Journal of Molecular Graphics and Modelling 28 (2009) 297–303

Contents lists available at ScienceDirect

Journal of Molecular Graphics and Modelling

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

InCa-SiteFinder: A method for structure-based prediction of inositol and carbohydrate binding sites on proteins

Mahesh Kulharia [a,b], Stephen J. Bridgett [a], Roger S. Goody [b], Richard M. Jackson [a,*]

[a] Institute of Molecular and Cellular Biology and Astbury Centre for Structural Molecular Biology, University of Leeds, Leeds, LS2 9JT, UK
[b] Department of Physical Biochemistry, Max Planck Institute for Molecular Physiology, Otto Hahn Strasse 11, Dortmund, 44227, Germany

ARTICLE INFO

Article history:
Received 30 June 2009
Received in revised form 17 August 2009
Accepted 18 August 2009
Available online 27 August 2009

Keywords:
Binding site prediction
Prediction of function
Q-SiteFinder
Glycobiology
Carbohydrate recognition
Glyco-bioinformatics

ABSTRACT

Carbohydrate binding sites are considered important for cellular recognition and adhesion and are important targets for drug design. In this paper we present a new method called InCa-SiteFinder for predicting non-covalent inositol and carbohydrate binding sites on the surface of protein structures. It uses the van der Waals energy of a protein–probe interaction and amino acid propensities to locate and predict carbohydrate binding sites. The protein surface is searched for continuous volume envelopes that correspond to a favorable protein–probe interaction. These volumes are subsequently analyzed to demarcate regions of high cumulative propensity for binding a carbohydrate moiety based on calculated amino acid propensity scores.

InCa-SiteFinder[1] was tested on an independent test set of 80 protein–ligand complexes. It efficiently identifies carbohydrate binding sites with high specificity and sensitivity. It was also tested on a second test set of 80 protein–ligand complexes containing 40 known carbohydrate binders (having 40 carbohydrate binding sites) and 40 known drug-like compound binders (having 58 known drug-like compound binding sites) for the prediction of the location of the carbohydrate binding sites and to distinguish these from the drug-like compound binding sites. At 73% sensitivity the method showed 98% specificity. Almost all of the carbohydrate and drug-like compound binding sites were correctly identified with an overall error rate of 12%.

1. Introduction

The function of a protein is largely defined by the nature of the molecules it interacts with. Therefore the prediction of protein binding sites and their characterization remain important goals for the biologist. The number of known structures of proteins has grown rapidly in recent years [1] and a large number of protein–ligand interaction sites remain uncharacterized [2]. A number of approaches have been developed to make predictions about the function of a protein from its structure [3]. Some of these methods look for motifs or domains associated with specific functions [3], others look for characteristic arrangement of functionally important or conserved residues [4]. The function of a protein depends upon the nature of ligands it can interact with, hence demarcation of the ligand binding sites and identification of the type of ligands it can bind is important for the assignment of function to the protein structure as well as for rational structure-based drug design.

Non-covalent carbohydrate binding proteins play an important role in cellular processes. Carbohydrates are involved in energy flow, cellular recognition and adhesion [5]. Carbohydrate binding proteins are however very diverse in structure and function [6]. They are increasingly being considered as putative drug targets because of their role in intra- and inter-cellular communication [6]. Experimentally carbohydrate binding sites have been extensively studied in the past [7]. However, only a few approaches have been developed for the prediction of carbohydrate binding sites from structure [8–10]. Taroni et al. ranked the surface patches on the basis of the average propensity of the patch residues to bind carbohydrates. The patches having an average propensity score above a specific threshold were considered as carbohydrate binders. This method was tested on two datasets. The first test set (comprising of 3 lectins and 4 enzymes) consisted of proteins non-homologous to the training dataset whereas the members of second dataset (19 enzymes and 14 lectins) were homologous to the training dataset. The method was 89% successful for identification of the carbohydrate binding sites in the homologous enzymes whilst the method correctly predicted 29% of cases in the homologous lectins. Shionyu-Mitsuyama et al. developed a set of rules from a dataset of 80 protein–carbohydrate binding sites that depicted the probable positions of carbohydrate-interacting

* Corresponding author. Tel.: +44 0113 343 2592; fax: +44 0113 343 3167.
E-mail address: r.m.jackson@leeds.ac.uk (R.M. Jackson).

[1] Access to InCa-SiteFinder is freely available at: http://www.modelling.leeds.ac.uk/InCaSiteFinder/.

1093-3263/$ – see front matter © 2009 Elsevier Inc. All rights reserved.
doi:10.1016/j.jmgm.2009.08.009

protein atoms. Using a set of 10 atom types they created a three-dimensional probability density map wherein each point on this map represented the probability of occurrence of a protein atom which could interact with a carbohydrate. Using these interaction maps they predicted the carbohydrate binding sites with a success rate of 66% and 50% in enzymes and lectins, respectively. Malik et al. trained a neural network using amino acid propensities for the prediction of carbohydrate binding sites. The training set comprised of 40 protein–carbohydrate complexes and the level of redundancy was reduced by removing protein sequences with more than 50% sequence identity. This method achieved only 23% specificity at 87% sensitivity.

Here the development of a new computational method for predicting carbohydrate binding sites is presented. The overall aim was to develop a new computational method for predicting carbohydrate binding sites with high sensitivity and specificity. The method differs from the previous carbohydrate binding site prediction methods in two important aspects. Firstly it uses 375 non-covalent protein–carbohydrate complexes for the derivation of amino acid propensity scores, which is more than that used in the previous studies. Secondly it uses a two-step procedure to identify sites. In step one; it uses an energetic grid-based approach to identify putative sites on the protein with a high probability of being a binding site, using the method of Laurie and Jackson [2]. In step two; it uses these sites and amino acid propensity scores to predict the location of carbohydrate binding sites. The aim of developing InCa-SiteFinder was to produce a method that could perform two functions: (1) locate likely ligand binding sites and (2) distinguish the nature of the binding site, to ascertain if the site can preferentially bind a carbohydrate ligand.

2. Methods

2.1. Construction of dataset for propensity calculation

Nearly 30,000 protein–ligand complexes present in PDBSUM [11–13] with structural information were extracted from the PDB [14]. Of these only protein–carbohydrate complexes having experimentally determined X-ray crystal structures with a resolution greater than 2.5 Å were retained. In addition, complexes were further removed if they had either: a covalently bound ligand; involved a drug-like compound ligand; had metallic ions; or had no classification in SCOP (version 1.69) [15]. A ligand was classified as non-covalently bound to the protein if none of its atoms were within the covalent interaction distance (see supporting information). The covalent interaction distance for a specific protein and ligand atom pair was the sum of their atomic radii plus a 10% tolerance limit.

A non-redundant dataset was constructed by considering the protein chain/s (containing a domain) with a bound carbohydrate ligand for each SCOP superfamily representative. The SCOP domain code is unique at the superfamily level in the carbohydrate binding domain for each entry and the best resolution structural representative was chosen. Thus the final dataset comprised a non-redundant dataset (NRD) with only one carbohydrate representative for each SCOP superfamily. Hydrogen atoms were added to these protein–carbohydrate complexes using the QuacPac software (OpenEye).

2.2. Calculation of amino acid propensities

For a non-redundant database of over 375 protein–carbohydrate complexes, propensities for a given amino acid to occur in a carbohydrate binding sites were calculated as the ratio of its relative contribution to the carbohydrate binding site area to its relative contribution to the complete protein surface area. The area contributed by an amino acid, *i*, to the carbohydrate binding site was considered as the difference in its solvent accessible surface area between the carbohydrate bound and unbound states. The propensity of an amino acid, *i*, to occur in a carbohydrate binding site (P_i^{CBP}) and drug-like compound binding site (P_i^{DBP}) are given by:

$$P_i^{CBP} = \frac{\Delta SASA_i^{CBS} / \sum_{j=1}^{20} \Delta SASA_j^{CBS}}{SASA_i / \sum_{j=1}^{20} SASA_j} \quad (1)$$

$$P_i^{DBP} = \frac{\Delta SASA_i^{DBS} / \sum_{j=1}^{20} \Delta SASA_j^{DBS}}{SASA_i / \sum_{j=1}^{20} SASA_j} \quad (2)$$

where $\Delta SASA_i^{CBS}$ is the solvent accessible surface area of amino acid *i* buried in the carbohydrate bound state. $\sum \Delta SASA_i^{CBS}$ is the total solvent accessible surface area of all amino acids buried in carbohydrate bound complexes. $\Delta SASA_i^{DBS}$ is the solvent accessible surface area of amino acid *i* buried in the drug-like ligand bound state. $\sum \Delta SASA_i^{DBS}$ is the total solvent accessible surface area of all amino acids buried in drug-like ligand bound complexes. $SASA_i$ is the solvent accessible surface area contributed by a specific amino acid *i* to the protein surface. $\sum SASA_i$ is the total solvent accessible surface area of all amino acids of the protein. For comparison the amino acid propensities of drug-like compound binding sites were also determined in the same way. These were calculated from a nonredundant database of 358 complexes of protein–drug-like compounds. The ligands were considered as drug-like if they conformed to Lipinski's rule of 5 [16] and did not contain a carbohydrate moiety.

2.3. InCa-SiteFinder

The process of calculating the protein–probe van der Waals interaction energy is described in detail in Laurie and Jackson [2]. Briefly, the protein atoms are placed in a three-dimensional box, which is divided into a cubic grid of resolution 0.9 Å. Using the program Liggrid the van der Waals energy of interaction is calculated between the protein and a methylene ($-CH_3$) probe placed at each grid point. The energy is calculated using the GRID force-field parameters as described in Ref. [17]. Grid points with a "protein–probe interaction" energy more favorable (negative) than a predetermined threshold are retained (Fig. 1). For these grid

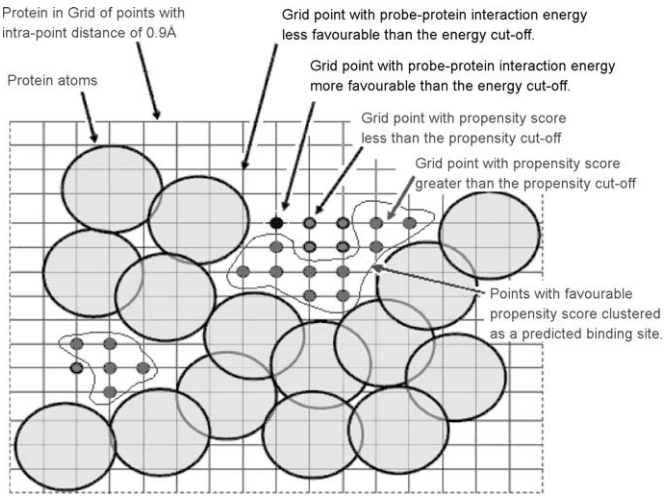

Fig. 1. An initial van der Waals energy cut-off is used to retain grid points in energetically favorable binding regions (small filled circles). A carbohydrate binding site occurrence propensity score cut-off is used to remove grid points in regions of low CBP score (small grey circles). Neighbouring favorable propensity score grid points are finally clustered to form the predicted sites (lines).