Protein Motif Analysis in Health and Diseases:

Protein-sugar interaction in Obesity

SUMMARY

Proteins are essential biological molecules involved in virtually every process within a living cell. In a living cell, these biological molecules function in concert with other molecules through protein-protein interaction, protein-nucleic acid interaction and protein-lipid interaction. For those proteins, which act as enzymes, such biological function is mediated through protein-ligand interaction. In most of these interactions, it is a small stretch of amino acids of the protein, termed motif, which is responsible for its biological function in its native conformation. Thus, analysis of a protein motif provides a better understanding on many aspects of protein function and protein interaction. Its thorough analysis at the molecular level plays a key role in discovery research for small molecule drug development and monoclonal antibody based vaccine development. Motif analysis also reveals evolutionary relationships between protein sequences, an important aspect in drug/vaccine discovery research. However, such analysis is inherently computationally demanding because of the exponential growth of the protein databases and the combinatorial number of ways in which protein motifs interact. In collaboration with others, we are involved in protein motif analysis particularly in the *Glycosyltransferase* enzyme family using Protein-probe MotifNetwork, which has been built on biologically grid enabled worksflows involving multiple computational steps for high-throughput search and supported by *cyberinfrastructure* (CI) to meet the computational demand.

The members of the *Glycosyltransferase* enzyme family catalyze the transfer of a sugar moiety from an activated donor sugar onto saccharide and non-saccharide acceptors (usually proteins/lipids). Such products are often found to play a key role in many pathological conditions. As for example, among the glycosyltransferases, presence of sialylmotifs is unique to the sialyltransferase family of enzymes that transfers sialic acid to the non-reducing end of a glycoconjugate. This 9-carbon sialyl moeity has been recognized as the key determinant of a diverse oligosaccharide structures involved in a variety of pathological events including the recent H1N1 (2009) pandemic flu. Members of Sialytransferase Family, therefore, have been recognized as target proteins for drug discovery research. Earlier, we identified and characterized the motifs of a mammalian sialytransferase and found those to particpate in binding with the donor or/and acceptor substrates. Therefore, drug discovery research is targeted to these motifs. A thorough analysis is warranted to better understand the nature of such motifs across the species for drug discovery research.

Such high-throughput analysis of protein motifs involves multiple computational steps. In the first step, Molecular Science Student Workbench (www.bsw-uiuc.net), which is a gateway providing bioinformatics tools and technologies, is used for searching of non-redundant protein sequence database. Once fetched and annotated, the motif analysis begins with a locally installed version of the InterProScan application, which is used to perform the basic analysis by processing each input sequence individually. InterProScan used in such analysis accesses a locally installed version of the InterPro dataset. The results of these runs are then processed. To begin with, the results for each run are individually analyzed to return identified domains including their score (eScore), start and end positions (bps), any known description, the domain ID, and optionally the InterProScan database “matchID”. These individual results are then

assembled into large data matrices from which several levels of analysis or subsequent computations are performed. In order to orchestrate the grid services, a Taverna-based workflow engine has been used in the MotifNetwork environment for enactment that generates several Cytoscape compatible files for displaying the generated domain-webs, which exploit one of the Cytoscape plugins named GenePro. The supporting grid-enabling services used to wrap and invoke the computational applications are implemented with the Generic Service Toolkit (GST). The ultimate results of this environment are data products, organized as matrices, and visualization files suitable for quick analysis. This approach, which is showing promising results, will also be applied for analyzing sugar-protein interactions. At this point, we are interested to find out the evolutionary nature of sialic acid. In that context, we will use this approach analyzing sialyltransferases, trans-sialidases and sialic acid binding proteins.

Such analytical approach will also help us understanding the role of sugar (Glyco-) molecules in obesity. Sugar molecules have significant roles in obesity and diabetes[1]. Yet, nothing is known about the role of glycosylation/glycosyltransferases in obesity. Searching Pubmed on April 18 (2011) with the key words 'childhood obesity and glycosyltransferases' returned no relevant articles other than on Visfatin. Motifs analysis in the glycosyltransferase family will be used to search this relationship. Moreover, Obesity is a metabolic disorder that imbalances the energy expenditure causing an individual to have higher BMI. Protein-sugar interactions in such metabolic network in obesity is one aspect that has already drawn such interests. Recently, it has been found that human microbiome may also have a significant role, particularly the oral and gut microbiome. So, an interesting question that needs to be answered is whether oral microbiota that uses sugar molecule for host-microbe interaction can 'predict' the at-risk individual before the BMI shows the trend. Our motif analysis tool will be used to explore these possibilities. We have already deployed CHOIS, Childhood Obesity Informatics System[2], that are gathering relevant data for such analysis.

RELEVANT PUBLICATIONS

High-throughput sialylmotif analysis in the glycosyltransferase protein family. Arun K. Datta, Jeffrey L. Tilson, Gloria Rendon, Eric Jakobsson. *TeraGrid '09*, Arlington (VA), June 22 - 25, 2009. [http://archive.teragrid.org/tg09/index.php?option=com_content&task=view&id=71]

High-throughput motif analysis for drug discovery research. Arun K. Datta. *Proceedings of the BIT's 8th Annual Congress of International Drug Discovery Sciences and Technology* (IDDST 2010; http://www.iddst.com/2010), p137, October 23-26, 2010, Beijing, China.

Cyberinfrastructure for Glycome Research. Arun K. Datta. *Proceedings of the 20th International Symposium on Glycoconjugates (Glyco XX)*. San Juan (PR), November 29 - December 4, 2009.

Comparative sequence analysis in the sialyltransferase protein family: Analysis of Motifs. Arun K. Datta (2009), *Current Drug Targets*, 10(6): 483-98.[Review article]

Functional Analysis of Motifs in the Sialyltransferase Protein Family. Arun K. Datta. *Proceedings of the Automated Function Prediction*, p87-88, Aug 30 – Sept. 1, 2006, San Diego.

[1] http://www.ncbi.nlm.nih.gov/pubmed/11229668
[2] http://www.nucri.org/mychois

SeqComp – A sequence alignment program for comparative genomics. Arun K. Datta, John Fischer, Jose Paz, and Yumiko Iwai in *Proceedings of the International Conference on Computer Science and its Applications* (ICCSA), p 315-319, July 01-02, 2003, San Diego, CA. US Education Service. ISBN 0-9742-4480-5.

Conserved Cysteines in the sialyltransferase sialylmotifs form an essential disulfide bond. Arun K. Datta, R. Chammas, and J. C. Paulson (2001). *J. Biological Chem*, 276:15200-7.

Mutations of the sialyltransferase S-sialylmotif alters the kinetics of the donor and acceptor substrates. Arun K. Datta, Abhishek Sinha, and James C. Paulson (1998). *J. Biological Chemistry*, 273: 9608 – 9614.

The sialyltransferase sialylmotif participates in binding the donor substrate CMP-NeuAc. Arun K. Datta, and James C. Paulson (1995). *J. Biological Chemistry*. 270: 1497 – 1500.

Both potential dolichol recognition sequences of hamster GlcNAc-1-P transferase are necessary for normal enzyme function. Arun K. Datta, and Mark A. Lehrman (1993). *J. Biological Chemistry*, 268: 12663 – 12668.

CHOIS: Enabling grid technologies for obesity surveillance and control. Arun K. Datta, Victoria Jackson, Radha Nandkumar, Jill Sproat, Weimo Zhu , Heidi Krahling (2010). In 'Healthgrid Applications and Core Technologies', vol 159, p191-202 (T. Solomonides et al., Eds), IOS Press, Washington D.C. ISBN 978-1-60750-582-2. Presented at the HealthGrid 2010, Paris, June 28-30, 2010.

Cyberinfrastructure for CHOIS - a Global Health initiative for obesity surveillance and control. Arun K. Datta, Victoria Jackson, Radha Nandkumar, and Weimo Zhu. Proceedings of the PRAGMA 18 (http://www.pragma-grid.net/), San Diego (CA), March 3 -4, 2010.