

# **MAP\_CHANNELS: a computation tool to aid in the visualization and characterization of solvent channels in macromolecular crystals**

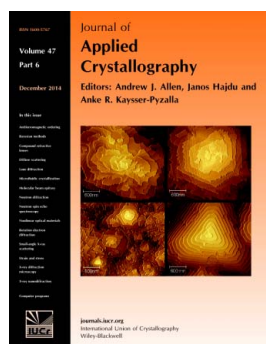
**Douglas H. Juers and Jon Ruffin**

*J. Appl. Cryst.* (2014). **47**, 2105–2108

Copyright © International Union of Crystallography

Author(s) of this paper may load this reprint on their own web site or institutional repository provided that this cover page is retained. Republication of this article or its storage in electronic databases other than as specified above is not permitted without prior permission in writing from the IUCr.

For further information see <http://journals.iucr.org/services/authorrights.html>



Many research topics in condensed matter research, materials science and the life sciences make use of crystallographic methods to study crystalline and non-crystalline matter with neutrons, X-rays and electrons. Articles published in the *Journal of Applied Crystallography* focus on these methods and their use in identifying structural and diffusion-controlled phase transformations, structure-property relationships, structural changes of defects, interfaces and surfaces, etc. Developments of instrumentation and crystallographic apparatus, theory and interpretation, numerical analysis and other related subjects are also covered. The journal is the primary place where crystallographic computer program information is published.

**Crystallography Journals Online** is available from [journals.iucr.org](http://journals.iucr.org)

# MAP\_CHANNELS: a computation tool to aid in the visualization and characterization of solvent channels in macromolecular crystals

Douglas H. Juers\* and Jon Ruffin

Department of Physics, Whitman College, Walla Walla, WA 99362, USA. Correspondence e-mail: juersdh@whitman.edu

A computation tool is described that facilitates visualization and characterization of solvent channels or pores within macromolecular crystals. A scalar field mapping the shortest distance to protein surfaces is calculated on a grid covering the unit cell and is written as a map file. The map provides a multiscale representation of the solvent channels, which when viewed in standard macromolecular crystallographic software packages gives an intuitive sense of the solvent channel architecture. The map is analysed to yield descriptors of the topology and the morphology of the solvent channels, including bottleneck radii, tortuosity, width variation and anisotropy.

© 2014 International Union of Crystallography

## 1. Introduction

The determination of high-resolution three-dimensional structures of macromolecules (*i.e.* proteins and nucleic acids) usually involves crystallization of the macromolecule into a highly ordered but porous material. Although the primary interest is normally in the macromolecule itself, there is often motivation for structural biologists to visualize and characterize the pores or the solvent channels within the crystal. As size exclusion can prevent the access of ligands to binding sites, pore diameters and their proximity to particular regions of the macromolecule are useful information. Knowledge of the solvent channels sheds light on crystal packing, which can often include biologically relevant interactions. The diameter of solvent channels also affects the behaviour of the crystals during cryogenic cooling (Weik *et al.*, 2001).

Outside of structural biology, macromolecular crystals have seen use as nanoporous materials, usually after being cross-linked to improve stability (Margolin & Navia, 2001; Vilenchik *et al.*, 1998). Applications include catalysis (Margolin & Navia, 2001), biosensing (Laothanachareon *et al.*, 2008), chiral separations (Vilenchik *et al.*, 1998) and drug delivery. Solvent channel architecture impacts many of these applications as it governs diffusion of solutes through the crystals (Cvetkovic *et al.*, 2005; Malek *et al.*, 2004). Both experimental and computational studies have examined relationships between diffusion and various aspects of pore architecture, including pore size (Cvetkovic *et al.*, 2005), surface roughness (Malek & Coppens, 2001) and chemical composition (Malek *et al.*, 2004).

For single proteins, there are many tools designed for characterizing surfaces, channels and pockets (Ho & Marshall, 1990; Levitt & Banaszak, 1992; Connolly, 1993; Kleywegt & Jones, 1994; Laskowski, 1995; Sanner *et al.*, 1996; Smart *et al.*, 1996; Petřek *et al.*, 2006, 2007; Ho & Gruswitz, 2008; Yaffe *et al.*, 2008; Le Guilloux *et al.*, 2009; Tan *et al.*, 2013; Benkaidali *et al.*, 2014). However, characterizing channels at the level of the crystal requires accounting for space-group symmetry and periodic boundary conditions associated with unit-cell translations. For macromolecular crystallographers, a common approach to visualize solvent channels is to display the macromolecule and its symmetry mates in molecular display programs, which is cumbersome and may be quite slow for large molecules.

MAP\_CHANNELS was, therefore, written with two goals in mind. Our first aim was for straightforward visualization of the pores within macromolecular crystals to get an intuitive sense for the pore architecture. Our second aim was to characterize the pores with metrics relevant for the diffusion of ligands into and out of the crystals, as well as for the response of the crystals to the large temperature change associated with cryogenic cooling.

## 2. MAP\_CHANNELS strategy

MAP\_CHANNELS involves three distinct conceptual steps. First, the information about the crystal is gleaned from the PDB file header, including space group, cell parameters, diffraction limit and information about the biological units present. Solvent content is determined from SEQRES records and in their absence from atom types.

Second, the distance field is calculated. The atoms are read and the unit cell is divided into a grid. Accounting for space group and noncrystallographic symmetry, the distance from each grid point to the surface of each atom in the cell is calculated (Li & Nussinov, 1998), keeping the shortest distance. The result is a list of gridpoints with an associated distance – the shortest distance from the gridpoint to any protein surface in the crystal. Spherical objects with radius identical to this distance may sit centred on the gridpoint without steric clash with protein atoms. The scalar field is written as a CCP4 (Collaborative Computational Project, Number 4, 1994) format map which can be displayed in standard macromolecular graphics packages. Adjusting the contour level of the map yields an intuitive multiscale view of the solvent channels.

Third, the distance field is analysed. If a gridpoint distance is greater than a user-defined cutoff radius, it is designated as a channel. The channel gridpoints are then broken into groups based on nearest neighbour connectivity, using a procedure essentially identical to an algorithm reported for identifying connected sets in protein envelopes (Hunt *et al.*, 1997). Space-group symmetry and unit-cell translation are accounted for so that unit-cell faces do not interrupt connected groups. Each connected group is then analysed to determine if it is a closed pocket or a pipe that traverses the unit cell, and if the latter along how many unit-cell axes the cell is crossed. For pipes

that traverse the cell in just one direction, tortuosity and width variation are calculated to give some sense of the departure of the pipe geometry from being perfectly uniform and straight. Tortuosity is the ratio of the contour length of the channel to the direct end-to-end distance. Width variation is the ratio of the widest point of the channel to the narrowest point of the channel. By analysing the distance field over a range of cutoff radii, a multiscale view of the channel structure is developed, yielding bottleneck radii for transit of the unit cell in each of the three unit-cell directions.

## 3. Language and machine requirements

*MAP\_CHANNELS* was written in GFortran (Stallman & the GCC Developer Community, 2013) and is currently available for Mac OS X 10.6–10.9. A Python plugin was written allowing *MAP\_CHANNELS* to be run from within *COOT* (Emsley *et al.*, 2010). The program can be also run from the command line. Both programs may be downloaded from <http://www.whitman.edu/~juersdh/>. The source code for *MAP\_CHANNELS* is available by contacting the authors.

## 4. Input requirements

The input to *MAP\_CHANNELS* is a PDB-format file. At a minimum, for the solvent channel map calculation the file must include ATOM records and unit-cell and space-group information (*i.e.* a CRYST1 record). Symmetry operators present in REMARK 290 section will be used, in which case *MAP\_CHANNELS* is self-contained. Otherwise the symmetry operators will be read from the CCP4 symmetry library using the space-group information in the CRYST1 record, in which case CCP4 (Collaborative Computational Project, 1994) must be present.

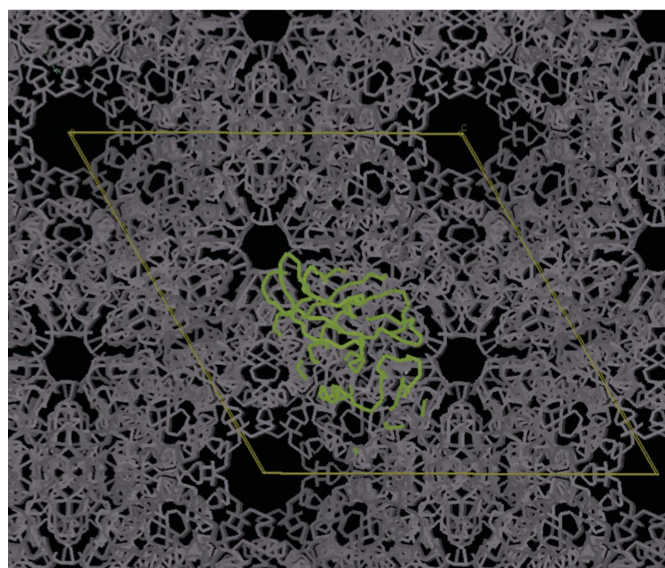
## 5. Results

### 5.1. Solvent channel visualization

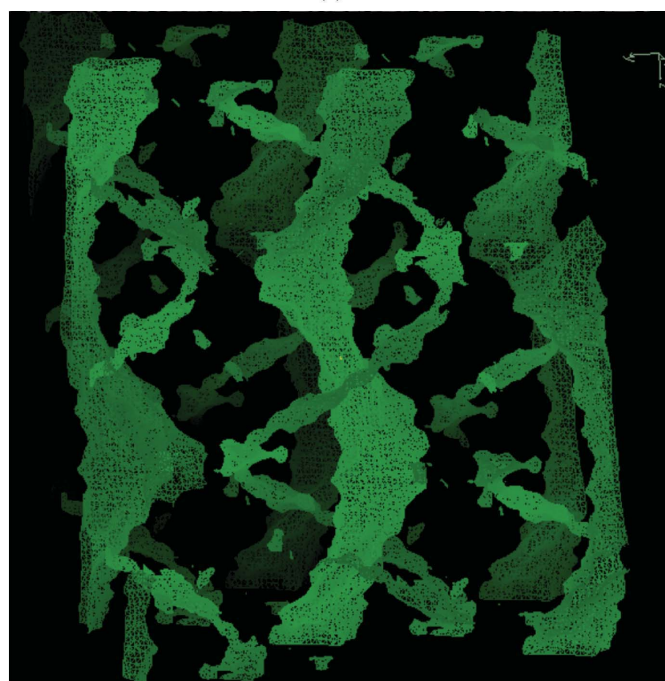
Fig. 1 shows solvent channels in hexagonal thermolysin (PDB structure 1thl; Holland *et al.*, 1994). Fig. 1(*a*) is a view achieved by displaying symmetry mates with a backbone representation in *COOT* (Emsley *et al.*, 2010), looking down the *c* axis. There appear to be two channels of radii 4 and 1.5 Å, but the view is essentially a projection and little three-dimensional information can be gleaned from the picture. Fig. 1(*b*) shows the distance map view of the same pore structure resulting from the *MAP\_CHANNELS* calculation, contoured at 6 Å with the *c* axis running up–down. The 6 Å contour level indicates that a spherical object of radius 6 Å could be placed with its centre anywhere on the map surface without steric clash with the macromolecule. The three-dimensional aspect of the pores is apparent, with two entwined helical pores (radii are about 11.5 and 6 Å). Comparing with Fig. 1(*a*) shows in this case that the projection view is somewhat misleading. The apparent 4 Å pore is formed by projecting the helical 11.5 Å pore along the *c* axis. Movie S1<sup>1</sup> illustrates a multiscale view of the pores in PDB structure 8tlh (Holland *et al.*, 1992) by showing the distance map at decreasing contour levels.

Examination of the map indicates active site access is from the 6 Å helical pore, *via* channels of somewhat smaller radius (~3.5 Å, Fig. 2). Thermolysin-inhibitor structures are commonly determined *via* both soaking and co-crystallization. The inhibitors are typically non-

<sup>1</sup> Supporting information for this paper is available from the IUCr electronic archives (Reference: JO5006).



(a)



(b)

**Figure 1**

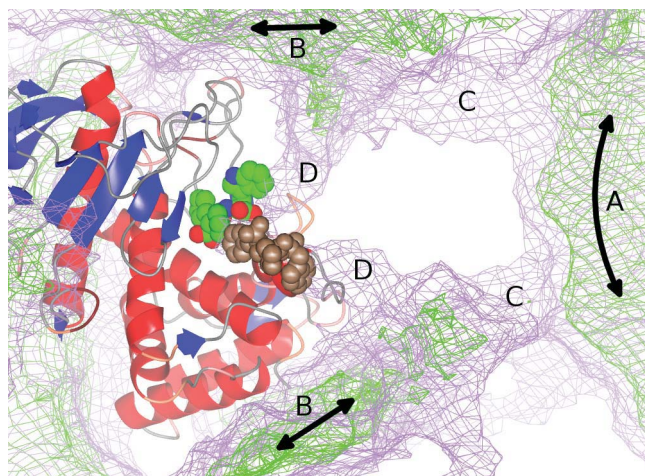
Two views of solvent channels in hexagonal thermolysin. (*a*) View down the *c* axis by displaying symmetry mates as backbone traces in *COOT*. A single protein and the unit-cell outline are both shown in yellow. There are two apparent pores of radii 4 and 1.5 Å. (*b*) Distance map view contoured at 6 Å. The *c* axis runs up–down. Two entwined helical pores are visible. The larger helical pore (radius 11.5 Å) when projected along the *c* axis yields the apparent 4 Å pore in (*a*).

hydrolyzable peptide mimics and can be quite long, but are usually narrower than ~7 Å, permitting diffusion to the active site.

### 5.2. Solvent channel characterization

Fig. 3 shows a subset of output from *MAP\_CHANNELS* characterizing the pores in the thermolysin crystal (1thl). The map is analysed using increasing cutoff radii, which in each case gives a view of how the channel structure ‘looks’ to a spherical object of radius equal to the cutoff radius. The multiscale cluster analysis lists for each





**Figure 2**

Schematic showing access of an inhibitor (green/blue/red spheres) to the active site of thermolysin (red/blue ribbons). The view is roughly with the *c* axis running vertically, as in Figs. 1(b) and Movie S1. The distance map is shown contoured at two levels: 6 Å (green) and 4 Å (lilac). A section of the larger, 11.5 Å radius pore (A) and sections of two of the smaller, 6 Å radius pores (B) are shown. At the 4 Å contour level (lilac), the 11.5 and 6 Å pores are connected by bridges (C). Access to the active site is *via* fingers (D) extending from the smaller pores. A symmetry-related ligand is shown in brown. Not shown, but in the field of view and occupying mainly the white space, are three other protein molecules.

#### Multiscale cluster analysis ...

Rad	Pipe	Pock	%Pi	%Po	Pipe/Pocket Data	Avg	SD	Tort	WV	(Average)
1.4	1	0	100	0	1.00 3D ABC 12.9	4.0	2.3			
2.4	1	0	100	0	1.00 3D ABC 12.9	4.9	2.1			
3.4	1	0	100	0	1.00 3D ABC 12.9	5.7	2.0			
4.4	2	0	100	0	0.66 1D ..C 12.9	7.2	2.0	1.2	1.1	( 1.6 1.2 )
					0.34 1D ..C 8.8	5.6	0.9	2.3	1.3	
5.4	2	0	97	3	0.77 1D ..C 12.9	7.9	1.8	1.2	1.1	( 1.5 1.1 )
					0.23 1D ..C 8.8	6.3	0.7	2.3	1.3	
6.4	1	1	86	14	1.00 1D ..C 12.9	8.5	1.5	1.2	1.1	( 1.2 1.1 )
					0.93 0D ... 8.8	7.1	0.5			
7.4	1	0	95	5	1.00 1D ..C 12.9	9.2	1.3	1.2	1.1	( 1.2 1.1 )
8.4	1	0	100	0	1.00 1D ..C 12.9	9.9	1.1	1.2	1.1	( 1.2 1.1 )
9.4	1	0	100	0	1.00 1D ..C 12.9	10.6	0.9	1.2	1.1	( 1.2 1.1 )
10.4	1	0	100	0	1.00 1D ..C 12.9	11.3	0.6	1.2	1.1	( 1.2 1.1 )
11.4	1	0	100	0	1.00 1D ..C 12.9	12.0	0.4	1.2	1.1	( 1.2 1.1 )
12.4	0	27	0	100	0.10 0D ... 12.7	12.7	0.0			
					0.10 0D ... 12.8	12.6	0.2			
					0.08 0D ... 12.5	12.5	0.0			
					0.07 0D ... 12.5	12.5	0.0			

#### SUMMARY DATA

Solvent content   Vm	:	0.50		2.44
Maximum distance to protein (Å)	:	12.91		
from	:	85.00	5.20	120.01
to	:	ASP	261	OD2 (radius= 1.38)
Maximum radius for 3D   2D   1D	:	3.40		3.40   11.40
Direction for 1D	:	..c		
Min   max tortuosity	:	1.21		1.58
Min   max width variation	:	1.08		1.17
Bottleneck tensor:				
3.40	3.40	3.40		
3.40	3.40	3.40		
3.40	3.40	11.80		

**Figure 3**

Section of output from *MAP\_CHANNELS* showing the channel characterization metrics. Details of the multiscale cluster analysis are shown, followed by summary information. Multiscale cluster analysis: for each cutoff radius, the number of pipes and pockets are listed with information about the major pipes and pockets; Rad is the cutoff radius (Å); Pipe is the number of pipes; Pock is the number of pockets; %Pi/%Po are % of gridpoints that are pipes or pockets; Frac. is the fraction of the pipe (pocket) gridpoints occupied by that pipe (pocket); xD is the dimensionality of the pipe/pocket (*i.e.* along how many of the unit-cell directions the pipe/pocket traverses the unit cell); Typ is the type of pipe (*e.g.* ABC = crosses cell in all three cell directions; .. C represents 'crosses cell only in the *c* direction'); Max/Min/SD are maximum, minimum and standard deviations of all of the distances for the pipe or pocket. For one-dimensional pipes, tortuosity and width variation information is listed. Tort is the tortuosity, WV is the width variation, (Average) is the average of tortuosity and width variation if multiple pipes are present, weighted by the number of pipe gridpoints.

cutoff radius the number of pipes (Pipe) (*i.e.* channels that traverse the unit cell), the number of pockets (Pock) and the percentage of channel gridpoints that are pipe (%Pi) or pocket (%Po) (also see caption for Fig. 3). For cutoff radii less than 4.4 Å, there is just one contiguous pipe transiting the unit cell in all three directions (*i.e.* a three-dimensional pipe of type ABC, meaning it traverses the unit cell along each unit-cell direction). This implies that spherical objects smaller than 4.4 Å (including water molecules) should be able to traverse the unit cell in all three directions. At 4.4 and 5.4 Å, there are two one-dimensional pipes (shown in Fig. 1), both traversing the cell in the *c* direction, one occupying 2–3 times as many gridpoints as the other. The larger pipe (maximum radius = 12.9 Å) has a tortuosity of 1.2, while the smaller pipe (maximum radius = 8.8 Å) has a tortuosity of 2.3. At 6.4 Å there is only one pipe (which is the larger pipe of Fig. 1), and this remains until 12.4 Å, at which point the pipe has broken into a few pockets (0D).

The summary data list the solvent content and show that the widest part of any channel is about 13 Å in radius, with bottlenecks in the *c* and *a/b* directions of 11.4 and 3.4 Å, respectively. This means that spherical objects smaller than 3.4 Å in radius (including water molecules, for example) should be able to transit the crystal in all three directions. Objects between 3.4 and 11.4 Å in size may transit the crystal only along the *c* cell edge, while objects larger than 11.4 Å should be size excluded from entering the crystal. Although objects between 11.4 and 13 Å could fit inside the crystals, they may be prohibited from doing so by the 11.4 Å bottleneck. However, this analysis assumes that the residues lining the surface of channels are static, while thermal motion of these residues would be expected to impact these cutoff radii. The off-diagonal elements of the bottleneck tensor indicate size exclusion for a change in diffusion direction.

### 5.3. Other output options

In addition to the basic channel characterization, other output options exist. Coordinates at the centre of one-dimensional channels may be output in PDB format with the *B*-factor field containing the channel radius at that point. Coordinates expanded to occupy one complete unit cell may be output, which can be useful to compare with the distance maps as well as the channel centre coordinates.

Coordinates can be output with the *B*-factor field containing the distance to the nearest solvent channel or the residue depth which is correlated with several physical properties of proteins (Pintar *et al.*, 2003). Increasing residue depth in the context of the crystal was recently shown to correlate with reduced radiation damage, possibly *via* protection by outer residues from damaging free radicals that form in the solvent channels (Juers & Weik, 2011). Finally, other output options are currently in development, including a binary search to rapidly find the location in the unit cell furthest from the protein, and a Reeb graph representation of the solvent channel architecture (Ushizima *et al.*, 2012; Hilaga *et al.*, 2001).

### 5.4. Other input options and control of execution

The user can control several aspects of the program execution. Concerning which atoms are used in the distance map calculation, options exist to include water molecules and other heteroatoms (excluded by default) and to exclude high-*B*-factor atoms. Concerning the length of time for the program to run, the calculation time is roughly linear with the number of atoms and cubic with 1/gridsize. The expected time, therefore, varies over all of the structures in the PDB by several orders of magnitude. The gridsize may be controlled explicitly or implicitly by specifying a maximum time for the calculation. The default mode is to initially set the gridsize to

2.0 Å and then increase the gridsize until the calculation is expected to take three minutes or less. Using a 2.0 Å gridsize, calculation times were 0.13 s, 20 s, 5 min and 2.5 h for monoclinic crambin, tetragonal hen lysozyme, tetragonal thermolysin and orthorhombic beta-galactosidase, respectively, for  $10^{6.3}$ ,  $10^{8.4}$ ,  $10^{9.2}$  and  $10^{10.9}$  distance calculations on a late 2013 Apple MacBook Pro with 2.6 GHz Intel Core i5 processor. The required gridsize will depend on the intended use of the program. For a qualitative view of the solvent channels, the default three minute maximum time mode will probably be effective. One may also choose to coarse-grain the protein using a pseudoatom with the amino acid residue's centre of mass and radius of gyration to represent the residue location and size. This speeds up the calculation about 10×, and still gives a qualitatively reasonable outline of the solvent channels. For more accurate determination of channel parameters, repeated executions may be required until the gridsize is at least smaller than the radius of the smallest channel.

## 6. Summary

We have described a computer program, *MAP\_CHANNELS*, aimed at facilitating the visualization and objective description of solvent channels in macromolecular crystals. The program may also be of use to the nanoporous materials community, when crystal structures are available (e.g. zeolites). Future work aims at summarizing solvent channel architecture of crystals present in the Protein Data Bank (<http://www.rcsb.org/pdb>), as well as improving descriptors of channel morphology.

This work was supported in part by a grant from the National Institutes of Health (GM090248).

## References

- Benkaidali, L., André, F., Maouche, B., Siregar, P., Benyettou, M., Maurel, F. & Petitjean, M. (2014). *Bioinformatics*, **30**, 792–800.  
 Collaborative Computational Project, Number 4 (1994). *Acta Cryst.* **D50**, 760–763.  
 Connolly, M. L. (1993). *J. Mol. Graph.* **11**, 139–141.

- Cvetkovic, A., Picioreanu, C., Straathof, A. J. J., Krishna, R. & van der Wielen, L. A. M. (2005). *J. Am. Chem. Soc.* **127**, 875–879.  
 Emsley, P., Lohkamp, B., Scott, W. G. & Cowtan, K. (2010). *Acta Cryst.* **D66**, 486–501.  
 Hilaga, M., Shinagawa, Y., Kohmura, T. & Kunii, T. L. (2001). *SIGGRAPH '01. Proceedings of the 28th Annual Conference on Computer Graphics and Interactive Techniques*, edited by L. Pocock, pp. 203–212. New York: ACM.  
 Ho, B. K. & Gruswitz, F. (2008). *BMC Struct. Biol.* **8**, 49.  
 Ho, C. M. W. & Marshall, G. R. (1990). *J. Comput. Aided Mol. Des.* **4**, 337–354.  
 Holland, D. R., Barclay, P. L., Danilewicz, J. C., Matthews, B. W. & James, K. (1994). *Biochemistry*, **33**, 51–56.  
 Holland, D. R., Tronrud, D. E., Pley, H. W., Flaherty, K. M., Stark, W., Jansonius, J. N., McKay, D. B. & Matthews, B. W. (1992). *Biochemistry*, **31**, 11310–11316.  
 Hunt, J. F., Vellieux, F. M. D. & Deisenhofer, J. (1997). *Acta Cryst.* **D53**, 434–437.  
 Juers, D. H. & Weik, M. (2011). *J. Synchrotron Rad.* **18**, 329–337.  
 Kleywegt, G. J. & Jones, T. A. (1994). *Acta Cryst.* **D50**, 178–185.  
 Laothanachareon, T., Champreda, V., Sritongkham, P., Somasundrum, M. & Surareungchai, W. (2008). *World J. Microbiol. Biotechnol.* **24**, 3049–3055.  
 Laskowski, R. A. (1995). *J. Mol. Graph.* **13**, 323–330.  
 Le Guilloux, V., Schmidtke, P. & Tuffery, P. (2009). *BMC Bioinformatics*, **10**, 268.  
 Levitt, D. G. & Banaszak, L. J. (1992). *J. Mol. Graph.* **10**, 229–234.  
 Li, A. J. & Nussinov, R. (1998). *Proteins Struct. Funct. Genet.* **32**, 111–127.  
 Malek, K. & Coppens, M. O. (2001). *Phys. Rev. Lett.* **87**, 125505.  
 Malek, K., Odijk, T. & Coppens, M. O. (2004). *ChemPhysChem*, **5**, 1596–1599.  
 Margolin, A. L. & Navia, M. A. (2001). *Angew. Chem. Int. Ed.* **40**, 2204–2222.  
 Petřek, M., Košinová, P., Koča, J. & Otyepka, M. (2007). *Structure*, **15**, 1357–1363.  
 Petřek, M., Otyepka, M., Banáš, P., Košinová, P., Koča, J. & Damborský, J. (2006). *BMC Bioinformatics*, **7**, 316.  
 Pintar, A., Carugo, O. & Pongor, S. (2003). *Biophys. J.* **84**, 2553–2561.  
 Sanner, M. F., Olson, A. J. & Spehner, J. C. (1996). *Biopolymers*, **38**, 305–320.  
 Smart, O. S., Neduvellil, J. G., Wang, X., Wallace, B. A. & Sansom, M. S. P. (1996). *J. Mol. Graph.* **14**, 354–360.  
 Stallman, R. M. & the GCC Developer Community (2013). *Using the GNU Compiler Collection*. GNU Press.  
 Tan, K. P., Nguyen, T. B., Patel, S., Varadarajan, R. & Madhusudhan, M. S. (2013). *Nucleic Acids Res.* **41**, W314–W321.  
 Ushizima, D. M., Morozov, D., Weber, G. H., Bianchi, A. G. C., Sethian, J. A. & Bethel, E. W. (2012). *IEEE Trans. Vis. Comput. Graph.* **18**, 2041–2050.  
 Vilenchik, L. Z., Griffith, J. P., St Clair, N., Navia, M. A. & Margolin, A. L. (1998). *J. Am. Chem. Soc.* **120**, 4290–4294.  
 Weik, M., Kryger, G., Schreurs, A. M. M., Bouma, B., Silman, I., Sussman, J. L., Gros, P. & Kroon, J. (2001). *Acta Cryst.* **D57**, 566–573.  
 Yaffe, E., Fishelovitch, D., Wolfson, H. J., Halperin, D. & Nussinov, R. (2008). *Proteins*, **73**, 72–86.