Title: Computational Analysis and Prediction of Carbohydrate Binding Sites in Proteins.

Abstract Submitted

By

Adeel Malik
Dept. of Biosciences,
Jamia Millia Islamia,
ND-25

Supervisor
Prof. (Dr.) SF Basir
Dept. Of Biosciences
Jamia Millia Islamia
ND-25

Co-Supervisor
Dr. Shandar Ahmad
Research Scientist,
NIBIO, Japan.

Abstract

Protein-Carbohydrate interactions are crucial in many biological processes with implications to drug targeting and gene expression. Nature of protein-carbohydrate interactions may be studied at individual residue level by analyzing local sequence and structure environments in binding regions in comparison to non-binding regions, which provide an inherent control for such analyses. Since carbohydrates assume a large variety of configurations, many carbohydrate-binding proteins are being considered as targets for new medicines.

In view of the above, accurate *in silico* identification of carbohydrate-binding sites is a key issue in genome annotation and drug targeting. The information about the factors, which prevent or support carbohydrate binding of an amino acid, is expected to be present in the evolutionary profile of the sequence as well as the identity and structure of amino acid residues in the neighborhood of potential carbohydrate binding sites. A number of reviews have been published on protein- carbohydrate interactions. Different aspects of protein carbohydrate recognition have also been extensively studied. However, bioinformatics approaches with a predictive goal are relatively rare. Compared with the abundance of methodologies developed for protein-nucleic acid or protein-protein interactions, there are still very few methods for predicting carbohydrate-protein interactions. Shionyu-Mitsuyama has developed a program that uses the empirical rules of the spatial distribution of protein atoms at known carbohydrate-binding sites for prediction. In another work, an analysis of the characteristic properties of sugar binding

sites was performed on a set of 19 sugar binding proteins. For each site six parameters were evaluated viz. solvation potential, residue propensity, hydrophobicity, planarity, protrusion and relative accessible surface area. Three of the parameters were found to distinguish the observed sugar binding sites from the other surface patches. These parameters were then used to calculate the probability for a surface patch to be a carbohydrate-binding site. These prediction methods are based on local structural descriptors of proteins and cannot be used if complete 3 dimensional structures are not available. Artificial neural networks have been used in the prediction of N-Linked and O-Linked Glycosylation sites. However, these studies are restricted to only one type of protein-carbohydrate interactions and therefore do not capture all protein-carbohydrate interactions, as sought out in this work.

In this work we explored the exact contribution from different sequence, evolutionary and structural attributes of proteins in determining their carbohydrate binding regions. Propensity of each of the 20 amino acid residues in binding regions has been calculated and compared with non-binding regions. Solvent accessibility, secondary structure and packing density of binding sites have been analyzed in a similar way. A neural network was designed to model sequence and evolutionary information (obtained by Position Specific Scoring Matrices) in addition to structural features, and to determine their role in the predictability of carbohydrate binding sites. Additionally, a web server, CBS-Pred [http://cbs-pred.netasa.org], has been developed, which predicts the carbohydrate binding sites from sequence and evolutionary information using neural networks. We also studied the binding sites of other protein-ligand and protein-DNA complexes and compared the propensity scores of all residues and their secondary structures with protein-carbohydrate complexes.

We have also compiled a database of diverse carbohydrate binding proteins used in this study, which might provide information to the scientific community globally. Additionally, an automated homology based modeling of a large number of N- and O-Linked glycoproteins has also been carried out and as a case study; the project based comparative modeling of the Latex allergen glycoprotein Hev b 4 was also carried out. All these models and protein carbohydrate complexes are available at http://www.procarb.org/procarbdb.