High-throughput sialylmotif analysis in the glycosyltransferase protein family

Arun K. Datta[1], Jeffrey L. Tilson[2], Gloria Rendon[3], Eric Jakobsson[3,4]

[1]National University, San Diego, [2]Renaissance Computing Institute, [3]National Center for Supercomputing Applications, [4]Center for Biophysics and Computational Biology at the University of Illinois at Urbana-Champaign

For correspondence: Arun K. Datta, adatta@nu.edu

ABSTRACT:

Sialic acid and its derivatives, often found as a constituent of the extracellular glycoconjugates, are increasingly recognized as the key determinants of a diverse oligosaccharide structures involved in a large variety of biological events as diverse as animal cell-cell interaction to oncogenic transformation (Varki, 1997, Hakomori, 1991). The transfer of sialic acid to such diverse carbohydrate structures is mediated by sialyltransferase (ST; Datta, 2008; Datta, and Paulson, 1997). Earlier L-sialylmotif and S-sialylmotif were shown to confer substrate specificity for the enzyme activity of a sialyltransferase (Datta, and Paulson, 1995; Datta, et al., 1998). While the L-sialylmotif of about 55 amino acids contributes to the binding of the donor substrate (Datta, and Paulson, 1995), the S-sialylmotif of about 22 amino acids contributes to the binding of both the donor and acceptor substrates (Datta, et al., 1998). A conserved disulfide linkage between these two brings these motifs closer together for catalytic activity (Datta, et al., 2001). Comparative sequence analysis also indicated that the conserved peptide sequence flanking these sialylmotifs apparently determine either the carbohydrate or the linkage specificity (Datta, 2008; Datta, 2006; Sujatha, and Balaji, 2006). So far, for mammalian sialyltransferase family, total 20 cloned enzymes with distinct substrate specificity have been determined by experimental analysis (Datta, 2008). However, when the subsequence,
CRRCAVVGNSGNLRESSYGPEIDSHDFVLRMNKAPTAGFEADVGTKTTHHLVYPE,
accounting for L-sialylmotif of hST3Gal I (Kitagawa, and Paulson, 1994) was used for searching *refseq_protein* database using BLASTP (Altschul, et al., 1997), it returned 29

blast hits for the human sequence database alone. On the other hand, similar search in the non-redundant protein sequence database returned 338 hits.

Analysis of a protein motif provides a better understanding on many aspects of protein function, protein interaction, and gene/protein and organism evolution (Tilson, et. al., 2007). It also reveals evolutionary relationships between protein sequences that are too distantly related (Tilson, et. al., 2007; Bjorklund, et. al., 2005). However, such analysis is inherently highly computationally intensive because of the exponential growth of the protein databases and the combinatorial number of ways in which motifs interact with each other. MotifNetwork environment, built on biologically oriented grid-enabled workflows, was shown to serve this purpose enabling the researchers conducting protein motif analysis in a systematic, multilevel, unified, and high throughput way (Tilson, et. al., 2007).

High-throughput analysis of sialylmotifs has been carried out following an approach involving multiple steps (Tilson, et. al., 2007). In the first step, *Molecular Science Student Workbench* (www.bsw-uiuc.net), which is a gateway providing bioinformatics tools and technologies, was used for searching of non-redundant protein sequence database that yielded 338 hits for the L-Sialylmotif of hST3Gal I. Once fetched and annotated, the motif analysis began with a locally installed version of the InterProScan application (Zdobnov, et. al., 2001), which is used to perform the basic analysis by processing each input sequence individually. InterProScan used in this analysis accesses a locally installed version of the InterPro (Apweiler, et. al., 2001; Mulder, et al., 2005) 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" (Tilson, et. al., 2007). These individual results are then assembled into large data matrices from which several levels of analysis or subsequent computations were performed. In addition, the Protein-Probe MotifNetwork Workflow, designed by using Taverna (Oinn, et. al., 2004) for orchestration and enactment, generated several Cytoscape (Shannon, et. al., 2003)

compatible files for displaying the generated domain-webs, which exploit one of the Cytoscape plugins named GenePro (Vlasblom, et. al., 2006). The supporting grid-enabling services used to wrap and invoke the computational applications are implemented with the Generic Service Toolkit (GST) (Kandaswamy, et. al., 2006). The ultimate results of this environment are data products, organized as matrices, and visualization files suitable for quick analysis. Details of these methods and the data output will be presented.

References:

Altschul, S. F., Madden, T. L., Schaffer, A. A., Zhang, J., Zhang, Z., Miller, W., Lipman, D. J. (1997). Gapped BLAST and PSI-BLAST: a new generation of protein database search programs. *Nucleic Acids Res*, 25: 3389-3402.

Apweiler, R., Attwood, T.K., Bairoch, A., Bateman, A., Birney, E., Biswas, M., Bucher, P., Cerutti, L., Corpet, F., Croning, M.D., Durbin, R., Falquet, L., Fleischmann, W., Gouzy, J., Hermjakob, H., Hulo, N., Jonassen, I., Kahn, D., Kanapin, A., Karavidopoulou, Y., Lopez, R., Marx, B., Mulder, N.J., Oinn, T.M., Pagni, M., Servant, F., Sigrist, C.J., and Zdobnov, E.M. (2001). The InterPro database, an integrated documentation resource for protein families, domains and functional sites. *Nucleic Acids Res*, 29: 37-40.

Bjorklund, A.K., Ekman, D., Light, S., Frey-Skott, J., and Elofsson, A. (2005). Domain Rearrangements in Protein Evolution. *Journal of Molecular Biology*, 353: 911-923.

Datta, A.K. (2008). Comparative sequence analysis in the sialyltransferase protein family: Analysis of motifs. *Current Drug Targets*, in press.

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

Datta, A.K., Chammas, R., and Paulson, J. C. (2001) Conserved cysteines in the sialyltransferase sialylmotifs form an essential disulfide bond. *J Biol Chem*, 276: 15200-15207.

Datta, A.K. and Paulson, J.C. (1995) The sialyltransferase "sialylmotif" participates in binding the donor substrate CMP-NeuAc. *J Biol Chem*, 270: 1497-1500.

Datta, A.K. and Paulson, J.C. (1997) Sialylmotifs of sialyltransferases. *Indian J Biochem Biophys*, 34: 157-165.

Datta, A.K., Sinha, A., and Paulson, J. C. (1998) Mutation of the sialyltransferase S-sialylmotif alters the kinetics of the donor and acceptor substrates. *J Biol Chem*, 273: 9608-9614.

Hakomori, S. (1991) Possible functions of tumor-associated carbohydrate antigens. *Curr Opin Immunol*, 3: 646-653.

Kandaswamy, G., Fang, L., Huang, Y., Shirasuna, S., Marru, S., and Gannon, D. (2006). Building Web Services for Scientific Grid Applications. *IBM Journal of Research and Development*, 50: 249-260.

Kitagawa, H. and Paulson, J.C. (1994) Differential expression of five sialyltransferase genes in human tissues. *J Biol Chem*, 269: 17872-17878.

Mulder, N. J., Apweiler, R., Attwood, T. K., Bairoch, A., Bateman, A., Binns, D., Bradley, P., Bork, P., Bucher, P., Cerutti, L., Copley, R., Courcelle, E., Das, U. et. al., (2005) InterPro, progress and status in 2005. *Nucleic Acids Res*, 33: D201-205.

Oinn, T., Addis, M., Ferris, J., Marvin, D., Greenwood, M., Carver, T., Pocock, M.R., Wipat, A., and Li, P. (2004). Taverna: a tool for the composition and enactment of bioinformatics workflows. *Bioinformatics*, 20: 3045-3054.

Shannon, P., Markiel, A., Ozier, O., Baliga, N.S., Wang, J.T., Ramage, D., Amin, N., Schikowski, B., and Ideker, T. (2003). Cytoscape: A software environment for integrated models of biomolecular interaction networks. *Genome Research*, 13: 2498-2504.

Sujatha, M.S. and Balaji, P.V. (2006) Fold recognition and comparative modeling of human alpha2,3-sialyltransferases reveal their sequence and structural similarities to CstII from Campylobacter jejuni. *BMC Struct Biol*, 6: 9.

Tilson, J.L., Rendon, G., Mao-Feng, G., and Jakobsson, E. (2007). MotifNetwork: A Grid-enabled workflow for high-throughput domain analysis of biological sequences: Implications for study of phylogeny, protein interactions, and intraspecies variation. *Proceedings of the 7th International Symposium on Bioinformatics and BioEngineering*.

Varki, A. (1997). Sialic acids as ligands in recognition phenomena. *FASEB J*, 11: 248-255.

Vlasblom, J., Wu, S., Pu, S., Superina, M., Liu, G., Orsi, C., and Wodak, S.J. (2006) GenePro: a Cytoscape plug-in for advanced visualization and analysis of interaction networks. *Bioinformatics*, 22: 2178-2179.

Zdobnov, E.M.a.A., R. (2001) InterProScan - an integration platform for the structure-recognition methods in InterPro. *Bioinformatics*, 17: 847-848.