Introduction to Glycoinformatics and Computational Applications

Kiyoko F. Aoki-Kinoshita

Department of Bioinformatics, Faculty of Engineering, Soka University,
1 – 236 Tangi-machi, Hachioji, Tokyo, 192 – 8577 Japan

E-Mail: kkiyoko@soka.ac.jp

Received: 11th December 2011/Published: 11th July 2012

Abstract

Glycoinformatics involves the development of computational methods and tools for the understanding of glycan function, including the development of databases, which use these methods and tools for validation of results. Thus several databases have been developed, storing a plethora of data from various analytical perspectives. These invaluable resources provide the data which theoretical computer scientists and data mining experts can use to develop new models and tools. This paper will describe the various carbohydrate-related databases that have become rather stable in this field, along with some of the methods that have been developed for analysing the glycan data from various perspectives. In particular, we focus on glycan biomarker and glycan binding pattern prediction.

Introduction

Glycoinformatics deals with the development of computational methods and tools as well as the construction of databases in order to gain insight into the functions of glycoconjugates and complex carbohydrates. As such, the glycoinformatics field has taken off with the development of various glycan structure databases and related information, as well as theoretical methods and practical tools for analysing glycans. In particular, with the development of major carbohydrate resources in Japan, the U.S. and Europe, it has become easier to develop and test these methods for practical applications. Here, we describe in detail some of these data resources and methods for beginners to get an idea of this newly emerging field of glycoinformatics.

Glycan Databases

The major databases that may be considered rather stable at the time of this writing are listed in Table 1. The GlycomeDB database [1] may be considered the main resource for glycan structure information, which is linked with several other glycan structure resources. Glycan binding data can be obtained from the CFG (Consortium for Functional Glycomics) [2] and JCGGDB (Japanese Consortium for Glycobiology and Glycotechnology Databases) which includes CabosDB [3]. Glycan profile data can be obtained from the CFG and EuroCarbDB [4], although the latter is still in the developmental stages. UniCarbKB [5] is also based on the EuroCarbDB infrastructure, so this resource will also provide profiling data once it is made available.

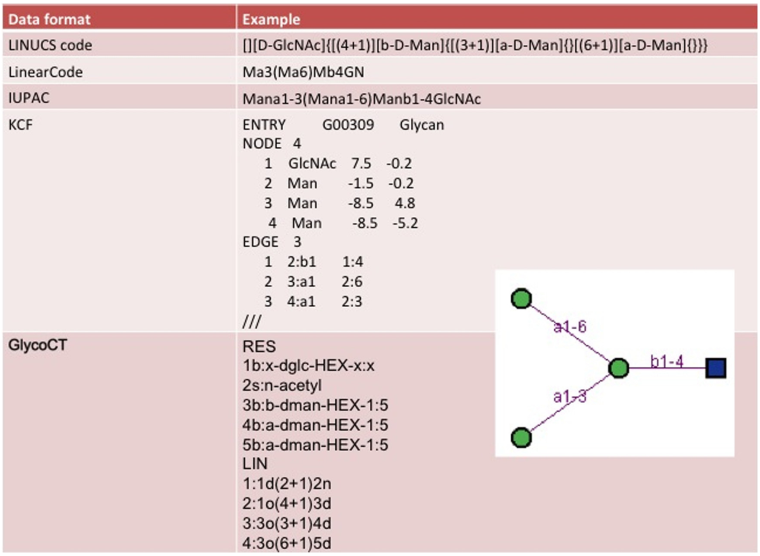
Table 1. Major glyco-databases.

Database	Description	URL
GlycomeDB	Portal for glycan structures that have been integrated from several of the major glycan-related databases	http://www.glycome-db.org
GLYCOSCIENCES.de	One of the earliest databases of glycan structure data, mainly extracted from the PDB (Protein DataBank). In addition to structural data, also includes NMR data and literature references.	http://www.glycosciences.de/
Consortium for Functional Glycomics (CFG)	A major database containing not only glycan structures, but also glycan binding affinity data from their glycan arrays, glycan profiling data from MALDI-TOF analysis of various samples, knock-out mouse phenotype data and glyco-enzyme expression data.	http://www.functionalglycomics.org
Japanese Consortium for Glycobiology and Glycotechnology Database (JCGGDB)	A comprehensive database portal for major glyco-related databases in Japan, including mass spectral data of glycan profiles, lectin array data, glycoprotein data, glycogene information including disease information, etc.	http://jcggdb.jp
KEGG GLYCAN	As one of the databases of the KEGG resource, glycan structures and their pathway data, including glycogene information as organized by the KEGG ORTHOLOGY can be obtained.	http://www.genome.jp/kegg/glycan/
EuroCarbDB	A database infrastructure for a distributed database of experimental data for glycobiology, including (tandem) MS, HPLC and NMR data and the related glycan structures.	http://www.eurocarbdb.org http://www.ebi.ac.uk/eurocarb/home.action
UniCarbKB	A new database of experimental data for glycobiology, linked with glyco-related UniProtKB data.	http://unicarb-db.biomedicine.gu.se/unicarbkb

Glyco-resources refer to those web sites that provide computational tools for glyco-analysis. The database GLYCOSCIENCES.de [6] provides such tools as pdb-care for checking PDB files for accurate carbohydrate information, CARP for generating Ramachandran plots of glycosidic linkages, GlyVicinity and GlySeq for analysing statistical information around glycosylation sites and carbohydrate residues, and several glycan modelling tools based on molecular dynamics (MD) and molecular mechanics (MM) in addition to glycosylation prediction tools. GLYCAM Web [7] is another web-based tool for generating structural conformations of glycan structures, based on the GLYCAM force field parameters. Finally, RINGS [8] is a web-based resource providing a variety of tools for data mining glycan structures.

Glycan Data Formats

One issue in working with the various carbohydrate databases available is the handling of different glycan structure formats. An example of some of the major formats for a given glycan structure is illustrated in Figure 1. Each of these formats is used in different databases, and since there was no coordination between these database projects previously, different formats emerged to satisfy different requirements [9]. Even though a consensus has been obtained for GLYDE-II to be used as the standard format for carbohydrate data exchange between databases [10], it is not necessarily user-friendly or human-readable. Thus, some tools for converting between different glycan structure formats have been developed in RINGS, which provides utilities to convert between IUPAC, GlycoCT, Linear Code®, LINUCS and KCF. However, collaborations have emerged to link data between different databases based on GLYDE-II, and resources have started to either provide tools for converting between different formats or allowing different formats for input. Consequently, there is less concern now for which format to use to store glycan structure data.



Data format	Example
LINUCS code	[][D-GlcNAc]{[(4+1)][b-D-Man]{[(3+1)][a-D-Man]{}[(6+1)][a-D-Man]{}}}
LinearCode	Ma3(Ma6)Mb4GN
IUPAC	Mana1-3(Mana1-6)Manb1-4GlcNAc
KCF	ENTRY G00309 Glycan NODE 4 1 GlcNAc 7.5 -0.2 2 Man -1.5 -0.2 3 Man -8.5 4.8 4 Man -8.5 -5.2 EDGE 3 1 2:b1 1:4 2 3:a1 2:6 3 4:a1 2:3 ///
GlycoCT	RES 1b:x-dglc-HEX-x:x 2s:n-acetyl 3b:b-dman-HEX-1:5 4b:a-dman-HEX-1:5 5b:a-dman-HEX-1:5 LIN 1:1d(2+1)2n 2:1o(4+1)3d 3:3o(3+1)4d 4:3o(6+1)5d

Figure 1. Illustration of various glycan structure formats for the same glycan structure (illustrated). While many formats are linear sequences, some like KCF and GlycoCT use graph-like notations across multiple lines.

Computational Applications

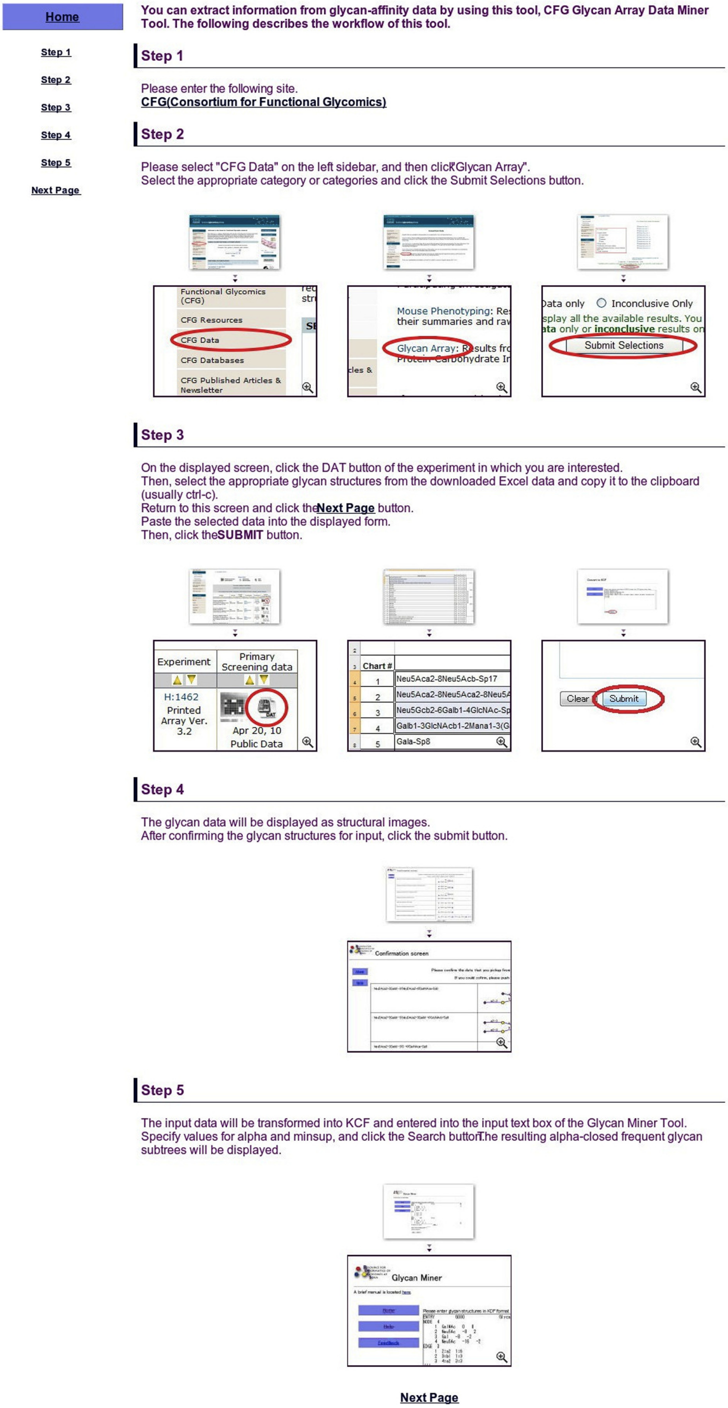
Two of the major biological problems facing glycan structure analysis involve the discovery of glycan biomarkers and the understanding of glycan recognition patterns. The former can be obtained from the glycan profiles that are being generated by various databases, and the latter can be analysed from glycan binding and structural data.

Glycan profiles provide different sets of numerous glycan structures that are claimed to exist in particular cells or tissues. At times, there may be hundreds of glycan structures in a particular sample. It is understood that these glycan structures would consist of partially synthesized structures and portions of glycoproteins and glycolipids, these profiles may be considered snapshots of a particular sample. By comparing different snapshots of control versus target tissues, one may infer potential glycan biomarkers. In fact, classification methods from the machine learning field have been applied to predict such glycan biomarkers in the past [11 – 14].

However, another approach for the analysis of glycan profiles by computing is called the *frequent subtrees* that are found within the data set. There are several frequent subtree mining algorithms, but the one that has been applied to glycans is called the α-closed frequent subtree algorithm [15]. A parameter α is specified in order to provide the user with some leeway in computing the results. This method has been implemented as a web-based tool called Glycan Miner in RINGS, taking as input a set of glycan structures in any format in addition to the α and support parameter to specify frequentness.

The frequent subtree algorithm can also be applied to glycan binding data, to predict the glycan substructures that are mostly involved in binding, by specifying those glycan structures that bind strongly with a particular glycan binding protein (GBP). In fact, the Glycan Miner tool comes with a useful helper application for those using the CFG's glycan array data, as shown in Figure 2. Snapshots of the CFG web page and indications of where to retrieve the necessary data are illustrated. Additionally, the conversion tool is linked to the Glycan Miner Tool such that the converted data is automatically inputted and ready to be computed with the specified parameters.

In contrast, another method to predict glycan binding patterns of GBPs is a probabilistic model that computes disconnected patterns from within glycan data. This model was developed due to the understanding that some GBPs such as lectins (and Siglecs in particular) may recognize not only the residues at the terminal end, but also non-terminal residues further in the chain. Thus, probabilistic sibling-dependent tree Markov models called PSTMM [16] and OTMM (ordered tree Markov model) [17] were developed to capture such potential glycan binding patterns.



You can extract information from glycan-affinity data by using this tool, CFG Glycan Array Data Miner Tool. The following describes the workflow of this tool.

Step 1

Please enter the following site.
CFG(Consortium for Functional Glycomics)

Step 2

Please select "CFG Data" on the left sidebar, and then click"Glycan Array".
Select the appropriate category or categories and click the Submit Selections button.

Functional Glycomics (CFG)
CFG Resources
CFG Data
CFG Databases
CFG Published Articles & Newsletter
Mouse Phenotyping: Re
their summaries and raw
Glycan Array: Results fro
Protein-Carbohydrate In
Data only
○ Inconclusive Only
splay all the available results. You
ata only or inconclusive results on
Submit Selections

Step 3

On the displayed screen, click the DAT button of the experiment in which you are interested.
Then, select the appropriate glycan structures from the downloaded Excel data and copy it to the clipboard (usually ctrl-c).
Return to this screen and click the**Next Page** button.
Paste the selected data into the displayed form.
Then, click the**SUBMIT** button.

Step 4

The glycan data will be displayed as structural images.
After confirming the glycan structures for input, click the submit button.

Step 5

The input data will be transformed into KCF and entered into the input text box of the Glycan Miner Tool.
Specify values for alpha and minsup, and click the Search button.The resulting alpha-closed frequent glycan subtrees will be displayed.

Figure 2. Snapshots of the CFG web page and indications of where to retrieve the necessary data are illustrated. Additionally, the conversion tool is linked to the Glycan Miner Tool such that the converted data is automatically inputted and ready to be computed with the specified parameters.

However, in order to allow general glycobiologists to use these models a method to extract the glycan patterns directly from the models was required. This was implemented in what was called the Profile PSTMM model [18], which is currently available as a tool in RINGS. The difficulty in extracting these patterns using Profile PSTMM came in the fact that the shape of the pattern that should be extracted must be specified beforehand. This is similar to the fact that the number of amino acids that must be extracted from Profile HMMs must be first specified when training the model. However, this is a bigger challenge for the branched structures of glycans. This difficulty is illustrated in Figure 3 in comparison with amino acid sequence profiles. Whereas the size (length) of amino acid sequence profiles can be specified directly based on the lengths of the input sequences, it is more difficult to specify the branched shape of a profile from a set of branched glycan structures. This is especially difficult to do automatically without any user input.

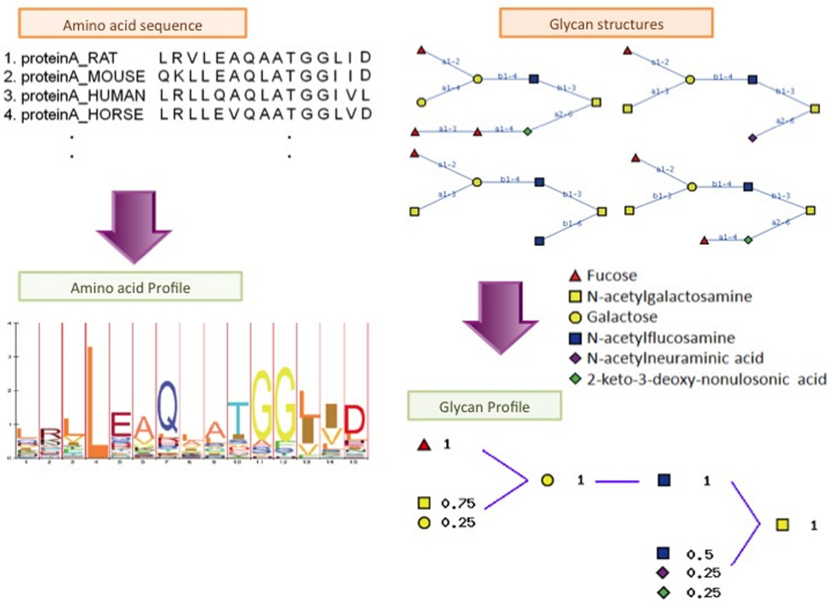


Figure 3. An illustration comparing amino acid sequence profiles and glycan structure profiles (patterns). Whereas the size (length) of amino acid sequence profiles can be specified from the dataset beforehand, it is more difficult to specify the branched shape of a profile from a set of branched glycan structures.

At first, the Maximum Common Subtree of all of the glycans were computed and used automatically as the default shape. However, if the input contained even one small structure, then the resulting glycan pattern would become very small. Thus, in order to enable larger sized patterns, the concept of aligning the input glycan structures and extracting a portion of the glycans that overlap the majority of times was devised. Multiple sequence alignment is actually considered a difficult problem even for linear sequences, so a heuristic was developed based on the popular multiple protein sequence alignment method ClustalW. This new method, called MCAW, for multiple carbohydrate alignment with weights, was then developed as a tool in RINGS. This tool is actually the basis of a number of potential analytical tasks, including the computation of glycan score matrices [19] for assessing glycosidic linkage similarity, in addition to the original intention of aiding probabilistic models in determining the glycan pattern to learn.

Thus, based on such fundamental computational methods, it may become possible to apply scoring functions to indicate the "similarity" of monosaccharides and glycosidic linkages. Such methods may also be applicable to spatial and structural analyses of glycans compared to protein structures, and even different patterns between organisms and species may be considered for future analysis.

Conclusion

In addition to the analysis of glycan structures, glycoinformatics also involves the understanding of glycan biosynthesis pathways, the functions of glycogenes and systems analysis of glyco-related pathways [20]. There is also a lack of glycoprotein data, but once a sufficient amount of such data is accumulated, more complex methods for analysing glycosylation and their functions may be developed. Additionally, another difficulty is the very large area of conformational analysis of glycans. Overall, it is probably safe to assume that this new area of glycoinformatics is sure to continue to progress in the years to come.

References

[1] Ranzinger, R., Herget, S., von der Lieth, C.-W., Frank, M. (2011) GlycomeDB – a unified database for carbohydrate structures. *Nucl. Acids Res.* **39** (Suppl 1):D373-D376.
doi: http://dx.doi.org/10.1093/nar/gkq1014.

[2] Rillahan, C.D., Paulson, J.C. (2011) Glycan microarrays for decoding the glycome. *Annu. Rev. Biochem.* **80:**797–823, 2011.
doi: http://dx.doi.org/10.1146/annurev-biochem-061809-152236.

[3] Kikuchi, N., Taniguchi, N., Suzuki, A., Ito, Y., Narimatsu, H. CabosDB: Carbohydrate Sequencing Database (Editors Kawasaki., T., Hase, S.). *Experimental Glycoscience.* Springer Japan, 426–428, 2008.

[4] von der Lieth, C.-W., Freire, A.A., Blank, D., Campbell, M.P., Ceroni, A., Damerell, D.R., Dell, A., Dwek, R.A., Ernst, B., Fogh, R., Frank, M., Geyer, H., Geyer, R., Harrison, M.J., Henrick, K., Herget, S., Hull, W.E., Ionides, J., Joshi, H.J., Kamerling, J.P., Leeflang, B.R., Lütteke, T., Lundborg, M., Maass, K., Merry, A., Ranzin-

ger, R., Rosen, J., Royle, L., Rudd, P.M., Schloissnig, S., Stenutz, R., Vranken, W.F., Widmalm, G., Haslam, S.M. (2011) EUROCarbDB: An open-access platform for glycoinformatics. *Glycobiology* **21**(4):493 – 502.
doi: http://dx.doi.org/10.1093/glycob/cwq188.

[5] Campbell, M.P., Hayes, C.A., Struwe, W.B., Wilkins, M.R., Aoki-Kinoshita, K.F., Harvey, D.J., Rudd, P.M., Kolarich, D., Lisacek, F., Karlsson, N.G., Packer, N.H. (2011) UniCarbKB: putting the pieces together for glycomics research. *Proteomics* **11**(21):4117 – 21.
doi: http://dx.doi.org/10.1002/pmic.201100302.

[6] Lütteke, T., Bohne-Lang, A., Loss, A., Goetz, T., Frank, M., von der Lieth, C.-W. (2006) GLYCOSCIENCES.de: an Internet portal to support glycomics and glycobiology research. *Glycobiology* **16**(5):71R-81R.
doi: http://dx.doi.org/10.1093/glycob/cwj049.

[7] Woods Group. (2005 – 2011) GLYCAM Web. Complex Carbohydrate Research Center, University of Georgia, Athens, GA. (http://www.glycam.com)

[8] Akune, Y., Hosoda, M., Kaiya, S., Shinmachi, D., Aoki-Kinoshita, K.F. (2010) The RINGS resource for glycome informatics analysis and data mining on the Web. *OMICS* **14**(4):475 – 86.
doi: http://dx.doi.org/10.1089/omi.2009.0129.

[9] Aoki-Kinoshita, K.F. *Glycome Informatics: methods and applications.* CRC Press, 2010.

[10] Packer, N.H., von der Lieth, C.-W., Aoki-Kinoshita, K.F., Lebrilla, C.B., Paulson, J.C., Raman, R., Rudd, P., Sasisekharan, R., Taniguchi, N., York, W.S. (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**(1):8 – 20.
doi: http://dx.doi.org/10.1002/pmic.200700917.

[11] Hizukuri, Y., Yamanishi, Y., Nakamura, O., Yagi, F., Goto, S. *et al.* (2005) Extraction of leukemia-specific glycan motifs in humans by computational glycomics. *Carbohydr. Res.* **340:**2270 – 2278.
doi: http://dx.doi.org/10.1016/j.carres.2005.07.012.

[12] Kuboyama, T., Hirata, K., Aoki-Kinoshita, K.F., Kashima, H., Yasuda, H. (2006) A gram distribution kernel applied to glycan classification and motif extraction. *Genome Inform.* **17:**25 – 34.

[13] Yamanishi, Y., Bach, F., Vert, J.P. (2007) Glycan classification with tree kernels. *Bioinformatics* **23:**1211 – 1216.
doi: http://dx.doi.org/10.1093/bioinformatics/btm090.

[14] Li, L., Ching, W.K., Yamaguchi, T., Aoki-Kinoshita, K.F. (2010) A weighted q-gram method for glycan structure classification. *BMC Bioinformatics* **11**(Suppl 1):S 33. doi: http://dx.doi.org/10.1186/1471-2105-11-S1-S33.

[15] Takigawa, I., Hashimoto, K., Shiga, M., Kanehisa, M. and Mamitsuka, H. (2010) Mining Patterns from Glycan Structures. In: *Proceedings of the International Beilstein Symposium on Glyco-Bioinformatics*, Logos-Verlag Berlin, pp. 13 – 24.

[16] Ueda, N., Aoki-Kinoshita, K. F., Yamaguchi, A., Akutsu, T., and Mamitsuka, H. (2005) A probabilistic model for mining labeled ordered trees: capturing patterns in carbohydrate sugar chains. *IEEE Transactions on Knowledge and Data Engineering* **17**(8):1051 – 1064.
doi: http://dx.doi.org/10.1109/TKDE.2005.117.

[17] Hashimoto, K., Aoki-Kinoshita, K.F., Ueda, N., Kanehisa, M., Mamitsuka, H. (2008) A new efficient probabilistic model for mining labeled ordered trees applied to glycobiology. *ACM Trans. on Knowledge Discovery from Data (TKDD)*, **2**(1), Article No. 6.

[18] Aoki-Kinoshita, K.F., Ueda, N., Mamitsuka, H., and Kanehisa, M. (2006) ProfilePSTMM: capturing tree-structure motifs in carbohydrate sugar chains. *Bioinformatics* 22:e25-e34.
doi: http://dx.doi.org/10.1093/bioinformatics/btl244.

[19] Aoki, K. F., Mamitsuka, H., Akutsu, T., and Kanehisa, M. (2005) A score matrix to reveal the hidden links in glycans. *Bioinformatics* **21**(8):1457 – 1463.
doi: http://dx.doi.org/10.1093/bioinformatics/bti193.

[20] Puri, A., Neelamegham, S. (2011) Understanding Glycomechanics Using Mathematical Modeling: A Review of Current Approaches to Simulate Cellular Glycosylation Reaction Networks. *Ann. Biomed. Eng.*, Springer Netherlands, pp. 1 – 12.
doi: http://dx.doi.org/10.1007/s10439-011-0464-5.