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

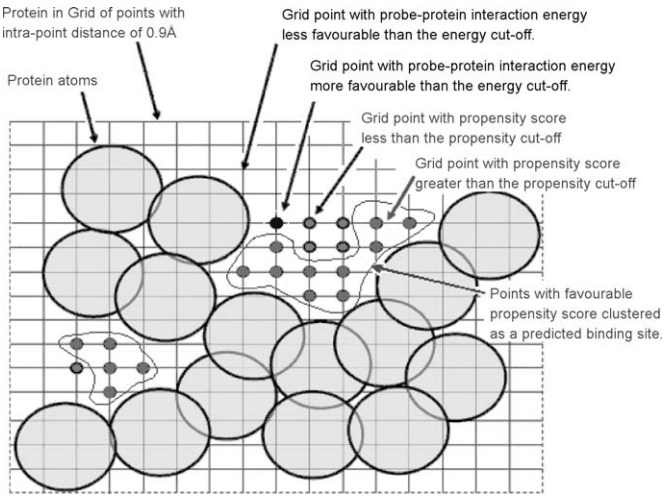

Fig. 1. An initial van der Waals energy cut-off is used to retain grid points in energetically favorable binding regions (small filled circles). A carbohydrate binding site occurrence propensity score cut-off is used to remove grid points in regions of low CBP score (small grey circles). Neighbouring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

points, carbohydrate binding propensity ($Score_k^{CBP}$) and drug-like compound binding propensity ($Score_k^{DBP}$) scores are calculated by considering the type of amino acid residue whose atoms are occluded from solvent exposure due to the predicted grid points. An amino acid is considered to be interacting with a grid point if at least one of its atoms is within 1.6 Å of the grid point. The overall carbohydrate binding propensity ($Score_k^{CBP}$) and drug-like compound binding propensity ($Score_k^{DBP}$) scores of the grid point, k, are defined as:

$$Score_k^{CBP} = \frac{\sum_{i=1}^{20}(n_i)(P_i^{CBS})}{N} \quad (3)$$

$$Score_k^{DBP} = \frac{\sum_{i=1}^{20}(n_i)(P_i^{DBS})}{N} \quad (4)$$

where n_i is the number of atoms of a specific amino acid (i) within 1.6 Å of the grid point; N is the total number of atoms interacting with the grid point k.

The grid points having a propensity score below a predetermined threshold values are removed (Fig. 1). The remaining grid points are clustered on the basis of their spatial proximity. A cluster is defined as the group of grid points wherein none of the grid points has its centre farther than 1.0 Å from the centre of the nearest grid point. For each cluster i a sum of $Score_k^{CBP}$ and $Score_k^{DBP}$ scores of the grid points ($j \rightarrow 1, n$) are calculated according to the following equations:

$$PSSBC_i = \sum_{j=1}^{n} Score_j^{CBP} \quad (5)$$

$$PSSBD_i = \sum_{j=1}^{n} Score_j^{DBP} \quad (6)$$

where $PSSBC_i$ and $PSSBD_i$ are the propensity scores of the cluster (which is a putative ligand binding site) i, to bind carbohydrates and drug-like compounds, respectively; n is the total number of the grid points in the cluster. The sites are then subjected to a threshold, whereby the sites having scores less than a given cut-off are removed. The remaining sites are ranked in order of their PSSBC score. For each of the ranked sites a differential propensity score ($DiffPS_i$) is calculated as the difference between $PSSBC_i$ and $PSSBD_i$. This represents the overall preference of the predicted site for carbohydrate over drug-like compound ligands and is defined as:

$$DiffPS_i = PSSBC_i - PSSBD_i \quad (7)$$

where $DiffPS_i$ is the differential propensity of site i. A cut-off is applied on the predicted sites in order to optimally identify carbohydrate binding sites over drug-like compound binding sites.

2.4. Parameterization and optimization of InCa-SiteFinder

To assess the performance of the InCa-SiteFinder a number of parameters were calculated. *Precision* is a measure of the correspondence of the predicted site and actual ligand volume. This measure was calculated to assess the predictive capacity of Q-SiteFinder [2]. It is calculated by taking the percentage of the volume of the predicted binding site that is occupied by the ligand atoms. The second parameter calculated was *coverage*, which is the percentage of the ligand atoms that are covered by the predicted site. Precision and coverage individually do not depict the actual success of the prediction method. Hence, to give a single parameter for performance assessment the precision was multiplied by coverage ($P \times C$) to obtain a single parameter, tau (τ). The ability of InCa-SiteFinder to predict carbohydrate binding sites was optimized and evaluated on a training dataset of 50 protein–carbohydrate complexes (none of these belonged to the SCOP superfamilies of the dataset used to derive the amino acid propensities) by 10-fold cross-validation (see supporting information for further details).

2.5. Determination of PSSBC cut-off for classifying a putative site

To determine the cut-off value of PSSBC, a test set of 45 protein–carbohydrate complexes having 45 carbohydrate binding sites was created (see supporting information). These complexes had zero SCOP superfamily level overlap with the 50 complexes used for optimizing the energy and amino acid propensity cut-offs. Using InCa-SiteFinder the top 30 putative ligand binding sites were predicted for each member of the dataset. These sites were ranked in decreasing order of their PSSBC values. The overall percentage success rate for the j^{th} ranked prediction was calculated by dividing the total number of correctly predicted true carbohydrate binding sites (NTCBS), at the j^{th} rank, by the total number of true carbohydrate binding sites present in the entire database (NTBS). The value of j is incremented by a factor of 1 each time to get a series of success rates for all of the 30 ranks. The success rate for rank, j, was defined as:

$$\text{Success rate}_j = 100\sum_{i=1}^{n}\frac{NTCBS_i}{NTBS} \quad (8)$$

where $NTCBS_i$ is the number of true carbohydrate binding sites correctly predicted for test set member i, at rank, j. n is the total number of protein–carbohydrate complexes and the NTBS is the number of total true carbohydrate binding sites in the dataset. In addition, an average value of PSSBC ($AvePSSBC_j$) for each rank, j, was similarly calculated as:

$$AvePSSBC_j = \frac{\sum_{i=1}^{n} PSSBC_k}{n} \quad (9)$$

where $PSSBC_k$ is defined in Eq. (5) and n is the total number of complexes in the dataset. A $PSSBC_k$ cut-off was determined (from the plots of the average value of PSSBC and success rate of predictions versus site ranking) such that none of the true carbohydrate binding sites scored less than the cut-off.

2.6. Determination of differential propensity score cut-off for carbohydrate binding site

A second test set of 40 protein–carbohydrate complexes and 40 complexes of protein–drug-like compounds was created (see supporting information). This dataset did not overlap with the training dataset. The values for the van der Waals energy of probe–protein interaction and the probe propensity score cut-offs of the training set were used to predict the top 30 sites for each member of the dataset. For each of the predicted sites a differential propensity score (DiffPS) was calculated. The success rate of the method was calculated according to Eq. (8). Average differential propensity score for the ranked sites were calculated as:

$$AveDiffPS_j = \frac{\sum_{i=1}^{n} DiffPS_i}{n} \quad (10)$$

where $DiffPS_i$ is defined in Eq. (7) and n is total number of complexes in the dataset. This was used to determine an effective cut-off value for DiffPS which allows differentiation of carbohydrate and drug-like compound binding sites.

2.7. Dataset for evaluation of the ability of the InCa-SiteFinder to distinguish between the carbohydrate binding sites and drug-like compound binding sites

A third independent dataset (see supplementary information) was prepared for the evaluation of the ability of the method to

classify the carbohydrate and drug-like compound binding sites. It comprised 40 protein–carbohydrate complexes and 40 protein–drug-like compound complexes. In order to be included in this dataset, members were not permitted to have SCOP superfamily representatives in the training or previously used test sets. Also no two members of this dataset were permitted to belong to the same SCOP superfamily. This dataset had two mutually exclusive subsets: (1) 40 drug-like compound binders which had 58 drug-like compound binding sites and (2) 40 carbohydrate binders which had 40 carbohydrate binding sites. For each of the protein–ligand complexes the top 30 sites were predicted and scored for PSSBC, PSSBD and DiffPS. On the basis of DiffPS the sites were predicted to be either carbohydrate binding or drug-like compound binding. The predictive capacity of InCa-SiteFinder was evaluated in terms of specificity and sensitivity (see supporting information).

3. Results and discussion

3.1. Amino acid propensity to interact with carbohydrate molecule

A number of statistical analyses were carried out to identify a property that has maximum potential for differentiating carbohydrate from other types of ligand binding site. This included secondary structure type, amino acid polarity, and amino acid propensity (see supporting information). The profile of amino acid propensities for occurrence in the carbohydrate binding site, calculated as a function of solvent accessible surface area, is very different from drug-like ligand binding sites (Fig. 2).

The propensities of arginine, aspartic acid, cystine, glutamic acid, isoleucine, leucine, lysine, methionine, phenylalanine and tryptophan showed maximum differentiation for carbohydrate and drug-like compound binding sites. For carbohydrate binding sites a high occurrence of arginine, lysine, glutamic and aspartic acid, along with the reduced presence of isoleucine, leucine, methionine, phenylalanine and cystine relative to drug-like sites is seen (Fig. 2). For both carbohydrate binding sites and drug-like binding sites tryptophan has the highest propensity. In a number of studies tryptophan has been identified as an important residue for the ligand binding [18].

3.2. Parameterization and evaluation of InCa-SiteFinder

3.2.1. Probe–protein interaction energy and grid point's $Score^{CBP}$ cut-off value determination and validation

The performance of InCa-SiteFinder was optimized by 10-fold cross-validation of carbohydrate ligand binding site precision and coverage, using τ (where $\tau = P \times C$) as the optimization metric (see Section 2). The optimization matrices calculated for identification of the best possible combination of protein–probe van der Waals interaction energy and the probe's carbohydrate binding propensity, are very similar. The optimal cut-off values are tabulated in Table 1 for the 10-fold cross-validation (see Figure II in supporting information). An average τ-matrix was calculated by taking the average of the 10 τ-matrices, giving a combination of van der Waals energy cut-off of −1.0 kcal/mol and propensity score cut-off of 0.25.

Table 1
The pairs of cut-offs for probe's propensity score and protein–probe van der Waals interaction which produced best τ values during 10-fold cross-validation.

Set no.	Probe propensity score	Protein–probe interaction energy (kcal/mol)
1	0.25	−1.0
2	0.25	−1.0
3	0.25	−1.1
4	0.375	−1.0
5	0.125	−1.0
6	0.25	−1.1
7	0.25	−1.0
8	0.25	−1.0
9	0.25	−1.0
10	0.25	−1.0

The cut-off values obtained in the optimization set were tested to predict the carbohydrate binding sites for the evaluation set members. The results are plotted in receiver operating characteristic curves (Fig. 3). The area under the dashed-curve (AUC) is 82%. Even though InCa-SiteFinder reaches a sensitivity of 86% with the specificity of 83%, 7 out of 50 carbohydrate binding sites could not be predicted. Visual examination of the structures revealed that in all of the 7 cases the carbohydrate interaction with the protein receptor was limited to just a few atoms of the ligand molecules which occupied peripheral regions on the protein surface with few atomic contacts. One example is depicted in Fig. 4. Such sites are difficult to predict and are of limited interest due to their small size and likely lack of functional importance. Functionally important sites generally occur in deep pockets with considerable coverage of the ligand molecule. If these peripheral binders are excluded from the test dataset the AUC increases from 82% to 96%. The method showed absolute sensitivity at 83% specificity (1-FPR; see supporting information). The ROC curve for the reduced test set is shown (solid curve) in Fig. 3.

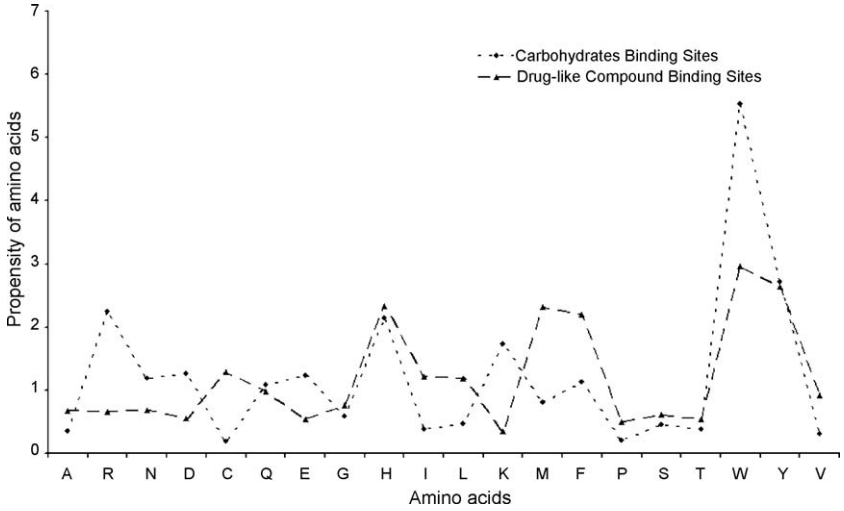

Fig. 2. The propensity of amino acids to occur in the binding site of carbohydrates and drug-like compounds.

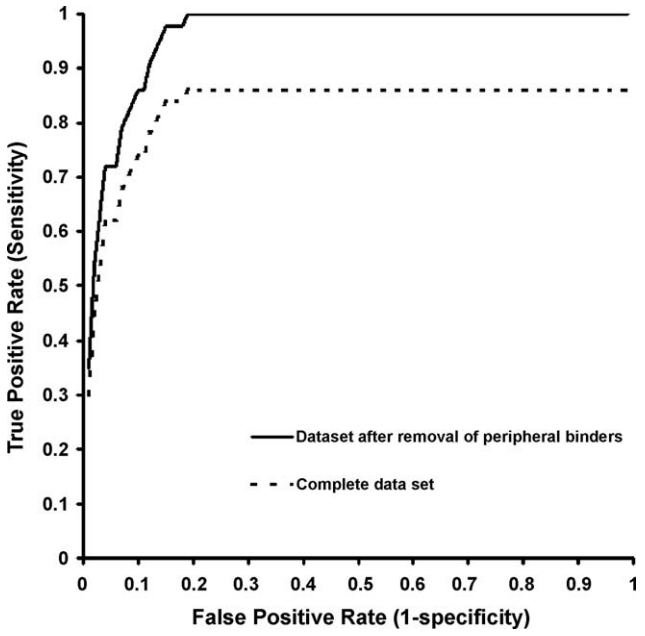

Fig. 3. Receiver operating characteristic curve illustrating the success of the carbohydrate binding site prediction. The dashed-curve represents the performance of the method over the entire test dataset. Solid curve represents the performance of the method after peripheral carbohydrate binders were removed from the dataset. These peripheral protein–carbohydrates had only marginal atomic interactions.

3.2.2. PSSBC threshold determination and validation

$PSSBC_i$ is the propensity score for probe cluster (putative ligand binding site) *i*, to bind carbohydrates. The cut-off values obtained for probe–protein van der Waals interaction energy and PSSBC were used for the prediction of 45 ligand binding sites for the 45 members of the second validation set (see Section 2). The predicted ligand binding sites were ranked in decreasing order of their PSSBC (calculated according to Eq. (5)) and these values were used as the basis for determining the PSSBC cut-off. The "ranked predicted ligand binding sites success rate" and the average score for each rank were calculated using Eqs. (8) and (9), respectively. A plot of these values for each rank is given in Fig. 5. The PSSBC cut-off was determined such that all of the true carbohydrate binding sites scored more than the cut-off. The success rate reaches the value of 100% when the average value of PSSBC for the predicted site is around 60. However, as the test dataset is small, a conservative cut-off value of 30 was chosen, to prevent losing any low PSSBC scoring carbohydrate binding sites.

3.2.3. DiffPS threshold determination and validation

The success rate (calculated according to Eq. (8)) in identifying the potential carbohydrate binding sites was plotted against the

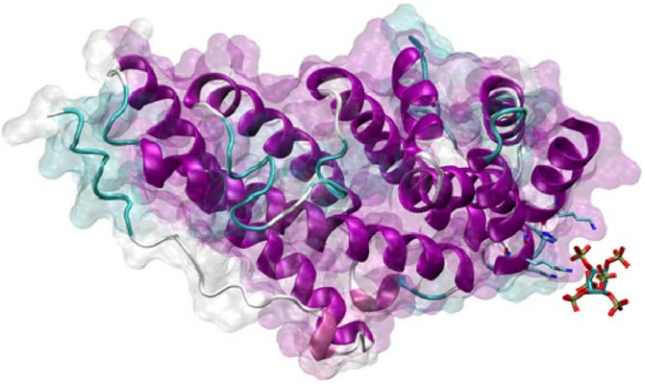

Fig. 4. Clathrin assembly protein in complex with inositol hexakisphosphate.

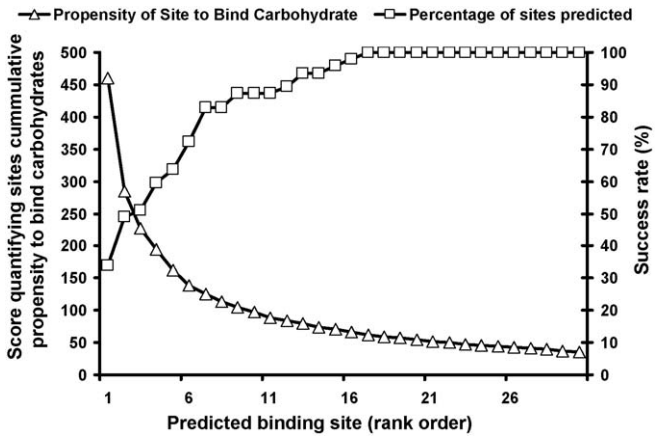

Fig. 5. Success rate (%) plotted along with average score (average PSSBC values) versus predicted binding site (ranked order).

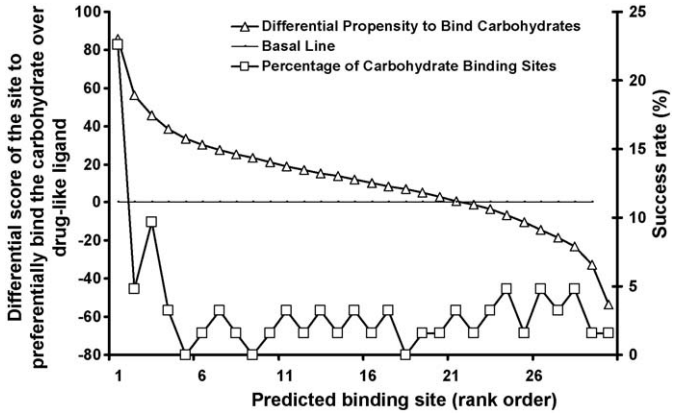

Fig. 6. Success rate (%) plotted along with DiffPS for the determination of its cut-off values for carbohydrate binding sites.

average DiffPS values (calculated according to Eq. (10)) for each of the top 30 sites (Figs. 6 and 7) ranked according to PSSBC (the propensity score of a site to bind carbohydrates). The carbohydrate binding sites have more positive DiffPS scores and the majority of the carbohydrate binding sites are concentrated in the top 4 ranks. "Drug-like compound" binding sites have more negative DiffPS scores and are concentrated in the last 4 ranks. In the middle region of the DiffPS range both carbohydrate and drug-like compound binding sites are present. From Figs. 6 and 7 the threshold values for classifying a ligand binding site as a carbohydrate binder or

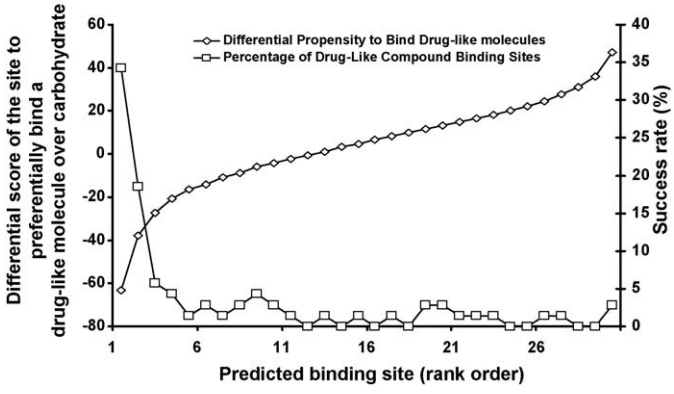

Fig. 7. Success rate (%) plotted along with DiffPS for the determination of its cut-off values for drug-like compound binding sites.

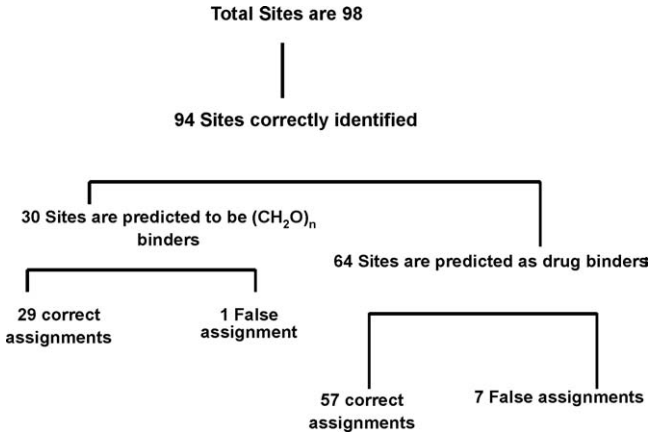

Fig. 8. Summary of InCa-SiteFinder evaluation.

drug-like compound binder were determined. A site having a DiffPS of more than 10 was considered to be purely a carbohydrate binding site. A site was considered to be a drug-like compound binding site if its DiffPS value was less than −20. These values have been adopted in the final testing of InCa-SiteFinder. Sites with DiffPS values between 10 and −20 can be considered to be of dual nature and able to interact with both types of ligands.

3.2.4. Final testing of InCa-SiteFinder

The proteins in the DiffPS validation dataset (see Section 2) included 58 drug-like sites and 40 carbohydrate binding sites. Out of these 98 ligand binding sites InCa-SiteFinder identified 30 as carbohydrate binding sites and 64 as drug-like compound binding sites. For the remaining 4 proteins (all carbohydrate binders) no site could be identified as a true carbohydrate binding site or a drug-like compound binding site. Among the 30 sites predicted to be carbohydrate binding sites 29 were true carbohydrate binding sites. The single remaining site was actually a drug-like compound binding site wrongly classified as a carbohydrate binding site. Among the 64 sites predicted to be drug-like compound binding sites 57 were true drug-like compound binding sites and 7 were actually carbohydrate binding sites wrongly predicted to be drug-like compound binding sites (Fig. 8). The overall specificity of the method was calculated to be 0.983 and the sensitivity was 0.725. The average volume of all of the sites predicted for the carbohydrates on the surface of the test set was 143 $Å^3$. The average volume of the correctly predicted sites was 920 $Å^3$.

Two examples of the correct predictions are shown in Fig. 9. These sites show the variation in prediction. The precision of predicted carbohydrate binding sites depends upon the ligand and binding site-character. Some of the predicted sites have high precision and high coverage (Fig. 9a) whereas in other cases the site may have higher coverage of the ligand but with less precision (Fig. 9b). Sites with high precision and low coverage are generally smaller predicted sites that occupy only part of ligand binding site. Such sites are not considered by InCa-SiteFinder to be ligand binding sites and so are removed because their PSSBC value is less than the threshold of the carbohydrate propensity score for a site (see Section 2).

4. Conclusions

We have presented a method, InCa-SiteFinder, for the identification of carbohydrate binding sites by first locating the energetically favorable pockets followed by identifying the regions with high cumulative propensity for binding carbohydrates. It is able to correctly predict carbohydrate binding sites as seen in the ROC plots (Fig. 3). This ability can be attributed to the combination of using van der Waals energy of protein–probe interaction developed in the Q-SiteFinder method [2] combined with the ability of amino acid propensity to correctly distinguish carbohydrate binding sites. Carbohydrate binding sites have been shown to be rich in aromatic residues like tryptophan [18]. These residues are thought to form CH/pi interactions with the carbohydrates by orienting their planar surface for the stacking arrangement [19]. Also as noted by Taroni et al. [8] an increased occurrence of residues like arginine, aspartic acid, and glutamic acid, give the site a greater potential for making bidentate hydrogen bonds with adjacent hydroxyls on the sugar. There is also a relative reduced presence of residues like glycine, leucine, isoleucine and cysteine with no sidechain hydrogen bonding capacity. The van der Waals energy alone cannot discriminate between different types of ligands, hence, the use of the propensity scores in combination with protein–probe interaction energetic criteria yields better results. The method is also able to distinguish the carbohydrate binding sites from drug-like compound binding sites with very high specificity (0.983) whilst retaining high sensitivity (0.725).

The value of the differential propensity (DiffPS) score to distinguish the preference of the predicted site for carbohydrate over drug-like compound ligands is a key factor in the success of InCa-SiteFinder. Sites with high positive values are almost always carbohydrate binders. Conversely, the sites with greater negative values are mostly drug-like compound binding sites. This is valuable information for identifying the carbohydrate binding sites that do not bind drugs-like molecules. More remarkable still is the identification of drug-like binding sites with greater success than the identification of carbohydrate binding sites. Though our aim was the prediction of carbohydrate binding sites, we have noted the potential use of InCa-SiteFinder in identification of drug-like binding sites. However, we have not attempted to optimize the method for this purpose. Sites with dual propensity to bind carbohydrate and drug-like compounds have DiffPS values between 10 and −20. The tool developed here may form the basis for a method that could not only discriminate between different types of functional site, but also

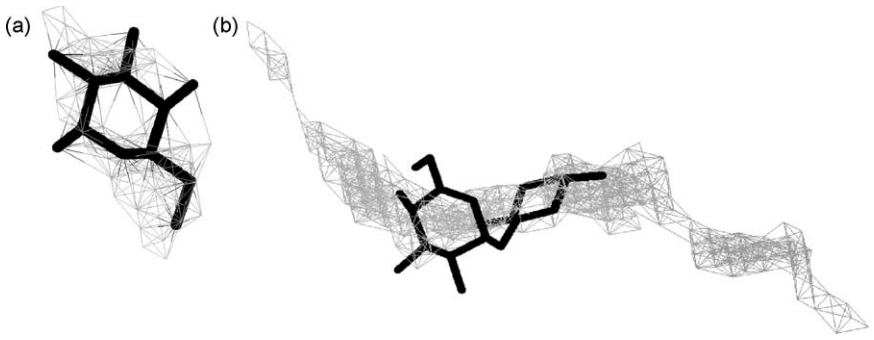

Fig. 9. Galactose molecule (5abp) covered by the predicted site (a) and dideoxy-4-amino glucopyranoside from 1hx0 inside the predicted site (b).

facilitate the process of structure-based drug design. In this later case an ability to characterize sites that are amenable to binding drug-like molecules would be of great interest for medicinal applications, including blocking protein–protein interactions and for design of competitive inhibitors.

Acknowledgements

MK carried out the work whilst working as a visiting research student at the University of Leeds. MK was funded by the IMPRS-CB and Prof. Roger S. Goody in the form of a PhD studentship. SJB was funded by a BBSRC masters studentship.

Appendix A. Supplementary data

Supplementary data associated with this article can be found, in the online version, at doi:10.1016/j.jmgm.2009.08.009.

References

[1] M. Tagari, et al., E-MSD: improving data deposition and structure quality, Nucleic Acids Res. 34 (2006) D287–D290.
[2] A.T.R. Laurie, R.M. Jackson, Q-SiteFinder, an energy-based method for the prediction of protein–ligand binding sites, Bioinformatics 21 (9) (2005) 1908–1916.
[3] R.A. Laskowski, J.D. Watson, J.M. Thornton, ProFunc: a server for predicting protein function from 3D structure, Nucleic Acids Res. 33 (Web Server issue) (2005) W89–W93.
[4] N.J. Burgoyne, R.M. Jackson, Predicting protein interaction sites: binding hot-spots in protein–protein and protein–ligand interfaces, Bioinformatics 22 (11) (2006) 1335–1342.
[5] B.K. Brandley, R.L. Schnaar, Cell-surface carbohydrates in cell recognition and response, J. Leukoc. Biol. 40 (1) (1986) 97–111.
[6] C.R. Bertozzi, L.L. Kiessling, Chemical glycobiology, Science 291 (5512) (2001) 2357–2364.
[7] W.I. Weis, K. Drickamer, Structural basis of lectin–carbohydrate recognition, Annu. Rev. Biochem. 65 (1996) 441–473.
[8] C. Taroni, S. Jones, J.M. Thornton, Analysis and prediction of carbohydrate binding sites, Protein Eng. 13 (2) (2000) 89–98.
[9] C. Shionyu-Mitsuyama, et al., An empirical approach for structure-based prediction of carbohydrate-binding sites on proteins, Protein Eng. 16 (7) (2003) 467–478.
[10] A. Malik, S. Ahmad, Sequence and structural features of carbohydrate binding in proteins and assessment of predictability using a neural network, BMC Struct. Biol. 7 (2007) 1.
[11] R.A. Laskowski, PDBsum: summaries and analyses of PDB structures, Nucleic Acids Res. 29 (1) (2001) 221–222.
[12] R.A. Laskowski, V.V. Chistyakov, J.M. Thornton, PDBsum more: new summaries and analyses of the known 3D structures of proteins and nucleic acids, Nucleic Acids Res. 33 (Database issue) (2005) D266–D268.
[13] R.A. Laskowski, et al., PDBsum: a Web-based database of summaries and analyses of all PDB structures, Trends Biochem. Sci. 22 (12) (1997) 488–490.
[14] H.M. Berman, et al., The protein data bank, Nucleic Acids Res. 28 (1) (2000) 235–242.
[15] A.G. Murzin, et al., SCOP: a structural classification of proteins database for the investigation of sequences and structures, J. Mol. Biol. 247 (4) (1995) 536–540.
[16] A.K. Ghose, V.N. Viswanadhan, J.J. Wendoloski, A knowledge based approach in designing combinatorial and medicinal chemistry libraries for drug discovery. 1. Qualitative and quantitative definitions of a drug like molecule, Abstr. Pap. Am. Chem. Soc. 217 (1999) U708–U1708.
[17] R.M. Jackson, Q-fit: a probabilistic method for docking molecular fragments by sampling low energy conformational space, J. Comput. Aided Mol. Des. 16 (1) (2002) 43–57.
[18] S. Gao, et al., Effect of amino acid residue and oligosaccharide chain chemical modifications on spectral and hemagglutinating activity of *Millettia dielsiana* Harms. ex Diels. lectin, Acta Biochim. Biophys. Sin. (Shanghai) 37 (1) (2005) 47–54.
[19] H. Petrokova, et al., Crystallization and preliminary X-ray diffraction analysis of cold-active beta-galactosidase from *Arthrobacter* sp. C2-2, Collect. Czech. Chem. Commun. 70 (1) (2005) 124–132.