An ILP Approach to Model and Classify Hexose Binding Sites

Houssam Nassif[1,2], Hassan Al-Ali[3], Sawsan Khuri[4,5], Walid Keirouz[6], and David Page[1,2]

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

Abstract. Hexoses play a key role in many cellular pathways and in the regulation of development and disease mechanisms. As such, hexose binding proteins and their binding properties are of great interest to biomedical researchers. Current protein-sugar computational models are based, at least partially, on prior biochemical knowledge. We investigate the empirical support for biochemical findings by comparing ILP-induced rules to actual biochemical results. Our method matches lab findings, is able to identify hexose binding sites, and in addition reveals a Trp-Glu dependency. Our classifier achieves a similar accuracy as other black-box classifiers while providing insight into the discriminating process.

Key words: ILP, Aleph, rule generation, hexose, protein-carbohydrate interaction, binding site, substrate recognition

1 Introduction

Inductive Logic Programming (ILP) has been shown to perform well in predicting various substrate-protein bindings (e.g., [2, 13]). In this paper we apply ILP to a different and well studied binding task.

Hexoses are 6-carbon sugar molecules that play a key role in different biochemical pathways, including cellular energy release, signaling, carbohydrate synthesis, and the regulation of gene expression [11]. Hexose binding proteins belong to diverse functional families that lack significant sequence or, often, structural similarity [4]. Despite this fact, these proteins show high specificity to their hexose ligands. The few amino acids present at the binding site play a large role in determining the binding site's topology and biochemical properties and hence the ligand type and the protein's functionality.

Many researchers have investigated protein-sugar binding sites. From the biochemical perspective, Rao et al. [9] fully characterized the architecture of sugar

Presented at the 19th International Conference on Inductive Logic Programming (ILP'09).

binding in lectins and identified conserved loop structures within the protein as essential for sugar recognition. Later, Quiocho and Vyas [8] 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, and Phe, as well as 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. [16] 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 hydrophilic side group which establishes hydrogen bonds and a hydrophobic core that is able to stack against aromatic residues. Sugar binding sites are thus neither strictly hydrophobic nor strictly hydrophilic, due to the dual nature of sugar docking [16].

Some of this biochemical information has been used in computational work with the objective of accurately predicting sugar binding sites in proteins. Examples include Taroni et al. [16], who devised a probability formula by combining individual attribute scores, and Shionyu-Mitsuyama et al. [10] who used atom type densities within binding sites to develop an algorithm for predicting carbohydrate binding. Other groups formulated a signature for characterizing galactose binding sites based on geometric constraints, pyranose ring proximity and hydrogen bonding atoms [14], and implemented a 3D structure searching algorithm, COTRAN, to identify galactose binding sites. Other researchers used a neural network to predict general carbohydrate and specific galactose binding sites [5]. More recently, Nassif et al. [7] used support vector machines [17] to model and predict glucose binding sites in a wide range of proteins.

Biochemical findings are arrived at through wet lab experiments; computational hexose classifiers incorporate different parts of these findings in black-box models and use these models as the bases of predictions. No work to date has taken the opposite approach: given hexose binding sites data, what biochemical rules can we extract with no prior biochemical knowledge and what is the performance of the resulting classifier based solely on the extracted rules?

This work presents an ILP classifier that extracts rules from the data without prior knowledge. It classifies binding sites based on the extracted biochemical rules, clearly specifying the rules used to discriminate each instance. This inductive data-driven approach validates the biochemical findings and allows a better understanding of the black-box classifiers' output.

2 Materials and Methods

The Protein Data Bank (PDB) [1] is the largest repository of experimentally determined three-dimensional structures of biological macromolecules. We mine it for the three most common hexoses, Galactose, Glucose and Mannose. Using PISCES [18], we impose a 30% overall sequence identity as a cut-off. We examine

the remaining structures at close range using the Swiss-PDBViewer program [3] and discard several proteins 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-redundant positive data set of 80 protein-hexose binding sites. We also extract an equal number of negative examples. The negative set is composed of non-hexose binding sites and of non-binding surface grooves.

We use the ILP engine Aleph [12] to learn first-order rules. Rule learning is especially appealing because of its easy-to-understand format. A set of if-then rules describing a certain concept is highly expressive and readable [6]. We estimate the classifier's performance using 10-fold cross-validation.

Given the molecule and the binding site, we extract the coordinates of every atom within a distance of 10 Å from the binding site center. We define a Euclidean distance measure between any two atoms. We only extract atoms using their PDB atomic names; we do not consider residues. For every atom we compute its charge, hydrogen bonding, and hydrophobicity properties as done by Nassif et al. [7].

We restrict the clause length to a maximum number of 8 literals, with only one in the head. The consequent of any rule is $bind(A)$, where A is predicted to be a hexose binding site. Literals describing individual atoms are of the form:

$$point(A, B, C, D, E, F, G, H, I, J) \tag{1}$$

where A is the binding site and B is the atom number. C, D, & E specify the Cartesian coordinates. F is the charge, G the hydrogen-bonding, and the H hydrophobicity. Lastly, I and J refer to the atomic element and its name.

Clause bodies can also use distance literals:

$$dist(A, B1, B2, M, N) \tag{2}$$

where A is the binding site and $B1$ and $B2$ two atom numbers. M is their distance apart and N the error, resulting in $M \pm N$. The error N is set to 0.5 Å.

No literal can contain terms pertaining to different binding sites. As a result, A is the same in all literals in a clause. For each clause, coverage of up to 5 training-set negative examples is tolerated. The cost function to minimize is:

$$cost = (\#\ covered\ negatives) - (\#\ covered\ positives) \tag{3}$$

To speed the search, we use Aleph's heuristic search.

3 Results

ILP's accuracy over the 10 folds is 67.5%. Even though Aleph was only looking at the atoms according to their PDB atomic name, valuable information regarding amino acids can be inferred. For example ND1 atoms are only present within His residues and a rule requiring the presence of ND1 is actually requiring His.

We present the rules' English translation, with residue substitution. We sort the rules by their coverage. The queried site is considered hexose binding if any of these rules apply:

1. It contains a TRP residue and a GLU with an OE1 atom that is 8.53 Å away from a negatively charged Oxygen atom.
   [Pos cover = 22, Neg cover = 4]
2. It contains a PHE or TYR residue and an ASP with an OD1 atom that is 5.24 Å away from an ASP or ASN's OD1.
   [Pos cover = 21, Neg cover = 3]
3. It contains a branching aliphatic residue (LEU, VAL, ILE), an ASP and an ASN. ASP and ASN's OD1 atoms are 3.41 Å away.
   [Pos cover = 15, Neg cover = 0]
4. It contains a hydrophilic non-hydrogen bonding Nitrogen atom (PRO, ARG, HIS) with a distance of 7.95 Å away from a HIS ND1 nitrogen, and 9.60 Å away from a branching aliphatic residue's CG1.
   [Pos cover = 10, Neg cover = 0]
5. It has a hydrophobic CD2 atom, a hydrophilic PRO backbone or HIS ND1 nitrogen and two GLU (or two GLN) distant by 11.89 Å.
   [Pos cover = 11, Neg cover = 2]
6. It contains an ASP B, two identical atoms Q and X, and a hydrophilic hydrogen-bonding atom K. Atoms K, Q and X have the same charge. B's ODE1 oxygen share the same Y-coordinate with K and the same Z-coordinate with Q. Atom X is 8.29 Å away from atom K.
   [Pos cover = 8, Neg cover = 0]
7. It contains a SER, and two GLN and/or HIS, with NE2 atoms that are 3.88 Å apart.
   [Pos cover = 8, Neg cover = 2]
8. It contains an ASN and a PHE, TYR or HIS residue, with a CE1 atom that is 7.07 Å away from a Calcium.
   [Pos cover = 5, Neg cover = 0]
9. It contains a LYS or ARG, a PHE or TYR, a PRO or HIS, and a Sulfate or a Phosphate.
   [Pos cover = 3, Neg cover = 0]

4 Discussion

Aleph's error of 32.5% has a p-value < 0.0002, according to a two-sided binomial test. This error rate is comparable to other general sugar binding site classifiers, although each was run on a different dataset (Table 1). Contrary to black-box classifiers, ILP provides a number of interesting insights. It infers most of the established biochemical information about residues and relations just from the PDB atom names and properties.

The first two rules, covering the highest number of positives, rely on the aromatic residues (TRP, TYR, PHE). This highlights the docking interaction between the hexose and the aromatic residues [15]. The aromatic residues stack against the apolar sugar pyranose ring which stabilizes the bound hexose. HIS is mentioned in many of the rules, this is also an aromatic amino acid that would provide a similar docking mechanism to TRP, TYR and PHE [8].

Table 1. Comparison of general sugar binding site classifiers. Not meant as a direct comparison since the datasets are different.

Program	Error (%)	Method and Dataset
ILP hexose predictor	32.50	10-fold cross-validation, 80 hexose and 80 non-hexose binding sites.
Shionyu-Mitsuyama et al. [10]	31.00	Test set of 61 polysaccharide binding sites.
Taroni et al. [16]	35.00	Test set of 40 carbohydrate binding sites.
Malik and Ahmad [5]	39.00	Leave-one-out method, 40 carbohydrate and 116 non-carbohydrate binding sites.

All rules require the presence of a planar polar residue (Asn, Asp, Gln, Glu, Arg). These residues have been identified as the most frequently involved in the hydrogen-bonding of the hexose [8]. The hydrogen bond is probably the most relevant interaction in protein binding in general.

A high residue-sugar propensity value reflects a high tendency of that residue to be involved in sugar binding. The residues having high propensity values are the aromatic residues, including histidine, and the planar polar residues [16]. This fact is reflected by several of the rules above.

Some rules require hydrophobic atoms/residues, while others require hydrophilic ones. Rule 5 requires both and reflects the dual nature of sugar docking, composed of a polar-hydrophilic aspect establishing hydrogen bonds and a hydrophobic aspect responsible for the pyranose ring stacking [16].

Rules number 8 and 9 require the presence of different ions (Calcium, Sulfate, Phosphate), confirming the relevance of ions in hexose binding [8].

Finally, rule 2 suggests a dependency between Phe/Tyr and Asn/Asp. Such a relation has been proven in lectins [9]. Similarly, rule 1 suggests a dependency between Trp and Glu, a link not previously identified in the literature.

Whereas these rules provide a deep insight into the requirements for a hexose binding site, they would ideally need to be validated by running a set of mutated samples, substituting the atoms referred to in the rules above with other atom types and running the analysis again. The Protein Data Bank is continually growing with more proteins added on a daily basis and these rules could be further refined by testing the algorithm again once more hexose binding structures have been deposited into the database.

5 Conclusion

ILP achieves a similar accuracy as other general sugar black-box classifiers, while offering insight into the discriminating process. With no prior biochemical knowledge, Aleph was able to induce most of the known hexose-protein interaction biochemical rules. The results of our data-driven computational methods therefore correspond to wet lab findings and could form the basis for a predictive software tool. ILP finds a previously unreported dependency between Trp and Glu, a relationship that merits further investigation.

Acknowledgments. This work was partially supported by US NIH grant R01CA127379-01.

References

1. H.M. Berman, J. Westbrook, Z. Feng, G. Gilliland, T.N. Bhat, H. Weissig, I.N. Shindyalov, and P.E. Bourne. The Protein Data Bank. *Nucleic Acids Res.*, 28:235–242, 2000.
2. P. Finn, S. Muggleton, D. Page, and A. Srinivasan. Pharmacophore Discovery using the Inductive Logic Programming System PROGOL. *Mach. Learn.*, 30:241–273, 1998.
3. N. Guex and M.C. Peitsch. SWISS-MODEL and the Swiss-PdbViewer: An Environment for Comparative Protein Modeling. *Electrophoresis*, 18:2714–2723, 1997.
4. S. Khuri, F.T. Bakker, and J.M. Dunwell. Phylogeny, Function and Evolution of the Cupins, a Structurally Conserved, Functionally Diverse Superfamily of Proteins. *Mol. Biol. Evol.*, 18:593–605, 2001.
5. A. Malik and S. Ahmad. Sequence and Structural Features of Carbohydrate Binding in Proteins and Assessment of Predictability Using a Neural Network. *BMC Struct. Biol.*, 7:1, 2007.
6. T.M. Mitchell. *Machine Learning.* McGraw-Hill, Singapore, 1997.
7. H. Nassif, H. Al-Ali, S. Khuri, and W. Keirouz. Prediction of Protein-Glucose Binding Sites Using Support Vector Machines. *Protein. Struct. Func. Bioinf.*, DOI: 10.1002/prot.22424, 2009.
8. F.A. Quiocho and N.K. Vyas. Atomic Interactions Between Proteins/Enzymes and Carbohydrates. In S.M. Hecht, editor, *Bioorganic Chemistry: Carbohydrates*, chapter 11, pages 441–457. Oxford University Press, New York, 1999.
9. V.S.R. Rao, K. Lam, and P.K. Qasba. Architecture of the Sugar Binding Sites in Carbohydrate Binding Proteins—a Computer Modeling Study. *Int. J. Biol. Macromol.*, 23:295–307, 1998.
10. C. Shionyu-Mitsuyama, T. Shirai, H. Ishida, and T. Yamane. An Empirical Approach for Structure-Based Prediction of Carbohydrate-Binding Sites on Proteins. *Protein Eng.*, 16(7):467–478, 2003.
11. E. Solomon, L. Berg, and D.W. Martin. *Biology.* Brooks Cole, Belmont, CA, 8th edition, 2007.
12. A. Srinivasan. *The Aleph Manual*, 2007. http://www.comlab.ox.ac.uk/activities/machinelearning/Aleph/aleph.html.
13. A. Srinivasan, R.D. King, S. Muggleton, and M.J.E. Sternberg. Carcinogenesis Predictions Using ILP. In *Proceedings of the 7th International Workshop on Inductive Logic Programming*, pages 273–287, 1997.
14. M.S. Sujatha and P.V. Balaji. Identification of Common Structural Features of Binding Sites in Galactose-Specific Proteins. *Protein. Struct. Func. Bioinf.*, 55:44–65, 2004.
15. M.S. Sujatha, Y.U. Sasidhar, and P.V. Balaji. Energetics of Galactose- and Glucose-Aromatic Amino Acid Interactions: Implications for Binding in Galactose-Specific Proteins. *Protein Sci.*, 13:2502–2514, 2004.
16. C. Taroni, S. Jones, and J.M. Thornton. Analysis and Prediction of Carbohydrate Binding Sites. *Protein Eng.*, 13(2):89–98, 2000.
17. V.N. Vapnik. *Statistical Learning Theory.* John Wiley & Sons, New York, 1998.
18. G. Wang and R.L. Dunbrack. PISCES: A Protein Sequence Culling Server. *Bioinf*, 19:1589–1591, 2003.