Final Report
OTKA 83857 project

1. Introduction, state-of-art

It is well known that glycosylation of proteins is both important and widespread. At the time of writing the project proposal, techniques studying protein glycosylation were limited in scope, limited in analytical figures of merit and were also expensive. In consequence, our knowledge on protein glycosylation was also limited. The aims of the project were to advance state-of-art methodologies, and to apply them to real-world samples.

In the past 5 years (i.e. between writing the proposal and finishing the project) the state-of-art of analytical methods for studying glycosylation improved tremendously. In spite of this progress, glycoprotein analysis remains difficult and expensive (especially compared to proteomics). For this reason advances in the last 5 years have not been sufficient to get a breakthrough in the biological/biochemical understanding of glycosylation, and understanding the role of glycosylation in pathological processes.

2. Difficulties encountered and changes in the project

In the course of accomplishing the project, a number of changes and modifications have occurred, and various difficulties were encountered. Some were administrative (approved by OTKA); some were due to objective complications (e.g. relocation of MTA TTK); and some were related to unforeseen developments in science and methodology. The most important items are listed below:

a) Administrative changes: Some junior people left the team (change of employment, M. Hegedus, B. Gyorgy), their role was taken over by new personal and PhD students joining the team (D. Bene, B. Bobaly, A. Jeko, E. Toth), approved by OTKA.
b) Delays in sample collection occurred due to delays obtaining in ethical permits, but sample collection was successfully finished by the end of the project (these delays were reported in the annual reports).
c) Unforeseen relocation and reorganization of MTA TTK, and also the financial difficulties of MTA TTK caused further delays in the project. These prompted us to ask for and obtain permission to prolong the project by one year (with no extra funding).
d) Improvements in state-of-art mass spectrometry technology and capabilities of our aging mass spectrometer (Waters QTOF, from 2004) created a gap between the required and available figures of merit. In particular, sensitivity of our WATERS QTOF was 100 or 1000 times worse that that available in the market in 2014/5; and this was a crucial disadvantage in the glycoproteomics field. We did manage to get good results and publish them in high quality research papers; but had to delay publication of our most important

discoveries. In mid-2015 we got a state-of-art Bruker mass spectrometer. Now we are reproducing our results obtained on the Waters QTOF instrument at a much higher sensitivity (and better accuracy, and lower false discovery rate).. We think these new measurements are very important, and we plan to publish these in very prestigious journals (we are aiming at a Nature paper in 2016/7). We consider this to be a better choice (even though it is outside the timeframe of the present OTKA project) than publishing our available "preliminary" results obtained using an old instrument in a middling journal.

e) Both our own and the state-of-art understanding of glycosylation processes were hindered by i) smaller than expected differences in glycosylation patterns in healthy and diseased persons and ii) a still present gap between required and available (state-of-art) figures of merit in analytical methods for glycosylation analysis.

For these reasons we were not able to fully realize our aims characterizing the connection between glycosylation and cancer.

3. Developed methodologies

A major part of the OTKA project was dedicated to developing novel, mass spectrometry based methodologies for glycosylation analysis. The aim was to develop methods, which yield site-specific glycosylation pattern for N-glycoproteins and which are sufficiently sensitive and selective, so that glycosylation profiles of individual proteins can be determined from a small amount of blood plasma. Achieving this aim required developing complex (multi-step) analytical procedures. The main steps of analysis are the following: 1) Enrichment of proteins/glycoproteins/glycopeptides; 2) Sample preparation, digestion 3) Nano.UHPLC-MS/MS experiment design; 4) Evaluation of glycopeptide (tandem) mass spectra; 5) Glycoform quantitation; 6) Protocol established for studying glycosylation patterns.

We have developed/optimized methods for each of these steps. In several cases various alternatives were evaluated. Often the developed methods/techniques reached beyond the state-of-art, and these have been published in prestigious journals (most publications listed below are in Q1 quartile in SCIMago in analytical chemistry and in Q1 or Q2 in biochemistry). Here we summarize the major methodologies developed in the course of the OTKA project.

1) *Enrichment of proteins/glycoproteins/glycopeptides from biological fluids.*

This is the first step in most analytical schemes; indispensable for glycoprotein analysis. The most common enrichment techniques utilize affinity chromatography; often in sequence. The first stage is usually "depletion", selective removal of the most abundant plasma proteins (in our case IgG and albumin) by affinity capture. In a subsequent stage we have compared the efficiency of strong cation exchange; lectin affinity capture and the use of boronic acid stationary phase in order to enrich glycoproteins and glycopeptides (the latter obtained by proteolytic digestion of plasma). The results showed boronic acid chromatography as most promising; but even this shown insufficient selectivity, so we did not follow this route. We have pioneered a different direction, based on RP chromatography of intact proteins. This is a fairly new direction, made possible by the development of large pore diameter columns. A main feature that proteins/glycoproteins are resolved based on the properties of the polypeptide chain, while the type and size of the attached sugar units have very small influence on retention. This does not allow separating various glycoforms of a given glycoprotein. This is, in fact, a major advantage, as all glycoforms of a given glycoprotein are isolated in the same fraction; allowing unbiased characterization of the glycosylation pattern is subsequent analysis. Detailed comparison of various enrichment techniques; characteristics of RP-HPLC of intact proteins; and performance of RP-HPLC isolation of plasma fractions are primarily described in the following publications: *Ozohanics et al, Rapid Comm Mass Spectrom, 2012; Bobaly et al, J. Chrom A, 2014; Toth et al, J. Pharm Biomed Anal, 2014.*

2) *Sample preparation, digestion.*

After enrichment of glycoproteins the next step is proteolytic digestion and preparation of the sample for subsequent nano.UHPLC-MS/MS analysis. Although there are many such protocols available, the main challenge was to adapt one for small amount of sample (in the order of 10 uL). This is often needed, partly because the available amount of sample is limited; sample dilution results in loss of sensitivity and increased bias; and working with large sample amounts (if available) increases the cost of consumables significantly. We have developed such a protocol, which is now widely used, and was published: *Turiak et al, J. Proteomics, 2011*.

3) *nano.UHPLC-MS/MS experiment design.*

This is the core technology used in the OTKA project. The methodology has been optimized, the most important variables/features being: a) nano.UHPLC peak capacity (2 hour gradient was considered best); b) loadable sample amount (glycopeptide analysis is sensitivity limited, and glycopeptides are typically small peaks, so the column had to be overloaded for identifying/quantifying glycopeptides); c) CID-MS/MS collision energy had to be optimized to a value far from typical tuning conditions; this aspect will be described in detail below); d) glycopeptide identification and quantitation required different conditions, the experimental protocol therefore was optimized as two separate experiments. Various aspects of experiment design were published in *Toth et al, J. Pharm Biomed Anal, 2014; Ozohanics et al, J. Chrom A, 2012.*

4) *Evaluation of glycopeptide (tandem) mass spectra.*

This is a critical step of glycopeptide MS/MS analysis. Experiments are performed automatically (data dependent analysis, inclusion lists, etc.); but evaluation is not trivial. The stability of the molecular ion (protonated or cationized molecule) has a major role in fragmentation, and this has to be considered. At the start of the project it was not possible to automatize the evaluation of glycopeptide MS/MS spectra (in contrast to peptides and proteomics, where data evaluation is well automatized). First we have manually evaluated glycopeptide spectra; than we established fragmentation rules and optimized CID collision energy; we have developed a data evaluation algorithm; and finally we have developed/modified our software (Glycominer) to perform the calculations. Now, partly based on our studies, commercial software have also become available for this purpose. Fragmentation characteristics of glycopeptides have been published: *Kuki et al, Rapid Comm. Mass Spectrom., 2012; Vekey et al, Int. J. Mass Spectrom, 2013; 2014; Rondeau et al, Rapid Comm. Mass Spectrom, 2014.* Our software (Glycominer) is available at our webpage: http://www.szki.ttk.mta.hu/ms/glycominer/index.php

5) *Glycoform quantitation.*

Glycoform quantitation is an essential part determining glycosylation patterns; and is crucial for characterizing glycosylation of biological samples. Quantitation is based on the peak abundances of various glycopeptide glycoforms. Manual data evaluation, in principle, is straightforward; but to perform it on a large number of glycopeptides/glycoforms in a large number of samples is unfeasible. Automatic procedures were (and to a large degree still are) unavailable. For this reason we have developed our own software to identify glycopeptide-related signals; evaluate and quantify them. Quantitation relies heavily on the reproducibility and robustness of analysis; which are the weak aspects of nano.HPLC technology. We have studied reproducibility in long series of analysis (taking as long as a week). We have identified various sources of error, and devised a simple algorithm to compensate for these errors, which improved reproducibility and removed a significant source of potential bias. Results have been published primarily in *Ozohanics et al, J. Chrom A, 2012; Thoth et al, J. Mass Spectrom, 2015.* The program (Glycopattern) is available at our webpage (http://www.szki.ttk.mta.hu/ms/glycopattern/index.html).

6) *Protocol established for studying glycosylation patterns.*

Based on the various, individually studied and optimized steps of glycopeptide analysis we have developed an overall protocol, both for our “old” Waters and a slightly modified version for our “new” Bruker QTOF instruments. These protocols have been used to determine glycosylation patterns for various proteins in plasma samples. As projected in the planning stage of the project; many samples studied in the initial phases have been re-measured by the established protocol, so that all comparisons should be compared using identical experimental protocols. The concise version of the “overall” protocol is published in *Toth et al, J. Pharm Biomed Anal, 2014*. The detailed protocols are available in our intranet webpage, but these are not in the public domain.

4. Glycosylation studies, glycosylation patterns

Based on the methods and protocols developed in the course of the OTKA project (described above) we have determined the detailed glycosylation patterns of various plasma proteins. The glycosylation pattern of some glycoproteins are published in *Ozohanics et al, J. Chrom A, 2012* and *Toth et al, J. Pharm Biomed Anal, 2014*. Typical glycosylation patterns of alpha-1-acid glycoprotein and serotransferrin; the two proteins studied in most detail are shown in Fig. 1:

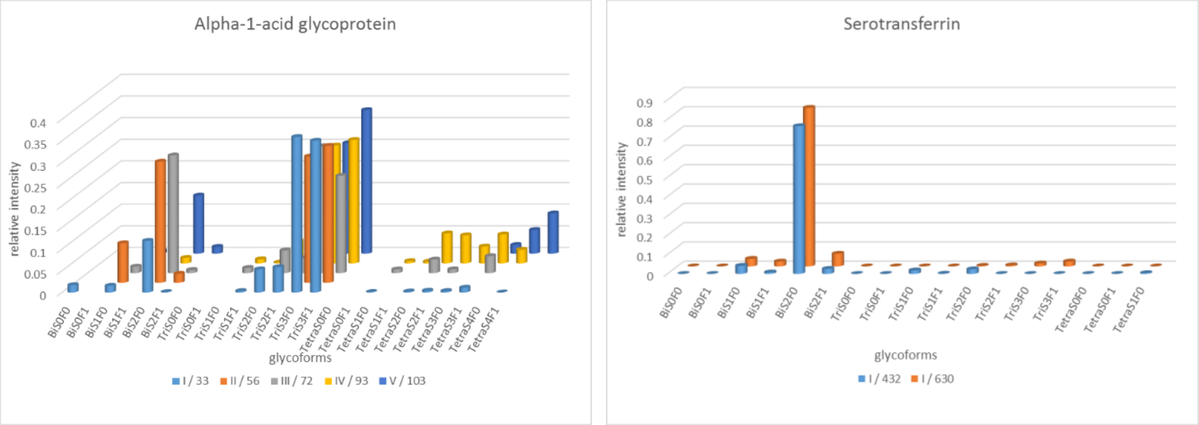

Fig 1. Glycosylation patterns of 5 glycosylation sites in alpha-1-acid glycoprotein and 2 glycosylation sites in serotransferrin

We have studied the glycosylation patterns of a large number of individuals (healthy, cancerous and suffering from other diseases), and found various correlations. The most important findings were the following:

1) Glycosylation patterns at different sites of the same protein are different (as shown above for two proteins).
2) We have determined that the abundance of individual glycoforms (like BiS2 at site 33 for AGP); and derived parameters, like the degree of fucosylation, sialysation and the number of antennae at a given site are all useful parameters for characterizing glycosylation.
3) The parameters characterizing glycosylation are different for various individuals. For example there are individuals showing large or small degree of fucosylation. For example, Fig. 2a shows a good linear correlation between the degree of fucosylation at two different glycosylation sites in AGP, for various individuals. Some other glycosylation features vary, but do not seem to correlate with each other. Such an example is shown in Fig. 2b, indicating the degree of sialylation between between serotransferrin and haptoglobin, for several individuals.

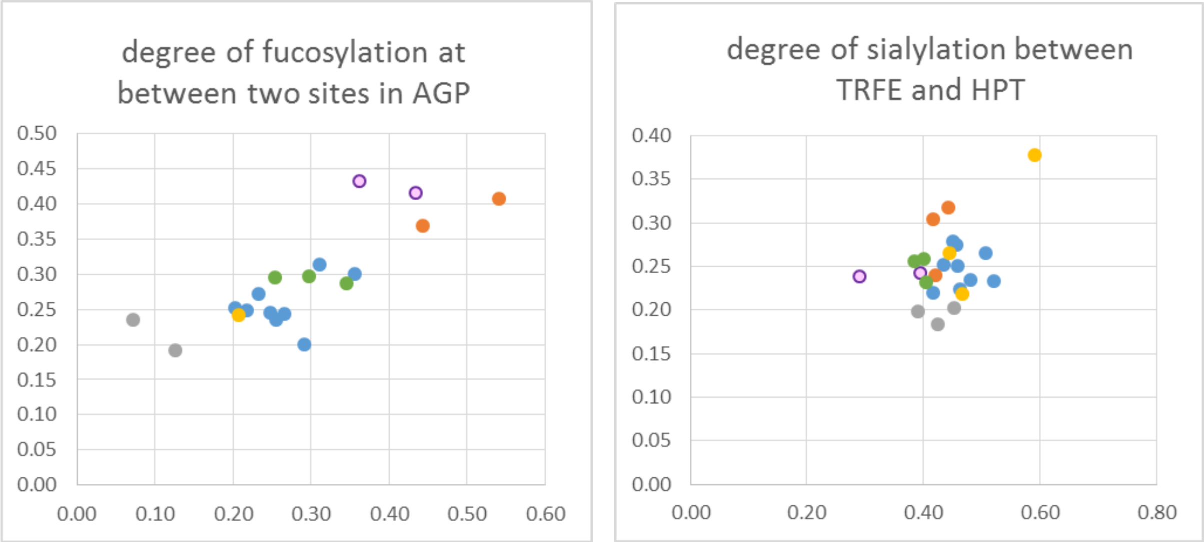

Fig. 2. Correlations between glycosylation features for different individuals.

4) It was found that glycosylation of some proteins and at some glycosylation sites vary significantly; while other proteins and other sites glycosylation is conservative.

The results show that the biochemical machinery performing glycosylation is very complex. Different proteins and various sites not only show different glycosylation patterns, but some features correlate well (e.g. high or low fucosylation of all proteins and at all sites); while some others are markedly different (e.g. a given individual may show high sialylation of one protein, but low degree of sialylation for another protein). The results shown above represent the state-of-art at present. Science at present is descriptive, understanding the biological significance is in the future. The results described above are in the course of publication, at present we are reproducing the results on the new Bruker Maxis instrument to allow us publications in the most prestigious journals.

5. Proteomics, glycosylation and pathological processes

The methodologies developed in the course of the OTKA project allowed us to determine site specific glycosylation patterns for various proteins. These in turn allowed us to study the connection between proteomics, glycosylation and pathological diseases; as foreseen in the OTKA project. Beside the main objective of identifying cancer biomarkers, several other "mini" projects have been performed. Performing these studies were based on the major methodology developments described above. Publications brought light to various aspects of proteomics, glycosylation with respect to disease pathology and the functional roles of proteins and protein PTMs in normal and pathological biochemical processes; some of which are implicated in cancer development as well: *Turiak et al, J. Proteomics, 2011; Di Stefano et al, J. Chrom. A, 2012; Leveles* *et al, Acta Cryst. D. Biol Cryst., 2013; Wysocka-Kapcinska et al, PLOS ONE, 2013; Nagy et al, Angew Chem Int Ed, 2014; Toth et al, J. Pharm Biomed Anal, 2016.* All of these journals are in Q1 quartile in SCIMago in their respective fields.

Of particular relevance is the study of protein glycosylation and radiotherapy of cancer patients (*Toth et al, J. Pharm Biomed Anal, 2016*). The results clearly suggest that 1) changes in the glycosylation of plasma proteins occur in the timeframe of weeks; 2) effect of external stress on cells (like radiotherapy) influence protein glycosylation for months; and 3) changes in glycosylation depends in a large degree on the individual – i.e. glycosylation changes are highly individual.

The long-term objective of the present OTKA project was to establish if and how glycosylation features may characterize pathological processes related to cancer, thereby opening the way to develop a novel, glycosylation related cancer marker. This aim has been approached, but not completely fulfilled, due to problems and unexpected difficulties described above. Research in this direction is still going on (financed from the Academy budget), and we still think this will lead to a breakthrough connecting cancer and glycosylation. At present we are repeating our "preliminary" experiments on the new, high sensitivity Bruker Maxis instrument; and plan to publish the results in 2016/17 in very prestigious journals. Here we show some groundbreaking, but yet unpublished results.

We have studied various groups of patients with lung disease (asthma, pneumonia, COPD and lung cancer), and compared these with healthy individuals. The healthy individuals were further divided into cohorts of men and women, and also cohorts based on age. All patient cohorts were balanced for age and sex.

We have studied individual glycosylation features (at various sites of various proteins), for various individuals and patient groups. Two such distributions/correlations are shown in Fig. 3. These indicate correlations between various glycosylation features. Fig 3 also shows that various glycosylation features are distinctly differ for various patient groups. Data for lung cancer patients are characteristically different from the results of all other groups; but some other patient cohorts can also be distinguished based on these glycoform abundances. For example the red dots (COPD patients) are closest to the cancer patients; while women (yellow dots) are farthest.

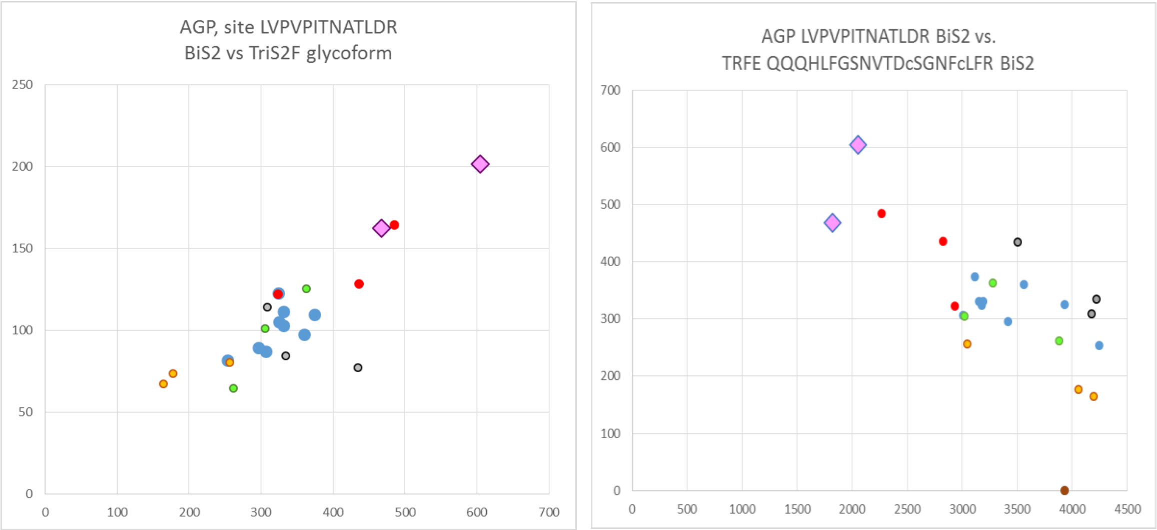

Fig. 3 correlations between individual glycoforms in a number of selected individuals. Left hand site two different glycoforms at a given AGP site; on the right hand the BiS2 glycoform on two different proteins (AGP and TRFE). Large pink diamonds indicate results for lung cancer patients.

Similar comparisons can be made not only based on individual glycoform abundances, but also on the type of glycosylation processes (e.g. high degree or high rate of fucosylation, or sialylation. This comparison represents the state of the overall glycosylation state of an individual; which may also be called as the state of the glycosylation (Golgi) machinery. Such results are shown in Fig. 4. This shows that cancer patients (large pink diamonds) are characterized by an average amount of triantennary glycanes and a high degree of fucosylation. Here, just as before, lung cancer (red dots) and COPD patients behave in a similar manner.

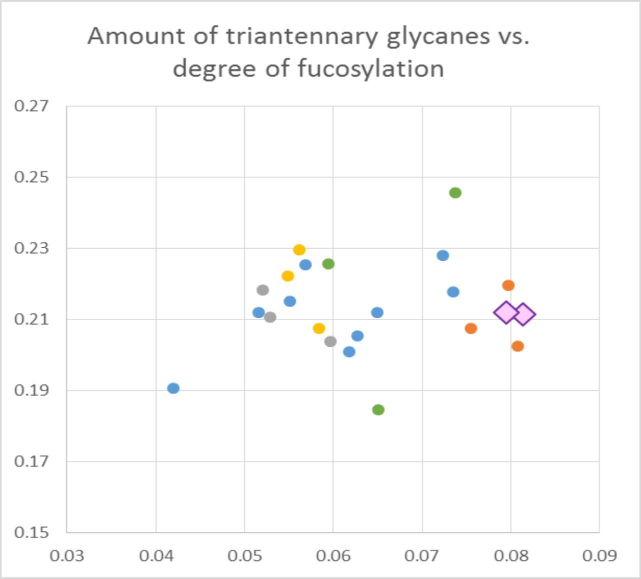

Fig. 4 Overall degree of triantennary glycanes vs. the degree of fucosylation

The last result is the evaluation of all glycosylation features studied, evaluated by a statistical method (principle component analysis). Note, this is an unsupervised method, i.e. the method does not use information on the patients, only data on glycosylation are considered. Fig. 5 shows the distribution of individuals based on the studied glycosylation features.

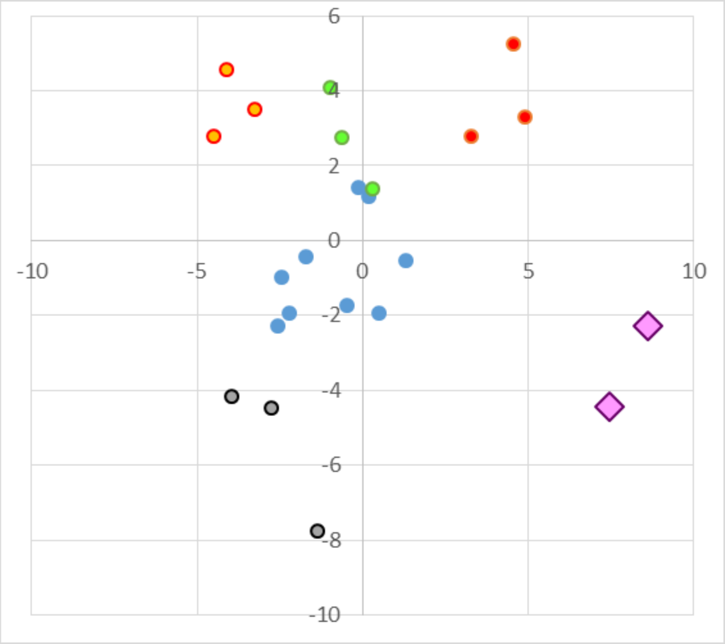

Fig. 5: PCA separation (PC1 and (PC2+PC3) dimensions) of individuals based on the glycosylation features. Pink diamonds show cancer patients.

Note that most patient groups can be identified (although there are some overlaps), indicating that glycosylation highly depends on the type of illness or pathology. Note that 1) the cancer patients are the best identified group, farthest away from any other group of patient. Note that old women (orange) and young man (gray) are farthest from the cancer group in the PC1 direction (x axis). The y axis (sum of PC2 and PC3 values) shows good separation of lung cancer and COPD patients, which were quite similar in the previous figures. Pneumonia and asthma patients can also be well distinguished.

6. Summary and outlook

In the course of the OTKA project we have developed the key technologies and methods for determining site specific glycosylation patterns from the blood plasma of various individuals. These are state-of-art methodologies, published in high quality journals. We have used these methods in various applications; in order to get information on the glycosylation characteristics (and also on the proteomics) of various individuals. These results have also been published in high quality journals. We have excellent and very promising preliminary results on glycosylation markers for lung cancer; but these have not yet been published. We are currently preforming/repeating experiments on our new Bruker Maxis instrument to improve data quality in order to publish the results in very prestigious journals in the near future. This delay is, in part, due to unforeseen difficulties encountered (detailed above). The other reason for the delay is that studies on glycosylation advanced at a slow pace worldwide; due to the fact, that understanding glycosylation is a much harder nut to crack, than expected.