Alexander Behne[1,2]
Thilo Muth[1,2]
Matthias Borowiak[1,2]
Udo Reichl[1,3]
Erdmann Rapp[1,2]

[1]Department of Bioprocess Engineering, Max Planck Institute for Dynamics of Complex Technical Systems, Magdeburg, Germany
[2]glyXera GmbH, Magdeburg, Germany
[3]Department of Bioprocess Engineering, Otto-von-Guericke University, Magdeburg, Germany

Received December 21, 2012
Revised March 7, 2013
Accepted March 8, 2013

glyXalign: High-throughput migration time alignment preprocessing of electrophoretic data retrieved via multiplexed capillary gel electrophoresis with laser-induced fluorescence detection-based glycoprofiling

Glycomics has become a rapidly emerging field and monitoring of protein glycosylation is needed to ensure quality and consistency during production processes of biologicals such as therapeutic antibodies or vaccines. Glycoanalysis via multiplexed CGE with LIF detection (xCGE-LIF) represents a powerful technique featuring high resolution, high sensitivity as well as high-throughput performance. However, sample data retrieved from this method exhibit challenges for downstream computational analysis due to intersample migration time shifts as well as stretching and compression of electropherograms. Here, we present glyXalign, a freely available and easy-to-use software package to automatically correct for distortions in xCGE-LIF based glycan data. We demonstrate its ability to outperform conventional algorithms such as dynamic time warping and correlation optimized warping in terms of processing time and alignment accuracy for high-resolution datasets. Built upon a set of rapid algorithms, the tool includes an intuitive graphical user interface and allows full control over all parameters. Additionally, it visualizes the alignment process and enables the user to readjust misaligned results. Software and documentation are available at http://www.glyxera.com.

1 Introduction

The importance of glycosylation with respect to critical quality attributes of proteins used in therapy and prevention of disease is a widely recognized fact [1–5] and serves as one of the major driving forces behind the recent boom of glycoanalysis in basic research, medicine, and pharmaceutical industry alike. While glycoprofiling is used to make major progress in clinical biomarker discovery [6–8], the need for efficient control of antigen glycosylation of novel cell culture based vaccines [9–12] and well-established biologicals such as monoclonal antibodies and erythropoietin, but also other biopharmaceuticals is equally on the rise [13, 14]. Traditional analytical methods based on LC or MS provide the means for analyzing moderately sized sample sets in adequate time, yet, the emerging practice of employing multiplexed CGE with LIF detection (xCGE-LIF)—originally proposed for the automated sequencing of DNA molecules [15]—promises to be a more cost-effective high-throughput approach [16–18]. This method electrophoretically separates fluorescently labeled N-glycans by size as well as electric charge and measures their fluorescence as they pass the LIF detector. This results in characteristic electropherograms in which signal peaks correspond to the different glycan structures contained in the sample.

Generally, every separation technique suffers from certain distortions in its measured data and the xCGE-LIF method is no exception. Specifically, the phenomenon of intersample migration time shifts presents one of the most prominent hurdles for comparing similar electropherograms, for example, those retrieved from different patients in a clinical study or for quality monitoring of antibodies produced in different batches.

Therefore, in order to preserve the high-throughput principle, a procedure to rapidly and automatically align electropherograms is required as a preprocessing step for further data analysis. For this purpose, several algorithms have been proposed in the past decades, from now well-established methods such as dynamic time warping (DTW) [19] and correlation optimized warping (COW) [20] to more recent variations such as PAGA (Peak Alignment by Genetic Algorithm) [21], PARS (Peak Alignment using Reduced Set mapping) [22], or alignDE [23].

Common to all of these approaches is that they employ mathematically sophisticated or otherwise complex solutions to correct distorted data in a generalized fashion, which often comes at the cost of decreased user friendliness and increased processing time for large datasets. However, based upon our experience, intersample variations in similar electropherograms do not appear to exhibit a high degree of complexity. Therefore, in the following we propose a comparatively simple, yet powerful algorithm implemented in MATLAB to adequately align batches of similar data while still maintaining reasonable running times. Furthermore, a graphical user interface (GUI) is provided aiming to facilitate and visualize the preprocessing during its execution. The glyXalign software can be downloaded from http://www.glyxera.com.

2 Materials and methods

2.1 Experimental data

As validation dataset, we used xCGE-LIF measurements of N-glycans released from whole human plasma glycoproteome obtained from healthy donors. Human plasma exhibits a characteristic complex composition of various N-glycans [24], which is stable in all samples. Therefore, any distortions observed in the data can only result from the measurement itself.

Sample preparation was performed as published in [17]. A 3130 Genetic Analyzer (Applied Biosystems) was used for the measurements. Obtained raw data were converted into XML format using the freely available *Data File Converter* (http://www.appliedbiosystems.com).

2.2 Algorithm

The core feature of the implemented alignment algorithm constitutes the detection of positions of characteristic data points in electropherograms. These landmark vectors serve as representatives for the electropherograms, meaning that all transformations applied to these vectors will also be applied to the original data at the end of the alignment process. The most striking landmarks inside electropherograms are their peaks that are usually well separated from the baseline. The simple peak detection procedure employed in the algorithm involves initial smoothing of the electropherograms using Savitzky–Golay filtering [25], application of a baseline correction algorithm described in [26], and subsequent screening of the smoothed data for local maxima by comparing intensities of neighboring data points.

The peak position vectors retrieved in this manner can be warped arbitrarily by applying a power series of the general form

$$P(x) = \sum_{n=0}^{\infty} a_n (x - b)^n \tag{1}$$

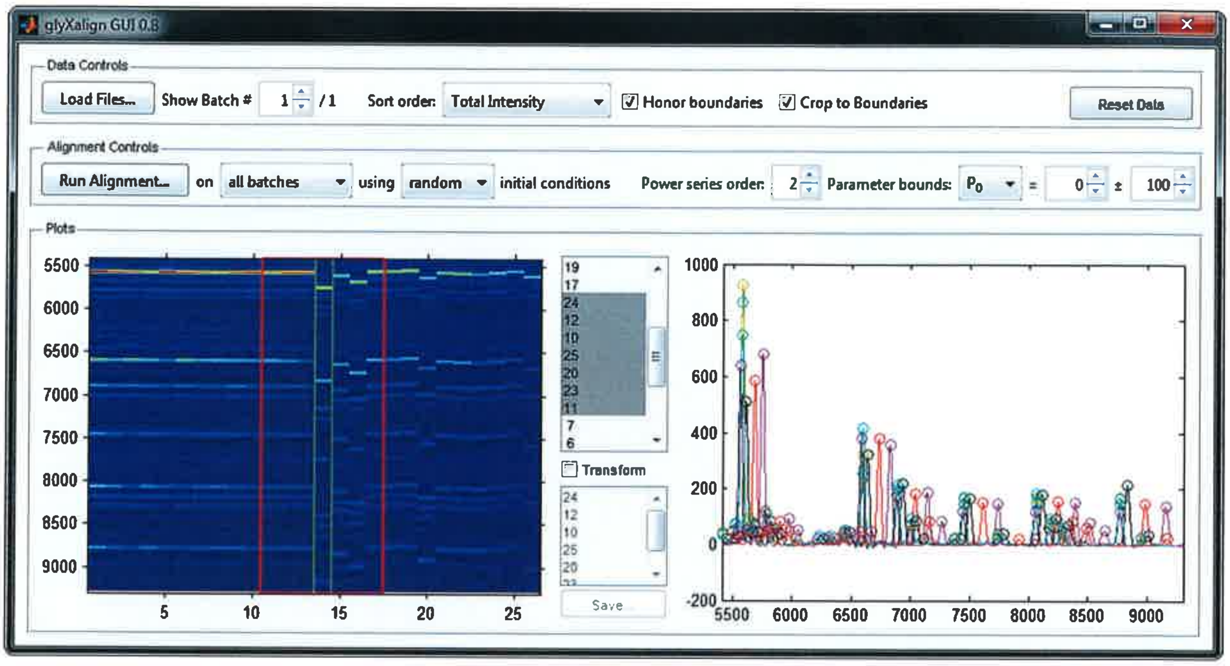
to each of their elements. The constant coefficient a_0 can be interpreted as a defined shift along the migration time axis, whereas a_1 serves the purpose of linearly stretching or compacting peak positions. Higher order coefficients in turn represent nonlinear distortions.

As with any alignment method, a suitable reference needs to be picked from the selected dataset. Other electropherograms will conform to the shape of this reference when suitable power series coefficients for warping their peak vectors can be determined and applied to them in a similar fashion. This is realized in the GUI (see Fig. 1) by applying a series of parameter optimizations using an implementation of the Nelder–Mead simplex algorithm [27] for minimizing the simple pairwise distance metric

$$d(A, B) = \sum_{i=1}^{m} \min_{j \in [1,n]} \left| B_j - A_i \right| \tag{2}$$

in which A and B represent warped peak vectors of two different electropherograms.

Since this objective function frequently contains several local minima, the optimal solution to the alignment problem is not always inherently apparent and multiple strategies to improve upon this situation have been devised and implemented. For instance, each optimization will be repeated a defined number of times with random starting conditions in an attempt to find the global minimum. Furthermore, the aforementioned peak detection procedure is sensitive to the general signal intensity of the data, which results in fewer detectable peaks in low-intensity electropherograms. The resolution with respect to the migration time, however, has less of an impact on alignment accuracy. For example, the common case of two adjacent peaks, which cannot be resolved via CGE, produces shorter peak vectors, yet accuracy is not affected significantly as the dominant peaks remain. Peak vectors with significantly different lengths tend to align worse to each other than those with similar lengths. Therefore, the dataset may be sorted with respect to certain intensity-dependent qualities, for example, total signal intensity, maximum intensity, or the number of peaks inside a specified migration time interval. By doing so, the algorithm is provided with a defined order of pairwise alignment operations in which the preceding neighbor serves as a temporary reference. To avoid gradual shifts attributable to accumulating minor



glyXalign GUI 0.8

Data Controls

Load Files... Show Batch # 1 /1 Sort order: Total Intensity ☑ Honor boundaries ☑ Crop to Boundaries Reset Data

Alignment Controls

Run Alignment... on all batches using random initial conditions Power series order: 2 Parameter bounds: P0 = 0 ± 100

Plots

Figure 1. Screenshot of the glyXalign GUI application during the process of alignment. The left-hand side features a heat map view of all 26 electropherograms of the currently selected batch in which each column represents the signal intensities along the migration time axis of a single electropherogram. The single lane framed in green marks the electropherogram that is currently in the process of being aligned. The electropherograms of the lanes framed in bold red are displayed in the right hand side panel together with their detected peaks (circular markers).

misalignments, the electropherograms are aligned to the primary global reference in a second optimization step using the parameters established in the first step as starting conditions.

Despite these strategies, the batch-wide alignment process remains prone to errors due to the random nature of the optimization or due to limitations in the overall number of features being detectable in some samples. As a consequence, some electropherograms may refuse to align properly after repeated attempts. For these cases, the GUI allows the user to interactively assign suitable starting conditions for subsequent optimizations by hand.

After the entire batch has been processed and all alignment coefficients have been determined, the migration time vectors of the respective original electropherograms will be warped and their signal values subsequently sampled at the original grid positions by way of linear interpolation.

3 Results and discussion

To evaluate the performance of the proposed algorithm in comparison to established methods, we carried out a series of tests to determine average running times as well as alignment accuracy. In addition to the proposed approach, we used MATLAB implementations of the DTW and COW algorithms by Tomasi et al. [28].

The initial state of the test dataset before any alignment preprocessing was applied is shown in Fig. 2A–C. Characteristic distortions in the form of migration time shifts as well as stretching or compression of interpeak areas are clearly evident. The result of successful alignment is displayed in Fig. 2D–F.

Table 1. Comparison of the different alignment algorithms applied to two different parameter sets (see Supporting Information)

Method	Average distance	Average correlation	Average running time (s)
DTW 1	106.20 ± 269.49	0.966 ± 0.018	8441.0 ± 283.4
DTW 2	141.40 ± 62.24	0.934 ± 0.045	230.0 ± 0.4
COW 1	85.44 ± 58.74	0.971 ± 0.013	1567.7 ± 11.5
COW 2	97.00 ± 76.13	0.968 ± 0.015	8.3 ± 0.1
glyXalign 1	44.74 ± 41.62	0.969 ± 0.029	17.6 ± 0.5
glyXalign 2	30.37 ± 21.71	0.977 ± 0.012	28.5 ± 0.4

Each alignment method was repeated with different parameter sets in order to measure their impact on alignment quality for each of the test runs (Table 1). The alignment accuracy for each method was determined by computing the average distance of all electropherograms to the reference and the average of their respective Pearson correlation coefficients. The results obtained from the DTW and COW implementations exhibit correlation values that are on par with or better than those of the glyXalign method, but perform worse in terms of the average peak distance metric.

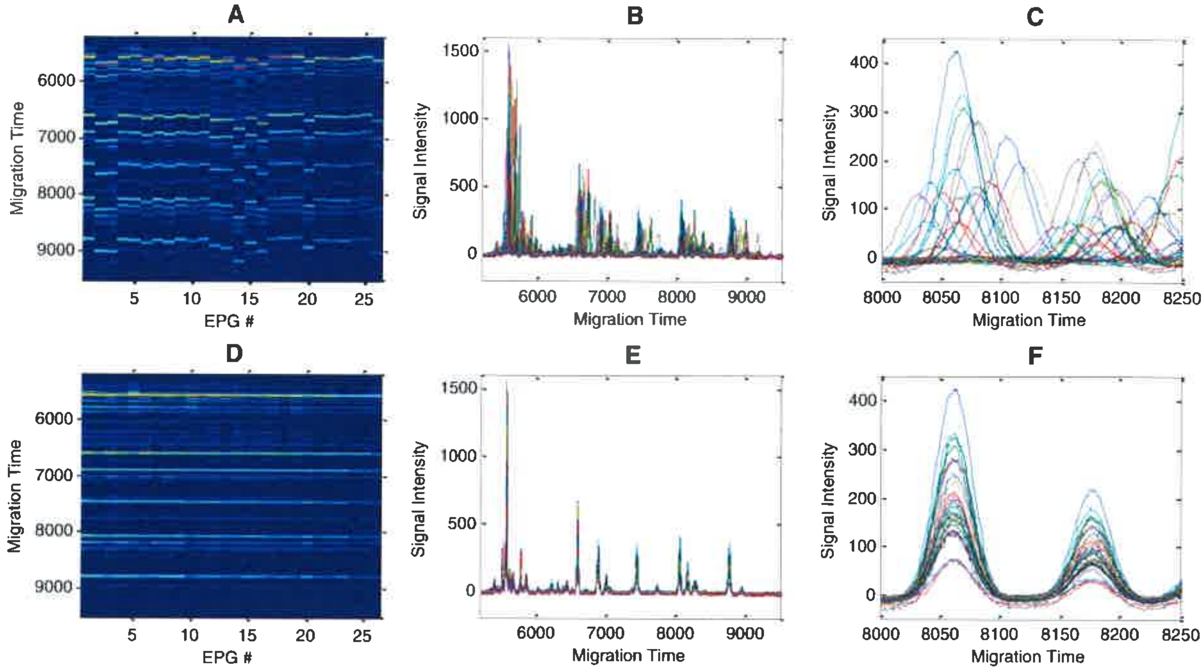


Figure 2. Graphical representations of the test dataset before processing (A–C) and after successful alignment (D–F). (A) Heat map of the dataset cropped to the dimensionless migration time range containing the peaks of interest (each column represents one electropherogram; the leftmost column contains the reference). The entire dataset is sorted with respect to the total signal intensity in descending order. (B) Intensity versus migration time plot of all electropherograms in the dataset. (C) Detailed view of (B) in the migration time range 8000–8250. (D–F) Same as (A–C), after alignment.

Due to the architecture of the DTW and COW algorithms, several concessions in terms of data preparation had to be made when measuring their performance. Most significantly, both algorithms are designed to work on data with a coarser temporal resolution than present in the test dataset. To maintain reasonable running times, only every fifth data point was sampled to reduce the overall complexity of the alignment problem for these algorithms. No such treatment was necessary as data reduction constitutes an integral part of glyXalign method in the form of the described peak detection routine.

The DTW and COW implementations were tested both on the original and the down-sampled dataset. Suitable parameters for aligning the down-sampled dataset could be determined (see Supporting Information), but applying these to the original data resulted in processing times exceeding several hours, rendering this approach impractical. It has to be noted, that the DTW and COW algorithms are able to perform much faster under favorable conditions, but determining the optimal set of parameters is rarely a trivial task and highly depends on the composition of the used dataset.

The glyXalign method was tested using first-order and second-order power series settings. The results show that the majority of electropherograms will align reasonably well using simple first-order warping, but subtle distortions remain, such as partial misalignments near the edges of the chosen migration time interval. At the cost of slightly increased processing time, these imperfections can be dealt with adequately by applying the alignment technique using second-order parameters. Higher orders will result in significantly longer running time while yielding marginally better accuracy. On top of that, this type of configuration potentially comes with the undesirable side effect of destabilizing the optimization routine as local minima possibly occur more often (Supporting Information Fig. 1). Yet, the effort required to readjust the affected electropherograms is minimal when using the GUI, which renders the process easy and comfortable.

4 Concluding remarks

A software tool is presented providing a GUI for interactively aligning batches of xCGE-LIF-based data, which can be used even by nonexperts. The algorithm implemented enables accurate alignment of electropherograms and efficient preprocessing of large sample sets, for example, gained from clinical studies or industrial quality monitoring, thereby significantly extending the scope of toolboxes for glycan analysis [18]. Finally, the proposed method is not limited to electrophoretic data and may be applied to measurements derived from other analytical methods such as chromatography, spectrometry, or spectroscopy.

The authors would like to thank Dirk Benndorf for his support. This work has been supported by funding from the European

Union's Seventh Framework Programme (FP7-Health-F5-2011) under grant agreement no. 278535 (HighGlycan).

The authors have declared no conflict of interest.

5 References

[1] Apweiler, R., Hermjakob, H., Sharon, N., *Biochim. Biophys. Acta* 1999, *1473*, 4–8.

[2] Varki, A., *Glycobiology* 1993, *3*, 97–130.

[3] Shental-Bechor, D., Levy, Y., *Curr. Opinion Struct. Biol.* 2009, *19*, 524–533.

[4] Ruhaak, L. R., Uh, H. W., Beekman, M., Koeleman, C. A., Hokke, C. H., Westendorp, R. G., Wuhrer, M., Houwing-Duistermaat, J. J., Slagboom, P. E., Deelder, A. M., *PloS One* 2010, *5*, e12566.

[5] Hütter, J., Rödig, J., Höper, D., Seeberger, P. H., Reichl, U., Rapp, E., Lepenies, B., *J. Immunol.* 2013, *190*, 220–230.

[6] Packer, N. H., von der Lieth, C. W., Aoki-Kinoshita, K. F., Lebrilla, C. B., Paulson, J. C., Raman, R., Rudd, P., Sasisekharan, R., Taniguchi, N., York, W. S., *Proteomics* 2008, *8*, 8–20.

[7] Ohtsubo, K., Marth, J. D., *Cell* 2006, *126*, 855–867.

[8] Hart, G. W., Copeland, R. J., *Cell* 2010, *143*, 672–676.

[9] Rödig, J., Rapp, E., Kampe, M., Kaffka, H., Bock, A., Genzel, Y., Reichl, U., *Biotechnol. Bioeng.* 2013, *110*, 1691–1703.

[10] Rödig, J., Rapp, E., Djeljadini, S., Lohr, V., Genzel, Y., Jordan, I., Sandig, V., Reichl, U., *J. Carbohydr. Chem.* 2011, *30*, 281–290.

[11] Schwarzer, J., Rapp, E., Hennig, R., Genzel, Y., Jordan, I., Sandig, V., Reichl, U., *Vaccine* 2009, *27*, 4325–4336.

[12] Schwarzer, J., Rapp, E., Reichl, U., *Electrophoresis* 2008, *29*, 4203–4214.

[13] Kawasaki, N., Itoh, S., Hashii, N., Takakura, D., Qin, Y., Huang, X., Yamaguchi, T., *Biol. Pharm. Bull.* 2009, *32*, 796–800.

[14] Hecht, M. L., Stallforth, P., Silva, D. V., Adibekian, A., Seeberger, P. H., *Curr. Opin. Chem. Biol.* 2009, *13*, 354–359.

[15] Callewaert, N., Geysens, S., Molemans, F., Contreras, R., *Glycobiology* 2001, *11*, 275–281.

[16] Laroy, W., Contreras, R., Callewaert, N., *Nat. Protoc.* 2006, *1*, 397–405.

[17] Ruhaak, L. R., Hennig, R., Huhn, C., Borowiak, M., Dolhain, R. J., Deelder, A. M., Rapp, E., Wuhrer, M., *J. Proteome Res.* 2010, *9*, 6655–6664.

[18] Rapp, E., Hennig, R., Borowiak, M., Kottler, R., Reichl, U., *Glycoconjug. J.* 2011, *28*, 234–235.

[19] Wang, C. P., Isenhour, T. L., *Anal. Chem.* 1987, *59*, 649–654.

[20] Nielsen, N. P. V., Carstensen, J. M., Smedsgaard, J., *J. Chromatogr. A* 1998, *805*, 17–35.

[21] Forshed, J., Schuppe-Koistinen, I., Jacobsson, S. P., *Anal. Chim. Acta* 2003, *487*, 189–199.

[22] Torgrip, R. J. O., Åberg, M., Karlberg, B., Jacobsson, S. P., *J. Chemom.* 2003, *17*, 573–582.

[23] Zhang, Z. M., Chen, S., Liang, Y. Z., *Talanta* 2011, *83*, 1108–1117.

[24] Gornik, O., Wagner, J., Pucic, M., Knezevic, A., Redzic, I., Lauc, G., *Glycobiology* 2009, *19*, 1547–1553.

[25] Savitzky, A., Golay, M. J. E., *Anal. Chem.* 1964, *36*, 1627–1639.

[26] Eilers, P. H. C., Boelens, H. F. M., *Technical Report, Baseline Correction with Asymmetric Least Squares Smoothing*, Leiden University Medical Centre, Leiden, The Netherlands 2005, http://www.science.uva.nl/~hboelens/publications/draftpub/Eilers_2005.pdf.

[27] Lagarias, J. C., Reeds, J. A., Wright, M. H., Wright, P. E., *SIAM J. Optim.* 1998, *9*, 112–147.

[28] Tomasi, G., van den Berg, F., Andersson, C., *J. Chemom.* 2004, *18*, 231–241.