Modelling the biosynthesis of glycan molecules from existing glycan structures towards the automation of glycan profiling
Problem statement

Alessio Ceroni
Imperial College
a.ceroni@imperial.ac.uk

1 Problem background

Glycosylation is the most common post-translation modification of proteins and glycans are increasingly recognised as having important biological functions related to development and diseases.

The process of protein glycosylation starts with the addition of a precursor glycan molecule to either an asparagine (N-linked) or a threonine/serine residue (O-linked). The precursor is trimmed down (in case of N-linked glycosylation) and then the molecule is enlarged by the sequential addition of further monosaccharide components carried out by specific enzymes. The final molecule is composed by a number of monosaccharides ranging from 2 to 20-30. The monosaccharides can be linked at several positions around their ring, thus resulting in branched structures.

Mass spectrometry is the main technique used to identify the structures of the glycan molecules expressed by a cell or a group of cells in a tissue. Normally the glycans are first detached from their protein counterparts and then a mass spectrum of this mixture of molecules is produced. By the interpretation of the mass signals a glycan profile for the analysed sample can then be derived.

CS translation

After RNA is translated into proteins, Glycosylation happens. Glycans are molecules which are attached to the proteins in this process. Glycans consist of atomic components called monosaccharides, and grow in a tree shape (the tree getting a size up to 20-30).

Mass spectrometry is a (noisy) process used to retrieve the mass of a (fragment of) a glycan. This can be used as a hint for the (unknown) structure of the glycan.

2 Available data

The Consortium for Functional Glycomics (http://www.functionalglycomics.org) is a large research initiative formed to define the paradigms by which protein-carbohydrate interactions mediate cell communication. The strategy to achieve this goal is to work with the scientific community to create unique resources and services that Participating Investigators can utilize in their own research. Data produced by the Scientific Cores, much of which is generated using samples provided by investigators, are uploaded into the CFG relational database and are available from the project website.

Glycan profiling experiments performed by the Analytical Glycotechnology Core (C) identify the presence of various N- and O-linked glycans in human and mouse tissues. For each species and tissue type there are several profiles available. Each profile is essentially a list of proposed glycan structures matching the measured mass signals. In the website each profile is pictured as a spectra annotated with cartoons representing the glycan structures, but the profiles are also available as a list of structures encoded using a linear string format.

CS translation

Glycan molecules are represented as labelled trees. Vertices have discrete labels for the monosaccharide type and real labels for the individual monosaccharide mass (mass is a deterministic function of type). Profiles are histograms of the mass of glycan molecules which is computed as the sum of the mass of the vertices. Each tissue-specie pair is associated with a specific set of glycan molecules.

3 Sources of uncertainty and noise in the problem

The various arrangements of monosaccharide in the tree structure of the glycan molecules and the existence of various classes of monosaccharide having the same atomic compositions result in a high number of possible glycan structures with the same mass. In order to assign a limited set of structures to a given mass signal, additional information about the biosynthesis of the glycans must be used. This information consists of the specificity of the enzymes which recognise a part of the growing molecule and attach a defined monosaccharide to it. Information about the biosynthetic process is not complete and is mostly known only for human/mammalian organisms. For the specific purposes of this exercise, this information is supposed to be totally unknown although the way experts use it may inspire the proposed solution. Additional experiments can be carried out to disambiguate between the possible structures but this data will not supposed to be available in this setting. Other basic knowledge such as the monosaccharide masses is given.

CS translation

Information on the total mass is not enough to unambiguously derive the tree structure of the glycan. This is because each monosaccharide does not have a unique mass value and vertices order and hierarchical arrangement information is just not derivable (as the total mass is the result of a commutative operation). Glycan molecules that have the same mass contribute to the same bin in the histogram profile.

The tree structures are built in an incremental fashion (i.e. no revision of previous building decisions is allowed at later steps). Specific biological entities (enzymes) can attach new vertices at the actual frontier of the tree (the glycan as it is being synthesized). Each enzyme can be therefore interpreted as a rule of a generative mechanism. The attachment is dependent on the actual status of the tree frontier. The enzyme set is tissue-specie specific. There exist similarity relationships across tissue-species pairs (i.e. species phylogenetically close are likely to share the same set of enzymes). Data is available only for two species: human and mouse. The rules and the enzyme set is supposed to be unknown in this setting.

The tissue-specie specific abundance of each enzyme determines the weight or probability assigned to each rule. This in turn determines the probability (the ratio of the masses) of the trees (glycan molecules) in the given (tissue-specie specific) histogram (profile).

4 Questions the domain experts would like to have answered

Given a set of available profiles derived by experts:

1. use the assigned structures to create a model of the biosynthetic process and produce the minimal set of all possible glycans that can be produced by a given organism; this set should contain all the assigned structures and should not contain any isomer of the existing structures which has never been previously assigned; the quality of the model will be validated leaving out from the training set all the molecules with a specific mass;

2. use the model to produce a complete profile for a given set of mass values; in this setting the information given by the complete list of mass values should be used to correlate the assignments of the single signals, given that the presence of a particular structure would reinforce the possibility of finding all the molecules that can be produced from it by addition of monosaccharides.

CS translation

1. Determine the specie specific generative model for the tree (glycan molecules) set. Evaluate the model by allowing training over a subset of the trees

(e.g. all glycan molecules with total mass in a specific range) and testing the generalization properties of the model over the remaining instances.

2. Determine the tissue-specie specific generative model for the tree (glycan molecules) set. Evaluate the model by allowing training over a subset of the trees (e.g. all glycan molecules with total mass in a specific range) and testing the generalization properties of the model over the remaining instances.