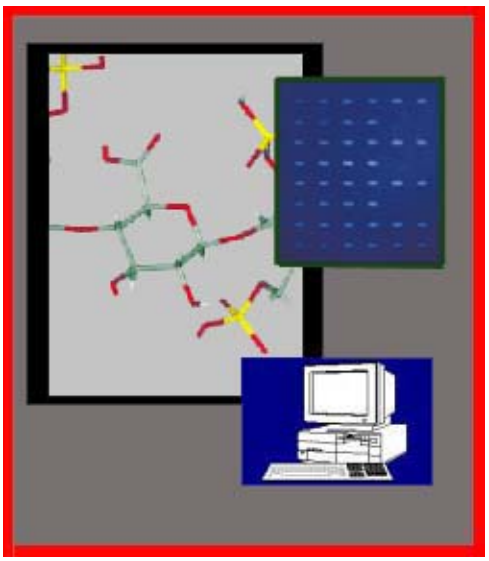
Using Bioinformatics to Spearhead the Glycomics Revolution



Srinivas R. Viswanathan
Genomics & Bioinformatics (MB&B 452a)
TA: Haiyuan Yu
December 13, 2002

1. THE IMPORTANCE OF GLYCOSYLATION

Once dismissed as nuisances that hampered protein purification[1], glycosylations are now appreciated as functionally important co- and post-translational modifications. Glycosylations occur in five distinct groups, with N-linked and O-linked glycosylations being the most common[2]. N-glycosylations are found on Asn and O-glycosylations are found on Ser or Thr[2]. Unlike proteins, DNA, and RNA, glycosylations can have residues connected by various linkage types and can be highly branched, resulting in structures of astounding complexity[3].

Up to 70% of all proteins may be N-glycosylated[19] at strict consensus sequences (N-X-S/T)[4]. Accurate prediction of O-glycosylation sites has proven considerably harder[2], although neural network models have had reasonable success[5,6]. Glycosylation is important in protein folding, oligomerization, sorting, and transport[7], and deficiencies in glycosylation have been implicated in several diseases[8,9,10]. Bioinformatics has been crucial for elucidating glycan structures and will likely become even more prevalent in glycobiology as the glycomics revolution begins to demand organization and manipulation of carbohydrate data on a large scale.

2. MODELING GLYCOSYLATION STRUCTURES

The determination of glycosylation structures can be broken down into two problems: determining the primary structure of the glycan (e.g. the manner in which the individual residues are linked) and determining the most probable conformation of the glycan (e.g. the three-dimensional structure). Bioinformatics has been successful in integrating experimental techniques (e.g. NMR) and theoretical approaches (e.g. molecular modeling) to help determine both the primary structure and conformation of glycans.

A. Determining the Primary Structure of Glycans

Currently, NMR is the only technique that can unambiguously determine both the anomericities (α or β) and linkage-types (1,4-; 1,6-; etc.) of residues in an oligosaccharide[11]. The specific ^{13}C and ^{1}H NMR shift patterns of a particular oligosaccharide are widely used to determine its primary structure[12]. Several computational methods allow for efficient primary structure determination, and four main ones are described below.

i. Structural-Reporter-Group Approach

The structural-reporter-group approach relies on the SUGABASE database, which combines ^{13}C and ^{1}H chemical shifts with CarbBank Complex Carbohydrate Structure Data[13]. This approach is based on the fact that specific linkage compositions and structural motifs display characteristic shifts outside of the 3-4 ppm range. The user enters proton or carbon chemical shifts and receives a list of all possible structural hits[12]. However, the list must still be manually inspected to see which hits are consistent with experiment.

ii. Computer-assisted spectrum evaluation of regular polysaccharides (CASPER)

CASPER performs primary structure analysis on both linear and multi-branched oligosaccharides[14, 15]. The user must enter chemical shifts, coupling constants ($^1J_{CH}$ or $^3J_{HH}$), sugar composition, and linkage composition obtained from biochemical analyses[12]. CASPER will then generate all possible structures and simulate their proton and carbon NMR spectra. Structures incompatible with the coupling constants are removed. The generated spectra of the remaining structures are compared against the experimental data and ranked based on lowest total difference in chemical shifts[12, 14,15].

iii. Computer-Assisted Structure Determination With ^{13}C-NMR Data

The computer-assisted structure determination approach originally predicted the structures of only unbranched polysaccharides[16] but has now been applied to predict branched glycan structure as well[17]. The program, BIOPSEL, utilizes a spectral database that stores the ^{13}C chemical shifts of monomers, dimers, and trimers. The user must enter the experimental ^{13}C spectrum and monosaccharide composition for the glycan of interest. BIOPSEL then uses the values in the database to calculate theoretical ^{13}C NMR spectra for all possible structures given the monosaccharide composition. (The theoretical spectra are produced by generating subspectra for each residue with substitution effects and combining these subspectra into the whole spectrum[17].) Theoretical spectra are then compared to the experimentally obtained spectrum and ranked in order of similarity. This approach can predict the primary structure of oligosaccharides up to six units in length with high accuracy[17].

iv. Artificial Neural Networks (ANNs)

While the three approaches above look at individual chemical shift-structure relationships, ANNs have the capacity to look at the entire proton spectrum as a whole and apply pattern-matching techniques. An ANN can be trained on a subset of spectra with known structures, and can then be used to determine the corresponding structures for unknown spectra[18]. If spectra with noise are included in the training set, the ANN performs particularly well[12]. The disadvantage of neural network approaches in general is that they remain "black boxes" with respect to methodology.

B. Determining the Three-Dimensional Structure (Conformation) of Glycans

Primary structure alone is often inadequate for understanding the functional roles of glycans; for this, conformational information is needed. Unfortunately, however, glycoproteins are notoriously hard to crystallize[19, 20], largely because of the conformational flexibility of their glycan antennae[19]. In the rare instances where glycoproteins have been successfully crystallized, electron density of the glycan has usually been so low that no defined spatial conformation beyond the rigid core region can be determined[21]. Furthermore, glycans may exist in several different conformations in solution[22]. Thus, finding secondary and tertiary structural motifs from crystallography has proven difficult[19], even though such motifs may exist[23].

The structural conformation of a glycosylation depends mainly on the particular ϕ, ψ, and ω torsion angles about the glycosidic bond, since this is where most rotation occurs[11, 28]. Glycans may form hydrogen bonds both internally and with the solvent[23], and differences in the nature of hydrogen bonding may help to choose between two possible conformations. Steric

and hydrophobic factors such as van der Waals interactions and the *exo*-anomeric effect (a stereoelectronic effect of lone pairs on the linkage oxygen) [23, 24,11] also affect glycan conformation.

i. Molecular Dynamics (MD) Approaches

By using approximations from classical mechanics, molecular modeling can be used to calculate the energy of a macromolecule in a particular conformation. MD simulations model dynamic behavior of the molecule[11]. Each atom is assigned a velocity that can change based on the forces present. Tiny (~1 fs) steps are then taken and atom positions are updated[25]. This approach has been widely used to model proteins, polynucleotides, and carbohydrates[25, 26]. Several MD protein simulation packages such as AMBER, DISCOVER, CHARMm, and GROMOS[25] have also been applied to carbohydrate modeling using new parametrizations[27, 28, 29]. These packages calculate the energy of the glycan in a particular conformation with a potential energy function E(x,y,z), in which each point has a contribution to the energy based on its value for bond, angle, torsion, Lennard Jones, and electrostatic parameters[26]. The term "force field" is given to the set of parameters used to generate the potential energy function for a molecule.

Many force fields currently exist for modeling carbohydrates, and there is disagreement as to which ones produce models most consistent with experimental data [25, 26]. Several force fields model carbohydrate structure *in vacuo* and consider primarily steric factors[22]. The dielectric may be set at $\varepsilon=1$(vacuum) or $\varepsilon=80$ (to account for the charge screening of water)[23]. While such force fields often produce structures that are fairly consistent with experimental data[30], the interactions taken into account are not truly indicative of those present in solution. For example, the Hard Sphere exo-anomeric (HSEA) force field takes only van der Waals interactions and the ϕ torsion angle into account, completely neglecting hydrogen bonding and dipolar effects[22]. Other force fields (e.g. AMBER, CHARMm and GROMOS) take energetic contributions from bond stretching, angle bending, torsion, and non-bonded interactions into account[22]. These force fields are appealing because they can model explicit water molecules in MD simulations[22, 28]. However, parameter sets must be carefully chosen because each set yields a different degree of approximation and is appropriate in a different context. Still other force fields such as MM2/MM3 use complex mathematical formulae to accurately model subtle sugar ring puckering and bond length variations[19, 22]. The weakness of MM2/MM3 is that it cannot accurately account for hydrogen bonding, and so is currently limited to gas-phase and crystal simulations[22].

MD carbohydrate simulations employ various methods to arrive at the conformation with the global free energy minimum without becoming trapped in a local minimum[25]. Steepest descent minimization follows the energy gradient down into energy-minimum valleys[25]. Monte Carlo methods employ a random number seed to randomly move through states and sample a large conformational space, ultimately arriving at the global minimum[29]. A third technique, simulated annealing,

involves heating the system to a very high temperature (e.g. 750K), then running MD at incrementally lower temperatures and minimizing (e.g. by steepest descent) to find the global minimum[20].

ii. NMR-Based Approaches

NMR measurements of nuclear Overhauser effects (NOEs) are often used to confirm structures predicted by molecular dynamics[31]. NOEs, which give the distance between two nuclei, can be used to find the distance between two protons across a glycosidic bond[11]. Any distance usually yields several possible conformations. If enough such distance constraints are present, a single conformation can be identified. The main disadvantage to using NMR in three-dimensional structure prediction is that it yields time-averaged conformations[11]; this is particularly a problem when a glycan is present in several very different conformations in solution.

3. Extending Glycomics to a Large Scale

A. Current Databases and Shortcomings

The ability of Bioinformatics to organize large stores of information has been invaluable in proteomics and genomics. Sadly, however, glycomics has been slow in moving toward large-scale analysis of data[3]. Funding for SUGABASE[13] and CarbBank[32] has been discontinued, the BOLD database[33] is limited to O-glycans, and the GlycoSuite[34] curated relational Database contains only about 7000 sources. Furthermore, glycan databases have lacked the organization and visibility of protein and gene banks. A promising new database, SWEET-DB[35], integrates information on structure (from CarbBank), NMR shifts (from SUGABASE), and 3D coordinates (generated with SWEET-II) and could become the dominant database in glycobiology. Still, for the glycomics revolution to take off, the stores of data in databases such as SWEET-DB must be increased, and this data must be kept organized with the help of bioinformatics.

B. Applying Proteomics and Genomics Techniques to Glycomics

Glycobiology could benefit from a large, collaborative effort to find the sequences and coordinates of as many glycans as possible, using the techniques described above. A glycan bank comparable in scale to GenBank or PDB would make bioinformatics as important in glycomics as it is in genomics and proteomics, and would allow the answering of many new types of questions (see Appendix). For example, structural diversity seen in glycans is reminiscent of that seen in proteins. Thus, it should be possible to structurally align glycans by iterative programming, identify structural motifs, and correlate these motifs to function. If analysis showed that certain sequence classes of glycans perform certain functions, one could develop Hidden Markov Models to generate novel sequences of a desired class. From sequence, one could then obtain glycan structure using a Molecular Builder program (e.g. POLYS[36]). Lastly, one could perform large-scale screens for certain types of glycans using carbohydrate arrays[37] (such technology is currently in development). In sum, once large stores of glycan data become available, many of the techniques developed for proteomics and genomics can be directly applied to glycomics.

REFERENCES

[1] Voet D and Voet JG. *Biochemistry*. 2nd Edition. Wiley: New York, 1995. pgs. 266-274.

[2] Spiro RG (2002) **Protein glycosylation: nature, distribution, enzymatic formation, and disease implications of glycopeptide bonds.** *Glycobiology*, **12**, 43R – 56R.

[3] Von der Leith C-W (2002) **Expanding proteomics to glycobiology: biocomputing approaches undestanding the function of sugar.** *Pac. Symp. Biocomput, 7: 283-4*.

[4] Marshall RD (1974) **The nature and metabolism of carbohydrate peptide linkages of glycoproteins**. *Biochem. Soc. Symp* , **40**, 17–26.

[5] Hansen JE et al. (1998) **NetOglyc: prediction of mucin type O-glycosylation sites based on sequence context and surface accessibility.** *Glycoconj J,* **15**, 115-130.

[6] Gupta R et al. (1999) **Scanning the available Dictostelium discoideum proteome for O-linked GlcNAc glycosylation sites using neural networks.** *Glycobiology*, **9**, 1009-22.

[7] Helenius A and Aebi M (2001) **Intracellular functions of N-linked glycans.** *Science*, **291**, 2364-9.

[8] Freeze, H.H. and Westphal, V. (2001) **Balancing *N*-linked glycosylation to avoid disease.** *Biochimie*, **83**, 791–9.

[9] Schachter, H. (2001) **Congenital disorders involving defective *N*-glycosylation of proteins.** *Cell Mol. Life Sci.*, **58**, 1085–1104.

[10] Lübke, T., Marquardt, T., von Figura, K., and Korner, C. (1999) **A new type of carbohydrate-deficient glycoproteins syndrome due to a decreased import of GDP-fucose into the Golgi.** *J. Biol. Chem.*, **274**, 25986–25989.

[11] Wormald MR et al. (2002) **Conformational studies of oligosaccharides and glycopeptides: complementarity of NMR, X-ray crystallography, and molecular modeling.** *Chem. Rev., **102**, 371-86.*

[12] Duus JO et al. (2000) **Carbohydrate structural determination by NMR spectroscopy: modern methods and limitations.** *Chem. Rev., **100**, 4589-614*.

[13] SUGABASE Carbohydrate Database. Online. Available: http://www.boc.chem.uu.nl/sugabase/databases.html

[14] Jansson PE et al. (1991) **CASPER: a computer program used for structural analysis of carbohydrates.** *J. Chem. Inf. Comput. Sci.,* **31**, 508-16.

[15] Stenutz R et al. (1998) **Computer-assisted structural analysis of oligo- and polysaccharides: an extension of CASPER to multibranched structures.** *Carbohydr. Res,* **306,** 11-17.

[16] Lipkind GM et al. (1988) **A computer-assisted structural analysis of regular polysaccharides on the basis of 13C-n.m.r. data.** *Carbohydr Res.,* **175,** 59-75.

[17] Toukach FV and Shashkov AS (2001) **Computer-assisted structural analysis of regular glycopolymers on the basis of 13C NMR data.** *Carbohydr Res.,* **335**, 101-14.

[18] Meyer B et al. (1991) **Identification of the ^{1}H-NMR spectra of complex oligosaccharides with artificial neural networks.** *Science, **251,*** 542-4.

[19] Bohne A and von der Lieth CW. (2002) **Glycosylation of proteins: a computer based method for the rapid exploration of conformational space of N-glycans.** *Pac. Symp. Biocomput.*, **7**, 285-96.

[20] Rutherford T et al. (1995) **Influence of the extent of branching on solution conformations of complex oligosaccharides: a molecular dynamics and NMR study of a penta-antennary 'bisected' N-glycan.** *Biochemistry*, **34,** 14131-7.

[21] Petrescu A et al. (1999) **A statistical analysis of N- and O-glycan linkage conformations from crystallographic data.** *Glycobiology*, ***9,*** 343-52.

[22] Woods RJ (1998) **Computational carbohydrate chemistry: what theoretical methods can tell us.** *Glycoconj. J.,* ***15,*** 209-16.

[23] Woods RJ (1995) **Three-dimensional structures of oligosaccharides.** *Curr. Opin. Struct. Biol.,* ***5,*** *591-8.*

[24] Wolfe S et al. (1979) **On the magnitudes and origins of the 'anomeric effects', 'exo-anomeric effects', 'reverse anomeric effects', and C-X and C-Y bond lengths in XCH_2YH molecules.** *Carbohydr. Res.,* ***69,*** *1-26.*

[25] Gerstein M (2002) **Simulation.** Class Notes. MB&B 452a, Genomics & Bioinformatics.

[26] Harvey S (2002) **Quanta Tutorial 3.** Online. Available: http://uracil.cmc.uab.edu/~harvey/Tutorials/bmg759/QUANTAtut3.html

[27] Homans SW (1990) **A molecular mechanical force field for the conformational analysis of oligosaccharides: comparison of theoretical and crystal structures of Man alpha 1-3Man beta 1-4GlcNAc.** *Biochemistry,* **29**, 9110-8.

[28] Kuttel M et al. (2002) **Carbohydrate solution simulations: producing a force field with experimentally consistent primary alcohol rotational frequencies and populations.** *J. Comput. Chem.,* ***23,*** *1236-43.*

[29] Zhang H et al. (1996) **Conformational analysis of two glycoproteins: a Monte Carlo simulated annealing approach using a soft-sphere potential.** *Carbohydr. Res.,* ***284,*** *25-34.*

[30] Rutherford TJ et al. (1994) **Restrained vs free dynamics simulations of oligosaccharides: application to solution dynamics of biantennary and bisected biantennary N-linked glycans.** *Biochemistry,* **33,** 9606-14.

[31] Imberty A (1997) **Oligosaccharide structures: theory versus experiment.** *Curr. Opin. Struct. Biol., 7, 716-23.*

[32] CarbBank and the CCSD. Online. Available: http://bssv01.lancs.ac.uk/gig/pages/gag/carbbank.htm.

[33] Cooper CA et al. (1999) **BOLD – a biological O-linked glycan database.** *Electrophoresis*, **20**, 3589-98.

[34] Cooper CA et al. (2000) **GlycoSuiteDB: a new curated relational database of glycoprotein glycan structures and their biological sources.** *Nucleic Acids Res.,* ***29,*** *332-5.*

[35] Lob A et al. (2002) **SWEET-DB: an attempt to create annotated data collections for carbohydrates.** *Nucleic Acids Res.,* **30***,* 405-8.

[36] Engelson SB et al. (1996) **A molecular builder for carbohydrates: application to polysaccharides and complex carbohydrates.** *Biopolymers,* **39**, 417-33.

[37] Houseman BT and Mrksich M (2002) **Carbohydrate arrays for the evaluation of protein binding and enzymatic modification.** *Chem. Biol.,* **9**, 443-54.