Data and text mining

Chemical shift-based identification of monosaccharide spin-systems with NMR spectroscopy to complement untargeted glycomics

Piotr Klukowski1,* and Mario Schubert2,*

1Department of Computer Science, Faculty of Computer Science and Management, Wrocław University of Science and Technology, 50-370 Wrocław, Poland and 2Department of Biosciences, University of Salzburg, 5020 Salzburg, Austria

*To whom correspondence should be addressed.

Received on February 23, 2018; revised on May 25, 2018; editorial decision on June 6, 2018; accepted on June 10, 2018

Abstract

Motivation: A better understanding of oligosaccharides and their wide-ranging functions in almost every aspect of biology and medicine promises to uncover hidden layers of biology and will support the development of better therapies. Elucidating the chemical structure of an unknown oligosaccharide remains a challenge. Efficient tools are required for non-targeted glycomics. Chemical shifts are a rich source of information about the topology and configuration of biomolecules, whose potential is however not fully explored for oligosaccharides. We hypothesize that the chemical shifts of each monosaccharide are unique for each saccharide type with a certain linkage pattern, so that correlated data measured by NMR spectroscopy can be used to identify the chemical nature of a carbohydrate.

Results: We present here an efficient search algorithm, GlycoNMRSearch, which matches either a subset or the entire set of chemical shifts of an unidentified monosaccharide spin system to all spin systems in an NMR database. The search output is much more precise than earlier search functions and highly similar matches suggest the chemical structure of the spin system within the oligosaccharide. Thus, searching for connected chemical shift correlations within all electronically available NMR data of oligosaccharides is a very efficient way of identifying the chemical structure of unknown oligosaccharides. With an improved database in the future, GlycoNMRSearch will be even more efficient deducing chemical structures of oligosaccharides and there is a high chance that it becomes an indispensable technique for glycomics.

Availability and implementation: The search algorithm presented here, together with a graphical user interface, is available at http://glyconmrssearch.nmrhub.eu.

Contact: piotr.klukowski@pwr.edu.pl or mario.schubert@sbg.ac.at

Supplementary information: Supplementary data are available at Bioinformatics online.

1 Introduction

Carbohydrates are essential for life. Beside their function as energy storage molecules and their nutritional value, glycans—well-defined oligosaccharides attached to proteins—represent the most abundant posttranslational modification of proteins (Khoury et al., 2011) with crucial functions ranging from protein folding to cell–cell recognition and the distinction between self and non-self (Varki and Gagneux, 2017). Smaller oligosaccharides attached to lipids—glycolipids—also play a crucial part in cell–cell interactions and immune responses (Cummings, 2009). In contrast to nucleic acids and
proteins, which are linear chains consisting of four or 20 building blocks, respectively, glycans can be linked in different ways. This variation is caused by different linkage positions, the two possible configurations at the anomic carbon and a large amount of known building blocks (Seeberger, 2017). Glycans can also be branched. The different ways of linking monosaccharides result in an enormous variety of glycans. This is exploited by nature, which uses a large amount of different glycopolymers to decorate proteins and cell surfaces, of which we just start to understand their function. For example, glycans are crucial for distinguishing species from one another and also play an important role in fertilization (Pang et al., 2011). There seems to be a code—the glycome (Pilbello and Mahal, 2007) or sugar code (Sols et al., 2015).

Essential for understanding the glycome is the ability to determine the exact composition of certain glycans, or parts of glycans—glycoepitopes, before their function can be studied on a cellular level. Due to the abundance of protein glycosylation, glycoproteomics developed as a branch of proteomics, which is mainly driven by mass spectrometry (MS) techniques. Traditionally a very tedious protocol is followed in which glycans are cleaved from proteins and analyzed by a battery of techniques including digestion of glycans into monosaccharides, analysis of monosaccharide composition, and certain neighbors. A complete match of all chemical shifts between two datasets strongly indicates identical saccharide types and similar substitutions.

2 Results
2.1 Algorithm and web interface
To search for any given chemical shifts and correlations thereof within a chemical shift database, we developed a search function to validate by a second independent method.

NMR spectroscopy is used since decades to analyze oligosaccharides, especially by 2D NMR spectroscopy (Duus et al., 2000), which is very powerful for the determination of the exact chemical structure. In addition to NMR and MS a monosaccharide analysis consisting of a digestion and HPLC separations and chemical modifications of certain functional groups were traditionally used. Combinations of these techniques, especially combining NMR and MS are very powerful, e.g. for the characterization of glycoconjugate vaccines (Yu et al., 2018). Pioneering efforts extracted typical chemical shifts, called structural reporter groups, from extensive collections of 1D NMR spectra and chemical shift tables, which helped to identify the chemical structure of a carbohydrate (Vliegenthart and Kamerling, 2007; Vliegenthart et al., 1980). However, considering the enormous amount of NMR data the use of tables became rather impractical. Early databases and computer algorithms were developed, e.g. SugaBase (van Kuik et al., 1992) with simple search functions, which was later implemented in SWEET-DB (Loss et al., 2002) and Glycosciences.de (Lütteke et al., 2006). First combinations of chemical shifts included the elegant analysis of cross-peak patterns in 2D 1H–1H TOCSY spectra to distinguish hexapyranose spin systems (Gheysen et al., 2008). However, despite todays strong contribution of NMR spectroscopy to many aspects of glycomics (Kato and Peters, 2017; Marchetti et al., 2016), the methodology to identify unknown saccharides is cumbersome and did not change much in the last decades.

We hypothesize that there is a short-cut to obtain the chemical structure of an oligosaccharide just from experimental chemical shifts. Completely assigned chemical shifts of a spin-system (e.g. 13C values in the case of hexose consisting of 6 carbon and 7 proton chemical shifts) contain an enormous amount of information. Since these shifts reflect the local environment of each nucleus, they are influenced by the saccharide type, the configuration (e.g. pyranose versus furanose, α and β), the conformation/pucker, the substitution patterns, effects of neighboring saccharides and the electronic configuration. The combination of all 1H and 13C chemical shifts is typically so distinct that each monosaccharide with a certain linkage within an oligosaccharide gives a unique set of chemical shifts. For a hexose the chemical shifts of all carbons C1 to C6 and protons H1 to H6 could be considered as point in a 13-dimensional space that is distinct for the different hexoses with a certain type of substitution and certain neighbors. A complete match of all chemical shifts between two datasets strongly indicates identical saccharide types and similar substitutions.

This uniqueness of a completely chemical shift-assigned spin system could be exploited for explorative glycomics, to match measured NMR data of an unknown oligosaccharide against data of known glycans. However, several roadblocks prevented such a strategy: none of the existing NMR databases combine chemical shift values to spin systems and even more important a search function to match chemical shift correlations of spin systems did not exist. Not even pairs of 1H–13C chemical shifts could be searched. The only search functions with chemical shifts as input available so far, are based on providing a list of either 1H or 13C chemical shifts that are matched to any of the chemical shifts of an oligosaccharide entry (Kapaev and Toukach, 2018; Loss and Lütteke, 2015), no matter whether they are from the same spin-system or not. This makes searches within the current NMR databases of oligosaccharides very inefficient because they produce large output lists, which mainly consist of false-positive hits. Predictions of chemical shifts for a given oligosaccharide structure have been developed, for example CASPER (Kapaev and Toukach, 2015; Lundborg and Widmalm, 2011). There is even a tool to simulate 2D NMR spectra (Kapaev and Toukach, 2016). However for using these tools, one or several oligosaccharide structures are required as input, which makes it not suitable to study unknown oligosaccharides.

Here, we present an efficient search algorithm that matches correlations of any chemical shifts within an unknown spin-system against all spin systems in an NMR database. From given peak coordinates (in ppm units), the search procedure finds the best fitting monosaccharide entries, as presented in Figure 1. The algorithm is made publicly available as a web application. Extensive tests revealed that it can be used to predict the monosaccharide type, linkage type and often the local chemical environment beyond the next neighbor based on the chemical shifts of an experimental, initially unknown spin system.
Typically the anomeric chemical shifts $C_1$ and $H_1$ are well visible in 2D NMR spectra and expected correlations indicated as arrows in a hexose spin system. For reasons of clarity some correlations are omitted. At the bottom typical correlations are further illustrated by the pyranose form of $\beta$-d-glucose. (b) Rationale of the approach that is described in the paper. The goal is to match unassigned or assigned spin systems by the search algorithm to data of an NMR database to obtain a list of best matching entries, for which links to the chemical composition of the oligosaccharides are provided.

Fig. 1. Schematic illustration of assigning carbohydrate spin systems by 2D NMR spectra and their use to search an NMR database for similar structures. (a) Typically applied 2D NMR spectra and expected correlations indicated as arrows in a hexose spin system. For reasons of clarity some correlations are omitted. At the bottom typical correlations are further illustrated by the pyranose form of $\beta$-d-glucose. (b) Rationale of the approach that is described in the paper. The goal is to match unassigned or assigned spin systems by the search algorithm to data of an NMR database to obtain a list of best matching entries, for which links to the chemical composition of the oligosaccharides are provided.

The algorithm was tested using chemical shift data of several tri- and tetrascarbohydrates that were assigned by us previously and whose chemical shifts are not contained in the glycosciences.de database (Lütteke et al., 2006), namely $\alpha$-1,3-fucosylated LaciNAc (LDNF), lactosamine and the amphibian trisaccharide Bv9 (Aeschbacher et al., 2017; Zierke et al., 2013). A thorough spin system analysis with NMR spectroscopy associated each observed chemical shift exactly with a certain proton or carbon within the spin system. Due to the distinct $^{13}C$ chemical shifts of anomeric carbons, the anomic $C_1$ and the corresponding $H_1$ can be easily assigned. $^1H$–$^1H$ COSY spectra reveal $H_2$ and the corresponding $C_2$ using a $^1H$–$^{13}C$ HSQC spectrum.

The first search variant applied the complete set of $^1H$ and $^{13}C$ chemical shifts of a spin system with assignments to individual proton and carbon positions within the spin system (Supplementary Fig. 3a). A second search used the same $^1H$ and $^{13}C$ chemical shifts as $^1H$–$^{13}C$ pairs unassigned to a position (Supplementary Fig. 3b). The third and fourth search strategy uses a smaller set of resonances, because the complete chemical shift assignment of an entire spin system might sometimes not be available or might not be required. Typically the anomeric chemical shifts $C_1$ and $H_1$ are well visible in a natural abundance $^{13}C$-HSQC. By using in addition a 2D COSY, a 2D HMHC–COSY or other similar experiments the neighboring resonances of $C_2$–$H_2$ and $C_3$–$H_3$ can often be assigned. Therefore, we tested the search algorithm with the three $^{13}C$ resonances of $C_1$, $C_2$, $C_3$ and their associated proton resonances $H_1$, $H_2$ and $H_3$ (Supplementary Fig. 3c). The fourth strategy consisted of a search of a $C_1$–$H_1$ chemical shift pair together with all $^1H$ chemical shifts that are extractable from $H_1$-correlations in a 2D TOCSY—data that are easy to collect (Supplementary Fig. 3d).

The results of all search strategies using the three oligosaccharides are summarized in Figure 3. In all cases the top hits presented the correct monosaccharide type and surprisingly the linkage and
the chemical environment of the hits was in most cases either identical or similar to the query. The results for searching six unassigned C-H correlations and for searching the completely assigned spin system consisting of six $^{13}$C (C1–C6) and six $^1$H (H1–H6) resonances were identical for the given examples. However, the duration of the search differed by several orders of magnitude (Supplementary Fig. S4): it was typically in the milliseconds range for the assigned spin system while the search took several seconds when using unassigned pairs. The third search using three $^{13}$C (C1–C3) and three $^1$H (H1–H3) resonances gave similar good results. Only the fourth search strategy using C1, H1 and two to four additional $^1$H resonances extracted from a 2D TOCSY made the limitations visible: in many cases it also resulted in matches of identical or similar glycoepitopes but in rare cases the few chemical shifts are similar to other saccharides and the top hits might contain quite different saccharides, for example Gal instead of GlcNAc as in the case of Bv9 (terminal Gal instead of branched GlcNAc shown in Fig. 3).

2.2 Influence of the $^{13}$C offset

We noticed in initial searches a discrepancy in $^{13}$C referencing between our experimental data referenced to DSS and the data of glycosciences.de. For $^{13}$C referencing we used an external Bruker standard sample containing 2 mM sucrose and 0.5 mM DSS in H2O/D2O (9:1, v/v) and applied a $^{13}$C/$^1$H factor of 0.251449530 for calibrating $^{13}$C following the recommendations of IUPAC, IUBMB and IUPAB for proteins and nucleic acids (Markley et al., 1998; Wishart et al., 1995). Because the data in glycosciences.de should be referenced to TMS, a difference of 2.7 ppm is expected (Aeschbacher et al., 2012). To systematically analyze the offset between the database and values referenced to DSS, we repeated one of the searches for fucose of Bv9 and plotted the score value of the top hit as a function of the $^{13}$C offset (Supplementary Fig. S5). Interestingly, the optimum was not around the expected value of −2.7 ppm but between −1.8 and −2.0 ppm. Therefore we used −1.8 ppm in all our searches. An offset of −1.7 ppm was reported to be caused by dioxane as internal reference and indirect referencing to TMS leading to $^{13}$C chemical shift discrepancies in protein data (Wang and Wishart, 2005). Considering the widespread use of dioxane for referencing carbohydrate spectra, this is likely the cause for the −1.8 ppm offset. Especially for studies of glycoproteins and protein–carbohydrate interactions we recommend to use DSS as a common standard.

2.3 Application example: sialic acids in glycoproteins

GlycoNMRSearch can be applied to carbohydrates of unknown composition. An example is provided in Figure 4. We recently detected two signal sets of sialic acid spin systems of N-glycans in a commercial sample of human serum albumin (Schubert et al., 2015). This finding was unexpected, since the amino acid sequence of albumin lacks N-glycosylation sequons. As was later revealed by...
mass spectrometry, the commercial protein sample contained small amounts of at least four other proteins, namely haptoglobin, hemoglobin, serum transferrin and α1B glycoprotein (Grunwald-Gruber et al., 2017). Each of these proteins contains two to five N-glycans, contributing to the carbohydrate signals visible by NMR spectroscopy. Although their amount is estimated to contribute only 4–7% of total protein content, their high glycosylation density could lead to quite intense signals especially for terminal non-reducing saccharide ends.

Whereas the assignment was achieved at the time by the hypothesis that the two sets could originate from Neu5Ac2, 6Gal and Neu5Aca2, 6Gal and their typical chemical shifts reported in the literature were compared to the experimental ones, GlycoNMRSSearch provides directly a clear answer: one spin system matches Neu5Ac2, 6Gal, the other to Neu5Aca2, 6Gal (Fig. 4c–f).

The current search strategy may be limited by several factors: the database might lack certain sialic acid containing glycans, and neighboring antenna in multiantenna N-glycans could potentially influence the shifts as well.

2.4 Additional feature: chemical shift correlations
As a supplementary feature we provide statistical 2D plots of 1H–13C chemical shift pairs, corresponding to cross-peak correlations in a 13C-HSQC spectrum for each monosaccharide, e.g. for α-1-Fuc pyranose as shown in Supplementary Figure S6. This will also help to improve the database in the future by pinning down incorrect data. It may also reveal various clusters of chemical shifts reflecting interesting structure-chemical shift relationships similar to the recently found secondary structure element (Aeschbacher et al., 2017). In this particular case a non-conventional hydrogen bond lead to a characteristic down-field 1H chemical shift that appears as a separate cluster in a chemical shift statistics. After the origin of this separate cluster is determined, the characteristic chemical shift range can then be used as an indicator for a particular structure.

3 Materials and methods
3.1 The database
The core of the proposed search routine is a standardized dataset, developed on the basis of the entire NMR part of the glycosciences.de database (Loss and Lütteke, 2015; Lütteke et al., 2006). The dataset is composed of 16 465 monosaccharides, characterized by a 27-element chemical shift vector, a unique LinucID identifier (the entry number of glycosciences.de), linkage and monosaccharide name. To enable glycan searches, we performed a data cleaning step, which involved disambiguation of glycosciences.de chemical shift names, standardization of monosaccharide names and filtering out information that is not relevant for the proposed routine.

In total 271 types of monosaccharides are covered (Supplementary Fig. S7). Unfortunately, they are not distributed uniformly. The three most common monosaccharides (β-3-GlcNAc, β-3-Galp, α-2-Manp) constitute 7345 out of 19 093 records. Eighty-five types of glycans appear in the database more than 5 times and 97 are represented by just one record.

A crucial factor for performance of the search engine is the quality of the data, measured in completeness of chemical shifts. As shown in Section 2, typically six chemical shifts are sufficient to reliably identify a monosaccharide spin system. Fortunately, glycosciences.de contains thousands of such records. After the data cleaning step, we were able to identify 4200 records having at least 3 carbon shifts, 7232 monosaccharides with at least 3 hydrogen shifts, 2956 with at least 6 chemical shifts (3× C + 3× H) and 2746 complete entries (at least 5 carbon and 5 hydrogen shifts).

3.2 Integer programming formulation of the search routine
The proposed method calculates a ranking of monosaccharides, sorted in ascending order by a score, which is a measure of dissimilarity between the observed chemical shifts and the record from the glycosciences.de database (Lütteke et al., 2006). Consequently, the
top hits in the ranking suggest the saccharide type and substitutions of the observed spin system.

To start a search procedure, the spin system observations are provided as input. They adopt three forms: (i) chemical shift values (e.g., one of the atoms has a chemical shift of 63 ppm), (ii) linkage (e.g., an atom with a chemical shift of 63 ppm is next to an atom with a chemical shift of 70 ppm) and (iii) known assignments within a spin system (e.g., H3 has a chemical shift of 4 ppm). Those observations are automatically converted into an instance of a discrete optimization problem, and solved with the Branch and Bound algorithm (Junger et al., 2009).

More formally, a query vector \( q \in \mathbb{R}^{27} \) stores the observed resonance frequencies (ppm units) of the \(^1\)C, \(^1\)H and \(^1\)H atoms belonging to a monosaccharide spin system. In a general case, the chemical shift assignment is unavailable, thus there is no atom label (e.g., C1 or H1) associated with the observed resonance \( q_i \). We define \( q = [c_1, c_2, \ldots, c_9, h_1, h_2, \ldots, h_9, h_{1}', h_2', \ldots, h_{9}'] \), where \( c_i, h_i, h_i' \in \mathbb{R} \) are resonance frequencies of the \( i \)th carbon, hydrogen and second hydrogen (indicated by a prime in case of a CH2 group), respectively. Therefore each element of \( q \) is associated with a known chemical shift. Both vectors are related by \( q = Xq' \), where \( X \in \{0, 1\}^{27 \times 27} \) is a permutation matrix (Fig. 5).

The goal of the search procedure is to find a ranking of \( N \) database entries \( e = \{e^{(1)}, e^{(2)}, \ldots, e^{(N)}\} \), \( e^{(k)} \in \mathbb{R}^{27} \), such that the following implication holds:

\[
m \leq n \Rightarrow \ell(q, e^{(m)}) \leq \ell(q, e^{(n)})
\]

where \( \ell \) is a custom loss function, and \( m \) and \( n \) denote positions in the ranking.

The key problem in the search procedure is to find, for each database entry \( e^{(k)} \) and unassigned query vector \( q \), the permutation matrix \( X^{(k)} \), which minimizes the loss function under constraints. It is formulated as the following discrete optimization problem:

\[
\min_{X^{(k)}} \ell(X^{(k)}q, e^{(k)})
\]

subject to:

\[
\forall_j \sum_i x^{(k)}_{ij} = 1
\]

\[
\forall_j \sum_i x^{(k)}_{ij} = 1
\]

\[
\exists (i) \in \mathcal{P}_e \quad \sum_{j=1}^N x^{(k)}_{ij} = 1
\]

\[
\forall (i) \in \mathcal{P}_e \quad \exists m, n \in \mathcal{C} \quad x^{(k)}_{mi} + x^{(k)}_{mj} = 2 \wedge |m - n| = 1
\]

\[
\forall (i) \in \mathcal{P}_e \quad \exists m, n \in \mathcal{N} \quad x^{(k)}_{mi} + x^{(k)}_{mj} = 2 \wedge |m - n| = 1
\]

\[
\forall (i) \in \mathcal{P}_c \quad \exists m, n \in \mathcal{C} \quad x^{(k)}_{mi} + x^{(k)}_{mj} = 2 \wedge |m - n| = 9
\]

where \( \mathcal{P}_e \) is a set of resonance frequencies that are not permuted (chemical shifts are assigned), \( \mathcal{P}_c \) and \( \mathcal{P}_n \) are sets that store relative positioning of homonuclear (C-C, H-H) pairs of atoms, \( \mathcal{P}_{CH} \) is a set that constraints relative positioning of heteronuclear atoms, \( \mathcal{C} = \{1, 2, \ldots, 9\} \), \( \mathcal{N} = \{10, 11, \ldots, 27\} \) are indices of the vector \( q \) associated with carbon and hydrogen atoms.

Consider a simple instance of this minimization problem, where the chemical shift of H1 is known (5.3 ppm). In addition, the chemical shifts of a single C-H group with unknown position have been determined (56 ppm and 4 ppm). These assumptions yield the following initial values: \( q_1 = 5.3, q_3 = 56, q_4 = 4, \mathcal{P}_c = \{10, 11\}, \mathcal{P}_N = \{\}, \mathcal{P}_C = \{\}, \mathcal{P}_{CH} = \{1, 2, 3\} \). Thus \( q \) stores values of observed resonance frequencies. One element of the permutation matrix is known, because of the available H1 chemical shift assignment (\( \mathcal{P}_A \)). Since the pair \( (q_1, q_3) \) belongs to the same C-H group, it can be mapped to \( (q_2, q_1) \) \( (q_2, q_{20}) \) \( (q_1, q_{12}) \) \( (q_3, q_{21}) \) \( (q_4, q_{27}) \).

It was found that the algorithm works well in practice, if \( \ell \) is weighted least squares error, defined as:

\[
\ell(q, e^{(k)}) = \sum_{i \in \mathcal{C}} I(q_i \neq 0)(q_i - c_i^{(k)})^2
\]

\[
+\omega_N \sum_{i \in \mathcal{N}} I(q_i \neq 0)(q_i - c_i^{(k)})^2
\]

where \( \omega_N \in \mathbb{R}^+ \) is a weight associated with the hydrogen shifts mismatch and \( I(\cdot) \) is an indicator function, which ensures that zero elements of \( q \) do not contribute to the final score. In the experimental section, we set \( \omega_N = 10^2 \) to weight squared chemical shift mismatches of protons, in ppm, 100-fold higher than carbons.

Although a weighted least squares error is used by default, our search engine can perform an optimization with arbitrary loss \( \ell \). Some examples of \( (\ell, \omega_N) \) are available in the web site implementation. It is worth mentioning that results of the optimization task (Equation 1) are presented in a normalized form:

\[
y^{(k)} = \frac{y_{\max}^{(k)} - \ell(X^{(k)}q, e^{(k)})}{y_{\max}^{(k)}} \times 100\%
\]

where \( y^{(k)} \) is a normalized score for the \( k \)th database entry, \( y_{\max}^{(k)} \) is a constant, which defines the maximum allowed value of the loss function \( \ell \). Database entries \( e^{(k)} \), where \( \ell(q, e^{(k)}) > y_{\max}^{(k)} \) are filtered out by Branch and Bound, since they have no practical significance.

Generating a ranking \( e \) is possible by solving the above optimization problem for each of the 16 465 records in the database. Taking advantage of the fact that the value of the objective function (Equation 1) rises rapidly, if either two saccharides of different types are compared, or the permutation matrix is incorrect, it is possible to exclude many partial solutions early in the search procedure. Consequently, in empirical studies, the search time was less than 15 seconds in all cases, and frequently below 3 seconds (Supplementary Fig. S4).
4 Conclusion

GlycoNMRSearch is a significant improvement over earlier search functions for carbohydrate NMR data that allowed only searches for lists of either $^1$H or $^{13}$C chemical shifts (Kapaev and Toukach, 2018; Loss and Lütteke, 2015). GlycoNMRSearch predicts with high precision the monosaccharide type and even the environment of the monosaccharide as long as matching chemical shifts with similar epitopes are in the database. This complements new strategies to explore unknown oligosaccharides. The key of the approach is to link observed chemical shifts to spin systems, which are subsequently compared to a chemical shift database. As a result, the best matching oligosaccharides very rapidly suggest the nature of the oligosaccharide. GlycoNMRSearch is well-suited for the analysis of unknown carbohydrates either in glycoproteins (untargeted glycomics), glycolipids or natural products. We are convinced that GlycoNMRSearch will become an essential tool for analyzing oligosaccharides by NMR spectroscopy in the near future.

The power of the search algorithm depends of course on the quality and the coverage of the chemical shift database. If the database does not contain glycoepitopes that are similar to the query, the search algorithm cannot find any appropriate match. In such a case, the results may reveal somehow similar structures or also completely wrong ones, however with a significantly lower score value. Unfortunately the NMR part of glycosciences.de is not updated or corrected anymore. There is large room for improvements, for example implementing chemical shift data from other databases like the BioMagRes database (Ullrich et al., 2008), or the Carbohydrate structure database (CSDB) (Toukach and Egorova, 2016). In addition data from publications and background data from the CASPER chemical shift prediction platform (Lundborg and Widmalm, 2011) could be included. The best way to expand the database in the future, would be an upload and verification system for users that like to contribute to the database.

In summary, GlycoNMRSearch gives excellent results for the common saccharides that are well covered in the chemical shift database. The precision of GlycoNMRSearch will further increase, the more chemical shift data of high quality is available covering a wider range of glycoepitopes. An important task of the near future is to extend and improve the chemical shift database taking advantage of the vast amount of data that has already been published, but is not included in any database and to provide a tool that allows uploading and quality control of such data.

Acknowledgements

We thank Dr. Thomas Lütteke for providing the complete NMR data of the glycosciences.de database, Dr. Adam Gonzalez for valuable discussions, Dr. Wolfgang Skala, Dr. Peter Güntert, Dr. Alvaro Mallagaray and Arthur Hinterholzer for valuable comments on the manuscript.

Author contributions

P.K. and M.S. designed the search algorithm. P.K. coded the search algorithm and created the online web interface. M.S. and P.K. tested the algorithm. M.S. and P.K. wrote the manuscript.

Funding

This work was supported by Wroclaw University of Science and Technology and by the University of Salzburg.

Conflict of Interest: none declared.

References


