BIOINFORMATICS ORIGINAL PAPER
Vol. 24 no. 23 2008, pages 2740–2747
doi:10.1093/bioinformatics/btn515

Systems biology

Systems-level modeling of cellular glycosylation reaction networks: O-linked glycan formation on natural selectin ligands

Gang Liu[1,†], Dhananjay D. Marathe[1,†], Khushi L. Matta[2] and Sriram Neelamegham[1,3,*]

[1]Chemical and Biological Engineering, State University of New York, Buffalo, NY 14260, [2]Cancer Biology, Roswell Park Cancer Institute, Buffalo, NY 14263 and [3]NY State Center for Excellence in Bioinformatics and Life Sciences, State University of New York, Buffalo, NY 14260, USA

Received on July 11, 2008; revised on September 3, 2008; accepted on October 2, 2008

Advance Access publication October 7, 2008

Associate Editor: Martin Bishop

ABSTRACT

Motivation: The emerging field of Glycomics requires the development of systems-based modeling strategies to relate glycosyltransferase gene expression and enzyme activity with carbohydrate structure and function.

Results: We describe the application of object oriented programming concepts to define glycans, enzymes, reactions, pathways and compartments for modeling cellular glycosylation reaction networks. These class definitions are combined with current biochemical knowledge to define potential reaction networks that participate in the formation of the sialyl Lewis-X (sLe^X) epitope on *O*-glycans linked to a leukocyte cell-surface glycoprotein, P-selectin Glycoprotein Ligand-1 (PSGL-1). Subset modeling, hierarchical clustering, principal component analysis and adjoint sensitivity analysis are applied to refine the reaction network and to quantify individual glycosyltransferase rate constants. Wet-lab experiments validate estimates from computer modeling. Such analysis predicts that sLe^X expression varies directly with sialyltransferase α2,3ST3Gal-IV expression and inversely with α2,3ST3Gal-I/II.

Availability: SBML files for all converged models are available at http://www.eng.buffalo.edu/~neel/bio_reaction_network.html

Contact: neel@eng.buffalo.edu

Supplementary information: Supplementary data are available at *Bioinformatics* online.

1 INTRODUCTION

Glycosylation is the process by which saccharides are added to proteins and lipids. A systems-level understanding of this process is important for studies of human pathophysiology. Interest in the coordinated regulation of various glycosylation-related enzyme families (transferases, isomerases, synthases and transporters) is enhanced due to the observation that deficiency in the function of one or more enzymes involved in glycosylation results in a cluster of genetic diseases called congenital defects of glycosylation (CDG) (Jaeken and Matthijs, 2007). Aberrant glycosylation is also associated with cancer (Hakomori, 2001) and with a cell adhesion disorder called LAD-II (Leukocyte Adhesion Deficiency II, CDG IIc) (Karsan *et al*., 1998). In the last case, defects in the synthesis of fucosylated glycostructures results in deficiency of a carbohydrate epitope, expressed on white blood cells (leukocytes), called sialyl Lewis-X [sLe^X, NeuAcα2,3Galβ1,4 (Fucα1,3) GlcNAc-][1]. Reduction in sLe^X results in reduced leukocyte adhesion to blood vessel wall and patient immunodeficiency. Besides the disease context, systems-level understanding of glycosylation can aid biotechnology, including antibody development and glycan engineering (Amano *et al*., 2008).

Novel, recently developed tools are providing new opportunities for the development and application of systems-level modeling in the field of Glycomics (Packer *et al*., 2008). These tools include improved 1-/2-D NMR, novel chromatography methods, advanced tandem mass spectrometry coupled with algorithms for spectral analysis, glycan/lectin-based microarrays, and expanded sets of glycan standards and libraries. Together these tools are yielding new insight both on the monosaccharide composition and structure of complex glycoconjugates. Some of these studies provide quantitative information on the distribution of O-linked glycans in human glycoproteins glycophorin A, neutrophil gelatinase B (Royle *et al*., 2002), breast cancer mucin MUC1 (Muller and Hanisch, 2002) and L-selectin ligand CD34 (Satomaa *et al*., 2002).

Although, experimental technologies and biochemical knowledge developed in recent years suggest that we may be able to establish a quantitative link between cellular gene/enzyme levels and corresponding carbohydrate structures, few models of glycosylation reaction networks exist (Krambeck and Betenbaugh, 2005; Shelikoff *et al*., 1996; Umana and Bailey, 1997). The model by Umana and Bailey (1997) examines N-linked glycosylation based on experimentally determined rate constants, but it does not attempt to relate model output with experimentally determined glycan distribution. Krambeck and Betenbaugh (2005) make this comparison with experimental data. However, their model is large with several thousand reactants and reactions, all of which cannot be independently measured.

*To whom correspondence should be addressed.

†The authors wish it to be known that, in their opinion, the first two authors should be regarded as joint First Authors.

[1]NeuAc, sialic acid; Gal, Galactose; Fuc, Fucose; GlcNAc, *N*-acetyl-D-glucosamine; GalNAc, *N*-acetyl-D-galactosamine; GlycoT, glycosyltransferase; SialylT/ST3Gal, sialyltransferase; FT/FucT, fucosyltransferase; GalT, galactosyltransferase; GlcNAcT, *N*-acetylglucosaminyl transferase.

Table 1. Core-2 O-glycans of PSGL-1[a]

Glycan group	Glycan number	Structure[b]	Percent composition[c] (adjusted, %)
1	S20	NeuAcα2,3(Galβ1,4GlcNAcβ1,3)$_2$Galβ1,4GlcNAcβ1,6 (NeuAcα2,3Galβ1,3) GalNAc-O	3.5 (6.0)
	S18	NeuAcα2,3(Galβ1,4GlcNAcβ1,3)$_2$Galβ1,4GlcNAcβ1,6 (Galβ1,3) GalNAc-O	
	S19	(Galβ1,4GlcNAcβ1,3)$_2$Galβ1,4GlcNAcβ1,6 (NeuAcα2,3Galβ1,3) GalNAc-O	
2	S17	**NeuAcα2,3Galβ1,4(Fucα1,3)GlcNAcβ**1,3Galβ1,4GlcNAcβ1,6 (NeuAcα2,3Galβ1,3) GalNAc-O	2.8 (4.8)
	S14	**NeuAcα2,3Galβ1,4(Fucα1,3)GlcNAcβ**1,3Galβ1,4GlcNAcβ1,6 (Galβ1,3) GalNAc-O	
3	S13	NeuAcα2,3Galβ1,4GlcNAcβ1,3Galβ1,4GlcNAcβ1,6 (NeuAcα2,3Galβ1,3) GalNAc-O	6.4 (11.0)
	S10	NeuAcα2,3Galβ1,4GlcNAcβ1,3Galβ1,4GlcNAcβ1,6 (Galβ1,3) GalNAc-O	
	S11	Galβ1,4GlcNAcβ1,3Galβ1,4GlcNAcβ1,6 (NeuAcα2,3Galβ1,3) GalNAc-O	
4	S7	NeuAcα2,3Galβ1,4GlcNAcβ1,6 (NeuAcα2,3Galβ1,3) GalNAc-O	23.2 (40.0)
5	S5	Galβ1,4GlcNAcβ1,6 (NeuAcα2,3Galβ1,3) GalNAc-O	15.0 (25.9)
	S4	NeuAcα2,3Galβ1,4GlcNAcβ1,6 (Galβ1,3) GalNAc-O	
6	S2	Galβ1,4GlcNAcβ1,6 (Galβ1,3) GalNAc-O	7.1 (12.3)

[a]PSGL-1 contains 71 Ser/Thr sites where O-linked glycosylation may occur. Aeed *et al*. (1998) measured the distribution of glycans on this protein (data presented above).
[b]All species listed here are not necessary in each subset pathway. However, a minimum of at least one glycan from each glycan group (left column) must be present in each subset model. Species **S14** and **S17** bear the tetrasaccharide epitope, sLe^X (denoted in **bold**).
[c]% Composition with respect to all *O*-glycan structures of PSGL-1. Of these 42.1% do not have the core-2 structure: NeuAcα2,3Galβ1,3GalNAc-O = 12.5%; Galβ1,3GalNAc-O = 23.8%; GalNAc-O = 3.9%; NeuAcα2,3Galβ1,4(Fucα1,3)GlcNAc …. GalNAc-O = 1.9%. These are omitted in our analysis since they are either core-1 (Galβ1,3GalNAcα-) glycans that do not undergo significant modifications in the trans-Golgi, or they are low abundance species. Composition is 'adjusted' as a percentage of total core-2 compounds.

While previous papers have focused on *N*-glycan biosynthesis, we present here, the first reaction network model for O-linked glycosylation. In order to bring a greater 'systems approach' to bear on this field, we suggest the use of defined class structures and object-oriented programming concepts for the systematic construction of glycosylation reaction networks. Incorporation of class definitions in this manner can facilitate the introduction of XML-based notations in future models of glycosylation reaction networks. A concept called 'subset-modeling' is also introduced in order to accommodate for the group specificity properties of glycosyltransferases (GlycoTs), i.e. the ability of enzymes to act specifically on a class of substrates but not other closely related molecules. The above are combined with genetic algorithm (GA)-based optimization strategies, hierarchical clustering and principal component analysis (PCA) to define both reaction pathways and corresponding GlycoT rate constants that lead to the formation of the sLe^X epitope on *O*-glycans of a human leukocyte glycoprotein called P-Selectin Glycoprotein Ligand-1 (PSGL-1). In our case, these reaction pathways and enzyme rate constants are inferred by fitting individual subset models with structural data that quantify the distribution of O-linked glycans on PSGL-1 (Aeed *et al*., 1998, Table 1). Such an effort that critically examines the features regulating sLe^X expression on PSGL-1 is important due to the role of this epitope in regulating leukocyte cell-adhesion processes that accompany tissue inflammation, and consequently the interest to modulate this 'druggable target'. *In silico*[2] rate constants estimated from computational modeling compare favorably with our wet-lab estimates. Sensitivity analysis applied to determine the effect of perturbing specific rate constants on a system output (sLe^X in our case) also yield experimentally testable hypothesis. Together, the coupling of simulation and experimental tools, as presented here, represents a promising approach for the application of molecular systems biology concepts in the emerging field of Glycomics.

2 SYSTEM AND METHODS

2.1 Modeling overview and assumptions

Figure 1A presents an overview of our approach. The goal is to determine the reaction pathway/significant reactions and corresponding rate constants (bottom left, Fig. 1A) that allow fitting of semi-quantitative experimental data (top left, Fig. 1A) on glycan structures, with computational models that are represented by biochemical reaction networks (right side, Fig. 1A). With the goal of illustrating this modeling approach, here, we fit the experimental data of Aeed *et al*. (1998). These authors assayed the distribution of O-linked glycans of PSGL-1 in HL-60 cells (Table 1, Supplementary Material provides a brief description of their experimental methods). The unknowns that are fit by the model are the rate constants that quantify the enzyme activity of five GlycoTs: β1,4GalT-IV, β1,3GlcNAcT, α2,3ST3Gal-I/II, α2,3ST3Gal-IV and α1,3FT-VII (bottom left, Fig. 1A). These five rate constants were also experimentally measured by performing wet-lab experiments in our laboratory (Marathe *et al*., 2008). Rate constants determined by fitting glycan distribution data (Table 1) in the current article were compared with our wet-lab rate constant measurements.

The portion of the code dealing with pathway construction is written using FORTRAN 90 (Intel, Santa Clara, CA) (top right, Fig. 1A). To facilitate modeling, five classes are defined namely 'GLYCAN', 'ENZYME', 'REACTION', 'PATHWAY' and 'COMPARTMENT'. Each of these classes

[2]GlycoT rate constants estimated from simulations is denoted k_{silico}. k_{situ} is the estimated GlycoT rate constant in Golgi based on wet-lab experiments conducted with cell lysate and calculations elaborated upon in Supplementary Material. [A], [E_0], [M] and V refer to acceptor, enzyme, monosaccharide concentration and reaction velocity in Golgi. $[A]_{wet}$, $[E_0]_{wet}$, $[M]_{wet}$ and V_{wet} refer to the same parameters in wet-lab experiments.

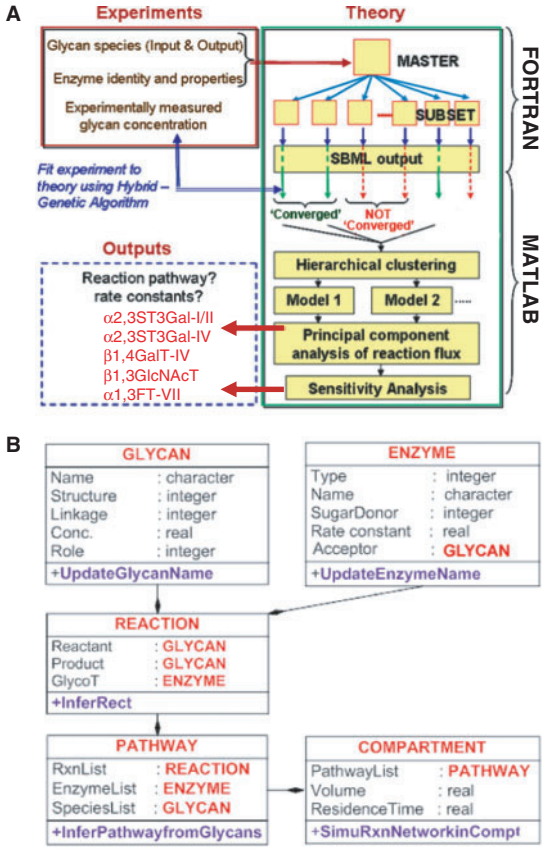

Fig. 1. Modeling overview: (**A**) experimental data (top-left box) on the distribution of *O*-glycans on PSGL-1 are fit to a mathematical model (right-side box) that is described as a series of biochemical reactions. Five GlycoT rate constants listed in the bottom-left box are determined. (**B**) The class diagram describes selected fields and methods in five classes: GLYCAN, ENZYME, REACTION, PATHWAY and COMPARTMENT.

contains the properties of the particular component. The association map between the classes is shown in Figure 1B. Supplementary Material provides details about these classes including sample FORTRAN code and examples for the case of O-linked glycans of PSGL-1 (Supplementary Tables S1 and S2). Sections 2.2 and 2.3 below provide details on the strategy to generate both the master pathways (*MP*s) and subset pathways (*SP*s). The output from this section is a series of SBML-files containing all possible *SP*s.

Model fitting is performed for each subset pathway. The GA toolbox in MATLAB (Mathworks, Natick, MA, USA) is used for this purpose and the fitness criteria, f, is defined to quantify the ability of each subset model to fit experimentally determined glycan distribution. Models with $f < 0.19$ are said to 'converge'. The number of such subset models that converged is designated l. Converged models are further analyzed using three post-simulation analysis methods: (i) Clustering is performed to group the l converged models based on similarity in reaction network structure. (ii) PCA is applied to determine significant reactions and corresponding rate constants in each cluster (Liu *et al*., 2005). The output from this analysis is the PCA flux parameter (PCA_j), which is a measure of the importance of the j-th reaction flux across multiple models in a given cluster. (iii) Adjoint sensitivity analysis is employed to define potential experiments that can be used to test and refine the reaction network model (Cao *et al*., 2003; Liu and Neelamegham, 2008). This analysis yields the adjoint sensitivity coefficient W_j $(\sim(\Delta[\mathrm{sLe^X}]/\Delta k_j)\cdot(k_j/[\mathrm{sLe^X}]))$, which is a measure of the effect of infinitesimal perturbation in system parameter (j-th *in silico* enzyme rate constant k_j in our case) on system variable/output (concentration of species **S14** and **S17**, $[\mathrm{sLe^X}]$, in our case). Details on parameter estimation and post-simulation analysis methods are provided in Supplementary Material.

The following are modeling assumptions: (i) Only five GlycoTs are considered in this reaction network. This is reasonable since these are the dominant enzyme activities detected in our biochemical assays (Marathe *et al*., 2008). Mouse knockout experiments also suggest that these are the critical enzyme activities involved in selectin-ligand formation (Lowe, 2003). (ii) All reactions are modeled to reside in a single compartment that represents the trans-Golgi. This is consistent with experimental data showing that all five enzymes reside in the trans-Golgi (Supplementary Table S3). This compartment is assumed to be well mixed with uniform enzyme, reactant and product concentrations. Compartment-specific parameters, including compartment volume, residence time and initial glycan concentration at the start of the simulation (Supplementary Table S4). Since glycan composition is expressed as percentage of total glycans in Table 1, residence time is the only parameter affecting simulation results in this article. (iii) The reaction in Golgi is limited by substrate/acceptor ([A]), not sugar-nucleotide ([M]) availability. This is reasonable since donor sugar-nucleotide availability in the Golgi lumen is high (1.0–10 mM) (Monica *et al*., 1997; Tomiya *et al*., 2001), and it is typically much greater than the corresponding Michaelis–Menten constant ($K_{M,mono}$ =10–200μM range, see Table 2); Further, [A] is less than the corresponding Michaelis–Menten constant, $K_{M,acc}$ (Table 2). Taken together, these observations suggest that the GlycoT rate expression in cells (V) can be expressed as a first-order reaction with respect to [A]:

$$V = \frac{k_{\mathrm{cat}}[E_0][\mathrm{A}]}{(K_{\mathrm{M,acc}}+[\mathrm{A}])} \sim \frac{k_{\mathrm{cat}}[\mathrm{E_0}][\mathrm{A}]}{K_{\mathrm{M,acc}}} = k_{silico}\cdot[\mathrm{A}] \qquad (1)$$

Here, the rate constant for product formation (k_{cat}) and the Golgi enzyme concentration ($[\mathrm{E_0}]$) along with $K_{M,acc}$ are combined to define the apparent first-order glycosyltransferase rate constant, k^4_{silico}.k_{silico} is thus a lumped parameter containing enzyme activity and expression data. (iv) Product inhibition is ignored since Golgi residing phosphatases cleave nucleotide-byproduct (like UDP), thus minimizing their effect on transferase reactions.

2.2 *MP* construction

The *MP* is a collection of all possible reactions joining the 'initial' and 'product' glycan(s). Here, the 'initial glycan' represents the initial carbohydrate structure that enters the Golgi compartment during the *in silico* computations. In the example presented here, since our primary focus is on chain extension and termination and not initiation, we defined the core-2 trisaccharide structure, Galβ1,3(GlcNAcβ1,6)GalNAc- (**S1**), to be the 'initial glycan' (Fig. 2). The 'product' glycan(s) defines the repertoire of carbohydrate structures resulting from a biochemical reaction network. In our case, these product glycans are listed in Table 1 and shown schematically in Figure 2 without an enclosing box.

For given data on initial and product glycan(s), the program automatically generates the *MP* (Fig. 2). This is visualized using OpenGL-based subroutines in our code. The algorithm for *MP* construction involves choosing one initial and product glycan pair at a time (see flowchart in Supplementary Fig. S1). It is verified based on the structure of these two molecules that the product glycan can indeed be formed starting with the initial precursor, i.e. the structure of the initial glycan is a subset of the final molecule. For example, while a glycan like **S7** can be formed from **S1**, **S7** cannot be formed from a hypothetical initial glycan like Galβ1,4(Fucα1,3)GlcNAcβ1,6GalNAc-. For the chosen pair of initial-product glycans, the sequence of steps required to determine the connecting reactions involves: (i) Elimination of a single, terminal monosaccharide from the product glycan based on rules defined in the 'InferRect' method of the REACTION class, i.e. it is verified that an enzyme exists in that compartment that can yield the product glycan from the reduced entity. (ii) The above step is repeated iteratively by eliminating additional monosaccharides, one at a

Table 2. Estimation of *in situ* enzyme rate constant from wet-lab experiments

Enzyme (Abbr.)	Monosaccharide added	V_{wet}[a] (μM/h)	$[A]_{wet}$[a](μM)	$[M]_{wet}$[a](μM)	[M][b](μM)	[A][b](μM)	$K_{M,mono}$[c](μM)	$K_{M,acc}$[c](μM)
β1,4GalT-IV (GT)	Gal(●)	12.44243	500	957.88	3760	6.6	31	290
β1,3GlcNAcT (GnT)	GlcNAc(■)	0.00127	5000	0.18	7010	6.6	200	710
α2,3ST3Gal-I/II (I/II)	NeuAc(◆)	0.00097	7500	0.2966	1350	6.6	23	400
α2,3ST3Gal-IV (STIV)	NeuAc(◆)	0.00015	7500	0.2967	1350	6.6	74.1	300
α1,3FT-VII (FT)	Fuc(△)	0.14685	3000	8.47	901	6.6	16.4	3080

[a]Parameters/data are from our wet-lab experiments.
[b]*In situ* concentration of monosaccharides and acceptors (PSGL-1 glycans) are calculated based on Golgi volume and other literature data ((Tomiya *et al*., 2001) and Supplementary Table S4).
[c]K_M data are from literature cited in Supplementary Table S5.

time, in a stepwise manner and generating new 'intermediate glycans' until the number of monosaccharides in the 'intermediate' formed by progressive erosion of the product glycan is the same as the initial glycan. The structure of the initial glycan is then compared with the final reduced 'intermediate'. If the structures of these two molecules are identical, then all reactions determined above are added to the REACTION list. Otherwise, these are removed from further consideration. During the generation of the *MP*, the new glycans that are part of the reaction list are also generated (shown in dashed boxes in Fig. 2) and these are classified to have the role of 'intermediate glycan'. (iii) The above process is repeated iteratively by degrading the product glycan via all possible pathways that are allowed based on enzyme properties. (iv) For g_i initial and g_f product glycans, the above steps (i)–(iii) are repeated $g_i \times g_f$ times. (v) In the final step, reactions that are duplicated in the above analysis are deleted. A consolidated list of enzymes and glycans in the pathway are also included in the PATHWAY class. The final *MP* thus generated is said to have dimensions of m reactants $\times n$ reactions.

2.3 *SP* generation

SPs are generated by deleting one or more species in the *MP* at a time, along with associated biochemical reactions. This reduced pathway is said to be a *SP* provided the 'role' of glycans in the CLASS definition is preserved, i.e. 'initial glycans' in *MP* remain 'initial glycans' in *SP* , the 'intermediate glycans' in *MP* and *SP* are identical and so on. This test for *SPs* is termed 'subset criterion'. Depending on the number of species and reactions deleted, *SP* is of dimensions $m' \times n'$. In our simulations, a total of ${}^{m}C_j$ pathways are generated upon deleting j species from a network with m total reactants. Each of these was tested using the 'subset criterion' to determine the subset pathways. Independent SBML format files in the form of initial value problems was written for each subset model:

$$\frac{d\boldsymbol{C}}{dt} = \boldsymbol{\alpha}^{\boldsymbol{T}} \cdot \boldsymbol{v}, \boldsymbol{C}(\text{at } t=0) = \boldsymbol{C_0} \quad (2)$$

Here, the vector $\boldsymbol{v}$ contains the n individual reaction velocities [= k_{silico}·[A], Equation (1)], $\boldsymbol{C}$ contains the concentration of the m reactants, and the $m \times n$ matrix α^T contains the stoichiometric coefficients of the network. The vector $\boldsymbol{C_0}$ contains initial reactant concentration at $t = 0$.

2.4 GlycoT rate constant: *in silico* versus wet-lab data

A comparison between *in silico* estimated rate constants (k_{silico}) versus corresponding wet-lab measured values (k_{situ}) is made. 'Wet-lab' experiments were performed using lysates of HL-60 cells. In these studies, GlycoT reaction velocities (V_{wet}) were measured using enzymology-based assays (Marathe *et al*., 2008). Since these experiments were performed with cell lysates, they do not measure enzyme rate constants directly *'in situ'* in the cellular Golgi. Mathematical expressions were thus derived to estimate this rate constant (k_{situ}) from V_{wet} data [Equation (S13), Supplementary Material]. All parameter values in this equation are either fixed based on our experimental conditions or are K_M values from published literature (Table 2).

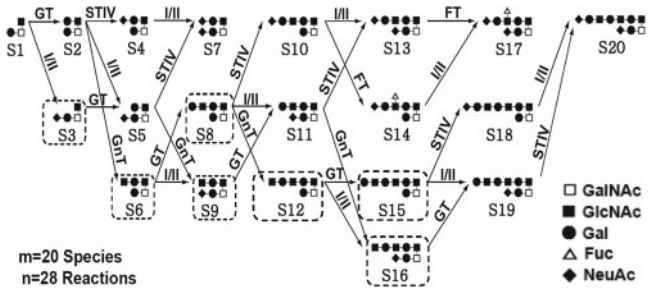

Fig. 2. *MP*: Trisaccharide **S1** is the 'initial glycan'. Additional monosaccharides are added by one of five enzymes shown on reaction arrows (α2,3ST3Gal-I/II or I/II; α2,3ST3Gal-IV or ST-IV, β1,4GalT-IV or GT; β1,3GlcNAcT or GnT and α1,3FT-VII or FT). This results in products with up to 10-monosaccharides. **S14** and **S17** which bear the sLe^X epitope are designated as 'system output'. Species in dashed boxes are 'intermediate' glycans inferred by the pathway generation algorithm. The remaining are either products or intermediates defined in Table 1. Legends present symbolic notation used to designate monosaccharides.

A description of the wet-lab assay procedure, along with mathematical derivation is provided in Supplementary Material.

3 IMPLEMENTATION AND RESULTS

The structure of carbohydrates linked to the glycoprotein PSGL-1/CD162 has been determined by Aeed *et al*. (1998) using the human promyelocytic HL-60 cells as a model system (Table 1). We performed detailed studies to measure GlycoT activities in these same cells under normal growth conditions and upon cell perturbation/stimulation (Marathe *et al*., 2008). We now demonstrate that systems biology-based modeling can be applied to bridge the gap between *in vitro* enzyme measurements and glycan structure data.

3.1 Master and subset pathways

The first step in our analysis involves the generation of the *MP*. This is done for the glycan distribution data in Table 1 (Fig. 2). This reaction network contains 20 species and 28 reactions ($m = 20$, $n = 28$). The reaction network includes all possible enzymatic reactions and reactants based on the list of enzymes provided, the starting material and the final products.

While multiple pathways can link given starting reactants and products, many of these pathways may have but minor contributions

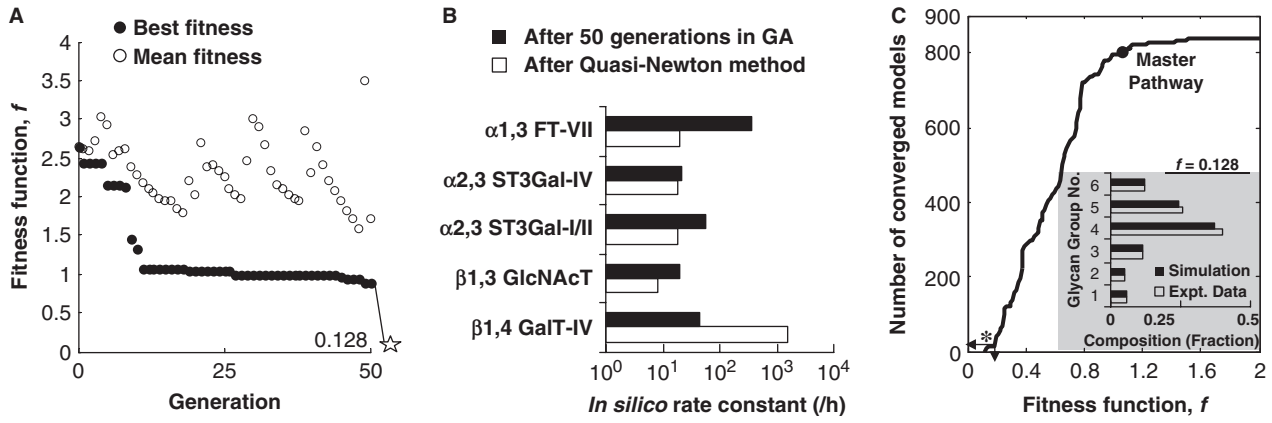

Fig. 3. Parameter estimation: (**A**) 'Best fit' (filled circle) of the GA finds the region of the local minima within 10–15 generations. Large variations in the mean GA fit (open circle) indicate that the system continues to explore a wide space range throughout the convergence process. Quasi-Newton local minimization (indicated by open star) is applied after 50 generations to determine the local minimum of the 'best fit'. (**B**) Glycosyltransferase rate constants after the GA step and after application of the quasi-Newton method. (**C**) The number of the 'converged models' chosen for further analysis depends on the choice of fitness function, f. A total 15 models 'converged' at $f < 0.19$ (indicated by asterisk). The *MP* with all possible reactions did not converge ($f = 1.04$, indicated by closed circle). Inset to (C) compares simulation fits with experimental data for glycan groups shown in Table 1, and for fitness value $f = 0.128$.

to system output. In addition, due to the substrate specificity of individual enzymes, some reactions may not take place at all. Thus, inclusion of all reactions of the *MP* for parameter estimation may not be appropriate. To address this feature, *SPs* ($m < 20, n < 28$) were automatically generated. In our case, the number of species that could be deleted from the *MP* varied from 1 to 10 (Supplementary Table S6). Deletions of 11 or more species resulted in pathways where a direct link between the initial reactant and final products was absent. Of the 616 665 pathways tested, 837 met the 'subset criterion' described in Section 2. SBML format output files were generated for all subset models for further analysis.

3.2 Model convergence

A variety of approaches were attempted to fit the 837 subset models to the experimental data. Results from the hybrid-GA are presented here, since it represents a reasonable strategy. In this regard, while GA allows determination of approximate, global solutions for non-linear complex problems in large solution space, the quasi-Newton local non-linear optimization algorithm efficiently determines the local solution using the GA result as an initial guess. Five unknown rate constants corresponding to the GlycoTs of interest (β1,4GalT-IV, β1,3GlcNacT, α2,3ST3Gal-I/II, α2,3ST3Gal-IV and α1,3FT-VII) were fit for the comprehensive *MP* and each subset model.

A representative simulation is shown in Figure 3A. Here, the best fit for GA model is seen to rapidly converge in the first 10–15 generations, while the mean fit continues to sample a wide range of variables in the parameter search space during all 50 generations of the computation. After 50 generations, the best fit from GA has likely reached the neighborhood of the global minimum. Thus, the local quasi-Newton method was applied to converge the solution to the local minima. Glycosyltransferase kinetic rate constants estimated following the GA step, and after local minimization are shown in Figure 3B. While enzyme rate constants do not change for some enzymes during the quasi-Newton step, dramatic changes are observed for others, especially β1,4GalT-IV.

The number of 'converged' models and the structure of these models depend on the definition of the fitness function, f (Fig. 3C). In our case, all models with $f < 0.19$ are said to 'converged'. This corresponds to 15 of the 837 models tested (i.e. $l = 15$). The *MP* has poor convergence ($f = 1.04$), and this observation supports the need to analyze subset models when studying glycosylation reaction networks. The number of glycan species in these converged models varied from 14 to 17 (Supplementary Table S6). A comparison of experiment versus simulation glycan distribution results for one of the converged models ($f = 0.128$) shows good agreement (inset to Fig. 3C). Here, glycan composition is collated into one of six groups according to structural information in Table 1. Glycan composition data for other models corresponding to $f = 0.436$, 0.784 and 1.10 display greater deviation from experimental results (Supplementary Fig. S2). In particular, the increase in relative error with f is evident in group 6 at $f = 0.784$ and 1.10.

3.3 Model cluster and PCA

The 15 converged models were grouped into two clusters using hierarchical analysis (Fig. 4A). The first eight of the 15 converged models clustered into one group (Cluster-I, $l_1 = 8$). These models did not have either species **S4** or **S9**. The next seven models (Cluster-II, $l_2 = 7$) displayed the absence of species **S3** and **S5**. **S13** was prominently absent in all 'converged' models.

While cluster analysis allows the cataloging of pathways into groups of related networks, such analysis does not provide a measure of relative importance of individual reactions. To address this issue, PCA of the reaction velocity matrix ($\boldsymbol{V}$) was performed (Fig. 4B). Such analysis quantifies the relative importance of individual flux in each cluster in terms of the PCA flux parameter (PCA_j). PCA_j values corresponding to individual reactions are shown in Figure 4B. Supplementary Material presents similar analysis for Cluster-II. Such analysis reveals that the flux for the **S1→S3** reaction was small compared to that for **S1→S2** in Cluster-I (Fig. 4B). Similarly, in Cluster-II, flux via species **S9** is low. Based on this analysis, it is

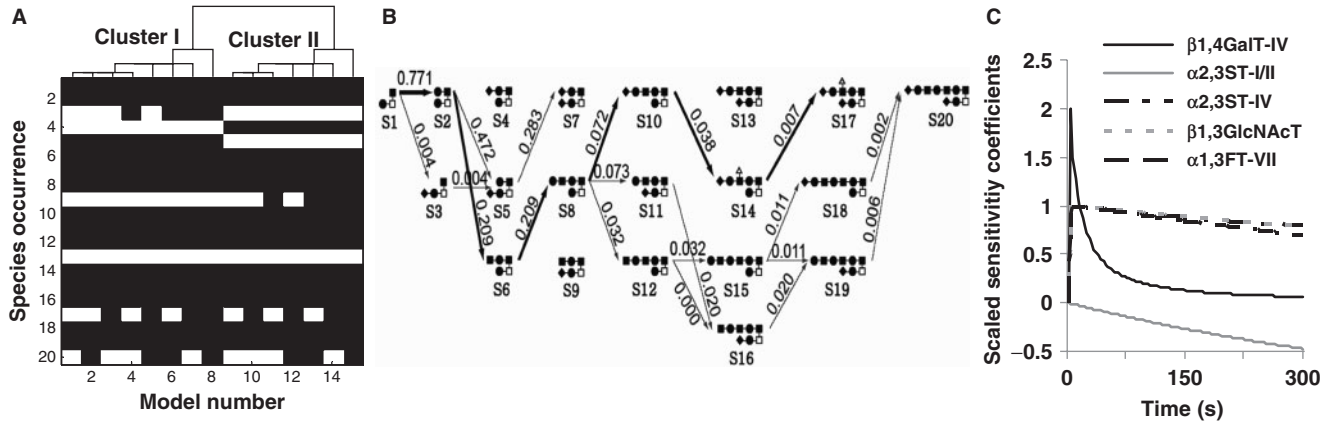

Fig. 4. Hierarchical Clustering, PCA and Sensitivity Analysis: (**A**) 15 'converged' models are clustered into two groups of 8 (Cluster-I) and 7 (Cluster-II). Species present in each model is represented in black while missing species appear as white. Dendritic tree is shown on top of the figure. (**B**) All reactions present in Cluster-I are plotted. Species (**S4**, **S9** and **S13**) that do not appear in any model of this cluster is represented as an isolated species without linking reactions. The PCA flux parameters (PCA_j) for individual reactions are provided on corresponding reaction arrows. Large PCA_j correspond to the more important reactions. (**C**) Sensitivity analysis results for Model number-8 in Cluster-I. α2,3ST3Gal-I/II expression varies inversely with cell surface sLe^X expression. Scaled sensitivity coefficient values of ∼1 for α2,3ST3Gal-IV, α1,3FT-VII and β1,3GlcNAcT suggest that variation in these enzyme activities have a proportional effect on PSGL-1 sLe^X expression.

likely that the contribution of **S3** and **S9** to this reaction network is small and further model reduction may be possible.

The above findings suggest the following propositions. First, in order to match the experimental observations of Aeed *et al.* (1998), the most reasonable sequence of enzyme action to form the sLe^X epitope on PSGL-1 involves β1,4GalT-IV (**S1→S2**), β1,3GlcNAcT (**S2→S6**), followed by β1,4GalT-IV for a second time (**S6→S8**), α2,3ST3Gal-IV (**S8→S10**), α1,3FT-VII (**S10→S14**) and finally α2,3ST3Gal-I/II (**S14→S17**). This pathway is highlighted by bold/dark arrows in Figure 4B. The dominant sLe^X form in all 15 of the converged models was **S14** and this suggests that the monosialylated glycan represents the dominant sLe^X bearing physiological ligand. Second, in both clusters, **S11** prefers to form **S16** rather than **S13**. Thus, the data suggest that **S11** is a preferred substrate for β1,3GlcNAcT (**S11→S16**) rather than α2,3ST3Gal-IV (**S11→S13**). Third, only one of the two pathways that lead to the formation of the disialylated structure **S7** may be operational in HL-60: while α2,3ST3Gal-I/II (**S2→S5**) acts prior to α2,3ST3Gal-IV (**S5→S7**) on the core-2 tetrasaccharide **S2** in cluster-II, the exact opposite (**S2→S4** followed by **S4→S7**) is the case in Cluster-I. Overall, these simulations suggest a role for α2,3ST3Gal-I/II in addition to α2,3ST3Gal-IV in regulating the formation of selectin ligands.

3.4 Sensitivity analysis

Adjoint sensitivity analysis was performed to assess the effect of infinitesimal enzyme perturbation on system output, i.e. sLe^X which is represented by the sum of **S14** and **S17** (Fig. 4C). Consistent with the findings of PCA, sensitivity analysis also suggests that α2,3ST3Gal-I/II correlates negatively with sLe^X expression on PSGL-1. One-fold reduction in the expression of these genes is expected to increase the sLe^X expression and selectin-binding function by 30–50%. In contrast, 1-fold reduction in several other enzymes (α2,3ST3Gal-IV, α1,3FT-VII, β1,3GlcNAcT) is expected to proportionally reduce the formation of sLe^X-type structures and selectin-binding function. Similar results were obtained upon subjecting models in Cluster-I and -II to such analysis (Supplementary Fig. S3). Besides sensitivity analysis which studies the effect of infinitesimal enzyme activity change, we also computed sensitivity coefficients in response to larger changes (50%) since larger perturbations are more likely in wet-lab experiments that apply silencing RNA or over-expression strategies to test specific hypothesis. These larger perturbation simulations predict similar hypothesis, that sLe^X expression varies directly with α2,3ST3Gal-IV expression and inversely with α2,3ST3Gal-I/II.

3.5 Comparison of *in silico* rate constants with wet-lab measurements

Absolute values of GlycoT rate constants estimated *in silico* (k_{silico}) were compared with corresponding values derived from experimental data (k_{situ}) (Table 3). These comparisons were based on measurements of reaction velocities in enzymology-based experiments (see Section 2.4 and Table 2 for details). As seen, all the rate constants in the two systems lie within the same order of magnitude. At the very extremes, simulation values for β1,3GlcNAcT activity were 25% that of wet-lab estimates, and α1,3FT-VII rate constants were 6.5-fold higher in the simulations compared with *in situ* estimates. We note that there are no loose parameters in these calculations and this provides confidence that computer simulations can be designed for cellular glycosylation reaction networks in order to provide realistic estimates of rate constants.

4 DISCUSSION AND CONCLUSION

By acting as the natural ligands for the selectin family of adhesion molecules, the *O*-glycans of PSGL-1 play a critical role during cell adhesion interactions between leukocytes and the inflamed endothelium that line all blood vessels (Neelamegham, 2004). The

Table 3. Comparison between experimental (situ) and simulation (silico) rate constant

Enzyme	k_{silico} (/h)[a]	k_{situ} (/h)
β1,4GalT-IV	2191.7±1148.6[b]	1505.7
β1,3GlcNAcT	7.9±0.1	32.4
α2,3ST3Gal-I/II	18.1±0.2	4.1
α2,3ST3Gal-IV	17.8±0.1	2.8
α1,3FT-VII	20.1±0.2	3.1

[a]Data are presented as mean ± SD for all 15 converged models.
[b]Individual simulation rate constants range from 1235.0/h–5161.4/h.

dominant *O*-glycan of PSGL-1 that mediates selectin recognition is expressed near the N-terminus of the protein. This glycan bears the sLe^X-epitope (**S14** and **S17**, Fig. 2), and it constitutes 2.8% of all PSGL-1 glycans (Table 1). Our modeling effort focuses on the pathways and reaction rates that control sLe^X expression. Three key outcomes include: (i) Determination of the potential reaction pathway that leads to sLe^X expression on PSGL-1. (ii) Quantitation of enzyme kinetics for GlycoTs participating in *O*-glycan formation, and comparison of these estimates with wet-lab data. (iii) Generation of experimentally testable hypothesis.

With regard to the first outcome, our modeling effort fits a reaction network of protein glycosylation with experimental data that quantify the relative abundance of various glycan species. Instead of fitting the entire *MP* with all theoretically possible reactions present, we suggest that it is more reasonable to delete selected reactions in the network to create 'subset models' prior to parameter estimation. The rationale is that all reactions predicted in a network model based on the mere presence of a given enzyme/GlycoT may not be realistic in a real, complex reaction network. This is because GlycoTs exhibit unique substrate specificity that may preclude the existence of selected reactions. Since the presence or absence of all individual reactions cannot be determined based on laboratory experiments alone, subset modeling may represent an alternate strategy that provides insight into the behavior of GlycoTs *in situ* within the cellular Golgi. In support of this, we note that convergence of the *MP* yields large fitness function values ($f = 1.04$) compared with subset models that exhibit remarkably better fits. A combination of individual data fits along with cluster, PCA and sensitivity analysis, as implemented here, may facilitate arrival at a consensus reaction network model.

With regard to the second outcome, experimental data were collected to validate enzyme rate constants predicted *in silico*. Reasonable quantitative agreement was observed between simulation and wet-lab experiments. This finding supports the overall approach proposed in this article. Here, it was noted that enzyme activities that enable *O*-glycosylation in HL-60 follow the sequence β1,4GalT-IV$\gg$β1,3GlcNAcT$\sim$α2,3ST3Gal-I/II$\sim$α2,3ST3Gal-IV$\sim$α1,3FT-VII. This observation that GalT activity is enhanced compared with other enzyme activity is consistent with studies that report on the structure of *O*-glycans on other leukocyte glycoproteins (Fukuda *et al*., 1986; Maemura and Fukuda, 1992). Here, it is reported that either Galβ or NeuAcα2,3Galβ, rather than GlcNAcβ, appears as the terminal unit at the non-reducing end of *O*-glycans.

With regard to the third outcome, the model predicts that the expression of sLe^X on the selectin-ligand PSGL-1 varies directly with sialyltransferase α2,3ST3Gal-IV expression and inversely with α2,3ST3Gal-I/II. This possibility, that the interplay between two different enzymes may regulate sLe^X expression and thus the selectin-mediated adhesion function of PSGL-1, is currently being tested using short hairpin RNA (shRNA) directed against individual sialyltransferases (A. Buffone Jr., DDM, S. A. Patil and S.N., unpublished data).

In the current article, individual enzyme activity and competition between multiple enzymes for a single substrate (like the simultaneous action of α2,3ST3Gal-IV and β1,3GlcNAcT on Galβ1,4GlcNAcβ) are considered to be key parameters regulating output glycan composition. This proposition that GlycoT activity and substrate specificity are the dominant features regulating glycan synthesis is consistent with our recent work (Marathe *et al*., 2008). Here, by systematically monitoring a range of GlycoT transcript levels, enzyme activities and corresponding carbohydrate structures, we demonstrate the existence of a semi-quantitative link between enzyme activity and glycan abundance.

Indeed, additional parameters may play a role in regulating glycosylation and these can be incorporated in the proposed object-oriented modeling framework. For example, in simulations where we adapted the current modeling strategy to analyze another experimental dataset that profiles the glycans of PSGL-1 from HL-60 (Wilkins *et al*., 1996), we noted a potential role for peptide backbone sequence in addition to the glycosyltransferase enzyme activities in regulating glycan structure (data not shown). These last computations accounted for the spatial distribution of enzymes in the Golgi by using the multi-compartment simulation features available in the modeling framework. Consistent with these findings, experimental data from other groups also suggest a role for the polypeptide sequence in regulating the initiation of O-linked glycosylation at Serine/Threonine sites (Gerken *et al*., 2002). Overall, future investigations that introduce additional model parameters and experimental data may evaluate the role of peptide sequence and other features in regulating glycan profiles.

The current article represents a starting point that can lead to further development of systems-based models in the field of Glycomics. While GAs are used for sampling the global parameter space here, this is a time consuming task. Simulations of each subset model takes ~20 min on a Pentium 4 3.0 GHz processor with 2 GB RAM. This large computational time can likely be reduced in the future by using a more efficient GA algorithm or alternative parameter estimation methods. Additional development of Monte Carlo-based sampling methods to test only selected subset models for 'convergence' can allow analysis of larger, more complex reaction networks in reasonable time. Further, while we describe here a reaction network for glycosylation with focus on PSGL-1, it is apparent that this approach can be extended to studies of other proteins for which glycan profiles are being measured by the Glycomics community (Fukuda *et al*., 1986; Muller and Hanisch, 2002; Royle *et al*., 2002; Satomaa *et al*., 2002). In the long run, definition of XML-based notations to describe glycans and the incorporation of this into our codes will be beneficial. Progress has been initiated in this direction with some investigators suggesting XML-based notation for glycans (Kikuchi *et al*., 2005; Sahoo *et al*., 2005).

In summary, we conclude that simulation methods as described in this article may provide a novel strategy to link systems-based modeling methods with experimental data on enzyme rates and glycan structural information.

Funding: National Institutes of Health (HL63014).

Conflict of Interest: none declared.

REFERENCES

Aeed,P.A. *et al*. (1998) Characterization of the O-linked oligosaccharide structures on P-selectin glycoprotein ligand-1 (PSGL-1). *Glycoconj. J.*, **15**, 975–985.

Amano,K. *et al*. (2008) Engineering of mucin-type human glycoproteins in yeast cells. *Proc. Natl Acad. Sci. USA*, **105**, 3232–3237.

Cao,Y. *et al*. (2003) Adjoint sensitivity analysis or differential-algebraic equations: the adjoint DAE system and its numerical solution. *Siam J. Sci. Comput.*, **24**, 1076–1089.

Fukuda,M. *et al*. (1986) Structures of O-linked oligosaccharides isolated from normal granulocytes, chronic myelogenous leukemia cells, and acute myelogenous leukemia cells. *J. Biol. Chem.*, **261**, 12796–12806.

Gerken,T.A. *et al*. (2002) Mucin core O-glycosylation is modulated by neighboring residue glycosylation status. Kinetic modeling of the site-specific glycosylation of the apo-porcine submaxillary mucin tandem repeat by UDP-GalNAc:polypeptide N-acetylgalactosaminyltransferases T1 and T2. *J. Biol. Chem.*, **277**, 49850–49862.

Hakomori,S. (2001) Tumor-associated carbohydrate antigens defining tumor malignancy: basis for development of anti-cancer vaccines. *Adv. Exp. Med. Biol.*, **491**, 369–402.

Jaeken,J. and Matthijs,G. (2007) Congenital disorders of glycosylation: a rapidly expanding disease family. *Annu. Rev. Genomics Hum. Genet.*, **8**, 261–278.

Karsan,A. *et al*. (1998) Leukocyte Adhesion Deficiency Type II is a generalized defect of de novo GDP-fucose biosynthesis. Endothelial cell fucosylation is not required for neutrophil rolling on human nonlymphoid endothelium. *J. Clin. Invest.*, **101**, 2438–2445.

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

Krambeck,F.J. and Betenbaugh,M.J. (2005) A mathematical model of N-linked glycosylation. *Biotechnol. Bioeng.*, **92**, 711–728.

Liu,G. and Neelamegham,S. (2008) In silico Biochemical Reaction Network Analysis (IBRENA): a package for simulation and analysis of reaction networks. *Bioinformatics*, **24**, 1109–1111.

Liu,G. *et al*. (2005) Sensitivity, principal component and flux analysis applied to signal transduction: the case of epidermal growth factor mediated signaling. *Bioinformatics*, **21**, 1194–1202.

Lowe,J.B. (2003) Glycan-dependent leukocyte adhesion and recruitment in inflammation. *Curr. Opin. Cell. Biol.*, **15**, 531–538.

Maemura,K. and Fukuda,M. (1992) Poly-N-acetyllactosaminyl O-glycans attached to leukosialin. The presence of sialyl Le(x) structures in O-glycans. *J. Biol. Chem.*, **267**, 24379–24386.

Marathe,D.D. *et al*. (2008) Systems-level studies of glycosyltransferase gene expression and enzyme activity that are associated with the selectin binding function of human leukocytes. *FASEB J*. [Epub ahead of print; August 26, 2008], doi: 10.1096/fj.07-104257.

Monica,T.J. *et al*. (1997) A mathematical model of sialylation of N-linked oligosaccharides in the trans-Golgi network. *Glycobiology*, **7**, 515–521.

Muller,S. and Hanisch,F.G. (2002) Recombinant MUC1 probe authentically reflects cell-specific O-glycosylation profiles of endogenous breast cancer mucin. High density and prevalent core 2-based glycosylation. *J. Biol. Chem.*, **277**, 26103–26112.

Neelamegham,S. (2004) Transport features, reaction kinetics and receptor biomechanics controlling selectin and integrin mediated cell adhesion. *Cell Commun. Adhes.*, **11**, 35–50.

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.

Royle,L. *et al*. (2002) An analytical and structural database provides a strategy for sequencing O-glycans from microgram quantities of glycoproteins. *Anal. Biochem.*, **304**, 70–90.

Sahoo,S.S. *et al*. (2005) GLYDE-an expressive XML standard for the representation of glycan structure. *Carbohydr. Res.*, **340**, 2802–2807.

Satomaa,T. *et al*. (2002) O-glycans on human high endothelial CD34 putatively participating in L-selectin recognition. *Blood*, **99**, 2609–2611.

Shelikoff,M. *et al*. (1996) A modeling framework for the study of protein glycosylation. *Biotechnol. Bioeng.*, **50**, 73–90.

Tomiya,N. *et al*. (2001) Determination of nucleotides and sugar nucleotides involved in protein glycosylation by high-performance anion-exchange chromatography: sugar nucleotide contents in cultured insect cells and mammalian cells. *Anal. Biochem.*, **293**, 129–137.

Umana,P. and Bailey,J.E. (1997) A mathematical model of N-linked glycoform biosynthesis. *Biotechnol. Bioeng.*, **55**, 890–908.

Wilkins,P.P. *et al*. (1996) Structures of the O-glycans on P-selectin glycoprotein ligand-1 from HL-60 cells. *J. Biol. Chem.*, **271**, 18732–18742.