Metrics for Quantifying Differences in Influenza A Virus Hemagglutinin Site-specific Glycosylation

Jacquelyn Turcinovic[1,2], Joshua Klein[3,4], Joseph Zaia[4]

[1]Boston University Bioinformatics BRITE REU Program, Summer 2017; [2]St. Edward's University, Department of Bioinformatics;
[3]Boston University Program in Bioinformatics; [4]Boston University School of Medicine, Department of Biochemistry

Abstract

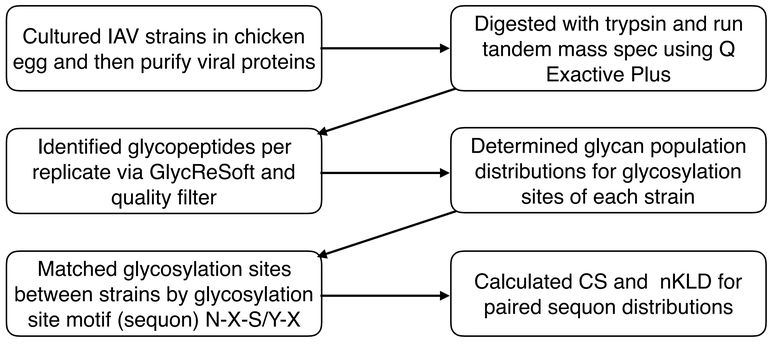
In the 2015-2016 flu season, the CDC estimates five million influenza A virus (IAV) infections in the United States were prevented by vaccination. The IAV protein involved in yearly vaccine production, hemagglutinin (HA), is heavily glycosylated, meaning it has carbohydrate chains of varying composition and topology called glycans attached at several sites. With the recent development of mass spectrometric methods for assignment of site-specific glycosylation of IAV glycoproteins, there is need for metrics for comparison of strain-specific differences. By treating site-specific glycan populations as discrete probability distributions, we hypothesized that glycosylation differences between IAV strains can be quantified using distance metrics. To test this, we compared the performance of cosine similarity (CS) and normalized Kullback-Leibler divergence (nKLD) in quantifying the differences in glycosylation between the Philippines-1982 and Philippines-1982-BS IAV strains. Overall, we found nKLD to be the more sensitive metric.

Background and Motivation

Despite 60 years of influenza vaccine development, a long-lasting, broadly neutralizing vaccine does not exist; instead, vaccines are developed yearly. Researchers ensure a vaccine does not interact with previous years' vaccines using the concept of antigenic distance. Hemagglutinin (HA), the influenza A virus (IAV) protein involved in testing these interactions, is heavily glycosylated (Figure 1). Researchers postulate that HA glycosylation is an evolutionary mechanism affecting antibodies' ability to bind to IAV, called antigenicity. Therefore, understanding how HA glycosylation influences antigenicity differences between IAV strains may better inform vaccine development and eventually help improve vaccine effectiveness and longevity.

Methods

Strain Name	UniProt Accession ID
Philippines-1982	AFG99160
Philippines-1982-BS	AAC79579



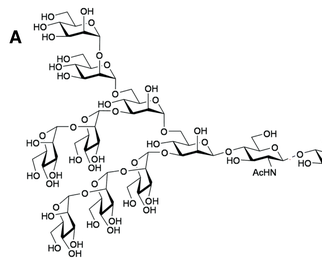
CS is the cosine of the angle between two vectors and compares probability distributions by modeling the distributions as n-dimensional vectors. CS values range from negative one to positive one. Zero means the vectors are orthogonal and have no entries in common. One means the vectors are identical.

$$\cos\theta = \frac{\vec{a}\cdot\vec{b}}{||\vec{a}|| \times ||\vec{b}||}$$

nKLD quantifies divergence of an experimental probability distribution from an expected probability distribution in a range from zero to one. Zero means the probability distributions behave in the same manner. One means we cannot relate the experimental distribution's behavior to the expected distribution's behavior.

$$KLD = \sum_i P(i) \times \log\frac{P(i)}{Q(i)}$$

$$nKLD = 1 - e^{KLD}$$



C **0,9,2,0**
0 Fucose
9 Hexose
2 N-Acetylated Hexose
0 Sialic acid

Figure 1. Representations of a glycan in a chair diagram (A), symbolic diagram (B), and monosaccharide abbreviation (C).

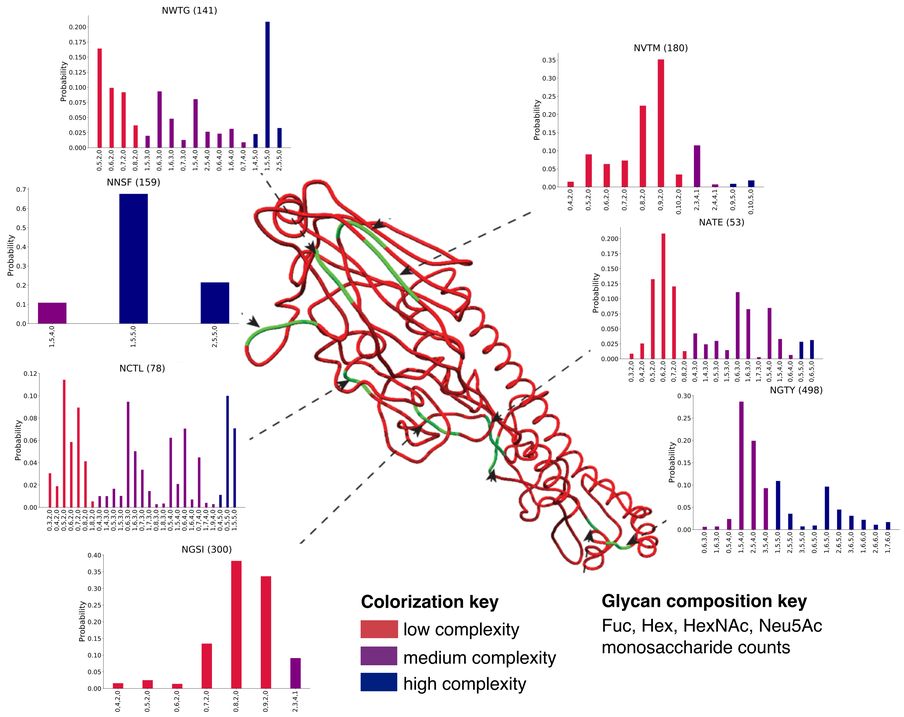


Figure 2. A structural image of the AFG99160 HA protein is shown with ten glycosylation sites marked in green. Bar graphs show probability distributions of site glycan populations. Bars are colored by complexity, which classifies glycans by monosaccharide compositions. Three sites had no data. As can be seen, the number and proportion of different glycans can be highly variable.

Results

The CS and nKLD values calculated for matched sequons are shown in Table 1.

- Cosine similarity values occupy range 0.949-0.997, which we read as high similarity.
- nKLD values occupy range 0.048-0.754, which we read as highly similar to highly divergent.

Sequon	AFG99160 (1982)	AAC79579 (1982-BS)	Cosine	nKLD
NATE	53	37	0.977	0.084
NCTL	78	62	0.949	0.155
NNSF	159	143	0.942	0.754
NGSI	300	284	0.997	0.048
NGTY	498	482	0.951	0.161

Table 1. Cosine similarity and nKLD values for all AFG99160 and AAC79579 glycosylation site pairs with coverage.

Discussion

Why do we observe a discrepancy between CS and nKLD?

- CS is high if vectors are dominated by high-scoring entries, even if there are differences in paired values.
- nKLD is sensitive to differences in vector entry pairs, no matter the entry size.

When probabilities of all glycans are close, nKLD shows high similarity (Figure 2A), but as probabilities diverge, nKLD grows faster than CS (Figure 2B). Since nKLD reflects differences in glycan population distributions without being affected by the probability of individual glycans, it is less forgiving to differences in individual glycan probability than CS. Therefore, we believe nKLD to more accurately represent site-specific glycosylation differences.

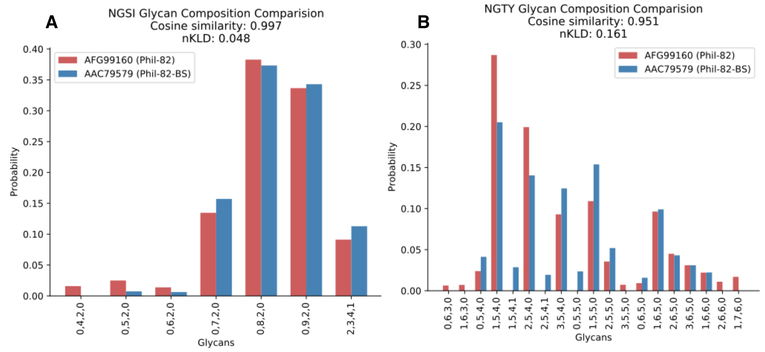


Figure 3. Cosine similarity values do not always agree with nKLD values because nKLD is more sensitive to differences in probability distribution. A. NGSI glycan distribution comparison. The cosine similarity value is 0.997, and the nKLD value is 0.328. B. NGTY glycan distribution comparison. The cosine similarity value is 0.951, and the nKLD value is 0.717.

Conclusions

Defining HA shape is crucial to IAV antigenicity research, as it determines which antibodies can bind to IAV in an immune response. Understanding the glycan population at each HA glycosylation site will improve the definition of IAV strain shape and thus antigenicity. An accurate, standardized method for comparing glycosylation site populations between strains is essential for this understanding. Because nKLD is more sensitive than CS at detecting population differences, we believe nKLD to be the more useful metric for future glycan-focused antigenicity research.

References and Acknowledgements

- Bailey E, Long L-P, Zhao N, et al. (2016). Antigenic Characterization of H3 Subtypes of Avian Influenza A Viruses from North America. Avian Diseases, 60, 346-353.
- Cai Z, Zhang T, Wan X-F (2012). Antigenic Distance Measurements for Seasonal Influenza Vaccine Selection. Vaccine, 30, 448-453.
- Kshitij K, Klein, JA, White MR, et al. (2016). Integrated Omics and Computational Glycobiology Reveal Structural Basis for Influenza A Virus Glycan Microheterogeneity and Host Interactions. Molecular and Cellular Proteomics, 15, 1895-1912.
- Rolfes MA, Foppa IM, Garg S, et al. (2016). Estimated Influenza Illnesses, Medical Visits, Hospitalizations, and Deaths Averted by Vaccination in the United States. CDC, NCIRD.
- Varki A, Cummings RD, Esko JD, et al. (2009). Essentials of Glycobiology. 2nd edition. NCBI Bookshelf.

This work was funded, in part, by NSF grant DBI-1559829, awarded to the Boston University Bioinformatics BRITE REU program. Additional funding was provided by NIH grant P41GM104603.