Published online 26 May 2015 *Nucleic Acids Research, 2015, Vol. 43, Web Server issue* ***W383–W388***
doi: 10.1093/nar/gkv561

PatternQuery: web application for fast detection of biomacromolecular structural patterns in the entire Protein Data Bank

David Sehnal[1,2,3,†], Lukáš Pravda[1,2,†], Radka Svobodová Vařeková[1,2], Crina-Maria Ionescu[1] and Jaroslav Koča[1,2,*]

[1]CEITEC - Central European Institute of Technology, Masaryk University Brno, Kamenice 5, 625 00 Brno, Czech Republic, [2]National Centre for Biomolecular Research, Faculty of Science, Masaryk University, Kotlářská 2, 611 37 Brno, Czech Republic and [3]Faculty of Informatics, Masaryk University Brno, Botanická 68a, 602 00 Brno, Czech Republic

Received March 13, 2015; Revised May 01, 2015; Accepted May 17, 2015

ABSTRACT

Well defined biomacromolecular patterns such as binding sites, catalytic sites, specific protein or nucleic acid sequences, etc. precisely modulate many important biological phenomena. We introduce PatternQuery, a web-based application designed for detection and fast extraction of such patterns. The application uses a unique query language with Python-like syntax to define the patterns that will be extracted from datasets provided by the user, or from the entire Protein Data Bank (PDB). Moreover, the database-wide search can be restricted using a variety of criteria, such as PDB ID, resolution, and organism of origin, to provide only relevant data. The extraction generally takes a few seconds for several hundreds of entries, up to approximately one hour for the whole PDB. The detected patterns are made available for download to enable further processing, as well as presented in a clear tabular and graphical form directly in the browser. The unique design of the language and the provided service could pave the way towards novel PDB-wide analyses, which were either difficult or unfeasible in the past. The application is available free of charge at http://ncbr.muni.cz/PatternQuery.

INTRODUCTION

In the past years an overwhelming volume of biomacromolecular structures have been deposited in the worldwide deposition system Protein Data Bank (PDB) (1). The amount of data which was available 20 years ago is nowadays released every week, and this rapid pace is maintained. Small high-resolution protein structures are deposited, as well as extensive ribosomes or viral capsids. The whole scientific community can benefit from this abundance of biomacromolecular structures, being enabled to carry out experiments and analyses which were not feasible before (2–4). Such richness of 3D data accents the immense need for structural bioinformatics tools and services to help in reasoning out a variety of structural properties, which often go hand in hand with biological function.

Presently, various computational tools and frameworks exist for the definition of molecular (sub)structure, such as SMILES (5), MQL (6), or SLN (7), which are mainly focused on small organic compounds. There are also tools that enable the definition and analysis of more general structural patterns, some of which rely on an internal molecular language (8–14). A structural pattern can, in principle, be any part of a biomacromolecule, i.e. protein backbone, ligands or metals together with their binding sites or surroundings, specific amino acids or nucleotide sequences, and sets of atoms or residues satisfying given criteria (distance, composition, intramolecular connectivity, etc.). Nevertheless, these tools are designed to operate either on a low number of structures, or their functionality is focused on very specific and narrow applications. Furthermore, some of the most popular services and databases use structure information for defining or inferring structure-function relationships (15,16). Even critical interaction sites are defined at the primary and secondary structure level (17,18), mainly because of the large structural variation of biomacromolecules. Ultimately, to our knowledge, there is no tool available for the general and systematic description and extraction of 3D structural patterns from biomacromolecules tailored for the mining of structural databases.

In this article, we address the general philosophy of describing 3D structural patterns, and present an approach

*To whom correspondence should be addressed. Tel: +420 54949 4947; Fax: +420 54949 2556; Email: Jaroslav.Koca@ceitec.muni.cz
†These authors contributed equally to the paper as first authors.

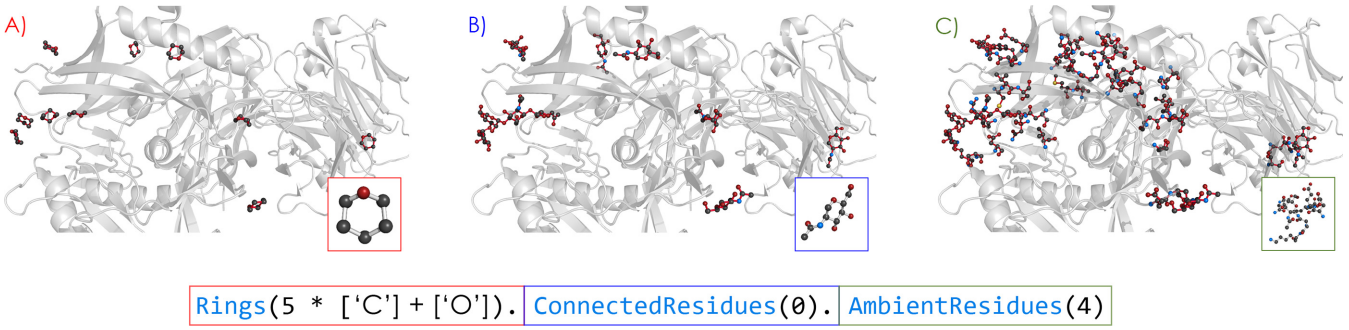

Figure 1. The query recognizes the binding pocket of any residue containing a pyranose moiety in the envelope glycoprotein gp160 from *Human immunodeficiency virus 1* in complex with *Homo sapiens* immunoglobulins (3u7y). One of the recognized patterns is highlighted in the box. (**A**) First, the query identifies a pyranose moiety (a ring composed of 5 carbons and an oxygen atom). (**B**) Then, all residues which include this pattern in their structure are identified. (**C**) Finally, all the residues that are at most 4Å from any of the pyranose containing residues are detected as well. This ensures all the potential coordination partners are recognized properly. The molecules were visualized using PyMOL.

for their effective identification and extraction from individual biomacromolecules, as well as from the PDB archive. This approach is implemented as the user-friendly web service PatternQuery (PQ). The service is built on a simple yet powerful language for the description of any molecular structural patterns based on the nature and relationship between atoms, residues and other structural elements. The unique design of PQ allows the user to simultaneously operate at the primary, secondary and tertiary level of biomacromolecular structure.

The results provided by PQ can serve as a source of input data in further analyses, such as structural and functional assignment of uncharacterized proteins, analysis of newly determined structures, comparative structural analysis, design and engineering of novel functional sites, etc.

DESCRIPTION OF THE TOOL

PatternQuery is an interactive web application for the optimal definition of biomacromolecular structural patterns, followed by their fast detection and extraction from the entire PDB or user defined datasets. These patterns are described by unique expressions based on the Python programming language, which are designed to define biomacromolecular structural patterns based on the nature and relationship between atoms, residues and other structural elements. These expressions define the composition, topology, connectivity, and 3D structure of a pattern. By composing these expressions into a query, 3D structural patterns can be identified inside biomacromolecules. Figure 1 gives the PQ query example that identifies and extracts a 3D pattern made up of a residue containing a pyranose moiety, together with its immediate surroundings.

The PatternQuery application can be used in two modes. The PQ Explorer mode (Figure 2) is useful for real-time investigation of smaller datasets (either user-uploaded or a small subset of the PDB), and tuning the queries prior to searching the whole PDB. The PQ Service mode (Supplementary Figure S1) is optimized for querying the entire PDB archive. Finally, a command-line version of the PQ application is available for processing in-house databases of 3D biomacromolecular structure data.

The PQ web pages contain several interactive guides, which explain the features and give an easy walkthrough the application, along with plenty of tips. Rich documentation is provided as well, in the form of a Wiki user manual with many examples.

PatternQuery workflow

The procedure of using the PatternQuery application involves four steps: (i) query definition; (ii) input data specification; (iii) running the PQ query; (iv) visualization and analysis of retrieved patterns.

(i) Query definition

First, it is necessary to build a query that optimally describes the structural pattern(s) of interest. The PatternQuery language is well documented, and its usage is richly illustrated on many examples and several case studies. Detailed knowledge of the language is not required, since the integrated high performance coding editor (ACE, http://ace.c9.io) provides syntax suggestions and relevant query examples. Multiple queries can be defined for a single run.

(ii) Input data specification

Second, the queried data has to be specified. Small subsets of the PDB, or the user's own datasets can be queried in the PQ Explorer mode. Large custom databases can be queried using the command line version of PatternQuery. In the PQ Service mode, the default queried dataset is a weekly updated mirror of the latest release of the PDB stored in the PDBx/mmCIF file format. Alternatively, a subset of the PDB can be specified based on a list of PDB entry IDs, or on various metadata criteria. By specifying a subset of the PDB as input, it is ensured that only patterns from relevant structures are retrieved, and the query can be executed in a more time efficient manner. For example one may restrict the search only to biomacromolecules including a DNA chain from *Homo sapiens,* determined by X-ray diffraction of resolution better than 2Å and published in the past 3 years.

Optionally, all the patterns identified while running PQ may be subjected to the structural validation of annotation (19). During this process, which is briefly described in the supplementary materials in the section SI Structure Validation, all ligands and non-standard residues larger than six heavy atoms are inspected for their completeness and chirality correctness. Possible discrepancies or structural inconsistencies are highlighted. This may aid further processing of the results by discarding low-quality patterns.

(iii) Running the PQ query

After setup, the specified data set is queried with all the defined PQ expressions. This process involves generating the structure's internal representation, together with proper bond identification based on the intramolecular atomic distances, and then attempting to match the PQ query with any suitable substructure. The theoretical framework behind this process is given in the supplementary materials (Theoretical Background section).
Depending on the complexity of the defined queries and the number of dataset entries, running the queries may take from a few seconds (for a few hundred small to medium-sized entries), up to approximately one hour for 100 000 PDB entries. Most types of queries have $O(N)$ or $O(N \log N)$ time complexity (where N is the number of atoms in the structure), meaning that doubling the number of structures being processed will roughly double the running time. A benchmark of the application is available in the supplementary materials (section Performance Overview).

(iv) Visualization and analysis of retrieved patterns

The PQ results consist of structure files with the patterns, and statistics about their origin and composition. All the results are made available for inspection or download under a unique web address for at least a month, in both the PQ Explorer and PQ Service modes.

The PatternQuery output provides a straightforward and rich report in both tabular and graphical form, including summary and detailed information about each pattern identified. The summary includes the number of detected patterns and PDB entries that the patterns were extracted from, together with possible errors and warnings, often caused by discrepancies either in the biomacromolecular structure, or in the file format. The detailed report provides a pattern view, focused on each individual pattern identified, and a PDB entry view, focused on each PDB entry queried. Additionally, in the PDB entry view, the results for all patterns identified in that particular PDB entry can be accessed together.

Useful statistics in the form of the atom and residue composition are given for each extracted pattern, along with all the metadata from the parent data set entry (PDB entry). These can serve for further filtering of interesting results. Each extracted pattern can be visualized interactively (ChemDoodle, http://www.chemdoodle.com). Optionally, the validation report can be readily accessed.

Limitations

The setup of the PatternQuery web application, particularly in the PQ Service mode, is limited to 10 queries to be executed during a single run. The maximum number of results that can be returned by a single query execution on our server is one million patterns or ten million atoms, whichever is reached first. This limitation is not present in the command line tool. Additional limitations are discussed in detail in the supplementary materials (Limitations section).

RESULTS AND DISCUSSION

We provide two case studies, which demonstrate the possible usage of the PatternQuery web application. Additional biologically relevant examples, together with the corresponding PQ queries, are available on our wiki pages. All the queries used in the case studies can be found in the supplementary materials.

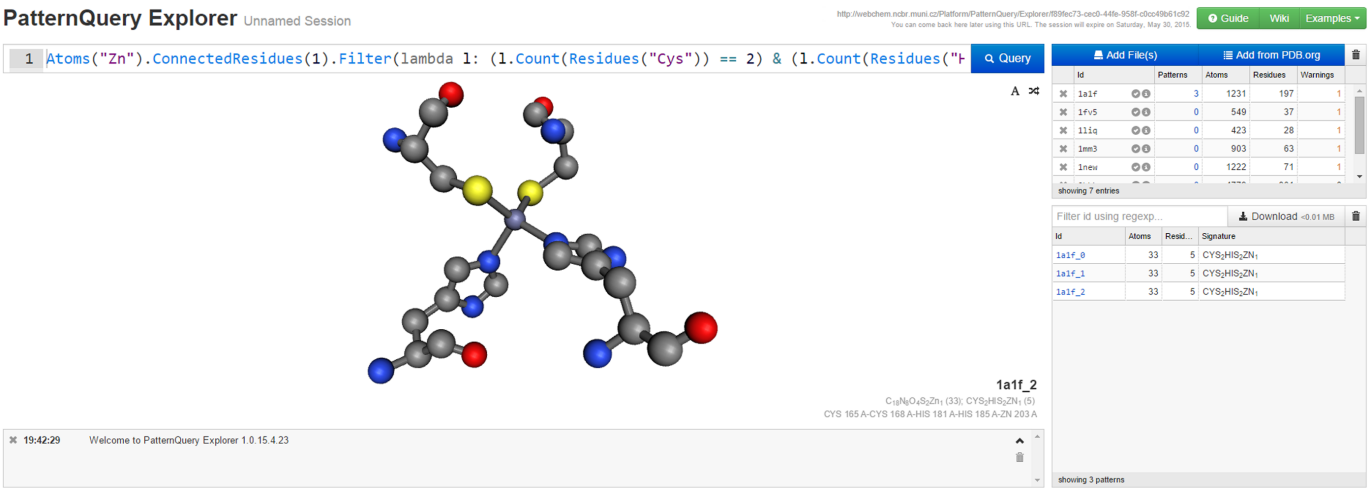

Figure 2. The PatternQuery Explorer mode is tailored for querying smaller user-defined datasets (up to 100 entries) uploaded in one of the supported formats. Additionally, a subset of the PDB archive can be queried as well, based on PDB ID or a variety of metadata.

Case study I - LecB sugar binding sites

Pseudomonas aeruginosa is an opportunistic pathogen associated with a number of chronic infections. This pathogen forms a biofilm enabling it to survive both the response of the host immune system, and antibiotic treatment (20). One of the cornerstones of biofilm formation, in the case of *P. aeruginosa,* is the presence of sugar-binding proteins on the outer cell membrane — LecA (PA-IL) and LecB (PA-IIL). Their inhibition is considered to be a promising approach for anti-pseudomonadal treatment (21).

LecB binds with the highest affinity to L-fucosides and D-mannosides (22), however, other monosaccharides are recognized as well (23). The sugar-binding domain is calcium dependent, with two calcium ions stabilizing the binding site. We employed PQ in the discovery of sugar binding sites of similar geometry as the tetrameric LecB entry in the PDB. Specifically, we have searched for 2 calcium ions at most 4Å apart, and all the residues with direct interaction with either of these ions. Furthermore, just the molecular patterns containing a residue with a furan or pyran ring were preserved. The complete PQ query which identifies such patterns is given as SI Query 1. Due to the fact that the sugar-binding domain is calcium dependent, we were able to restrict the search only to the biomacromolecules having a calcium ion in their structure, and containing a pyranose or furanose moiety (3074 PDB entries as of 25.4.2015), which tremendously reduced query-running time. The initial analysis of the PDB archive revealed 355 different patterns originating from 231 PDB entries. However, the majority of the sugar moieties originated from nucleotides. To filter them out, a simple filter was employed (SI Query 2), which provided 108 distinct patterns originating from 36 PDB entries of 7 different organisms. The majority of them originated from *P. aeruginosa*, however other pathogens such as *R. solanacearum*, *B. cenocepacia* or *C. violaceum* were identified among the organisms of origin. The sugar-binding domain in 87 of the patterns are composed of 3x Asp, 2x Asn and Glu and Gly residues, which is the binding site referred to as the sugar binding motif in the literature (24) for a total of 24 PDB entries from 3 organisms. In 12 further patterns a glycine residue was not present due to the fact that the structure stored in the PDB is only the asymmetric unit, rather than the expected biological unit, which is a tetramer. Finally, the remaining 9 patterns, originating from 6 different pectate lyase (EC: 4.2.2.2) structures, exhibited a different binding motif in comparison to the LecB protein. These patterns contained α-D-galactopyranuronic acid and its derivatives rather than a fucose or mannose derivative. A detailed list of these sugar ligands is given in the Supplementary Tables S1 and S2.

Finally, the quality of the 3D structure of the patterns was examined. A total of 9 patterns originating from 3 PDB entries exhibited a serious structural issue, i.e. half of the α-L-fucose ligands in complex with the 1oxc PDB entry exhibit incorrect chirality at the C1 carbon atom. The details of this analysis can be found in the supplementary materials (SI Query Validation 1).

Case study II - C_2H_2 zinc fingers

The class of zinc finger DNA-binding proteins is the most abundant across all biology (25). They fulfill a remarkable range of diverse functions, including DNA recognition, transcriptional activation, regulation of apoptosis or lipid binding (25). Due to their specificity and modular architecture, they often serve as a rational engineering target for binding a wide range of DNA sequences to activate, repress, cut or paste genes (26). The classical C_2H_2 zinc finger domain is composed of a simple ββα fold, which is stabilized by a zinc ion coordinated by two histidine and two cysteine residues. The fold is often described by the pattern of X_2-C-X_{2-4}-C-X_{12}-H-X_{3-5}-H, where X stands for any amino acid, C is cysteine and H is histidine. Nevertheless, atypical variations also exist, which differ from the consensus profile (27) (e.g. UniProt ID (28): P47043). The X_{12} region of the consensus profile is usually further decomposed into the sequence X_3-[F|Y]-X_5-Ψ-X_2, where [F|Y] represents either a phenylalanine or tyrosine residue, and Ψ denotes a hydrophobic residue (29).

We have queried the whole PDB archive (access date 25.4.2015) using several different PQ queries. At first, we searched just for patterns with primary sequences which satisfy the basic consensus profile of the typical C_2H_2 zinc finger domain, without further specification of the X_{12} region (SI Query 3). We identified 595 patterns in 342 different PDB entries. The results of such a query will inevitably be plagued by a number of false positive hits, i.e. patterns satisfying the primary sequence criteria, but which are not zinc fingers. This is due to the fact that no further checks for the presence of a zinc ion or stabilizing residues were included in the query. Closer inspection of the results revealed that the above-defined primary sequence corresponds not only to the C_2H_2 zinc finger fold, but also to a variety of fumarate reductases and hydrolases. In order to filter out these false positive hits, we adjusted the query so that the pattern must contain a zinc ion stabilized by two cysteine and two histidine residues from the consensus profile (SI Query 4). This final query resulted in 461 different patterns originating from 278 PDB entries. The majority of the results (356 patterns in 239 PDB entries) also satisfied the special pattern of the X_{12} region between the second cysteine and the first histidine (SI Query 5). The largest number of structures was isolated from Eukaryotes, mainly *Homo sapiens,* and determined by solution nuclear magnetic resonance spectroscopy. However, a few structures originating from viruses and bacteria were found as well. No residues relevant for validation were detected inside the input patterns, and therefore validated.

Furthermore, it has been reported that the zinc finger fold may also be stabilized by other metals (30). We have modified the query so that possible substitutions of zinc with other metals can also be considered (SI Query 6). Running this query returned five additional patterns from two PDB entries, where the zinc ion was substituted by cobalt (31) and cadmium (32). Although the cobalt-binding protein contains 5 zinc finger domains, just 4 patterns were identified, due to the alternate primary sequence in one of the patterns. These primary sequence modifications can be ac-

counted for by modifying the regular expression in the PQ query.

CONCLUSION

In this article, we presented PatternQuery, a novel web application for rapid definition and extraction of 3D structural patterns from the entire PDB. The web application is easy to use and platform-independent. Results are presented in a clear graphical and tabular form. Rich documentation regarding both the underlying language and the features of the web application, along with several biologically relevant case studies are available at http://ncbr.muni.cz/PatternQuery.

The innovative approach described in the present study enables mining large databases (entire PDB or in-house structural databases), a task which was unfeasible in the past, or was difficult for patterns with more complex structure.

SUPPLEMENTARY DATA

Supplementary Data are available at NAR Online.

FUNDING

This work was funded by the Ministry of Education,Youth and Sports of the Czech Republic [contract number LH13055], the European Community's Seventh Framework Programme [CZ.1.05/1.1.00/02.0068] from the European Regional Development Fund, and the Grant Agency of the Czech Republic [14-29577S]. Support to CMI was provided via the project "Employment of Newly Graduated Doctors of Science for Scientific Excellence" [CZ.1.07/2.3.00/30.0009], co-financed from the European Social Fund and the state budget of the Czech Republic. Funding for open access charge: Institutional budget of the National Centre for Biomolecular Research, Masaryk University, Czech Republic.

Conflict of interest statement. None declared.

REFERENCES

1. Gutmanas,A., Alhroub,Y., Battle,G.M., Berrisford,J.M., Bochet,E., Conroy,M.J., Dana,J.M., Fernandez Montecelo,M.A., van Ginkel,G., Gore,S.P. *et al.* (2014) PDBe: Protein Data Bank in Europe. *Nucleic Acids Res.*, **42**, D285–D291.
2. Smith,K.P., Gifford,K.M., Waitzman,J.S. and Rice,S.E. (2014) Survey of phosphorylation near drug binding sites in the Protein Data Bank (PDB) and their effects.*Proteins Struct. Funct. Bioinforma.*, **83**, 25–36.
3. Gavenonis,J., Sheneman,B.A., Siegert,T.R., Eshelman,M.R. and Kritzer,J.A. (2014) Comprehensive analysis of loops at protein-protein interfaces for macrocycle design. *Nat. Chem. Biol.*, **10**, 1–8.
4. Steinkellner,G., Gruber,C.C., Pavkov-Keller,T., Binter,A., Steiner,K., Winkler,C., Lyskowski,A., Schwamberger,O., Oberer,M., Schwab,H. *et al.* (2014) Identification of promiscuous ene-reductase activity by mining structural databases using active site constellations. *Nat. Commun.*, **5**, 4150.
5. Weininger,D. (1988) SMILES, a chemical language and information system. 1. Introduction to methodology and encoding rules. *J. Chem. Inf. Comput. Sci.*, **28**, 31–36.
6. Proschak,E., Wegner,J.K., Schüller,A., Schneider,G. and Fechner,U. (2007) Molecular query language (MQL)–a context-free grammar for substructure matching. *J. Chem. Inf. Model.*, **47**, 295–301.
7. Homer,R.W., Swanson,J., Jilek,R.J., Hurst,T. and Clark,R.D. (2008) SYBYL line notation (SLN): A single notation to represent chemical structures, queries, reactions, and virtual libraries. *J. Chem. Inf. Model.*, **48**, 2294–2307.
8. Humphrey,W., Dalke,A. and Schulten,K. (1996) VMD: Visual molecular dynamics. *J. Mol. Graph.*, **14**, 33–38.
9. Biasini,M., Schmidt,T., Bienert,S., Mariani,V., Studer,G., Haas,J., Johner,N., Schenk,A.D., Philippsen,A. and Schwede,T. (2013) OpenStructure: an integrated software framework for computational structural biology. *Acta Crystallogr. D. Biol. Crystallogr.*, **69**, 701–709.
10. Kalev,I., Mechelke,M., Kopec,K.O., Holder,T., Carstens,S. and Habeck,M. (2012) CSB: a Python framework for structural bioinformatics. *Bioinformatics*, **28**, 2996–2997.
11. *The PyMOL Molecular Graphics System*. Version 1.7.4, Schrödinger, LLC.
12. Täubig,H., Buchner,A. and Griebsch,J. (2006) PAST: fast structure-based searching in the PDB. *Nucleic Acids Res.*, **34**, W20–W23.
13. Nadzirin,N., Willett,P., Artymiuk,P.J. and Firdaus-Raih,M. (2013) IMAAAGINE: a webserver for searching hypothetical 3D amino acid side chain arrangements in the Protein Data Bank. *Nucleic Acids Res.*, **41**, W432–W440.
14. Samson,A.O. and Levitt,M. (2009) Protein segment finder: an online search engine for segment motifs in the PDB. *Nucleic Acids Res.*, **37**, D224–D228 .
15. Andreeva,A., Howorth,D., Chothia,C., Kulesha,E. and Murzin,A.G. (2014) SCOP2 prototype: a new approach to protein structure mining. *Nucleic Acids Res.*, **42**, D310–D314.
16. Sillitoe,I., Lewis,T.E., Cuff,A., Das,S., Ashford,P., Dawson,N.L., Furnham,N., Laskowski,R.A., Lee,D., Lees,J.G. *et al.* (2014) CATH: comprehensive structural and functional annotations for genome sequences. *Nucleic Acids Res.*, **43**, D376–D381.
17. Furnham,N., Holliday,G.L., De Beer,T.A.P., Jacobsen,J.O.B., Pearson,W.R. and Thornton,J.M. (2014) The Catalytic Site Atlas 2.0: cataloging catalytic sites and residues identified in enzymes. *Nucleic Acids Res.*, **42**, 1–5.
18. Higurashi,M., Ishida,T. and Kinoshita,K. (2009) PiSite: a database of protein interaction sites using multiple binding states in the PDB. *Nucleic Acids Res.*, **37**, D360–D364.
19. Sehnal,D., Svobodová Vařeková,R., Pravda,L., Ionescu,C.-M., Geidl,S., Horský,V., Jaiswal,D., Wimmerová,M. and Koča,J. (2015) ValidatorDB: database of up-to-date validation results for ligands and non-standard residues from the Protein Data Bank. *Nucleic Acids Res.*, **43**, D369–D375.
20. Hauck,D., Joachim,I., Frommeyer,B., Varrot,A., Philipp,B., Möller,H.M., Imberty,A., Exner,T.E. and Titz,A. (2013) Discovery of two classes of potent glycomimetic inhibitors of Pseudomonas aeruginosa LecB with distinct binding modes. *ACS Chem. Biol.*, **8**, 1775–1784.
21. Ernst,B. and Magnani,J.L. (2009) From carbohydrate leads to glycomimetic drugs. *Nat. Rev. Drug Discov.*, **8**, 661–677.
22. Winzer,K., Falconer,C., Garber,N.C., Diggle,S.P., Camara,M. and Williams,P. (2000) The Pseudomonas aeruginosa lectins PA-IL and PA-IIL are controlled by quorum sensing and by RpoS. *J. Bacteriol.*, **182**, 6401–6411.
23. Sabin,C., Mitchell,E.P., Pokorná,M., Gautier,C., Utille,J.-P., Wimmerová,M. and Imberty,A. (2006) Binding of different monosaccharides by lectin PA-IIL from Pseudomonas aeruginosa: thermodynamics data correlated with X-ray structures. *FEBS Lett.*, **580**, 982–987.
24. Mitchell,E., Houles,C., Sudakevitz,D., Wimmerova,M., Gautier,C., Pérez,S., Wu,A.M., Gilboa-Garber,N. and Imberty,A. (2002) Structural basis for oligosaccharide-mediated adhesion of Pseudomonas aeruginosa in the lungs of cystic fibrosis patients. *Nat. Struct. Biol.*, **9**, 918–921.
25. Laity,J.H., Lee,B.M. and Wright,P.E. (2001) Zinc finger proteins: new insights into structural and functional diversity. *Curr. Opin. Struct. Biol.*, **11**, 39–46.
26. Gersbach,C.A., Gaj,T. and Barbas,C.F. (2014) Synthetic zinc finger proteins: the advent of targeted gene regulation and genome modification technologies. *Acc. Chem. Res.*, **47**, 2309–2318.
27. Wang,Z., Feng,L.S., Matskevich,V., Venkataraman,K., Parasuram,P. and Laity,J.H. (2006) Solution structure of a Zap1 zinc-responsive

domain provides insights into metalloregulatory transcriptional repression in Saccharomyces cerevisiae. *J. Mol. Biol.*, **357**, 1167–1183.
28. Activities at the Universal Protein Resource (UniProt). (2014) *Nucleic Acids Res.*, **42**, D191–D198.
29. Pabo,C.O., Peisach,E. and Grant,R.A. (2001) Design and selection of novel Cys2His2 zinc finger proteins. *Annu. Rev. Biochem.*, **70**, 313–340.
30. Hartwig,A. (2001) Zinc finger proteins as potential targets for toxic metal ions: differential effects on structure and function. *Antioxid. Redox Signal.*, **3**, 625–634.
31. Pavletich,N.P. and Pabo,C.O. (1993) Crystal structure of a five-finger GLI-DNA complex: new perspectives on zinc fingers. *Science*, **261**, 1701–1707.
32. Malgieri,G., Zaccaro,L., Leone,M., Bucci,E., Esposito,S., Baglivo,I., Del Gatto,A., Russo,L., Scandurra,R., Pedone,P.V. *et al.* (2011) Zinc to cadmium replacement in the A. thaliana SUPERMAN Cys 2His 2 zinc finger induces structural rearrangements of typical DNA base determinant positions. *Biopolymers*, **95**, 801–810.