BIOINFORMATION
Discovery at the interface of physical and biological sciences

open access www.bioinformation.net Hypothesis

Volume 8(24)

Prediction of protein-mannose binding sites using random forest

Harshvardan Khare[1], Vivek Ratnaparkhi[1], Sonali Chavan[1] & Valadi Jayraman[2]*

[1]Bioinformatics centre, University of Pune, Pune, India; [2]Centre for Development of Advanced Computing (C-DAC), Pune, India; Valadi Jayraman – Email: vkjayaram@yahoo.com; Phone: +91-20–25704228; Fax: +91-20–25694004; *Corresponding author

Received November 16, 2012; Accepted November 19, 2012; Published December 08, 2012

Abstract:
Mannose is an abundant cell surface monosaccharide and has an important role in many biochemical processes. It binds to a great diversity of receptor proteins. In this study we have employed Random Forest for prediction of mannose binding sites. Mannose-binding site is taken to be a sphere around the centroid of the ligand and the sphere is subdivided into different layers and atom wise and residue wise features were extracted for each layer. The method achieves 95.59 % of accuracy using Random Forest with 10 fold cross validation. Prediction of mannose binding site analysis will be quite useful in drug design.

Keywords: Binding site prediction, Carbohydrate binding site prediction, Mannose binding site prediction, Machine learning, Random Forest.

Background:
There is an exponential increase in genome sequence and protein structure data in last few years. Comparatively less availability of experimental assays of carbohydrate binding and discoveries of essential roles of some of the protein-carbohydrate interactions in various metabolic processes suggests the necessity for prediction algorithms. It is known that carbohydrate-binding proteins share low sequence and structural similarity **[1].** Despite this low similarity in sequence as well as structure, their binding sites are very specific. This specificity can be attributed to the local characteristics of binding sites such as hydrogen bonding patterns, presence of stacking interactions **[2].** Another proof for presence of local characteristic features comes from biochemical studies of sugar binding in lectins by Rao et al **[3].** They found conserved loop structures to be important for sugar binding. Hydrophobic stacking interactions have also been found to be specific for carbohydrate binding **[2].** Such features constitute a multidimensional feature space. Prediction of mannose binding site employing Random Forest is carried out under the assumption that from such a space enough informative features can be extracted and employed for supervised classification of binding and nonbinding sites.

Mannose binding proteins cover a diverse range of functions. They can be broadly classified into two classes, viz. 1) those having N or O-glycosylation bonds with sugars and 2) those exhibiting non-covalent interactions with sugars. In this work only the proteins with non-covalent interactions are considered. In the literature there exist a few studies pertaining to prediction of carbohydrate binding sites. Shionyu Mitsuyama *et al* **[4]** first derive empirical rules based on the similarity of spatial distribution of amino acid residues in known binding protein structure and subsequently employ the derived rules for identification of positive sites. Taroni *et al* **[5]** used amino acid propensity at carbohydrate binding sites. Sujatha and Balaji **[6]** developed a COTRAN algorithm to identify Galactose binding sites. Malik and Ahmad **[7]** employed neural network to predict carbohydrate-binding sites. Nassif *et al* **[8]** used different types of atomic and residue features to predict glucose-binding sites.

Methodology:
For the purpose of extracting different features we need to provide a rational method of representing the binding sites of different structures **[8].** As in the earlier work binding site has been represented as concentric spherical shells around its centre

[8] and this centre is taken as the centroid of the atoms excluding hydrogen atoms present in the ligand. This approach was used by Nassif *et al* **[8]** for extracting different shell features. The shells are started from a distance of 3 Angstroms from the shell centre and continued up to 10 Angstroms from the shell centre, each shell with a width of 1 Angstrom. The radius of outermost shell is chosen to be 10 Angstrom unit. The radius of mannose pyranose ring is 1.5 Angstrom unit. The molecular interactions are significant to a range of 7 Angstrom unit **[9].** Therefore the radius of outermost sphere is kept at 10 Angstrom unit.

Preparation of Dataset

A non-redundant list of 11 proteins, which bind to mannose non-covalently, where structures of protein-mannose complexes are known, was taken from PDB. This dataset of 11 proteins is kept as the final dataset. The positive dataset consists of 55 mannose-binding sites derived from 11 mannose-binding proteins. True binding sites for non-mannose ligands have been included as the negative data; Non-mannose ligands include non-mannose hexoses, non-sugar organic molecules and metal ions. These comprise of 78 Glucose binding sites, 40 Galactose binding sites and 69 other ligand-binding sites. All these 187 binding sites along with 55 positive sites form our input data for the experiments.

Extraction of Shell Features

Separate features

15 such features **Table 1 (see supplementary material)** were extracted for each shell. These consist of both atomic features and residue wise features. The first 14 features are the same as those employed by Nassif *et al* **[8].** The first eight features are based on the number of atoms of a particular type. Features 1, 2 and 3 define charge. Features 4, 5 and 6 define hydrophobic character. Features 7 and 8 define the ability of forming hydrogen bonds. Features 9 to 14 are based on the number of atoms of a particular type of residue. The usefulness of such features and their relevance to prediction of glucose binding sites has been discussed in detail by Nassif *et al* **[8].** In addition to these features accessible surface area has also been included in our experiments.

Combined features

These features comprise of different combinations of independent features **Table 2 (see supplementary material).** First eight features are used for generating combinations. These eight features fall in three categories viz. charge, hydrophobicity and hydrogen bonding property. Charge has three possible values; hydrophobicity has three possible values while hydrogen bonding property has two possible values. These are combined in all possible ways to obtain 18 combinations. Out of these 18 combinations only 7 combinations are physically possible. These seven features are calculated for each shell. The idea is that specific combinations of independent features can have better discriminative capabilities. The seven combinations employed in our work are shown in **Table 2.**

Feature selection

Feature selection is needed to reduce the feature space by filtering out unwanted features that reduce the classification performance. Feature selection is useful to know relatively more informative features from a collection of features that might contain redundant and non-informative features increasing the confidence of classification. For the selection of the attributes, information gain attribute evaluator from Weka software was employed.

Classification

Random Forests are an ensemble of randomly constructed independent decision trees. In each decision tree a randomly chosen fixed subset of features are employed to build a classification model. Bootstrapping technique is used in each tree for selection of training set. Due to this about one third of the examples are left unused and are known as out of bag examples. It is customary to use this out of bag examples as validation set for tuning the algorithm parameters. Hence a separate test data is not normally required in RF for checking the overall accuracy of the forest. After all individual trees are built a majority vote is then taken to decide on the class label for each case.

Discussion:

Separate versus Combined features

Separate features refer to the all possible values of various properties taken together as separate features. For example, for a property called 'charge', there are three possible values viz. positive, zero and negative. These three properties taken separately can be considered as three different features. Thus, here the feature 'positive charge' shows the number of atoms with positive charge. Combined features refer to the combination of values of more than one property. Advantage of using combined features is that, the combination of more than one property avoids the redundancy in the features. Since the feature values considered here are the counts of atoms of a particular property, using different values of the properties will give redundant counts for some of the properties. Clubbing the properties together to form a new property will automatically reduce the redundant counts. Thus the combined features give more realistic properties rather than the separate features. Another advantage of the combined features is the reduction in the feature space. Here only the atom wise features are used and residue wise features are omitted from the combinations. The results **Table 3 (see supplementary material)** indicate that with separate features there is a slight decrease in MCC and slight increase in accuracy with feature selection. The reversal in this trend is observed for combined features. The maximum accuracy is found to be 95.59 % and 94.11% for separate and combined features respectively.

Conclusion:

In this work ligand centroid approaches were employed for prediction of mannose-binding sites. The tuned classifier model with most informative features provides an accuracy of more than 90 % percent. The developed model can be used to predict the mannose binding sites with a high degree of confidence.

Acknowledgement:

Dr. V.K. Jayaraman gratefully acknowledges funding from Department of Science and Technology, New Delhi for financial assistance

References:

[1] Khuri S *et al*. *Mol Biol Evol*.2001 **18**: 593 [PMID: 11264412]

[2] García-Hernández E *et al*. *Glycobiology.* 2000 **10**: 993 [PMID: 11030745]
[3] Rao VS *et al*. *Int J Biol Macromol.* 1998 **24**: 295 [PMID: 9849627]
[4] Shionyu Mitsuyama C *et al. Protein Eng.* 2003 **16**: 467 [PMID: 12915724]
[5] Taroni C *et al*. *Protein Eng.* 2000 **13:** 89 [PMID: 10708647]
[6] Sujatha MS & Balaji PV, *Proteins.* 2004 **55**: 44 [PMID: 14997539]
[7] Malik A & Ahmad S, *BMC Struct Biol.* 2007 **7**: 1 [PMCID: PMC1780050].
[8] Nassif H *et al*. *Proteins.* 2009 **77**: 121 [PMID: 19415755]
[9] Bobadilla *et al*. *Advances in bioinformatics and its applications.* 2004 pp. **307**: 318.

Edited by P Kangueane
Citation: Khare *et al.* Bioinformation 8(24): 1202-1205 (2012)

BIOINFORMATION

open access

Supplementary material:

Table 1: List of separate features

Sr. No	Feature Name
1	Number of atoms of negative charge
2	Number of atoms of zero charge
3	Number of atoms of positive charge
4	Number of atoms of hydrophilic nature
5	Number of atoms of hydroneutral nature
6	Number of atoms of hydrophobic nature
7	Number of atoms that can form hydrogen bonds
8	Number of atoms that can not form hydrogen bonds
9	Number of atoms of residues of aromatic nature
10	Number of atoms of residues aliphatic nature
11	Number of atoms of residues acidic-carboxylic nature
12	Number of atoms of residues basic nature
13	Number of atoms of residues neutral nature
14	Number of atoms of histidine
15	Average solvent accessible area per shell

Table 2: List of combined features

Sr. No	Feature Name
1	Negative charge and Hydrophilic and Hydrogen bonding
2	Zero charge and Hydrophilic and Hydrogen bonding
3	Zero charge and Hydrophilic and Non hydrogen bonding
4	Zero charge and Hydroneutral and Non hydrogen bonding
5	Zero charge and Hydrophobic and Non hydrogen bonding
6	Positive charge and Hydrophilic and Hydrogen bonding
7	Positive charge and Hydrophilic and Non hydrogen bonding

Table 3: Result of Mannose binding site prediction using both separate and combined features

Feature Type	Accuracy	MCC
Separate	95.59	0.83
Combined	94.11	0.91