ACS chemical biology

Letters

pubs.acs.org/acschemicalbiology

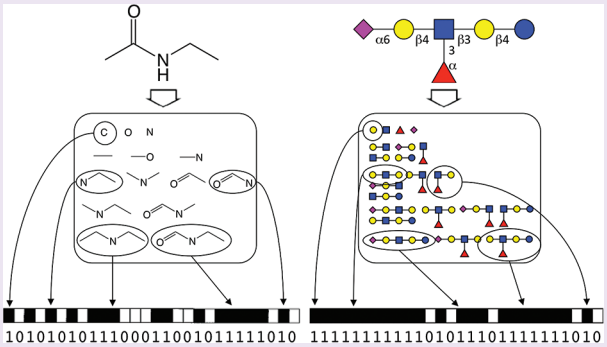
Glycan Fingerprints: Calculating Diversity in Glycan Libraries

Christoph Rademacher*,† and James C. Paulson

Departments of Chemical Physiology and Molecular Biology, The Scripps Research Institute, 10550 North Torrey Pines Road, La Jolla, California 92037, United States

Supporting Information

ABSTRACT: Carbohydrate libraries printed in glycan micorarray format have had a great impact on the high-throughput analysis of the specificity of a wide range of mammalian, plant, and bacterial lectins. Chemical and chemo-enzymatic synthesis allows the construction of diverse glycan libraries but requires substantial effort and resources. To leverage the synthetic effort, the ideal library would be a minimal subset of all structures that provides optimal diversity. Therefore, a measure of library diversity is needed. To this end, we developed a linear representation of glycans using standard chemoinformatic tools. This representation was applied to measure pairwise similarity and consequently diversity of glycan libraries in a single value. The diversities of four existing sialoside glycan arrays were compared. More diverse arrays are proposed reducing the number of glycans. This algorithm can be applied to diverse aspects of library design from target structure selection to the choice of building blocks for their synthesis.



Glycosylation is the most abundant post-translational modification of proteins and displays a much higher diversity than any other class of biopolymers.[1−3] Glycosylation and its information content underlies many biochemical and cell biological fundamentals important for life.[4] For instance, many glycan binding proteins regulate signaling pathways, thereby resembling glycans as a major component of the communication system. However, the storage capacity of information involved in communication has still to be deciphered. Recent efforts in the design of glycan arrays to elucidate the carbohydrate specificity of many lectins have advanced our understanding significantly.[5] Still the rational design of glycan libraries is currently driven by intuition and manual inspection of carbohydrate databases to choose a diverse subset of glycans, rather than an objective measure, and the outcome may be ambiguous.

Diversity lies in the heart of the rational design of chemical libraries, and the need to cover a broad chemical space has driven many theoretical research efforts since the early 1980s.[6,7] With the advent of automated chemical synthesis, chemo-enzymatic synthesis, and community efforts to establish large collections of glycan structures, carbohydrate chemistry groups are faced with questions similar to those the combinatorial screening community was asking more than a decade ago: What is the diversity of a given selection of compounds? Which structures are to be selected for diversification allowing only a limited number of building blocks or precursors?

The most basic question toward maximizing diversity is how similar are two molecules? While diversity is a description of a group of molecules, similarity is a measure of two members. The pairwise similarity of molecules can be used to measure diversity of a set[8] and is therefore essential for library diversification.[9] For a limited set of glycans, one may visually judge similarity, but this demands expert knowledge and is not manageable for a large number of structures. The definition of a clear concept of similarity and consequently the ability to compare molecules more objectively is fundamental.[10]

Determining molecular similarity has two major components: a descriptor that translates the molecular structure into a machine-readable format and a similarity coefficient that quantifies the level of similarity between pairs of molecules. Thousands of different descriptors have been developed since the 1980s, ranging but not limited to 2D descriptors such as molecular graphs and fingerprints to 3D descriptors. Despite the fact that 3D descriptors have intuitively superior information content, they often perform less well than 2D descriptions and are computationally more demanding.[7,11] Structural fingerprints such as Daylight fingerprints represent a simplified, linear representation of molecules.[12] These fingerprints are sequences of bits ("1"s and "0"s) representing the presence of connectivity pathways in a molecule. Using these so-called bit strings has proven to be very effective establishing a relationship between molecules and their biological properties.[13]

In the field of biomolecules, methods have been developed that are based on the linear structure of nucleic acids and peptides. In contrast, the analysis of the third major biopolymer, glycans, is much more complex as these structures

Received: January 4, 2012
Accepted: February 27, 2012
Published: February 28, 2012

 © 2012 American Chemical Society

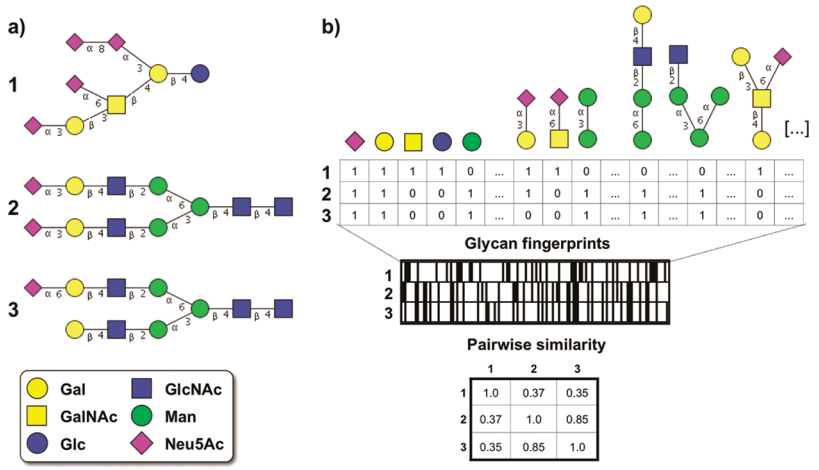


Figure 1. Principle of glycan fingerprints. (a) Complex glycan structures are presented in CFG annotation with simple representations of monosaccharides and glycosidic bonds. Three example glycans are depicted (1−3). (b) All unique fragments of a glycan are determined up to given path length (here: heptasaccharides) and encoded into a linear bit string using "1" for the presence of a fragment. Example fragments are shown along with a short representation of the corresponding bit string. The full glycan fingerprint given below has black areas for 1s and white areas for 0s. A pairwise comparison of the bit strings between 1−3 reflects their respective similarities given as a value between 0 and 1, with 1 being identical.

are branched. Previous approaches calculating the structural similarity of glycans have adopted methods from protein and nucleic acid analysis employing scoring matrices[14] or pattern extraction and kernel methods[15] or use shared disaccharide fragments as a descriptor to calculate similarity.[16]

Chemoinformatics solutions have inspired us to develop an orthogonal approach to this complex biological problem. The major abstraction in this approach is the treatment of complex glycan structures similar to small molecules. Monosaccharide units replace atoms, and glycosidic bonds substitute conventional chemical bond representations. Glycans are therefore treated as degenerate chemical graphs lacking circular assemblies of monosaccharide units. In this analogy, it is important to note that glycans exceed the diversity of small molecules of comparable size. Glycosidic bonds are more diverse than annotations of single, double, or triple bonds, and a larger alphabet of monosaccharides is found in glycans, which surpasses the average number of atom types found in small molecules[1] (Supplementary Figure 1).

We chose structural fingerprints of glycans as a way to transform a glycan structure into a linear representation, so-called "glycan fingerprints". To exemplify the principle of the method, three glycans were selected representing structures that are found on glycolipids (**1**) and *N*-glycans (**2**, **3**) (Figure 1). Each glycan can be represented by a bit string encoding for the presence of all of its fragments, *e.g.*, up to the size of a heptasaccharide. Herein, a fragment is defined as a subtree of n monosaccharides connected by $n - 1$ glycosidic bonds and is stored only once per occurrence. Heptasaccharide fragments were chosen as the upper limit of the path length in analogy to Daylight fingerprints, with path length being the number of monosaccharide connected in a fragment[12] (Supplementary Figure 1).

If a certain pattern is found in a glycan, a pseudorandom bit in the fingerprint is set to 1. In the example glycans (**1**−**3**), the monosaccharide *N*-acetyl neuraminic acid (Neu5Ac) is found in all glycans, while the disaccharide Neu5Acα(2−6)GalNAc is only found in the ganglioside (**1**). The resulting sequences of 1s and 0s can then be used to measure the similarity between two glycan fingerprints using a similarity coefficients. Because of the known dependency of many similarity coefficients on the relative size of the query and the comparison molecule, the modified Tanimoto (S_{MT}) coefficient has been proven to be a reliable measure and was chosen throughout this study. The Tanimoto coefficient is defined as the number of fragments present in both glycans divided by the sum of unique fragments of both. Its modification includes contribution of unset bits and thereby reduces the size bias.[17,18]

During the process of encoding the glycan into a sequence of bits, it is important to properly select the size of the sequence for sufficient information storage. If not set carefully, glycans become either too similar (false positives) because the bit string is filled with 1s or become artificially dissimilar (false negatives) because the number of 1s is low compared to blank bits.[12] We then empirically tested various conditions changing the initial bit string length and the number of times the string is folded using a selected test set of 8413 glycans from the KEGG database, a depository focused on mammalian carbohydrates.[19] With increasing size of a glycan there is an increase in the number of fragments that needs to be encoded. For instance, most tetrasacchrides can be deconvoluted into 10 unique fragments ranging from mono- to the parent tetrasaccharide. Up to 60 unique fragments can be found in octasaccharides (Figure 2a). Therefore, we were concerned whether the descriptor was able to cover larger structures without exceeding its storage capacity. The bit string density, being the fraction of 1s in the sequence, is a valid measure to assess if the encoding is still suitable for the analysis. Theoretical considerations suggest that bit strings can have a density up to 0.2−0.4 without losing specificity.[12] Hence, we varied the length and folding empirically and chose bit strings of a length of 1024 bits that are folded once to fulfill these demands (Figure 2b and Supplementary Figure 2a,b). When not applied to the analysis of the KEGG glycans the settings were adjusted according to the demands of the structures analyzed (Supplementary Figure 2c,d).

As a first validation, we used the pairwise similarity of a 33-member library of sialic acid terminated glycans, called sialosides, to calculate a similarity dendrogram (Figure 3a). We chose this particularly challenging example because these glycans all terminate with sialic acids. Thereby, the structural requirements for terminal sialylation render these glycans very

dx.doi.org/10.1021/cb300003z | *ACS Chem. Biol.* 2012, 7, 829−834

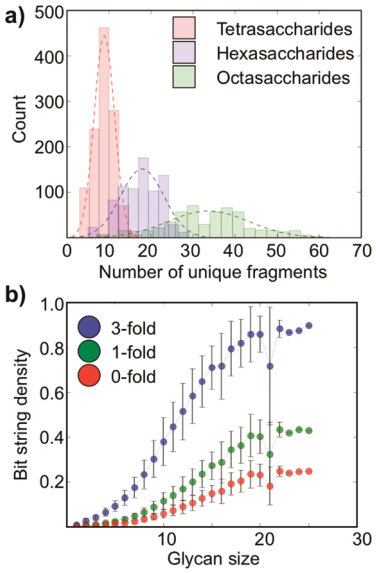


Figure 2. Analysis of the bit string representation of 8413 glycan structures from the KEGG database. (a) The average number of unique fragments increases with increasing size of a glycan. The distribution of unique fragments of tetra-, hexa-, and octasaccharides is shown in histograms. (b) Increasing size of a glycan results in an increasing number of 1s in the string (bit string density) and thereby raises the information density. Here, the bit string density is shown for strings with the length of 1024 bits being either 0-, 1-, or 3-times folded.

similar. We found very good agreement with our intuitive clustering. For instance, all *N*-glycans are clustered in a single branch of the dendrogram as they are similar due to the core nonasaccharide. It is also not surprising to see Neu5Gc containing glycans defining their own subtree, because the algorithm considers Neu5Gc as different from Neu5Ac as any other monosaccharide. Even though a single hydroxyl group discriminates the two monosaccharides, their recognition in a biological context is very different.[20] As another validation, we also applied the clustering to the mammalian glycan array from the Consortium for Functional Glycomics (CFG, version 3.1) and found that it leads to the identification of distinct activity islands (Supplementary Figure 3).

A more objective method for the validation of a descriptor rather than visual inspection is to investigate its neighborhood behavior.[21] This method is based on the similarity principle, which states, "Similar molecules have similar properties". Conversely, dissimilar molecules are expected to have high biological activity differences. Thus, a plot of dissimilarity *versus* biological activity difference should exhibit a characteristic trapezoidal distribution. Publicly available glycan array data through the CFG gateway were used, and the neighborhood behavior was assessed for 30 lectins. A low dissimilarity correlates well with a low activity difference. Conversely, a large structural difference may lead to either small or large activity differences. A statistically significant enhancement of data points in the lower right triangle was found, identifying glycan fingerprints as a valid diversity descriptor (Figure 3b and Supplementary Figure 4).

To show one application of the method, we chose four recently published glycan arrays to address the following questions: What are their diversities, and which structures are missing in the compound selection? We focused on sialic acid terminated glycans and used all sialosides from the current

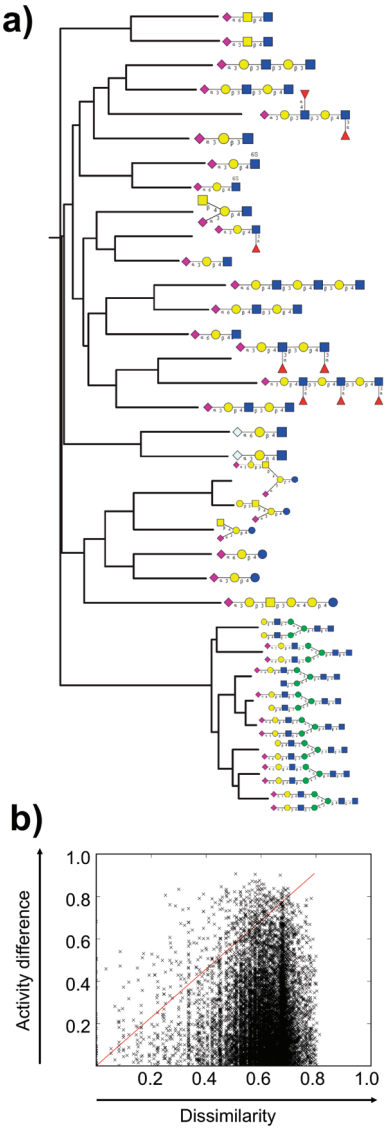


Figure 3. Validation of the similarity measure. (a) A set of 33 sialosides and one non-sialylated N-glycan was clustered on the basis of a pairwise calculated distance matrix. The clustering agrees well with intuitive similarity assessment. (b) Data from 30 lectins evaluated against 377 glycans was retrieved from the CFG database. Each data point represents a pairwise comparison for dissimilarity (1-SMT) of two glycans plotted against their normalized activity difference. The highest fluorescence intensity of each lectin was used for normalization of the individual data sets. High dissimilarity between a pair of glycans favors a higher activity difference. Hence, an optimal diversity descriptor has the highest density of data points in the lower right triangle (diagonal marked in red; for details see Methods in Supporting Information).

CFG mammalian glycan array (version 5.0) and three focused libraries targeting Influenza A viruses from the Wong group[22] and the Feizi group[23] and a custom sialoside library of ours[24,25] covering 155, 30, 70, and 56 sialosides, respectively. Their diversity index, given by the mean pairwise dissimilarity,[8] clearly identifies the CFG array to comprise the most diverse set (Figure 4a). The calculated values are in the expected range, taking into account that these sets of glycans are all sialosides, which by itself imposes restrictions not only to the terminal residues but also on the core structures. Inspection of the diversity of our own sialoside library highlights those classes of

dx.doi.org/10.1021/cb300003z | *ACS Chem. Biol.* 2012, 7, 829–834

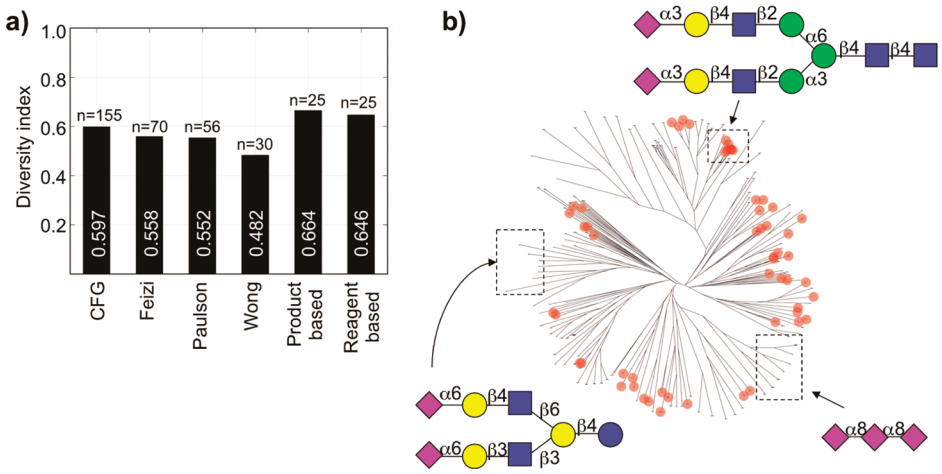


Figure 4. Glycan diversity of four current and two designed microarrays. (a) The diversity index of four current sialoside microarrays was calculated highlighting the CFG array (version 5.0) being most diverse. Two 25-membered custom arrays are proposed following either a product-based or a reagent-based design strategy. (b) A root-less dendrogram depicting all sialosides covered by the four libraries with structures from our own library marked in red. Three structures are shown explicitly as representatives of certain clusters.

glycans that would diversify the selection such as α2,8-linked sialic acids and branched glycosphingolipid structures (Figure 4b). It is important to recall that except for the CFG array, the other glycan arrays are designed for the specific purposes of Influenza A virus studies and thereby restrict the pool to those of highest biological relevance.

We next wanted to construct a small 25-member array with equal or greater diversity than the present formats without the constraint of a specific application such as Influenza A studies. The construction of a diverse library can follow two principles with either a product-based selection mechanism or a reagent-based rational underlying the decision of which compounds to include in the library.[9] We chose the structures present on the current arrays as the pool of all possible sialosides, as these structures were synthesized in the past in contrast to sialosides from the KEGG database. Moreover, the database structures may contain incomplete or false entries that may obscure the analysis. With this set we succeeded in defining an array of 25 glycans that has a diversity index of 0.664, higher than the parent libraries of larger size (Figure 4b, "product-based" and Supplementary Figure 5a). Thus, a more concise sialoside library with higher diversity in absence of a biological target can be constructed using a genetic algorithm for selection using the diversity coefficient as the scoring function. We also found that diversification did not result in a bias toward the selection of small glycans, emphasizing the impact of the modified Tanimoto coefficient (data not shown). However, a more realistic setting for a carbohydrate chemistry laboratory toward the construction of a diverse library follows a reagent-based approach. Therefore, according to the proposal of Werz *et al.*, we used their 20 building blocks sufficient to chemically synthesize 50% of the mammalian glycome. With this set of reagents it was feasible to chemically synthesize 72% of all sialosides from the pool. This excludes all structures with internal GalNAc such as LacDiNAc structures, α2,8-linked sialosides, structures with sulfate groups on the GlcNAc, and β-galactoses having attached glycans in positions other than in 2, 3, or 6. Under these limitations, an excellent diversity of 0.646 was achieved for a 25-member library (Supplementary Figure 5b). We believe that a diversity library like this will be helpful in the design of more focused libraries for biological targets to ensure that major structural classes are not inadvertently omitted.

Here we report a fast and versatile method that can easily handle very large data sets to calculate the similarity between a pair of glycans based on "glycan fingerprints". This method was developed in analogy to chemoinformatics approaches used to address the same problem in the field of combinatorial chemistry, where much time has been devoted solving many practical and theoretical problems over the last decades. We therefore based our model on existing and robust measures that now will enter the field of carbohydrate chemistry allowing many aspects of carbohydrate chemistry and biology to be approached in the future such as library design, activity prediction and database searches.

METHODS

Encoding of Glycan Structures and Data Handling. Glycan structures were stored in xml file format with a monosaccharide having the attributes monosaccharide identifier (*e.g.*, "Glc"), linkage (*e.g.*, "4"), configuration of the anomeric center (*e.g.*, "α"), and any modification (*e.g.*, "6OSO3").[26−28] The xml data structure was built using the ElementTree toolkit (version 1.3a).[29] Glycans from the KEGG database,[19] the CFG depository (array versions 3.1 and 5.0), and the sialoside arrays from Childs *et al.*,[23] Liao *et al.*,[22] and Nycholat *et al.*[24,25] were converted (for selection of the glycans see Supporting Information). Filters were applied to remove incomplete data, and the generated files were analyzed carefully to ensure high quality data sets. All programming was done in python 2.5.1 using numpy, Scipy, and matplotlib libraries. All computation was performed on a regular desktop computer.

Calculating the Glycan Fingerprints. Glycans were fragmented into mono- to heptasaccharides by systematic identification of the respective unique subtrees, with a subtree being a fragment of the glycan comprising all monosaccharides plus their connecting glycosidic bond information and modifications. The glycan fingerprints were stored in hashed and folded string: A bit was set to 1 using a pseudorandom number generator in a bit string for the presence of a unique fragment. The resulting bit string was then folded by logical OR operations of the two halves to increase the information content.[10]

Calculation of Glycan Similarity and the Diversity Index. The pairwise similarity of two glycans A and B encoded in a glycan

dx.doi.org/10.1021/cb300003z | *ACS Chem. Biol.* 2012, 7, 829−834

fingerprint of the length n was calculated using the modified Tanimoto coefficient S_{MT}:[18]

$$S_{MT}(A, B) = \left(\frac{2 - \rho_0}{3}\right)S_T(A, B) + \left(\frac{1 - \rho_0}{3}\right)S_{T0}(A, B)$$

with

$$S_{T0}(A, B) = \left(\frac{d}{n - c}\right)$$

and S_T(A,B) being the Tanimoto coefficient given by

$$S_T(A, B) = \left(\frac{c}{a + b - c}\right)$$

and ρ_0 being the average bit string occupancy over all structures in the entire data set, the number of bits set to 1 in glycan A being a and in glycan B being b, and c being the number of bits set in both fingerprints, while d represents the number of bits set in neither. The glycan dissimilarity matrix was calculated as dissimilariy = $1 - S_{MT}$. The matrix was then converted using a neighbor-joining algorithm,[30] and dendrograms were generated in DrawTree using default parameters for rooted and rootless trees,[31] respectively.

The diversity index D(GA) of a glycan array GA was calculated by the mean pairwise intermolecular dissimilarity:[8]

$$D(GA) = 1 - \frac{\sum_{A=1}^{N(GA)}\sum_{B=1}^{N(GA)} S_{MT}(A, B)}{N(GA)}$$

with S_{MT}(A,B) being the pairwise similarity of glycans A and B, and N(GA) being the total number of glycans.

Glycan Array Data. Lectins binding data to the printed glycan array (version 3.1) were retrieved from the CFG web server for 30 different lectins (see Supporting Information for full list). Inconclusive and redundant data was not included in the analysis. Moreover, for the purpose of descriptor evaluation, noisy data or data with very low number of hits were removed. Only data from active glycans that had more than one standard deviation of fluorescence above the average were used.

Product-Oriented and Reagent-Oriented Library Design. The union of all glycans present on the four sialoside arrays studied in this communication served as a pool of all potential products. A genetic algorithm was used to optimize the set of 25 sialosides that represent the most diverse subset. Briefly, 200 individuals were picked at random for the starting populations and evolution took place under a mutation rate of 0.25% and a one-point crossover rate of 1.0 for 500 generations using the Pyevolve package including a penalty for choosing identical members.[32] For the reagent-based design, following an automated chemical synthesis approach, the 20 building blocks proposed by Werz and Coworkers were implemented. The implementation did not take into account if protecting groups were orthogonal, and it was assumed the glycans were linked to a solid support at the reducing end. Only sialosides were then considered for the generation of a custom glycan array that were feasible in the boundaries of this framework. The same genetic algorithm was used as described above. All scripts are available upon request from the authors.

■ ASSOCIATED CONTENT

Supporting Information

This material is available free of charge *via* the Internet at http://pubs.acs.org.

■ AUTHOR INFORMATION

Corresponding Author

*E-mail: Christoph.Rademacher@mpikg.mpg.de.

Present Address

[†]Department of Biomolecular Systems, Max Planck Institute of Colloids and Interfaces, Am Mühlenberg 1, 14476 Potsdam, Germany.

Notes

The authors declare no competing financial interest.

■ ACKNOWLEDGMENTS

This work was supported in part by the National Institutes of Health grant AI058113 and CDC project 200-2009-32562. C.R. was supported by an EMBO long-term fellowship. The authors thank C. Nycholat, J. Landström, and A. Tran-Crie for critical review and help in preparation of the manuscript. We thank the reviewers for helpful comments to improve the manuscript.

■ REFERENCES

(1) Werz, D. B., Ranzinger, R., Herget, S., Adibekian, A., von der Lieth, C. W., and Seeberger, P. H. (2007) Exploring the structural diversity of mammalian carbohydrates ("glycospace") by statistical databank analysis. *ACS Chem. Biol. 2*, 685−691.

(2) Adibekian, A., Stallforth, P., Hecht, M.-L., Werz, D. B., Gagneux, P., and Seeberger, P. H. (2011) Comparative bioinformatics analysis of the mammalian and bacterial glycomes. *Chem. Sci. 2*, 337−344.

(3) Laine, R. A. (1994) Invited Commentary: A calculation of all possible oligosaccharide isomers both branched and linear yields 1.05 × 10^{12} structures for a reducing hexasaccharide: the Isomer Barrier to development of single-method saccharide sequencing or synthesis systems. *Glycobiology 4*, 759−767.

(4) Varki, A. (2009) *Essentials of Glycobiology*, 2nd ed., Cold Spring Harbor Laboratory Press, Cold Spring Harbor, NY.

(5) Rillahan, C. D., and Paulson, J. C. (2011) Glycan microarrays for decoding the glycome. *Annu. Rev. Biochem. 80*, 797−823.

(6) Martin, E. J., Blaney, J. M., Siani, M. A., Spellmeyer, D. C., Wong, A. K., and Moos, W. H. (1995) Measuring diversity: experimental design of combinatorial libraries for drug discovery. *J. Med. Chem. 38*, 1431−1436.

(7) Stumpfe, D., and Bajorath, J. (2011) Similarity searching. *Wiley Interdiscip. Rev.: Comput. Mol. Sci. 1*, 260−282.

(8) Turner, D. B., Tyrrell, S. M., and Willett, P. (1997) Rapid quantification of molecular diversity for selective database acquisition. *J. Chem. Inf. Comput. Sci. 37*, 18−22.

(9) Burke, M. D., and Schreiber, S. L. (2004) A planning strategy for diversity-oriented synthesis. *Angew. Chem., Int. Ed. 43*, 46−58.

(10) Randic, M., and Wilkins, C. L. (1979) Graph theoretical approach to recognition of structural similarity in molecules. *J. Chem. Inf. Comput. Sci. 19*, 31−37.

(11) Brown, R. D., and Martin, Y. C. (1997) The information content of 2D and 3D structural descriptors relevant to ligand-receptor binding. *J. Chem. Inf. Comput. Sci. 37*, 1−9.

(12) James, C. A., Weininger, D. (1994) *Daylight Theory Manual*, Daylight Chemical Information Systems, Inc, Irvine, CA.

(13) Gillet, V. J., Willett, P., and Bradshaw, J. (2003) Similarity searching using reduced graphs. *J. Chem. Inf. Comput. Sci. 43*, 338−345.

(14) Aoki, K. F., Mamitsuka, H., Akutsu, T., and Kanehisa, M. (2005) A score matrix to reveal the hidden links in glycans. *Bioinformatics 21*, 1457−1463.

(15) Kuboyama, T., Hirata, K., Aoki-Kinoshita, K. F., Kashima, H., and Yasuda, H. (2006) A gram distribution kernel applied to glycan classification and motif extraction. *Genome Inform. 17*, 25−34.

(16) Ranzinger, R., Frank, M., von der Lieth, C. W., and Herget, S. (2009) Glycome-DB.org: a portal for querying across the digital world of carbohydrate sequences. *Glycobiology 19*, 1563−1567.

(17) Holliday, J. D., Salim, N., Whittle, M., and Willett, P. (2003) Analysis and display of the size dependence of chemical similarity coefficients. *J. Chem. Inf. Comput. Sci. 43*, 819−828.

(18) Fligner, M. A., Verducci, J. S., and Blower, P. (2002) A modification of the Jaccard−Tanimoto similarity index for diverse

selection of chemical compounds using binary strings. *Technometrics 44*, 110−119.
(19) Kanehisa, M., Goto, S., Kawashima, S., Okuno, Y., and Hattori, M. (2004) The KEGG resource for deciphering the genome. *Nucleic Acids Res. 32*, D277−280.
(20) Varki, A. (2010) Colloquium paper: uniquely human evolution of sialic acid genetics and biology. *Proc. Natl. Acad. Sci. U.S.A. 107* (Suppl 2), 8939−8946.
(21) Patterson, D. E., Cramer, R. D., Ferguson, A. M., Clark, R. D., and Weinberger, L. E. (1996) Neighborhood behavior: a useful concept for validation of "molecular diversity" descriptors. *J. Med. Chem. 39*, 3049−3059.
(22) Liao, H. Y., Hsu, C. H., Wang, S. C., Liang, C. H., Yen, H. Y., Su, C. Y., Chen, C. H., Jan, J. T., Ren, C. T., Cheng, T. J., Wu, C. Y., and Wong, C. H. (2010) Differential receptor binding affinities of influenza hemagglutinins on glycan arrays. *J. Am. Chem. Soc. 132*, 14849−14856.
(23) Childs, R. A., Palma, A. S., Wharton, S., Matrosovich, T., Liu, Y., Chai, W., Campanero-Rhodes, M. A., Zhang, Y., Eickmann, M., Kiso, M., Hay, A., Matrosovich, M., and Feizi, T. (2009) Receptor-binding specificity of pandemic influenza A (H1N1) 2009 virus determined by carbohydrate microarray. *Nat. Biotechnol. 27*, 797−799.
(24) Nycholat, C. M., McBride, R., Ekiert, D. C., Xu, R., Rangarajan, J., Peng, W., Razi, N., Gilbert, M., Wakarchuk, W., Wilson, I. A., and Paulson, J. C. (2012) Recognition of sialylated poly-LacNAc on N- and O-linked glycans by human and avian influenza A virus hemagglutinins. *Angewandte Chemie*, DOI: 10.1002/anie.201200596 and 10.1002/ange.201200596.
(25) Xu, R., McBride, R., Nycholat, C. M., Paulson, J. C., and Wilson, I. A. (2012) Structural characterization of the hemagglutinin receptor specificity from the 2009 H1N1 influenza pandemic. *J. Virol. 86*, 982−990.
(26) Herget, S., Ranzinger, R., Maass, K., and Lieth, C. W. (2008) GlycoCT-a unifying sequence format for carbohydrates. *Carbohydr. Res. 343*, 2162−2171.
(27) Sahoo, S. S., Thomas, C., Sheth, A., Henson, C., and York, W. S. (2005) GLYDE-an expressive XML standard for the representation of glycan structure. *Carbohydr. Res. 340*, 2802−2807.
(28) Kikuchi, N., Kameyama, A., Nakaya, S., Ito, H., Sato, T., Shikanai, T., Takahashi, Y., and Narimatsu, H. (2005) The carbohydrate sequence markup language (CabosML): an XML description of carbohydrate structures. *Bioinformatics 21*, 1717−1718.
(29) Lundh, F. (2007) The ElementTree toolkit; http://effbot.org/.
(30) Gascuel, O. (1997) BIONJ: an improved version of the NJ algorithm based on a simple model of sequence data. *Mol. Biol. Evol. 14*, 685−695.
(31) Felsenstein, J. (1993) PHYLIP (Phylogeny Inference Package) version 3.5c. Distributed by the author, Department of Genetics, University of Washington, Seattle.
(32) Perone, C. S. (2009) Pyevolve: a Python open-source framework for genetic algorithms. *ACM SIGEVOlution 4*, 12−20.