JOURNAL OF
CHEMICAL INFORMATION
AND MODELING

Article

pubs.acs.org/jcim

Solvent Accessible Surface Area-Based Hot-Spot Detection Methods for Protein−Protein and Protein−Nucleic Acid Interfaces

Cristian R. Munteanu,[†] António C. Pimenta,[‡] Carlos Fernandez-Lozano,[†] André Melo,[‡] Maria N. D. S. Cordeiro,[‡] and Irina S. Moreira*[,‡,§]

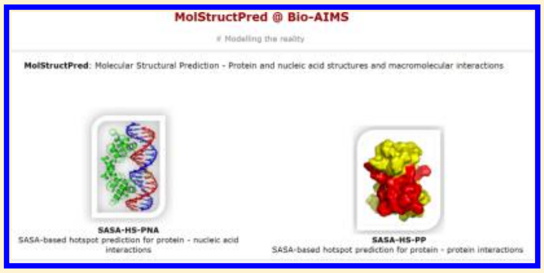
[†]Information and Communication Technologies Department, Computer Science Faculty, University of A Coruna, Campus de Elviña s/n, 15071 A Coruña, Spain

[‡]REQUIMTE/Departamento de Química e Bioquímica, Faculdade de Ciências da Universidade do Porto, Rua do Campo Alegre s/n, 4169-007 Porto, Portugal

[§]CNC—Center for Neuroscience and Cell Biology, Universidade de Coimbra, Rua Larga, FMUC, Polo I, 1°andar, 3004-517 Coimbra, Portugal

Ⓢ *Supporting Information*

ABSTRACT: Due to the importance of hot-spots (HS) detection and the efficiency of computational methodologies, several HS detecting approaches have been developed. The current paper presents new models to predict HS for protein−protein and protein−nucleic acid interactions with better statistics compared with the ones currently reported in literature. These models are based on solvent accessible surface area (SASA) and genetic conservation features subjected to simple Bayes networks (protein−protein systems) and a more complex multi-objective genetic algorithm−support vector machine algorithms (protein−nucleic acid systems). The best models for these interactions have been implemented in two free Web tools.



MolStructPred @ Bio-AIMS

Modelling the reality

MolStructPred: Molecular Structural Prediction - Protein and nucleic acid structures and macromolecular interactions

SASA-HS-PNA
SASA-based hotspot prediction for protein - nucleic acid interactions

SASA-HS-PP
SASA-based hotspot prediction for protein - protein interactions

■ INTRODUCTION

Proteins are essential macromolecules in several biochemical processes due to their ability to perform multiple tasks such as enzymatic catalysis, transport, and signal transduction among others.[1,2] To perform these tasks, proteins have to form complexes with other biomolecules, which is a fundamental step for several biochemical processes. It is therefore crucial to attain a complete understanding of the structural atomistic details of these interactions to better develop methods to influence the binding. Although protein-based interfaces usually comprehend a high number of residues, it has been proved that the majority of the binding energy can be accounted for by the interaction of a small number of residues known as hot-spots (HS).[3−5] In order to investigate the contribution of a residue to the binding energy, the residue of interest is mutated to an alanine, and the binding free energy difference ($\Delta\Delta G_{\text{binding}}$) is calculated. The definitions of HS vary among authors, but it is most commonly accepted that HS are defined as residues with $\Delta\Delta G_{\text{binding}} \geq 2.0$ kcal mol^{-1}; the ones with $\Delta\Delta G_{\text{binding}} < 2.0$ kcal mol^{-1} are called null-spots (NS).[5]

HS can be found experimentally, using molecular biology and thermodynamic methods upon alanine scanning mutagenesis (ASM), but these are not only expensive but often complex and time-consuming. Due to these difficulties, computational approaches with a higher relation efficiency/cost and lower experimental time have been developed. These can be generally described as empirical functions or knowledge-based models, all atom methods and feature-based approaches.[6−10] Although fully atomistic models were shown to accurately predict HS and be able to fully characterize this type of interactions, the time and complexity involved is often very high.[9,11] In our previous work we investigated feature-based methods that combine solvent accessible surface area (SASA) descriptors calculated from static structures and molecular dynamics (MD) ensembles, which were analyzed by a support vector machine (SVM) algorithm.[6] We presented a new HS predictive model: SASA-based hot-spot detection (SBHD). However, at the time our method was only applied to a small number of complexes, and it has been demonstrated to have a high number of false positives (incorrectly detect NS as HS). To improve this aspect and achieve a greater overall performance, we have added an extra feature (residue genomic conservation), significantly extending the data as well as the number of different machine learning (ML) techniques used. We have also tested two different separated data sets: (i) protein−protein and (ii) protein−nucleic acid complexes.

With these additions to the model, we attained a more accurate and time-efficient HS detection methodology. Moreover, our method can be applied not only to protein−protein but also to protein−nucleic acid complexes. This is the first time that such a method has been applied solely to this type of

Received: December 22, 2014
Published: April 6, 2015

DOI: 10.1021/ci500760m

highly challenging interface. Web servers were also constructed and made available for the scientific community at BioAIMS portal (http://bio-aims.udc.es/MolStructPred.php). The code of the Web tools that includes Python,[12] VMD,[13] and Weka scripts[14] for SASA calculation and ML model prediction is available as a pySBHD repository (https://github.com/muntisa/pySBHD).

The data sets S1 (http://dx.doi.org/10.6084/m9.figshare.1300069), S2 (http://dx.doi.org/10.6084/m9.figshare.1300071), S3 (http://dx.doi.org/10.6084/m9.figshare.1348809), and Pronit (http://dx.doi.org/10.6084/m9.figshare.1300072) are also available for download.

■ MATERIALS AND METHODS

1. Data Sets Construction. The data sets are constituted by protein complexes for which there simultaneously exist experimental alanine scanning mutagenesis data, genetic conservation scores, and tridimensional crystallographic structures of the bounded complex. Following other authors procedure,[15−17] protein sequences in each data set were filtered to ensure that a maximum of 35% sequence identity could be found for at least one protein in each interface. The identity sequence matrix was calculated using the CLUSTAL OMEGA web server.[18] Four different data sets were studied, three for the protein−protein case and one for the protein−nucleic acid systems. For protein−protein complexes, the data set was split between quantitative (data sets S1 and S3) and qualitative data (data set S2). Table S1 in the Supporting Information lists the complexes with quantitative HS information gathered from the Alanine Scanning Energetics database (ASEdb).[19] This table comprises 477 mutations from 15 complexes. The complexes present in Table S2 represent qualitative data retrieved from the Binding Interface Database (BID)[20] and from Zhu et al.[17] It comprises 91 mutations from 15 complexes, and only mutations identified with "strong" were considered HS. Table S3 comprises 222 mutations from 28 complexes retrieved from SKEMPI[21] and PINT[22] data sets. Table S1 was used as training/test set and Tables S2 and S3 as independent test sets. Table S4 lists the information for 177 mutation existent at 28 protein−DNA/RNA complexes gathered from the Protein−Nucleic Acid Complex Database (Pronit)[23−25] concerning single alanine mutations. Since $\Delta\Delta G_{\text{binding}}$ values were not explicitly available, these were calculated by $\Delta\Delta G_{\text{binding}} = \Delta G_{\text{mutant}} - \Delta G_{\text{wild-type}}$. For cases with different experimental data for the same residue, we have use the median value for the construction of our data set. Table S4 was used as training/test set for the protein−nucleic acid case.

2. PDB Structures Preparation. The crystal structures were retrieved from the Protein Data Bank (PDB),[26] and all water molecules, ions, and other small ligands were removed. Only monomeric chains involved in the interfaces identified in Tables S1−S4 were left in the PDB files. In the presence of symmetrical models, only one protein−protein complex was selected. In the cases with modified residues, these were changed back to the original amino acid as identified in the PDB.

3. Sequence/Structural Features. Various authors have demonstrated that the evolutionary rate of the interfacial residues depends of the degree of solvent exposition upon binding.[27,28] They concluded that residues not exposed to solvent as well as interfacial residues have lower evolutionary rates, and therefore are universally important to protein stability.[28] This way, HS are thought as conserved sites as the degree of conservation of residues is similar to the inverse of the rate of evolution.[29] Taking this fact into account, we focused our attention not only in the SASA features already used by us in a previous work[6] but also on the genetic conservation of residues at protein-based interfaces. The CONSURF server[29] was used to calculate the conservation score for each amino acid at an interfacial position for all complexes. We have also calculated 12 different SASA descriptors from the PDB complexes. SASA was initially defined by Lee and Richards[30] and is generally calculated using the algorithm developed by Shrake and Rupley.[31] It can be described as the area of the surface traced by the center of a probe sphere, whose radius is the nominal radius of the solvent, as it rolls over the van der Waals surface of the molecule. The VMD[13] built-in *measure sasa* command was used to calculate he SASA value for all residues of all protein using the assigned radius for each atom extended by 1.4 Å to find points on a sphere that is exposed to solvent. ${}_{\text{comp}}\text{SASA}_i$ is the solvent accessible surface area of residue *i* in complex form, while ${}_{\text{mon}}\text{SASA}_i$ is the residue SASA in the monomer form. ΔSASA_i, the SASA variation upon complexation, is determined using these features (eq 1). ${}_{\text{rel}}\text{SASA}_i$ is determined using the results from ΔSASA for each residue and dividing it by the ${}_{\text{mon}}\text{SASA}_i$ value for the same residue, providing a differentiation between residues with equal ΔSASA but different absolute monomer SASA values (eq 2). Further four features (${}_{\text{comp/res}}\text{SASA}_i$, ${}_{\text{mon/res}}\text{SASA}_i$, ${}_{\Delta/\text{res}}\text{SASA}_i$, and ${}_{\text{rel/res}}\text{SASA}_i$), defined by eqs 3, 4, 5, and 6, were determined employing amino acid standardization by dividing the previous features by approximate average protein ${}_{\text{res}}\text{SASA}_{\text{r}}$ values as determined by Miller and colleagues,[32,33] Gly = 85, Ala = 113, Cys = 1 40, Asp = 151, Glu = 183, Phe = 218, His = 194, Ile = 182, Lys = 211, Leu = 180, Met = 204, Asn = 158, Pro = 143, Gln = 189, Arg = 241, Ser = 122, Thr = 146, Val = 160, Trp = 259, and Tyr = 229, with "r" being the respective residue type. These values determined by Miller were then replaced by our own protein average ${}_{\text{ave}}\text{SASA}_{\text{r}}$ values for each amino acid type in its respective protein and used to assert amino acid standardization, resulting in ${}_{\text{comp/ave}}\text{SASA}_i$, ${}_{\text{mon/ave}}\text{SASA}_i$, ${}_{\Delta/\text{ave}}\text{SASA}_i$, and ${}_{\text{rel/ave}}\text{SASA}_i$ defined in eqs 7−10.

$$\Delta\text{SASA}_i = |{}_{\text{comp}}\text{SASA}_i - {}_{\text{mon}}\text{SASA}_i| \tag{1}$$

$${}_{\text{rel}}\text{SASA}_i = \frac{\Delta\text{SASA}_i}{{}_{\text{mon}}\text{SASA}_i} \tag{2}$$

$${}_{\text{comp/res}}\text{SASA}_i = \frac{{}_{\text{comp}}\text{SASA}_i}{{}_{\text{res}}\text{SASA}_{\text{r}}} \tag{3}$$

$${}_{\text{mon/res}}\text{SASA}_i = \frac{{}_{\text{mon}}\text{SASA}_i}{{}_{\text{res}}\text{SASA}_{\text{r}}} \tag{4}$$

$${}_{\Delta/\text{res}}\text{SASA}_i = \frac{\Delta\text{SASA}_i}{{}_{\text{res}}\text{SASA}_{\text{r}}} \tag{5}$$

$${}_{\text{rel/res}}\text{SASA}_i = \frac{{}_{\text{rel}}\text{SASA}_i}{{}_{\text{res}}\text{SASA}_{\text{r}}} \tag{6}$$

$${}_{\text{comp/ave}}\text{SASA}_i = \frac{{}_{\text{comp}}\text{SASA}_i}{{}_{\text{ave}}\text{SASA}_{\text{r}}} \tag{7}$$

DOI: 10.1021/ci500760m

Table 1. Statistical Performance of the 10 Best Models for Detection of HS at Protein−Protein Complexes

			data set S1 − Training					data set S1 − Test/Cross Validation					
Weka classifier/model		features used	TPR	FPR	specificity	precision	F1 score	AUROC	TPR	FPR	specificity	precision	F1 score	AUROC
FS	**BayesNet**	**ConSurf score, $\Delta SASA_i$, ${}_{rel/res}SASA_i$, ${}_{rel/ave}SASA_i$**	**0.79**	**0.21**	**0.79**	**0.87**	**0.83**	**0.85**	**0.78**	**0.26**	**0.72**	**0.86**	**0.82**	**0.81**
	BayesNet	ConSurf score, ${}_{comp}SASA_i$, ${}_{rel/res}SASA_i$, ${}_{rel/ave}SASA_i$	0.81	0.29	0.71	0.86	0.83	0.85	0.78	0.34	0.66	0.84	0.81	0.81
	BayesNet	$\Delta SASA_i$, ${}_{rel}SASA_i$, ${}_{comp/res}SASA_i$, ${}_{rel/res}SASA_i$, ${}_{rel/ave}SASA_i$	0.76	0.20	0.80	0.87	0.81	0.85	0.75	0.31	0.69	0.84	0.79	0.80
	BayesNet	$\Delta SASA_i$, ${}_{rel}SASA_i$, ${}_{comp/res}SASA_i$, ${}_{rel/ave}SASA_i$	0.78	0.24	0.76	0.86	0.82	0.85	0.76	0.36	0.64	0.83	0.79	0.81
	BayesNet	ConSurf score, ${}_{comp}SASA_{i,}$ ${}_{comp/res}SASA_i$, ${}_{rel/res}SASA_i$	0.82	0.31	0.69	0.86	0.84	0.85	0.79	0.38	0.62	0.83	0.81	0.79
noFS	ADTree	pool	0.89	0.41	0.59	0.88	0.89	0.91	0.88	0.51	0.49	0.86	0.87	0.82
	RulesDTNB	pool	0.90	0.44	0.56	0.89	0.90	0.87	0.87	0.55	0.45	0.85	0.86	0.82
	Bayesnet	pool	0.84	0.31	0.69	0.87	0.85	0.87	0.81	0.35	0.65	0.85	0.83	0.81
	LADtree	pool	0.91	0.37	0.63	0.91	0.91	0.90	0.87	0.49	0.51	0.85	0.86	0.81
	Metabagging	pool	0.91	0.42	0.58	0.91	0.91	0.94	0.88	0.58	0.42	0.86	0.87	0.81
			data set S2					data set S3					
Weka classifier/model		features used	TPR	FPR	specificity	precision	F1 score	AUROC	TPR	FPR	specificity	precision	F1 score	AUROC
FS	**BayesNet**	**ConSurf score, $\Delta SASA_i$, ${}_{rel/res}SASA_i$, ${}_{rel/ave}SASA_i$**	**0.67**	**0.35**	**0.65**	**0.69**	**0.68**	**0.75**	0.63	0.52	0.48	0.65	0.65	0.60
	BayesNet	ConSurf score, ${}_{comp}SASA_i$, ${}_{rel/res}SASA_i$, ${}_{rel/ave}SASA_i$	0.64	0.36	0.64	0.66	0.65	0.66	0.67	0.53	0.47	0.67	0.67	0.60
	BayesNet	$\Delta SASA_i$, ${}_{rel}SASA_i$, ${}_{comp/res}SASA_i$, ${}_{rel/res}SASA_i$, ${}_{rel/ave}SASA_i$	0.65	0.33	0.67	0.68	0.66	0.74	0.63	0.51	0.49	0.66	0.64	0.58
	BayesNet	$\Delta SASA_i$, ${}_{rel}SASA_i$, ${}_{comp/res}SASA_i$, ${}_{rel/ave}SASA_i$	0.70	0.31	0.69	0.71	0.70	0.70	0.64	0.53	0.47	0.65	0.64	0.58
	BayesNet	ConSurf score, ${}_{comp}SASA_i$, ${}_{comp/res}SASA_i$, ${}_{rel/res}SASA_i$.	0.61	0.46	0.54	0.60	0.60	0.66	0.66	0.55	0.45	0.66	0.66	0.59
noFS	ADTree	pool	0.69	0.45	0.55	0.78	0.73	0.78	0.70	0.62	0.38	0.66	0.67	0.61
	RulesDTNB	pool	0.65	0.51	0.49	0.74	0.69	0.71	0.71	0.66	0.34	0.65	0.66	0.60
	Bayesnet	pool	0.64	0.40	0.60	0.74	0.69	0.73	0.62	0.53	0.47	0.65	0.64	0.60
	LADtree	pool	0.65	0.48	0.52	0.71	0.68	0.71	0.68	0.61	0.39	0.64	0.65	0.60
	Metabagging	pool	0.67	0.59	0.41	0.70	0.68	0.68	0.72	0.69	0.31	0.65	0.65	0.62

Note: FS = filter feature selection (before the classification to select the best features), BayesNet = Bayesian network classifier, ADTree = alternative decision tree classifier, RulesDTNB = decision table/Naïve Bayes hybrid classifier, LADtree = multiclass alternating decision tree using the LogitBoost strategy, Metabagging = classifier that use bagging to reduce variance; pool = all features.

$$_{\text{mon/ave}}\text{SASA}_i = \frac{_{\text{mon}}\text{SASA}_i}{_{\text{ave}}\text{SASA}_\text{r}} \tag{8}$$

$$_{\Delta\text{/ave}}\text{SASA}_i = \frac{\Delta\text{SASA}_i}{_{\text{ave}}\text{SASA}_\text{r}} \tag{9}$$

$$_{\text{rel/ave}}\text{SASA}_i = \frac{_{\text{rel}}\text{SASA}_i}{_{\text{ave}}\text{SASA}_\text{r}} \tag{10}$$

As the SASA features described from eqs 3 to 10 have very low order of magnitude, the results presented here were multiplied by a factor of 10^3. To summarize, the features 1−13 used in this work were the following: CONSURF score, $_{\text{comp}}\text{SASA}_i$, $_{\text{mon}}\text{SASA}_i$, ΔSASA_i, $_{\text{rel}}\text{SASA}_i$, $_{\text{comp/res}}\text{SASA}_i$, $_{\text{mon/res}}\text{SASA}_i$, $_{\Delta\text{/res}}\text{SASA}_i$, $_{\text{rel/res}}\text{SASA}_i$, $_{\text{comp/ave}}\text{SASA}_i$, $_{\text{mon/ave}}\text{SASA}_i$, $_{\Delta\text{/ave}}\text{SASA}_i$, and $_{\text{rel/ave}}\text{SASA}_{\text{i.}}$

4. Baseline Machine Learning Techniques. ML techniques are widely used in very different difficult problems ranging from accurately classifying microbial communities in the microbiome,[34] to predicting protein complex biological functions,[35,36] to estimating body pose of a human from the point cloud obtained from a depth sensor.[37] We started our experiments using Weka software[14] ML methods for each classification problem (protein−protein and protein−nucleic acid) in order to establish the baseline false positive rate (FPR) and area under the receiver operator curve (AUROC) performance and be able to compare the experiments.

All the classification models used a 10-fold cross validation of the training set in order to avoid over-fitting and to obtain model's generalization error. Thus, cross validation is able to indicate how well the model is expected to deal with new unknown data. Data are separated randomly into 10 isolated parts, using 9 of the 10 parts to train the model and taking the remaining fold of data to test the final performance of the model. It is necessary to repeat this process nine times more, using each time other part of the data as the testing data. We obtained a measure of the final performance of each model by averaging the ten runs of the cross validation process. We evaluated the accuracy of the models with the AUROC, the true positive rate (TPR/recall/sensitivity, eq 11), the FPR (fall out, eq 12), the precision (positive predictive value/PPV, eq 13), the specificity (true negative rate/TNR, eq 14), and the F1 score (eq 15).

$$\text{TPR} = \frac{\text{TP}}{\text{TP} + \text{FN}} \tag{11}$$

$$\text{FPR} = \frac{\text{FP}}{\text{FP} + \text{TN}} \tag{12}$$

$$\text{PPV} = \frac{\text{TP}}{\text{TP} + \text{FP}} \tag{13}$$

$$\text{specificity} = \frac{\text{TN}}{\text{FP} + \text{TN}} \tag{14}$$

$$\text{F1} = \frac{2\text{TP}}{2\text{TP} + \text{FP} + \text{FN}} \tag{15}$$

In all these equations, TP stand for true positive (predicted hot-spots that are actual hot-spots), FP stands for false positive (predicted hot-spots that are not actual hot-spots), FN stands for false negative (non-predicted hot-spots that are actual hot-spots), and TN stands for the true negatives (correctly predicted null-spots).

4.1. Feature Selection Approaches. In order to improve the performance of a ML algorithm and to reduce the number of features in the input space, we also performed a feature selection (FS) approach, as the number and relevance of the input variables can affect the performance of the model. The aim of this process was to find the best subset of features describing as well or better the structure of the data. The FS process can answer the main question of how many and which are the most important features to discriminate between HS and NS. In this kind of problems, it is usual to perform a FS approach in order to improve the performance of the classification and to choose the best features to solve the classification problem. There are mainly three approaches for FS: filter (before to search for a model), wrapper, and embedded (inside the classification method).[38] In this study, we used the filter and wrapper methods. Please refer to the Results section in order to find the particular approaches implemented for this study.

5. Comparison with Other Software. The validity and performance of the method were determined by analyzing similar HS prediction methodologies in terms of overall AUROC, F1, TPR, FPR, TNR, and precision over the same data used by us. The available HS prediction methodologies focus solely on protein−protein interfaces, and therefore we could only compared our method develop for this particular systems. The main servers tested were PredictProtein (former ISIS[39]), ROBETTA,[40] KFC2-A,[41] KFC2-B,[41] and HotPoint.[42]

6. Web Tool Implementation. The best models have been implemented as free online tools: SASA-HS-PNA (SASA-based hot-spot prediction for protein−nucleic acid interactions) at http://bio-aims.udc.es/SASA-HS-PNA.php and SASA-HS-PP (SASA-based hot-spot prediction for protein−protein interactions) at http://bio-aims.udc.es/SASA-HS-PP.php. The tools are based on Python,[12] HTML/PHP, and they use VMD[13] inputs and outputs. The user is able to predict HS for protein−protein and protein−nucleic acid interactions by uploading PDB files or CONSURF data files and by defining the interaction monomer chains.

■ RESULTS

1. Protein−Protein Complexes, Final Model Description. The models were trained and tested on a data set (Table S1) and comprehensively tested on two independent test sets (Tables S2 and S3). The statistical performances of the best ten models are presented in Table 1. In this particular classification problem, we found that the best FS approach is a filter with BayesNet classifier. This FS approach assess the relevance of features by looking only at the intrinsic properties of the data set.[38] The best classifier uses four features: CONSURF score, ΔSASA_i, $_{\text{rel/res}}\text{SASA}_i$, and $_{\text{rel/ave}}\text{SASA}_i$. When used as training set, it shows the best predictive power (TPR = 0.79, FPR = 0.21, precision = 0.87, F1 score = 0.83, and AUROC = 0.85). So, 79% of true HS are correctly predicted, and 87% of the predicted HS are identified as true HS. The F1 score and AUROC values are especially higher than the ones reported in literature. After testing the model by cross-validation it remained as the highest predictive model (TPR = 0.78, FPR = 0.26, precision = 0.86, F1 score = 0.82, and AUROC = 0.81), which was particularly true when further applied to the two independent test sets, S2 (TPR = 0.67, FPR = 0.35, precision = 0.69, F1 score = 0.68, and AUROC = 0.75) and S3 (TPR =

Table 2. Statistical Performance of the Best Model for Detection of HS at Protein−Protein Complexes (Bayes Network on Four Features: ConSurf score, $\Delta SASA_i$, $_{rel/res}SASA_i$, and $_{rel/ave}SASA_i$) by Type of Amino Acid

	TPR	FPR	specificity	precision	F1 score	AUROC
		Data Set S1				
total residues (80 HS/434 NS)	0.79	0.21	0.79	0.87	0.83	0.85
charged residues (28 HS/165 NS)	0.78	0.17	0.83	0.90	0.82	0.90
negatively charged residues (14 HS/68 NS)	0.85	0.09	0.91	0.91	0.86	0.95
positively charged residues (14 HS/97 NS)	0.78	0.17	0.83	0.90	0.82	0.90
polar residues (25 HS/152 NS)	0.83	0.48	0.52	0.92	0.87	0.81
nonpolar residues (15 HS/75 NS)	0.75	0.26	0.74	0.82	0.77	0.83
aromatic residues (10 HS/37 NS)	0.65	0.22	0.78	0.80	0.68	0.81
		Data Set S2				
total residues (35 HS/60 NS)	0.67	0.35	0.65	0.69	0.68	0.75
charged residues (13 HS/33 NS)	0.61	0.48	0.52	0.64	0.62	0.72
negatively charged residues (5 HS/11 NS)	0.56	0.42	0.58	0.63	0.58	0.81
positively charged residues (8 HS/22 NS)	0.63	0.53	0.47	0.65	0.64	0.69
polar residues (5 HS/17 NS)	0.82	0.20	0.80	0.85	0.83	0.90
nonpolar residues (16 HS/9 NS)	0.68	0.33	0.67	0.70	0.69	0.74
aromatic residues (6 HS/8 NS)	0.64	0.31	0.69	0.70	0.64	0.71
		Data Set S3				
total residues (79 HS/223 NS)	0.63	0.45	0.55	0.67	0.65	0.62
charged residues (24 HS/91 NS)	0.70	0.60	0.37	0.70	0.70	0.55
negatively charged residues (10 HS/39 NS)	0.71	0.52	0.48	0.74	0.73	0.65
positively charged residues (14 HS/52 NS)	0.67	0.67	0.33	0.67	0.67	0.47
polar residues (23 HS/64 NS)	0.72	0.41	0.59	0.73	0.72	0.66
nonpolar residues (23 HS/56 NS)	0.53	0.55	0.45	0.58	0.55	0.54
aromatic residues (17 HS/36 NS)	0.51	0.55	0.45	0.53	0.53	0.51

Table 3. Statistical Performance of Various Algorithms Tested for Data Sets S1, S2, and S3

	Data Set S1					Data Set S2					Data Set S3				
algorithm	TPR	FPR	specificity	precision	F1 score	TPR	FPR	specificity	precision	F1 score	TPR	FPR	specificity	precision	F1 score
SBHD2	0.79	0.21	0.79	0.87	0.83	0.67	0.35	0.65	0.69	0.68	0.63	0.52	0.48	0.65	0.65
Robetta	0.36	0.06	0.94	0.72	0.48	0.46	0.28	0.72	0.69	0.55	0.25	0.13	0.87	0.73	0.37
KFCA	0.38	0.09	0.91	0.54	0.44	0.58	0.24	0.76	0.66	0.61	0.48	0.12	0.88	0.56	0.52
KFCB	0.38	0.12	0.88	0.31	0.34	0.65	0.29	0.71	0.43	0.52	0.52	0.16	0.84	0.31	0.39
HotPoint	0.36	0.11	0.89	0.41	0.38	0.67	0.28	0.72	0.46	0.54	0.37	0.14	0.86	0.51	0.43
PredictProtein	0.21	0.13	0.87	0.31	0.25	0.50	0.37	0.63	0.14	0.22	0.24	0.20	0.80	0.18	0.21

0.63, FPR = 0.52, precision = 0.65, F1 score = 0.65, and AUROC = 0.60). We have also analyzed HS predictions by amino acid type (Table 2). We separated the amino acids into different sets according to their hydrophilicity/hydrophobicity character: negatively charged (Asp and Glu), positively charged (His, Lys, and Arg), charged (Asp, Glu, Lys, Arg, His), polar (Ser, Thr, Asn, Gln, Tyr), nonpolar (Val, Ile, Leu, Met, Phe, Trp), and aromatic (Phe, Trp, Tyr, His). Special residues as Cys, Gly, and Pro were not included in these groups. Our algorithm was also assessed against some of the state-of-the-art methods available on Web servers (Tables S5−S7). The overall prediction performances of all softwares/servers are listed in Table 3 and by amino acid type group in Table S8. All residues types have a good predictive value (AUROC) ranging between 0.81 and 0.95 for data set S1, between 0.71 and 0.90 for data set S2, and between 0.50 and 0.62 for data set S3. It suggests that our model is not biased toward a single amino acid type. It is specially rewording the high statistical value attained for the charged residues, typical problematic residues at the majority of algorithms as shown in Table S8. The F1 values for this type of residues range from 0.20 to 0.56 in S1 (0.24 and 0.62 in S2; 0.21 and 0.48 in S3) for the tested software solutions, which clearly contrast with the 0.82 value at S1 (0.62 in S2 and 0.72 in S3) attained by our algorithm. The amino acid composition of HS has shown not to be equally distributed, with Trp, Tyr, and Arg residues being the most common ones due to their size and conformation.[43] Tables 2 and S8 shows that this type of aromatic residues (such as Trp and Tyr, common HS) are also well predicted (AUROC = 0.81, AUROC = 0.71, AUROC = 0.51 for data sets S1, S2, and S3, respectively) by our algorithm and with higher performance than by other available software solutions.

ROBETTA[44] is an energy-based method tested on 380 mutants and with a reported average error ($\Delta\Delta G_{\text{binding-theor}} - \Delta\Delta G_{\text{binding-exptl}}$) of 0.83 kcal/mol. The residue degree of exposure to the solvent was taken into account. KFCA is a hybrid method that combines KFC and Robetta. KFC[15−17] was trained in a variety of features such as residue size, atomic contacts, hydrogen bonds or chemical type. It was based on various SASA features and different structural characteristics and trained on 265 residues. HotPoint[45] uses an empirical formula, and it is based on the SASA value of the complex

Table 4. Statistical Performance of the Three Models for Detection of HS at Protein−Nucleic Acid Complexes

		training						test − cross validation					
model	features	TPR	FPR	specificity	precision	F1 score	AUROC	TPR	FPR	specificity	precision	F1 score	AUROC
GA-SVM-Full[a]	**ConSurf score, $_{\mathrm{del/res}}\mathbf{SASA}_i$**	**0.82**	**0.30**	**0.70**	**0.82**	**0.85**	**0.83**	**0.82**	**0.37**	**0.63**	**0.81**	**0.81**	**0.78**
GA-SVM-Linear (svm_C = 38.169228)	pool	0.67	0.41	0.39	0.76	0.71	0.71	0.61	0.31	0.69	0.74	0.64	0.70
Random Forest	ConSurf score, $_{\mathrm{comp}}\mathrm{SASA}_i$, $_{\mathrm{mon}}\mathrm{SASA}_i$	0.94	0.55	0.45	0.92	0.90	0.87	0.73	0.40	0.60	0.74	0.74	0.69

[a]Best model: GA-SVM, svm_C: 110.373487 | svm_gamma: 18.573694.

between other features. PredictProtein (former ISIS[46]) is a sequence-based method that uses a neural network and was assessed on 296 mutants. It should be noted that PredictProtein was developed for finding binding site residues rather than HS, and that this software is based solely on sequence, and so no structural information is used, which makes a comparison less straightforward. As perceived (Tables 2 and S8), our method suppresses by far all the remaining ones for these particular data sets. F1 scores are much higher on data set S1 (0.82 against 0.61 for the second best), on data set S2 (0.68 against 0.60 for KFCA and 0.65 on data set S3). TPR and precision are also particularly high. However, Table 2 also shows that our method has higher Specificity values for all data sets, which shows that the performance of our algorithm in the negative data set, its ability to predict NS, is a bit inferior to the other available software solutions.

2. Protein−Nucleic Acid Complexes, Final Model Description. In this particular classification problem (Table 4), we found that the best FS approach is a wrapper with a SVM as a decision function. Genetic algorithms (GA)[47] are a bioinspired meta-heuristics finding for the best subset[36] of input features that better reproduce the original structure of the data. Wrapper approaches treat the problem of finding the best subset in the same step as the model selection as these approaches embed the model hypothesis search within the feature subset search.[38] Based on GAlib,[48] we enhanced it in two different ways. First we enhanced it in order to add a real number representation and to find the best parameter for a simple linear kernel (parameter C) using all the available information (GA-SVM-Linear). Second we modified it in order to add a binary representation for each individual as a feature mask for the input feature space and a real-number representation of the parameters of a radial basis function (RBF) kernel (parameters C and γ). Thus, we were able to perform a feature selection process while we were seeking for the best combination of parameters for the kernel according to this particular subset of features (GA-SVM-Full), such approach is considered a multi-objective GA in which different criteria are optimized simultaneously.[49] SVM decision function was obtained by LIBSVM.[50] Fitness function was developed in order to reduce FPR (type I error) and maximize AUROC. We performed the same set of experiments with a particle swarm optimization (PSO) approach with a SVM as a decision function based on the latest Standard PSO implementation (SPSO-2011)[51,52] and also following a filter approach with Weka classifiers (best results were achieved with Random Forest[53]), but we were not able to improve the GA-based results.

In protein−nucleic acid systems, the HS are particularly enriched with Trp, Phe, and Tyr. However, there is a higher occurrence of positively charged residues and a reduced occurrence of hydrophobic and negatively charged residues.[54] We also assessed the performance of our method by amino acid type. From Table 5 we conclude that the method works for all types of residues, particularly for the polar ones. However, the statistical values attained are a bit worse than the ones for the protein−protein complexes as protein−nucleic acids are particularly challenging interfaces much more polar due to the phosphate groups on the DNA strands and the presence of a higher number of positively charged groups on the protein monomer.

Table 5. Statistical Performance of the Best Model for Detection of HS at Protein−Nucleic Acid Complexes (Genetic Algorithm-Support Vector Machine Algorithm Based on Two Features: ConSurf score and $_{\mathrm{del/res}}\mathrm{SASA}_i$) by Type of Amino Acid

Data Set − Pronit	TPR	FPR	specificity	precision	F1 score	AUROC
total residues (43 HS/128 NS)	0.68	0.42	0.38	0.76	0.71	0.65
charged residues (24 HS/77 NS)	0.63	0.49	0.51	0.68	0.65	0.61
negatively charged residues (2 HS/23 NS)	0.76	0.48	0.52	0.88	0.81	0.51
positively charged residues (22 HS/54 NS)	0.60	0.49	0.51	0.63	0.61	0.61
polar residues (11 HS/30 NS)	0.76	0.15	0.85	0.84	0.77	0.83
nonpolar residues (7 HS/17 NS)	0.96	0.10	0.90	0.96	0.96	1.00
aromatic residues (15 HS/20 NS)	1.00	0.00	1.00	1.00	1.00	1.00

Considering different types of input data (protein−protein and protein−nucleic acid), we can therefore make a *decision function* dependent on each one of them and construct a new classifier encoding each type of data in different kernels following a multiple kernel learning (MKL) approach.[55] Thus, we are able to construct a new composite kernel which is a linear combination of base kernels derived from each type:

$$K_{ij} = \sum_{p}^{l=1} \lambda^{(l)} K_{ij}^{(l)} \qquad (16)$$

where $K_{ij}^{(l)}$ is the simple kernel derived from each type of data l, assuming p data sets. Thus, the kernel coefficients weight the importance of each kernel in the final solution. To the best of our knowledge this is first time that protein and DNA information is evaluated in that way considering a feature selection−multiple kernel learning (FSMKL) approach.[56,57] In our experiments, we achieved in test a 0.67 (AUROC) and 0.34 (FPR), so we obtained a great performance with a low cost regarding type I error. This particular classification method deals particularly well with these unbalanced data sets (in training: TPR = 0.92, FPR = 0.12, AUROC = 0.92). It was necessary to enhance the FSMKL capacities in order to choose the best features, according to a ranking of those most statistically aligned with the class, for solving a classification problem within a MKL framework in a cross validation process. We considered a large number of kernels, each one of them with a variable number of those statistically aligned features per kernel.

3. Web Server Implementation. The Web server (http://bio-aims.udc.es/MolStructPred.php) automates the analysis of the HS present at protein−protein or protein−nucleic acid systems in a user-friendly way. In the submission page, the user provides a protein structure (PDB format), defines the two monomers (by chain identifiers) and uploads a CONSURF file with the Conservation score for the residues at the system. The submitted job enters the server, and the SASA features are calculated for all residues. The output file gives the final classification (HS/NS) of all interfacial residues using the algorithms described in previous sections. All files are available for download. Our servers show an improved predictive power and efficiency, and are free and open to the scientific community.

DISCUSSION

The performance in ML is usually measured using predictive accuracy, which could be problematic if the data are unbalanced.[58] Table S1 comprises 71 HS/406 NS, Table S2 35 HS/56 NS, Table S3 60 HS/162 NS, and Table S4 20 HS/80 NS, which demonstrates that our data sets are highly unbalanced (classes are not equally represented as HS are less represented in nature). This way, we evaluated the performance of each model by taking into account recall (TPR), precision, specificity, and FPR as well as F1 score and AUROC. The F1 score gives us a balance between recall and precision rates. A ROC curve is plotted with true positive rates versus false positive rates for different classification thresholds. The AUROC can measure the classification's performance and is independent of any decision threshold. We aimed to compare type I errors of each model in order to avoid further misinterpretations of the results, and for that purpose we aimed to attain a high rate of correctly predicted examples in the minority class (HS) to avoid predictive errors. For example, in an unbalanced data set with 99% of the majority class examples, a simple ML approach would give a predictive accuracy of 99%, which clearly would not be appropriate. The F1 score is highly used for unbalanced data, and it should be higher than the % of HS at the data sets (14.9% at S1, 38.5% at S2, 27.0% at S3, and 20.0% at S4). As showed in our results, the F1 scores are much higher demonstrating their predictive power.

We showed that simple Bayes networks are able to classify HS for protein−protein interactions but only complex methods such as GA-SVM-Full can classify HS for protein−nucleic acid interactions. A Bayesian network[59] model is a type of acyclic graph (probabilistic/statistical graph model) where each node represents a set of random variables and each connection of the directed acyclic graph (DAG) represents conditional dependencies in the relationship between them. Each of the graph's nodes is associated with probability function.[60] For more information about Bayesian network classifiers in Weka please refer to[61] and for a general discussion about Bayes methods.[62] Despite the fact of its simplicity Bayesian classifiers performs very well in many complex real-world situations[63] and in this particular biological problem: protein−protein systems. The predictive power was tested in a large set of protein−protein mutations and gave F1 score = 0.82 and AUROC = 0.81 for the cross-validated S1 data set; F1 score = 0.68 and AUROC = 0.71 for an independent test set S2; and F1 score = 0.65 and AUROC = 0.60 for S3. We compared our proposed method with other ML models and an energy-based approach, displaying a clear higher predictive power.

For the protein−nucleic acid systems we had to use some more complex algorithms. One of the most commonly used ML technique due to its robustness is the SVM based on the Vapnik-Chervonenkis dimension.[64−66] The basic implementation (data are linearly separable) deals with a binary classification problem in which data are separated by an oriented hyperplane with a maximum margin of separation against classes. Thus, the hyperplane separates examples (positive and negative classes) in such a way that the distance between the boundary and the nearest data point (belonging to one class) is maximal. Those data points are known as support vectors[67] and are the points which have most influence in the final position/orientation of the hyperplane. For non-linearly separable data, the kernel trick arise in order to solve these kind of problems.[68,69] Using −1/+1 as labels for each class, the separating hyperplane is given as $w{\cdot}x + b = 0$ where w is the weight and b the bias, denoting the scalar product with $(\cdot)$. We therefore consider the *decision function* following $f(x_i) = sign(w{\cdot}x_i + b)$ being $(i = 1 \cdots n)$ the data points. We would like to maximize the margin, which is equivalent to

$$\min \frac{1}{2} \| w^2 \| \quad \text{subject to the constraints}$$

$$y_i(w{\cdot}x_i + b) \geq 1, \forall\, i \tag{17}$$

Training a SVM requires the solution of a quadratic programming (QP) optimization problem which is usually a very computational cost- and time-consuming constrained problem. Above formulation can be reduced by minimization to get a new formulation using Lagrange multipliers α_i with respect to w and b and maximized with respect to $\alpha_i \geq 0$. This is a convex QP[70]; therefore, making use of the Karush-Kuhn-Tucker conditions we can solve the problem getting the following *dual formulation*

$$W(\alpha) = \sum_{n}^{i=1} \alpha_i - \frac{1}{2} \sum_{n}^{i,j=1} \alpha_i \alpha_j y_i y_j (x_i{\cdot}x_j) \tag{18}$$

subject to the constraints $\alpha_i \geq 0$, $\sum_{i=1}^{n} \alpha_i y_i = 0$. Thus, the size of the problem depends on the number of samples, not in the number of input dimensions. This QP problem could be solved by the sequential minimal optimization (SMO) algorithm by decomposing the overall QP problem into QP sub-problems similar to Osuna's method choosing the smallest possible optimization problem at every step.[71] We found that the best approach for protein−nucleic case was GA with a SVM as a decision function. The statistical values were a bit lower, when assessed by type of residue, as these are particularly daunting systems due the large electrostatics involved at the interface.

Until now not a single feature has stand out as enough to produce an accurate method. However, in our approaches we have used two of the most crucial ones: genetic conservation and SASA. SASA relates to the fact that solvent occlusion is a critical characteristic as stated by the "O-ring theory", recently supported by a variety of different sources.[72−77] Our servers/methods available via Web interface were developed to provide a fast estimation of HS that contribute to the stability of protein−protein and protein−nucleic acid complexes. This estimation could be the basis for more detailed studies of the structural effects of punctual mutations by the use of all atom methods such as thermodynamical integration (TI), free energy perturbation (FEP), and molecular mechanics-generalized Born/Poisson−Boltzmann surface area (MM-GB/PBSA),[78−86] which allows deriving free binding energies from structural ensembles retrieved from MD simulations.

■ ASSOCIATED CONTENT

Ⓢ Supporting Information

Table S1, $\Delta\Delta G_{\text{binding}}$ experimental values obtained from essential the ASEdb database; Table S2, HS/NS classification retrieved from the Binding Interface Database (BID) (HS = hot-spot; NS = null-spot); Table S3, $\Delta\Delta G_{\text{binding}}$ experimental values obtained from the PINT and SKEMPI data sets; Table S4, $\Delta\Delta G_{\text{experimental}}$ values for protein/acid nucleic systems at the Pronit database; Table S5, HS/NS classification of other softwares for the complexes at the ASEdb database; Table S6, HS/NS classification for the complexes in the BID ; Table S7, HS/NS classification for the complexes PINT and SKEMPI data sets (for clarity in Tables S5, S6, and S7, only the HS were marked; all other residues were considered NS); and Table S8, statistical performance of various algorithms tested for data sets S1, S2, and S3 by amino acid group. This material is available free of charge via the Internet at http://pubs.acs.org.

■ AUTHOR INFORMATION

Corresponding Author

*E-mail: irina.moreira@cnc.uc.pt. Phone: (+351) 239 820 190. Fax: (+351) 239 822 776.

Funding

This work was funded by FEDER funds through the Operational Programme for Competitiveness Factors (COMPETE) and by National Funds through FCT (Foundation for Science and Technology) under the projects PEST-C/EQB/LA0006/2013 and COMPETE FCOMP-01-0124-FEDER-37285. This work was further supported by "Collaborative Project on Medical Informatics (CIMED)" PI13/00280, funded by the Carlos III Health Institute from the Spanish National plan for Scientific and Technical Research and Innovation 2013−2016, and the European Regional Development Funds (FEDER). The authors acknowledge the support by the Galician Network of Drugs R+D REGID (Xunta de Galicia R2014/025). I.S.M. was supported by FCT Ciência 2008 and FCT Investigator programmes - IF/00578/2014 (cofinanced by European Social Fund and Programa Operacional Potencial Humano).

Notes

The authors declare no competing financial interest.

■ REFERENCES

(1) Janin, J. Elusive Affinities. *Proteins: Struct. Funct. Genet.* **1995**, *21*, 30−39.

(2) Jones, S.; Thornton, J. M. Principles of Protein-Protein Interactions. *Proc. Natl. Acad. Sci. U.S.A.* **1996**, *93*, 13−20.

(3) Clackson, T.; Ultsch, M. H.; Wells, J. A.; de Vos, A. M. Structural and Functional Analysis of the 1:1 Growth Hormone: Receptor Complex Reveals the Molecular Basis for Receptor Affinity. *J. Mol. Biol.* **1998**, *277*, 1111−1128.

(4) DeLano, W. L.; Ultsch, M. H.; de Vos, A. M.; Wells, J. A. Convergent Solutions to Binding at a Protein-Protein Interface. *Science* **2000**, *287*, 1279−1283.

(5) Bogan, A. A.; Thorn, K. S. Anatomy of Hot Spots in Protein Interfaces. *J. Mol. Biol.* **1998**, *280*, 1−9.

(6) Martins, J. M.; Ramos, R. M.; Pimenta, A. C.; Moreira, I. S. Solvent-Accessible Surface Area: How Well Can Be Applied to Hot-Spot Detection? *Proteins* **2013**, *82*, 479−490.

(7) Zhu, X.; Mitchell, J. C. Kfc2: A Knowledge-Based Hot Spot Prediction Method Based on Interface Solvation, Atomic Density, and Plasticity Features. *Proteins* **2011**, *79*, 2671−83.

(8) Carl, N.; Hodošček, M.; Vehar, B.; Konc, J.; Brooks, B. R.; Janežič, D. Correlating Protein Hot Spot Surface Analysis Using Probis with Simulated Free Energies of Protein−Protein Interfacial Residues. *J. Chem. Inf. Model.* **2012**, *52*, 2541−2549.

(9) Massova, I.; Kollman, P. Combined Molecular Mechanical and Continuum Solvent Approach (Mm-Pbsa/Gbsa) to Predict Ligand Binding. *Perspect. Drug Disc. and Design* **2000**, *18*, 113−135.

(10) Tuncbag, N.; Keskin, O.; Gursoy, A. Hotpoint: Hot Spot Prediction Server for Protein Interfaces. *Nucleic Acids Res.* **2010**, *38*, W402−W406.

(11) Moreira, I. S.; Fernandes, P. A.; Ramos, M. J. Computational Alanine Scanning Mutagenesis—an Improved Methodological Approach. *J. Comput. Chem.* **2007**, *28*, 644−54.

(12) van Rossum, G. Python Tutorial. , Technical Report Cs-R9526; Centrum voor Wiskunde en Informatica (CWI), Amsterdam, 1995.

(13) Humphrey, W.; Dalke, A.; Schulten, K. Vmd: Visual Molecular Dynamics. *J. Mol. Graphics* **1996**, *14*, 33−38.

(14) Hall, M.; Frank, E.; Holmes, G.; Pfahringer, B.; Reutemann, P.; Witten, I. H. The Weka Data Mining Software: An Update. *SIGKDD Explor. Newsl.* **2009**, *11*, 10−18.

(15) Darnell, S. J.; LeGault, L.; Mitchell, J. C. Kfc Server: Interactive Forecasting of Protein Interaction Hot Spots. *Nucleic Acids Res.* **2008**, *36*, W265−W269.

(16) Darnell, S. J.; Page, D.; Mitchell, J. C. An Automated Decision-Tree Approach to Predicting Protein Interaction Hot Spots. *Proteins: Struct., Funct. Bioinf.* **2007**, *68*, 813−823.

(17) Zhu, X.; Mitchell, J. Kfc2: A Knowledge-Based Hot Spot Prediction Method Based on Interface Solvation, Atomic Density and Plasticity Features. *Proteins* **2011**, *79*, 2671−2683.

(18) Sievers, F.; Wilm, A.; Dineen, D.; Gibson, T. J.; Karplus, K.; Li, W.; Lopez, R.; McWilliam, H.; Remmert, M.; Söding, J.; Thompson, J. D.; Higgins, D. G. Fast, Scalable Generation of High-Quality Protein Multiple Sequence Alignments Using Clustal Omega. *Mol. Syst. Biol.* **2011**, *7*, 539.

(19) Thorn, K. S.; Bogan, A. A. Asedb: A Database of Alanine Mutations and Their Effects on the Free Energy of Binding in Protein Interactions. *Bioinformatics* **2001**, *17*, 284−285.

(20) Fischer, T. B.; Arunachalam, K. V.; 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.

(21) Moal, I. H.; Fernández-Recio, J. Skempi: A Structural Kinetic and Energetic Database of Mutant Protein Interactions and Its Use in Empirical Models. *Bioinformatics* **2012**, *28*, 2600−2607.

(22) Kumar, M. D. S.; Gromiha, M. M. Pint: Protein−Protein Interactions Thermodynamic Database. *Nucleic Acids Res.* **2006**, *34*, D195−D198.

(23) Kumar, M. D. S.; Bava, K. A.; Gromiha, M. M.; Prabakaran, P.; Kitajima, K.; Uedaira, H.; Sarai, A. Protherm and Pronit: Thermodynamic Databases for Proteins and Protein−Nucleic Acid Interactions. *Nucleic Acids Res.* **2006**, *34*, D204−D206.

(24) Prabakaran, P.; An, J.; Gromiha, M. M.; Selvaraj, S.; Uedaira, H.; Kono, H.; Sarai, A. Thermodynamic Database for Protein-Nucleic Acid Interactions (Pronit). *Bioinformatics* **2001**, *17*, 1027−1034.

(25) Sarai, A.; Gromiha, M. M.; An, J.; Prabakaran, P.; Selvaraj, S.; Kono, H.; Oobatake, M.; Uedaira, H. Thermodynamic Databases for Proteins and Protein-Nucleic Acid Interactions. *Biopolymers* **2001**, *61*, 121−126.

(26) Bernstein, F. C.; Koetzle, T. F.; Williams, G. J.; Meyer, E. F.; Brice, M. D.; Rodgers, J. R.; Kennard, O.; Shimanouchi, T.; Tasumi, M. The Protein Data Bank. A Computer-Based Archival File for Macromolecular Structures. *Eur. J. Biochem.* **1977**, *80*, 319−324.

(27) Eames, M.; Kortemme, T., Structural Mapping of Protein Interactions Reveals Differences in Evolutionary Pressures Correlated to Mrna Level and Protein Abundance. *Structure* 15, 1442−1451.

(28) Franzosa, E. A.; Xia, Y. Structural Determinants of Protein Evolution Are Context-Sensitive at the Residue Level. *Mol. Biol. Evol.* **2009**, *26*, 2387−2395.

(29) Ashkenazy, H.; Erez, E.; Martz, E.; Pupko, T.; Ben-Tal, N. Consurf 2010: Calculating Evolutionary Conservation in Sequence and Structure of Proteins and Nucleic Acids. *Nucleic Acids Res.* **2010**, *38*, W529−W533.

(30) Lee, B.; Richards, F. M. The Interpretation of Protein Structures: Estimation of Static Accessibility. *J. Mol. Biol.* **1971**, *55*, 379−IN4.

(31) Shrake, A.; Rupley, J. A. Environment and Exposure to Solvent of Protein Atoms. Lysozyme and Insulin. *J. Mol. Biol.* **1973**, *79*, 351−371.

(32) Miller, S.; Janin, J.; Lesk, A. M.; Chothia, C. Interior and Surface of Monomeric Proteins. *J. Mol. Biol.* **1987**, *196*, 641−656.

(33) Miller, S.; Lesk, A. M.; Janin, J.; Chothia, C. The Accessible Surface Area and Stability of Oligomeric Proteins. *Nature* **1987**, *328*, 834−836.

(34) Beck, D.; Foster, J. A. Machine Learning Techniques Accurately Classify Microbial Communities by Bacterial Vaginosis Characteristics. *PLoS One* **2014**, *9*, No. e87830.

(35) Fernandez-Lozano, C.; Fernandez-Blanco, E.; Dave, K.; Pedreira, N.; Gestal, M.; Dorado, J.; Munteanu, C. R. Improving Enzyme Regulatory Protein Classification by Means of Svm-Rfe Feature Selection. *Mol. BioSyst.* **2014**, *10*, 1063−1071.

(36) Fernandez-Lozano, C.; Gestal, M.; González-Díaz, H.; Dorado, J.; Pazos, A.; Munteanu, C. R. Markov Mean Properties for Cell Death-Related Protein Classification. *J. Theor. Biol.* **2014**, *349*, 12−21.

(37) Ugolotti, R.; Cagnoni, S. Differential Evolution Based Human Body Pose Estimation from Point Clouds. *Proceedings of the 15th Annual Conference on Genetic and Evolutionary Computation*; ACM: New York, 2013; pp 1389−1396

(38) Saeys, Y.; Inza, I.; Larrañaga, P. A Review of Feature Selection Techniques in Bioinformatics. *Bioinformatics* **2007**, *23*, 2507−2517.

(39) Ofran, Y.; Rost, B. Isis: Interaction Sites Identified from Sequence. *Bioinformatics* **2007**, *23*, e13−e16.

(40) Kim, D. E.; Chivian, D.; Baker, D. Protein Structure Prediction and Analysis Using the Robetta Server. *Nucleic Acids Res.* **2004**, *32*, W526−W531.

(41) Zhu, X.; Mitchell, J. C. Kfc2: A Knowledge-Based Hot Spot Prediction Method Based on Interface Solvation, Atomic Density, and Plasticity Features. *Proteins: Struct., Funct. Bioinf.* **2011**, *79*, 2671−2683.

(42) Tuncbag, N.; Gursoy, A.; Keskin, O. Identification of Computational Hot Spots in Protein Interfaces: Combining Solvent Accessibility and Inter-Residue Potentials Improves the Accuracy. *Bioinformatics* **2009**, *25*, 1513−1520.

(43) Bogan, A.; Thorn, K. Anatomy of Hot Spots in Protein Interfaces. *J. Mol. Biol.* **1998**, *280*, 1−9.

(44) Kortemme, T.; Baker, D. A Simple Physical Model for Binding Energy Hot Spots in Protein-Protein Complexes. *Proc. Natl. Acad. Sci. U.S.A.* **2002**, *99*, 14116−14121.

(45) Tuncbag, N.; Keskin, O.; Gursoy, A. Hotpoint: Hot Spot Prediction Server for Protein Interfaces. *Nucleic Acids Res.* **2010**, *38*, W402−W406.

(46) Ofran, Y.; Rost, B. Protein-Protein Interaction Hotspots Carved into Sequences. *PLoS Comput. Biol.* **2007**, *3*, No. e119.

(47) Holland, J. H. *Adaptation in Natural and Artificial Systems: An Introductory Analysis with Applications to Biology, Control, and Artificial Intelligence*; University of Michigan Press: Oxford, UK, 1975.

(48) Wall, M. *Galib: A C++ Library of Genetic Algorithm Components*, Version 2.4; Mechanical Engineering Department, Massachusetts Institute of Technology, 1996

(49) Kalyanmoy, D. Multi-Objective Genetic Algorithms: Problem Difficulties and Construction of Test Problems. *Evol. Comput.* **1999**, *7*, 205−230.

(50) Chang, C.-C.; Lin, C.-J. Libsvm: A Library for Support Vector Machines. *ACM Trans. Intell. Syst. Technol.* **2011**, *2*, 1−27.

(51) Clerc, M. Beyond Standard Particle Swarm Optimisation. *Int. J. Swarm Intell. Res. (IJSIR)* **2010**, *1*, 46−61.

(52) Zambrano-Bigiarini, M.; Clerc, M.; Rojas, R. Standard Particle Swarm Optimisation 2011 at Cec-2013: A Baseline for Future Pso Improvements, 2013 IEEE Congress on Evolutionary Computation (CEC), June , 20−23 2013; 2013; pp 2337−2344.

(53) Breiman, L. Random Forests. *Mach. Learn.* **2001**, *45*, 5−32.

(54) Ahmad, S.; Keskin, O.; Sarai, A.; Nussinov, R. Protein-DNA Interactions: Structural, Thermodynamic and Clustering Patterns of Conserved Residues in DNA-Binding Proteins. *Nucleic Acids Res.* **2008**, *36*, 5922−5932.

(55) Gönen, M.; Alpaydin, E. Multiple Kernel Learning Algorithms. *J. Mach. Learn. Res.* **2011**, *12*, 2211−2268.

(56) Seoane, J. A.; Day, I. N. M.; Gaunt, T. R.; Campbell, C. A Pathway-Based Data Integration Framework for Prediction of Disease Progression. *Bioinformatics* **2014**, *30*, 838−845.

(57) Fernandez-Lozano, C.; Seoane, J.; Gestal, M.; Gaunt, T.; Dorado, J.; Campbell, C. Texture Classification Using Feature Selection and Kernel-Based Techniques. *Soft Computing* **2015**, 1−12.

(58) Chawla, N. V.; Bowyer, K. W.; Hall, L. O.; Kegelmeyer, W. P. Smote: Synthetic Minority over-Sampling Technique. *J. Artif. Int. Res.* **2002**, *16*, 321−357.

(59) Pearl, J. *Probabilistic Reasoning in Intelligent Systems: Networks of Plausible Inference*; Morgan Kaufmann Publishers Inc., 1988; p 552.

(60) Christofides, N. *Graph Theory: An Algorithmic Approach (Computer Science and Applied Mathematics)*; Academic Press, Inc., 1975.

(61) Bouchkaert, R. R. *Bayesian Network Classifiers in Weka*, 2007.

(62) Zhang, H. Exploring Conditions for the Optimality of Naïve Bayes. *Int. J. Pattern Recognit. Artif. Intell.* **2005**, *19*, 183−198.

(63) Hand, D. J.; Yu, K. Idiot's Bayes? Not So Stupid after All? *Int. Stat. Rev.* **2001**, *69*, 385−398.

(64) Cortes, C.; Vapnik, V. Support-Vector Networks. *Machine Learning* **1995**, 273−297.

(65) Vapnik, V. N. *The Nature of Statistical Learning Theory*; Springer-Verlag: New York, 1995; p 188.

(66) Vapnik, V. N. *Estimation of Dependences Based on Empirical Data* [in Russian]; Nauka, 1979; English translation Springer Verlang, 1982.

(67) Burges, C. J. C. A Tutorial on Support Vector Machines for Pattern Recognition. *Data Mining Knowledge Disc.* **1998**, *2*, 121−167.

(68) Shawe-Taylor, J.; Cristianini, N. *Kernel Methods for Pattern Analysis*; Cambridge University Press, 2004.

(69) Cristianini, N.; Shawe-Taylor, J. *An Introduction to Support Vector Machines: And Other Kernel-Based Learning Methods*; Cambridge University Press, 2000; p 189.

(70) Alpaydin, E. *Introduction to Machine Learning*; The MIT Press: 2010; p 584.

(71) Platt, J. C., Fast Training of Support Vector Machines Using Sequential Minimal Optimization. In *Advances in Kernel Methods*; MIT Press: 1999; pp 185−208.

(72) Li, J.; Liu, Q. 'Double Water Exclusion': A Hypothesis Refining the O-Ring Theory for the Hot Spots at Protein Interfaces. *Bioinformatics* **2009**, *25*, 743−750.

(73) Ramos, R. M.; Fernandes, L. F.; Moreira, I. S. Extending the Applicability of the O-Ring Theory to Protein−DNA Complexes. *Comput. Biol. Chem.* **2013**, *44*, 31−39.

(74) Moreira, I. S.; Fernandes, P. A.; Ramos, M. J. Computational Determination of the Relative Free Energy of Binding—Application to Alanine Scanning Mutagenesis. In *Molecular Materials with Specific Interactions—Modeling and Design*; Sokalski, W. A., Ed.; Springer: Amsterdam, 2007; Vol. *4*, pp 305−339.

(75) Martins, J. M.; Ramos, R. M.; Pimenta, A. C.; Moreira, I. S. Solvent-Accessible Surface Area: How Well Can Be Applied to Hot-Spot Detection? *Proteins: Struct., Funct. Bioinf.* **2014**, *82*, 479−490.

(76) Moreira, I. S.; Ramos, R. M.; Martins, J. M.; Fernandes, P. A.; Ramos, M. J. Are Hot-Spots Occluded from Water? *J. Biomol. Struct. Dyn.* **2013**, *32*, 186−197.

(77) Xia, J.; Zhao, X.; Song, J.; Huang, D. Apis: Accurate Prediction of Hot Spots in Protein Interfaces by Combining Protrusion Index with Solvent Accessibility. *BMC Bioinformatics* **2010**, *11*, 174.

(78) Huo, S.; Massova, I.; Kollman, P. A. Computational Alanine Scanning of the 1:1 Human Growth Hormone−Receptor Complex. *J. Comput. Chem.* **2002**, *23*, 15−27.

(79) Kollman, P. A.; Massova, I.; Reyes, C.; Kuhn, B.; Huo, S.; Chong, L.; Lee, M.; Lee, T.; Duan, Y.; Wang, W.; Donini, O.; Cieplak, P.; Srinivasan, J.; Case, D. A.; Cheatham, T. E. Calculating Structures and Free Energies of Complex Molecules: Combining Molecular Mechanics and Continuum Models. *Acc. Chem. Res.* **2000**, *33*, 889−897.

(80) Massova, I.; Kollman, P. A. Computational Alanine Scanning to Probe Protein−Protein Interactions: A Novel Approach to Evaluate Binding Free Energies. *J. Am. Chem. Soc.* **1999**, *121*, 8133−8143.

(81) Ramos, R. M.; Moreira, I. S. Computational Alanine Scanning Mutagenesis—an Improved Methodological Approach for Protein−DNA Complexes. *J. Chem. Theory Comput.* **2013**, *9*, 4243−4256.

(82) Martins, S. A.; Perez, M. A. S.; Moreira, I. S.; Sousa, S. F.; Ramos, M. J.; Fernandes, P. A. Computational Alanine Scanning Mutagenesis: Mm-Pbsa Vs Ti. *J. Chem. Theory Comput.* **2013**, *9*, 1311−1319.

(83) Moreira, I. S.; Martins, J. M.; Ramos, R. M.; Fernandes, P. A.; Ramos, M. J. Understanding the Importance of the Aromatic Amino-Acid Residues as Hot-Spots. *Biochim. Biophys. Acta (BBA)—Proteins Proteomics* **2013**, *1834*, 404−414.

(84) Moreira, I. S.; Fernandes, P. A.; Ramos, M. J. Computational Alanine Scanning Mutagenesis—an Improved Methodological Approach. *J. Comput. Chem.* **2007**, *28*, 644−654.

(85) Moreira, I.; Fernandes, P.; Ramos, M. Unravelling Hot Spots: A Comprehensive Computational Mutagenesis Study. *Theor. Chem. Acc.* **2007**, *117*, 99−113.

(86) Lafont, V.; Schaefer, M.; Stote, R. H.; Altschuh, D.; Dejaegere, A. Protein−Protein Recognition and Interaction Hot Spots in an Antigen−Antibody Complex: Free Energy Decomposition Identifies "Efficient Amino Acids. *Proteins: Struct., Funct. Bioinf.* **2007**, *67*, 418−434.

DOI: 10.1021/ci500760m