Similarity Assessment by Multivariate Statistics Method Based on Distance Between Biosimilar and Originator

Jian Xu, Zhihui Shao, Xiaoxiong Han, Yingfeng Huang, Xun Zou, Yaling Shen*

*State Key Laboratory of Bioreactor Engineering, Shanghai Collaborative Innovation Center for Biomanufacturing Technology, East China University of Science and Technology, Shanghai, People’s Republic of China

CAS Key Laboratory of Synthetic Biology, CAS Center for Excellence in Molecular Plant Sciences, Institute of Plant Physiology and Ecology, Chinese Academy of Sciences, Shanghai, China

Process Development Department, Dragon Sail Pharmaceutical, Shanghai, China

R&D Laboratories, Dragonboat Biopharmaceutical, Shanghai, China

Shanghai Sanjin Bioscience and Technology, Shanghai, China

* Corresponding authors

Yaling Shen (ylshen@ecust.edu.cn),

Tel: +86-2164253156, Fax: +86-2164250068,

Address: 130 Mei Long Road, Shanghai 200237, China.

Abstract

The development of biosimilar products or follow-on biologics has been flourishing in recent years because of lower price than originators. In this study, a multivariate data analysis method based on JMP software was proposed to assess the glycosylation pattern similarity of antibody candidates from different conditions in optimization experiment with reference. A
specific distance was generated by this method and indicated the glycoform similarity between biosimilar and reference. This method can be applied to analyze the similarity of other physicochemical and function characters between follow-on biologics and originators. Then the design of experiments method can be realized to optimize culture conditions of cell culture to attain similar antibody candidates.

**Key words:** biosimilar, similarity, glycosylation pattern, multivariate statistics method, cluster analysis.

### 1. Introduction

During past two decades, biologic products (also termed “biopharmaceuticals”) have been developed intensively by companies for treating cancer and autoimmune diseases such as rheumatoid arthritis. 6 of the top pharmaceutical products sales in 2019 were biologic products (Table 1). So, there has been an increasing trend toward development of biosimilars considering market interest and improving access of biologics. Biosimilar products can offer lower price so as to optimize efficiencies across healthcare systems. However, biopharmaceuticals typically involve expression of the gene using living cell followed by purification and formulation to acquire stable drug product. They are large, complex and heterogeneous compared to small molecular produces. It is impossible to manufacture identical copies of biologic products (WHO 2016). To define biosimilar, different regulatory body has different criteria. EMA define a biotherapeutic product that is similar in terms of quality, safety, and efficacy to an already licensed reference product (reference medicinal product) in the EMA (EMA 2015). While, FDA define a biological product that is highly similar to a US-licensed reference product notwithstanding minor differences in clinically
inactive components, and for which there are no clinically meaningful differences between the biological product and the reference product in terms of safety, purity, and potency of the product (FDA 2015b). WHO define a biotherapeutic product that is similar in terms of quality, safety, and efficacy to an already licensed reference product (WHO 2016). In broad terms, a biosimilar is highly similar to a reference product in terms of structure and function. And the high similarities of physicochemical and function characters are main aim in pre-clinical development phase.

The structural and functional elements of therapeutic antibody include primary structure, purity, charge heterogeneity, glycosylation and other post-translational modifications, as well as target and receptor bindings activity and bioactivity features (Kirchhoff et al. 2017). A variety of analytic techniques are developed to demonstrate these elements. However, many analytic results are multivariate data such as charge heterogeneity, glycosylation and size heterogeneity. It is hard to use simple method to assess the similarity between biosimilar and originator (or reference) based on these multivariate outcomes.

Identification of the glycosylation pattern is a key consideration during the development of mono-clonal antibody (mAb) biosimilars, since the glycan chains in the Fc region can substantially alter protein activity and the PK profile, and in some case, antigenicity (Kirchhoff et al. 2017). Because many cell culture conditions such as nutrient availability, pH, dissolved oxygen (DO), ammonia, cell viability, growth phase, temperature will affect glycosylation, the culture process parameters should be well monitored and controlled in manufacture phase (Patrick et al. 2019). Chemical supplements such as metal ion and substrate for glycan chain synthesis have been described in the literature as effective
glycosylation modulators during up-stream process development (Crowell et al. 2006, Gramer et al. 2011). To find the optimal addition amount, design of experiment (DOE) is used intensively during screening experiment. But a specific parameter indicating the similarity should be defined as response before modelling and ANOVA analysis. An effective method that can output a specific parameter to assess the similarity of glycan profile is needed.

In this work, a multivariate data analysis method was first applied to generate a specific index that can represent the similarity of glycan profiles from different antibody candidates. By this method, the similar antibodies were clustered by the glycan distribution and identified the most similar antibody easily. And the effect of supplements could be qualified statistically.

2. Materials and methods

2.1 Cell line and reagents

The cell line expressing recombinant IgG1 was derived from CHO DG44. The basal media is CD DF1 (Shanghai BasalMedia Technologies Co., LTD., Shanghai, China) supplemented with 6 mM Glutamine (Sigma, Shanghai, China). Feed media is Efficient Feed™ C+ (Thermofisher, Shanghai, China) and add additives as the design. The additives, N-Acetyl-D-glucosamine (GlcNAc) and MnCl₂ were purchased from Sigma–Aldrich Shanghai Trading Co., Ltd (Shanghai, China).

2.2 Cell culture conditions and process design

The cells were cultured in shaker flasks (Corning, New York, USA) with vented cap at 5% CO₂, 70% humidity and 37°C (shift to 35°C from day 4 to harvest) condition. The volume of beginning culture was 30 mL with 120 rpm of shaking rate. Inoculum density was adjusted to
1 million cells/mL equally. GlcNAc and MnCl$_2$ were supplemented into culture by two ways: adding in feed media, or adding separately into basal media on day 4. The supplemented amount was designed as the table 2. The first 14 runs and second 14 runs were duplicated, and the last 2 runs were control runs without supplemental. The feed was started from day 3 and the feed volume was calculated by the following equation.

$$V_{\text{feed}} = V_{\text{current}} \times Q_{\text{feed}} \times \frac{(VCD_1 + VCD_2)}{2 \times Glc_{\text{feed}} \times 1000}$$

In which, $V_{\text{feed}}$ (mL) is the feed volume of current day; $V_{\text{current}}$ (mL) is the culture volume before feeding; $Q_{\text{feed}}$ (75) is the feed rate factor; $VCD_1$ ($10^6$ cell/mL) is the viable cell density of current day; $VCD_2$ ($10^6$ cell/mL) is calculated viable cell density of next day; and $\text{Glc}_{\text{feed}}$ (g/L) is the glucose concentration in feed media.

Sampling was conducted every day from day 3. The residual glucose was analyzed by Glucose Test Kit (Beihai, China). Additional glucose was calculated by the following equation and added to maintain the concentration of glucose at 3 g/L when viable cell density less than $1.0 \times 10^7$ cell/mL, and shift to 4 g/L when viable cell density higher than $1.0 \times 10^7$ cell/mL.

$$V_{\text{Glc}} = \left[ (\text{Glc}_{\text{target}} - \text{Glc}_{\text{test}}) \times V + (\text{Glc}_{\text{target}} - \text{Glc}_{\text{test}}) \times V_{\text{feed}} \right] / \text{Glc}_{\text{solution}}.$$

In which, $V_{\text{Glc}}$ (mL) is the addition volume of glucose solution; $\text{Glc}_{\text{feed}}$ (g/L) is the glucose concentration in feed media; $\text{Glc}_{\text{test}}$ (g/L) is the glucose concentration in culture; $\text{Glc}_{\text{target}}$ (g/L) is the target glucose concentration after feeding; and $\text{Glc}_{\text{solution}}$ (g/L) is the concentration of glucose solution.

After 11 days culture, the culture was harvested by centrifugation to remove cell debris and
purified by Protein A resin.

2.3 Glycan Analysis

2.3.1 Enzymolysis of glycan chains from antibody

Firstly, replaced the buffer of the antibody sample with 50 mmol/L NH₄HCO₃ (pH 8.0). Then 500 μg antibody was hydrolyzed with 2000 U PNGase F (New England Biolabs, Beijing, China) at 37°C for 24 h. Pre-cooled ethanol was added to the final concentration of 75% (v/v). The mixture was blended and stand for 0.5 h at -20°C. After centrifugation at 13000 rpm for 15 min, the supernatant was vacuum dried.

2.3.2 Fluorescence labeling of glycan chains

Add 5 mg 8-aminopyrene-1,3,6-trisulfonic acid trisodium salt (APTS, AB Sciex Pte. Ltd., Framingham, USA) into 0.5 mL aqueous solution containing 15% acetic acid and vortex. Then add 15 μL APTS solution and 5 μL tetrahydrofuran solution containing 1 mol/L sodium cyanoborohydride to the vacuum-dried glycan chains. After fluorescent labeling at 55°C for 2h, 400 μL ultrapure water was added and analyzed by PA800 Plus capillary electrophoresis apparatus (AB Sciex Pte. Ltd., Framingham, USA).

2.3.3 Analysis by capillary electrophoresis apparatus

Beckman N-Cho coated capillary with total length of 60.5 cm and effective length of 50 cm and inner diameter of 50 μm and electrophoresis buffer was purchased from Beckman Coulter Life Sciences (Indianapolis, USA). The capillary temperature was 20°C. Sample was injected at 2.0 psi for 10 s. Then the sample was separated at 30 kV for 20 min. Fluorescence detection was implemented at excitation and emission wavelengths of 488 and 520 nm, respectively.
2.4 Multivariate Statistics

The peak area percentages from glycan analysis are used for assigning the similarity in this case. Seven peaks are identified and coded numbers by isoform. In this way all the data is normalized for each sample. A symmetrical matrix of the data is built and a Hierarchical Cluster Analysis is performed with JMP software (SAS Institute Inc.). A dendrogram can be generated and save the distance matrix to another data table.

3. Results

3.1 Glycan Analysis

The glycan peak distribution of 30 samples from different shake flask culture are shown in Fig. 1. All the samples’ glycan shows similar pattern that peak 3 is the highest portion, however the percentages of each peak are different. At a glance, the supplements of GlcNAc and MnCl₂ should affect the glycan distribution of antibodies. But it is difficult to test the effect by statistic method, because a specific number to assign the similarity is lack to indicate the variance of these peaks.

3.2 Cluster analysis

Perform the cluster analysis tool in JMP as described in method, a hierarchical clustering tree is shown in Fig. 2. The single samples are the leaves and the similar samples are clustered on one branch. We have set the duplicate runs same color, so it is shown that almost all duplicate samples are in third or fourth branch from main trunk. It is very clear to see that SF-7, 8, 13, 14, 21, 22, 27, 28 are located in same third branch with reference, which illustrates these samples’ glycan distribution is very close to that of reference.

3.3 Distance
The distance is to show the similarity between samples and reference. They can be generated by run saving the distance matrix option in menu. The distances are plot to run label is shown in Fig. 3. Here is very clear to see that the antibody from SF-14 and 28 are the nearest to reference antibody in aspect of glycan distribution, while SF-6 and 20 are the farthest. So as to confirm the real peak distribution similarity, only SF-14, SF-20 and reference are shown in Fig. 4. It is easy to find that SF-14 is similar to reference and the difference between SF-20 and reference is significant.

3.4 Effect Significance Analysis

Use the “Fit model” in JMP, set the Distance as the response and put GlcNAc, MnCl₂ and adding them in basal or feed media in model effect window, run the fit program. The ANOVA results are shown in Table 3. In the effect test, GlcNAc and add in basal or feed media are significant factor and MnCl₂ is not. By this way, the significance of factors can be identified in a statistic method.

4. Discussion

In developing a biosimilar drug, a stepwise approach is needed beginning with chemistry manufacturing controls (CMC) and bioanalytical characterization (Burchiel et al. 2019). In CMC development phase, because the structure of antibody is very complex, the quality attributes will be characterized by multiple analysis method such as glycan isoform, capillary isoelectric focusing (cIEF), cation-exchange chromatography (CEX) and peptide map. We know that these methods output multivariance results rather than a single data. Clustering is a ubiquitous data analysis tool to divide complex data into groups of similar items (Andreas et al. 2019). Therefore, it can be utilized to reveal the similarity of these multivariance test
results effectively. Seung-Ho Kang et al (2013) proposed a three-arm parallel design to assess biosimilarity between a biosimilar product and an innovator biological product based on relative distance of means observed from the test and reference products. In the proposed design, if the relative distance is less than a prespecified margin, they claim that the two products are claimed to be biosimilar. This method’s merit is to provide a specific standard to access similarity. Beyond this method, we do not find other strategy to access biosimilar quality by a statistic pathway in publish.

The clustering program in JMP can output hierarchical tree, making the cluster results visualized and easy to find the similar group with reference. This can be utilized to control the batch quality in manufacture phase. If the third branch is set as the similar margin, the leaves of different batches data can be judged as qualified batches. Once a batch’s quality is clustered to other third branch, deviation investigation can be triggered and to recall or destroy the batch based on risk assessment. The distance between samples and reference can be used as response value in optimization experiment. By this way, quality by design (QbD) strategy is feasible to character the process effect and optimize or define operation space of critical process parameter (CPP) by DOE. In future, lots of batches would be clustered, a distance can be set as margin to decide qualified batch too.

5. Conclusion

In this paper, a multivariate statistics method is proposed to access the similarity of antibodies to reference from different conditions in optimization experiment. The multivariance test results can be grouped by this method and a specific distance can be generated. The distance value indicates the similarity between biosimilar and reference. In this study, the glycoform
with shortest distance shows highest similar to the reference glycoform. When a specific
value can indicate the similarity, the DOE method can be realized to evaluate the effect of
factors and optimize culture conditions.

Abbreviations
WHO: World health organization; EMA: European medicines agency; FDA: U.S. Food and
drug administration; mAb: mono antibody; PK: Pharmacokinetics; DO: Dissolved oxygen;
DOE: Design of experiment; ANOVA: Analysis of Variance; GlcNAc: N-Acetyl-D-glucosamine;
CMC: Chemical, manufacture and control; cIEF: Capillary isoelectric focusing; CEX: Cation-exchange chromatography; QbD: Quality by design; CPP:
Critical process parameter;

Declarations
Ethics approval and consent to participate
Not applicable
Consent for publication
Not applicable
Availability of data and materials
Not applicable
Competing interest
The authors confirm that there are no conflicts of interest.
Funding
Authors' contributions
Jian Xu wrote the main manuscript text. Zhihui Shao and Xiaoxiong Han analyzed the data by statistic tool. Yingfeng Huang, Xun Zou and Yaling Shen insturcted the experiment.

Acknowledgements
The authors wish to acknowledge the Quality Analytical Science group in Dragonboat biopharmaceutical Co., for testing the glycosylation.

Authors' information (optional)
Not Applicable

References
Andreas A, Margareta A, Naomi C. B (2019) To cluster, or not to cluster: An analysis of clusterability methods. Pattern Recognition. 88:13–26. https://doi.org/10.1016/j.patcog.2018.10.026
Burchiel S W, Aspbury R, Munday J (2019) The search for biosimilars and biobetters. Drug Discovery Today. 24(5):1087-1091. https://doi.org/10.1016/j.drudis.2019.03.016
Crowell C K, Grampp G E, Rogers G N, Miller J, Scheinman R I (2006) Amino acid and manganese supplementation modulates the glycosylation state of erythropoietin in a CHO culture system. Biotechnology & Bioengineering. 96(3):538-549. https://doi.org/10.1002/bit.21141
EMA. (2015) Guideline on similar biological medicinal products containing biotechnology-derived proteins as active substance: Non-clinical and clinical issues. Available online at:
FDA. (2015b) Guidance for industry: Scientific considerations in demonstrating biosimilarity to a reference product. Available online at: http://www.fda.gov/downloads/Drugs/GuidanceComplianceRegulatoryInformation/Guidances/UCM291128.pdf

Gramer M J, Eckblad J J, Donahue R, Brown J, Shultz C, Vickerman K, Priem P, Bremer E, Gerritsen J, Berkel P (2011) Modulation of antibody galactosylation through feeding of uridine, manganese chloride, and galactose. Biotechnology & Bioengineering. 108(7):1591-1602. https://doi.org/10.1002/bit.23075

Kirchhoff C F, Wang X Z M, Conlon H D, Anderson S, Ryan A M, Bose A (2017) Biosimilars: Key regulatory considerations and similarity assessment tools. Biotechnology and Bioengineering. 114(12):2696–2705. https://doi.org/10.1002/bit.26438

Patrick H, Khattak S F, Li Z J (2009) Optimal and consistent protein glycosylation in mammalian cell culture. Glycobiology. 19(9):936-949. https://doi.org/10.1093/glycob/cwp079

Kang S H, Chow S C (2013) Statistical assessment of biosimilarity based on relative distance between follow-on biologics. Stats in Medicine. 32(3):382-392. https://doi.org/10.1002/sim.5582

WHO. (2016) Guidelines on evaluation of monoclonal antibodies as similar biotherapeutic products (SBPs). Available online at: https://www.who.int/biologicals/biotherapeutics/WHO_TRS_1004_web_Annex_2.pdf.
Table 1 The top 10 pharmaceutical products sold in 2019*.

| Rank | Medicine | Sales (Billion $) | Company       | Type            |
|------|----------|-------------------|---------------|-----------------|
| 1    | Humira   | 19.17             | Abbvie        | Antibody        |
| 2    | Eliquis  | 12.15             | BMS/Pfizer    | Small Moledular |
| 3    | Keytruda | 11.08             | MSD           | Antibody        |
| 4    | Revlimid | 7.171             | Celgene       | Small Moledular |
| 5    | Imbruvica| 8.09              | J&J/Abbvie    | Small Moledular |
| 6    | Opdivo   | 8.00              | BMS           | Antibody        |
| 7    | Eylea    | 7.44              | Bayer/Regeneron | Fusion Protein |
| 8    | Avastin  | 7.12              | Roche         | Antibody        |
| 9    | Xarelto  | 6.93              | Bayer         | Small Moledular |
| 10   | Enbrel   | 6.92              | Amgen/Pfizer  | Fusion Protein  |

Note: * Data from the annual statement of the company.
Table 2 The design of experiment for investigating glycan profile adjustment.

| Run | GlcNAc (mM) | MnCl2 (μM) | Add in Basal or Feed Media |
|-----|-------------|------------|---------------------------|
| SF-1 | 120         | 30         | Feed                      |
| SF-2 | 80          | 30         | Feed                      |
| SF-3 | 40          | 30         | Feed                      |
| SF-4 | 120         | 15         | Feed                      |
| SF-5 | 80          | 15         | Feed                      |
| SF-6 | 40          | 15         | Feed                      |
| SF-7 | 120         | 0          | Feed                      |
| SF-8 | 80          | 0          | Feed                      |
| SF-9 | 40          | 0          | Feed                      |
| SF-10 | 20       | 0          | Basal                     |
| SF-11 | 10         | 0          | Basal                     |
| SF-12 | 5          | 0          | Basal                     |
| SF-13 | 10         | 30         | Basal                     |
| SF-14 | 10         | 15         | Basal                     |
| SF-15 | 120        | 30         | Feed                      |
| SF-16 | 80         | 30         | Feed                      |
| SF-17 | 40         | 30         | Feed                      |
| SF-18 | 120        | 15         | Feed                      |
| SF-19 | 80         | 15         | Feed                      |
| SF-20 | 40         | 15         | Feed                      |
| SF-21 | 120        | 0          | Feed                      |
| SF-22 | 80         | 0          | Feed                      |
| SF-23 | 40         | 0          | Feed                      |
| SF-24 | 20         | 0          | Basal                     |
| SF-25 | 10         | 0          | Basal                     |
| SF-26 | 5          | 0          | Basal                     |
| SF-27 | 10         | 30         | Basal                     |
| SF-28 | 10         | 15         | Basal                     |
| SF-29 | 0          | 0          | N/A                       |
| SF-30 | 0          | 0          | N/A                       |
Table 3 Analysis of four factors effecting on glycan similarity using distance value as the response.

| Source                        | Nparm | DF | Sum of Squares | F Ratio | Prob>F |
|-------------------------------|-------|----|----------------|---------|--------|
| GlcNAc                        | 1     | 1  | 173.1          | 27.2    | <0.0001*|
| MnCl₂                         | 1     | 1  | 17.6           | 2.8     | 0.1083 |
| Add in Basal or Feed Media    | 2     | 2  | 474.2          | 37.3    | <0.0001*|
Figure Captions

Fig. 1 Glycan peak distribution of 30 samples for different shake flask culture.

Fig. 2 Hierarchical clustering tree analyzed by JMP software.

Fig. 3 Euclidean distance of 30 antibody candidates to reference.

Fig. 4 Glycan peak distribution similarity comparison of SF-14, SF-20 and reference.
Fig. 1

Fig. 2
