PROTEINS
STRUCTURE ■ FUNCTION ■ BIOINFORMATICS

Prediction of protein-glucose binding sites using support vector machines

Houssam Nassif,[1] Hassan Al-Ali,[2] Sawsan Khuri,[3,4] and Walid Keirouz[5]*

[1] Department of Computer Sciences, University of Wisconsin-Madison, Madison, Wisconsin
[2] Department of Biochemistry and Molecular Biology, University of Miami Miller School of Medicine, Miami, Florida
[3] Center for Computational Science, University of Miami Miller School of Medicine, Miami, Florida
[4] Dr. John T. Macdonald Foundation Department of Human Genetics, University of Miami Miller School of Medicine, Florida
[5] Department of Computer Science, American University of Beirut, Beirut, Lebanon

ABSTRACT

Glucose is a simple sugar that plays an essential role in many basic metabolic and signaling pathways. Many proteins have binding sites that are highly specific to glucose. The exponential increase of genomic data has revealed the identity of many proteins that seem to be central to biological processes, but whose exact functions are unknown. Many of these proteins seem to be associated with disease processes. Being able to predict glucose-specific binding sites in these proteins will greatly enhance our ability to annotate protein function and may significantly contribute to drug design. We hereby present the first glucose-binding site classifier algorithm. We consider the sugar-binding pocket as a spherical spatio-chemical environment and represent it as a vector of geometric and chemical features. We then perform Random Forests feature selection to identify key features and analyze them using support vector machines classification. Our work shows that glucose binding sites can be modeled effectively using a limited number of basic chemical and residue features. Using a leave-one-out cross-validation method, our classifier achieves a 8.11% error, a 89.66% sensitivity and a 93.33% specificity over our dataset. From a biochemical perspective, our results support the relevance of ordered water molecules and ions in determining glucose specificity. They also reveal the importance of carboxylate residues in glucose binding and the high concentration of negatively charged atoms in direct contact with the bound glucose molecule.

Proteins 2009; 77:121–132.

Key words: hexose; carbohydrate; protein-carbohydrate interaction; substrate recognition; binding site signature; feature vector; SVM; random forests.

INTRODUCTION

Hexoses are 6-carbon sugar molecules that play a key role in several different biochemical pathways, including cellular energy release, signaling, carbohydrate synthesis, and the regulation of gene expression.[1] Glucose and galactose are the two most commonly found hexoses in nature. Proteins that bind these sugars are implicated in several human diseases, including diabetes, various metabolic disorders, and Huntington disease. The biochemical and molecular pathways for these disease mechanisms have not all been elucidated and much work remains to be done.

In parallel, genome sequencing of a wide range of species has yielded sequence knowledge of a large number of proteins whose biochemical functions are still unknown. The three-dimensional structures of many of these proteins were elucidated. Some of these proteins have been shown to be members of certain pathways, but they lack sufficient sequence or structural similarity to any other protein with a known function.

The functional annotation of these "unknown" proteins is of paramount importance. One approach to tackle this problem is to predict what these proteins may bind to. Prediction of glucose-specific binding sites will significantly improve our understanding of protein structure-function relationships and enable us to assign possible functions to some of the many genomic proteins whose function remains unknown (e.g., Patra and Mandal[2]). This, in turn, will allow us to better understand disease mechanisms that may involve some of these proteins and be better placed for either diagnosis or treatment of these diseases.

Proteins that bind hexoses belong to diverse functional families that lack significant sequence or, often, structural similarity.[3] Despite this dissimilarity in the binding site architecture between protein families, these proteins show high specificity to their hexose

Additional Supporting Information may be found in the online version of this article.
Supported by the American University of Beirut (H.N.).
*Correspondence to: Walid Keirouz, Department of Computer Science, American University of Beirut, P.O. Box 11-0236, Riad El-Solh, Beirut 1107 2020, Lebanon. E-mail: walid@aub.edu.lb
Received 2 October 2008; Revised 6 February 2009; Accepted 9 March 2009
Published online 19 March 2009 in Wiley InterScience (www.interscience.wiley.com).
DOI: 10.1002/prot.22424

ligands. Assuming that common recognition principles exist for the identification of common substrates,[4] these hexose binding sites have unique distinguishing biochemical and spatial features. The amino acids constituting a binding site determine its topology and biochemical properties and may be selected, at least in part, for mediating intermolecular interactions, perhaps at the expense of protein stability.[5] Based on this argument, the spatial and biochemical properties of a binding site should enable us to predict the ligand type and, therefore, to speculate on the function of this ligand-binding protein domain. This article addresses the issue of identifying glucose binding sites by building a program that correctly classifies a protein pocket into glucose binding or non-glucose binding.

Researchers have investigated protein-sugar binding sites for several years. From the biochemical perspective, Rao *et al.*[6] fully characterized the architecture of sugar binding in lectins and identified conserved loop structures within the protein as essential for sugar recognition. Later, Quiocho and Vyas[7] presented a review of the biochemical characteristics of carbohydrate-binding sites and identified the planar polar residues (Asn, Asp, Gln, Glu, Arg) as the most frequently involved residues in hydrogen bonding. They found that the aromatic residues Trp, Tyr, Phe, and His, stack against the apolar surface of the sugar pyranose ring. Quiocho and Vyas also pinpointed the role of ordered water molecules and metal ions in determining substrate specificity and affinity. Taroni *et al.*[8] analyzed the characteristic properties of sugar binding sites and described a residue propensity parameter that best discriminates sugar-binding sites from other protein-surface patches. Simple sugars typically have a polar-hydrophilic end which establishes hydrogen bonds and a hydrophobic end which is responsible for the pyranose ring stacking. Sugar binding sites are thus neither hydrophobic nor hydrophilic, due to the dual nature of sugar docking.[8] In fact, as Garcia-Hernandez *et al.*[9] showed, some polar groups in the protein-carbohydrate complex behave hydrophobically. Furthermore, Zhang *et al.*[10] reported that the hydrogen bonds between the hexose ligand and certain amino acids in galactosyltransferases are crucial for the orientation of the ligand and the correct function of the protein.

Some of this biochemical information has been used in computational work with the objective of accurately predicting protein sugar-binding sites. Shionyu-Mitsuyama *et al.*[11] were some of the first to use atom type densities within binding sites to develop an algorithm for predicting carbohydrate-binding. Sujatha and Balaji[4] formulated a signature for characterizing galactose-binding sites based on solvent accessibility and secondary structure types. They implemented a three-dimensional structure searching algorithm, COTRAN, to identify galactose-binding sites. Chakrabarti *et al.*[12] modeled one glucose-binding site and one galactose-binding site by optimizing their binding affinity under geometric and folding free energy constraints. More recently, Malik and Ahmad[13] used a neural network to predict general carbohydrate and specific galactose binding sites.

On a broader scale, Gold and Jackson[14] compiled the SitesBase database of precalculated protein-ligand binding site similarities. Given a ligand-binding site, SitesBase returns all database entries with similar binding sites, ranked by a similarity score. They did this by performing an all-against-all geometric hashing over the Protein Data Bank[15] (PDB). Although the primary use of this database is to examine structural similarities between related binding sites, it can also provide evidence of functional similarity for unclassified binding sites.

None of the previous work specifically targeted glucose-protein interactions. Some targeted a specific protein family,[6,10,12] while others focused on galactose[4,13] or on general hexoses.[7,9] Attempts to develop computational classifiers for carbohydrate binding sites[8,11,13] led to moderate results. Structural comparisons[14] give similarity scores, but require additional analysis to determine a site's functionality.

With this in mind, this work builds and trains a support vector machines (SVM) classifier to predict glucose binding sites. SVMs are used successfully in a wide area of biological domains[16]; they are straightforward and computationally inexpensive. We define a protein binding site as a collection of atoms within a macromolecule with a central docking-pocket and a geometrically defined chemical neighborhood, as performed by Bobadilla *et al.*[17] In this simplified model, charge, polarity, mobility, and hydrophobicity can all be considered as determinant features of a binding site. We thus follow Bagley and Altman's stipulation that "by temporarily abandoning a view of sites as groups of amino acids, and instead concentrating on the chemical milieu in important locations, we may gain insight into the critical factors that define a site[18]".

Our classifier may be a useful filter to identify potential binding sites that can then be verified using molecular dynamics simulations or biochemical experiments. It can also act as a discriminating layer on top of structural similarity search engines, like Gold and Jackson,[14] for glucose binding sites annotation.

MATERIALS AND METHODS

In this work, we extract multiple chemical, amino-acid and spatial features from the binding-site and concatenate them as a feature vector. We then analyze the features and select a relevant subset. Finally, we train the SVM classifier.

Binding site representation

We represent the binding site as a sphere centered at the ligand, as portrayed in Figure 1. The sphere is subdi-

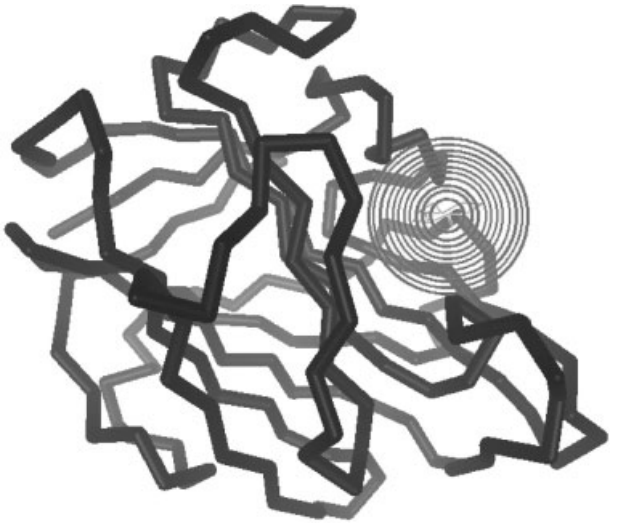

Figure 1

Importance of charge features according to RF. CNEUT stands for neutral charge feature, NEG for negative charge feature and L# for the layer number.

vided into concentric shells as suggested by Bagley and Altman.[18] The center of the glucose-binding site is the center of the glucose pyranose ring and is computed as the centroid of the coordinates of the ring's 6 atoms (C1, C2, C3, C4, C5, O5). For sites that do not bind glucose, the center is taken as the center of the cavity or the respective ligand's center point.

The pyranose ring radius is 1.5 Å, the farthest glucose atom (O6) is 3.5 Å away and the molecular interactions are significant to a range of 7 Å as suggested by Bobadilla *et al.*[17] Therefore, the radius of the sphere is fixed at 10 Å. The first layer width is fixed to 3 Å while the width of the subsequent seven layers is 1 Å each.

Preparation of the dataset

The SVM classifier needs to be trained with both glucose-binding (positives) and nonglucose binding (negatives) sites. We mine the PDB[15] for protein-glucose complexes. We consider the heterogen (HET) group name of all glucose forms and derivatives and use the following HETs: GLC for D-glucose, AGC for α-D-glucopyranose, and BGC for β-D-glucopyranose. We remove theoretical structures and redundancies, as well as files older than PDB format 2.1. Using PISCES,[19] we impose a 30% overall sequence identity as a cut-off. We examine the remaining structures at close range using the Swiss-PDBViewer program.[20] To remain true to the specificity of a glucose binding site, we discard several proteins at this point due to the proximity of other ligands in the binding pocket or the fact that the same binding pocket can bind to multiple ligands. The final outcome is a non-

Table I

Inventory of Positive Training and Cross-Validation Glucose-Binding Sites

PDB ID	Glucose	PDB ID	Glucose	PDB ID	Glucose
1BDG	GLC-501	1I8A	GLC-189	1PWB	GLC-405
1EX1	GLC-617	1ISY	GLC-1461	1Q33	GLC-400
1GJW	GLC-701	1ISY	GLC-1471	1S5M	AGC-1001
1GWW	GLC-1371	1J0Y	GLC-1601	1UA4	GLC-1457
1H5U	GLC-998	1JG9	GLC-2000	1V2B	AGC-1203
1HIZ	GLC-1381	1K1W	GLC-653	1WOQA	GLC-290
1HIZ	GLC-1382	1KME	GLC-501	2BQP	GLC-337
1HKC	GLC-915	1MMUA	GLC-1	2BVW	GLC-602
1HSJ	GLC-671	1NF5	GLC-125	2BVW	GLC-603
1HSJ	GLC-672	1NSZ	GLC-1400		

The cavity center is computed as the centroid of the glucose's pyranose ring.

redundant data set of 43 protein-glucose binding sites. We use 29 for training and cross validation (Table I) and the remaining 14 for testing (Table II).

The negative dataset consists of three groups of sites that do not bind glucose. The first group of 36 sites, labeled "nonsugar binding", bind ligands other than hexoses. We consider the ligand's centroid to be the center of the binding site. The second group of 15 sites, labeled "sugar-binding", bind nonglucose hexoses and other sugars, namely galactose, mannose, fructose, sucrose, and glucose-derivatives. The cavity center is computed in a manner similar to glucose positives (see Fig. 1), using the pyranose or furanose ring's centroid. Finally, the third group of 17 sites, labeled "nonbinding", includes sites that are not known to bind any ligand. We used two thirds of the dataset for training and cross-validation (Table III), and kept a third for testing (Table IV).

Physio-chemical properties used as descriptors

Charge, hydrogen bonding and hydrophobicity are the three properties that define most chemical bonding. We therefore use them to determine glucose recognition and binding. These chemical properties are each assigned nominal values: the measure of charge per atom is positive, neutral, or negative; atoms are either able to form hydrogen bonds or not; hydrophobicity measures are considered as hydrophobic, hydroneutral, or hydrophilic (Table V). Table V shows a detailed listing of each atom

Table II

Inventory of Positive Testing Glucose-Binding Sites

PDB ID	Glucose	PDB ID	Glucose	PDB ID	Glucose
1S5M	AGC-1001	2E2O	GLC-400	2IPL	BGC-501
1SZ2	BGC-1001	2F2E	AGC-401	2O9T	GLC-500
1SZ2	BGC-2001	2FH6	GLC-1097	2PWF	BGC-9998
1U2S	GLC-1	2FVY	GLC-307	3F9M	GLC-500
1Z8D	GLC-901	2H3H	BGC-1500		

The cavity center is computed as the centroid of the glucose's pyranose ring.

Table III
Inventory of Negative Training and Cross-Validation Sites That do Not Bind Glucose

Non-sugar binding sites			Sugar binding sites	
PDB ID	Cavity center	Ligand	PDB ID	Ligand
11GS	1672–1675	MES-3	1AXZ	GLA-401
1A42	2054–2055	BZO-555	1BQP	MAN-402
1A50	4939–4940	FIP-270	1D1W	GAL-1400
1A53	2016–2017	IGP-300	1DZQ	GAL-502
1AA1	4472–4474	3PG-477	1EUU	GAL-2
1AJN	6074–6079	AAN-1	1KLF	MAN-1500
1AJS	3276–3281	PLA-415	1LBY	F6P-295
1AL8	2652	FMN-360	1TJ4	SUC-1
1BOB	2566	ACO-400	2GAM	NGA-502
1D09	7246	PAL-1311	4PBG	BGP-469
1DY1	1423	ZN-401		
1EQY	3831	ATP-380	Nonbinding sites	
1F8I	13237	MG-451	PDB ID	Cavity center
1FI2	1493	MN-202	11AS	5132
1IOL	2674–2675	EST-400	1A7W	351
1J1L	2246	FE2-1001	1BSI	103–114
1JTV	2136–2137	TES-500	1C3P	1089–1576
1KF6	16674–16675	OAA-702	1C5K	605–871
1NX8	6104–6109–6110	N7P-290	1DXJ	867–1498
1TVO	2857	FRZ-1001	1EVT	2149–2229
1UK6	2142	PPI-1300	1FSZ	2048–2190
2BIW	15,171	FE-1492	1KLM	4113–4373
2PAH	5318	FE-453	1KWP	1212
3PCB	3421–3424	3HB-550	2BG9	1237

The cavity center is computed as the centroid of the given atoms or ring. The atoms are listed by their PDB serial number.

and its features. For example, the first entry in the table, the amide peptide linkage oxygen of an amino acid, is denoted as O. It does not bear a partial atomic charge, has a hydrophilic tendency, and is capable of forming a hydrogen bond with a ligand. Several atoms or molecules present in the protein cavity are not part of the protein structure, but may have an effect on ligand specificity,[7] such as water, sulfate, phosphate, calcium, magnesium, and zinc. We include these atoms and molecules in the features list presented in Table V. With regard to hydrogen atoms, we use them only in the context of the heavy atom they are attached to such as nitrogen amide or carboxyl oxygen.

In addition to chemical features, we incorporate residue features in our algorithm. Amino acids are generally categorized into subgroups, based on the structural and chemical properties of their side chains.[21,22] Accordingly, we classify the amino acids into the following subgroups: aromatic (Phe, Trp, Tyr), aliphatic (Ala, Ile, Leu, Met, Val), acidic-carboxylate (Asp, Glu), basic (Arg, Lys), neutral (Asn, Cys, Gln, Gly, Pro, Ser, Thr), and histidine (His). We classify histidine as its own group because it is often found in locations which are part buried and part exposed, and it can have roles which are unique among the aromatic hydrophobic amino acids, such as in metal binding.[23]

Feature extraction

The algorithm obtains the spatial distribution of the different atoms in a spherical region of a certain radius and divides them into concentric layers. For each layer, the program samples all the atoms and residues contained in this layer, then samples all the properties of each atom and creates a feature vector for every layer. As such, each feature has a measure in each concentric layer.

The binding site feature vector is the concatenation of the feature vectors of the layers. To put equal initial weight on the different features, the data is standardized by scaling and centering.[24]

Feature selection

Feature representation is an important step in designing classifiers. Some feature combinations can effectively partition the input space, while others are completely irrelevant. Knowing that an increase in the feature dimensionality tends to increase measurement cost and decrease classification accuracy,[25] good designers aim at minimizing the number of features while maximizing the classification performance.

Random Forests[26] (RF) is a classification algorithm based on multiple classification trees which provides measures of feature importance and is used as a feature selection tool. The RF feature selection method is robust to noise, can be used when the number of features is much greater than the number of observations, incorporates feature interactions and returns a direct feature importance measure based on information gain.[26] The RF feature importance score measures the decrease of classification accuracy when values of a feature are randomly permuted.[27] The higher the score, the more important the feature is. Random Forests matches or outperforms

Table IV
Inventory of Negative Testing Sites That do Not Bind Glucose

Nonsugar binding sites			Sugar binding sites	
PDB ID	Cavity center	Ligand	PDB ID	Ligand
1A0J	6985	BEN-246	1ISZ	GAL-471
1ATG	1751	WO4-250	1KWK	GAL-701
1B8A	7224	ATP-500	1KZB	MAN-1501
1BO5	7811	GOL-601	1LIU	FBP-580
1RTK	3784–3787	GBS-300	1Y9G	FRU-801
1W8N	4573–4585	DAN-1649	Nonbinding sites	
1ZYU	1284–1286	SKM-401	PDB ID	Cavity center
2C9Q	777	CU-1103	1AM2	1277
2D7S	3787	GLU-1008	1QZ7	2509–3592
2GP4	12	MSE-1	1ZT9	1056–1188
2GSH	6260	MG-402	2DN2	749–1006
2GSV	1126	SO4-102	2GRK	369–380
			2GSE	337–10,618

The cavity center is computed as the centroid of the given atoms or ring. The atoms are listed by their PDB serial number.

Table V
Chemical Features

Atom type	Functional group	Residue	PDB atom symbol	Charge	Hydrophobicity	Hydrogen bonding
Oxygen	Amide peptide linkage	All amino acids	O	0	HPHIL	HB
Oxygen	Carboxyl C- terminus	All amino acids	OXT	−ve	HPHIL	HB
Oxygen	Carboxyl	GLU, ASP	OE1, OE2, OD1, OD2	−ve	HPHIL	HB
Oxygen	Amide	GLN, ASN	OE1, OD1	0	HPHIL	HB
Oxygen	Hydroxyl	SER, THR, TYR	OG, OG1, OH	0	HPHIL	HB
Nitrogen	Amide peptide linkage	All amino acids except PRO	N	0	HPHIL	HB
Nitrogen	Amide peptide linkage	PRO	N	0	HPHIL	NHB
Nitrogen	Amide	GLN, ASN	NE2, ND2	0	HPHIL	HB
Nitrogen	Amide	LYS	NZ	+ve	HPHIL	HB
Nitrogen	Guanidino	ARG	NE	+ve	HPHIL	NHB
Nitrogen	Guanidino	ARG	NH1, NH2	+ve	HPHIL	HB
Nitrogen	Imidazole	HIS	ND1, NE2	0	HPHIL	HB
Nitrogen	Indole	TRP	NE1	0	HNEUT	NHB
Carbon	Amide peptide linkage	All amino acids	C	0	HNEUT	NHB
Carbon	C-alpha	All amino acids	CA	0	HNEUT	NHB
Carbon	Aliphatic chain (neutral)	ALA, SER, THR, CYS, ASP, ASN, GLU, GLN, ARG, LYS, PRO	CB, CG, CD, CE	0	HNEUT	NHB
Carbon	Aliphatic chain (hydrophobic)	LEU, VAL, ILE, MET	CB, CG, CD, CE	0	HPHOB	NHB
Carbon	Aliphatic branch	LEU, VAL, ILE	CG1, CG2, CD1, CD2, CD1	0	HPHOB	NHB
Carbon	Aromatic	PHE, TYR, TRP	CG, CD1, CD2, CE1, CE2, CZ, CG,CD1, CD2, CE2, CE3, CZ2, CZ3, CH2	0	HPHOB	NHB
Carbon	Imidazole	HIS	CG, CD2, CE1	0	HPHOB	NHB
Sulfur	Sulfhydril	CYS	SG	0	HPHIL	HB
Sulfur	Thioether	MET	SD	0	HNEUT	NHB
Oxygen	Sulfate	SO4	O1, O2, O3, O4	−ve	HPHIL	HB
Oxygen	Phosphate	2HP	O1, O2, O3, O4	−ve	HPHIL	HB
Oxygen	Water	HOH	O	0	HPHIL	HB
Calcium	Ion	CA	CA	+ve	HPHIL	HB
Magnesium	Ion	MG	MG	+ve	HPHIL	HB
Zinc	Ion	ZN	ZN	+ve	HPHIL	HB

HPHIL, hydrophilic; HPHOB, hydrophobic; HNEUT, hydroneutral; HB, Hydrogen bonding; NHB, nonhydrogen bonding.

other feature selection approaches, such as F-ratio, Wilcoxon statistic, and Shrunken Centroids.[28]

The use of Random Forests as a feature selection tool is ideal in our case because the number of features we are monitoring is greater than the number of examples. Furthermore, its use for feature selection, coupled with SVM for classification, outperforms SVM alone.[29] We perform feature selection in the R statistical computing environment[30] with the varSelRF package.[28]

Classification method

Support vector machines[31] is a parametric statistical linear classifier that performs a nonlinear mapping of the input space to a new (potentially higher dimensional) feature space to which a linear machine can be applied. SVM constructs a hyperplane separating the positive examples from the negative ones in the new space representation. To avoid overfitting, SVM chooses the Optimal Separating Hyperplane that maximizes the margin in feature space.[32] The margin is defined as the minimal distance between the hyperplane and the training examples. The selected data points that support the hyperplane are called support vectors. A smaller number of support vectors reflects a better generalization.[33] SVMs achieve good performance when applied to real problems.[16,34]

We use the R LIBSVM implementation[35,36] and opted for the nonlinear soft margin implementation together with the RBF kernel as suggested by Hsu *et al.*[37] The SVM gamma and cost parameters are tuned independently for each run. They are incremented exponentially: gamma ranges from 2^{-14} to 2^{2} and cost from 2^{-4} to 2^{14}.

Assessment of predictions

We estimate the classifier performance by a leave-one-out cross-validation technique, also known as jackknife cross-validation, performed over the datasets in Tables I and III. A classifier is designed using (n-1) samples and evaluated on the one remaining sample. This process is repeated n times, once per sample. Leave-one-out cross-validation derives an unbiased error estimate[38] and is a method of choice when the data is scarce: all the data is used, in turn, for training and testing. Model selection and parameter tuning is performed independently in each fold to avoid selection bias. Leave-one-out has been

Table VI
Testing the Importance of Water and Ions to Glucose Binding

Properties	Error over dataset (%)	Error over subset of dataset (%)
Include water and ions	18.92	7.81
Discard water	18.92	10.94
Discard ions	20.27	7.81
Discard water and ions	20.27	12.5

successfully integrated with SVMs and kernel classifiers.[39,40]

In addition to cross-validation we use a hold-out independent testing set. The testing set (Tables II and IV) has a train-to-test ratio of 2:1. The hold-out method has a pessimistically biased estimate since different partitionings will give different estimates, but has a lower variance and time complexity than leave-one-out.[41]

We use the percentage of misclassified samples as an estimate of the generalization error rate.[41] We also report our algorithm's sensitivity (ratio of true positives over the sum of true positives and false negatives) and specificity (ratio of true negatives over the sum of true negatives and false positives). We use the R IPRED package[42] to perform error estimation.

RESULTS

Inclusion of water and ions

Before running the main classifier, we test the need to include water and ion features in our algorithm. Ordered water molecules and ions, such as sulfate, phosphate, calcium, magnesium, and zinc are known to have an effect on ligand specificity of carbohydrate binding sites.[7] We perform leave-one-out SVM cross-validation experiments while discarding water molecules, ions or both (Table VI). We also show the results of a preliminary experiment performed over a subset of our dataset which lacks the sugar-binding sites negatives of Table II. We run these experiments while including charge, hydrogen bonding, hydrophobicity and residue properties. The results clearly show that the inclusion of water and ions together in the computation yields similar or better results than the exclusion of either one (see Table VI).

Physio-chemical descriptors

We perform glucose-binding site recognition over the training and cross-validation dataset. We use each physio-chemical descriptor alone and compare the SVM error with and without the RF feature selection. We plot the most relevant features of each property. Table VII shows the SVM error, sensitivity and specificity of each experiment. It also reports the number of support vector instances as a percentage of the total number of instances.

Charge

Charge performs poorly as a discriminating property on its own, with an error of 24.32%. However, after feature selection, it returns a much better result of 14.86% improving both sensitivity and specificity. The five selected features reflect a prominence of negatively charged groups in the glucose binding site (see Fig. 2). Negatively charged atoms are common in layer 3 and subsequent layers, they seemed to be a distinguishing characteristic of glucose binding sites. The negatively charged atoms of layer 3, which are in direct contact with the glucose molecule, constitute the most discriminating feature. In contrast, layers 1 and 2 overlap with the glucose molecule's own space. Atomic presence in layer 1, which is 3 Å wide, generates a large steric hindrance and specifies nonbinding sites. A high atom concentration at layer 2 specifies small moiety nonsugar binding negatives. Since charge-neutral atoms are much more abundant in a protein than charged atoms, RF selects charge-neutral as features for layers 1 and 2.

Hydrogen bonding

Applying feature selection to hydrogen bond alone leads to a small drop in error rate from 17.57 to 14.86%, while the support vector percentage increases from 41.89

Table VII
Comparison of SVM's Cross-Validation Performance on Chemical and Residue Properties With and Without RF Feature Selection

Property	RF	Number of features	SVM error (%)	Sensitivity (%)	Specificity (%)	Support vectors (%)
Charge	False	24	24.32	79.31	73.33	77.03
	True	5	14.86	86.21	84.44	44.59
H-Bond	False	16	17.57	82.76	82.22	41.89
	True	3	14.86	82.76	86.67	47.30
Hydro	False	24	16.22	72.41	91.11	67.57
	True	15	12.16	82.76	91.11	40.54
Residues	False	48	21.62	48.28	97.78	100.00
	True	19	9.46	93.10	88.89	41.89
Charge + H-Bond + Hydro + Residue	False	112	18.92	75.86	84.44	79.73
	True	24	8.11	89.66	93.33	40.54

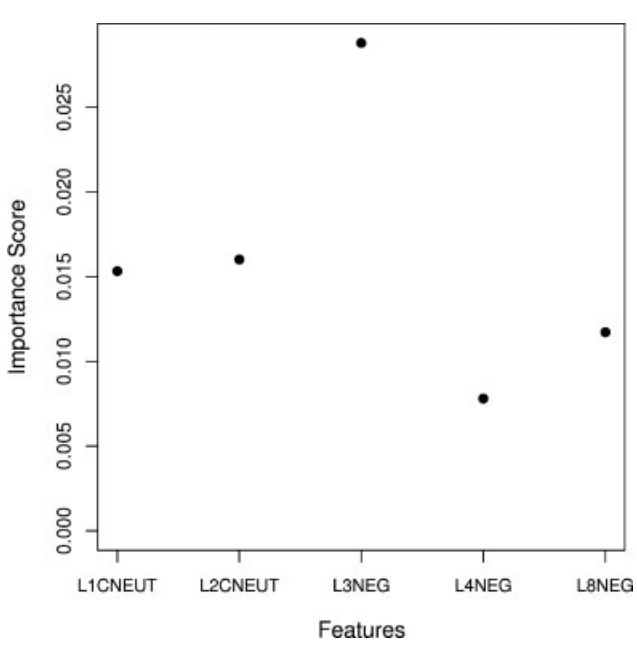

Figure 2

Importance of hydrogen bond features according to RF. HB stands for hydrogen-bonding feature and L# for the layer number.

to 47.30%. Since a lower percentage of support vectors reflects better generalization, the increase offsets the performance gain due to feature selection. Figure 3 shows a plot of the selected features. The positive and negative data sets differ in the distribution of their hydrogen bonding atoms. Glucose-binding and sugar-binding sites have more hydrogen-bonding atoms at layer 3 than sites that are not sugar-binding. A high atom concentration at layer 1 indicates a large steric hindrance and specifies nonbinding sites. We revisit the issue of hydrogen bonds later in this article.

Hydrophobicity

Hydrophobicity is the best discriminating chemical property with an error rate of 16.22% and a 91.11% specificity. Feature selection lowers the error rate to 12.16% while improving sensitivity to 82.76% and lowering the percentage of support vectors to 40.54%. Figure 4 plots the seven most relevant hydrophobicity features. The subsequent eight features have RF importance scores close to zero. Most hydrophobicity features have a low importance score, which reflects a small information gain. Although most relevant features are hydrophilic, we notice a hydrophobic feature at layer 7.

Residue

Before feature selection, the residue classifier has a 100% support vectors percentage (see Table VII). This classifier memorizes the data, but is not able to generalize. Feature selection cuts the support vectors to 41.89%,

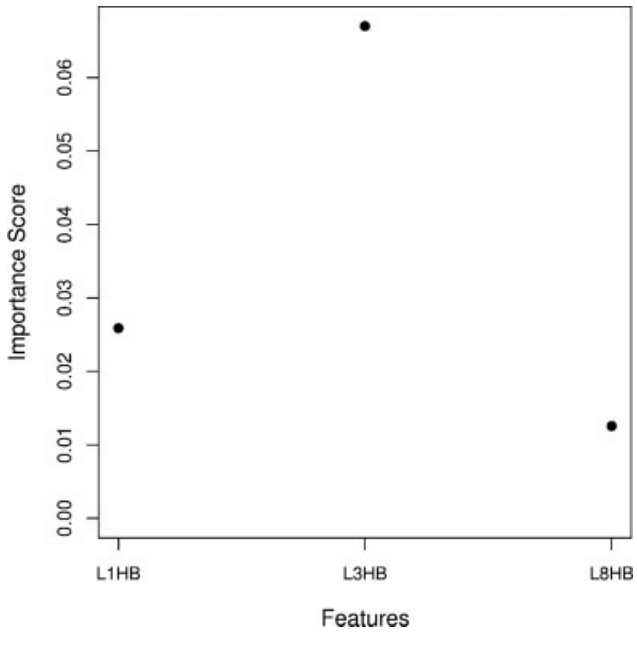

Figure 3

The seven highest hydrophobicity feature importance measures as returned by RF. HPHIL stands for hydrophilic feature, HNEUT for hydroneutral feature, HPHOB for hydrophobic feature, and L# for the layer number.

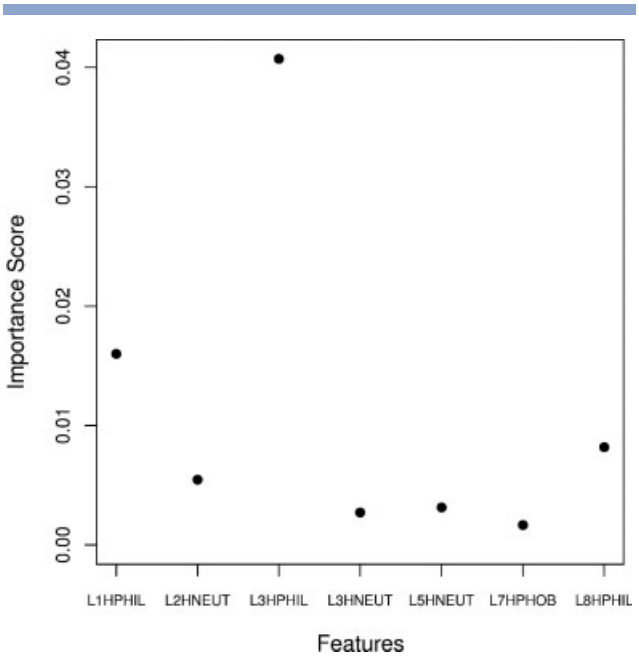

Figure 4

The eight highest residue feature importance measures as returned by RF. ALIPH stands for aliphatic residue feature, CAR for carboxylate residue feature, RNEUT for neutral residue feature, ARO for aromatic residue feature and L# for the layer number.

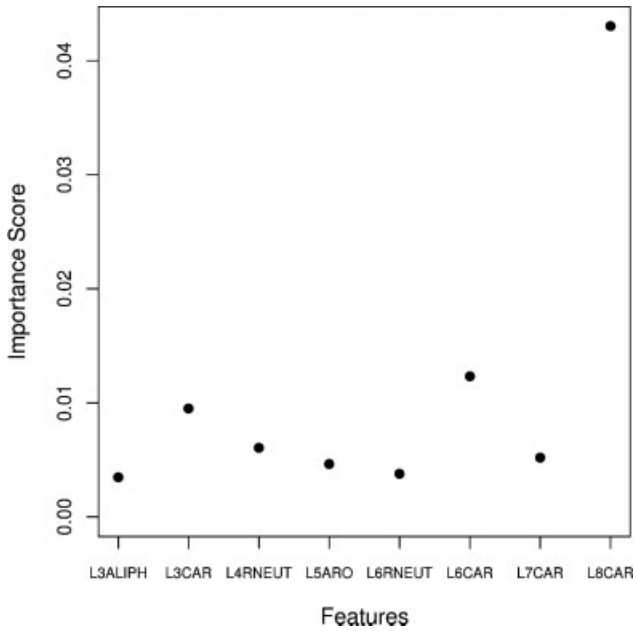

Figure 5

The seven highest atomic and residue feature importance measure as returned by RF. CNEUT stands for neutral charge, NEG for negative charge, HB for hydrogen-bonding, HPHIL for hydrophilic, CAR for carboxylate-bearing residues and L# for the layer number.

while achieving a 9.46% error rate. Sensitivity (93.10%) and specificity (88.89%) are also high, with a slight bias toward detecting true positives. Figure 5 plots the eight most relevant residue features. The subsequent 11 features have RF importance scores close to zero. The high information gain of carboxylate-bearing acidic residues suggests a high propensity of the negatively charged glutamate (Glu) and aspartate (Asp) amino acids to occur in the glucose binding sites. The most relevant residue features also include aromatic residues, known to play a role in glucose docking.

Combining chemical and residue features

The results of combining chemical and residue features confirm our previous findings. This combination yields the best discriminating feature subset. With 24 selected features, it drops the error rate to 8.11% and the support vectors percentage to 40.54%, the lowest levels in our analysis (see Table VII). It achieves a high specificity rate of 93.33% and a slightly lower sensitivity rate of 89.66%. Although the exclusive use of residue features improves the sensitivity to 93.10%, the specificity, error rate and support vectors percentage are worse than the corresponding values for combinations of chemical and residue features. Table VIII lists the 24 selected features. Figure 6 plots the seven most relevant features. As seen from these results, the layer eight acidic carboxylate residue feature ranks as the most important in the identification of a glucose binding site, closely followed by layer three hydrogen bonding and hydrophilic chemical features. All these features characterize glucose binding sites. The next clustering of RF importance scores are the layers 1 and 2 neutral charge features, characterizing nonsugar binding sites, and layer 3 negative charge and layer 6 acidic carboxylate residue features. To improve readability, Figure 6 does not show the remaining features of Table VIII, which follow at a lower RF importance score level.

Validation over the testing set

To validate the classifier of Table VIII, we test it using the separate and independent testing set (see Tables II and IV). Our model misclassifies two positive and two negative entries. We thereby achieve a 10.81% error rate, a 85.71% sensitivity and a 91.30% specificity. These results parallel the cross-validation estimates, despite a slightly lower performance. The hold-out testing method is known to produce pessimistically biased estimates.[41]

Table VIII

The Feature Combination Achieving the Best Results

Property	Features	L1	L2	L3	L4	L5	L6	L7	L8
Charge	Negative			X				X	X
	Neutral	X	X						
	Positive								
H-Bond	Non H-Bonding	X							
	H-Bonding	X		X					X
Hydrophobicity	Hydrophilic	X		X					X
	Neutral		X	X					
	Hydrophobic					X		X	
Residues	Aromatic [Phe, Tyr, Trp]								
	Aliphatic [Ala, Ile, Leu, Met, Val]								
	Neutral [Asn, Cys, Gln, Gly, Pro, Ser, Thr]				X	X		X	
	Acidic (Carboxylate) [Asp, Glu]			X		X	X	X	X
	Basic [Arg, Lys]								
	Histidine [His]								

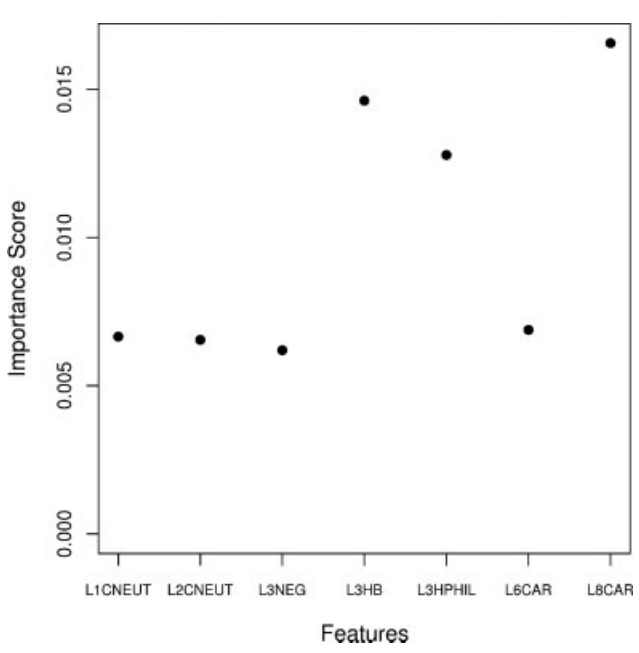

Figure 6

The 7 highest atomic and residue feature importance measure as returned by RF. CNEUT stands for neutral charge, NEG for negative charge, HB for hydrogen-bonding, HPHIL for hydrophilic, CAR for carboxylate-bearing residues and L# for the layer number.

DISCUSSION

Chemical properties

We first establish that ordered water molecules and ions that appear in the crystallized x-ray PDB structure are important features to add to our algorithm, thereby confirming the findings of Quiocho and Vyas.[7] We provide further evidence that prosthetic groups or crystallographic water molecules are a physical extension of the binding pocket, being as involved in coordinating the bound ligand as amino acids are.

Our next results show that, of the chemical properties, hydrophobicity outperforms both charge and hydrogen bonding in its importance in defining a glucose binding site. Charge has the same error as hydrogen bond (14.86%), but a smaller support vectors percentage. This is surprising since hydrogen bond would have been expected to be more important for glucose docking than charge and hydrophobicity.

Most protein-ligand binding requires the establishment of hydrogen bonds between the protein and the ligand, which means that atoms and residues capable of establishing hydrogen bonds abound in most binding-site grooves. Hence the hydrogen bond property may not be able to discriminate between our positive glucose-binding data set and our negative nonglucose-binding data set. To test this hypothesis, we build a classifier that exclusively uses, as a negative data set, domains that do not bind any ligand (such as surface residues, transmembrane helices, and others). These nonbinding sites are not biased with respect to hydrogen bonding atoms and such a classifier should be able to discriminate primarily according to the hydrogen bonding property. We use the same 29 positive entries from Table I and a negative set of 28 sites that do not bind any ligand (see Supporting Information). We perform the runs on each of the chemical properties, namely charge, hydrogen bond and hydrophobicity (Table IX).

As expected, the hydrogen bond feature outperforms charge and hydrophobicity for SVM classification error and for the support vector percentage. A comparison of the Tables VII and IX results reveals a great improvement in classification accuracy, simply by using a negative data set that is less similar in function and shape to the positive data set. Both the hydrogen bond high discrimination capacity and the sharp drop in the classification error confirm our suggestion. Hydrogen bonding is a key property in glucose-binding sites, but it is not a good discrimination criteria vis-a-vis other binding sites.

Feature selection findings

Feature selection pinpoints the difference in spatial configuration between positive and negative binding-site data sets. Negative nonbinding sites and small groove nonsugar binding-sites have a higher atomic presence in layers 1 and 2, both of which overlap with the bound glucose molecule's own space. Positive binding sites tend to have most of their discriminating chemical features in layer 3, in direct contact with the bound glucose.

Chemically, charge feature selection reveals the relative negativity of layers 3 and above (see Fig. 2). This finding is reflected by the high propensity of the negatively charged carboxylate-bearing residues in layers 3 and above (see Fig. 5). Both carboxylate-bearing residues, glutamate (Glu) and aspartate (Asp) were identified by Taroni *et al.*[8] as having a high sugar interface propensity level.

Feature selection over the combined chemical and residue properties shows that hydrogen bonding feature in layer 3 has one of the highest RF scores (see Fig. 6). In their review of glucose binding sites, Quiocho and Vyas[7] identify the planar polar residues (Arg, Asn, Asp, Gln, Glu) as involved in a network of hydrogen bonds with the docked glucose. We have already established the prominence of Asp and Glu. Asn and Gln are both part

Table IX

SVM Trained Using an Exclusively Nonbinding Sites Negative Set

Property	SVM error (%)	Support vectors (%)
Charge	5.26	73.68
Hydrogen bond	3.51	61.40
Hydrophobicity	5.26	68.42

Table X
Cross-Validation Entries Misclassified by Final Model

Dataset subgroup	Dataset size	Misclassified entries
Glucose binding sites	29	1H5U, 1HSJ (GLC-672), 1ISY (GLC-1471)
Non-sugar binding sites	24	1A53
Non-glucose sugar binding sites	10	1AXZ, 1BQP
Non-binding sites	11	None

of our "neutral" amino acid subgroup, a group that is relevant to glucose-binding classification (see Fig. 5 and Table VIII).

As Taroni *et al.*[8] report, the sugar binding sites are neither hydrophobic nor hydrophilic. Hexoses exhibit a dual hydrophobic-hydrophilic nature and both antagonist properties are involved in hexose docking. Many glucose binding sites have both a hydrophilic region which establishes hydrogen bonds and a hydrophobic region which is responsible for the pyranose ring stacking over aromatic residues. This fact may explain why both hydrophobic and hydrophilic chemical features have high RF scores in hydrophobicity feature selection (see Fig. 4).

A constant feature of carbohydrate docking is their pyranose ring docking against an aromatic residue.[7,13,43] In fact, the COTRAN galactose binding site predictor[4] is based on identifying exposed aromatic residues. Residue feature selection shows that the aromatic feature plays an important role in discrimination (see Fig. 5). Its presence in layer 5 helps discriminate glucose from nonglucose binding sites. However, this feature was not selected in the final model of Table VIII. Unlike other sugars, glucose stacks over an aromatic residue in most, but not all, glucose binding sites.[43] Although Sujatha *et al.*[43] found positional and energetic differences between glucose and galactose docking, they concluded that aromatic residues may not play a significant role in distinguishing glucose from galactose binding. Our classifier, too, does not rely on this feature.

Analysis of final model

Using feature selection and leave-one-out cross-validation, we select a subset of 24 features (Table VIII) that achieve a low 8.11% error. Our model misclassifies three positive glucose binding sites and three negative nonglucose binding sites (Table X). It correctly rejects all nonbinding sites, the easiest negative subgroup.

One false positive is a nonsugar binding site, 1A53, which binds indole-3-glycerol phosphate. This compound is ~9 Å, not much bigger than glucose, and this protein is a member of the β-barrel protein family which includes many glucose binding proteins (cf. SCOP database[44]). It is therefore not unlikely that our algorithm might misclassify it. The other two false positives are nonglucose sugar binding sites, a negative subgroup which we populated with structures that were highly similar to glucose-binding sites, as a more challenging test to the algorithm. In fact, 1AXZ is a galactose binding site, while 1BQP binds mannose. Galactose, mannose and glucose binding sites are very similar. Nevertheless, our model correctly rejected eight nonglucose sugar binding sites, namely five galactose and mannose binding sites, and all fructose, sucrose and glucose-derivative sites that were tested.

A closer look at the false negatives explains their misclassification. 1H5U contains a pyridoxal-5′-phosphate (PLP) molecule in the binding pocket as a coligand with the glucose, thus it is not strictly speaking binding only glucose. This highlights an idea for future expansion of our algorithm, namely to add chemical features for possible cofactors. In contrast, 1HSJ is crystallized with glucose, but the protein in vivo actually binds maltose, a disaccharide composed of two glucose units. The binding site is thus larger and topologically different from single glucose binding sites. Similarly, 1ISY in vivo binds xylan, even though here it has been crystallized with a glucose. Thus, even though glucose was used in the crystallization of the above two structures, they are not in essence glucose-specific binding proteins. In that sense, our algorithm demonstrated some ability to distinguish between true glucose-binding sites and "artificial" glucose binding sites resulting from crystallization conditions.

When applied to the hold-out testing set, the final model of Table VIII returns two false positives (1W8N, 1KZB) and two false negatives (1Z8D, 2O9T). 1W8N was used as a nonsugar binding site which binds 2-deoxy-2,3-dehydro-*N*-acetyl-neuraminic acid. However, this protein also binds galactose, albeit at a different binding site, and close inspection of its structure reveals that its two binding sites are not dissimilar, which would explain its misclassification. 1KZB binds mannose, a hexose very similar to glucose, as explained above.

As for the false negatives, both are unusual structures. 1Z8D is the structure of a glycogen phosphorylase in an intermediate state, and the binding pocket is closed in tight on the glucose molecule. In addition, 1Z8D has a modified lysine inside the binding pocket (LLP680) which would alter its conformation. Although this is not in contact with the glucose, it could affect the recognition of the binding site by our algorithm.

2O9T is an exo-β-glucosidase that attaches to the end of a large carbohydrate complex and nicks off the glucose molecules one by one. The crystal for that x-ray structure was generated by transferring a seed from another crystal into crystallization solution saturated with cellotetraose, and after 15 min was flash-frozen to 120 K°.[45] Glucose is thus present in the crystal structure as a product of substrate hydrolysis, and it might be that the crystallized form depicts the structure of the complex halfway between tight product binding and product release. This is supported by the observation that when crystallized

Table XI
Comparison of Carbohydrate-Binding Site Predictor Programs

Program	Error (%)	Sensitivity (%)	Specificity (%)	Dataset
Glucose binding site predictor	8.11	89.66	93.33	Leave-one-out method over 29 glucose binding and 35 nonglucose binding sites
COTRAN[4]	5.09	88.68	95.91	Overall performance over sixfolds, totaling 106 galactose-binding sites and 660 nongalactose binding sites
Malik and Ahmad[13]	29.00	63.00	79.00	Leave-one-out method over 18 galactose binding sites and 116 noncarbohydrate ligand-binding sites
Shionyu-Mitsuyama et al.[11]	31.00	Not reported	Not reported	Test set of 61 polysaccharide binding sites
Taroni et al.[8]	35.00	Not reported	Not reported	Test set of 40 carbohydrate binding sites

Not Meant as a Direct Comparison Since the Data Sets Used in These Programs Were Different.

with a covalently-bound glucose-like inhibitor (2F-DNPG), which in theory resembles a transition state intermediate, the glucose moeity forms slightly different contacts with the binding pocket.

Given these encouraging results using an algorithm that resolves the binding site into spherical layers, we expect that a second-generation algorithm that takes the analysis one step further by including the angular orientation of the sugars would perform even better. Shape-based approaches are frequently used in drug design,[46] and extrapolations from those techniques could be made for applications in glucose binding site predictions.

Comparison with other programs

This work develops the first glucose-binding site predictor program, which means that comparisons with other programs have to be based on different data sets. We present a list of carbohydrate binding site predictor programs and the data sets they used (Table XI). We add the performance scores simply to give the reader an idea of the accuracy of each program within its own data set. This list includes the COTRAN galactose binding site identifier of Sujatha and Balaji,[4] the galactose-binding site identifier of Malik and Ahmad,[13] and the two general carbohydrate binding site predictors of Taroni *et al.*[8] and Shionyu-Mitsuyama *et al.*[11] The last two predictors are not based on a binary classification scheme and only output predicted carbohydrate-binding sites without offering sensitivity and specificity values. Note that COTRAN is validated on a highly skewed data where the 660 negatives outnumber the 160 positives. This discrepancy may explain the high specificity and the low error score. From these published results, our SVM glucose classifier reports a low error rate and the highest sensitivity score based on our own cross-validated dataset of glucose binding and nonbinding protein domains.

CONCLUSIONS

This article reports on the first classifier program that aims to predict glucose binding sites. It demonstrates that SVMs can make an important contribution to this field, especially if coupled with Random Forests feature selection. Given the center of a protein surface groove, our system uses SVM to correctly detect a glucose binding site 89.66% of the time and correctly reject a nonglucose binding site 93.33% of the time. We use Random Forests to determine different chemical and residue properties that characterize a glucose binding site and showed that glucose binding sites can be modeled using a limited number of basic features. An SVM classification using a small subset of the initial features gives an error as low as 8.11%. Our results support the relevance of ordered water molecules and ions in determining glucose-binding specificity and highlight the importance of carboxylate residues in glucose binding. Finally we note a high concentration of negatively charged atoms in direct contact with the bound glucose.

A direct application of this work is the prediction of potential glucose binding sites. Our system can form the core of a glucose binding-site predictor package; such a program identifies protein groove centers and may feed the center coordinates to our program for functional prediction. Such a predictive system can even parse the whole PDB to predict and annotate potential glucose binding sites. The utility of this tool for functional predictions and eventually for serving genome annotation is potentially great.

REFERENCES

1. Solomon E, Berg L, Martin DW. Biology, 8th ed. Florence, KY: Brooks Cole; 2007.
2. Patra M, Mandal C. Search for glucose/galactose-binding proteins in newly discovered protein sequences using molecular techniques and structural analysis. Glycobiology 2006;16:959–968.
3. Khuri S, Bakker FT, Dunwell JM. Phylogeny, function and evolution of the Cupins, a structurally conserved, functionally diverse superfamily of proteins. Mol Biol Evol 2001;18:593–605.
4. Sujatha MS, Balaji PV. Identification of common structural features of binding sites in galactose-specific proteins. Protein Struct Funct Bioinf 2004;55:44–65.
5. Jaramillo A, Wernisch L, Héry S, Wodak SJ. Folding free energy function selects native-like protein sequences in the core but not on the surface. Proc Natl Acad Sci USA 2002;99:13554–13559.

6. Rao VSR, Lam K, Qasba PK. Architecture of the sugar binding sites in carbohydrate binding proteins—a computer modeling study. Int J Biol Macromol 1998;23:295–307.
7. Quiocho FA, Vyas NK. Atomic interactions between proteins/enzymes and carbohydrates. In: Hecht SM, editor. Bioorganic chemistry: carbohydrates. New York: Oxford University Press; 1999. pp 441–457.
8. Taroni C, Jones S, Thornton JM. Analysis and prediction of carbohydrate binding sites. Protein Eng 2000;13:89–98.
9. García-Hernández E, Zubillaga RA, Chavelas-Adame EA, Vázquez-Contreras E, Rojo-Domínguez A, Costas M. Structural energetics of protein-carbohydrate interactions: insights derived from the study of lysozyme binding to its natural saccharide inhibitors. Protein Sci 2003;12:135–142.
10. Zhang Y, Swaminathan GJ, Deshpande A, Boix E, Natesh R, Xie Z, Acharya KR, Brew K. Roles of individual enzyme-substrate interactions by alpha-1,3-galactosyltransferase in catalysis and specificity. Biochemistry 2003;42:13512–13521.
11. Shionyu-Mitsuyama C, Shirai T, Ishida H, Yamane T. An empirical approach for structure-based prediction of carbohydrate-binding sites on proteins. Protein Eng 2003;16:467–478.
12. Chakrabarti R, Klibanov AM, Friesner RA. Computational prediction of native protein ligand-binding and enzyme active site sequences. Proc Natl Acad Sci USA 2005;102:10153–10158.
13. Malik A, Ahmad S. Sequence and structural features of carbohydrate binding in proteins and assessment of predictability using a neural network. BMC Struct Biol 2007;7:1.
14. Gold ND, Jackson RM. Fold independent structural comparisons of protein-ligand binding sites for exploring functional relationships. J Mol Biol 2006;355:1112–1124.
15. 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.
16. Noble WS. Support vector machine applications in computational biology. In: Schoelkopf B, Tsuda K, Vert JP, editors. Kernel methods in computational biology. Cambridge, MA: MIT Press; 2004. pp 71–92.
17. Bobadilla L, Nino F, Narasimhan G. Predicting and characterizing metal-binding sites using support vector machines. Proc ICBA 2004;8:307–318.
18. Bagley SC, Altman RB. Characterizing the microenvironment surrounding protein sites. Protein Sci 1995;4:622–635.
19. Wang G, Dunbrack RL. PISCES: a protein sequence culling server. Bioinformatics 2003;19:1589–1591.
20. Guex N, Peitsch MC. SWISS-MODEL and the Swiss-PdbViewer: an environment for comparative protein modeling. Electrophoresis 1997;18:2714–2723.
21. Betts MJ, Russell RB. Amino acid properties and consequences of substitutions. In: Barnes MR, Gray IC, editors. Bioinformatics for Geneticists. UK: Wiley; 2003. pp 289–316.
22. Sullivan SA, Landsman D. Characterization of sequence variability in nucleosome core histone folds. Protein Struct Funct Genet 2003;52:454–465.
23. Leitgeb S, Nidetzky B. Structural and functional comparison of 2-His-1-carboxylate and 3-His metallocentres in non-haem iron(II)-dependent enzymes. Biochem Soc Trans 2008;36:1180–1186.
24. Duda RO, Hart PE, Stork DG. Pattern classification, 2nd ed. New York: Wiley-Interscience; 2001.
25. Jain AK, Chandrasekaran B. Dimensionality and sample size considerations in pattern recognition practice. In: Krishnaiah PR, Kanal LN, editors. Handbook of statistics 2. Amsterdam: North-Holland; 1982. pp 835–855.
26. Breiman L. Random forests. J Mach Learn Res 2001;45:5–32.
27. Breiman L, Cutler A. Random forests; 2005. Available at http://www.stat.berkeley.edu/users/breiman/RandomForests/cc_home.htm. Accessed on 2009-04-15.
28. Díaz-Uriarte R, de Andrés SA. Gene selection and classification of microarray data using random forest. BMC Bioinf 2006;7:3.
29. Chen YW, Lin CJ. Combining SVMs with various feature selection strategies. In: Guyon IM, Gunn SR, Nikravesh M, Zadeh L, editors. Feature extraction, foundations and applications. Berlin, Germany: Springer; 2006.
30. R Development Core Team. R: a language and environment for statistical computing. Vienna, Austria: R Foundation for Statistical Computing; 2005. Available at http://www.R-project.org.
31. Vapnik VN. Statistical learning theory. New York: Wiley; 1998.
32. Baldi P, Brunak S. Bioinformatics, the machine learning approach, 2nd ed. Cambridge, Massachusetts: MIT Press; 2001.
33. Cristianini N, Shawe-Taylor J. An introduction to support vector machines. UK: Cambridge University Press; 2002.
34. Scholkopf B. SVMs-a practical consequence of learning theory. IEEE Intell Syst 1998;13:18–21.
35. Chang CC, Lin CJ. LIBSVM: a library for support vector machines; 2001. Available at http://www.csie.ntu.edu.tw/∼cjlin/libsvm.
36. Meyer D. Support vector machines. R News 2001;1:23–26.
37. Hsu CW, Chang CC, Lin CJ. A practical guide to support vector classification, Tech. Report. National Taiwan University, 2003. http://www.csie.ntu.edu.tw/∼cjlin/papers/guide/guide.pdf.
38. Luntz A, Brailovsky V. On estimation of characters obtained in statistical procedure of recognition. Techicheskaya Kibernetica 1969;3 (in Russian).
39. Cawley GC, Talbot NLC. Fast exact leave-one-out cross-validation of sparse least-squares support vector machines. Neural Networks 2004;17:1467–1475.
40. Cawley GC, Talbot NLC. Efficient approximate leave-one-out cross-validation for kernel logistic regression. Mach Learn 2008;71:243–264.
41. Jain AK, Duin RPW, Mao J. Statistical pattern recognition: a review. IEEE Trans Pattern Anal 2000;22:4–37.
42. Peters A, Hothorn T. IPRED: improved predictors; 2008. Available at http://cran.r-project.org/web/packages/ipred/index.html.
43. Sujatha MS, Sasidhar YU, Balaji PV. Energetics of galactose- and glucose-aromatic amino acid interactions: implications for binding in galactose-specific proteins. Protein Sci 2004;13:2502–2514.
44. Andreeva A, Howorth D, Chandonia JM, Brenner SE, Hubbard TJ, Chothia C, Murzin AG. Data growth and its impact on the SCOP database: new developments. Nucleic Acids Res 2008;36 (Database issue):D419–D425.
45. Isorna P, Polaina J, Latorre-Garcia L, Canada FJ, Gonzalez B, Sanz-Aparicio J. Crystal structures of *Paenibacillus polymyxa* beta-glucosidase B complexes reveal the molecular basis of substrate specificity and give new insights into the catalytic machinery of family I glycosidases. J Mol Biol 2007;371:1204–1218.
46. Weisel M, Proschak E, Kriegl JM, Schneider G. Form follows function: shape analysis of protein cavities for receptor-based drug design. Proteomics 2009;9:451–459.