracy of predicting hot spots. KFCA recalls an impressive 72% of alanine scanning hot spots while maintaining a reasonable rate of false positive predictions. Finally, our analysis of the calmodulin/myosin kinase complex and the BMP2/BMP receptor complex indicates a strong correlation between KFC hot spot predictions and experimental mutations that significantly reduce binding affinity. In the case of BMP2, KFC highlights a large protrusion where the alanine mutation of both its residues is more disruptive than the sum of its individual mutations.

The FADE program, with source code, is available for download at http://www.mitchell-lab.org/. In the near future, we will offer an interactive web-based interface to run KFC and visualize its predictions. Individuals requiring local versions of the KFC codes should contact Julie Mitchell.

ACKNOWLEDGMENTS

We thank Tanja Kortemme from the University of California, San Francisco for her assistance using Robetta's computational alanine scanning service, and Christina Carlson and Kaitlin Martin from the University of Wisconsin-Madison for their preliminary efforts while visiting the Mitchell group.

REFERENCES

1. Li X, Keskin O, Ma B, Nussinov R, Liang J. Protein–protein interactions: hot spots and structurally conserved residues often locate in complemented pockets that pre-organized in the unbound states: implications for docking. J Mol Biol 2004;344:781–795.
2. Clackson T, Wells JA. A hot spot of binding energy in a hormone-receptor interface. Science 1995;267:383–386.
3. Bogan AA, Thorn KS. Anatomy of hot spots in protein interfaces. J Mol Biol 1998;280:1–9.
4. DeLano WL. Unraveling hot spots in binding interfaces: progress and challenges. Curr Opin Struct Biol 2002;12:14–20.
5. Burley SK, Petsko GA. Aromatic-aromatic interaction: a mechanism of protein structure stabilization. Science 1985;229:23–28.
6. Jones S, Thornton JM. Principles of protein–protein interactions. Proc Natl Acad Sci USA 1996;93:13–20.
7. Xu D, Tsai CJ, Nussinov R. Hydrogen bonds and salt bridges across protein–protein interfaces. Protein Eng 1997;10:999–1012.
8. Sheinerman FB, Norel R, Honig B. Electrostatic aspects of protein–protein interactions. Curr Opin Struct Biol 2000;10:153–159.
9. Chakrabarti P, Janin J. Dissecting protein–protein recognition sites. Proteins 2002;47:334–343.
10. Crowley PB, Golovin A. Cation-pi interactions in protein–protein interfaces. Proteins 2005;59:231–239.
11. Massova I, Kollman PA. Computational alanine scanning to probe protein–protein interactions: a novel approach to evaluate binding free energies. J Am Chem Soc 1999;121:8133–8139.
12. Guerois R, Nielsen JE, Serrano L. Predicting changes in the stability of proteins and protein complexes: a study of more than 1000 mutations. J Mol Biol 2002;320:369–387.
13. Kortemme T, Baker D. A simple physical model for binding energy hot spots in protein–protein complexes. Proc Natl Acad Sci USA 2002;99:14116–14121.
14. Hu Z, Ma B, Wolfson H, Nussinov R. Conservation of polar residues as hot spots at protein interfaces. Proteins 2000;39:331–342.
15. Ma B, Elkayam T, Wolfson H, Nussinov R. Protein–protein interactions: structurally conserved residues distinguish between binding sites and exposed protein surfaces. Proc Natl Acad Sci USA 2003; 100:5772–5777.
16. Mitchell JC, Shahbaz S, Ten Eyck LF. Interfaces in molecular docking. Mol Simulat 2004;30:97–106.
17. Towell G, Shavlik J. Knowledge-based artificial neural networks. Artif Intell J 1994;70:119–165.
18. Mangasarian O, Shavlik J, Wild E. Knowledge-based kernel approximation. J Machine Learn Res 2004;5:1127–1141.
19. Mitchell JC, Kerr R, Ten Eyck LF. Rapid atomic density methods for molecular shape characterization. J Mol Graph Model 2001;19: 325–330.
20. Kortemme T, Kim DE, Baker D. Computational alanine scanning of protein–protein interfaces. Sci STKE 2004; p 12. Available at http://robetta.bakerlab.org/.
21. Berman HM, Westbrook J, Feng Z, Gilliland G, Bhat TN, Weissig H, Shindyalov IN, Bourne PE. The Protein Data Bank. Nucleic Acids Res 2000;28:235–242. Available at http://www.rcsb.org/pdb/.
22. Thorn KS, Bogan AA. ASEdb: a database of alanine mutations and their effects on the free energy of binding in protein interactions. Bioinformatics 2001;17:284–285. Available at http://thornlab.cgr.harvard.edu/hotspot/.
23. Fischer TB, Arunachalam KV, Bailey D, Mangual V, Bakhru S, Russo R, Huang D, Paczkowski M, Lalchandani V, Ramachandra C, Ellison B, Galer S, Shapley J, Fuentes E, Tsai J. The binding interface database (BID): a compilation of amino acid hot spots in protein interfaces. Bioinformatics 2003;19:1453–1454. Available at http://tsailab.org/BID/.
24. Wang G, Dunbrack RL Jr. PISCES: a protein sequence culling server. Bioinformatics 2003;19:1589–1591. Available at http://dunbrack.fccc.edu/PISCES.php.
25. Quinlan JR. C4.5: Programs for machine learning. San Francisco: Morgan Kaufmann Publishers, 1993. Available at http://www.rulequest.com/.
26. Wei L, Altman RB, Chang JT. Using the radial distributions of physical features to compare amino acid environments and align amino acid sequences. In: Altman RB, Dunker AK, Hunter L, Klein TE, editors, Pacific symposium of biocomputing 97, Maui, HI, Singapore: World Scientific, 1997; p 497–508.
27. Wei L, Altman RB. Recognizing protein binding sites using statistical descriptions of their 3D environments. In: Altman RB, Dunker AK, Hunter L, Klein TE, editors, Pacific Symposium of Biocomputing 98, Maui, HI. Singapore: World Scientific, 1998; p 465–476.
28. Joachims T. Advances in Kernel methods: support vector learning, chapter Making large-scale SVM learning practical. Cambridge: MIT Press, 1999, pp 169–184. Available at http://svmlight.joachims.org/.
29. Witten IH, Frank E. Data mining: practical machine learning tools and techniques, 2nd ed. San Francisco: Morgan Kaufmann Publishers, 2005. Available at http://www.cs.waikato.ac.nz/ml/.
30. Bagley SC, Altman RB. Characterizing the microenvironment surrounding protein sites. Protein Sci 1995;4:622–635.
31. Vriend G. WHAT IF: a molecular modeling and drug design program. J Mol Graphics 1990;8:52–56.
32. Hooft RW, Sander C, Vriend G. Positioning hydrogen atoms by optimizing hydrogen-bond networks in protein structures. Proteins 1996;26:363–376.
33. Stone M. Cross-validatory choice and assessment of statistical predictions. J R Stat Soc B Met 1974;36:111–147.
34. Geisser S. The predictive sample reuse method with applications. J Am Stat Assoc 1975;70:320–327.
35. Kohavi R. A study of cross-validation and bootstrap for accuracy estimation and model selection. In: Proceedings of the Fourteenth International Joint Conference on Artificial Intelligence, IJCAI 95, Montrèal, Quèbec, Canada, Morgan Kaufmann, 1995; pp 1137–1145.
36. R Development Core Team. R: a language and environment for statistical computing. Vienna, Austria: R Foundation for Statistical Computing; 2004. Available at http://www.r-project.org/.
37. Chin D, Sloan DJ, Quiocho FA, Means AR. Functional consequences of truncating amino acid side chains located at a calmodulin-peptide interface. J Biol Chem 1997;272:5510–5513.
38. Bagchi IC, Huang QH, Means AR. Identification of amino acids essential for calmodulin binding and activation of smooth muscle myosin light chain kinase. J Biol Chem 1992;267:3024–3029.
39. Nickel J, Dreyer MK, Kirsch T, Sebald W. The crystal structure of the BMP-2:BMPR-IA complex and the generation of BMP-2 antagonists. J Bone Joint Surg Am 2001;83-A (Suppl 1):7–14.
40. Kirsch T, Nickel J, Sebald W. BMP-2 antagonists emerge from alterations in the low-affinity binding epitope for receptor BMPR-II. EMBO J 2000;19:3314–3324.
41. Matson JS, Nilles ML. Interaction of the Yersinia pestis type III regulatory proteins LcrG and LcrV occurs at a hydrophobic interface. BMC Microbiol 2002;2:16.