Available online at www.sciencedirect.com
SCIENCE DIRECT®
Carbohydrate Research 339 (2004) 1015–1020
Carbohydrate RESEARCH

Data mining the protein data bank: automatic detection and assignment of carbohydrate structures

Thomas Lütteke,* Martin Frank and Claus-W. von der Lieth

Central Spectroscopic Department, German Cancer Research Center, INF 280, D-69120 Heidelberg, Germany

Received 6 July 2003; accepted 15 September 2003

Abstract—Knowledge of the 3D structure of glycans is a prerequisite for a complete understanding of the biological processes glycoproteins are involved in. However, due to a lack of standardised nomenclature, carbohydrate compounds are difficult to locate within the Protein Data Bank (PDB). Using an algorithm that detects carbohydrate structures only requiring element types and atom coordinates, we were able to detect 1663 entries containing a total of 5647 carbohydrate chains. The majority of chains are found to be *N*-glycosidically bound. Noncovalently bound ligands are also frequent, while O-glycans form a minority. About 30% of all carbohydrate containing PDB entries comprise one or several errors. The automatic assignment of carbohydrate structures in PDB entries will improve the cross-linking of glycobiology resources with genomic and proteomic data collections, which will be an important issue of the upcoming glycomics projects. By aiding in detection of erroneous annotations and structures, the algorithm might also help to increase database quality.

Keywords: Data analysis; 3D structure database; Glycosylation; Bioinformatics; Algorithm

1. Introduction

Protein glycosylation is probably by far the most common and complex type of co- and posttranslational modifications encountered in proteins. Glycosylation differs from most other covalent protein modifications such as phosphorylation, acetylation and formylation with respect to the size and the complexity of the added group and the magnitude of the cellular machinery devoted to synthesis and modulation.[1,2] Inspection of protein databases reveals that as many as 70% of all proteins have potential *N*-glycosylation sites (Asn-X-Ser/Thr, X not proline) and *O*-glycosylation is even more ubiquitous.[3]

The oligosaccharide moieties cover a range of diverse biological functions. First of all, they are involved in the process of folding and subsequent conformational maturation in the rough endoplasmic reticulum. Without added glycans, many plasma membrane proteins and secretory proteins are not able to fold properly.[1,4] Simply because of their large size and hydrophilicity, glycans can alter the physico-chemical properties of a glycoprotein, making them more soluble, reducing backbone flexibility and therefore leading to increased protein stability, protecting them from proteolysis, etc.[3] Carbohydrates are absolutely required for the correct maturation, function and intracellular sorting of many glycoproteins. The failure of glycoproteins to be correctly processed, trafficked and degraded, leads to diseases in human. Certain carbohydrates found on the surface of cells can affect the onset and progression of disease states.[2]

Another major function of protein-linked glycans is to provide additional recognition epitopes for protein receptors. These are implicated in a variety of cell–cell and cell–matrix recognition events. By participating in the process of cellular recognition, the oligosaccharide moieties play a pivotal role in inflammation, immune response and cancer.[4–6] These recognition events involve specific carbohydrate binding proteins—the lectins.[4]

* Corresponding author. Fax: +49-6221-424554; e-mail: t.luetteke@dkfz.de

0008-6215/$ - see front matter © 2003 Elsevier Ltd. All rights reserved.
doi:10.1016/j.carres.2003.09.038

They depend on the precise three-dimensional shape of the glycan,[3] therefore knowledge of the 3D structure of the glycan is a prerequisite for a full understanding of the biological processes glycoproteins are involved in.

The progressing *glycomics* projects will dramatically accelerate the understanding of the roles of carbohydrates in cell communication and hopefully lead to novel therapeutic approaches for treatment of human disease. The MIT's magazine of innovation (21 January 2003) has identified *glycomics* as one of the top 10 technologies that will change the future. The development of new and advanced bioinformatic tools, algorithms and data collections for glycobiology is an absolute requirement to manage and analyse successfully the large amount of data, which will be produced by the upcoming glycomics projects. An important issue will be the cross-linking of glycobiology resources with genomic and proteomic data collections. The existing glycorelated databases are not reciprocally cross-referenced as are, for instance, the diverse gene and protein databases. Although the oligosaccharide databases make reference to major protein databases, no significant effort is made to link protein entries to their glycan structures.[7]

The largest source of biomolecular 3D structures publicly available is the Protein Data Bank (PDB).[8] In September 2003, the PDB consists of about 22,500 entries, most of which are proteins. Only 18 entries are pure carbohydrate structures, but many protein structures include carbohydrate compounds. The problem, however, is that in contrast to proteins or nucleic acids there is no standard nomenclature for carbohydrate residues within the pdb-files.[9] In some cases, entire carbohydrate chains are combined in one single residue (e.g., the residues ASL and ASF in the PDB entry 1agm). Furthermore, for many monosaccharide residues as defined in the PDB Het Group Dictionary (http://pdb.rutgers.edu/het_dictionary.txt) there is no distinction between α- and β-form. Information about how the single carbohydrate residues are linked to each other may be given within the LINK records of a pdb-file but is missing in most cases. For these reasons, it is difficult for glycobiologists to find a carbohydrate structure of their interest within the PDB. So far, only very few attempts to analyse special types of this carbohydrate related data were made.[3,4,9] Here, we present an algorithm that detects and assigns carbohydrate compounds—covalently attached glycans as well as noncovalently bound ligands—just on the basis of element types and atom coordinates. This approach has the advantage that it is not limited to the annotation given in pdb file format. The algorithm was implemented in the software *pdb2linucs*. The program analyses the input file and lists included carbohydrates using the LINUCS notation. LINUCS is a linear, unique nomenclature for carbohydrate structures that is well suited for use in data processing or computer algorithms.[10]

2. Materials and methods

2.1. Protein Data Bank

According to the PDB Holdings List of 9 September 2003, the PDB contains a total of 22,448 structures, 19,062 of which are solved by X-ray, the remaining 3386 are solved by NMR. The vast majority of the entries (20,262) are classified as Proteins, Peptides and Viruses, 1231 are nucleic acids, 937 represent protein/nucleic acid complexes and 18 entries are referred to as carbohydrate structures.

2.2. Algorithm

To detect carbohydrate compounds in a molecular 3D structure, a perception of the molecular topology given in the structure file is performed. In pdb-files, carbohydrates belong to the nonstandard residues. Therefore, information about carbohydrate structures can be found in the HETATM records of the pdb-files. For nonstandard residues, connectivity information should be given explicitly in the pdb-files within the CONECT records. Alternatively, bonds can be calculated from the distances between the atoms.

2.2.1. Detection of carbohydrate rings. In a first step, all atoms from the pdb HETATM records being part of a ring are identified and the ring size as well as the element types of the atoms forming the rings are evaluated. Potential carbohydrate structures are assigned to a ring size of five or six atoms, exactly one of which is an oxygen. All the other atoms forming the ring have to be carbons. To identify the anomeric carbon for each ring, it is checked if one of the two carbon atoms that are connected to the ring oxygen is attached to another oxygen, nitrogen of sulfur atom. If such an atom is found, the respective carbon atom is assigned as the anomeric carbon. Several pdb-files contain carbohydrate rings lacking the oxygen or nitrogen atom at the anomeric carbon. To be able to detect these rings as well, the two carbon atoms are checked again, this time for another attached exocyclic carbon atom. If such an atom is found, the respective ring carbon atom can be assigned as the C-5-atom or in furanose rings the C-4-atom, respectively, whereas the other one is assigned as the anomeric carbon. In many cases, there is an oxygen or nitrogen atom of another residue within the vicinity of the anomeric carbon, but too far away to be regarded as a normal covalent bond. Visual inspection of such structures often support the assumption that a certain site should be glycosylated or certain sugar units should be connected. Insufficient experimental resolution and/or inadequate procedures to optimise the carbohydrate moiety may be the reason for the improper positioning of the sugar relative to the protein. To assign these

'potential' linkages, the oxygen and nitrogen atoms near the anomeric carbon are searched for possible connections. Asparagine nitrogens are only considered if the amino acid is part of a sequon. For each possible linkage, a penalty score is calculated from deviations of bond length and bond angles from the standard values. Linkages with a penalty score above a threshold are omitted. If more than one possible linkage is found by this procedure, the one with the lowest penalty score is selected.

2.2.2. Chain detection. In the next step, the rings forming the reducing end of carbohydrate chains are identified. This is accomplished by finding those carbohydrate rings where the anomeric carbon is not glycosidically connected to another carbohydrate residue. In glycan structures, this atom is bound to an amino acid, in ligands, it is connected to a noncarbohydrate group or just to a hydroxyl oxygen. After the carbohydrate type of this ring is assigned (see below), all residues belonging to the respective carbohydrate chain have to be found. To do this, the nonanomeric carbons are checked for attached carbohydrate residues. In the case such residues are found, they are recursively handled in the same way. By this procedure, the entire carbohydrate chain is assigned. The collected data about monosaccharide types and linkages is stored in a tree-like graph, from which the LINUCS, code is built subsequently.

2.2.3. Residue type assignment. To detect the type of a carbohydrate residue, a coding scheme is derived based on ring size and a description of the monosaccharide's stereo-centres. The spatial orientation of the hydroxyl groups is evaluated using a virtual torsion angle formed by the four atoms connected to each ring carbon atom. This angle is calculated from the virtual bonds C_{n-1}–O_n–H_n–C_{n-1} (Fig. 1a). If the hydroxyl group points down relatively to the plane spanned by the respective ring carbon and the two ring atoms connected to that carbon, the virtual torsion angle is positive. In this case, a '1' is added to the code. In the opposite case, that is the hydroxyl group points up, the torsion angle is negative, which results in a '2' in the code (Fig. 1b). If the carbon atom assigned as the anomeric carbon is connected to another exocyclic carbon atom, the ring is a ketose or a sialic acid. In this case, the value representing the stereochemistry of this ring carbon is increased by 3. The derived code can be easily compared to a lookup table where the stereochemistry of carbohydrates occurring in the PDB is stored.

After the basic residue type is detected, modifications like acetylation, methylation or sulfatation are assigned by checking the groups that are connected to the carbon or the oxygen atoms, respectively.

2.3. Web interface

To make the software accessible to the public, a web interface was installed at http://www.dkfz-heidelberg.de/spec/pdb2linucs/. There, the user can directly access the PDB reservoir and input the PDB ID of an entry of interest or upload a file in pdb format. If carbohydrate compounds are identified, the entry can be visualised using the freely available *Chime* plugin (http://www.mdlchime.com). The LINUCS codes of the carbohydrate chains found in the entry are listed on the results page. For those chains where there is an entry available in the carbohydrate database SWEET-DB,[11] a direct link to that entry is offered. The PDB server can be directly accessed to recall further data like references, experimental conditions and similar protein structures. Additionally, chains consisting of two or more residues can be displayed using the IUPAC-nomenclature or as symbolic trees. In the case N- or O-glycans are detected, the amino acid sequence of the respective protein is also displayed with the residues of glycosylation sites highlighted in colour.

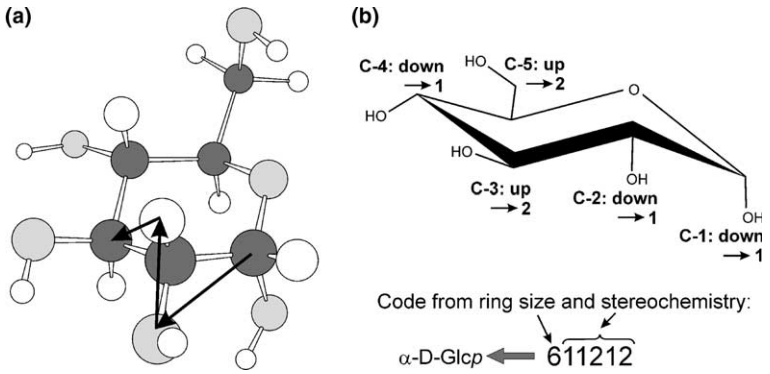

Figure 1. Assignment of monosaccharides: (a) To detect the stereochemistry of the ring carbon atom C_n, a virtual torsion angle (C_{n-1}–O_n–H_n–C_{n+1}) is defined; (b) Depending on the spatial orientation of the hydroxyl group, this angle becomes either positive or negative, which is encoded as 1 or 2, respectively, in the coding scheme.

2.4. Implementation

The algorithm was implemented in the program *pdb2linucs*. The software was written in C, parameter files like the carbohydrate lookup table are realised as ASCII texts based on the extensible mark-up language XML. The web interface is implemented in PHP, which is a script language that is interpreted on the web server and generates HTML output. For visualisation, the freely available *Chime* plugin (http://www.mdlchime.com) is used.

3. Results and discussion

3.1. Cross-referencing with other data collections

The automatic annotation of carbohydrate structures and their conversion to a unique linear notation enables an efficient cross-referencing with other data collections containing glyco-related data. As a showcase, direct links to SWEET-DB[11] were established. For each glycan chain detected in a selected PDB file, SWEET-DB is automatically searched and a direct link is established in case of success. The data contained in SWEET-DB can be retrieved by clicking on the corresponding button. In case the selected PDB ID is available as an entry in the 3D lectine database (http://www.cermav.cnrs.fr/lectines/), a direct link to that database is established as well.

3.2. Distribution of carbohydrate compounds

Among the about 22,500 entries available in the PDB (September 2003), we were able to detect 1663 entries containing carbohydrate compounds. Carbohydrate chains (5647) consisting of 10,726 residues were found (Table 1). About half of the chains are found to be *N*-glycosidically bound. Noncovalently bound ligands are also frequent, while O-glycans form a minority. Within all glycan chains present in the PDB, 88% are N-glycans, 10% are *O*-glycosidically bound to serine or threonine, while only 2% are connected to glutamate or aspartate.

In nearly all of the chains bound to aspartate or glutamate, the first residue is 2-deoxy, 2-deoxy-2-fluoro or there is an aliphatic group like glycerol inserted between the amino acid and the first carbohydrate ring (data not shown).

Compared to the estimation that more than 50% of all proteins are glycosylated, which is based on an analysis of well-characterised and annotated glycoproteins,[12,13] 856 out of 22,500 protein structures containing glycosidically bound carbohydrates seem to be few. This can be explained by the fact that the majority of the proteins deposited in the pdb are either recombinantly expressed in *E. coli* or originate from bacteria. In both cases, the proteins will not be glycosylated, even if the original mammalian or plant protein is. The carbohydrate compounds, which are often located on the surface of the glycoproteins, are quite flexible. Therefore, only a few entries in the PDB contain X-ray diffraction data with sufficient electron density for detecting an entire oligosaccharide chain attached to a glycoprotein in crystalline form.[9] Furthermore, the conformational flexibility of glycan chains on the protein surface may hamper crystal growth.[4] In most cases where large regions of the glycan are available in the structure, either the glycan lies along the protein surface or bridges between adjacent protein molecules. Both instances lead to an immobilisation of the glycan.[9]

Table 1. Distribution of carbohydrate compounds found in the PDB

Type	Entries	Chains	Residues
Glycans	856	3255	6249
Asn	770	2852	5745
Ser	62	194	237
Thr	47	131	152
Glu	34	67	95
Asp	11	11	20
Ligands	957	2392	4477
Total	1663	5647	10,726

For each chain type, the number of PDB entries containing one or more chains of this type, the total number of chains found in the PDB and the number of residues the chains consist of are given. The sum of the single values in the 'entries' column exceeds the total number of carbohydrate containing PDB entries since there are many files containing different glycan types or both glycan and ligand chains.

3.3. Distribution of monosaccharide types

N-Acetylglucosamine is by far the most common monosaccharide type within the PDB (Table 2). This can be explained by the fact that the first two residues of a *N*-glycosidically bound chain are of this type,[1,2] and the average chain length of N-glycan structures in the PDB is 1.99. N-Glycans make up 88% of the glycans and 51% of all chains found in the PDB (Table 1). O-Glycans and ligands also contain *N*-acetylglucosamine residues. All glucose-based residues together form more than 60% of all monosaccharides in the PDB. Together with mannose, galactose and fucose residues, more than 90% of all monosaccharides are covered. For some monosaccharide types, like those glucose residues that are not *N*-acetylglucosamine, alpha and beta forms are almost equally distributed, while others show an obvious preference for one of these forms. Apart from the fucose residues, all frequently occurring residues appear mainly in the D chiral form (Table 2). Only 12 residues present in the L chiral form were found here. Among the less common residue types, the L chiral form is preferred by Rib*f*, Ido*p*, All*p* and Ara*p* (data not shown).

Table 2. Distribution of monosaccharide types in the PDB

Type	Number	Ratio (%)	α-D/β-D[a]	Ratio (%)
Glc*p*NAc	4766	44.4	410/4298	8.7/91.3
Glc*p* (not NAc)	2026	18.9	1004/996	50.2/49.8
Man*p*	2008	18.7	1321/670	66.3/33.6
Gal*p*	795	7.4	248/541	31.4/68.6
Fuc*p*	346	3.2	223/89/18/10[a]	65.6/26.2/5.3/2.9
Xyl*p*	224	2.1	30/188	13.8/86.2
Fru*f*	158	1.5	46/188	19.7/80.3
Neu*p*5NAc	140	1.3	134/6	95.7/4.3
Other	263	2.5	143/39/42/34[a]	55.4/15.1/16.3/13.2

Monosaccharide residues found in the PDB are listed in the order of occurrence. Ratio of the basic type is given relative to the total number of residues, ratio of subtypes (α-D/β-D) relative to the basic type, respectively. Occurrence of subtypes do not always sum up to the occurrence of the basic type since some residues found in the PDB lack an oxygen or respective atom connected to the anomeric carbon and thus cannot be assigned to alpha/beta form.

[a] Fuc*p*, other: α-L/β-L/α-D/β-D.

3.4. Erroneous entries

About 30% of all carbohydrate containing PDB entries comprise one or several errors. The most common type of errors found in the PDB is a wrong assignment of the α-/β-isoforms. For example, there are two different PDB residue names for mannoses: MAN encodes for α-D-Man*p*, BMA for β-D-Man*p*. We have found 263 entries that contain at least one β-D-Man*p* named MAN. The opposite case, α-D-Man*p* named BMA, was found in 10 entries only. Another frequent error is the lack of an oxygen, nitrogen or sulfur atom connected to the anomeric carbon. A wrong assignment of the residue type is also quite common. The number of entries with missing connections almost equals the quantity of entries comprising surplus connections, while there are more than five times as many entries lacking atoms than containing surplus ones. Other kinds of errors like interchanged coordinates or overlapping residues occur infrequently (Table 3).

Table 3. Types of errors detected in pdb entries

Type of error	Number of PDB entries comprising the type of error
Wrong assignment of α-/β-isoform	286[a]
No oxygen or respective atom connected to anomeric carbon	96
Wrong residue type assignment	50
Wrong assignment of D-/L-form	14
Missing connections	37
Surplus connections	34
Missing atoms	38
Surplus atoms	7
Other	19

The errors detected during the PDB analysis can be classified into eight categories. Each category is counted only once per entry. The sum of the category values exceeds the total number of erroneous entries since there can be errors of different categories within one entry.

[a] 263 of which are β-D-Man*p* named α-D-Man*p* residues.

Two examples of erroneous structures are shown in Figure 2: In the PDB entry 1b3j, the coordinates of the C-2 atom and the *N*-acetyl nitrogen atom of the residue NAG 302B are interchanged (Fig. 2a). A plenty of wrong linkages within single residues as well as between different residues are found within the PDB entry 1dzg (Fig. 2b). In that entry, atoms, which are distant more than 60 Å are marked to be connected.

4. Outlook

The large frequency of wrong annotation and other errors in the carbohydrate structures in PDB files points up the need of a check software. In the area of protein

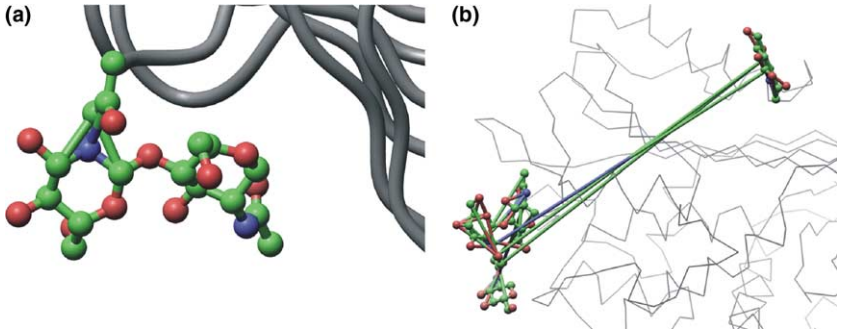

Figure 2. Erroneous structures (examples). (a) PDB entry 1b3j. The coordinates of the C-2 atom and the *N*-acetyl nitrogen atom of residue NAG 302B are interchanged. (b) PDB entry 1dzg contains a series of wrong connection within single residues as well as between different residues.

structures, several quality control tools like WhatCheck[14] or ProCheck[15] are available. Frequent use of such programs leads to an increasing database quality.[16] Similarly, the algorithm presented here could aid to improve the reliability of the annotation of carbohydrate compounds stored in PDB entries.

The automatic, IUPAC-conform annotation of carbohydrate structures on the basis of atomic coordinates, atoms and bonds as provided in PDB entries will improve the possibilities to cross-link glycorelated data with other genomic and proteomic data collections. This will be an important issue of the upcoming glycomics projects. The presented automatic algorithm to extract and annotate glyco-related information contained in the PDB allows easy establishing of self-acting procedures so that all cross-references in linked data collections can be routinely updated. Such automated procedures are necessary since the number of new entries deposited in the PDB is rapidly increasing due to the rapid progress to automate the experimental procedures to determine 3D structure.[17]

Acknowledgements

This work was supported by a grant from the German Research Council (DFG) within the digital library program.

References

1. Helenius, A.; Aebi, M. *Science* **2001**, *291*, 2364–2369.
2. Charlwood, J.; Bryant, D.; Skehel, J. M.; Camilleri, P. *Biomol. Eng.* **2001**, *18*, 229–240.
3. Wormald, M. R.; Petrescu, A. J.; Pao, Y.-L.; Glithero, A.; Elliot, T.; Dwek, R. A. *Chem. Rev.* **2002**, *102*, 371–386.
4. Imberty, A.; Perez, S. *Protein Eng.* **1995**, *8*, 699–709.
5. Rudd, P.; Elliott, T.; Cresswell, P.; Wilson, I.; Dwek, R. *Science* **2001**, *291*, 2370–2376.
6. Kirschner, K. N.; Woods, J. R. *Proc. Natl. Acad. Sci. U.S.A.* **2001**, *98*, 10541–10545.
7. Marchal, I.; Golfier, G.; Dugas, O.; Majed, M. *Biochimie* **2003**, *85*, 75–81.
8. Berman, H. M.; Westbrook, J.; Feng, Z.; Gilliland, G.; Bhat, T. N.; Weissig, H.; Shindyalov, I. N.; Bourne, P. E. *Nucleic Acids Res.* **2000**, *28*, 235–242.
9. Petrescu, A.; Petrescu, S.; Dwek, R.; Wormland, M. *Glycobiology* **1999**, *9*, 343–352.
10. Bohne-Lang, A.; Lang, E.; Förster, T.; von der Lieth, C.-W. *Carbohydr. Res.* **2001**, *336*, 1–11.
11. Loss, A.; Bunsmann, P.; Bohne, A.; Loss, A.; Schwarzer, E.; Lang, E.; von der Lieth, C.-W. *Nucleic Acids Res.* **2002**, *30*, 405–408.
12. Apweiler, R.; Hermjakob, H.; Sharon, N. *Biochim. Biophys. Acta* **1999**, *1473*, 4–8.
13. Jung, E.; Veuthey, A.-L.; Gasteiger, E.; Bairoch, A. *Proteomics* **2001**, *1*, 262–268.
14. Hooft, R. W. W.; Vriend, G.; Sander, C.; Abola, E. E. *Nature* **1996**, *381*, 272.
15. Laskowski, R. A.; MacArthur, M. W.; Moss, D. S.; Thornton, J. M. *J. Appl. Crystallogr.* **1993**, *26*, 283–291.
16. Weissig, H.; Bourne, P. E. *Bioinformatics* **1999**, *15*, 807–831.
17. Oldfield, T. J. *Proteins* **2002**, *49*, 510–528.