A proposal for a hidden-Markov-model-based method of simulating control of glycan synthesis with a review of current trends in glycomics

Robert Brown. ***MBB452 Final Project*** **(10 Dec 2005)**

A recent *Nature Methods* review summarizes progress in the study of glycans and problems still faced in the field.[1] The authors observe that the study of glycomics is made more difficult than many other problems in biology due to the 'analog' action of glycan function. Additionally glycans are produced purely enzymatically, rather than on templates, thus they are expressed in a range of structures. The authors identify five major directions in which glycomics needs to progress: functional genomics, glyco-gene microarrays, glycan structure determination, understanding of glycan-protein biochemistry, and bioinformatics (Fig. 1).

Functional Genomics

Functional genetics approaches to glycomics are made much more difficult by the heterogeneity and enzymatic biosynthesis of glycans. Simple knockouts of glycan structures are impossible; instead, experimenters must knockout one or more of the approximately 200 genes involved in glycan biosynthesis. The review strongly stresses the importance of understanding the functions both the direct effects of all of these enzymes and their downstream effects on organismal phenotype. A brief survey of the KEGG database[2] shows that knowledge of direct function is generally very good; however, knowledge of regulation and downstream effects is much poorer. I would rather question whether knocking out synthetic genes and examining gross phenotype is the best way to proceed with glycomics though it is certainly an important tool.

Glyco-gene microarrays

The authors seem to strongly laud the creation of these microarrays, though they do not seem to be anything novel. Experimenters have simply taken Affymetrix array and printed on cDNA complementary to genes relevant to glycosynthesis.[3] While it is not nearly as novel as the authors indicate, it is certainly a very important tool in understanding the general pattern of glycosynthesis in a cell. I would, however, caution that reviewing the KEGG database shows that many of these enzymes are posttranscriptionally regulated by glycosylation. As such, if you have reasons to believe that a cell has abnormal glycosylation, you have no reason to believe that transcript levels correspond in any meaningful way with activity.

Glycan structure determination

I was rather confused when I first read this section, as it glosses over the use degradation and phosphorescent labeling for sequencing, focusing instead on mass spectrometry and NMR. Having experience with the problems both manual and high-throughput protein structure determination by mass spectroscopy, I am rather critical of reliance on MS in the much more difficult prospect of sequencing glycans. Additionally, a brief review of the literature will suggest that phosphorescent techniques, essentially running out glycogens by degratdative electrophoresis, are showing very promising results in both clinical and laboratory settings.[4,5,6,7] These are seeming very robust for simultaneous high-throughput and fine-structure determination. These techniques are claiming to be more accessible, require less material, be faster, and give at least as fine structure as the other techniques. Until I noticed the ‘competing interests statement,’ I could not explain the author’s spending only a couple of sentences discussing such methods. The authors have significant intellectual investment and expertise in the use of mass spectroscopy for this sequencing[89] and perhaps felt less capable of writing on techniques with which they had less familiarity. Furthermore, it appears that G. Venkataraman, the author most expert in analytical glycomics, is the founder of and chief investor in Momenta Pharmaceuticals,[10] a company holding patents heavily reliant on the continued use of mass spectroscopy and NMR in glycogenics. In fairness to the authors, they do advocate using all available methods orthogonally in order to arrive at a final fine structure. Ultimately having good methods of determining what glycans are actually present is probably the most important aspect of glycomics as it is essential for assessing the 'glycozome.'

Glycan-protein biochemistry

As glycan-protein interactions are thought to be very important, especially at the cell surface, the understanding of how glycans bind to proteins is obviously an essential piece of information in the determination of their function. The authors suggest two main methods for studying such interactions. The first is to synthesize a wide selection of proteins and do a competitive binding assay. This seems like a very good method, especially for finding highly interactive species as possible drugs, though it suffers from the problem that the glycans must be artificially synthesized in order to have sufficient quantities. It might also be interesting to attempt affinity co-electrophoresis on the mixture to get a good profile of the relative binding potentials of all species.[11] The amount of glycan that would be needed for this experiment is low enough that it could also be used with cell derived glycans, though the agarose might affect the binding properties. The other technique that the authors describe involves the creation of arrays of glycans and the use of ELISA-like double antibody screens to assay protein binding. This would be an ideal method for high-throughput binding studies except that it has been suggested that glycan structure leaves them especially vulnerable to changed binding properties when immobilized on a surface.[12] There are no other good high-throughput screens (other than, perhaps the above-mentioned co-electrophoresis), so this will have to do. It may, however, be wise to verify binding by other methods.

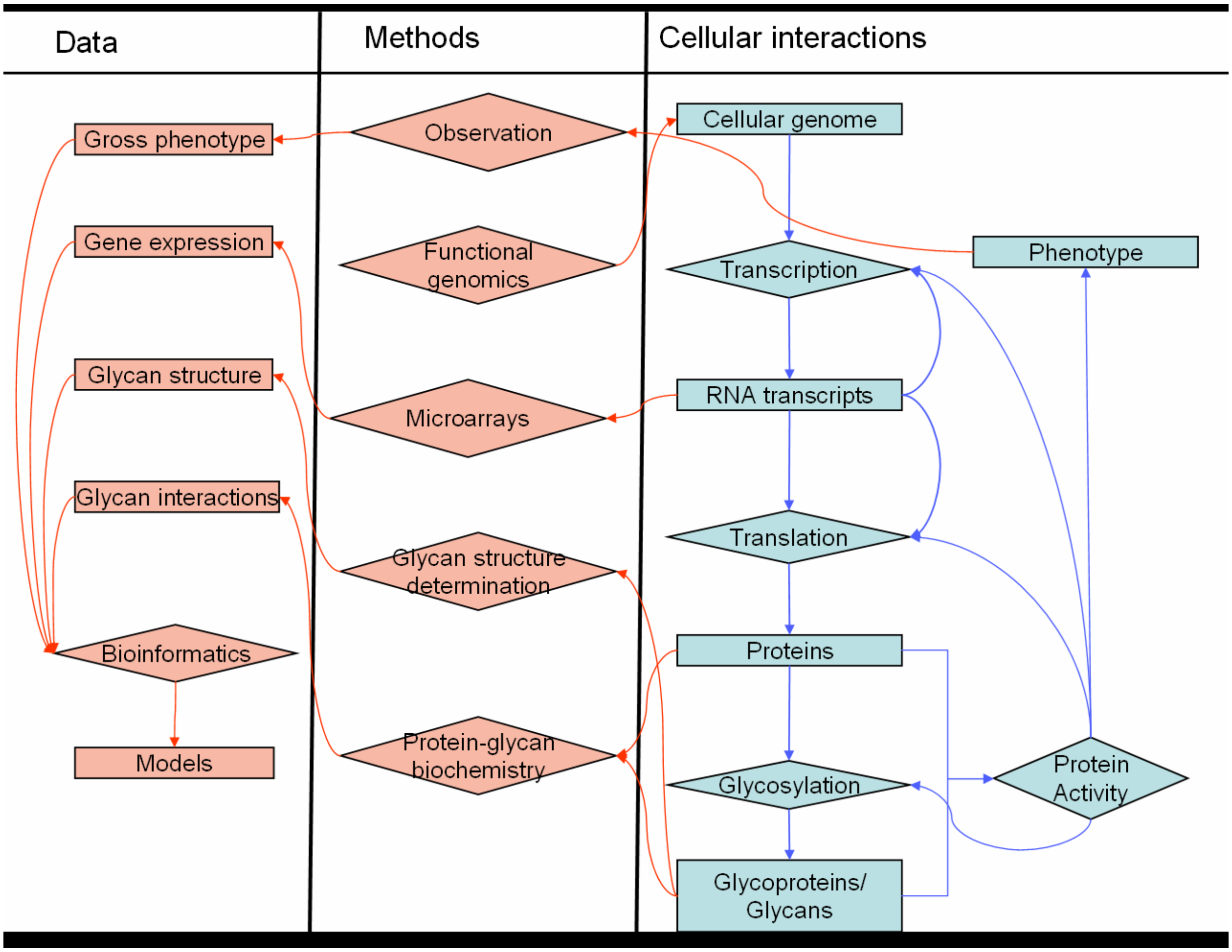

Figure 1: A basic schematic representation of the cellular processes effected by and affecting the glycozome and the glycomics techniques used to study them.
Rectangles represent objects and diamonds processes. Blue is naturally-occurring while red is experimenter-induced.

Bioinformatics

The synthesis of all the data above into good models of the glycome is obviously an important consideration. The authors understand this, but do not seem to have any particularly salient insight in what needs to be done. They simply advocate the creation of a user-friendly object-oriented relational database that will allow all the important data to be held in one location and quickly cross-referenced. A quick search through the KEGG and CFG databases indicates that this process is already well on its way (indeed it seems that three of the authors wrote the CFG data structure). They do not, however, give any real idea about what to do with the data once it is databased other than with broad concepts , such as ‘structure prediction.’

Despite its flaws, the review is generally a good overview of the current state of glycomics research. It is, however, very brief and probably only suitable for someone searching for the bare basics in the field. However, it seems to assume some knowledge of glycans. It is, thus, difficult to determine who the intended audience is. As the authors are all intimately tied to the Consortium for Functional Glymocis, the text is widely biased towards the work of this organization. Indeed, even their section headings correspond directly with the working groups of the consortium. A bit wider explanation of the field would have been a nice addition, but the authors are all experienced in the field and do a good job of summarizing recent advances and continuing difficulties.

A proposal for the creation and use of a hidden Markov model to rapidly estimate glycosynthetic protein activities and simulate glycosomal regulation

The synthesis of glycans is thought to be the process of competing reactions of about 200 glycosynthetic proteins. These act to sequentially ligate additional saccharides onto the end of the carbohydrate chain, perhaps occasionally branching. As such, the creation of a glycan can be viewed as a series of events in which enzymes compete, one attaches and, if the glycan terminus is the enzymes substrate, adds its characteristic structure to the glycan, it then disengages, and competition begins again. Though many other factors likely affect the chances, the odds of any given enzyme acting in any given step should be directly proportional to its cellular activity (or more correctly, its cellular activity relative to the other competing enzymes). This implies that the knowledge of a glycan's structure gives direct evidence of the activity levels of glycosynthetic proteins.

This begins to imply a hidden Markov model underlying each glycan structure. The model essentially equates to a hub, representing the glycan with no enzyme engaged, and *N* spokes where *N* is the number of enzymes in the model. The state at the end of each spoke represents the action of an enzyme. The hub itself does not emit a signal. The probability of transitioning from the hub to a spoke, *i,* is $\frac{c_i \rho_i}{\sum_{j=0}^{j=N} c_j \rho_j}$ where c_i is a constant related to the intrinsic activity of the enzyme in this glycan and ρ_i is the cellular activity level of enzyme *i.* The emission profile of each state will be a conditional stating that if the glycan is an acceptable substrate the enzyme will emit its activity, else this state will not emit. The probabilities of return to the hub or recycling through the state will be an intrinsic property of enzyme *I*, though in most cases it will always return to the hub. Thus, given an array of glycan structures, one can optimize the transition probabilities and solve for all ρ. This provides an estimate of the activity levels of all glyco-synthetic proteins in the cell. It is useful to know this, because, as I pointed out earlier, transcript data may be rendered useless by widespread and self-dependent post-transcriptional modifications. Furthermore, with both this and the transcript data it would be easy to compare the levels of expression with the levels of activity. Thus, one could gain a good idea of how the glycozome self-regulates.

In order to accomplish this study, we first need to find a variety of conserved glycosylation sites within the genome. Reports of such sites are abundant throughout the literature.[13,14] We must ensure that we get enough and different enough sites that all glycosynthetic enzymes will be well represented. The most obvious consideration here is to make sure that there are both *N*-linked and *O*-linked glycoproteins represented. I would hazard to say that ten to twenty well-chosen sites would be sufficient. We then need a source of cells. Cultured cell lines would generally be most practical, though there is no reason that biopsy or autopsy sections could not be used. Next, we will need antibodies against the proteins that contain these sites so that we can sequentially immunoprecipitate them from the whole cell extract. For this reason, it would greatly simplify the experiment, if we could find multiple glycosylation sites on any given protein. At this point, we simply must find the structure distribution of the glycans at these positions.

After the first batch of structures comes in, we must optimize the model to find the values of the activity constants $c_1 - c_N$. The most useful method for doing this would probably be applying either the combinatorial or the Lagrange multiplier method to optimize all transitions.[15] As we can, by definition, assume that all cellular activity levels are cell-wide, any difference between the figures at various glycosylation sites must be a result of the site-specific constant. Further rounds of training will likely be needed in order to refine these values, but a general framework will be in place. Future refinements could include removing the hub and the conditional emission from the states, then simply allowing any state to transition into any other state. This would certainly provide a more powerful model, but I would refrain for the minute as I would worry about over fitting with a matrix of about 40,000 possible transitions and it is much easier to error-check 200 elements than a 200 by 200 matrix. Once the constants were assigned values, any experiment that found the structure distribution at these glycosylation sites would indicate activity levels of all glycosynthetic proteins in the cell. Since another round of transition probability optimization, this time assuming the constants and optimizing across all sites, using the new experimental values would return values of $\rho_1 - \rho_N$. Further, any experiment that gave structures at these glycosylation sites and additional sites would indicate the values of $c_1 - c_N$ for all the additional sites. Knowing this, it is easy to imagine a rapidly expanding database of the constant values, eventually covering every glycosylation site in the genome. At this point, the knowledge of the structure distribution of glycans at any given glycosylation site would allow one to predict not only the activity levels of all glycosynthetic proteins but also the expected distribution of glycan structure at every glycosylation site in the genome.

If it were to work this well, this technique would be very powerful indeed. Predictive perfection would indicate almost complete knowledge of the glycome. This knowledge is even more useful as the variables involved are all firmly rooted in physical properties; as they are physical observables, they could be empirically tested. Indeed, testing could be a powerful method of model refinement. Admittedly, knowledge of those facors that contributed to the various constants would be nice. Presumably these constants are strongly correlated with the enzyme's binding affinity to the site of interest, though they could also relate to other cellular elements interacting with this site. I would argue that if this model were to come about, it would indicate perfect understanding of the self-regulation of the glycome. However, there are likely to be far more factors than I can conceive of involved so it is much more likely that this model would be in a constant state of refinement, growing more and more complex as it closer and closer approximates reality.

[1] R. Raman, S. Ragurram, G. Venkataraman, J. Paulson, and R. Sasiekharan. Glycomics: an integrated systems approach to structure-function relationships of glycans. *Nature Methods.* 2 (2005), 817-824. Review. (uncited facts are all drawn from this article).

[2] Kanehisa, M. and Goto, S.; KEGG: Kyoto Encyclopedia of Genes and Genomes. Nucleic Acids Res. 28 (2000), 27-30.

[3] E. Cornelli, M. Amado, S. Head, and J. Paulson. Custom microarray for glycobiologists: considerations for glycosyltransferase gene expression profiling. *Biochem Soc Symp.* (2002),135-42. Review.

[4] J. Khandurina, D. Blum, J. Stege, and A. Guttman. Automated carbohydrate profiling by capillary electrophoresis: A bioindustrial approach. *ELECTROPHORESIS*, 25 (2004), 2326-2331.

[5] P. Rudd, R. Dwek. Rapid sensitive sequencing of oligosaccharides from glycoproteins. *Curr. Opin. Biotechnol.* 8, 488-497.

[6] N. Callewaert, S. Geysens, F. Molemans, and R. Contreras. Ultrasensitive profiling and sequencing of N-linked oligosaccharides using standard DNA-sequencing equipment. *Glycobiology* (2001), 11, 275-281.

[7] N. Callewaert, H. Vlierberghe, A. Hecke, W. Laroy, J. Delanghe, and R. Contreras. Noninvasive diagnosis of liver cirrhosis using DNA sequencer−based total serum protein glycomics.
Nature Medicine, 10 (2004), 429 – 434.

[8] G. Venkataraman, Z. Shriver, R. Raman, R. Sasisekharan, Sequencing Complex Polysaccharides. Science, 15 (1999), 537-542.

[9] CFG, with which the authors are intimately tied, is also almost exclusively using MS. http://www.functionalglycomics.org/static/consortium/organization/sciCores/corec.shtml

[10] Company website: http://www.momentapharma.com/index.htm

[11] R. Nelson, A. Venot, M. Bevilacqua, R. Linhardt, and I. Stamenkovic. Carbohydrate-Protein Interactions in Vascular Biology. *Annual Review of Cell and Developmental Biology***,** 11 (1995), 601-631.

[12] Nelson, *et al.* (1995).

[13] M. No, M. Omary. Identification and mutational analysis of the glycosylation sites of human keratin 18. *J Biol Chem*, 270 (1995), 11820-11827.

[14] S. Bungert, L. Molday and R. Molday. Membrane Topology of the ATP Binding Cassette Transporter ABCR and Its Relationship to ABC1 and Related ABCA Transporters : IDENTIFICATION OF *N*-LINKED GLYCOSYLATION SITES. *J. Biol. Chem.*, 276 (2001), 23539-23546.

[15] X. Li, M. Parizeau, and R. Plamondon. Training Hidden Markov Models with Multiple Observations-A Combinatorial Method. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 22 (2000), 371-377.