MotiveValidator: interactive web-based validation of ligand and residue structure in biomolecular complexes

Radka Svobodová Vařeková[1,2,†], Deepti Jaiswal[1,†], David Sehnal[1,2,3], Crina-Maria Ionescu[1], Stanislav Geidl[1,2], Lukáš Pravda[1,2], Vladimír Horský[3], Michaela Wimmerová[1,2,*] 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

ABSTRACT

Structure validation has become a major issue in the structural biology community, and an essential step is checking the ligand structure. This paper introduces MotiveValidator, a web-based application for the validation of ligands and residues in PDB or PDBx/mmCIF format files provided by the user. Specifically, MotiveValidator is able to evaluate in a straightforward manner whether the ligand or residue being studied has a correct annotation (3-letter code), i.e. if it has the same topology and stereochemistry as the model ligand or residue with this annotation. If not, MotiveValidator explicitly describes the differences. MotiveValidator offers a user-friendly, interactive and platform-independent environment for validating structures obtained by any type of experiment. The results of the validation are presented in both tabular and graphical form, facilitating their interpretation. MotiveValidator can process thousands of ligands or residues in a single validation run that takes no more than a few minutes. MotiveValidator can be used for testing single structures, or the analysis of large sets of ligands or fragments prepared for binding site analysis, docking or virtual screening. MotiveValidator is freely available via the Internet at http://ncbr.muni.cz/ MotiveValidator.

INTRODUCTION

Validation arose as a major issue in the structural biology community when it became apparent that some published structures contained serious errors (1–6). Various tools for the validation of the protein and nucleic acid 3D structures are well established, such as WHAT_CHECK (7), PROCHECK (8), MolProbity (9) and OOPS (10).

An essential step in the validation process is checking the ligand structure. Ligands are chemical compounds which form a complex with a biomacromolecule (e.g. sugar, drug, heme) and play a key role in its function. The ligands are also the main source of errors in structures (11,12). Nonetheless, ligand validation is a very challenging task (13), because of the high diversity and nontriviality of their structure and the general lack of information about correct structures. Therefore, early validation tools focused on selected types of ligands (PDB-care (14) focused on carbohydrates) and their scope only widened later (ValLigURL (15)). Ligand validation features were recently added to existing software (e.g. Mogul (16), Coot (17)). New tools such as PHENIX (18) were developed to include ligand validation functionality. However, the functionality of some available tools (i.e. ValLigURL, Mogul, Coot, PHENIX) is aimed at the validation of selected properties (atom clashes, bond lengths, bond angles, etc.) or is limited to a selected type of molecules (e.g. PDB-care validates only carbohydrates).

This article presents the web-based application MotiveValidator, which offers a user-friendly, interactive and platform-independent environment for the validation of ligands and residues in PDB (http://www.wwpdb.org/docs. html) or PDBx/mmCIF (19) files provided by the user.

*To whom correspondence should be addressed. Tel: +420 54949 4947; Fax: +420 54949 2556; Email: Jaroslav.Koca@ceitec.muni.cz
Correspondence may also be addressed to Michaela Wimmerová. Tel: +420 54949 3805; Fax: +420 54949 2690; Email: michaw@chemi.muni.cz
†The authors wish it to be known that, in their opinion, the first two authors should be regarded as Joint First Authors.

Residues refer to any component of a biomacromolecule or a biomacromolecular complex (i.e. amino acids, nucleotides, ligands). Specifically, MotiveValidator is able to evaluate in a straightforward manner whether the ligand or residue under study has a correct annotation (3-letter code), i.e. if it has the same topology and stereochemistry as the model ligand or residue with this annotation. If not, MotiveValidator explicitly describes the differences. Validation is performed against so-called model residues, which can be either correct structures of the residue obtained from the wwPDB Chemical Components Dictionary (20) (accessed via the web interface provided by LigandExpo (21)), or against templates provided by the user. The output provides a report of the validation results, including summary and detailed information in both tabular and graphical form. MotiveValidator can process thousands of ligands or residues in a single validation run that takes no more than a few minutes.

MotiveValidator can be used for testing single structures, or the analysis of large sets of ligands or fragments prepared for binding site analysis, docking or virtual screening. A significant advantage of MotiveValidator is the ability to process structures obtained by any type of experiment and not requiring the user to have any additional knowledge in the field of X-ray crystallography or nuclear magnetic resonance.

DESCRIPTION OF THE TOOL

MotiveValidator incorporates several tools for the detection and extraction of residues (MotiveQuery; D. Sehnal *et al.*, unpublished work), motif superimposition (SiteBinder (22)), chirality verification (OpenBabel (23)), statistical evaluation of results (in-house program) and interactive visualization of 3D structures (ChemDoodle, http://www.chemdoodle.com). All these tools are integrated into a single program which runs on a server and is accessible under any operating system. The built-in 3D molecular visualizer requires an up-to-date web browser with WebGL enabled. In addition to running validations on the server, a command line version of MotiveValidator is also available.

MotiveValidator enables three kinds of validation to be performed, accessible via three modules. Residue Validation is the most general module, meant for any residue, including ligands. Sugar Validation is focused on carbohydrates and Motif/Fragment Validation on biomolecular fragments (motifs). A motif can in principle be any part of a biomacromolecule. Nonetheless, MotiveValidator is focused on the validation of residues, thus here motif generally refers to the residue under study, together with its immediate environment. Validation via any module involves three steps, namely setup, calculation and finally visualization and the analysis of results. We provide here an extensive description of the Residue Validation module and then briefly point out the differences for the other two modules.

Residue validation

Setup. Two kinds of input are required, namely the structure of a biomolecule or biomolecular complex to be validated and a model residue to serve as the reference template for validation (Supplementary Figure S1). The structure to be validated and model residue must be uploaded in PDB format, or can be retrieved in this format from the mirrors of the Protein Data Bank (24) and LigandExpo databases maintained on the MotiveValidator server and updated every week. The structure to be validated can also be uploaded in PDBx/mmCIF format. A single MotiveValidator run can validate multiple residues in multiple structures.

Calculation. After the setup, the validation proceeds in several steps. The sequence of steps performed during validation is as follows (see also Supplementary Figure S2 for a graphical dictionary of the main terms that appear in this section):

(i) In the structure(s) to be validated, find all instances of residues with the same 3-letter code as the model residue.

(ii) Extract the identified residues (i.e. residues to be validated) together with their immediate surroundings (i.e. atoms within one or two bonds of any atom of the residue to be validated), to obtain input motifs for validation.

(iii) For each input motif:

(a) Superimpose the input motif with the model residue to find the best atom pairing, i.e. the correspondence (mapping) between atoms from the model residue and from the input motif. Mathematically, it is the bijection which matches the most atoms from the input motif to the most atoms from the model residue and provides the lowest RMSD (root mean square deviation) for the structural superimposition. PDB names of atoms are not used in this step. The subset of atoms from the input motif paired with atoms in the model residue forms the validated motif. The atoms in a validated motif are checked for connectivity, to ensure that it is the same as in the model residue. Report any discrepancy between the inter-atomic bonds in the validated motif and in the model residue (section Processing Errors/Warnings).

(b) Establish the validated motif according to the best atom pairing identified in the previous step. Based on the validated motif, detect and report errors:

- missing atoms: atom in the model residue with no corresponding atom in the validated structure

- missing rings: missing atoms originating from cycles (rings)

- wrong chirality: atom from the validated motif with different chirality than the corresponding atom from the model residue;

and warnings:

- substitutions: atom from the validated motif with different chemical symbol than the corresponding atom in the model residue (e.g. O mapped to N)

- different atom name: atom from the validated motif with different PDB name than the corresponding atom from the model residue (e.g. the C1 atom mapped to the C7 atom)

- foreign atoms: atom from the model residue mapped to atom from outside the validated residue (i.e. from its surroundings).

Note: An occurrence of a warning does not mean that the validated motif is wrong. The warning serves only as information to the user.

Visualization and processing of results. All setup information, along with all input and output structures and files are deposited on the server in a unique directory, translated as a unique URL accessible for visualization and download for at least a month. The MotiveValidator output provides a straightforward report of the validation results, including a summary and detailed information in both tabular and graphical form, along with a 3D structure visualizer for closer inspection of the problematic structures.

The Summary section first provides a description of the validation process and then a validation report for each validated residue (Figure 1). The report contains information about the model residue (annotation, 2D structure) and an overview (table and pie chart) of issues found during validation, namely, the number of residues with missing atoms, missing (incomplete) rings, wrong chirality, correct chirality, substitutions, different atom names and foreign atoms. A list of specific issues and their localization within the residue (i.e. number of residues with particular missing atoms or atoms having wrong chirality) is also given.

The Details section (Figure 2, top) provides detailed information for each validated motif. It is organized into a table with one line per motif, containing basic identification of the motif inside the original input file and a list of all issues identified during validation. Each motif can be examined in the 3D space and a complete validation report is available in graphical form using the individual motif links (Figure 2, bottom).

The additional section Processing Errors/Warnings lists the issues found while processing the input files. Processing warnings are issues that may cause incorrect validation, such as atoms that are too close in the 3D space. Processing errors are major issues preventing the finalization of the validation, such as parts of the residue which are completely disconnected from the rest of the structure, probably due to missing atoms at multiple locations throughout the structure.

Sugar validation

A notable case of ligand validation is the analysis of carbohydrate structures because they have complex topology and many chiral atoms. Carbohydrates are involved in a variety of fundamental biological processes and have significant pharmaceutical and diagnostic potential. Additionally, more than 60% of nontrivial-sized ligands (>10 atoms) from the PDB contain a carbohydrate. For these reasons, MotiveValidator includes the mode Sugar Validation, which was developed specifically for the validation of carbohydrates. Unlike Residue Validation, the Sugar Validation setup stage requires only one input, namely the biomolecule(s) containing residues to be validated. This mode enables the automatic validation of all carbohydrate

residues identified in the input structure(s). Specifically, MotiveValidator identifies all motifs containing pyran or furan rings as saccharides and validates them against the corresponding model residues (same 3-letter code) retrieved from the LigandExpo mirror.

Motif/fragment validation

The Motif/Fragment Validation mode uses the model residue and fragments of biomolecules as the input, as opposed to entire biomolecules in the Residue Validation mode. The motifs (fragments) should contain the validated residue and its closest surrounding. The surrounding can include, e.g. atoms within one or two bonds of any atom of the validated residue or more. However, it must stay clear, which residue is the validated one. Therefore, the surrounding can contain just fragments of neighboring residues, but not the whole neighboring residues. It is very useful for the efficient processing of very large amounts of data, such as validating all instances of a residue in the entire PDB. The calculation skips steps (i) and (ii) related to residue detection and extraction, and instead starts directly with the superimposition [step (iii)] of the model residue and validated fragments. The fragments can be prepared manually or automatically. The MotiveValidator website also provides the utility MotifExtractor to enable automatic extraction of the desired motifs (residues and their surroundings) from large datasets of biomolecular structures.

RESULTS AND DISCUSSION

We provide examples of uses for MotiveValidator in the form of case studies for each of the three validation modes.

Residue validation: all proteins containing cholic acid

Cholic acid (CHD) is the best known bile acid and includes four rings and 11 chiral atoms. It contains three 6-member rings A, B and C in chair conformation and a 5-member ring D (Supplementary Figure S3A and B) (25). The PDB contains 299 instances of CHD as ligand in a total of 55 PDB entries (access date: 5.1.2014). We collected all 55 structures and validated all occurrences of CHD using the Residue Validation mode in MotiveValidator. The validation (Figure 1) took 15 s and showed that all 299 CHD instances are complete (no missing atoms). However, the validation revealed that almost 13% of the CHD ligands have incorrect chirality. The problematic molecules can be organized into three groups. The first group contains 18 ligands from nine PDB entries, with incorrect chirality at atoms C3, C8, C9, C12 and C14. The errors are caused by the unnatural boat conformation of rings A, B and C in these particular structures (Supplementary Figure S3C). All these structures come from bovine heart cytochrome c oxidase and were published by the same lab. The second group contains 18 ligands from the same nine PDB entries, with incorrect chirality at atoms C8, C9, C12, C14 and C17. The errors are caused by the unnatural twist-boat conformation of rings A, B and C (Supplementary Figure S3D). The third group contains two ligands from the H240A variant of human ferrochelatase (PDB ID 3AQI), with incorrect chirality at atom C20.
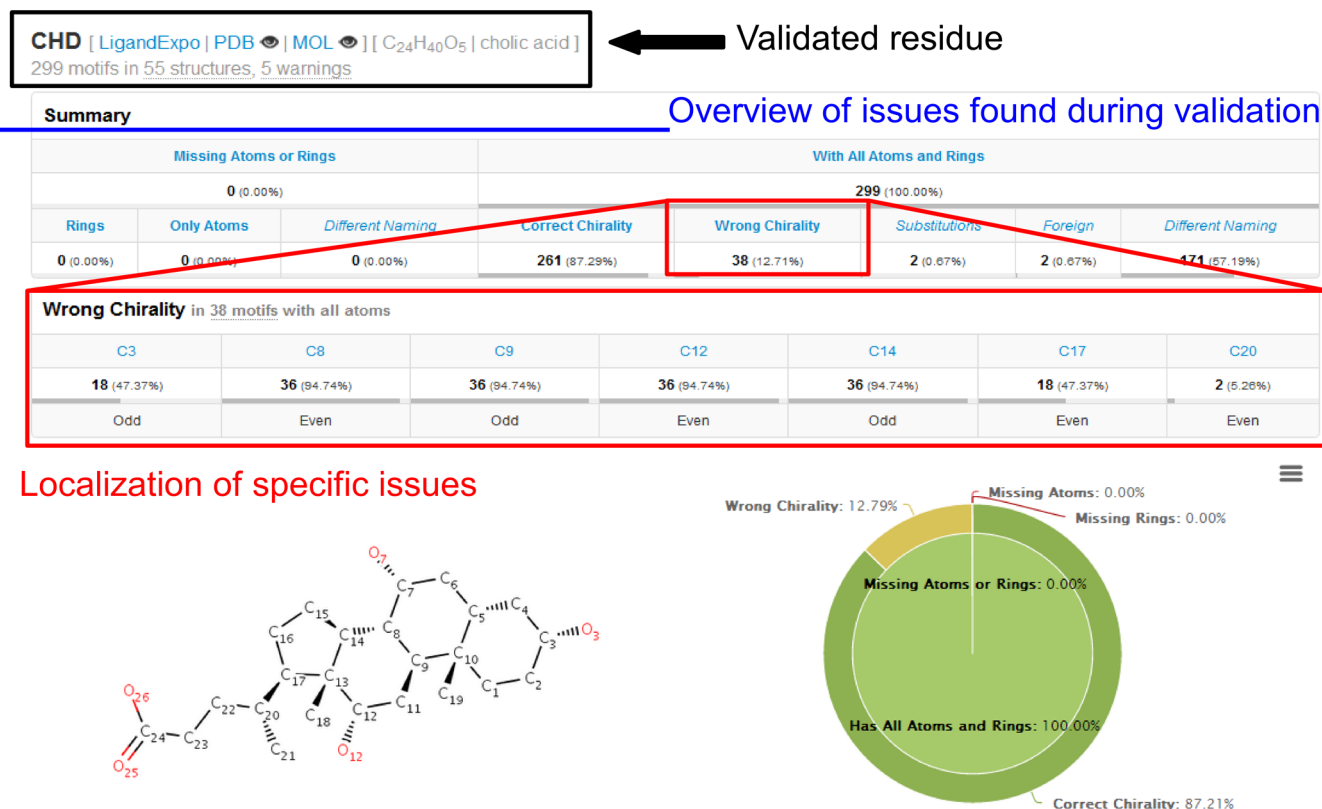
Figure 1. The Summary tab first provides a description of the validation process and a summary of the results in tabular and graphical form. An overview of the issues related to incomplete structure or incorrect chirality is given here, along with other useful notes. The problematic atoms are further highlighted to better localize the problems in the structures.

The complete results are available at the MotiveValidator website as a Sample calculation (http://ncbr.muni.cz/MotiveValidator/ProteinsWithCHD).

Sugar validation: nipah G attachment glycoprotein complexed with ephrin-B3

Nipah virus infection may lead to severe respiratory disease and fatal encephalitis in humans. The Nipah virus relies on the Nipah G attachment glycoprotein for host cell recognition. The crystal structure of the glycoprotein complexed with its receptor ephrin-B3 (PDB ID 3D12, (26)) contains 30 instances of 11 different carbohydrates, each with one ring and five chiral atoms: β-D-glucose (BGC), β-D-mannose (BMA), β-D-gulopyranose (GL0), α-D-glucose (GLC), α-L-galactopyranose (GXL), 2-(acetylamino)-2-deoxy-β-D-gulopyranose (LXB), 2-(acetylamino)-2-deoxy-α-D-idopyranose (LXZ), α-D-mannose (MAN), *N*-acetyl-D-glucosamine (NAG), *N*-acetyl-D-galactosamine (NGA) and 2-(acetylamino)-2-deoxy-α-L-glucopyranose (NGZ). Note that the names of the carbohydrates were obtained from LigandExpo and prefixes alpha- and beta- were replaced with α− and β− (see Supplementary Table TS1 for IUPAC systematic names). We validated all carbohydrate structures in this biomacromolecular complex using the Sugar Validator mode. The validation showed that 13 of these ligands had incorrect chirality (Supplementary Figure S4). In the few cases with GLC or NGA ligands, all five chi-

ral atoms exhibited incorrect chirality. Manual inspection of the structure showed further discrepancies in the ligand part. This is discussed in details in the Supplementary material (Supplementary Figure S5).

The complete results are available at the MotiveValidator website as a Sample calculation (http://ncbr.muni.cz/MotiveValidator/ComplexedGlycoprotein).

Motif/fragment validation: all *N*-acetyl-D-glucosamine residues from PDB

N-acetyl-D-glucosamine (NAG) is the second most frequent hetero-atom chemical component found in the PDB, amounting to 24 357 instances as ligands in a total of 3905 PDB entries (access date: 9.1.2014). NAG includes one pyran ring and five chiral atoms (Supplementary Figure S6A). We extracted all 24 357 NAG instances from the PDB using MotifExtractor. Each file contained one NAG motif, composed of a NAG residue and the atoms in its immediate surroundings (atoms within one or two bonds of the NAG residue). These motifs were validated using the Motif/Fragment Validation mode. The validation (Figure 2) took 195 s and revealed that 94% of NAG instances in the PDB are complete and have correct chirality. In addition, several issues were reported.

First, 16 NAG residues exhibit serious problems: some only contain a few atoms, others have errors in their bond information described by CONNECT keywords (see exam-

Figure 2. The Details tab enables the issues in selected groups of motifs to be inspected by specifying the residue name and type of issue. All information pertaining to a given motif is provided in a single row. Further, each motif can be examined in the 3D space and a complete validation report is accessible via the individual motif links.

ple in Supplementary Figure S6B). Second, approximately 3.5% of NAG residues are missing at least one atom. In most cases, the O1 atom is missing. Third, 2.7% of NAG residues have wrong chirality, mostly at C1, since that is the main site of covalent connection to other residues, which can cause a change in chirality. Some of the chirality errors are caused by incorrect placement of the ligand inside the electron density map. For example, residue NAG 2 A from the PDB entry 3A4X exhibits incorrect chirality at atom C2 (Supplementary Figure S6C). Using Coot and the corresponding electron density maps downloaded from the EDS server at Uppsala University (27), we found that NAG is not placed correctly in the electron density map, leading to a deformation in the vicinity of C2. New positioning leads to a conformation which fits the experimental 3D electron density

map markedly better and which has the correct chirality at position C2 (Supplementary Figure S6D).

Additionally, MotiveValidator found that over 60% of NAG residues in the PDB have a nitrogen substitution at O1, which indicates their participation in *N*-glycosylation. The ability to process and validate also residues with substitutions is an advantage of MotiveValidator.

The complete results are available at the MotiveValidator site as a Sample calculation (http://ncbr.muni.cz/MotiveValidator/MotifsNAG).

Limitations

MotiveValidator is limited in three main ways. First, there is the requirement to ensure that the model residue serv-

ing as the reference during validation is indeed correct. This limitation is overcome by using high-quality reference residues from LigandExpo. Second, the superimposition phase might not identify the optimal matching between the atoms of the model residue and those of the validated residue if their 3D structures are too different. Finally, software and data handling on the server currently limits the maximum size of the input file with structures to be validated (PDB or ZIP file) to 300 MB. We plan to minimize these limitations in the next version of MotiveValidator. For example, we will explore the use of additional metrics to improve the second limitation.

CONCLUSION

In this article we introduced MotiveValidator, a web-based interactive tool for validating ligand and residue structure in biomolecular complexes. The MotiveValidator interface is easy to use and platform-independent, enables interactive analyses with a high degree of automation, e.g. retrieving structures from local mirrors of the PDB and LigandExpo databases, automatic detection and extraction of sugars or selected residues, including their immediate surroundings. Results are presented in a clear graphical and tabular form, facilitating their interpretation and further processing.

SUPPLEMENTARY DATA

Supplementary Data are available at NAR Online.

ACKNOWLEDGMENTS

FUNDING

REFERENCES

1. Kleywegt,G.J. (2000) Validation of protein crystal structures. *Acta Crystallogr. D Biol. Crystallogr.*, **56**, 249–265.
2. Kleywegt,G.J. (2009) On vital aid: the why, what and how of validation. *Acta Crystallogr. D. Biol. Crystallogr.*, **65**, 134–139.
3. Kleywegt,G.J. (2007) Crystallographic refinement of ligand complexes. *Acta Crystallogr. D. Biol. Crystallogr.*, **63**, 94–100.
4. Davis,A.M., St-Gallay,S.A. and Kleywegt,G.J. (2008) Limitations and lessons in the use of X-ray structural information in drug design. *Drug Discov. Today*, **13**, 831–841.
5. Spek,A.L. (2009) Structure validation in chemical crystallography. *Acta Crystallogr. D. Biol. Crystallogr.*, **65**, 148–155.
6. Gutmanas,A., Oldfield,T.J., Patwardhan,A., Sen,S., Velankar,S. and Kleywegt,G.J. (2013) The role of structural bioinformatics resources in the era of integrative structural biology. *Acta Crystallogr. D. Biol. Crystallogr.*, **69**, 710–721.
7. Hooft,R.W., Vriend,G., Sander,C. and Abola,E.E. (1996) Errors in protein structures. *Nature*, **381**, 272.
8. Laskowski,R.A., MacArthur,M.W., Moss,D.S. and Thornton,J.M. (1993) PROCHECK: a program to check the stereochemical quality of protein structures. *J. Appl. Crystallogr.*, **26**, 283–291.
9. Chen,V.B., Arendall,W.B., Headd,J.J., Keedy,D.A., Immormino,R.M., Kapral,G.J., Murray,L.W., Richardson,J.S. and Richardson,D.C. (2010) MolProbity: all-atom structure validation for macromolecular crystallography. *Acta Crystallogr. D. Biol. Crystallogr.*, **66**, 12–21.
10. Kleywegt,G.J. and Jones,T.A. (1996) Efficient rebuilding of protein structures. *Acta Crystallogr. D. Biol. Crystallogr.*, **52**, 829–832.
11. Read,R.J., Adams,P.D., Arendall,W.B., Brunger,A.T., Emsley,P., Joosten,R.P., Kleywegt,G.J., Krissinel,E.B., Lütteke,T., Otwinowski,Z. *et al.*, (2011) A new generation of crystallographic validation tools for the protein data bank. *Structure*, **19**, 1395–1412.
12. Gore,S., Velankar,S. and Kleywegt,G.J. (2012) Implementing an X-ray validation pipeline for the Protein Data Bank. *Acta Crystallogr. D. Biol. Crystallogr.*, **68**, 478–483.
13. Kleywegt,G.J., Henrick,K., Dodson,E.J. and van Aalten,D.M.F. (2003) Pound-wise but penny-foolish: how well do micromolecules fare in macromolecular refinement? *Structure*, **11**, 1051–1059.
14. Lütteke,T. and von der Lieth,C.-W. (2004) pdb-care (PDB carbohydrate residue check): a program to support annotation of complex carbohydrate structures in PDB files. *BMC Bioinformatics*, **5**, 69.
15. Kleywegt,G.J. and Harris,M.R. (2007) ValLigURL: a server for ligand-structure comparison and validation. *Acta Crystallogr. D. Biol. Crystallogr.*, **63**, 935–938.
16. Bruno,I.J., Cole,J.C., Kessler,M., Luo,J., Motherwell,W.D.S., Purkis,L.H., Smith,B.R., Taylor,R., Cooper,R.I., Harris,S.E. *et al.*, Retrieval of crystallographically-derived molecular geometry information. *J. Chem. Inf. Comput. Sci.*, **44**, 2133–2144.
17. Debreczeni,J.É. and Emsley,P. (2012) Handling ligands with Coot. *Acta Crystallogr. D. Biol. Crystallogr.*, **68**, 425–430.
18. Adams,P.D., Afonine,P.V, Bunkóczi,G., Chen,V.B., Davis,I.W., Echols,N., Headd,J.J., Hung,L.-W., Kapral,G.J., Grosse-Kunstleve,R.W. *et al.*, (2010) PHENIX: a comprehensive python-based system for macromolecular structure solution. *Acta Crystallogr. D. Biol. Crystallogr.*, **66**, 213–221.
19. The Protein Data Bank exchange dictionary (PDF), Westbrook,J., Henrick,K., Ulrich,E.L. and Berman,H.M. (2005) *International Tables for Crystallography G. Definition and exchange of crystallographic data*. In: Hall,SR and McMahon,B (eds). Springer, Dordrecht, pp. 295–298.
20. Henrick,K., Feng,Z., Bluhm,W.F., Dimitropoulos,D., Doreleijers,J.F., Dutta,S., Flippen-Anderson,J.L., Ionides,J., Kamada,C., Krissinel,E. *et al.*, (2008) Remediation of the protein data bank archive. *Nucleic Acids Res.*, **36**, D426–D433.
21. Feng,Z., Chen,L., Maddula,H., Akcan,O., Oughtred,R., Berman,H.M. and Westbrook,J. (2004) Ligand Depot: a data warehouse for ligands bound to macromolecules. *Bioinformatics*, **20**, 2153–2155.
22. Sehnal,D., Vařeková,R.S., Huber,H.J., Geidl,S., Ionescu,C.-M., Wimmerová,M. and Koča,J. (2012) SiteBinder: an improved approach for comparing multiple protein structural motifs. *J. Chem. Inf. Model.*, **52**, 343–359.
23. O'Boyle,N.M., Banck,M., James,C.A., Morley,C., Vandermeersch,T. and Hutchison,G.R. (2011) Open Babel: an open chemical toolbox. *J. Cheminform.*, **3**, 33.
24. Berman,H., Henrick,K., Nakamura,H. and Markley,J.L. (2007) The worldwide Protein Data Bank (wwPDB): ensuring a single, uniform archive of PDB data. *Nucleic Acids Res.*, **35**, D301–D303.
25. Mukhopadhyay,S. and Maitra,U. (2004) Chemistry and biology of bile acids. *Curr. Sci.*, **87**, 1666–1683.

26. Xu,K., Rajashankar,K.R., Chan,Y.-P., Himanen,J.P., Broder,C.C. and Nikolov,D.B. (2008) Host cell recognition by the henipaviruses: crystal structures of the Nipah G attachment glycoprotein and its complex with ephrin-B3. *Proc. Natl. Acad. Sci. U.S.A.*, **105**, 9953–9958.

27. Kleywegt,G.J., Harris,M.R., Zou,J.Y., Taylor,T.C., Wählby,A. and Jones,T.A. (2004) The uppsala electron-density server. *Acta Crystallogr. D Biol. Crystallogr.*, **60**, 2240–2249.