3V: cavity, channel and cleft volume calculator and extractor

Neil R. Voss[1] and Mark Gerstein[2,3,4,*]

[1]Department of Cell Biology, The Scripps Research Institute, CB 129, La Jolla, CA, 92037, [2]Program in Computational Biology and Bioinformatics, [3]Department of Molecular Biophysics and Biochemistry and [4]Department of Computer Science, Yale University, Bass 432, New Haven, CT 06520, USA

ABSTRACT

As larger macromolecular structures become available, there is a growing need to understand their 'internal' volumes—such as deep clefts, channels and cavities—as these often play critical roles in their function. The 3V web server can automatically extract and comprehensively analyze all the internal volumes from input RNA and protein structures. It rapidly finds internal volumes by taking the difference between two rolling-probe solvent-excluded surfaces, one with as large as possible a probe radius and the other with a solvent radius (typically 1.5 Å for water). The outputs are volumetric representations, both as images and downloadable files, which can be used for further analysis.

The 3V server and source code are available from http://3vee.molmovdb.org.

INTRODUCTION

The number of large macromolecular structures available is increasing and these new structures contain internal features with biological relevance. There are several examples of biologically relevant channels in the Protein Data Bank (1). One example is the ribosomal exit tunnel, a large protruding channel ($\sim$22 000 Å^3 in channel volume) (2) in the 50S ribosomal subunit that all naturally synthesized proteins must pass through (3). Channels also play an important role in many membrane proteins, including ion channels ($\sim$8000 Å^3) (4) and mechanosensitive channels ($\sim$10 000 Å^3) (5). Additionally, large cavities are utilized as isolation chambers for protein folding chaperones, such as GroEL ($\sim$500 000 Å^3) (6). While it is interesting to study these channels, extracting them from a structure is not straightforward.

There are several programs involved in finding potential binding sites in macromolecules (7–18) and pores in membrane proteins (19,20). However, these programs and web servers are typically designed for small proteins, and are not necessarily useful for large macromolecular complexes. More recently, a few methods have been published to traverse known channels in structures (21–24). These tools determine the trajectory of channels from a starting location in protein structures. The output of these tools is the trajectory of the channel and the maximal radius at each point along the path. The *Voss Volume Voxelator* (3V) web server provides similar functionality, but using a different technique that can provide the overall shape of the channel. The 3V method requires no starting point and the probe radii are adaptable to the size of any structure and its channels, which is ideal for investigations relating to drug-binding sites.

FORMALISM AND METHODOLOGY

In order to calculate macromolecular volumes, the web server uses the rolling probe method (25–27). This method essentially works by rolling a virtual probe or ball of a given radius around the van der Waals surface of a macromolecule (Figure 1). There are two approaches of the rolling probe method to determine the volume of an enclosed surface: discrete and analytical. The 3V web server uses the discrete volume method, i.e. a 3D grid of voxels, rather than the analytical one [e.g. (28)]. The discrete method is scalable to structures of any size and complexity, and readily lends itself for subtracting any two volumes created using this method. This subtraction is required to obtain the channel information. The time complexity of the discrete volume method is linear, $O(n)$, with the number of particles, but it increases as the cube, $O(r^3)$, of the probe radius (in grid units) and the inverse cube of the 3D grid resolution, $O(g^{-3})$. Thus, macromolecules containing a large number of atoms can be computed with only small effect on the calculation time relative to smaller macromolecules for a given probe and grid size. The discrete volume method for the rolling probe allows for volumes to be subtracted and large macromolecules to be efficiently analyzed.

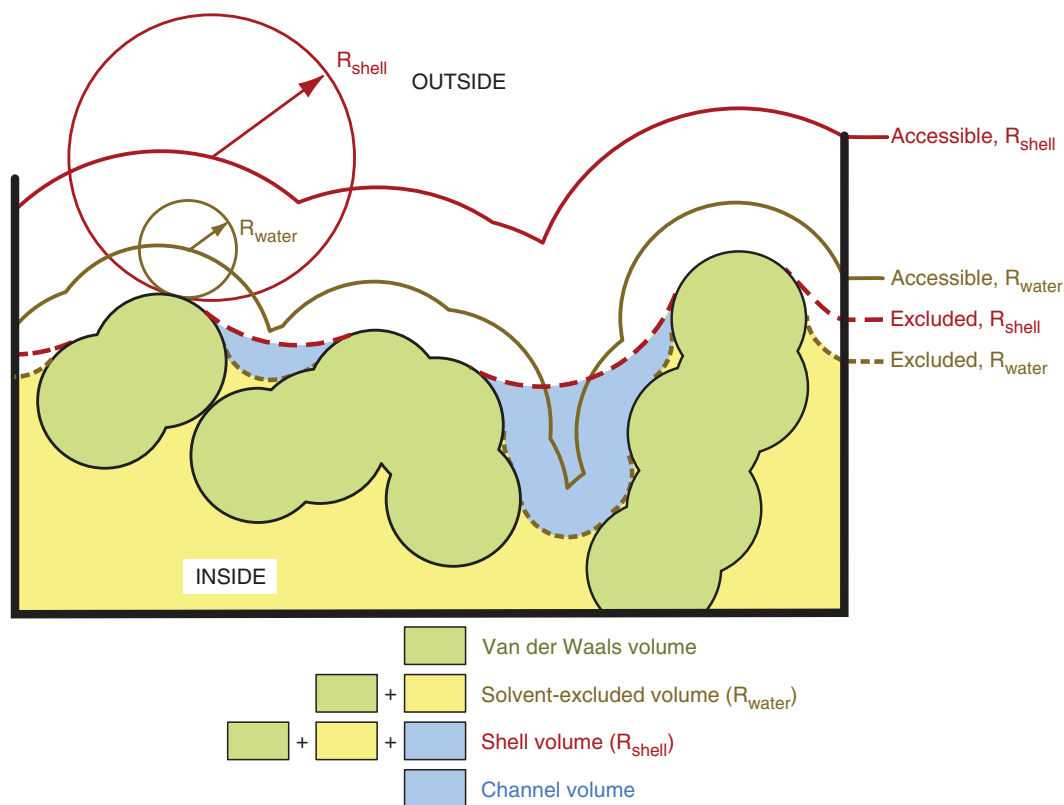*To whom correspondence should be addressed. Tel: +1 203 432 8189; Fax: +1 203 432 6946; Email: mark.gerstein@yale.edu

Figure 1. Richards' rolling probe definition. Two sphere of radius R_1 and R_2 are shown rolling across the surface of a macromolecule defined by the 12 atoms. The excluded surface is the surface defined by the exterior of the each sphere (dashed line) as they roll across the array of atoms By subtracting the enclosed volume of the smaller sphere (R_1) from the enclosed volume of the larger sphere (R_2) the solvent volume (blue) is obtained. Figure adapted from Richards, 1977 (26).

When using the rolling probe method, the size of the probe has a profound effect on the shape of the resulting surface. Figure 2A shows the surfaces of the macromolecule at the different probe radii for both the hammerhead ribozyme (29) and lysozyme (30) structures. When a probe of zero size is used, the van der Waals radius is obtained. However, as the probe radius increases, the surface features are filled in (as in Figure 2A, left). In the two-probe method for extracting internal volumes from a macromolecule, three types of volume are used:

- 'Solvent-excluded volume' or molecular surface (27) is the volume inside the excluded surface of the macromolecule determined using a spherical probe the size of a solvent molecule, typically 1.5 Å for water (Figure 2B, yellow). To put this volume into context, the water-excluded volume of a macromolecular complex is approximately equal to its Voronoi volume, which is 1.21 Å^3 Da for protein and 0.96 Å^3 Da for RNA (31).
- 'Shell volume' is the volume inside the limiting surface of the macromolecule, and is used to distinguish the interior of the macromolecule from its exterior (Figure 2B, red). While the solvent-excluded surface is strictly defined and straightforward to calculate, the shell volume is less precisely defined. The shell volume is defined here as the solvent-excluded

volume with a larger probe radius. This limiting radius is as large as possible to fill all internal cavities, but not too large that it includes external bulk solvent. In the limit of an infinitely-sized probe the convex hull is obtained, but this surface contains a significant amount of external volume that is not necessarily associated with the macromolecule. It has been found that a 10 Å probe is the optimal radius for the shell of the 50S ribosomal subunit (2), but 6 Å probe is a more reasonable value for smaller macromolecules, such as lysozyme or the hammerhead ribozyme (Figure 2A).

- 'Solvent volume' is the volume in the macromolecule that can be occupied by a solvent molecule, including both cavities and channels. It is determined by the difference between the shell volume and the solvent-excluded volume (Figure 2B, blue). Thus, the solvent-excluded volume plus the solvent volume is equal to the shell volume.

These two probe radii can be adjusted from their 1.5 and 6 Å default values. There is a large amount of freedom with these two probe radii that can be optimized for a specific problem depending on the size of the channel of interest. For example, the solvent-excluded probe could be adapted to correspond to the size of a small molecule or drug. A shell probe of 9 Å and a solvent probe of 3 Å was used to find the
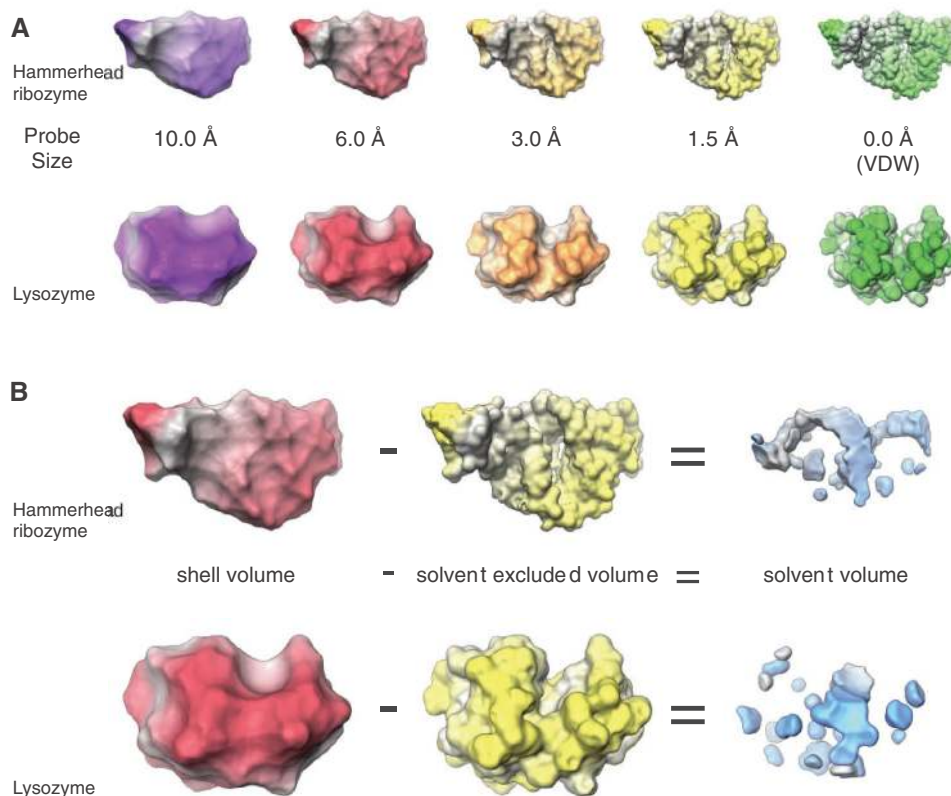
Figure 2. Effects of the rolling probe method. (**A**) The rolling probe volumes at different probe radii. From left to right, the excluded surface at probe radius 10.0, 6.0, 3.0, 1.5 (water-excluded surface) and 0.0 Å (van der Waals surface). (**B**) Shell volume (probe radius of 6.0 Å), solvent-excluded volume (probe radius of 1.5 Å), and solvent volume (blue). The solvent volume is determined by subtracting the solvent-excluded volume from the shell volume. In both (A) and (B), the top model is the hammerhead ribozyme structure (29) and the bottom is the lysozyme structure (30).

drug-binding pocket for the p-glycoprotein (Figure 4). For the ribosomal exit tunnel, a 10 Å shell probe radius and a ~3 Å solvent probe radius was ideal for extracting the exit tunnel (2).

The solvent volume can be further broken down into two types of internal volumes: cavities and channels. The cavity volume is all the volumes large enough to accommodate a solvent molecule, but that are not connected to the macromolecule exterior (Figure 3, red). Cavities are essentially isolated from the exterior, because solvent is unable to escape to the surface without any structural re-arrangements. The second type of internal volume is loosely defined as the 'channel volume'; this generalized channel volume includes surface invaginations, grooves, pockets, or clefts, as well as, long internal channels, large chambers and deep pores. In practice, it is difficult to segregate a channel from a cleft or a pocket without detailed analysis of the volume surface. The channel volume is, therefore, defined as the solvent volume minus the cavity volume, i.e. all volumes large enough to accommodate a spherical probe that are connected to the exterior surface (Figure 3, blue). In tightly packed structures, such as lysozyme, the solvent channels are merely surface invaginations (Figure 3A, blue); however, for larger structures such as the 50S ribosomal subunit, channels can act as large networks within the macromolecule (Figure 3B, blue).

WEB SERVER TOOLS AND RESULTS

There are six tools provided in the 3V web server: two external volume tools and four internal volume ones (Table 1). The first external volume tool is 'Volume Calculation', which is commonly used to study the surface of the macromolecule. Figure 4 (upper left) shows a typical result from this tool, for a excluded surface of the macromolecule. The second tool, 'Volume Range Calculation', is an extension of the first tool and performs the rolling probe calculation using a series of probe radii. This tool is useful for looking at how different probe radii affect the surface of the macromolecule. As shown in Figure 4 (left), for small probe radii, the gap between the two arms is completely defined; however, as the probe radius increases, the probe no longer fits between the protein arms and the volume is filled in.

Three tools are provided to locate and extract channels within a macromolecular structure (Table 1). All three internal volume tools employ the two-probe method to extract the solvent volume (Figure 2B) and, therefore, require both a small and large probe radius as input for their solvent-excluded and shell surfaces, respectively. The first internal volume tool is the 'Solvent Extract' tool. This tool extracts all of the solvent volume at once and requires no additional parameters (Figure 4). This can be useful for identifying all the channels of a macromolecule (Figure 4).
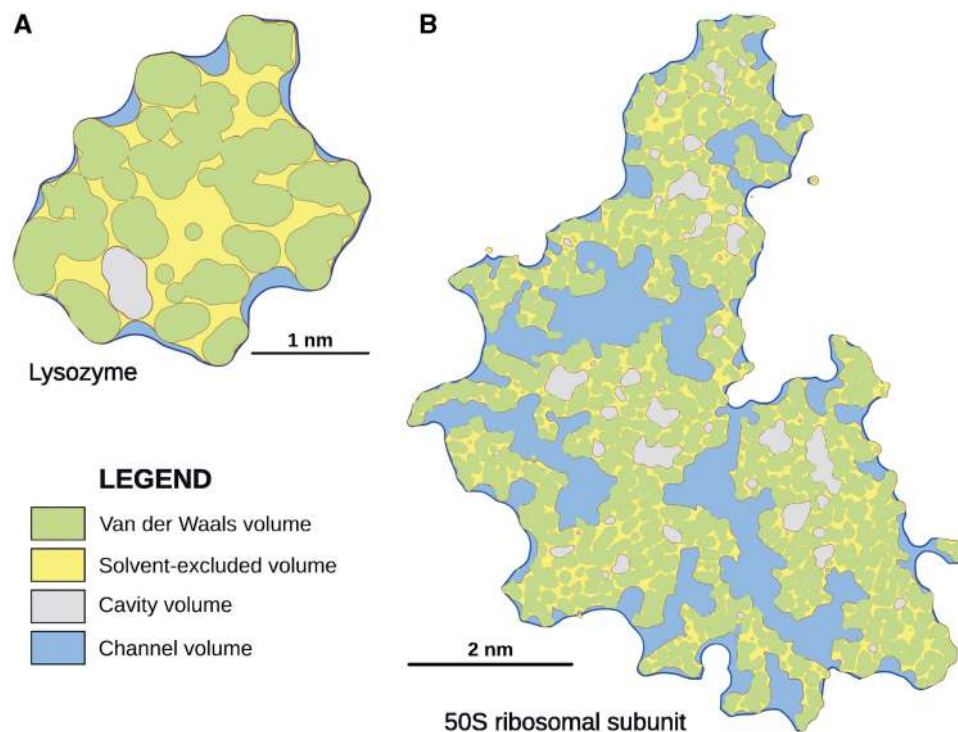
Figure 3. Slices through the macromolecule volumes showing the channels and cavities. The two macromolecules shown are (**A**) lysozyme (30) and (**B**) the 50S ribosomal subunit (33). Four volumes are shown: the van der Waals volume in green, the solvent-excluded surface in yellow, channel volume in blue, and the cavity volume in pink. The shell volume is equal to all the colored areas combined.

Table 1. Six tools provided by the 3V web server

	Tool name	Special input	Short description
External	Volume Calculation (specific)	Single probe radius	Calculate excluded surface for a given probe radius
	Volume range calculation	Min probe radius, max probe radius, radius step size	Calculate excluded surfaces for a range of probe sizes
Internal	Solvent extract (all)	Big probe radius, small probe radius	Extracts all solvent using two probe method
	Channel find (large)	Big probe radius, small probe radius, volume cutoff	Extracts all channels larger than a cutoff
	Channel extract (specific)	Big probe radius, small probe radius, x,y,z coordinate	Extracts solvent connected to a point
Special	Exit tunnel extract	Big probe radius, small probe radius	Extracts the exit tunnel from the 50S subunit

However, in many cases, it finds too many channels, making the 'Channel Find' tool the most useful for beginners. The 'Channel Find' tool calculates the solvent volume, locates all the individual channels in the solvent volume, and extracts any channels larger than a volume cutoff criteria. The volume cutoff criteria can be provided as a requested number of volumes, percentage of the shell volume or an absolute value in $\mathring{A}^3$. In most cases, the default values are sufficient to extract all significant channels from a structure. Figure 4 shows an example of a channel that was extracted using the 'Channel Find' tool. Once the location of a channel is known, then the 'Channel Extract' tool can be used. In the 'Channel Extract' tool, the coordinates of a point located within the channel are required. The 'Channel Extract' tool then extracts any

solvent volume connected to the provided coordinate (Figure 4). A final type of tool called the 'Exit Tunnel Extract' is a specific tool that can only extract the exit tunnel from the 50S ribosomal subunit (2) is provided for replicating the results from Voss *et al.* (2).

WEB SERVER INTERFACE

Tool selection page

The first page presented to the user is the tool selection page (Figure 5A, left). The tool selection page presents all six web server tools, which are divided into two categories: external volumes and internal ones (Table 1). The title and a brief description are given for each tool. To the right of
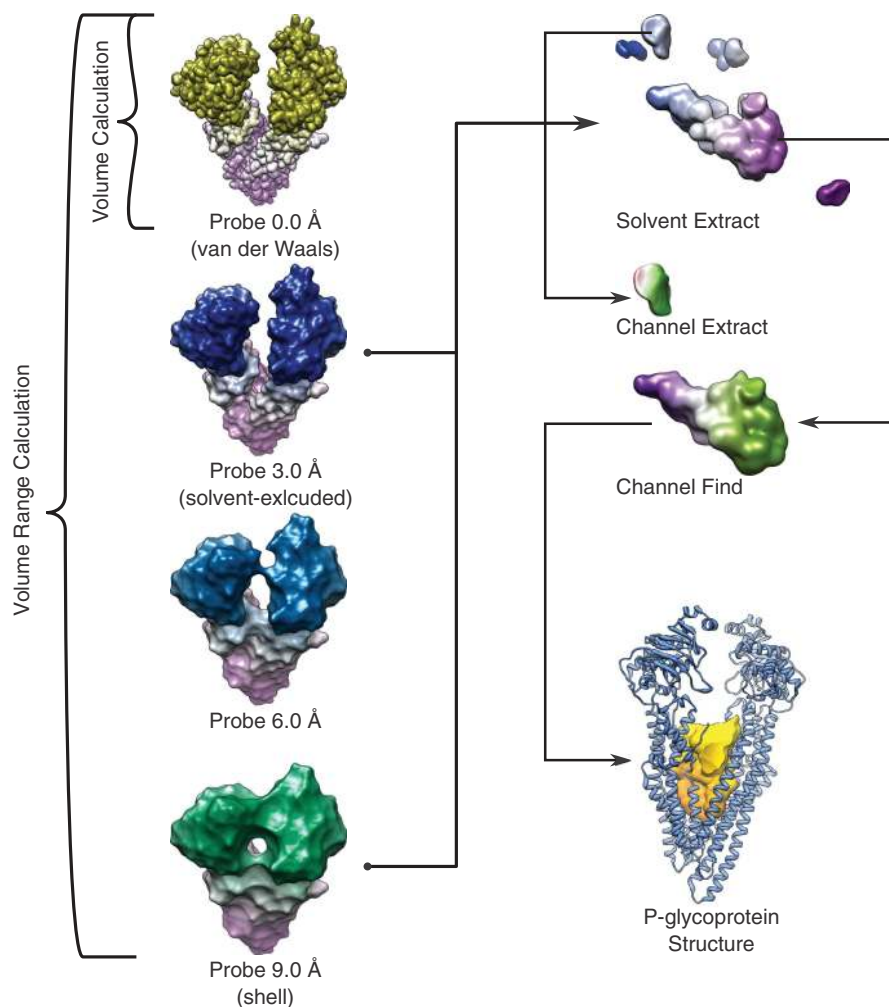
Figure 4. Results from the five general use tools. Results of running each of the five available general use tools on the p-glycoprotein structure (34). The *Volume Calculation* tool creates a single excluded surface of any molecule at any probe radius, shown is the van der Waals radius of 0.0 Å. The *Volume Range Calculation* tool calculates the excluded surface at several different probe radii, shown are probe radii of 0.0, 3.0, 6.0 and 9.0 Å with volumes ranging from 117 970 to 232 799 Å^3. The *Solvent Extract* finds all the solvent in a structure, independent of the size of the volume. In this case, the solvent occupied a volume of 13 487 Å^3. The *Channel Extract* tool produced the chosen cleft from the structure. The *Channel Find* tool found a single, large channel in the structure with a volume of 10 764 Å^3. At the bottom, right the channel found by the *Channel Find* tool is shown within a ribbons model of the PDB structure.

the tool selection table, the most recent runs of all of the tools are listed, providing both the PDB ID and tool name. Clicking on any recent runs allows the user to see the results page for the run. At the bottom of every page is a link to a glossary of terms, the source code, external links and reference information.

Interface for individual tools

The tool interface page is similar for all tools (Figure 5A). The white box on the left gives common parameters for each tool, while unique parameters are included in the gray box on the right. Each parameter comes with detailed help information, which can be accessed by dragging the mouse pointer over the parameter name. There are five common parameters for all tools (Table 2), including PDB input, biological unit, hetero atoms, grid size and allow public use. Each page also contains custom parameters that depend on which tool is being used. Typical custom parameters are rolling probe radii, but some require more

input. For example in the 'Channel Find' tool, a volume cutoff is required to skip any unimportant surface invaginations (Table 1). For users uncertain about how to use any tool, preset examples are provided. When a preset example is used, all the form entries are filled out with values for a given working example. The user can then either run the tool with the value unchanged or manipulate them and see the result.

Processing

After launching the web server tool, the page redirects to the job progress page with its associated job ID (Figure 5A). Each job submission is provided with a unique job ID based on the current date and three random hexadecimal characters (e.g. 10jan13.0e7) that serve as a permanent bookmarkable link to the data. Many web servers take an email address and send an email upon completion. It is a challenging issue to keep users up to date on long calculations through email.
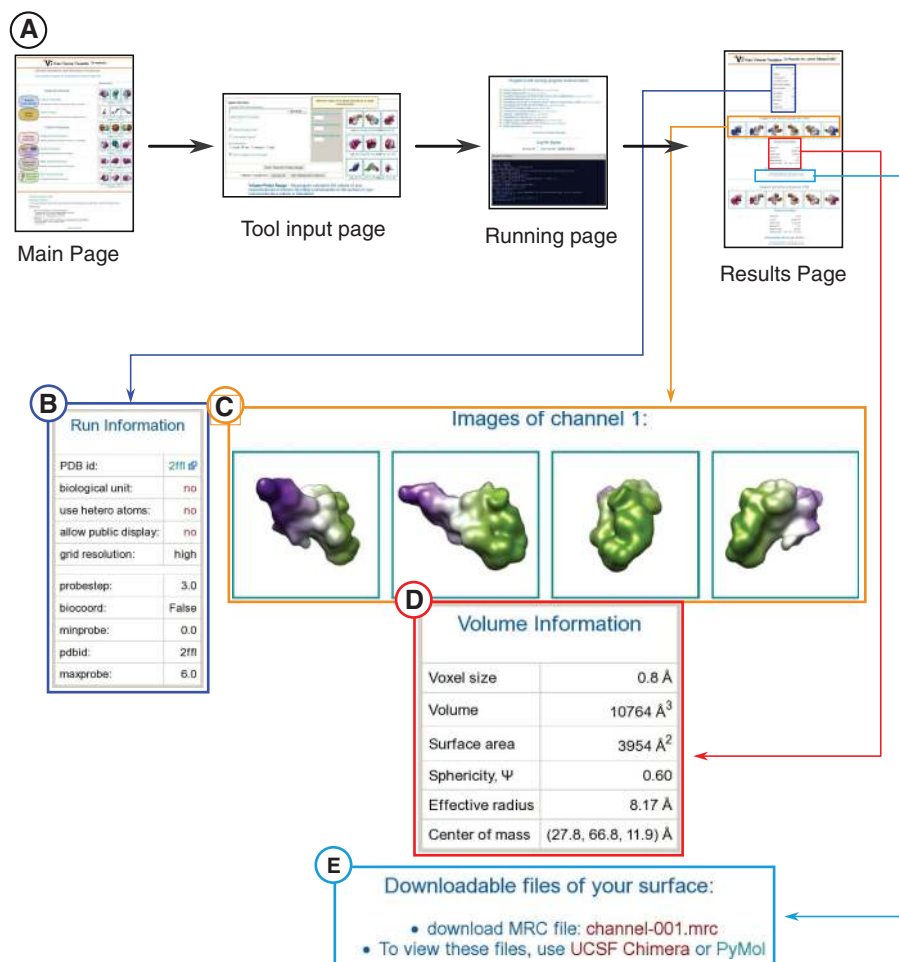
Figure 5. Web page structure. (**A**) The flow of the web page. The first web page presented is the tool selection page with links to run each of the six tools. A typical tool interface page, including the parameter input table and preset buttons. While the calculations are being performed, a progress page is presented with program output. After the program finishes, the progress page becomes the results page. The results page contains information about each calculated volume. (**B–E**) A detail of the results page provides the details of each calculated volume. Each results page contains four major components: (B) run information, such as the PDB ID and other input parameters, (C) image snapshots of the volume, (D) volume information, including the total volume, surface area, sphericity, and other statistics, and (E) links to download and display the volume.

Table 2. Common parameters to each tool

Parameter name	Type	Description
PDB upload	String of file	The only required input, it is either the path to a PDB file to upload or a PDB ID.
Biological unit	Boolean	Whether or not to use the biological unit of the PDB. This corresponds to the macromolecule that has been shown to be or is believed to be functional.
Hetero atoms	Boolean	Whether or not to use hetero atoms. Hetero atoms include water, metal ions, small molecules, chemically modified nucleotides or non-natural amino acids
Grid size	Choice	The choice of the fineness of the grid, which comes down to speed.
Allow usage	Boolean	Whether or not to share their results with the public.

Moreover, recent problems of emails getting lost in spam folders or having too many emails to track, render the email method unreliable. Therefore, 3V has been designed to use the unique job ID to track the progress of these longer processes. The job ID link then shows the progress of the calculations and, upon completion, presents the final results. By saving the web link for a particular job ID, users can close the web page and return at any time to see the results or check on the

progress. During processing, there are two ways to track the progress of the job. First, there is an information box at the top that describes the current processing step being carried out. Second, raw output from the programs is displayed in a terminal-like interface.

Output and results

The output pages have a common layout that depends on the number of extracted volumes (Figure 5A, right). At the

top of the page is information about the input parameters; this is followed by information about each output volume, including snapshot images, volume statistics and downloadable files with information to properly view the files. The input parameters are provided to help reproduce the output, if necessary (Figure 5B). Snapshot images provide a quick indication to the shape of the volume and are randomly colored for help visualizing the 3D shape and provide a reference for the different rotations of the volume (Figure 5C). Volume statistics include the voxel size, total volume, surface area, sphericity, effective radius and center of mass (Figure 5D). The total volume, surface area, and effective radius give one a general idea of the size of the volume. Sphericity, Ψ, is a measure of how much the volume resembles a sphere, defined as:

$$\psi = \frac{\pi^{1/3}(6V)^{2/3}}{A} \tag{1}$$

where V is the volume and A is the surface area. A sphericity of 1.0 implies that the volume is an exact sphere, whereas values of less than 1.0 are less spherical. For example, a cube has a sphericity of 0.806 (Figure 5D). Another way measure the shape of a volume is the effective radius. The effective radius is defined as the radius of a sphere with same surface area to volume ratio as the volume of interest, determined by:

$$r_{\text{eff}} = \frac{3V}{A} \tag{2}$$

The center of mass is most useful from the *Channel Find* tool for pinpointing the location of the channel. The center of mass for a given channel can then be input in the *Channel Extract* tool or used in other channel extraction tools (21–24). Finally, a downloadable file of the volume is available (Figure 5E) for visualization along with the original PDB file. A viewing guide is provided on the web server for new users to manipulate their volume in the available visualization software.

AVAILABILITY

The 3V collection of tools provides researchers with a simple method to investigate macromolecule volumes, including internal volumes that have biological significance. The web server is located at http://3vee .molmovdb.org. This intention is to use this server to store the results indefinitely. In addition, the source code for all the programs is available to download freely as packaged releases or from a SourceForge.net subversion repository and can be modified for more in-depth studies. The source uses only basic C libraries and optionally contains parallel code using the OpenMP library allowing it to be multi-platform.

ACKNOWLEDGEMENTS

FUNDING

Conflict of interest statement. None declared.

REFERENCES

1. 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.
2. Voss,N.R., Gerstein,M., Steitz,T.A. and Moore,P.B. (2006) The geometry of the ribosomal polypeptide exit tunnel. *J. Mol. Biol.*, **360**, 893–906.
3. Ban,N., Nissen,P., Hansen,J., Moore,P.B. and Steitz,T.A. (2000) The complete atomic structure of the large ribosomal subunit at 2.4 A resolution. *Science*, **289**, 905–920.
4. Doyle,D.A., Morais Cabral,J., Pfuetzner,R.A., Kuo,A., Gulbis,J.M., Cohen,S.L., Chait,B.T. and MacKinnon,R. (1998) The structure of the potassium channel: molecular basis of K+ conduction and selectivity. *Science*, **280**, 69–77.
5. Steinbacher,S., Bass,R., Strop,P. and Rees,D.C. (2007) Structures of the prokaryotic mechanosensitive channels MscL and MscS. *Curr. Topics Membranes*, **58**.
6. Braig,K., Otwinowski,Z., Hegde,R., Boisvert,D.C., Joachimiak,A., Horwich,A.L. and Sigler,P.B. (1994) The crystal structure of the bacterial chaperonin GroEL at 2.8 A. *Nature*, **371**, 578–586.
7. Binkowski,T.A., Naghibzadeh,S. and Liang,J. (2003) CASTp: computed atlas of surface topography of proteins. *Nucleic Acids Res.*, **31**, 3352–3355.
8. Ferre,F., Ausiello,G., Zanzoni,A. and Helmer-Citterich,M. (2004) SURFACE: a database of protein surface regions for functional annotation. *Nucleic Acids Res.*, **32**, D240–D244.
9. Hendlich,M., Rippmann,F. and Barnickel,G. (1997) LIGSITE: automatic and efficient detection of potential small molecule-binding sites in proteins. *J. Mol. Graph Model*, **15**, 359–363, 389.
10. Laskowski,R.A. (1995) SURFNET: a program for visualizing molecular surfaces, cavities, and intermolecular interactions. *J. Mol. Graph*, **13**, 323–330, 307–308.
11. Laurie,A.T.R. and Jackson,R.M. (2005) Q-SiteFinder: an energy-based method for the prediction of protein-ligand binding sites. *Bioinformatics*, **21**, 1908–1916.
12. Tseng,Y.Y., Dupree,C., Chen,Z.J. and Li,W. (2009) SplitPocket: identification of protein functional surfaces and characterization of their spatial patterns. *Nucleic Acids Res.*, **37**, W384–W389.
13. Venkatachalam,C.M., Jiang,X., Oldfield,T. and Waldman,M. (2003) LigandFit: a novel method for the shape-directed rapid docking of ligands to protein active sites. *J. Mol. Graph Model*, **21**, 289–307.
14. Yu,J., Zhou,Y., Tanaka,I. and Yao,M. (2010) Roll: a new algorithm for the detection of protein pockets and cavities with a rolling probe sphere. *Bioinformatics*, **26**, 46–52.
15. Kleywegt,G.J. and Jones,T.A. (1994) Detection, delineation, measurement and display of cavities in macromolecular structures. *Acta Crystallogr. D Biol. Crystallogr.*, **50**, 178–185.
16. Kinoshita,K. and Nakamura,H. (2003) Identification of protein biochemical functions by similarity search using the molecular surface database eF-site. *Protein Sci.*, **12**, 1589–1595.
17. Levitt,D.G. and Banaszak,L.J. (1992) POCKET: a computer graphics method for identifying and displaying protein cavities and their surrounding amino acids. *J. Mol. Graph*, **10**, 229–234.
18. Hernandez,M., Ghersi,D. and Sanchez,R. (2009) SITEHOUND-web: a server for ligand binding site identification in protein structures. *Nucleic Acids Res.*, **37**, W413–W416.

19. Coleman,R.G. and Sharp,K.A. (2009) Finding and characterizing tunnels in macromolecules with application to ion channels and pores. *Biophys. J.*, **96**, 632–645.

20. Smart,O.S., Neduvelil,J.G., Wang,X., Wallace,B.A. and Sansom,M.S. (1996) HOLE: a program for the analysis of the pore dimensions of ion channel structural models. *J. Mol. Graph*, **14**, 354–360, 376.

21. Pellegrini-Calace,M., Maiwald,T. and Thornton,J.M. (2009) PoreWalker: a novel tool for the identification and characterization of channels in transmembrane proteins from their three-dimensional structure. *PLoS Comput. Biol.*, **5**, e1000440.

22. Petrek,M., Otyepka,M., Banas,P., Kosinova,P., Koca,J. and Damborsky,J. (2006) CAVER: a new tool to explore routes from protein clefts, pockets and cavities. *BMC Bioinformatics*, **7**, 316.

23. Petrek,M., Kosinova,P., Koca,J. and Otyepka,M. (2007) MOLE: a Voronoi diagram-based explorer of molecular channels, pores, and tunnels. *Structure*, **15**, 1357–1363.

24. Yaffe,E., Fishelovitch,D., Wolfson,H.J., Halperin,D. and Nussinov,R. (2008) MolAxis: a server for identification of channels in macromolecules. *Nucleic Acids Res.*, **36**, W210–W215.

25. Lee,B. and Richards,F.M. (1971) The interpretation of protein structures: estimation of static accessibility. *J. Mol. Biol.*, **55**, 379–400.

26. Richards,F.M. (1977) Areas, volumes, packing and protein structure. *Annu. Rev. Biophys. Bioeng.*, **6**, 151–176.

27. Connolly,M.L. (1983) Analytical molecular surface calculation. *J. App. Crystallogr.*, **16**, 548.

28. Sanner,M.F., Olson,A.J. and Spehner,J.C. (1996) Reduced surface: an efficient way to compute molecular surfaces. *Biopolymers*, **38**, 305–320.

29. Martick,M. and Scott,W.G. (2006) Tertiary contacts distant from the active site prime a ribozyme for catalysis. *Cell*, **126**, 309–320.

30. Diamond,R. (1974) Real-space refinement of the structure of hen egg-white lysozyme. *J. Mol. Biol.*, **82**, 371–391.

31. Voss,N.R. and Gerstein,M. (2005) Calculation of standard atomic volumes for RNA and comparison with proteins: RNA is packed more tightly. *J. Mol. Biol.*, **346**, 477–492.

32. Pettersen,E.F., Goddard,T.D., Huang,C.C., Couch,G.S., Greenblatt,D.M., Meng,E.C. and Ferrin,T.E. (2004) UCSF Chimera–a visualization system for exploratory research and analysis. *J. Comput. Chem.*, **25**, 1605–1612.

33. Klein,D.J., Schmeing,T.M., Moore,P.B. and Steitz,T.A. (2001) The kink-turn: a new RNA secondary structure motif. *EMBO J.*, **20**, 4214–4221.

34. Aller,S.G., Yu,J., Ward,A., Weng,Y., Chittaboina,S., Zhuo,R., Harrell,P.M., Trinh,Y.T., Zhang,Q., Urbatsch,I.L. *et al.* (2009) Structure of P-glycoprotein reveals a molecular basis for poly-specific drug binding. *Science*, **323**, 1718–1722.