Nucleic Acids Research, 2024, 1–9
https://doi.org/10.1093/nar/gkae890
Database issue

DIONYSUS: a database of protein–carbohydrate interfaces

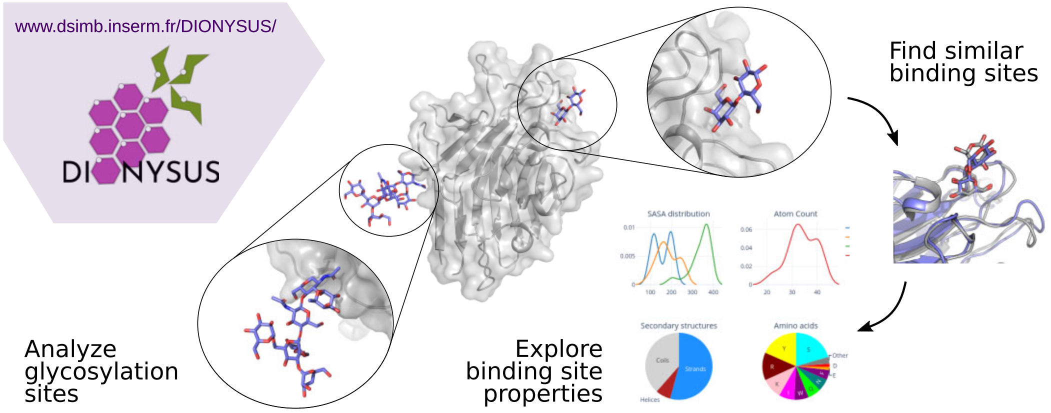
Aria Gheeraert[1], Thomas Bailly[1], Yani Ren[1,2], Ali Hamraoui[1,3], Julie Te[1], Yann Vander Meersche[1], Gabriel Cretin[1], Ravy Leon Foun Lin[1], Jean-Christophe Gelly[1], Serge Pérez[4], Frédéric Guyon[1] and Tatiana Galochkina[1,*]

[1]Université Paris Cité and Université des Antilles and Université de la Réunion, INSERM, BIGR, DSIMB, F-75015 Paris, France
[2]Université Paris-Saclay, INRAE, MetaGenoPolis, 78350 Jouy-en-Josas, France
[3]Institut de biologie de l'Ecole normale supérieure (IBENS), Ecole normale supérieure, CNRS, INSERM, PSL Universite Paris, 75005 Paris, France
[4]Centre de Recherches sur les Macromolécules Végétales, University Grenoble Alpes, CNRS, UPR, 5301 Grenoble, France
[*]To whom correspondence should be addressed. Tel: +33 1 81 72 43 30; Email: tatiana.galochkina@u-paris.fr

Abstract

Protein-carbohydrate interactions govern a wide variety of biological processes and play an essential role in the development of different diseases. Here, we present DIONYSUS, the first database of protein-carbohydrate interfaces annotated according to structural, chemical and functional properties of both proteins and carbohydrates. We provide exhaustive information on the nature of interactions, binding site composition, biological function and specific additional information retrieved from existing databases. The user can easily search the database using protein sequence and structure information or by carbohydrate binding site properties. Moreover, for a given interaction site, the user can perform its comparison with a representative subset of non-covalent protein-carbohydrate interactions to retrieve information on its potential function or specificity. Therefore, DIONYSUS is a source of valuable information both for a deeper understanding of general protein-carbohydrate interaction patterns, for annotation of the previously unannotated proteins and for such applications as carbohydrate-based drug design. DIONYSUS is freely available at www.dsimb.inserm.fr/DIONYSUS/.

Graphical abstract


www.dsimb.inserm.fr/DIONYSUS/
DIONYSUS
Find similar binding sites
Analyze glycosylation sites
Explore binding site properties
SASA distribution
Atom Count
Secondary structures
Amino acids
Coils
Strands
Helices
Other

Introduction

Carbohydrates are ubiquitous in enzymatic pathways and are the primary energy source in living organisms (1). Moreover, carbohydrates are abundantly present on the surface of living cells and thus are involved in cellular recognition and signaling (2–6). As a result, protein–carbohydrate interactions are related to various diseases, such as cancer tumor growth (7), and mediate a number of host–pathogen infections (8–11). Therefore, both carbohydrates and carbohydrate binding proteins are important targets for drug and protein design (12,13).

Despite the significant contribution of the glycoscience community to the annotation of carbohydrates and carbohydrate binding proteins (14–17), the comparison of protein-carbohydrate interfaces remains a complex task mostly due to the chemical diversity of carbohydrates, the flexibility of protein-carbohydrate interfaces and the experimental challenges of their resolution (18,19). The first database providing information on the resolved structures of protein-carbohydrate complexes, ProCarbDB (17), was released in 2021 but is no longer available at the time of publication of this article. Information on carbohydrate binding sites (CBS) can also be implicitly retrieved from other non-specific databases on protein-ligand interactions such as BioLiP (20,21) or Binding-MOAD (22). Still, they often miss

Received: July 3, 2024. **Revised:** September 3, 2024. **Editorial Decision:** September 25, 2024. **Accepted:** September 26, 2024

specific carbohydrate properties and do not distinguish between covalent and non-covalent interactions. Therefore, our primary goal in this study was an extensive systematic annotation and classification of all experimentally resolved protein–carbohydrate interfaces.

We have retrieved all the carbohydrate-containing entries available in the Protein Data Bank (23) and provided an exhaustive annotation of protein, carbohydrate and interaction site properties according to specialized databases and information extracted from 3D structure. Then, we performed a pairwise comparison and clustering of carbohydrate binding sites according to their geometry for a selected high quality subset of non-covalent protein-carbohydrate interfaces. DIONYSUS allows the user to efficiently perform comparative analysis of different sugar binding sites, to explore carbohydrate specificity for various protein types, as well as to evaluate complex composition and resolution quality through a user-friendly interface.

Materials and methods

Data extraction

Identification of carbohydrate-bringing compounds. We have extracted all the sugar-like compounds present in the Chemical Component Dictionary of the wwPDB (23). Our final list includes >3k different components, including nucleosides and their derivatives. Importantly, we have explicitly excluded RNA and DNA-forming polymer components from this study and invite users to consult specialized databases such as DNAproDB (24) for the analysis of corresponding molecular interfaces. Among identified compounds, 168 molecules with exhaustive carbohydrate-specific annotation in the PDB (25) were denoted as 'core dataset' (see Supplementary Data and Supplementary Figure S4 for the details).

Retrieval of general information. We used an API to download all mmCIF structure files with residue-level annotations containing at least one protein chain and one of the identified saccharides from the EMBL-EBI server (26). Each protein structure file was parsed using the GEMMI module (26). For structures containing several models and alternate locations, we treated each combination of a model and an alternate location individually. We report essential information such as PDB code, method of resolution, comments about the quality of the resolved structure, UniProt ID, organism, functional information, complete sequence, missing residues and missing atoms. For the web interface, we used the RCSB Saguaro Web Application (27) to recover protein information including secondary structure, missing residues, artifacts, mutations, metal coordination, hydropathy, disorder, and different domain annotations if they exist.

Interface analysis. For each carbohydrate residue, we analyzed its geometry and physico-chemical properties of the protein surface region in its proximity. Since protein–carbohydrate interactions are often flexible and contain resolution errors, we adopt a purely geometric approximation of a carbohydrate binding site for non-covalent protein-carbohydrate interactions, similar to that used in works on peptide–protein and drug–protein interfaces (28). For a given carbohydrate moiety (monomer or part of a more complex ligand), we consider its binding site to be formed by the heavy atoms located on the protein surface (solvent accessible surface area (SASA) >0 according to freeSASA (29)) and closer than 7 Å to any atom of the carbohydrate ring. According to this definition, if a complex glycoside interacts with a protein only through its aglycon part, it would not be classified as forming a protein-carbohydrate interface. In case of alternate locations of protein or carbohydrate residues, we generate distinct binding sites for each combination of model and alternate location. For each binding site we report chains involved in binding and cross-reference our binding site database with BioLiP2 (20,21), CAZy (30) (carbohydrate-active enzymes), UniLectin (14,31), SAbDab (32), LectomeXplore (33), Uniprot (34), GAG-DB (35). Finally, we use the same geometrical approximation to analyze protein surface region in the proximity of carbohydrates covalently attached to protein residues.

Annotation of protein–carbohydrate interfaces

PDB contains protein–carbohydrate complexes of different nature, including non-covalent ligand binding, glycosylated proteins and glycopeptides, and reaction intermediates in sugar enzymes. We attribute each carbohydrate residue to one of the three types according to their interaction mode: free carbohydrates (involving strictly non-covalent interactions), glycosylation heads (representing a sugar moiety covalently attached to a protein residue), and glycosylation bodies (referring to a carbohydrate portion covalently linked to the glycosylation head, potentially engaging in non-covalent interactions with protein residues). We use the same approximation for other covalent interactions, such as covalently linked aglycons (see Supplementary Data for the details).

We use both author-provided annotations and our computational assessments to validate the presence of covalent bonds between carbohydrates and proteins. The author's annotations can reveal inaccurately resolved glycosylations (e.g. 7QTV, where the N-glycosylation between N205 chain B and NAG401 chain B exhibits a C-N bond length of 3.4 Å). At the same time, sometimes they do not correspond well to the structural evidence (e.g. in structure 6S08, an O-glycosylation between T272, chain A, and FUC1 chain C was not reported, but distance calculation suggests its presence). For each bond, we assess its potential covalency by considering the sum of the upper bounds of experimental covalent radii (36) for the atoms involved, allowing for some margin (1.1 times the sum of upper bounds). Any structure containing a mismatch between the two methods is annotated accordingly. Additionally, we detect clashes between two atoms when their distance is less than half the sum of their experimental covalent radii, and we parse close contacts using the *_pdbx_validation_close_contact* section of the mmCIF file.

For glycosylation sites, we consider four categories: N-glycosylations involving a C-N bond between the carbohydrate and an asparagine or arginine residue, O-glycosylations defined as a C-O bond between the carbohydrate and either a threonine, serine, tyrosine, hydroxylysine or hydroxyproline, and C-glycosylations characterized by a C-C bond between the carbohydrate (typically mannose) and a tryptophan residue. Any other covalent bond between carbohydrates and proteins not fitting these categories is assigned to a separate class: undetermined. This category includes artifacts in PDB structures, covalent intermediates in sugar enzymes, covalently linked aglycons, or S-glycosylations (see section 'Difficult cases in annotation of glycosylation sites' of Supplementary Data, Supplementary Figures S1–S3).

For each CBS, we report its size in terms of atom number and solvent accessible surface area. Additionally, we provide information about the presence of other carbohydrates, nucleic acids or small molecules within the 7 Å cut-off, along with details regarding the chains and amino acids to which the carbohydrate binds. Carbohydrate binding sites often involve multiple protein chains (in particular, in antibodies and toxins). We consider all the chains involved in the interface formation and annotate the corresponding CBS accordingly. We report binding site composition in terms of amino acid types and secondary structure (helices, strands, and coils). Finally, we evaluate reliability of the resolved structure in terms of missing atoms in the carbohydrate ring/complete structure and incomplete occupancy of the carbohydrate.

The subset of protein-carbohydrate interfaces of high quality is denoted as a 'Refined dataset' and comprises structures with no artifacts detected by our analysis, i.e. satisfying the following conditions:

- The structure resolution is better than 3 Å for X-ray and EM structures
- No resolution problems are detected in the binding site proximity (no clashes/close contacts between any ligand atom and protein)
- There are no missing atoms in the protein binding site or in the carbohydrate compound
- At least 20 protein atoms participate in the binding site formation according to our definition
- The corresponding binding site is reported as biologically relevant according to BioLiP2

Such filters allow us to exclude from consideration a majority of crystallographic adjuvants, even though the automated protocol of BioLip2 can potentially miss out some binding sites of biological interest.

Information on carbohydrate and protein content is available at the 'About' page, which is updated dynamically.

Mapping to other databases

All the extracted PDB chains were mapped to UniProt (34), allowing us to obtain protein IDs in various databases including Gene Ontology (37), Enzyme Commission number in EXPASY (38), BRENDA (39) and information on the location of active sites and glycosylation sites and alternative PDB structures of the same protein. We have also retrieved residue-level annotations of protein domains according to Pfam (40), SCOP2 (41), CATH (42) and ECOD (43,44). Additionally, each entry was cross-referenced with several carbohydrate-specific databases such as UniLectin3D (14) for lectins, LectomeXplore (33) for predicted lectins, CAZy (30,45) for carbohydrate-active enzymes, SAbDab (32) for antibodies, GlyGen (46) and GlyConnect (47). Furthermore, we linked the binding site database to BioLiP (20,21), which additionally provides binding affinities for certain CBS through external databases like Binding MOAD (22), PDBBind (48) and BindingDB (49).

Finally, leveraging the resulting cross-annotations, we define four distinct categories of binding sites present in DIONYSUS (Supplementary Figure S3). First, we identify lectin binding sites as those situated within protein chains classified as lectins by Unilectin3D, whilst removing active sites based on Uniprot annotations. Then, we define enzyme active binding sites, as those located in structures reported in CAZy and aligned with active sites according to UniProt annotations. We distinguish different CAZy classes: Glycosyl Hydrolases, Glycosyl Transferases, Polysaccharide Lyases, Carbohydrate Esterases and Auxiliary Activities. Carbohydrate Binding Modules (CBMs) attached to enzymes and responsible for non-catalytic carbohydrate recognition are treated separately. Finally, we define antibody binding sites as those cross-referenced within a protein, denoted as heavy or light antibody chain in SAbDab. All the remaining binding sites are categorized as 'Others'.

Binding site alignment and similarity score

To explore structural similarities among CBS, we used a modification of the non-sequential structural alignment algorithm previously developed for off-target drug binding detection (50,51) and successfully applied to screening of protein-peptide interactions (28). The details of the method implementation are provided in Supplementary Data and in the 'About' section of the website. The code is available at: https://github.com/DSIMB/CompareCBS.

Redundancy elimination

Protein-ligand complexes are frequently crystallized as oligomers, and the exact same protein-ligand complex may have been resolved multiple times in the same pose. Still, the same protein chain can form multiple distinct binding sites for the same carbohydrate. We identified two CBS as identical if (i) the corresponding protein sequences share > 95% similarity, (ii) the interacting carbohydrate residue is the same and (iii) the non-sequential geometrical comparison of binding sites shows a score above 0.7 and coverage above 0.8 (see Supplementary Figure S6). Protein sequence identity clusters were retrieved from the PDB webserver (25) which uses the MMseqs2 (52) clustering algorithm. Recognition of identical oligosaccharides with different IUPAC names and subsequent glycan drawings were performed using Glycowork (53). Finally, to select a non-redundant representative subset of CBS we constructed a network where an edge links nodes representing identical sites (Supplementary Figure S6). We proceeded by iteratively choosing the node with the highest degree to be the 'representative' of its adjacent nodes. Subsequently, we eliminate these selected nodes from the network, continuing the process until the graph is empty.

Clustering of sugar binding sites

We performed clustering of protein-carbohydrate interfaces for protein complexes with the most common carbohydrates (see section 'Identification of carbohydrate-like compounds') and focused only on CBS of high quality ('Refined dataset' as defined above),

We then perform all-vs-all pairwise non-sequential alignments of the representative sites in each functional class, except for 'others' (Supplementary Table S1). Using the resulting similarity matrices, we performed a hierarchical variation of spectral clustering (54,55) which was demonstrated to be efficient in managing complex similarity matrices (56). Technical details of clustering implementation are provided in Supplementary Data section 'Hierarchical spectral clustering of carbohydrate binding sites' (Supplementary Figure S6 and Supplementary Table S2).

Search by keyword, protein sequence and structure

We have implemented database search by sugar- or protein-related keywords, by protein sequence and by protein structure. Keyword search analyzes all the protein, carbohydrate and binding site properties obtained as described above. Sequence search is performed using the *ggsearch36* tool in the fasta package (57) and searches among all the protein chains found in contact with at least one sugar moiety. Structure search is implemented using kpax (58), which aligns a target protein against a subset of protein domains with unique ECOD IDs selected based on their involvement in protein-carbohydrate interaction and the best resolution among proteins with the same fold. To obtain an exhaustive list of protein structures of similar fold, the user is invited to use the best hit ECOD ID as a parameter for the advanced database search.

Database interface and management

The DIONYSUS interface design is part of a standardized designed system developed by our team for our recently released web-servers and databases (www.dsimb.inserm.fr/pages/tools.html). DIONYSUS is developed using the Bootstrap framework and is fully responsive, ensuring an optimal user experience across various devices. DIONYSUS robustness was monkey tested using gremlins tool, which simulates random actions using JavaScript. Database creation is fully automated thanks to implementation of cron jobs and regular backups. Full database updates are programmed every several months, while the 'News' tab is constantly updated in sync with the PDB.

Results

As of June 2024, DIONYSUS contains over 330k carbohydrate moieties in interaction with proteins representing more than 4k of distinct sugar-like molecules. These interactions are found in ∼50k experimental structures involving 22k different proteins according to UniProt ID. Among those, ∼102k are glycosylation sites while ∼159k represent free ligand binding.

Web interface

DIONYSUS website provides five tools for database exploration, several examples of use in 'Help' page, methodological explanation and database statistics in 'About' and the used external databases in 'Resources'. The tools allow to: (i) search the database by protein sequence, structure or using various annotations, (ii) compare a binding site to all representatives found in DIONYSUS, (iii) explore clusters of CBS and (iv) analyze glycosylation patterns of a protein in different structures as compared to UniProt annotations. Finally, we provide information on the constantly updating database content at the 'News' section of the home page. DIONYSUS database organization and main possible ways of its exploration are provided in 'Help' with shortcuts to main functionalities implemented in the main page.

Search

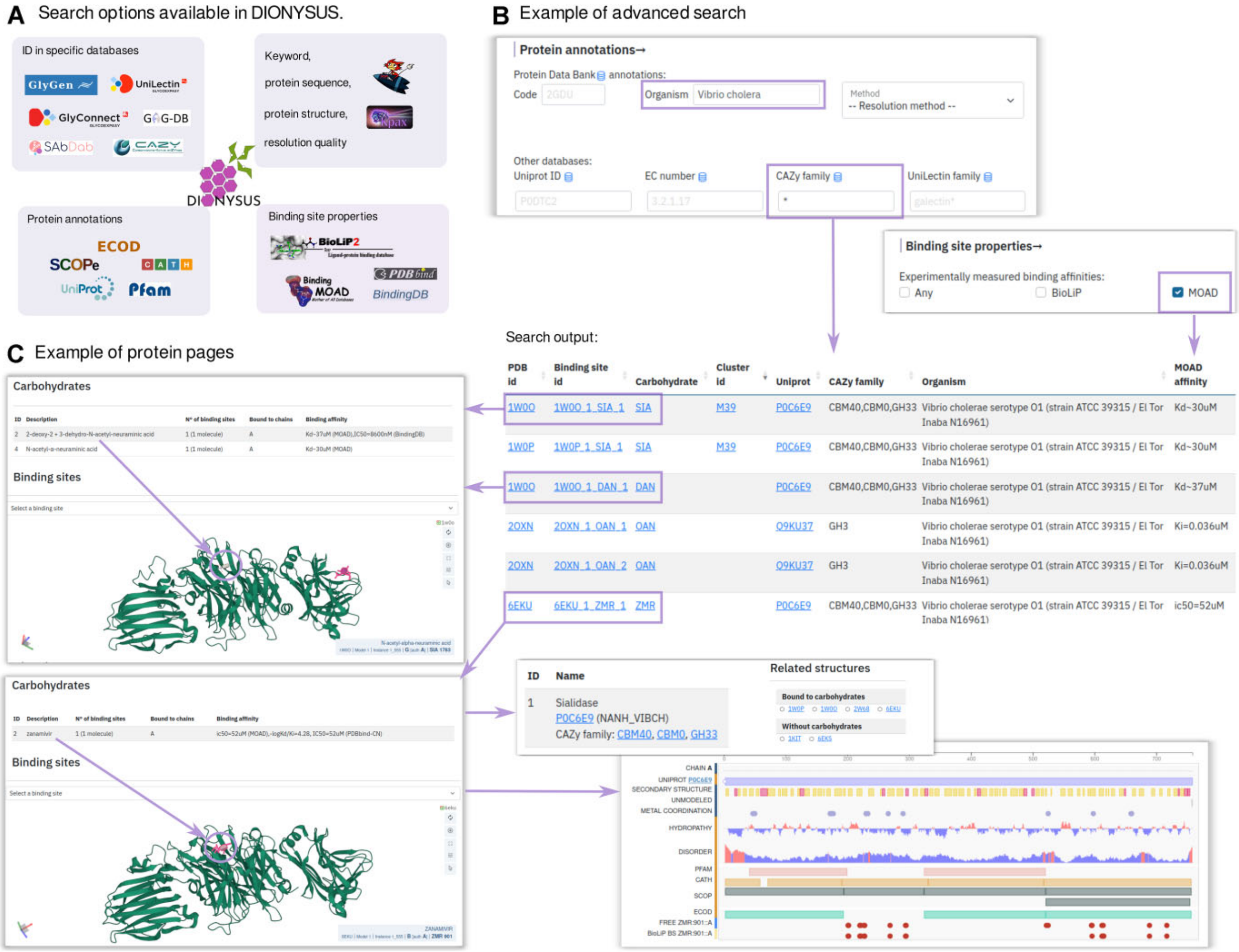
The user can both perform a quick search using a keyword or protein sequence/structure, and benefit from a more detailed advanced search (Figure 1). We provide four categories of parameters to search through the database: by protein, by binding site properties, by carbohydrate properties or according to data reliability. Protein properties include annotations extracted from the PDB and from the specialized databases as described in 'Materials and Methods' (Figure 1A). The binding site properties section provides search by physico-chemical properties, secondary structure content, presence (or absence) of other carbohydrates or ligands (e.g. ions) and availability of experimental CBS affinities. Moreover, the user can select a minimal number of protein atoms participating in the interface formation in order to filter out weak interactions. In the carbohydrate section, the user can query for specific carbohydrates, restraining the interface type (i.e. free ligand or glycosylation), and select only monosaccharides or carbohydrates with a specific chemical function. Interactive autocompletion is implemented in order to facilitate the choice of desired three-letter PDB code for a sugar ligand. Using 'Additional options', the user can make a more complex request by searching for PDB entries containing a desired combination of carbohydrate residues. The data reliability section allows the user to filter CBS based on different criteria such as missing atoms or mismatch between author and structural information on covalent bonds, as well as to select biologically relevant ligands according to BioLiP. In all text fields, the user can use the wildcard * for partial keywords. For example, a query for proteins from *Vibrio Cholerae* annotated in CAZy and having binding affinity information available in MOAD (Figure 1B) results in six different binding sites found in two different proteins (with distinct UniProt IDs) and crystallized in four different PDB entries. Interestingly, sialidase (UniProt ID: P0C6E9) was obtained in complexes with carbohydrates and carbohydrate mimetics bound to the catalytic site (Figure 1C, shown with an arrow) and/or to the lectin-like domain (Figure 1C top, highlighted in pink). Search results can be downloaded both in .csv, .pdb or .xls format as well as in the form of a .zip archive containing raw .mmCIF structures of the selected complexes.

Finally, the user can also eliminate redundancy in the search results according to different annotations, selecting for instance a single CBS per by UniProt ID or per DIONYSUS cluster. Moreover, the user can eliminate structures with several models, alternate locations or keep only structures with a single CBS.

Protein page

We provide multiple annotations for each protein chain of the PDB structure (Figure 1C). If a PDB entry contains several models (e.g. coming from NMR), the user can select each model using a drop-down menu. When the corresponding chain is bound to a carbohydrate, the user can select it to investigate individual chain properties according to different databases and to DIONYSUS annotation of the carbohydrate binding residues for each binding or glycosylation site (Figure 1C, right panel).

In the binding site section, the user can locate each CBS (per moiety) and recover information on its cluster, alternate locations if any, number of atoms and their SASA. In the 'Related structures' section, users can explore other resolved structures, which contain at least one chain with the same UniProt ID as the chain of interest. Those are grouped into two categories: structures containing a protein-carbohydrate interface (thus, annotated in DIONYSUS) and structures with no protein-



Protein annotations→

Protein Data Bank annotations:
Code 2GDU | Organism Vibrio cholera | Method -- Resolution method --

Other databases:
Uniprot ID P0DTC2 | EC number 3.2.1.17 | CAZy family * | UniLectin family galectin*

Binding site properties→

Experimentally measured binding affinities:
□ Any □ BioLiP ☑ MOAD

Search output:

PDB id	Binding site id	Carbohydrate	Cluster id	Uniprot	CAZy family	Organism	MOAD affinity
1W0O	1W0O_1_SIA_1	SIA	M39	P0C6E9	CBM40,CBM0,GH33	Vibrio cholerae serotype O1 (strain ATCC 39315 / El Tor Inaba N16961)	Kd~30uM
1W0P	1W0P_1_SIA_1	SIA	M39	P0C6E9	CBM40,CBM0,GH33	Vibrio cholerae serotype O1 (strain ATCC 39315 / El Tor Inaba N16961)	Kd~30uM
1W0O	1W0O_1_DAN_1	DAN		P0C6E9	CBM40,CBM0,GH33	Vibrio cholerae serotype O1 (strain ATCC 39315 / El Tor Inaba N16961)	Kd~37uM
2OXN	2OXN_1_OAN_1	OAN		Q9KU37	GH3	Vibrio cholerae serotype O1 (strain ATCC 39315 / El Tor Inaba N16961)	Ki=0.036uM
2OXN	2OXN_1_OAN_2	OAN		Q9KU37	GH3	Vibrio cholerae serotype O1 (strain ATCC 39315 / El Tor Inaba N16961)	Ki=0.036uM
6EKU	6EKU_1_ZMR_1	ZMR		P0C6E9	CBM40,CBM0,GH33	Vibrio cholerae serotype O1 (strain ATCC 39315 / El Tor Inaba N16961)	ic50=52uM

Carbohydrates

ID	Description	N° of binding sites	Bound to chains	Binding affinity
2	2-deoxy-2 + 3-dehydro-N-acetyl-neuraminic acid	1 (1 molecule)	A	Kd~37uM (MOAD),IC50=8600nM (BindingDB)
4	N-acetyl-α-neuraminic acid	1 (1 molecule)	A	Kd~30uM (MOAD)

Binding sites

Select a binding site

N-acetyl-alpha-neuraminic acid

Carbohydrates

ID	Description	N° of binding sites	Bound to chains	Binding affinity
2	zanamivir	1 (1 molecule)	A	ic50=52uM (MOAD),-logKd/Ki=4.28, IC50=52uM (PDBbind-CN)

Binding sites

Select a binding site

ZANAMIVIR

ID	Name
1	Sialidase P0C6E9 (NANH_VIBCH) CAZy family: CBM40, CBM0, GH33

Related structures

Bound to carbohydrates
○ 1W0P ○ 1W0O ○ 2W68 ○ 6EKU

Without carbohydrates
○ 1KIT ○ 6EKS

Figure 1. Different search options and examples of the search output. (**A**) Different databases linked to DIONYSUS and different types of database search implemented with logos of the corresponding engines (see Materials and methods for the details). (**B**) Example of the advanced search in DIONYSUS. Six binding sites were identified among all the protein-carbohydrate complexes expressed in *Vibrio cholera*, annotated in CAZy with available information on binding affinity in MOAD. (**C**) Protein pages of two sialidase complexes: with sialic acid (top, selected in pink), with Neu5Ac2en (top, pointed with arrow) and with a carbohydrate mimetic zanamivir (bottom, pointed with arrow). For PDB ID 6EKU we also provide additional information available on the protein pages such as identifiers in different databases, other structures of proteins sharing UniProt ID as well as sequence view with information on protein domain repartition and binding site location according to DIONYSUS and to BioLip.

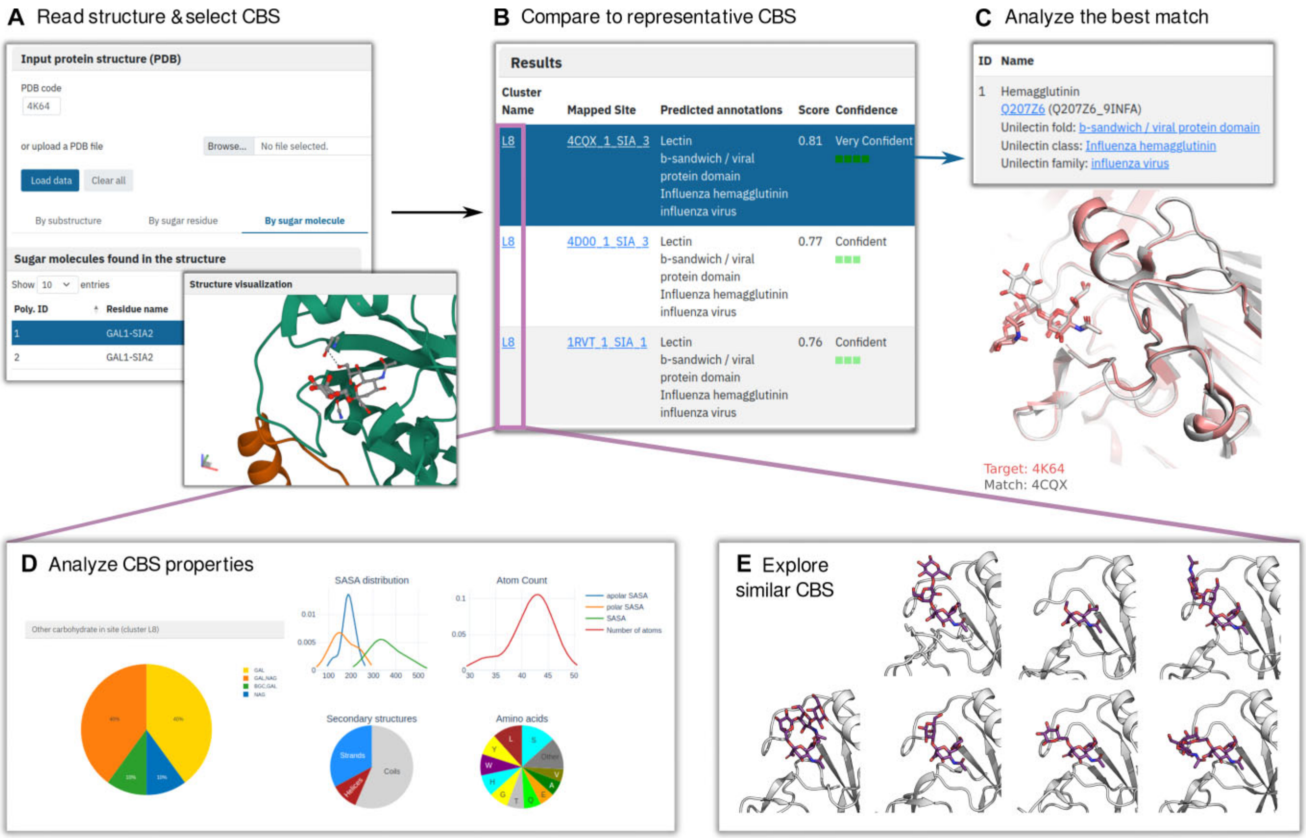
carbohydrate interface (with a link to the corresponding Protein Data Bank page).

Compare tool

A user can identify a cluster of similar protein-carbohydrate interfaces for a given PDB ID or for an uploaded PDB file. The submitted file is processed with an internal script and allows the user to select either a monosaccharide unit, or a complete carbohydrate ligand, as well as visualize the selection (Figure 2A). By clicking on 'Compare' button, the user can launch a two-step comparison process. First, we compare the CBS against all cluster representatives and select all clusters that scored higher than 0.5 (up to the top 5). Then, the target binding site is compared against every binding site from the selected clusters to refine our results. Finally, DIONYSUS ranks all the obtained scores and provides hyperlinks to the corresponding cluster pages for further exploration (Figure 2B) as well as a downloadable PyMOL session to investigate CBS superposition (Figure 2C, bottom). If a polysaccharide binding site is selected, the results are provided for each saccharide ring. For example, the best match for protein PDB ID 4K64 with no available annotation in UniLectin corresponds to sialic acid binding sites found in different hemagglutinin from Influenza (Figure 2C) suggesting its similar biological function.

Explore clusters

In the 'Clusters' page, we provide a t-SNE 2D projection of all CBS clusters depicted by a symbol of the most prevalent carbohydrate according to the glycan nomenclature (59,60). The symbol size is logarithmically proportional to the number of elements within the cluster (Supplementary Table S2). The user can separately visualize proteins of different functional families, as well as select clusters of different quality based on the ratio between the intra-cluster and the inter-cluster scores (for the details, see Supplementary Data). By



Input protein structure (PDB)

PDB code: 4K64

or upload a PDB file: Browse... No file selected.

Load data | Clear all

By substructure | By sugar residue | **By sugar molecule**

Sugar molecules found in the structure

Show 10 entries

Poly. ID	Residue name
1	GAL1-SIA2
2	GAL1-SIA2

Structure visualization

Results

Cluster Name	Mapped Site	Predicted annotations	Score	Confidence
L8	4CQX_1_SIA_3	Lectin b-sandwich / viral protein domain Influenza hemagglutinin influenza virus	0.81	Very Confident
L8	4D00_1_SIA_3	Lectin b-sandwich / viral protein domain Influenza hemagglutinin influenza virus	0.77	Confident
L8	1RVT_1_SIA_1	Lectin b-sandwich / viral protein domain Influenza hemagglutinin influenza virus	0.76	Confident

ID	Name
1	Hemagglutinin Q207Z6 (Q207Z6_9INFA) Unilectin fold: b-sandwich / viral protein domain Unilectin class: Influenza hemagglutinin Unilectin family: influenza virus

Figure 2. Example of comparison tool application to similar binding site search. (**A**) Compare tool identifies two carbohydrate molecules in the target structure (PDB ID: 4K64). We select the first one (shown in ‘Structure visualization’). (**B**) Results of comparison to representative CBS. The best match is obtained for the sialic acid binding site in the protein with PDB ID 4CQX. High scores are obtained for various proteins from cluster L8. (**C**) Information on the protein with the best match CBS and its superposition to the CBS in the target structure obtained using the downloadable PyMOL session. (**D**) Analysis of different properties of cluster L8. (**E**) Several other binding sites from cluster L8 in interaction with glycans of different composition. In accordance with panel D, most ligands bring galactose and/or *N*-acetyl-D-glucosamine residues in addition to sialic acid moiety.

clicking on each symbol, the user can access the page dedicated to the corresponding cluster for further exploration. The user can further explore such characteristics of CBS as carbohydrate properties, binding site properties and data reliability within a specific cluster (Figure 2D) as well as compare them to the average properties of CBS in our database. The page header provides essential information, including a concise summary of the most prevalent carbohydrates, protein functions, ECOD topologies, CAZy families, and UniLectin families/kingdom/folds (if any CBS within the cluster has such annotations). Furthermore, the user can directly visualize a representative CBS and corresponding protein structure and download a PyMOL session to explore the resulting local superposition of CBS (Figure 2E).

Glycosylation page

For a given UniProt ID, we provide all glycosylation patterns found in different resolved protein structures and in UniProt annotations. Glycosylations are depicted using the CPK color scheme with blue, red, and gray indicating N-linked, O-linked and C-linked glycans respectively. All other carbohydrate covalent binding sites are given in black. They correspond to either S-glycosylations, covalent intermediates or artifacts. The user can also hover over each dot to retrieve specific glycosylation annotations.

Discussion

DIONYSUS is the first open-access database gathering exhaustive annotations of all the resolved protein-carbohydrate interfaces and providing information on their properties at different levels. Unlike existing databases that focus on proteins primarily recognized for their carbohydrate-binding functions, we adopt an agnostic view of protein-carbohydrate interactions. This perspective enables us to identify carbohydrate-binding proteins that may lack proper annotations in current databases (as illustrated in Figure 2), as well as carbohydrate-binding proteins not typically included in specialized datasets, such as transporters and porins involved in carbohydrate transport. Additionally, growing evidence suggests that sugars play a role in general protein-protein interactions. For example, sugars seem to promote the dimerization of FSH (61) and play a crucial role in stimulation of the spike protein RBD in SARS-CoV-2 (62). Therefore, the presence and organization of carbohydrate binding sites can be crucial even for the proteins normally classified as having different functions than carbohydrate binding. Providing such information also falls within the key scope of DIONYSUS.

During the database construction we have taken into account a range of features specific for protein-carbohydrate interaction and, therefore, not taken into account in general protein-ligand interaction analysis. First, we correctly treat sugar binding sites formed by multiple chains, and clearly distinguish covalent and non-covalent protein-carbohydrate interactions. Then, we are also the first to systematically report data on the presence of other molecule types in carbohydrate binding sites, which can be crucial for the existence and stability of such interactions (for instance, calcium ions). Finally, we account for the intrinsic flexibility of the protein-carbohydrate interfaces by explicit consideration and interactive treatment of the NMR structures.

We retrieved all the available annotations for proteins referenced in the specialized databases and provide tools facilitating attribution of new annotations. For a given carbohydrate-binding protein, DIONYSUS search and comparison tools allow gaining information on the nature and specificity of the corresponding interactions at different levels. First, using sequence information only, one can identify protein homologs found in interaction with a carbohydrate by running sequence-based search. Next, using an experimental protein structure, or a deep learning-generated model, one can use a structure-based search to identify sugar-bound proteins sharing similar folds. Finally, for proteins experimentally characterized in complex with a carbohydrate, we offer the possibility to compare corresponding binding sites with common interaction patterns in our database to potentially identify their specificity.

The development of DIONYSUS provides an important contribution to the ongoing big data and artificial intelligence revolution in glycosciences (63). In recent years, we have observed an increased interest of the bioinformatics community to the problem of carbohydrate binding site and glycosylation site prediction using state-of-the-art deep learning methods (64–66). These tools have great potential in aiding annotation and analysis of various carbohydrate binding proteins. However, their development and performance evaluation crucially depends on quantity and quality of the available data. We believe that redundancy analysis and extensive annotations provided in DIONYSUS will contribute to further improvement of the existing methods and will allow us to better assess their current limitations.

Data availability

DIONYSUS is freely available to any user via the following link: www.dsimb.inserm.fr/DIONYSUS, and does not require any login or registration.

Supplementary data

Supplementary Data are available at NAR Online.

Funding

The study was performed in the framework of the SugarPred project funded by the French National Research Agency [grant no. ANR-21-CE45-0019] with the support of the Data Intelligence Institute of Paris (DiiP). A part of calculations was performed using high performance computing resources of TGCC (Très Grand Centre de Calcul) [grant no. A0140714131] funded by GENCI (Grand Equipement National de Calcul Intensif, France) as well as the IFB-core cluster (Institut Français de Bioinformatique, France). The authors also thank the Ministry of Research (France), Université Paris Cité (France), National Institute for Health and Medical Research (INSERM, France) and IdEx [grant no. ANR-18-IDEX-0001]. Funding for open access charge: French National Research Agency.

Conflict of interest statement

None declared.

References

1. He,X., Agnihotri,G. and Liu,H. (2000) Novel enzymatic mechanisms in carbohydrate metabolism. *Chem. Rev.*, **100**, 4615–4662.
2. Kannagi,R., Izawa,M., Koike,T., Miyazaki,K. and Kimura,N. (2004) Carbohydrate-mediated cell adhesion in cancer metastasis and angiogenesis. *Cancer Sci.*, **95**, 377–384.
3. Bendas,G. and Borsig,L. (2012) Cancer cell adhesion and metastasis: selectins, integrins, and the inhibitory potential of heparins. *Int. J. Cell Biol.*, **2012**, e676731.
4. Collins,B.E. and Paulson,J.C. (2004) Cell surface biology mediated by low affinity multivalent protein–glycan interactions. *Curr. Opin. Chem. Biol.*, **8**, 617–625.
5. Mythreye,K. and Blobe,G.C. (2009) Proteoglycan signaling co-receptors: roles in cell adhesion, migration and invasion. *Cell. Signal.*, **21**, 1548–1558.
6. Horacio,P. and Martinez-Noel,G. (2013) Sucrose signaling in plants: a world yet to be explored. *Plant Signal. Behav.*, **8**, e23316.
7. El Ghazal,R., Yin,X., Johns,S.C., Swanson,L., Macal,M., Ghosh,P., Zuniga,E.I. and Fuster,M.M. (2016) Glycan sulfation modulates dendritic cell biology and tumor growth. *Neoplasia*, **18**, 294–306.
8. Brabin,B.J., Romagosa,C., Abdelgalil,S., Menéndez,C., Verhoeff,F.H., McGready,R., Fletcher,K.A., Owens,S., d'Alessandro,U., Nosten,F., *et al.* (2004) The sick placenta—the role of malaria. *Placenta*, **25**, 359–378.
9. Brown,A. and Higgins,M.K. (2010) Carbohydrate binding molecules in malaria pathology. *Curr. Opin. Struct. Biol.*, **20**, 560–566.
10. Lin,B., Qing,X., Liao,J. and Zhuo,K. (2020) Role of protein glycosylation in host-pathogen interaction. *Cells*, **9**, 1022.
11. Sztain,T., Ahn,S.-H., Bogetti,A.T., Casalino,L., Goldsmith,J.A., Seitz,E., McCool,R.S., Kearns,F.L., Acosta-Reyes,F., Maji,S., *et al.* (2021) A glycan gate controls opening of the SARS-CoV-2 spike protein. *Nat. Chem.*, **13**, 963–968.
12. Kutzner,T.J., Gabba,A., FitzGerald,F.G., Shilova,N.V., García Caballero,G., Ludwig,A.-K., Manning,J.C., Knospe,C., Kaltner,H., Sinowatz,F., *et al.* (2019) How altering the modular architecture affects aspects of lectin activity: case study on human galectin-1. *Glycobiology*, **29**, 593–607.
13. García Caballero,G., Beckwith,D., Shilova,N.V., Gabba,A., Kutzner,T.J., Ludwig,A.-K., Manning,J.C., Kaltner,H., Sinowatz,F., Cudic,M., *et al.* (2020) Influence of protein (human galectin-3) design on aspects of lectin activity. *Histochem. Cell Biol.*, **154**, 135–153.
14. Bonnardel,F., Mariethoz,J., Salentin,S., Robin,X., Schroeder,M., Perez,S., Lisacek,F. and Imberty,A. (2019) UniLectin3D, a database of carbohydrate binding proteins with curated information on 3D structures and interacting ligands. *Nucleic Acids Res.*, **47**, D1236–D1244.
15. Cao,Y., Park,S.-J. and Im,W. (2021) A systematic analysis of protein–carbohydrate interactions in the Protein Data Bank. *Glycobiology*, **31**, 126–136.
16. Perez,S. and Makshakova,O. (2022) Multifaceted computational modeling in glycoscience. *Chem. Rev.*, **122**, 15914–15970.
17. Copoiu,L., Torres,P.H.M., Ascher,D.B., Blundell,T.L. and Malhotra,S. (2020) ProCarbDB: a database of carbohydrate-binding proteins. *Nucleic Acids Res.*, **48**, D368–D375.

18. Imberty,A. and Pérez,S. (2000) Structure, conformation, and dynamics of bioactive oligosaccharides: theoretical approaches and experimental validations. *Chem. Rev.*, **100**, 4567–4588.
19. Gajdos,L., Blakeley,M.P., Haertlein,M., Forsyth,V.T., Devos,J.M. and Imberty,A. (2022) Neutron crystallography reveals mechanisms used by Pseudomonas aeruginosa for host-cell binding. *Nat. Commun.*, **13**, 194.
20. Yang,J., Roy,A. and Zhang,Y. (2013) BioLiP: a semi-manually curated database for biologically relevant ligand–protein interactions. *Nucleic Acids Res.*, **41**, D1096–D1103.
21. Zhang,C., Zhang,X., Freddolino,P.L. and Zhang,Y. (2024) BioLiP2: an updated structure database for biologically relevant ligand–protein interactions. *Nucleic Acids Res.*, **52**, D404–D412.
22. Hu,L., Benson,M.L., Smith,R.D., Lerner,M.G. and Carlson,H.A. (2005) Binding MOAD (Mother Of All Databases). *Proteins Struct. Funct. Bioinforma.*, **60**, 333–340.
23. wwPDB consortium (2019) Protein Data Bank: the single global archive for 3D macromolecular structure data. *Nucleic Acids Res.*, **47**, D520–D528.
24. Sagendorf,J.M., Markarian,N., Berman,H.M. and Rohs,R. (2020) DNAproDB: an expanded database and web-based tool for structural analysis of DNA–protein complexes. *Nucleic Acids Res.*, **48**, D277–D287.
25. Berman,H.M., Westbrook,J., Feng,Z., Gilliland,G., Bhat,T.N., Weissig,H., Shindyalov,I.N. and Bourne,P.E. (2000) The Protein Data Bank. *Nucleic Acids Res.*, **28**, 235–242.
26. Wojdyr,M. (2022) GEMMI: a library for structural biology. *J. Open Source Softw.*, **7**, 4200.
27. Segura,J., Rose,Y., Westbrook,J., Burley,S.K. and Duarte,J.M. (2021) RCSB Protein Data Bank 1D tools and services. *Bioinformatics*, **36**, 5526–5527.
28. Guyon,F. and Moroy,G. (2023) Non-sequential alignment of binding sites for fast peptide screening. bioRxiv doi: https://doi.org/10.1101/2023.08.01.551496, 03 August 2023, preprint: not peer reviewed.
29. Mitternacht,S. (2016) FreeSASA: an open source C library for solvent accessible surface area calculations. *F1000Research*, **5**, 189.
30. Drula,E., Garron,M.-L., Dogan,S., Lombard,V., Henrissat,B. and Terrapon,N. (2022) The carbohydrate-active enzyme database: functions and literature. *Nucleic Acids Res.*, **50**, D571–D577.
31. Bonnardel,F., Perez,S., Lisacek,F. and Imberty,A. (2020) Structural database for lectins and the uniLectin web platform. In: Hirabayashi,J. (ed.) *Lectin Purification and Analysis: Methods and Protocols*. Methods in Molecular Biology. Springer US, NY, pp. 1–14.
32. Dunbar,J., Krawczyk,K., Leem,J., Baker,T., Fuchs,A., Georges,G., Shi,J. and Deane,C.M. (2014) SAbDab: the structural antibody database. *Nucleic Acids Res.*, **42**, D1140–D1146.
33. Bonnardel,F., Mariethoz,J., Pérez,S., Imberty,A. and Lisacek,F. (2021) LectomeXplore, an update of UniLectin for the discovery of carbohydrate-binding proteins based on a new lectin classification. *Nucleic Acids Res.*, **49**, D1548–D1554.
34. The UniProt Consortium (2023) UniProt: the Universal Protein Knowledgebase in 2023. *Nucleic Acids Res.*, **51**, D523–D531.
35. Pérez,S., Bonnardel,F., Lisacek,F., Imberty,A., Ricard Blum,S. and Makshakova,O. (2020) GAG-DB, the new interface of the three-dimensional landscape of glycosaminoglycans. *Biomolecules*, **10**, 1660.
36. Cordero,B., Gómez,V., Platero-Prats,A.E., Revés,M., Echeverría,J., Cremades,E., Barragán,F. and Alvarez,S. (2008) Covalent radii revisited. *Dalton Trans.*, **21**, 2832–2838.
37. The Gene Ontology Consortium, Aleksander,S.A., Balhoff,J., Carbon,S., Cherry,J.M., Drabkin,H.J., Ebert,D., Feuermann,M., Gaudet,P., Harris,N.L., *et al.* (2023) The Gene Ontology knowledgebase in 2023. *Genetics*, **224**, iyad031.
38. Gasteiger,E., Gattiker,A., Hoogland,C., Ivanyi,I., Appel,R.D. and Bairoch,A. (2003) ExPASy: the proteomics server for in-depth protein knowledge and analysis. *Nucleic Acids Res.*, **31**, 3784–3788.
39. Chang,A., Jeske,L., Ulbrich,S., Hofmann,J., Koblitz,J., Schomburg,I., Neumann-Schaal,M., Jahn,D. and Schomburg,D. (2021) BRENDA, the ELIXIR core data resource in 2021: new developments and updates. *Nucleic Acids Res.*, **49**, D498–D508.
40. Mistry,J., Chuguransky,S., Williams,L., Qureshi,M., Salazar,G.A., Sonnhammer,E.L.L., Tosatto,S.C.E., Paladin,L., Raj,S., Richardson,L.J., *et al.* (2021) Pfam: the protein families database in 2021. *Nucleic Acids Res.*, **49**, D412–D419.
41. Andreeva,A., Howorth,D., Chothia,C., Kulesha,E. and Murzin,A.G. (2015) Investigating protein structure and evolution with SCOP2. *Curr. Protoc. Bioinforma.*, **49**, 1.26.1–1.26.21.
42. Sillitoe,I., Bordin,N., Dawson,N., Waman,V.P., Ashford,P., Scholes,H.M., Pang,C.S.M., Woodridge,L., Rauer,C., Sen,N., *et al.* (2021) CATH: increased structural coverage of functional space. *Nucleic Acids Res.*, **49**, D266–D273.
43. Cheng,H., Schaeffer,R.D., Liao,Y., Kinch,L.N., Pei,J., Shi,S., Kim,B.-H. and Grishin,N.V. (2014) ECOD: an evolutionary classification of protein domains. *PLOS Comput. Biol.*, **10**, e1003926.
44. Schaeffer,R.D., Liao,Y., Cheng,H. and Grishin,N.V. (2017) ECOD: new developments in the evolutionary classification of domains. *Nucleic Acids Res.*, **45**, D296–D302.
45. Lombard,V., Golaconda Ramulu,H., Drula,E., Coutinho,P.M. and Henrissat,B. (2014) The carbohydrate-active enzymes database (CAZy) in 2013. *Nucleic Acids Res.*, **42**, D490–D495.
46. York,W.S., Mazumder,R., Ranzinger,R., Edwards,N., Kahsay,R., Aoki-Kinoshita,K.F., Campbell,M.P., Cummings,R.D., Feizi,T., Martin,M., *et al.* (2020) GlyGen: computational and informatics resources for glycoscience. *Glycobiology*, **30**, 72–73.
47. Alocci,D., Mariethoz,J., Gastaldello,A., Gasteiger,E., Karlsson,N.G., Kolarich,D., Packer,N.H. and Lisacek,F. (2019) GlyConnect: glycoproteomics goes visual, interactive, and analytical. *J. Proteome Res.*, **18**, 664–677.
48. Wang,R., Fang,X., Lu,Y., Yang,C.-Y. and Wang,S. (2005) The PDBbind database: methodologies and updates. *J. Med. Chem.*, **48**, 4111–4119.
49. Gilson,M.K., Liu,T., Baitaluk,M., Nicola,G., Hwang,L. and Chong,J. (2016) BindingDB in 2015: a public database for medicinal chemistry, computational chemistry and systems pharmacology. *Nucleic Acids Res.*, **44**, D1045–D1053.
50. Rasolohery,I., Moroy,G. and Guyon,F. (2017) PatchSearch: a fast computational method for off-target detection. *J. Chem. Inf. Model.*, **57**, 769–777.
51. Rey,J., Rasolohery,I., Tufféry,P., Guyon,F. and Moroy,G. (2019) PatchSearch: a web server for off-target protein identification. *Nucleic Acids Res.*, **47**, W365–W372.
52. Steinegger,M. and Söding,J. (2017) MMseqs2 enables sensitive protein sequence searching for the analysis of massive data sets. *Nat. Biotechnol.*, **35**, 1026–1028.
53. Thomès,L., Burkholz,R. and Bojar,D. (2021) Glycowork: a Python package for glycan data science and machine learning. *Glycobiology*, **31**, 1240–1244.
54. Shi,J. and Malik,J. (2000) Normalized cuts and image segmentation. *IEEE Trans. PATTERN Anal. Mach. Intell.*, **22**, 888–905.
55. Ng,A., Jordan,M. and Weiss,Y. (2001) On Spectral Clustering: analysis and an algorithm. In: *Advances in Neural Information Processing Systems*. MIT Press, Vol. **14**.
56. Sánchez-García,R.J., Fennelly,M., Norris,S., Wright,N., Niblo,G., Brodzki,J. and Bialek,J.W. (2014) Hierarchical Spectral Clustering of Power Grids. *IEEE Trans. Power Syst.*, **29**, 2229–2237.
57. Pearson,W.R. (2016) Finding protein and nucleotide similarities with FASTA. *Curr. Protoc. Bioinforma.*, **53**, 3.9.1–3.9.25.
58. Ritchie,D.W. (2016) Calculating and scoring high quality multiple flexible protein structure alignments. *Bioinformatics*, **32**, 2650–2658.

59. Varki,A., Cummings,R.D., Aebi,M., Packer,N.H., Seeberger,P.H., Esko,J.D., Stanley,P., Hart,G., Darvill,A., Kinoshita,T., *et al.* (2015) Symbol nomenclature for graphical representations of glycans. *Glycobiology*, **25**, 1323–1324.
60. Neelamegham,S., Aoki-Kinoshita,K., Bolton,E., Frank,M., Lisacek,F., Lütteke,T., O'Boyle,N., Packer,N.H., Stanley,P., Toukach,P., *et al.* (2019) Updates to the symbol nomenclature for glycans guidelines. *Glycobiology*, **29**, 620–624.
61. Fox,K.M., Dias,J.A. and Van Roey,P. (2001) Three-dimensional structure of human follicle-stimulating hormone. *Mol. Endocrinol.*, **15**, 378–389.
62. Díaz-Salinas,M.A., Jain,A., Durham,N.D. and Munro,J.B. (2024) Single-molecule imaging reveals allosteric stimulation of SARS-CoV-2 spike receptor binding domain by host sialic acid. *Sci. Adv.*, **10**, eadk4920.
63. Bojar,D. and Lisacek,F. (2022) Glycoinformatics in the artificial intelligence era. *Chem. Rev.*, **122**, 15971–15988.
64. Canner,S.W., Shanker,S. and Gray,J.J. (2023) Structure-based neural network protein–carbohydrate interaction predictions at the residue level. *Front. Bioinforma.*, **3**, 1186531.
65. Bibekar,P., Krapp,L. and Peraro,M.D. (2024) PeSTo-Carbs: geometric deep learning for prediction of protein–carbohydrate binding interfaces. *J. Chem. Theory Comput.*, **20**, 2985–2991.
66. He,X., Zhao,L., Tian,Y., Li,R., Chu,Q., Gu,Z., Zheng,M., Wang,Y., Li,S., Jiang,H., *et al.* (2024) Highly accurate carbohydrate-binding site prediction with DeepGlycanSite. *Nat. Commun.*, **15**, 5163.

Received: July 3, 2024. **Revised:** September 3, 2024. **Editorial Decision:** September 25, 2024. **Accepted:** September 26, 2024