BIOINFORMATICS APPLICATIONS NOTE Vol. 30 no. 23 2014, pages 3417–3418
doi:10.1093/bioinformatics/btu559

Structural bioinformatics

Advance Access publication August 20, 2014

GlycoPattern: a web platform for glycan array mining

Sanjay B. Agravat[1,2,*], Joel H. Saltz[2,3], Richard D. Cummings[1] and David F. Smith[1]

[1]National Center For Functional Glycomics, Emory University School of Medicine, Atlanta, GA 30322, USA, [2]Department of Mathematics and Computer Science, Emory University, Atlanta, GA 30322 and [3]Department of Biomedical Informatics, Stony Brook University, Stony Brook, NY 11794, USA

Associate Editor: Anna Tramontano

ABSTRACT

Summary: GlycoPattern is Web-based bioinformatics resource to support the analysis of glycan array data for the Consortium for Functional Glycomics. This resource includes algorithms and tools to discover structural motifs, a heatmap visualization to compare multiple experiments, hierarchical clustering of Glycan Binding Proteins with respect to their binding motifs and a structural search feature on the experimental data.

Availability and implementation: GlycoPattern is freely available on the Web at http://glycopattern.emory.edu with all major browsers supported.

Contact: sanjay.agravat@emory.edu

Received on March 1, 2014; revised on July 17, 2014; accepted on August 14, 2014

1 INTRODUCTION

Glycans form complex structures since the 10 monosaccharides found in animal glycans (Varki and Sharon, 2009) are coupled in two possible anomeric forms (α or β) and multiple glycosidic linkages and include branched sequences unlike nucleic acids and proteins. The number of glycans in the human glycome is unknown but estimated to be >7000 (Cummings, 2009). The interactions of glycans with glycan-binding proteins (GBP) are important in many biological functions, including cell adhesion, signaling and innate immunity, and most pathogens invade mammalian cells and tissues via initial protein–glycan interactions. Printed microarrays of defined glycan structures are interrogated with GBP detected by fluorescence to determine bound glycans. Knowing structures of glycans and the glycan determinants within glycans that are bound or not bound by a GBP provides valuable information about protein interaction and specificity.

GlycoPattern is a publicly available Web-based resource that allows investigators to analyze microarray data with the ability to discover motifs, cluster GBPs based on binding affinity to defined determinants, search the array for glycans or substructure using a text-based search, and perform a heatmap comparison on glycan array expression data from different experiments. GlycanMotifMiner (Cholleti *et al.*, 2012) is a frequent subtree mining algorithm for motif discovery without using predefined motifs that we incorporated into GlycoPattern with other features to help with the analysis, visualization and searching of glycan array data. Campbell *et al.* (2014) describe several glycomics databases and glycoinformatics tools, including Glycosciences.de, UnicarbKB, GlycomeDB and Resource for INformatics of Glycomes at Soka (RINGS). The Consortium for Functional Glycomics (CFG) Database (Raman *et al.*, 2006) mentioned in the review is the only resource we are aware of that contains Glycan Array data (among other types of data) but it is not considered a platform for mining glycan array data. RINGS (Akune *et al.*, 2010) is a Web portal providing algorithmic and data mining tools to aid glycobiology research that includes tools for drawing structures, mining glycan subtrees, pathway prediction and glycan profiles, but is limited in the tools available for mining across multiple glycan array experiments. GlycoPattern was developed to provide broader mining of glycan microarray data and to discover the specificity of GBP.

2 METHODS

The current version of GlycoPattern supports only the CFG glycan array, which reports GBP binding in relative fluorescence units (RFU) to defined glycans. Users can copy and paste the average RFU values corresponding to glycans from their experiments into GlycoPattern. The data are sorted according to glycan number as specified on the Mammalian Printed Arrays v4.0 through v5.1 (http://www.functionalglycomics.org/static/consortium/resources/resourcecoreh8.shtml). The application was developed using Python, HTML, Javascript, MySQL and the Pylons Web framework.

2.1 Glycan notation

Many notations representing textual glycan nomenclature are available, including International Union of Pure and Applied Chemistry (IUPAC) (McNaught, 1997), Linear Notation for Unique description of Carbohydrate Structures (LINUCS; Bohne-Lang *et al.*, 2001), Kyoto Encyclopedia of Genes and Genomes (KEGG) Chemical Function (Hattori *et al.*, 2003), LinearCode (Ehud *et al.*, 2002), GLYDE-II (Packer *et al.*, 2008) and GlycoCT (Herget *et al.*, 2008). We chose to use the modified IUPAC condensed nomenclature from the CFG because of the wide adoption of IUPAC in the Glycobiology community and the inclusion of the anomeric carbon. For structure searches, GlycoPattern can accept and display the modified IUPAC nomenclature for oligosaccharides. Glycan structures are also displayed in symbols (Varki and Sharon, 2009), and we have developed Javascript code to dynamically convert the modified IUPAC nomenclature into the symbolic representation using the HTML Canvas standard (Cabanier, 2014) without requiring the need to store an image file on disk.

*To whom correspondence should be addressed.

2.2 Motif discovery

A glycan array is interrogated with a GBP to determine the relative binding strengths of all the glycans on the array (Smith *et al*., 2010). The GlycanMotifMiner algorithm (Cholleti *et al*., 2012) functions by automatically finding frequently occurring patterns in binding glycans. The algorithm then incrementally discovers motifs of larger size and stops when it cannot find any motifs of next higher size. The algorithm uses a threshold for the minimum number of glycans that contain the motif along with a threshold for the maximum number of non-binding glycans that contain the motif. Motifs are discovered and reported automatically after data entry.

2.3 GBP-Motif hierarchical clustering

The protein–carbohydrate resource of the CFG has screened hundreds of GBPs to determine specificity for a glycan motif or determinant. GlycoPattern performs a hierarchical clustering from the results of GlycanMotifMiner on any number of experiments to generate a dendrogram showing the similarity of different GBPs for the determinant(s) they recognize.

The GBP Dendrogram is useful for selecting GBPs to do more fine-grained analysis to determine the similarities among related GBPs, and to rapidly screen a large database to identify reagents to detect specific glycans.

2.4 GBP–glycan heatmap

The GBP–glycan heatmap feature allows the user to select two or more experiments (including experiments from different CFG Array versions) and perform a side-by-side comparison of the glycans bound by any number of GBPs. GlycoPattern allows the user to hover over a cell in the Heatmap and display the glycan structure in symbolic representation. This feature allows the user to determine which features of a glycan structure are recognized by highly similar GBPs selected using the GBP-motif hierarchical clustering feature.

2.5 Glycan search

GlycoPattern allows users to search for glycan substructures within an experiment. The user interface requires the use of the modified IUPAC condensed nomenclature. The search supports hierarchical queries that represent the branching in certain glycan sequences where branching is noted with an opening and closing parenthesis around a sequence. The search function parses the modified IUPAC condensed nomenclature and converts it into Javascript Object Notation (JSON) format. When searching for a branched input sequence, the search function finds any match within a glycan structure for the input sequence rather than an exact match. This is useful when dealing with sequences that have more than two branch points from a single monosaccharide.

3 RESULTS

GlycoPattern is one of the few publicly available informatics resources available for mining Glycan Array data. At the time of writing, there are >50 international registered users that are not affiliated with Emory University with >400 experiments total. It was originally developed as a resource for the CFG, but we plan to expand its capabilities beyond the CFG to make it more accessible to the entire Glycobiology community. In our planned future work, GlycoPattern will be adapted to handle any defined glycan library using a format such as Glyde-II or CabosML (Kikuchi *et al*., 2005) to convert the structures into a GlycoPattern-readable format. Some of the authors are members of the joint international consortium for the Minimum Information Required for a Glycomics Experiment (York *et al*., 2014) and are actively working on developing standards to represent glycan microarrays similar to the efforts of the Minimum Information Required for a Microarray Experiment (Spellman *et al*., 2002) and (Rayner *et al*., 2006). Once standards are adopted, sharing, querying and analysis of data can be facilitated and will eventually lead to more maturity in the development of bioinformatics tools for the glycobiology community.

Funding: This work was supported by grants from the National Institutes of Health including U53 GM62116 and GM98791 (Consortium for Functional Glycomics), R01GM085447 (D.F.S.), and P41GM103694 (R.D.C.).

Conflict of interest: none declared.

REFERENCES

Akune,Y. *et al*. (2010) The RINGS resource for glycome informatics analysis and data mining on the web. *Omics*, **14**, 475–486.

Bohne-Lang,A. *et al*. (2001) LINUCS: linear notation for unique description of carbohydrate sequences. *Carbohydr. Res*., **336**, 1–11.

Cabanier,R. *et al*. (eds) (2014) HTML canvas 2D context. In: *W3C Candidate Recommendation (work in progress)* 21 August 2014, http://www.w3.org/TR/2014/CR-2dcontext-20140821/. Latest version available at http://www.w3.org/TR/2dcontext/

Campbell,M.P. *et al*. (2014) Toolboxes for a standardised and systematic study of glycans. *BMC Bioinformatics*, **15** (**Suppl. 1**), S9.

Cholleti,S.R. *et al*. (2012) Automated motif discovery from glycan array data. *Omics*, **16**, 497–512.

Cummings,R.D. (2009) The repertoire of glycan determinants in the human glycome. *Mol. Biosyst*., **5**, 1087–1104.

Ehud,B. *et al*. (2002) A novel linear code nomenclature for complex carbohydrates. *Trends Glycosci. Glycotechnol*., **14**, 127–137.

Hattori,M. *et al*. (2003) Development of a chemical structure comparison method for integrated analysis of chemical and genomic information in the metabolic pathways. *J. Am. Chem. Soc*., **125**, 11853–11865.

Herget,S. *et al*. (2008) GlycoCT-a unifying sequence format for carbohydrates. *Carbohydr. Res*., **343**, 2162–2171.

Kikuchi,N. *et al*. (2005) The carbohydrate sequence markup language (CabosML): an XML description of carbohydrate structures. *Bioinformatics*, **21**, 1717–1718.

McNaught,A.D. (1997) Nomenclature of carbohydrates (recommendations 1996). *Adv. Carbohydr. Chem. Biochem*., **52**, 43–177.

Packer,N.H. *et al*. (2008) Frontiers in glycomics: bioinformatics and biomarkers in disease. An NIH white paper prepared from discussions by the focus groups at a workshop on the NIH campus, Bethesda MD (September 11-13, 2006). *Proteomics*, **8**, 8–20.

Raman,R. *et al*. (2006) Advancing glycomics: implementation strategies at the consortium for functional glycomics. *Glycobiology*, **16**, 82R–90R.

Rayner,T.F. *et al*. (2006) A simple spreadsheet-based, MIAME-supportive format for microarray data: MAGE-TAB. *BMC Bioinformatics*, **7**, 489.

Smith,D.F. *et al*. (2010) Use of glycan microarrays to explore specificity of glycan-binding proteins. *Methods Enzymol. Glycobiol*., **480**, 417–444.

Spellman,P.T. *et al*. (2002) Design and implementation of microarray gene expression markup language (MAGE-ML). *Genome Biol*., **3**, RESEARCH0046.

Varki,A. and Sharon,N. (2009) Historical Background and Overview. In: Varki,A. *et al*. (eds) *Essentials of Glycobiology*. 2nd edn. Cold Spring Harbor Laboratory Press, Cold Spring Harbor, NY.

York,W.S. *et al*. (2014) MIRAGE: the minimum information required for a glycomics experiment. *Glycobiology*, **24**, 402–406.