MIRAGE: Minimum Information Required for a Glycomics Experiment

William S. York* and René Ranzinger

Complex Carbohydrate Research Center, The University of Georgia,
315 Riverbend Road, Athens GA, 30602-4712, U.S.A.

E-Mail: *will@ccrc.uga.edu

Abstract

The interpretation, evaluation, and reproduction of glycomics and glycoproteomics experiments are impeded by the failure to provide scientists who consume the results of these analyses with sufficient information describing the methods used to obtain the analytical data. With the enthusiastic support of glycoscientists and journal editors, a new initiative to specify the Minimum Information Required for A Glycomics Experiment (MIRAGE) has been established to address the unique data reporting aspects of glycoanalytic experiments. MIRAGE aims neither to dictate or control the experimental techniques used in a glyco-analysis nor to establish a metric for judging the quality of an experiment. Rather, it merely enumerates the information (data and metadata) that should be provided when the results of a glycoanalytic study are submitted to a journal or database.

Introduction

Glycobiology is an emerging discipline that focuses on the biosynthesis, structure and biological functions of carbohydrates [1], which exist as pure glycans or as key structural and regulatory components of proteins and lipids. Our understanding of the roles of glycosylation in development and disease has increased dramatically in recent years, due in part to advances in analytical methods for identifying and quantifying very small amounts of complex glycans that are present in biological samples. These advances have facilitated research that revealed critical roles of glycosylation in development [2, 3] homeostasis,

inflammation, vascular biology [4], neuromuscular disease [5] and a wide range of conge-
nital disorders of glycosylation [6]. Advances in glycochemistry show great promise for the
development of therapeutic and diagnostic tools [7], including prophylactic vaccines and
cancer detection technologies [8].

Diversity in the biological functions of glycoconjugates (such as glycoproteins and glyco-
lipids) is mirrored by the structural diversity of these complex molecules [9]. Glycan
biosynthesis is an elaborate process involving the transfer of individual sugar residues from
donor molecules such as nucleotide sugars to growing glycan chains [10]. These reactions
are catalysed by glycosyltransferases that recognize both the donor substrate and the accep-
tor substrate (the growing chain) to produce structures that are often branched. The glyco-
syltransferases direct the formation of glycosidic linkages, connecting each individual mono-
saccharide at a defined location (i. e., linked to one of several available oxygen atoms on the
acceptor residue). Formation of the glycosidic bond locks the oxygen forming the glycosyl
linkage in either the "alpha" or "beta" orientation, which defines the anomeric configura-
tion of the newly added monosaccharide residue. Glycosidic bond formation also locks the
newly added monosaccharide residue in a specific ring form (furanosyl or pyranosyl) with
either 5 or 6 atoms, respectively. Thus, in order to completely describe the structure of a
glycan, the following features must be specified: (1) the identity of each monosaccharide
residue; (2) the ring form (pyranosyl, furanosyl or open chain) of each monosaccharide
residue; (3) the anomeric configuration of each residue; (4) the type and position of mod-
ifications of the monosaccharide (such as deoxygenation, double bonds, acetylation, sulpha-
tion); (5) the attachment site for glycosyl linkages between residues; (6) the residue sequence
(and resulting molecular branching pattern). The extent to which each of these features can
be established depends on the analytical and data processing methods employed. Therefore,
a full description of the results of a glycomics analysis requires information regarding the
specific methods used in gathering and processing the data [11].

Glycomics involves the identification and quantification of all detectable glycan structures in
a sample derived from a particular organism, tissue, or cell type. Glycomics analyses are
often focused on determining how the populations of various glycans change as a result of
cell differentiation, tissue morphogenesis, disease progression or genetic manipulation. Due
to the molecular and biological complexity of the glycosylation process, thorough reporting
of the results of a glycomics experiment is highly challenging. The resulting data specifies
the identity and quantity of relatively large, complex structures, whose precise molecular
features are often inferred using prior knowledge, such as familiarity with a particular
biosynthetic mechanism. Specifying the exact methods and assumptions that were used to
assign and quantify reported structures will allow the biomedical scientist to appreciate the
scope and depth of the analysis and to evaluate correlations between defined structural
features and specific biological processes such as tissue development or disease progression.
Of course, such analytical data is only useful if it is interpreted in light of the biological and
experimental context in which the glycans were prepared for analysis.

Interpretation and reproducibility of glycomics data thus requires comprehensive meta-data, that is, information regarding the biological sample, sample handling, data acquisition and data processing. Standards specifying minimum meta-data sets have been established for several types of biomedical analysis, including microarray-based transcriptomics (MIAME) [12] and proteomics (MIAPE) [13]. MIAPE (Minimum Information About a Proteomics Experiment) consists of guidelines specifying "the minimum information that should be reported about a data set or an experimental process, to allow a reader to *interpret* and *critically evaluate* the conclusions reached, and to support their *experimental corroboration*" [14]. MIAPE specifies "neither the format in which information should be transferred nor the structure of any repository or document" [14] and is implemented as "a checklist of information that should be provided (for example about the protocols employed) when a data set is submitted to a public repository or when an experimental step is reported in a scientific publication" [14].

Taylor *et al.* [15] described the philosophical basis for the MIAPE standard:

> "It has always been a matter of policy that the PSI should neither attempt to produce standard operating procedures specifying how particular techniques should be performed nor attempt to establish quality assessment benchmarks. We do not believe it is the job of this body to dictate to the proteomics community how it should perform experiments or analyses."

Taylor and his co-authors also describe two fundamental criteria required for successful implementation of MIAPE: (1) *sufficiency* – "The ... guidelines should require sufficient information about a dataset and its experimental context to allow a reader to understand and critically evaluate the interpretation and conclusions, and to support their experimental corroboration." and (2) *practicability* – "Achieving compliance ... should not be so burden-some as to prohibit its widespread use."

The previously established standards, including as MIAME and MIAPE, are not independent, but have significant overlap. For example, much of the same information regarding the biological source of the analysed material is required for both microarray and proteomics analysis. Fortunately, a community of scientists has agreed to work together to coordinate diverse (but overlapping) checklists of the Minimum Information for Biological and Biomedical Investigations (MIBBI) [16]. Access to many of the domain-specific "minimum information" standards is available under the MIBBI umbrella (http://mibbi.org/).

Compliance with the MIAME and MIAPE standards is frequently stipulated as a prerequisite for publication of microarray and proteomics experiments, as including all of the data and metadata specified in these checklists helps ensure that the experiments can be understood, evaluated and reproduced. The same sets of information are required by databases that store microarray and proteomics data. Although the MIAME and MIAPE standards identify the

specific data and metadata that is required to effectively describe the experiments, they do not specify how these data should be represented, and digital standards for the representation and exchange of these data is beyond the scope of MIAME and MIAPE. Nevertheless, such digital standards are required for the effective data handling by software applications. Complementary initiatives that establish standard file formats to represent the meta information describing these experiments, along with dictionaries and ontologies to maintain semantic consistency among databases, will provide more extensive access to database searches (e. g., to find experiments that were performed using the same setup parameters), automated comparison of experimental results and mining of experimental data in the light of the meta information.

MIRAGE

An international group of glycobiologists (Will York, Catherine Costello, Hans Kamerling, Jonathan Bones, Joe Zaia, Niclas Karlsson, and Stuart Haslam) was organized at the Consortium for Functional Glycomics Conference in Washington 2009 (http://glycomics.scripps.edu/CFGWorkshopApril2009.html) to address the minimum information issue in the glycomics field. A key goal of this working group was to work towards the establishment of a **MIRAGE** (*Minimum Information Required for A Glycomics Experiment*) standard that is similar to the MIAPE standard for proteomics data, but that addresses the distinctive challenges of reporting glycomics analyses. The initial discussions have been very well received by scientists in the glycobiology community, who recognize the urgent need for a standard like MIRAGE. International leaders in the development and application of new techniques for glycomics analysis and the development of infrastructure and tools for glycoinformatics have been joined by the editors of the most influential journals that publish glycomics and glycoproteomics research to express their enthusiastic support for a MIRAGE standard.

Similar to the MIAPE standard for proteomics, the MIRAGE standard should focus on the analytical technologies and data interpretation issues associated with glycomics analysis. It should thus describe all of the information that would be required to reproduce and process data that was used to characterize a group of glycan structures or an individual structure. In this context, MIRAGE should specify the techniques and assumptions used to deduce each structural feature (e. g., composition, linkage, anomeric configuration, ring form, etc.) of the reported glycan or glycoconjugate.

The practical limitations of high-throughput glycomics analysis have been discussed by members of the glycobiology community at several meetings, including Workshops sponsored by the Consortium for Functional Glycomics (http://glycomics.scripps.edu/CFGWorkshopApril2009.html, http://glycomics.scripps.edu/SubgroupWorkshop/AnnouncementOct2009.pdf) and the Warren Workshop for Glycan Analysis (http://www.biomedicine.gu.se/biomedicine/Charles_Warren_Workshop_III).

High-throughput methods (such as mass spectrometry) do not always provide explicit information that can be used to make complete, unambiguous structural assignments of all the isomeric structures present in an analyte, and assignments are often made on the basis of biological considerations. The community consensus is that glycomics data reporting standards (such as MIRAGE) should *not* preclude the publication of such assignments. Rather, the standard should facilitate identification of the assumptions (such as "biosynthetic rules") that are invoked when explicit analytical data is not available. For example, when PNGase-F is used to release oligosaccharides from a mammalian glycoprotein, commonly accepted biochemical rules allow one to confidently propose that D-Glc*p*NAc (rather than just Hex-NAc) is the sugar residue at the reducing end of each of the oligosaccharides. By recognizing these issues, MIRAGE is intended to be a guide that improves the information content of a glycomics analysis report rather than an impediment to innovation or publication.
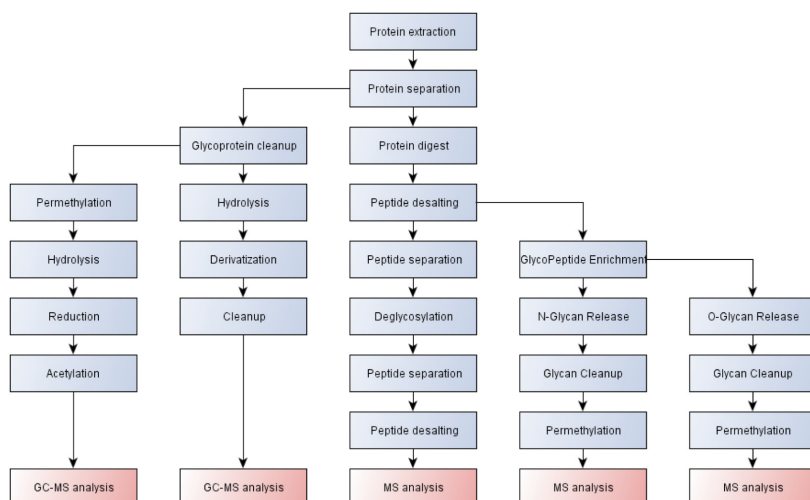

Figure 1. Typical experimental protocol for a glycomics analysis.

Due to the unique issues associated with glycomics analysis, a distinct MIRAGE standard (rather than an extension of MIAPE) is required. Many sample preparation issues are unique to glycomics (Figure 1). Furthermore, glycomics involves analysis of complex, branched molecules whose primary structures cannot (yet) be predicted *in silico* and that are often non-trivial to represent digitally or graphically. Another important aspect of glycomics is that it involves the analysis of molecules that have a high tendency for mass degeneracy.

Ideally, MIRAGE should allow each structural feature that is reported in a glycomics experiment to be *individually* identified along with the methods used to characterize or identify these features. It is important to note that each structure identified in a single glycomics or glycoproteomics analysis may be reported with a different confidence level. Special attention may be paid to particular glycans that are highly abundant or deemed

especially relevant. Thus, the amount and type of analytical data used to make structural assignments can vary considerably within a single analysis. For example, a structure may be assigned solely on its mass (deduced from mass profiling data), ruling out alternative structures on the basis of glycosyl residue composition data and/or structural assumptions based on the biology and history of the sample. Other structures may be based on multi-stage tandem mass spectrometry (MS^n), providing explicit information regarding the sequence, branching pattern and linkage pattern of the oligosaccharide. Different methods, such as NMR and High-Performance Liquid Chromatography (HPLC) can also provide information to assign structural details and/or quantify expression levels for individual glycans or glycoconjugates. Precisely identifying the method used for each structural and quantitative assignment makes it possible to reach informed conclusions when using these data to evaluate biological hypotheses. Thus, a key feature of MIRAGE is that it identifies analytical data and meta-data that, when provided, allows the depth of analysis to be evaluated for *each* reported structure.

The development and community-wide acceptance of a MIRAGE standard will be facilitated by reusing concepts and approaches that have been developed for other "minimum information" standards that are included in the MIBBI project. In keeping with the lessons learned during the development of these "minimum information" standards, MIRAGE should not dictate how experiments are performed or try to rank the quality of the experiment based on the provided information.

The importance of robust software for identifying, organizing and recording important metadata describing glycomics experiments cannot be overemphasized. Real-world implementation of the MIRAGE standard will be facilitated by making the collection of relevant metadata routine and consistent, especially if appropriate tools are available to help with this task. Furthermore, the data collection process will become easier once the analyst has done it a few times. Tools that read and parse data files generated by analytical instrumentation to automatically collect and record hardware configurations and data acquisition parameters are essential. User friendly and intuitive interfaces for template-based metadata entry will allow information collected for previous glycomics experiments to be reused, requiring the explicit entry of only those parameters that change from one experiment to the next. Templates describing frequently used analytical processes (experiments) can be shared, allowing a significant proportion of the required information to be specified by reusing information stored in these templates. The automatic generation of digital reports using well-defined data formats will facilitate the reuse and comparison of the data and meta-data. However, the development and implementation of such software tools is not part of the MIRAGE initiative. Rather, the defined checklists of meta information specified by MIRAGE will provide guidance for the software and database development.

SUMMARY

The definition and establishment of MIRAGE will require effort from the glycomics community to achieve consensus regarding the critical information that should be reported and effort from individual glycoanalysts to collect, enter and transfer this information. Nevertheless, providing this minimum set of meta-information will bring significant benefits. Based on the enthusiastic support of members of the glycobiology community, including editors of several high-impact journals, we envision that the MIRAGE standard will be used to encourage authors to collect and provide information that allows deep understanding and reproducibility of a manuscript's content. This will facilitate evaluation of the depth of glycomics data and the comparison of data sets. We believe that it is critical to provide a uniform specification of glycan structural data, meta-data and annotations. By identifying and naming the most important data and meta-data describing glycomics experiments, MIRAGE will facilitate the improvement of databases for carbohydrate structures and the development of software tools for data mining and logical inference using glycomics data. This will increase the information of these databases and provide more robust computational methods to generate and evaluate hypotheses regarding the roles of glycosylation in development and disease.

In summary, the motivation, philosophy and benefits of a MIRAGE standard are described below.

- Unique aspects of glycomics data acquisition, processing, interpretation and mining demand a new data-reporting standard – **MIRAGE.**

- MIRAGE should borrow extensively from and overlap with MIAPE and other reporting standards.

- MIRAGE will benefit glycoanalysts, glycobiologists and biomedical scientists in general in the evaluation and reproduction of published glycomics experiments.

- MIRAGE will not implement software but provide the framework for developers of software and database projects by defining the set of information that should be stored and/or handled by these applications.

- Bioinformaticians can facilitate implementation of MIRAGE by providing infrastructure for data processing, exchange, archiving and retrieval.

- Realization of MIRAGE depends on leadership from respected members of the glycoanalytic field. A MIRAGE initiative led primarily by computer scientists is unlikely to be fully accepted by the scientific community, making it difficult to obtain funding to develop the infrastructure required for its effective implementation.

OUTLOOK

As a result of the 2nd Beilstein Symposium on Glyco-Bioinformatics in 2011, a working group was formed with the mission to define the MIRAGE guidelines for reporting glycomics experiments. This group consists of informaticists and glycoanalysts (Sanjay Agravat, Daniel Kolarich, Masaki Kato, Joe Zaia, Matthew Campbell, Erdmann Rapp, Rene Ranzinger, Ryan McBride, Stuart Haslam, Weston Struwe, Will York). Development of MIRAGE will be overseen by a group of established leaders in the field of Glycomics. The Beilstein-Institut has expressed interest in assisting in the coordination and organization of this group, which will meet on a regular basis to define minimum information checklists for glycomics and glycoproteomics experiments. All results will be reported and made publicly available via the MIRAGE webpage (http://glycomics.ccrc.uga.edu/MIRAGE/). The group will work closely with the MIBBI project and other minimum information checklist initiatives.

ACKNOWLEDGEMENTS

REFERENCES

[1]   Varki, A. and Sharon, N. (2009) Historical Background and Overview. Chapter 1 in *Essentials of Glycobiology*, 2nd edition (Varki, A., Cummings, R.D., Esko, J.D., Freeze, H.H., Stanley, P., Bertozzi, C.R., Hart, G.W. and Etzler, M.E., eds.). Cold Spring Harbor (NY): Cold Spring Harbor Laboratory Press.

[2]   Haltiwanger, R.S. and Low, J.B. (2004) Role of Glycosylation in Development. *Annu. Rev. Biochem.* **73:**491 – 537.
       doi: http://dx.doi.org/10.1146/annurev.biochem.73.011303.074043.

[3]   Takeuchi, H. and Haltiwanger, R.S. (2010) Role of Glycosylation of Notch in Development. *Seminars in Cell and Developmental Biology* **21:**638 – 645.
       doi: http://dx.doi.org/10.1016/j.semcdb.2010.03.003.

[4]   Varki, A.P., Baum, L.G., Bellis, S.L., Cummings, R.D., Esko, J.D., Hart, G.W.,
        Linhardt, R.J., Lowe, J.B., McEver, R.P., Srivastava, A. and Sarkar, R. (2008) Work-
        ing Group Report: the Roles of Glycans in Hemostasis, Inflammation and Vascular
        Biology. *Glycobiology* **18:**747 – 749.
        doi: http://dx.doi.org/10.1093/glycob/cwn065.

[5]   Yoshida-Moriguchi, T., Yu, L., Stalnaker, S.H., Davis, S., Kunz, S., Madson, M.,
        Oldstone, M.B., Schachter, H., Wells, L. and Campbell, K.P. (2010) O-Mannosyl
        Phosphorylation of Alpha-Dystroglycan is Required for Laminin Binding. *Science*
        327:88 – 92.
        doi: http://dx.doi.org/10.1126/science.1180512.

[6]   Freeze, H. (2007) Congenital Disorders of Glycosylation: CDG-I, CDG-II, and
        beyond. *Current Molecular Medicin*e **7:**389 – 396.
        doi: http://dx.doi.org/10.2174/156652407780831548.

[7]   Finkelstein, J. (2007) Glycochemistry & Glycobiology. *Nature* **446:**999.
        doi: http://dx.doi.org/10.1038/446999a.

[8]   Wandall H.H., Blixt, O., Tarp, M.A., Pedersen, J.W., Bennett, E.P., Mandel, U.,
        Ragupathi, G., Livingston, P.O., Hollingsworth, M.A., Taylor-Papadimitriou, J.,
        Burchell, J. and Clausen, H. (2010) Cancer Biomarkers Defined by Autoantibody
        Signatures to Aberrant O-Glycopeptide Epitopes. *Cancer Res*earch **70:**1306 – 1313.
        doi: http://dx.doi.org/10.1158/0008-5472.CAN-09-2893.

[9]   Bertozzi, C.R. and Rabuka, D. (2009) Structural Basis of Glycan Diversity. Chapter 2
        in *Essentials of Glycobiology*, 2nd edition (Varki, A., Cummings, R.D., Esko, J.D.,
        Freeze, H.H., Stanley, P., Bertozzi, C.R., Hart, G.W. and Etzler, M.E., eds.). Cold
        Spring Harbor (NY): Cold Spring Harbor Laboratory Press.

[10]   Rini, J., Esko, J. and Varki, A. (2009) Chapter 5. Glycosyltransferases and Glycan-
        processing Enzymes. Chapter 5 in *Essentials of Glycobiology*, 2nd edition (Varki, A.,
        Cummings, R.D., Esko, J.D., Freeze, H.H., Stanley, P., Bertozzi, C.R., Hart, G.W.
        and Etzler, M.E., eds.). Cold Spring Harbor (NY): Cold Spring Harbor Laboratory
        Press.

[11]   Blow, N. (2009) Glycobiology: A Spoonful of Sugar. *Nature* **457:**617 – 620.
        doi: http://dx.doi.org/10.1038/457617a.

[12]   Brazma, A., Hingamp, P., Quackenbush, J., Sherlock, G., Spellman, P., Stoeckert, C.,
        Aach, J., Ansorge, W., Ball, C.A., Causton, H.C., Gaasterland, T., Glenisson, P.,
        Holstege, F.C., Kim, I.F., Markowitz, V., Matese, J.C., Parkinson, H., Robinson, H.,
        Robinson, A., Sarkans, U., Schulze-Kreme, S., Stewart, J., Taylor, R., Vilo, J. and

Vingron, M. (2001) Minimum Information about a Microarray Experiment (MIAME)-Toward Standards for Microarray Data. *Nature Genetics* **29:**365 – 371.
doi: http://dx.doi.org/10.1038/ng1201-365.

[13] Taylor, C.F., Paton, N.W., Lilley, K.S., Binz, P.A., Julian, R.K. Jr., Jones, A.R., Zhu, W., Apweiler, R., Aebersold, R., Deutsch, E.W., Dunn, M.J., Heck, A.J., Leitner, A., Macht, M., Mann, M., Martens, L., Neubert, T.A., Patterson, S.D., Ping, P., Seymour, S.L., Souda, P., Tsugita, A., Vandekerckhove, J., Vondriska, T.M., White-legge, J.P., Wilkins, M.R., Xenarios, I., Yates, J.R. 3 rd and Hermjakob, H. (2007) The Minimum Information about a Proteomics Experiment (MIAPE). *Nature Biotechnology* **25:**887 – 893.
doi: http://dx.doi.org/10.1038/nbt1329.

[14] Binz, P.-A., Barkovich, R., Beavis, R.C., Creasy, D., Horn, D.M., Julian, R.K., Seymour, S.L., Taylor, C.F. and Taylor, Y. (2008) Guidelines for Reporting the Use of Mass Spectrometry Informatics in Proteomics. *Nature Biotechnology* **26:**862.
doi: http://dx.doi.org/10.1038/nbt0808-862.

[15] Taylor, C.F., Binz, P.-A., Aebersold, R., Affolter, M., Barkovich, R., Deutsch, E.W., Horn, D.M., Hühmer, A., Kussmann, M., Lilley, K., Macht, M., Mann, M., Müller, D., Neubert, T.A., Nickson, J., Patterson, S.D., Raso, R., Resing, K., Seymour, S.L., Tsugita, A., Xenarios, I., Zeng, R. and Julian, R.K. Jr. (2008) Guidelines for Reporting the Use of Mass Spectrometry in Proteomics. *Nature Biotechnology* **26:**860 – 861.
doi: http://dx.doi.org/10.1038/nbt0808-860.

[16] Taylor, C.F., Field, D., Sansone, S.A., Aerts, J., Apweiler, R., Ashburner, M., Ball, C.A., Binz, P.A., Bogue, M., Booth, T., Brazma, A., Brinkman, R.R., Michael Clark, A., Deutsch, E.W., Fiehn, O., Fostel, J., Ghazal, P., Gibson, F., Gray, T., Grimes, G., Hancock, J.M., Hardy, N.W., Hermjakob, H., Julian, R.K. Jr., Kane, M., Kettner, C., Kinsinger, C., Kolker, E., Kuiper, M., Le Novère, N., Leebens-Mack, J., Lewis, S.E., Lord, P., Mallon, A.M., Marthandan, N., Masuya, H., McNally, R., Mehrle, A., Morrison, N., Orchard, S., Quackenbush, J., Reecy, J.M., Robertson, D.G., Rocca-Serra, P., Rodriguez, H., Rosenfelder, H., Santoyo-Lopez, J., Scheuermann, R.H., Schober, D., Smith, B., Snape, J., Stoeckert, C.J. Jr., Tipton, K., Sterk, P., Untergasser, A., Vandesompele, J. and Wiemann, S. (2008) Promoting Coherent Minimum Reporting Guidelines for Biological and Biomedical Investigations: the MIBBI project. *Nature Biotechnology* **26:**889 – 896.
doi: http://dx.doi.org/10.1038/nbt.1411.