Glycoproteomic Data Quality and Similarity in Influenza A Virus Variants

Olivia Marx[1,2], William Hackett[3], Joseph Zaia[3]

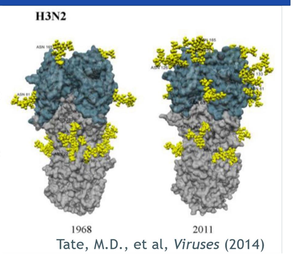
Southern Methodist University[1], Boston University Bioinformatics BRITE REU Program, Summer 2018[2], Boston University[3]

Abstract

This project aimed to determine what data quality is required to achieve a statistically significant result when comparing the glycosylation of two strains of the flu virus. Glycosylation is a post-translational modification where a sugar is added onto a protein. A glycosylated protein is called a glycoprotein and can have different sugars, or glycans, added to different places, which can change how a protein interacts with its environment. A mass spectrometer gathers data about the sample through analyzing particle masses. A program called GlycReSoft then interprets the mass spectrum data and identifies the glycoproteins present in the sample. We used the Jaccard and Tanimoto similarity metrics to assess similarity of glycosylation between virus strains, however, glycoproteomic data has many limitations. Preliminary analysis pointed to the conclusion that there may not be enough usable data to properly compare two datasets. To address this, we began development of a formula to determine a "quality score" for a dataset using statistics from GlycReSoft, including MS2 scores of how much the program trusts the glycopeptide is present, number of runs, and missing values in the data. We then assessed whether the statistic should make the score increase or decrease and integrated it into the scoring formula accordingly. Preliminary results indicate a correlation between the score and the similarity of the data to itself, and the more consistent a dataset is, the more likely it is reliable, suggesting that the score may be a valid assessment of data quality. Once the required data quality has been obtained, further analysis may be conducted, and similarity metrics may be used to compare glycosylation patterns. This work will deepen our understanding of how the flu virus mutates and the role of glycosylation in its evolution.

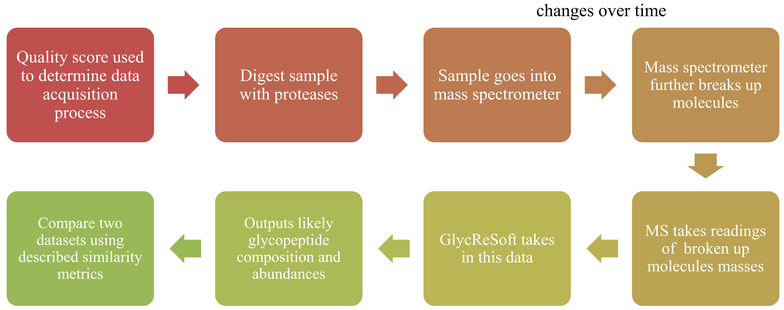
Background and Motivation

- The glycosylation on IAV changes as the virus mutates
- Glycans can shield antigenic sites on IAV so that the body cannot recognize them any more
- Previously, there has been no good way to assess the similarity between two sets of flu strains' glycoproteomic data
- We propose a function to determine the data quality needed to compare two datasets



Tate, M.D., et al, *Viruses* (2014)

Glycosylation (yellow) on the protein coat of H3N2 virus strain changes over time



Similarity Metrics

Jaccard Similarity Metric

$$\frac{M_{11}}{M_{01}+M_{10}+M_{11}}$$

- $M_{11} = number\ of\ common\ glycopeptides$
- $M_{01} = glycopeptides\ unique\ to\ set\ 1$
- $M_{10} = glycopeptides\ unique\ to\ set\ 2$

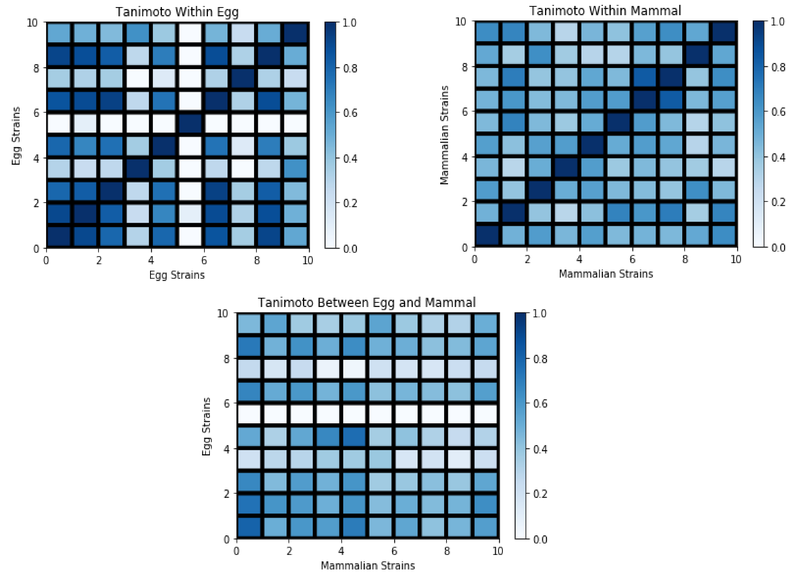
Tanimoto Similarity Metric

$$\frac{T_{11}}{T_{01}+T_{10}+T_{11}}$$

- $T_{11} = common\ glycopeptides\ abundance$
- $T_{01} = glycopeptides\ unique\ to\ set\ 1\ abundance$
- $T_{10} = glycopeptides\ unique\ to\ set\ 2\ abundance$

Similarity Metric Results

Tainimoto Similarities Between Mammalian and Avian IAV Cell Lines

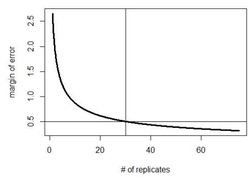


Quality Score Methods

Preliminary Analysis- Power Estimate

Power analysis to see how many replicates are needed to get a good result

Answer: about 30



Score Development

To develop the score, we analyzed the data output from GlycReSoft and decided whether an increase in the data should result in an increase or decrease in score

- Log terms were added based on whether a variable might be affecting the score too greatly

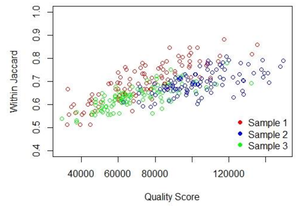
Number of replicates (N)	• Score increases
Missing values (mv%)	• Score decreases
False Discovery Rate (FDR)	• Score decreases
MS2 Score	• Score increases
Abundance Standard Deviation σ	• Score decreases
Search space over total possible glycoforms (Q/M)	• Score decreases

$$Score = \frac{N^2 * (\%presence)(1-FDR)}{-\log(\frac{Q}{M})} * mean(\frac{MS2}{\sigma})$$

Quality Score Results

Data Exploration

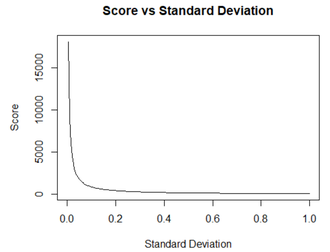
- To test the validity of the quality score, we checked the quality score against the "Within Jaccard"
- The positive trend could indicate that the score is a valid measurement of quality.



The quality score is still in its preliminary stages as it is still being tested and perfected as a means to accurately assess the data.

Score Effectors

- The score can be largely affected by small changes in standard deviation of normalized abundances
- Batch effects can influence standard deviation and cause an unexpected score distribution
- Sample with too much background noise can effect MS2 score
- Search space is incomplete



Conclusions and Future Work

We have :

- Used Jaccard and Tanimoto methods to assess data similarity
- Implemented quality scores in a known example and found batch effects
- Assessed Relationships between quality scores and similarity differences

In the Future:

- Implement score in GlycReSoft output pipeline
- Improve data quality using improved conditions from the quality score
- Use the score to help spot hidden variabilities in glycoproteomic assessment methods
- Compare many different strains of IAV and examine similarity to increase understanding of IAV glycosylation and improve flu vaccines

References and Acknowledgements

Joshua Klein, Luis Carvalho, Joseph Zaia; Application of network smoothing to glycan LC-MS profiling, *Bioinformatics*, Volume 34, Issue 20, 15 October 2018, Pages 3511–3518, https://doi.org/10.1093/bioinformatics/bty397

Sullivan, Lisa. "Sample Sizes for Two Independent Samples." *Power and Sample Size Determination*, 2016, sphweb.bumc.bu.edu/otlt/MPH-Modules/BS/BS704_Power/BS704 _Power5.html.

This work was funded, in part, by NSF grant DBI-1559829, awarded to the Boston University Bioinformatics BRITE REU program and NIH grant #U01CA221234.