Available online at www.sciencedirect.com
SCIENCE DIRECT®
Carbohydrate Research 340 (2005) 1059–1064
Carbohydrate RESEARCH

The use of a genetic algorithm search for molecular mechanics (MM3)-based conformational analysis of oligosaccharides

Abraham Nahmany,[a,†] Francesco Strino,[a,†] Jimmy Rosen,[a] Graham J. L. Kemp[b] and Per-Georg Nyholm[a,*]

[a]*Department of Medical Biochemistry, Göteborg University, Medicinaregatan 7B, Box 440, SE-405 30 Göteborg, Sweden*
[b]*Department of Computing Science, Chalmers University of Technology, SE-412 96 Göteborg, Sweden*

Received 24 September 2004; accepted 27 December 2004

Dedicated to Professor David A. Brant

Abstract—We have implemented a system called GLYGAL that can perform conformational searches on oligosaccharides using several different genetic algorithm (GA) search methods. The searches are performed in the torsion angle conformational space, considering both the primary glycosidic linkages as well as the pendant groups (C-5–C-6 and hydroxyl groups) where energy calculations are performed using the MM3(96) force field. The system includes a graphical user interface for setting calculation parameters and incorporates a 3D molecular viewer. The system was tested using dozens of structures and we present two case studies for two previously investigated O-specific oligosaccharides of the *Shigella dysenteriae* type 2 and 4. The results obtained using GLYGAL show a significant reduction in the number of structures that need to be sampled in order to find the best conformation, as compared to filtered systematic search.

Keywords: Genetic algorithms; Molecular mechanics; Oligosaccharide; O-antigen; *Shigella dysenteriae*

1. Introduction

Saccharides linked to proteins and lipids cover a large fraction of the surface area of most cells. Many of these saccharides are involved in specific recognition processes. To understand their biological function in detail it is necessary to have information about their three dimensional (3D) structure. Knowledge about the 3D structure of oligosaccharides also has medical applications, for example, in the design of vaccines targeted at surface saccharides of bacteria.[1–4]

The two main experimental techniques for determining the 3D structures of molecules—NMR spectroscopy and X-ray crystallography—are often difficult to apply to oligosaccharides. X-Ray is problematic since it is difficult to obtain crystals of suitable quality. NMR determination is difficult due to the paucity of well defined NOEs. An alternative approach for oligosaccharide conformational analysis is to search the space of possible conformations using computational methods to find favorable low energy conformations.

The O-specific oligosaccharide 3D structure of *Shigella dysenteriae* type 1 has been the subject of experimental as well as computational studies.[3,5] An earlier study by Rosen et al. used filtered systematic search to predict favorable conformations of the O-specific oligosaccharide of *S. dysenteriae* type 2 and *S. dysenteriae* type 4.[6,7]

A major problem with computational conformational analysis of oligosaccharides is achieving a good trade-off between the sampling of the conformational space and the required computation time.

Here we present the results obtained using the newly developed GLYGAL program coupled to MM3,[8]

* Corresponding author. Tel.: +46 (0)31 773 34 54; fax: +46 (0)31 823758; e-mail: nyholm@medkem.gu.se
† Contributed equally to this paper.

0008-6215/$ - see front matter © 2005 Elsevier Ltd. All rights reserved.
doi:10.1016/j.carres.2004.12.037

concerning *S. dysenteriae* type 2 and *S. dysenteriae* type 4 and a comparison with the results obtained by filtered systematic search.

2. Methods

The basic ideas for predicting oligosaccharide conformations using a standard genetic algorithm[9–11] are:

1. Initial population of randomly generated conformations ('individuals' or 'chromosomes') with respect to the torsion angles of the glycosidic linkages and pendant groups ('genes'). Each such 'chromosome' is represented as a vector of real-numbers for the torsion angles. The torsions of the rings are not randomized.
2. Evaluation using molecular mechanics MM3 as fitness function. Selection is carried out using the roulette wheel method, where individuals are evaluated as a function of their conformational energy.
3. Standard genetic operators like mutation and crossover are used to generate offspring.
4. Termination criteria are satisfied either after a fixed number of generations or when no improvement has occurred during several generations.

The major goal was to turn these ideas into easy-to-use software for oligosaccharide conformational search. The software which was developed was called GLYGAL. The program implements three different GAs for the purpose of oligosaccharide conformational search: standard GA, parallel GA, and an evolutionary programming algorithm.[9] All the GAs implemented can be used with local minimization and propagation of the minimized geometry to the progeny, namely Lamarckian GA.[9,12] Some default GA parameters, such as population size, number of iterations, etc. are suggested to the user and those can easily be set using the GLYGAL graphical user interface.

GLYGAL also requires a template file containing the coordinates. Pdb, xyz and MM3 files are currently supported. The following steps describe the course of events of the conformational search:

1. The torsion angles to be modified in the search are assigned automatically.
2. The template structure file is copied as many times as the size of the initial population of the GA. The copied structures are modified by torsion-angle adjustments to create the first randomly generated population of the GA.
3. The files are then sent to the MM3 program for evaluation and local minimization, and the results (i.e., energies and geometries) are sent back to GLYGAL.
4. The structures to be manipulated by the genetic algorithm operators are selected randomly using a roulette wheel method.
5. Genetic algorithm operators, such as mutation and crossover are performed on the structures and the next generation of structures is created. In the case of parallel GA a migration operator is also involved.
6. The structures are evaluated and termination criteria are checked. If not fulfilled, the process will resume from step 3.

As mentioned, GLYGAL uses the MM3 (96)[8,13] forcefield program for the energy evaluation and local minimization. The MM3 calculations are distributed on a Linux cluster (Csol Hoborg) of five nodes, each with dual 2200+ AMD processors.

One of the main problems with the existing methods for oligosaccharide modeling is the need of manual preprocessing, that is, to identify the torsion angles to be searched. In GLYGAL we solved this problem by developing an algorithm for identifying the torsion angles automatically. This algorithm creates a connection matrix containing the torsion angles for each glycosidic linkage and, on the diagonal, a ring vector containing all the torsion angles needed to identify the position of all atoms within a certain ring.

The pendant groups, that is, the C-5–C-6 and the hydroxyl groups, respectively, are included in the GA search just as the primary torsions of the glycosidic linkages. Since these minor torsions have less impact on the energy of the structure their sampling can be weighted accordingly, to achieve a thorough but efficient sampling of the conformational space. The systematic search[3,6,7] uses the simplification that the hydroxyls align in chains of weak hydrogen bonds[14,15] to reduce the number of dimensions to sample. This is not necessary when using GA search, as the algorithmic complexity scales well with higher dimensionality. It is thus possible to sample the hydroxyls individually within our computational capacity, which is to be preferred.[16]

The pendant groups add significantly to the number of sample dimensions. However, since the GA search scales well to increasing dimensionality it does not introduce too much of a computational burden. To further reduce the computational load and increase the sampling efficiency, the search space can be divided into layers and be successively refined as the search progresses. Another method, available in GLYGAL, is to attach different sampling 'weights' to the search dimensions, reflecting their impact on the conformational energy. Each individual (or 'chromosome') is represented as a vector of torsion angles of the glycosidic linkages and pendant groups ('genes'). For example, a vector representing a trisaccharide will contain five genes: one gene for each glycosidic linkage and one gene for the pendant groups of each residue. The weight values indicate the probability of the

gene being selected for genetic operators, namely, mutation or crossover. The values are proportional to the number of torsion angles within each gene. Higher values are assigned to genes representing glycosidic linkages and to genes involved in structural branches. The user interface of GLYGAL allows the user to modify the weight values, for example, the glycosidic linkages can be excluded from the search in order to perform quick and fine minimizations of the pendant groups without changing the starting conformation of the glycosidic linkages.

3. Results and discussion

We have performed tests on several dozens of different oligosaccharides and oligosaccharide fragments. We will present here the results obtained by GLYGAL on two O-specific oligosaccharides, *S. dysenteriae* type 2 and *S. dysenteriae* type 4, which have been previously studied with respect their 3D structures.[6,7] Both oligosaccharides contain repeating units of five residues. In total we investigated 20 different fragments of these oligosaccharides, that is, disaccharide, trisaccharide, tetrasaccharide, and the complete repeating units.

3.1. *Shigella dysenteriae* type 2

Rosen et al. investigated the O-specific oligosaccharide of *S. dysenteriae* type 2.[6] They started by investigating one repeating unit of the polysaccharide. The sequence of the repeating unit is illustrated in Figure 1. This pentasaccharide was first divided into fragments of disaccharides. A systematic search on the ϕ/Ψ space with 15° step size using the MM3 force field was performed on each one of the disaccharides. The results were energy maps providing a comprehensive overview of the ϕ/Ψ search space of each of the disaccharides. The energy maps obtained for the disaccharides are illustrated in the lower row of Figure 1. The energy maps were then used as filters for the systematic search performed on the trisaccharide of the branching point. The potential energy maps, which illustrate the trisaccharide at the branching point, are shown in the upper row of Figure 1. They could then construct the whole structure manually using the minimized fragments to get the favorable conformation. Using the filtered systematic search they had to sample approximately 350,000 structures of various fragments for this prediction.

The GLYGAL program was used to model on the structure of *S. dysenteriae* type 2 and 4 with a search scheme similar to the one suggested by Rosen et al.[6] We first divided the repeating unit oligosaccharide into fragments of disaccharides and used GLYGAL to search for favorable conformations. For each disaccharide we used a standard GA with 10 structures in the initial population. Running 10 generations on average we could find all minima for each of the disaccharide fragments found using the filtered systematic search. The dots in the A

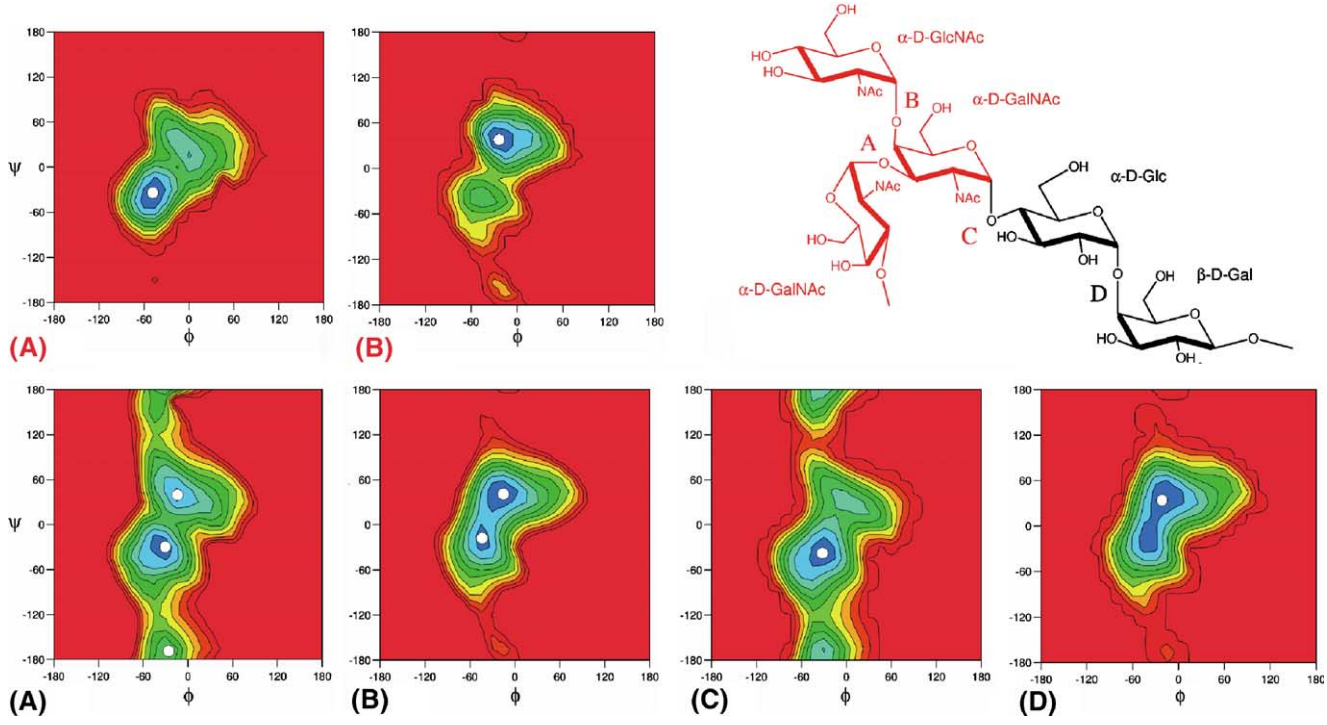

Figure 1. MM3 adiabatic energy maps generated by filtered systematic search for the different disaccharide moieties (lower row) as well as for the trisaccharide at the branching point of the repeating unit of the *Shigella dysenteriae* type 2 O-antigen.[6] Contour levels are shown at every kcal/mol with blue for low-energy regions and red for high-energy levels. The dots indicate the minima found using the GA search.

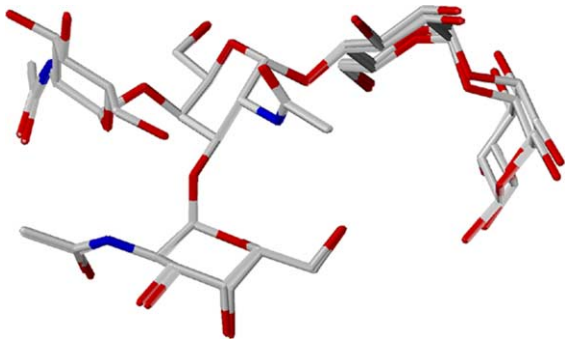

Figure 2. Superimposition of the MM3 minimum energy conformations of the *Shigella dysenteriae* type 2 repeating unit obtained by filtered systematic search and GLYGAL.

and B energy maps in the lower row of Figure 1 mark the minima found by the GA. This search was performed on the cluster in a couple of minutes.

The trisaccharide and the tetrasaccharide at the branching point were subsequently subjected to GLYGAL calculations yielding excellent agreement with the systematic search. In both cases a parallel GA was used. For the trisaccharide we used five populations each with 60 initial structures. After approximately 10 generations, that is, after approximately 3000 structures, the search converged to the low-energy minima found in the systematic search. The dots in the A and B energy maps in upper row and in the C and D maps of Figure 1 mark the minima found by the GA.

The complete pentasaccharide repeating unit was not tested using the filtered systematic search due to the huge CPU time needed for this kind of search. Using GLYGAL, on the other hand, we could conduct a search on the complete repeating unit and the result showed a very good fit to the structure obtained by Rosen et al. (see Fig. 2). Using the GLYGAL program we needed to sample approximately 8000 conformations.

3.2. *Shigella dysenteriae* type 4

The O-specific oligosaccharide of the *S. dysenteriae* type 4 was investigated in a similar way to the O-specific oligosaccharide of the *S. dysenteriae* type 2. Here again we used the filtered systematic search[7] for comparison. Figure 3 illustrates MM3 adiabatic energy maps for the repeating unit of *S. dysenteriae* type 4 generated by the filtered systematic search. The dots represent the minima found using the GLYGAL program. We observe that GLYGAL detected the minima for the linkage α-D-GlcNAc-(1→3)-α-D-GlcNAc shifting from one favorable minimum when investigating the disaccharide (energy map A in lower row) to another minimum when investigating the trisaccharide (energy map A in upper row) of the branching point. The minima found when investigating the trisaccharide are the favorable minima even when investigating the complete repeating unit using GLYGAL. We observe also that GLYGAL detected the change in minima for the linkage α-L-Fuc-(1→4)-α-D-

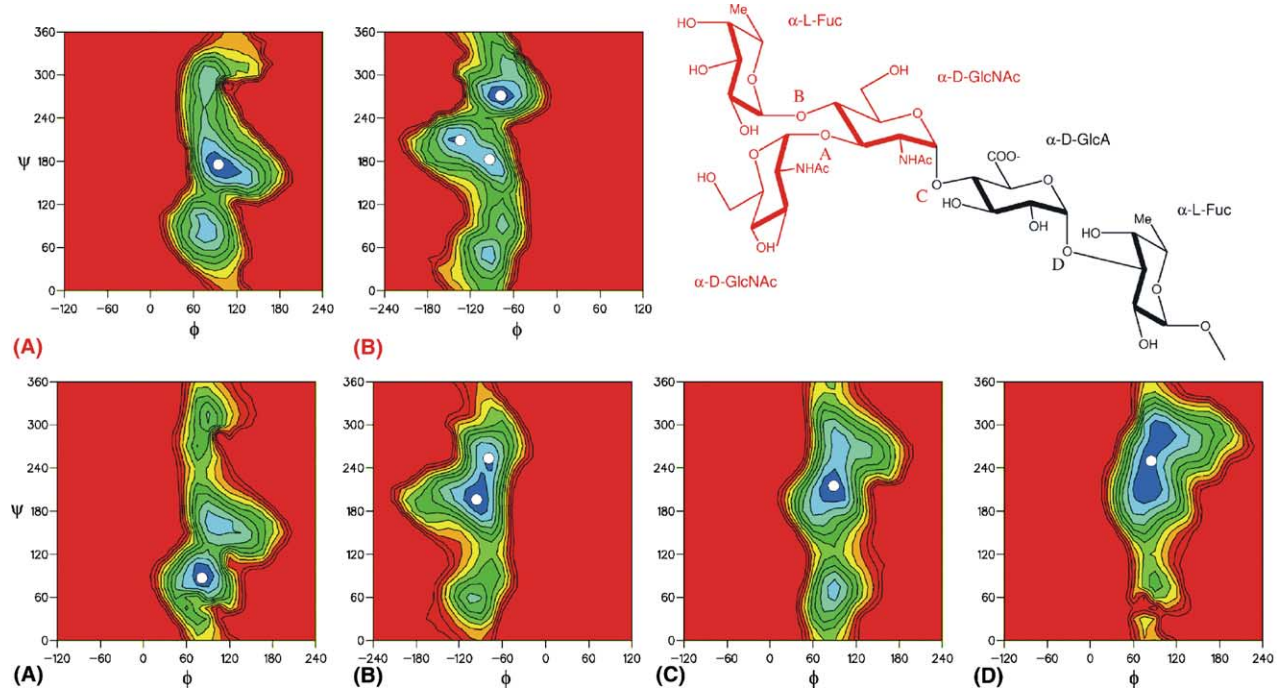

Figure 3. MM3 adiabatic energy maps generated by filtered systematic search for the different disaccharide moieties (lower row) as well as for the trisaccharide at the branching point of the repeating unit of the *Shigella dysenteriae* type 4 O-antigen.[7] The dots mark the minima found using the GLYGAL program.

GlcNAc (energy map B in lower and upper rows). Investigating the trisaccharide at the branching point, all the distinct minima were found using GLYGAL.

3.3. Pendant groups

In a systematic conformational search on saccharides, the C-5–C-6 torsions are usually sampled in their two favored conformations following from the *syn*-axial interaction.[15] Furthermore, the torsions of the secondary hydroxyl groups are often sampled using a simplification in which the hydroxyls align in chains of weak hydrogen bonds, giving rise to a clockwise and anticlockwise orientation.[15] The C–N single bond at the attachment of *N*-acetyl groups to the rings is often not sampled, although it is apparent that this torsion is fairly soft. Using GLYGAL, all the pendant groups were sampled, as described above, as part of the conformational search space. The hydroxyl groups converged to *C* or *R* conformations in most cases and seemed to adapt to the conformations of the glycosidic linkages. However, the *C* and *R* conformations only correspond to small energy differences, less than 1 kcal/mol at a dielectric constant of 80.

3.4. Program performance

In systematic search, one would have to cover the search space in a more thorough way having to pay a high price in the run time of the search. In contrast, one can ignore certain terms of the oligosaccharide energy in the force field, that is, by using a truncated force-field as in Glycan,[3] to achieve a faster result but with less accuracy. In Table 1 we summarize the performance of GLYGAL compared to the filtered systematic search.

Apart from the performance of GLYGAL in comparison to systematic search, we were interested in investigating the behavior of the different GAs implemented within GLYGAL, that is, standard GA, parallel GA, and evolutionary programming. Those were tested on different oligosaccharide structures having two to seven residues, linear or branched. The performance was measured by the number of structures needed to be sampled in order to find the best conformation.

We observed the following:

1. Almost no difference between the different GAs when investigating disaccharides. On average we need to sample 100 structures.
2. For larger structures, linear or branched, the parallel GA is preferable. We need to sample fewer structures using a parallel GA. One possible explanation is that a division into different populations that propagate in parallel, with individuals migrating between the populations, increases the diversity in the total population, thus resulting in faster convergence.

Table 1. Number of structures needed to be sampled using the GLYGAL program versus filtered systematic search for finding the lowest energy minima of oligosaccharide structures of different length

Search methods	Saccharide size	Sampled structures
Filtered systematic search	Disaccharide	576
	Trisaccharide	300,000
	Tetrasaccharide	Too many
	Pentasaccharide	Too many
Genetic algorithms (GLYGAL)	Disaccharide	100
	Trisaccharide	2500
	Tetrasaccharide	6000
	Pentasaccharide	10,000

The conformational searches used to generate the numbers in the table did not include a full optimization of the secondary hydroxyls.

3. Branched structures, especially those with vicinal branches, potentially have a more complex energy hypersurface than linear saccharides with the same number of residues. Therefore there is a need for more thorough sampling in the case of branched structures. However, it is our experience that the vicinally branched structures generally have a restricted flexibility with well defined favored conformations. Therefore the predictions on vicinally branched structures can give better insight into the structure–function relationship of the saccharide.
4. For saccharides up to a pentasaccharide in size, we prefer a fairly low number of generations and instead larger populations, as previously observed by others.[11] Thus to optimize the ϕ/Ψ of a pentasaccharide

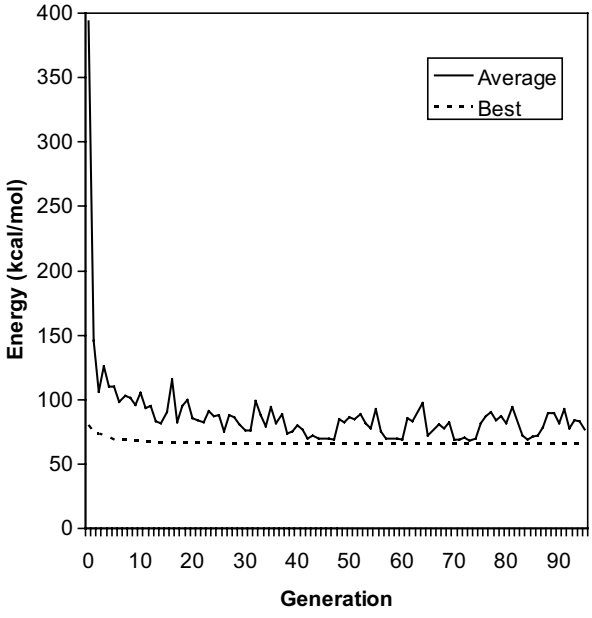

Figure 4. GA-MM3 optimization of the pentasaccharide of *Shigella dysenteriae* type 2. The average energy and the energy of the best structure are shown for each generation.

20 generations were successfully used with 4 populations of 60 structures each (Fig. 4). The optimization of the pendant groups required more generations. The details regarding the pendant groups will be described elsewhere. For structures larger than a pentasaccharide, future work is required to determine the optimal parameters with respect to number of generations and population size.

4. Conclusions

Using GLYGAL we were able to investigate different oligosaccharide structures showing that genetic algorithm search methods, though based on heuristics, are a very good tool for the task in hand. The major advantage of GA is that it scales better compared to systematic search. Thus for a trisaccharide the GA is about 100 times faster than the filtered systematic search, and for larger structures the difference is even more significant. This advantage allows the investigation of large saccharide structures and also the addition of pendant groups to the search.

To achieve a faster convergence to the minima it is possible to use a dynamic assignment of the weights for the operators. Thus it would be possible to automatically switch the focus from the ϕ/Ψ torsion angles to the pendant groups during the course of the calculations. Work to implement such schemes in the GA is currently in progress. The GLYGAL method was recently successfully applied to predict the structure of the repeating unit of the exopolysaccharide of *Burkholderia cepacia*.[17]

References

1. Kotloff, K. L.; Winickoff, J. P.; Ivanoff, B.; Clemens, J. D.; Swerdlow, D. L.; Sansonetti, P. J.; Adak, G. K.; Levine, M. M. *Bull. World Health Organ.* **1999**, *77*, 651–666.
2. Pozsgay, V.; Chu, C.; Pannell, L.; Wolfe, J.; Robbins, J. B.; Schneerson, R. *Proc. Natl. Acad. Sci. U.S.A.* **1999**, *96*, 5194–5197.
3. Nyholm, P. G.; Mulard, L. A.; Miller, C. E.; Lew, T.; Olin, R.; Glaudemans, C. P. *Glycobiology* **2001**, *11*, 945–955.
4. Clement, M. J.; Imberty, A.; Phalipon, A.; Perez, S.; Simenel, C.; Mulard, L. A.; Delepierre, M. *J. Biol. Chem.* **2003**, *278*, 47928–47936.
5. Coxon, B.; Sari, N.; Batta, G.; Pozsgay, V. *Carbohydr. Res.* **2000**, *324*, 53–65.
6. Rosen, J.; Robobi, A.; Nyholm, P. G. *Carbohydr. Res.* **2002**, *337*, 1633–1640.
7. Rosen, J.; Robobi, A.; Nyholm, P. G. *Carbohydr. Res.* **2004**, *339*, 961–966.
8. Allinger, N. L.; Yuh, Y. H.; Lii, J. H. *J. Am. Chem. Soc.* **1989**, *111*, 8551–8566.
9. Mitchell, T. *Machine Learning*; McGraw-Hill, 1997; pp 259–283.
10. Davis, L. *Handbook of Genetic Algorithms*; Van Nostrand Reinhold: New York, 1991.
11. Judson, R. S.; Jaeger, E. P.; Treasurywala, A. M.; Peterson, M. L. *J. Comput. Chem.* **1993**, *14*, 1407–1414.
12. Lamarck, J. B. *Philosophie Zoologique* **1809**.
13. Allinger, N. L.; Rahman, M.; Lii, J. H. *J. Am. Chem. Soc.* **1990**, *112*, 8293–8307.
14. Ha, S. N.; Madsen, L. J.; Brady, J. W. *Biopolymers* **1988**, *27*, 1927–1952.
15. French, A. D.; Tran, V. H.; Perez, S. In *Computer Modelling of Carbohydrate Molecules*; ACS Symposium Series 430; American Chemical Society: Washington, DC, 1990; pp 191–212.
16. Stortz, C. A. *Carbohydr. Res.* **1999**, *322*, 77–86.
17. Strino, F.; Nahmany, A.; Rosen, J.; Kemp, G. J. L.; Sacorreia, I.; Nyholm, P.-G. *Carbohydr. Res.* **2005**, *340*, see doi:10.1016/j.carres.2004.12.031.