Computational Biology and Chemistry 66 (2017) 36–43

Contents lists available at ScienceDirect

Computational Biology and Chemistry

journal homepage: www.elsevier.com/locate/compbiolchem

Research Article

Development of a sugar-binding residue prediction system from protein sequences using support vector machine

Masaki Banno[a], Yusuke Komiyama[b], Wei Cao[a], Yuya Oku[a], Kokoro Ueki[a], Kazuya Sumikoshi[a], Shugo Nakamura[a], Tohru Terada[a], Kentaro Shimizu[a,*]

[a] Graduate School of Agricultural and Life Sciences, The University of Tokyo, 1-1-1 Yayoi, Bunkyo-Ward, Tokyo 113-8657, Japan
[b] Digital Content and Media Sciences Research Division, National Institute of Informatics, 2-1-2 Hitotsubashi, Chiyoda-Ward, Tokyo 101-8430, Japan

ARTICLE INFO

Article history:
Received 8 March 2016
Received in revised form 5 October 2016
Accepted 23 October 2016
Available online 9 November 2016

Keywords:
Support vector machine
Sugar-binding proteins
Sugar-binding residue prediction
Carbohydrate
Machine learning

ABSTRACT

Several methods have been proposed for protein–sugar binding site prediction using machine learning algorithms. However, they are not effective to learn various properties of binding site residues caused by various interactions between proteins and sugars. In this study, we classified sugars into acidic and nonacidic sugars and showed that their binding sites have different amino acid occurrence frequencies. By using this result, we developed sugar-binding residue predictors dedicated to the two classes of sugars: an acid sugar binding predictor and a nonacidic sugar binding predictor. We also developed a combination predictor which combines the results of the two predictors. We showed that when a sugar is known to be an acidic sugar, the acidic sugar binding predictor achieves the best performance, and showed that when a sugar is known to be a nonacidic sugar or is not known to be either of the two classes, the combination predictor achieves the best performance. Our method uses only amino acid sequences for prediction. Support vector machine was used as a machine learning algorithm and the position-specific scoring matrix created by the position-specific iterative basic local alignment search tool was used as the feature vector. We evaluated the performance of the predictors using five-fold cross-validation. We have launched our system, as an open source freeware tool on the GitHub repository (https://doi.org/10.5281/zenodo.61513).

© 2016 The Authors. Published by Elsevier Ltd. This is an open access article under the CC BY-NC-ND license (http://creativecommons.org/licenses/by-nc-nd/4.0/).

1. Introduction

Interactions between sugar chains and proteins play essential roles in biological processes such as intercellular communication, immunity, and cellular recognition. The methods to empirically analyze such interactions include hemagglutination assays, which are employed in the discovery of novel lectins. In recent years, methods utilizing glycan arrays have been developed as high-throughput solutions, enabling researchers to obtain data on *in vitro* interactions between multiple sugar chains and proteins (Porter et al., 2010; Blixt et al., 2004; Gabius et al., 2011). Nevertheless, the bioinformatics-based prediction approaches can further reduce the time and effort involved in predicting such interactions, providing valuable clues for experimental work. Conventional methods are useful in determining protein–sugar chain interactions or identifying sugar chain recognition sequences. However, they cannot provide information on the binding residues in proteins. Methods such as X-ray crystallography and nuclear magnetic resonance have primarily been used to identify these binding residues. However, such techniques pose numerous challenges because they are generally cost- and labor-intensive, Moreover, the high motility of sugar chains renders the determination of their tertiary structures difficult (DeMarco and Woods, 2008). As partial solutions to such challenges, bioinformatics-based techniques have been attracting attention.

Docking simulation is a prediction method for sugar-binding residues based on their tertiary structures. To implement this method, many protein–ligand docking programs (Morris et al., 2009; Jones et al., 1995, 1997; Biesiada et al., 2011; Forli et al., 2016; Grinter et al., 2014) and molecular simulations are often employed. In a previous study involving sugar chain-binding residues, the heparin-binding residues have been predicted in an interleukin on the basis of its protein structure (DeMarco and Woods, 2008). The candidate residues were narrowed down via repeated docking with heparin monosaccharides and disaccharides. Then, the heparin hexasaccharides were docked to the remaining candidates to

[*] Corresponding author.
E-mail address: shimizu@bi.a.u-tokyo.ac.jp (K. Shimizu).

http://dx.doi.org/10.1016/j.compbiolchem.2016.10.009
1476-9271/© 2016 The Authors. Published by Elsevier Ltd. This is an open access article under the CC BY-NC-ND license (http://creativecommons.org/licenses/by-nc-nd/4.0/).

predict the heparin-binding residues in the interleukin. Another study has used machine learning to predict glucose-binding residues from tertiary structure of proteins. It has employed a learning model with a support vector machine (SVM), which used the occurrence rates of atoms appearing in the proximity of glucose-binding residues as the feature values (McDonald and Thornton, 1994). Tsai et al. (2012) developed a sugar-binding site prediction method based on three-dimensional probability density maps, representing the distributions of 36 non-covalent interacting atom types around protein surfaces. The method reported by Zhao et al. (2014) uses a structural alignment program, SPalign and binding affinity scores, according to a knowledge-based potential.

All of these methods rely on the tertiary structure of the target protein for the prediction of the binding residues, thus requiring the determination of the protein structure. The amino acid sequence of a protein is much easier to obtain than its tertiary structure. Thus, it is preferable for the high-throughput experiments such as genome-wide and glycan arrays analyses.

Some attempts have been made to build software applications capable of learning such features so that they can predict sugar-binding residues only from amino acid sequences. Malik et al. have developed a machine learning-based method using neural networks. They have constructed a prediction program using the position-specific scoring matrices (PSSMs) derived from the residue frequency and multiple alignments of 40 sugar-binding proteins and 18 galactose-binding proteins as the feature values. The performance of the program has been evaluated by leave-one-out cross-validation (CV) (Malik and Ahmad, 2007). Their results show that the prediction program performs more effectively when applied to a dataset of galactose-binding proteins than that when learning using to all sugar-binding proteins. Nassif et al. (2009) also developed a glucose-binding site prediction method. This method uses spatial features of binding pockets and amino acid and chemical features such as charge, polarity, mobility, and hydrophobicity as determinant features of a binding site. Recently, a mannose-binding site prediction program has been developed; it uses the composition profile of patterns as sequence features (Agarwal et al., 2011).

In this present study, we attempted a high-performance prediction by grouping the sugar-binding proteins depending on the characteristics of their binding residues and designing a predictor dedicated to each group. We analyzed the characteristics of the binding residues by clustering the sugars according to the residue composition at the binding sites, and thereby classified the sugars into different classes. Individual predictors for each sugar class made the learning of the propensities of the binding residues more effective. This, in turn, resulted in improved prediction performance of the predictor. Furthermore, our method uses only the amino acid sequences for prediction. SVM was employed because it is one of the representative techniques for the classification of the data into two categories with high generalization ability. SVM takes as input PSSMs around a target residue as feature values. It can improve the prediction capability further by extensive incorporation of the nature of homologous proteins coupled with sugar class-specific learning.

2. Materials and methods

2.1. Search for sugar-binding proteins in the protein data bank database

We targeted the sugars that frequently occur *in vivo*, namely aldoses and ketoses, and their derivatives in which the hydroxy group is oxidized or substituted with a methyl group, sulfonic group, phosphate group, acetyl group, amine group, or acetyl amide group. Fig. 1 illustrates the procedure for constructing the dataset used for prediction.

With sugar-binding residues defined as the residues within 4 Å of the sugar molecule. We performed an exhaustive search of protein data bank (PDB) for proteins with at least one sugar-binding residues. This study focused on noncovalent interactions between sugars and proteins and not on glycosylation sites at which sugars are covalently bonded with proteins. Therefore, the residues within the 1.5 Å distance from a sugar molecule, as well as the residues adjacent to a covalently bonded sugar molecule, were excluded from the search.

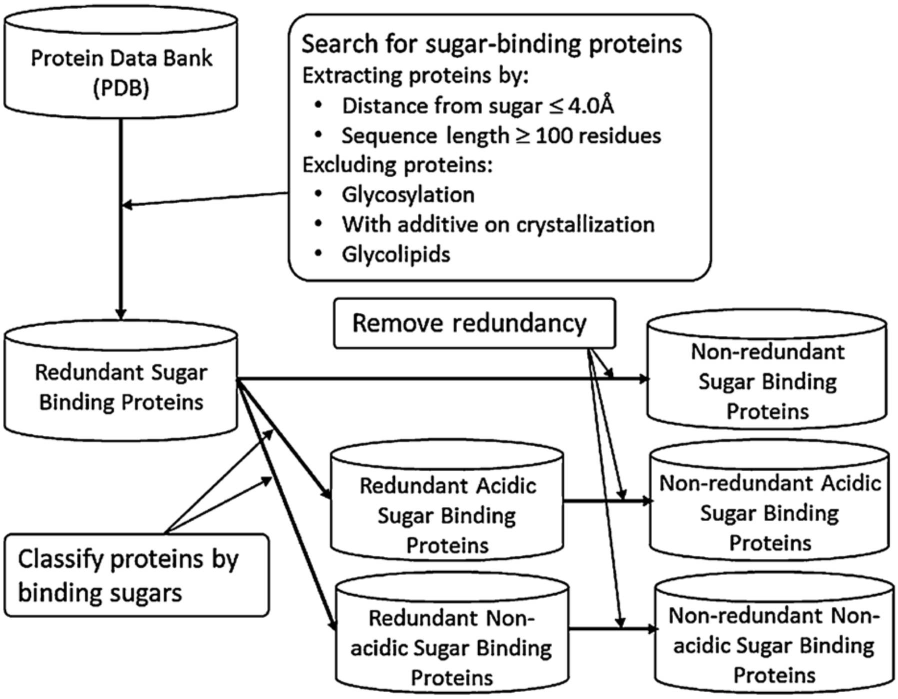
Protein Data Bank (PDB)
Search for sugar-binding proteins
Extracting proteins by:
• Distance from sugar ≤ 4.0Å
• Sequence length ≥ 100 residues
Excluding proteins:
• Glycosylation
• With additive on crystallization
• Glycolipids
Remove redundancy
Redundant Sugar Binding Proteins
Non-redundant Sugar Binding Proteins
Redundant Acidic Sugar Binding Proteins
Non-redundant Acidic Sugar Binding Proteins
Classify proteins by binding sugars
Redundant Non-acidic Sugar Binding Proteins
Non-redundant Non-acidic Sugar Binding Proteins

Fig. 1. Construction of the dataset used for prediction.

Download English Version:

https://daneshyari.com/en/article/6451380

Download Persian Version:

https://daneshyari.com/article/6451380

Daneshyari.com