J Mol Model (2010) 16:419–429
DOI 10.1007/s00894-009-0541-y

ORIGINAL PAPER

McVol - A program for calculating protein volumes and identifying cavities by a Monte Carlo algorithm

Mirco S. Till · G. Matthias Ullmann

Received: 31 March 2009 / Accepted: 23 May 2009 / Published online: 22 July 2009

Abstract In this paper, we describe a Monte Carlo method for determining the volume of a molecule. A molecule is considered to consist of hard, overlapping spheres. The surface of the molecule is defined by rolling a probe sphere over the surface of the spheres. To determine the volume of the molecule, random points are placed in a three-dimensional box, which encloses the whole molecule. The volume of the molecule in relation to the volume of the box is estimated by calculating the ratio of the random points placed inside the molecule and the total number of random points that were placed. For computational efficiency, we use a grid-cell based neighbor list to determine whether a random point is placed inside the molecule or not. This method in combination with a graph-theoretical algorithm is used to detect internal cavities and surface clefts of molecules. Since cavities and clefts are potential water binding sites, we place water molecules in the cavities. The potential water positions can be used in molecular dynamics calculations as well as in other molecular calculations. We apply this method to several proteins and demonstrate the usefulness of the program. The described methods are all implemented in the program McVol, which is available free of charge from our website at http://www.bisb.uni-bayreuth.de/software.html.

Keywords Cavities in proteins · Molecular volume · Monte Carlo · Water placement inside proteins

M. S. Till · G. M. Ullmann (✉)
Structural Biology/Bioinformatics, University of Bayreuth,
Universitätsstr. 30, BGI,
95447 Bayreuth, Germany
e-mail: Matthias.Ullmann@uni-bayreuth.de

Introduction

The identification of the surface of a protein has a long tradition in many fields of protein modeling and drug design [1–5]. The great interest in this subject is motivated by its importance for identifying ligand binding pockets and cavities in proteins. Moreover, protein crystal structures often show internal cavities that could be filled with water molecules. The identification of such water-filled cavities is important for the analysis of proton transfer networks in proteins, since these water molecules can play a role in hydrogen bond networks and therefore influence the long range proton transport within proteins [6–8]. Several methods have been developed to calculate the solvent accessible surface, molecular surface and molecular volume of a protein. Among them, algorithms based on the alpha shape theory are used in many approaches [2, 9, 10]. The alpha shape theory orders a subset of Delauny complexes with the aim of reducing the computational cost of an inclusion-exclusion formalism to calculate the protein surface and volume. An accurate computation of the molecular and solvent accessible surfaces and volumes is possible with this algorithm. However, the main drawbacks are numerical instabilities due to geometric degeneracy. The computation of the Delauny complexes are shown to be prone to such instabilities. A solution to this problem is found with the so-called “Simulation of Simplicity” [9] which is implemented for example in CASTp [2]. Other methods like LIGSITE [11], POCKET [12], or SURFNET [13] are grid based methods to define the protein surface and internal cavities or ligand binding sites. These methods are limited to the resolution of the grid they use. All these methods are basically methods for integrating the protein volume. Monte Carlo algorithms are known to be able to perform such integrations. A well-known textbook example is the integration of a circle area for the determination of the number π [14]. Such an algorithm can also be used for determining the volume of proteins.

In this paper, we describe an efficient Monte Carlo algorithm for calculating protein volumes and for identifying internal cavities. Our new algorithm is neither dependent on grid resolutions nor is the algorithm prone to geometric degeneracy at any point of the integration. Based on the identified cavities, we suggest possible positions for water molecules and place these water molecules. We apply this program to several proteins of different sizes and compare our results with experimentally identified water positions. The program is available from our website at http://www.bisb.uni-bayreuth.de/software.html.

Methods

Theory of the volume integration

In our algorithm, we consider the protein to consist of spherical atoms. In order to define the molecular volume (MV), we calculate the solvent accessible surface (SAS), which is defined by rolling a probe sphere over the atoms of the protein [15]. The probe sphere represents a solvent molecule. Therefore the probe sphere radius is adjustable to match the desired solvent molecule radius. Figure 1 shows a schematic drawing of the scenario. The MV consists of two parts: the volume of the protein atoms and the volume of the voids, i.e., the volume between the atoms which is not solvent accessible. The MV can be determined by a Monte Carlo integration: A point is randomly placed in a box with known dimensions that contains the whole molecule and it is determined whether this random point is in the solvent or in the MV. From the ratio between points inside the MV and the total number of points, the MV can be calculated. If the box has a volume V_{box}, then the MV is given by

$$MV = \frac{n_{inside}}{n_{tot}} V_{box} \tag{1}$$

where n_{inside} is the number of points inside the MV and n_{tot} is the total number of points.

Whether a point is inside the MV or not is determined by the following steps:

1. If the point is closer to one atom than the van der Waals radius of this atom, the point is inside the van der Waals volume and therefore inside the MV, else
2. If the distance of the point to any atom center is smaller than the van der Waals radius of the atom plus the probe sphere radius and the distance to the closest point of the SAS is larger than the probe sphere radius, the point belongs to a void and therefore to the MV.
3. In any other case, the point belongs to the solvent.

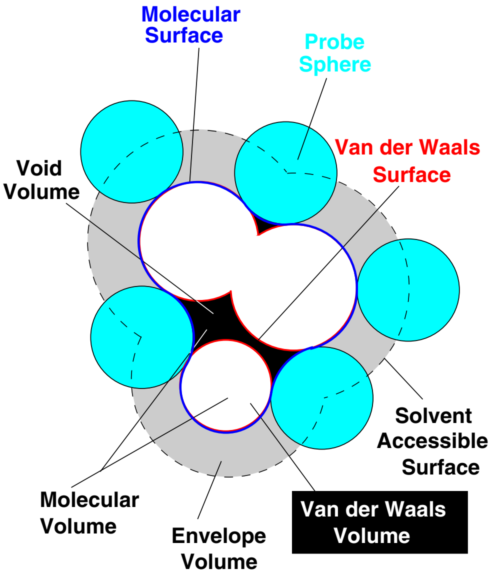

Fig. 1 Definition of volumes and surfaces of a molecule. The atoms of a molecule are represented as white spheres, the probe sphere as cyan spheres. The solvent accessible surface (dashed line) is defined by the center of the probe sphere when rolled over the atoms of the protein. The molecular surface (solid blue line) is defined by the surface points of the probe sphere closest to the protein atoms. The molecular volume consists of two parts, the Van der Waals volume of the atoms and the volume of the voids (shown in black) between these atoms. A void is defined as the space between atoms which is not solvent accessible.The molecular volume is represented by the area inside molecular surface (solid blue line). The solvent accessible surface encloses a volume that consists of three parts: the envelope region (gray), the Van der Waals volume (white), and the void volume (black)

For practical calculations, the SAS is represented by dots. The distance to the surface is than evaluated by calculating the distances to all surface points. In our implementation, we defined the surface points by the double cubic lattice method developed by Eisenhaber and coworkers [16]. This method can also be used to calculate the SAS by the following equation:

$$SAS = \sum_{i=1}^{N} 4\pi r_i^2 \frac{n_{surff,i}}{n_{tot,i}} \tag{2}$$

where N is the number of atoms, r_i is the radius of atom i, $n_{surf,i}$ is the number of dots on the SAS of atoms i and $n_{tot,i}$ is the number of dots placed on atom i, no matter whether they are on the SAS or not.

The pseudocode for determining whether a random point is inside the molecular volume or not is given in the following:

```
point.inside_solvent = true;
point.inside_prot = false;
point.inside_void = false;
for (all atoms(i))
    {if (distance(point,atom(i)) <= atom(i).radius)
        {point.inside_prot = true;
         point.inside_solvent = false;
         break;
        }
    }
if (point.inside_solvent == true)
   {for (all atoms (i))
        {if ((distance(point, atom(i)) <  (atom(i).radius + probe.radius))
             && (distance(point,surface) > probe.radius))
            {point.inside_void = true;
             point.inside_solvent = false;
            }
        }
   }
```

Implementation of the volume integration

A direct implementation of the algorithm described above would give correct results for the volume calculation. However, it would be quite slow, since many distances need to be evaluated. To reduce the number of distance calculations, we used two cell-based neighbor list [14] (see Fig. 2), one for the atoms and another one for the surface dots. Two steps are necessary to create the neighbor list with a given grid spacing. The first step is to place a grid on the protein, where the maximal and minimal Cartesian coordinates of the grid points are the maximal Cartesian coordinates of the protein atoms extended by the maximal radius of the atoms and the probe sphere radius. In our implementation, we allow that the grid cells can have negative indices [17]. Each grid point is initialized as an empty linked list. The second step is to fill the linked lists with the nearby atoms or surface dots. The assignment of atoms to grid cells is done by running over all coordinates, dividing them by the grid spacing and rounding these values to the nearest integer (using the standard C-function rint ()). The rounded coordinates give the indices of the grid cell to which the atom or surface dot is associated. A pointer to the atom or surface dot is appended to the linked list at this grid position. Calculating the distance of a random point to the closest atom or surface dot is then accomplished by the following steps: The coordinates of the random point are divided by the grid spacing and these values are rounded to the nearest integer (using the standard C-function rint()). This procedure gives the indices of the grid cell to which the point is assigned. Now only the distance to atoms or surface dots assigned to the neighboring

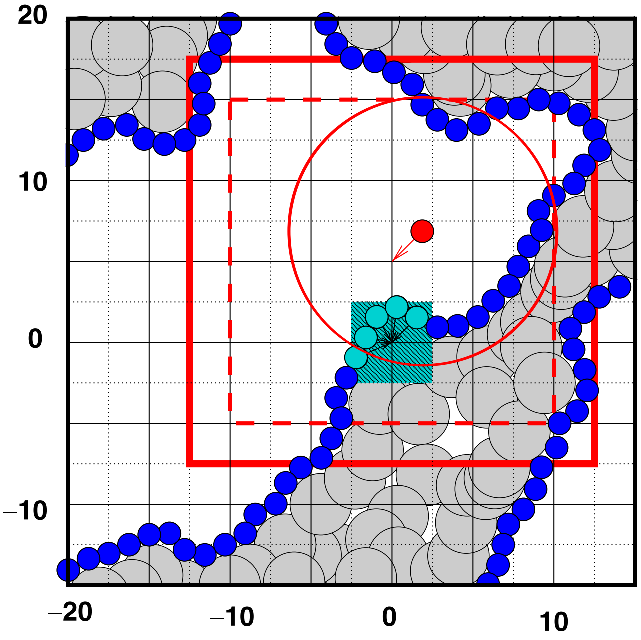

Fig. 2 Schematic drawing of the assignment of surface point to a neighbor list. The task is to find whether the distance of a random point (red) to a surface point (blue) is less than the probe sphere radius. Without a neighbor list, all surface points need to be evaluated until the first surface point within the probe sphere radius is found. In order to reduce the number of distance evaluations, a neighbor list is defined by mapping all surface points to a grid. For example, all cyan points are mapped to the grid point next to them (indicated by black arrows). All points within the cyan rectangle are mapped to this grid point. The random point is also assigned to a grid point. Now only the distance to the surface points in the neighboring grid cells (shown as the dashed red square) need to be evaluated. Only the surface points within the red circle have a distance to the random point that is smaller than the probe sphere radius. Using our neighbor list, only the distances to surface points that are within the solid read square are evaluated

grid cells needs to be calculated. How many neigboring grid cells need to be analyzed is determined as follows. All random points that are at least within a distance of the probe sphere radius plus the maximal atom radius need to be checked for determnінig whether the random point is within the void or envelope region. In order to check whether the point is not in the envelope region, it needs to have a distance from any surface point that is larger than the probe sphere radius. These distances are divided by the respective grid spacing and rounded to the next highest integer *h* (using the standard function ceil()). Then, all distances to the atoms and surface points in the neighboring grid cells are evaluated. Suppose the random point was assigned to the grid cell with the index (*i*, *j*, *k*), the distances to all atoms or surface dots assigned to the grid cells (*i* ± *h*, *j* ± *h*, *k* ± *h*) are calculated. By this procedure, the number of distance calculations is reduced by orders of magnitude. It should be noted, that the grid resolution influences the speed of the program but not the accuracy of the volume calculations, since the points to calculate the volume are placed randomly in the box.

Identification of cavities

The procedure described above allows not only to calculate protein volume but also identify internal cavities. We have two ways to identify internal cavities in our calculation. First, it is possible to identify cavities based on the dot surface and second, based on the volume integration. We describe both possibilities in the following.

First, the surface is defined based on surface points marking the accessibility to the probe sphere. The surface of an internal cavity is described in the same way as the outside surface of the protein. We applied a graph search algorithm to separate surface points defining the outside surface of the protein from surface points defining internal cavities. The undirected graph is generated by connecting surface dots which are less than a certain distance (ca. 1 to 2 Å) apart using a cell-based neighbor list. The basic idea is to divide the graph in unconnected subgraphs. Typically, the largest subgraph describes the outer surface of the protein and smaller subgraphs describe internal cavities. The graph search is implemented as a breadth first search (BFS) [18]. To save memory, both, searching and building the graph is implemented in one routine, since it is not necessary to keep the connectivity matrix in the memory. The BSF methods starts by placing all surface dots in one graph. A vector representing all surface dots shows the graph division. This vector is initialized with 0 as graph number for all elements. Starting from the first element *i* in this vector, we assign the subgraph number 1 to this element and identify all neighboring surface dots. These neighboring surface dots are considered as connected in our graph and therefore the subgraph number 1 is assigned to these points. Additionally, these points are placed on a stack. If all connections of i are evaluated, a loop is started with an empty stack as termination condition. Within this loop, the last dot placed on the stack is taken from the stack and the subgraph number 1 is assigned to all neighboring dots, which do not already have a subgraph number. These dots are also placed on the stack. In each loop iteration, one dot is taken from the stack and all neighboring dots, which are not already in a subgraph are placed on the stack. Therefore, if the stack becomes empty, no more dots are in the whole graph which are connected to subgraph 1 but are not assigned to subgraph 1. If all dots of the surface are placed in subgraph 1, the whole graph is not dividable into subgraphs. If there are dots with 0 as subgraph number remaining in the vector, one of these dots is taken as the next starting point i for subgraph number 2. This procedure is repeated until all dots are assigned to a subgraph. If more than one subgraph is found by the BFS algorithm, subgraphs not connected to the outer protein surface can be defined as internal cavities. The surface of each subgraph can be calculated using Eq. 2.

Second, we can map the random points placed during the MC integration on a grid with a given resolution. Saving the number of points on a grid reduces dramatically the memory requirements compared to saving all random points individually. In each grid cell, we count the number of random points that were placed inside an atom, inside a void, and inside the solvent. A grid cell is marked as solvent as soon as one random point mapped to this grid cell was evaluated to be in the solvent. All grid cells not marked as solvent are considered to be inside the protein. Searching for cavities is accomplished by separating solvent grid cells completely surrounded by protein grid cells from solvent grid cells which are connected to the borders of the box. This separation is achieved by a BFS algorithm as explained above. An undirected graph is build from all grid cells. Within this graph a grid cell has a connection to a neighboring cell, if both grid cells are marked as solvent. After evaluating all grid cells at least one subgraph is found, defining the solvent surrounding the whole protein. If additional subgraphs of solvent grid cells are found these subgraphs are internal cavities. The volume of the internal cavities is integrated again by a Monte-Carlo algorithm. This time with a box placed only around the cavity. The resulting volume is more exact, since more random points are placed in a smaller volume. The volume is again evaluated by Eq. 1.

Detecting surface clefts

One problem connected to the calculation of the surface of a protein is the detection of large clefts on the surface reaching deep into the protein. A cleft is a solvent accessible pocket on the protein surface surrounded by a given ratio of protein. By default our algorithm would treat a cleft with a connection to the solvent as solvent accessible and therefore this cleft is treated as solvent and not as cavity. Several attempts to detected surface clefts were made [1, 2, 4, 5, 11–13, 19–24]. Our method for detecting internal cavities led us to an algorithm which is capable of detecting clefts on the protein surface. For testing if a solvent grid point belongs to a cleft, we place a box on each solvent grid point. The volume of this box is checked for points belonging to the protein or cleft. If more than a given percentage of grid points in the box are protein or cleft points, the solvent point is marked as cleft. Figure 3 schematically depicts the evaluation of a solvent point. This algorithm runs iteratively until no more cleft points are found. The points marked as clefts are divided into subgraphs using the BFS method describe above. The determined clefts are treaded like cavities in the program flow, except that the cleft volume is not reevaluated with a smaller box.

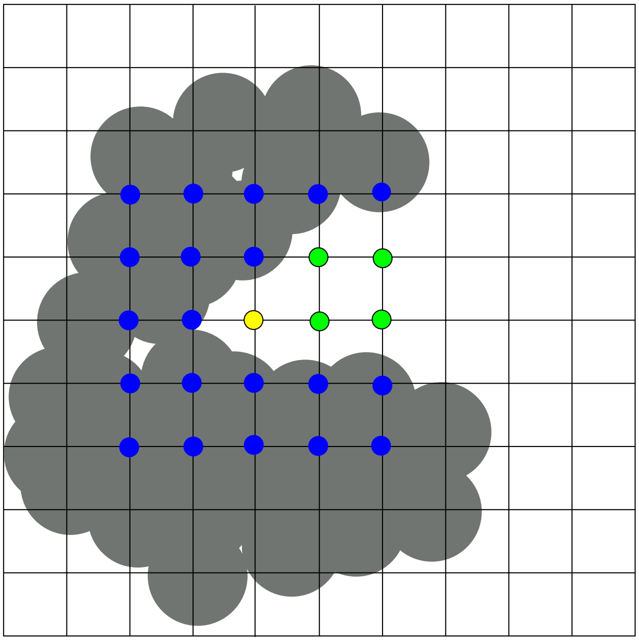

Fig. 3 Definition of clefts in proteins. The grey circles represent protein atoms. The yellow grid point (i,j,k) is a solvent grid point for which it is tested whether it is situated in a cleft or not. All grid cells in the two layers (i.e. i±2,j±2,k±2) are evaluated whether they are solvent grid points (green) or protein grid points (blue). The yellow grid point is considered to be situated in a cleft if a certain percentage of the surrounding grid points are protein grid points or cleft grid points

Placing water oxygen atoms

One reason for searching cavities in proteins is that they may contain water molecules. We place water molecules in all cavities with a volume larger than the volume of one water molecule. Based on the volume of each cavity, the number of water molecules each cavity can hold is determined by dividing the volume of the cavity by the volume of a water molecule. The result is rounded to the nearest integer. Initially, the atoms are place randomly inside the cavity by selecting a random solvent grid node that is far enough from the protein atoms. Starting from this configuration, a Monte Carlo method is applied to optimize the water positions on the grid.

We maximize the function D in Eq. 3

$$D=\sum_{i=1}^{K}\sum_{j=i+1}^{K}d(i,j)+\sum_{i=1}^{K}|x_i-x_{max}|$$
$$+\sum_{i=1}^{K}|y_i-y_{\mathrm{max}}|+\sum_{i=1}^{K}|z_i-z_{\mathrm{max}}|$$
$$+\sum_{i=1}^{K}|x_i-x_{\mathrm{mix}}|+\sum_{i=1}^{K}|y_i-y_{\mathrm{min}}|$$
$$+\sum_{i=1}^{K}|z_i-z_{\mathrm{min}}| \qquad (3)$$

where $d\,(i,j)$ is the distance between water molecule i and j and xyz_{min} and xyz_{max} are the minimal and maximal coordinates of the cavity, respectively. D is maximized by the Monte Carlo algorithm. Maximizing D ensures that the placed water molecules are as far apart from each other as possible and also as far apart as possible from the cavity borders. The algorithm moves one water molecule in a random direction at the grid and checks whether D has increased or not and if a water molecule at this position does not overlap with protein atoms. If the distance sum has increased, the new water position is accepted, otherwise, the move is discarded. The algorithm terminates after a given number of steps. By applying this algorithm, we ensure that the cavity is evenly filled with water molecules. Since no energy criteria are applied during the placement of water molecules, it is recommended to minimize the positions of the water molecules afterward.

Adding a membrane to membrane proteins

For electrostatic calculation on membrane proteins, it is often required to add dummy atoms around the protein representing the hydrophobic region of the membrane [25–27]. When such a membrane of dummy atoms is added, care must be taken, that internal cavities of the protein that are filled potentially by water molecules are not filled by dummy atoms. We implemented a procedure to add a dummy atom membrane in McVol to handle this problem.

Since the protein is placed in a box, all grid points of this box not assigned to a cavity or cleft are solvent grid points. On the basis of these grid points, McVol is capable of placing a membrane of dummy atoms around the protein. This membrane is built by defining an upper and lower border of the membrane. All solvent grid points within these borders (defined by the z-coordinates) are considered as membrane region. Grid points that are identified as cavities are not considered as membrane region in order to avoid that water filled cavities in the protein that are potentially important, for example for proton transfer, are filled with dummy atoms.

The overall flowchart of the program is given in Fig. 4.

Computational details

Structure preparation

All structures discussed in the following are derived from their pdb structures. Hydrogen atoms were added by the hbuild routine of CHARMM [28] and subsequently minimized. Atom radii were taken from Bondi [29] if not stated otherwise.

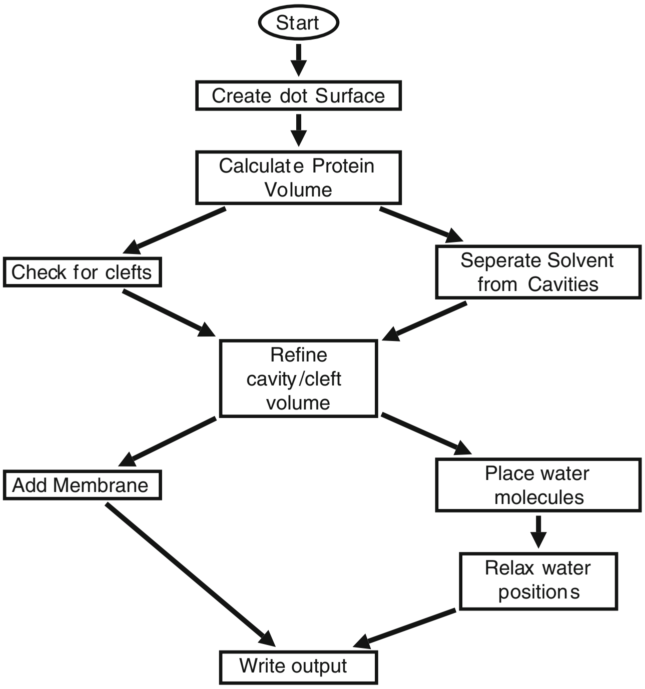

Fig. 4 Flowchart of the program McVol including the detection of potential water positions and adding a dummy atom membrane

Computational details

All calculations were done with 50 Monte Carlo steps per Å^3 of the box volume and 2500 surface dots unless stated otherwise. The probe sphere radius was initially set to 1.3 Å in accordance to the water volume. The grid resolution for the initial grid was set to 1 Å, the cavity volume refinement was done with a grid resolution of 0.5 Å. Water molecules were only placed in cavities larger than 18 Å^3. The number of water molecules per cavity was determined by dividing the cavity volume by the volume of a water molecule and rounding the result.

Results

Convergence of the Monte Carlo algorithm

We tested the convergence of the Monte Carlo algorithm for calculating the volume of a molecule by varying the number of Monte Carlo steps per cubic Å of the box volume between 50 and 250. Moreover, we varied the number of points placed initially on each atom for the creation of the dot surface by the double cubic lattice method [16] between 500 and 10,000 per atom. We use 3-hydrobenzoate hydrolase (pdb code 2dkh) [30] as a test case. Each calculation was repeated 10 times in order to get an error estimation. The results are shown in Fig. 5. We observed no influence of the number of Monte Carlo steps

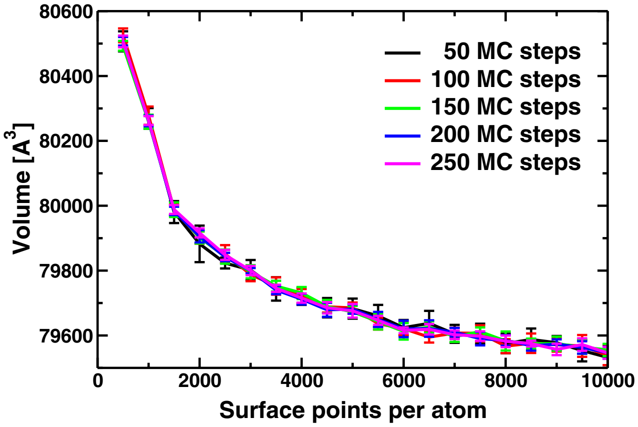

Fig. 5 Convergence of the protein volume determined by the program McVol in dependence on the number of Monte Carlo steps and surface points as atom. Molecular volume was calculated with 50 to 250 Monte Carlo steps per Å^3 box volume and 500 to 10000 surface points per atom. The protein 3-hydrobenzoate hydrolase (pdb code 2dkh) was used as an example

on the protein volume. All five calculations with the same number of surface points resulted in the same volume. The number of surface points influences the volume calculation, but only within a range of about 1%. Since the protein volume shows the strongest dependence for the increase in the number of surface points from 500 to 2000, we decided to take 2500 surface points for all further calculations unless otherwise stated. As shown in Fig. 5, the volume decreases with increasing number of surface points per atom. This behavior, which we term surface artifact, can be explained as follows. The decision, if a random point is inside a void or inside the envelope volume (see Fig. 1), is made based on the distance to the closest surface point. If the distance to the closest surface point is larger than the probe sphere radius, the point is inside a void. With fewer surface points, a random point which is located between two surface points might be treated as void point even if its real distance to the surface is less than the probe sphere radius and thus it should be considered as a point in the envelope volume. Since voids are included in the molecular volume, these misassigned points artificially increase the protein volume. However, as shown above, this effect only leads to a minor error. The number of Monte Carlo steps per Å^3 and the number of surface points per atom are the critical parameters for the runtime of the program. Table 1 gives a short overview of the runtime of the program in dependence of these two parameters. The runtime depends approximately linearly on the number of Monte Carlo steps with a slope of one. The dependence on the number of initial surface points is also linear but with a much smaller slope of about 0.01.

The relation between protein volume and number of atoms

We applied our algorithm to 15 enzymes between 896 and 20,835 atoms (see Table 2). In order to minimize the surface artifacts, we calculated the protein volume using 10,000 surface points per atom. For these proteins of different folds and molecular weights, we analyzed the volume of the voids, the volume of the protein and the ratios between these volumes. With one exception (2bgi) all structures show a similar ration between the protein volume and the number of atoms. The molecular volume is composed of the Van der Waals volume of the atoms and the volume of small voids between the atoms. Interestingly, the protein volume is directly correlated to the number of atoms, independent of the size or the folding of the protein (see Fig. 6). Linear regression leads to a slope of 8.04 Å^3/atom and a y-intercept of 102.9 Å^3. The y-intercept shows that the volume of the voids makes a significant contribution to the protein volume.

Cavities in proteins

The major goal of the above described algorithm is to find cavities in proteins. Identification of cavities in proteins is important for developing mechanistic models of the enzymatic activity, since cavities are often filled with water molecules that provide hydrogen bonds or are involved in proton transfer [31, 32]. The above described algorithm was applied to search cavities in three enzymes: Hen egg lysozyme, bacteriorhodopsin and the photosynthetic reaction center.

Table 1 Runtime of McVol (in seconds) for different parameter settings

MC steps per Å^3 box volume	Runtime [s] surface points per Atom								
	500	1000	1500	2000	2500	3000	4000	5000	10000
50	45	52	64	70	76	83	94	107	175
100	89	101	151	161	176	173	201	224	332
150	132	169	221	266	234	232	244	270	398
200	169	191	223	247	266	280	315	351	506
250	205	234	271	297	314	340	378	432	623

Table 2 Volume of 15 different proteins calculated by the program McVol

Protein	# atoms	Molecular volume [Å^3]	Volume/# atoms [Å^3]	vdW-Volume/ void-Volume
Bovine pancreatic tryp. inhibitor (1bpi) [40]	896	7325	8.175	3.648
Henn egg white Lysozyme (4lym) [34]	1967	16369	8.322	3.248
Bacterial BLUF photoreceptor (2byc) [41]	2262	17480	7.728	2.800
Bovine beta-lactoglobulin (1beb) [42]	2492	19668	7.892	2.646
Ferrodoxin NADP(H) reductase (2bgi) [43]	2716	31616	11.641	2.454
Bacteriorhodopsin (1c3w) [44]	3560	27483	7.720	2.788
Urate Oxidase (1r4u) [45]	4670	39054	8.363	3.155
Ammonuim transporter (2b2f) [46]	6140	45487	7.408	2.86
Alpha amylase (1bag) [47]	6446	53168	8.248	2.397
Cryptochrome (1np7) [48]	7842	62631	7.987	2.605
Glucose oxidase (1cf3) [49]	8803	73259	8.322	2.324
BM-40 FS/EC domain pair (1bmo) [50]	9145	72138	7.888	2.721
3-hydrobenzoate hydrolase (2dkh) [30]	9474	79876	8.431	3.027
Acetylene Hydratase (2e7z) [51]	11528	95304	8.267	2.363
Bacterial reaction center(2j8c) [26]	16738	138220	8.258	2.837
average			7.94±1.84	2.76±0.4

Hen egg lysozyme

NMR experiments identified three major cavities in hen egg lysozyme [33]. Each of these cavities is well defined by a set of amino acid side chains surrounding these cavities. We applied our algorithm to hen egg lysozyme (pdb-code 4lym [34]) using a probe sphere radius of 1.3 Å, 250 Monte-Carlo steps per Å^3 of the box volume and 2562 dots per atom on the dot surface. With this probe sphere radius we were not able to detect all of the experimentally reported cavities. Therefore we reduced the probe sphere radius to 1.1 Å. Applying our algorithm with the reduced probe sphere radius, we could reproduce the cavities proposed for hen egg lysozyme. The reduced probe sphere radius may be necessary since a water molecule is not a perfect sphere and the Bondi hydrogen radius may be too large for polar hydrogens.

The experimentally determined cavities were found as two internal cavities and one cleft. The volumes of these cavities and the solvent accessible surfaces are listed in Table 3. The calculated volume of the first cavity is only approximated, since cavity I and the “hydrated cavity” as proposed by Otting et. al. [33] are merged to one cleft in our calculation. This cleft has three main clusters, each of equal size (see Fig. 7). The whole cleft has a size of 114 Å^3 therefore, cavity I was approximated to 38 Å^3. The “hydrated cavity” contains the water molecules 65, 70, and 75 in the pdb file 4lym. If cavity I is subtracted from the large cleft detected by our algorithm, the remaining volume of the “hydrated cavity” is 76 Å^3, which perfectly fits the three water molecules (see Fig. 7).

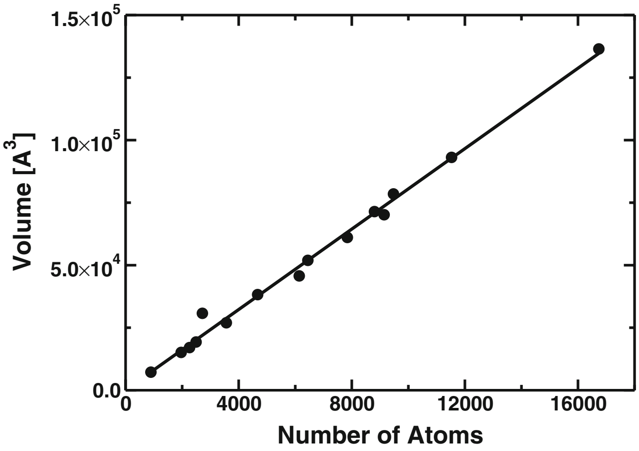

Fig. 6 Dependence of the molecular volume on the number of atoms. The red line is a regression of all points with a slope of 8.04 Å^3 and a y-intercept of 102.9 Å^3/atom

Table 3 Cavities found in the hen egg white lysozyme (4lym). The calculation was done with 250 MC steps per Å^3 box volume and 2500 surface points per atom

Cavity	Volume [Å^3]	SAS [Å^2]	Water molecules
I	38[a]	8.8	2
II	12	0.6	1
III	22	4.1	1
hydrated cavity	76	—	3

[a] Volume estimated from the cleft volume determined by McVol

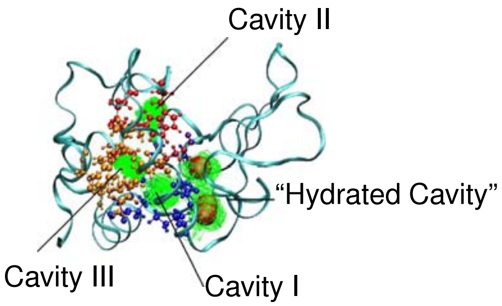

Fig. 7 Cavities found in Hen egg lysozyme. Colored residues show experimentally derived cavities. The large red speres represet three crystallographically resolved water molecules located in one large cleft

Bacteriorhodopsin

Water molecules are proposed in several proton transfer pathways through bacteriorhodopsin (BR) [35, 36]. Some of these water molecules are located near the retinal. We analyzed the cavities in the pdb-file 1c3w. We removed all experimentally derived water positions from the original file for this calculation. Our algorithm (applied with a probe sphere radius of 1.3 Å) was able to detect four cavities near the retinal. Cavity II perfectly fits the water molecules proposed to be involved in proton transfer. The calculated volumes and solvent accessible surfaces are shown in Table 4. The cavities are shown in Fig. 8. In addition, we compared the cavities found in BR with all experimentally derived water positions. Most of the experimentally derived water positions were also found as cavities by our algorithm. Lowering the probe sphere radius to 1.2 Å enabled us to find all experimentally derived water positions as cavities or clefts, except some positions which were on the surface of the protein and clearly not inside a cavity or a cleft. This result indicates that calculations with a probe sphere radius of 1.3 Å may not be able to identify all water filled cavities.

Photosynthetic reaction center

Many water molecules are participating in the proton transfer pathways in the photosynthetic reaction center [26, 37–39], but even in the x-ray structure with the highest resolution [26] not all cavities detected by McVol (using a probe sphere radius of 1.2 Å) are filled with water molecules. In addition to the crystallographically resolved water molecules, 35 cavities and surface clefts were found containing 103 water molecules. Some of these water molecules extend proposed proton transfer pathways connecting previously unconnected aminoacid sidechains participating in the proton transfer from the cytoplasmic site to the secondary quinone (Q_B). The location of the placed water molecules in the photosynthetic reaction center is shown in Fig. 9.

Table 4 Cavities found in the bacteriorhodopsin (1c3w) with a probe sphere radius of 1.3 Å. The calculation was done with 250 MC steps per $Å^3$ box volume and 2500 surface points per atom

Cavity	Volume [$Å^3$]	SAS [$Å^2$]	Water molecules
I	22	2.2	1
II	60	10.6	3
III	13	0.4	1
IV	43	9.0	2

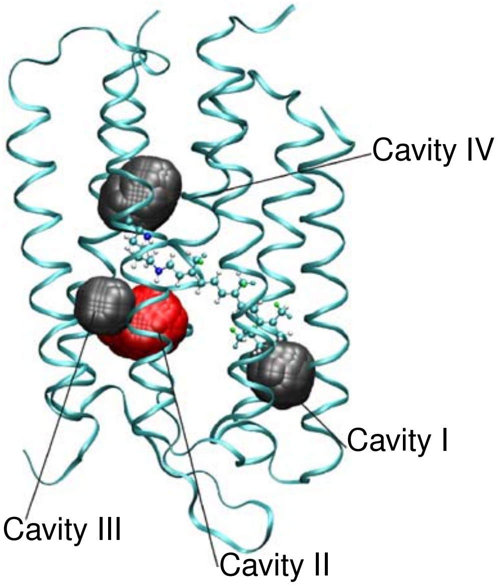

Fig. 8 Cavities found in bacteriorhodopsin. The red cavity fits three water molecules potentially involved in proton transfer in bacteriorhodopsin

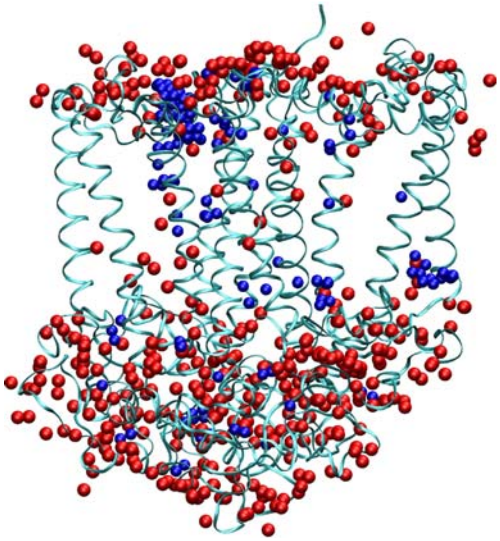

Fig. 9 Water molecules placed in the photosynthetic reaction center by the program McVol. Red spheres are crystallographically resolved water molecules, blue spheres are water molecules placed by McVol

Conclusion

In this work, we introduced a Monte Carlo algorithm for the calculation of protein volumes. Based on this algorithm, cavities inside the protein were located. The volume calculation are independent from any grid and therefore more accurate than the grid based methods developed so far.

The algorithm was applied to 15 proteins of different size. We found, that the ratio between the protein volume (including the volume of voids) and the number of atoms is almost the same for all sizes of proteins.

Our algorithm was able to reproduce experimentally derived cavities in the hen egg white lysozyme. Also the reported cavity volumes are in good agreement with our calculations. For bacteriorhodopsin, we could locate a cavity near the Schiff base maybe containing the water molecules important for the proton transfer process. An analysis of the cavities in the photosynthetic reaction center enabled us to place water molecules connecting originally separated proton transfer pathways through the protein. The Monte Carlo algorithm and the graph theoretical analysis of the protein volume, surfaces and cavities as well as the placement of water molecules is implemented in the program McVol. This program is able to calculate protein volumes, solvent accessible volumes and surfaces. McVol is available free of charge from our webpage http://www.bisb.uni-bayreuth.de/software.html.

Acknowledgements This work was supported by the DFG grant UL 174/7-1.

References

1. Kawabata T, Go N (2007) Detection of pockets on protein surfaces using small and large probe spheres to find putative ligand binding sites. Structure 68:516–529
2. Liang J, Edelsbrunner H, Woodward C (1998) Anatomy of protein pockets and cavities: measurement of binding site geometry and implications for ligand design. Prot Sci 7:1884–1897
3. Thornton JM, Todd AE, Milburn D, Borkakoti N, Orengo CA (2000) From structure to function: approaches and limitations. Nat Struct Biol 7:991–994
4. Kuntz ID, Blaney JM, Oatley SJ, Langridge R, Ferrin TE (1982) A geometric approach to macromolecule-ligand interactions. J Mol Biol 161:269–288
5. Delaney JS (1992) Finding and filling protein cavities using cellular logic operations. J Mol Graph 10:174
6. Warshel A (2002) Molecular dynamics simulations of biological reactions. Acc Chem Res 35:385–395
7. Till MS, Essigke T, Becker T, Ullmann GM (2008) Simulating the proton transfer in gramicidin a by a sequential dynamical Monte Carlo method. J Phys Chem, B 112:13401–13410
8. Bondar AN, Elstner M, Suhai S, Smith JC, Fischer S (2004) Mechanism of primary proton transfer in bacteriorhodopsin. Structure 12:1281–1288
9. Edelsbrunner H, Mucke EP (1990) Simulation of simplicity - a techique to cope with degenerate cases in geometric algorithms. ACM Trans Graph 9:66–104
10. Xie L, Bourne PE (2007) A robust and efficient algorithm for the shape description of protein structures and its application in predicting ligand binding sites. Bioinformatics 8
11. Hendlich M, Rippmann F, Barnickel G (1997) LIGSITE: automatic and efficient detection of potential small molecule-binding sites in proteins. J Mol Graph 15:359
12. Levitt DG, Banaszak LJ (1992) POCKET - A computer-graphics method for identifying and displaying protein cavities and their surrounding amino acids. J Mol Graph 10:229–234
13. Laskowski RA (1995) SURFNET - A program for visualizing molecular surfaces, cavities and intermolecular interactions. J Mol Graph 13:323
14. Allen MP, Tildesley DJ (1989) Computer simulation of liquids. Oxford University Press
15. Lee B, Richards FM (1971) Interpretation of protein structures - estimation of static accessibility. J Mol Biol 55:379
16. Eisenhaber F, Lijnzaad P, Argos P, Sander C, Scharf M (1995) The double cubic lattice method - efficient approaches to numerical-integration of surface-area and volume and to dot surface contouring of molecular assemblies. J Com Chem 16:273–284
17. Press WH, Teukolsky SA, Vetterling WT, Flannery BP (1992) Numerical recipes in C, 2nd edn. Cambridge University Press, Cambridge UK
18. Sedgewick (R) Algorithms in C++, part 5. Addison-Wesley, Boston
19. M. Masuya and J. Doi. Detection and Geometric Modeling of Molecular Surfaces and Cavities Using Digital Mathematical Morphological Operations. *J Mol Graph*, 13:331, 1995.
20. Peters KP, Fauck J, Frommel C (1996) The automatic search for ligand binding sites in proteins of known three-dimensional structure using only geometric criteria. J Mol Biol 256:201–213
21. Ruppert J, Welch W, Jain AN (1997) Automatic identification and representation of protein binding sites for molecular docking. Prot Sci 6:524–533
22. Brady GP, Stouten PFW (2000) Fast prediction and visualization of protein binding pockets with PASS. J Comput Aided Mol Design 14:383–401
23. Venkatachalam CM, Jiang X, Oldfield T, Waldman M (2003) Ligandfit: a novel method for the shape-directed rapid docking of ligands to protein active sites. J Mol Graph 21:289–307
24. Laurie ATR, Jackson RM (2005) Q-SiteFinder: an energy-based method for the prediction of protein-ligand binding sites. Bioinformatics 21:1908–1916
25. Calimet N, Ullmann GM (2004) The influence of a transmembrane ph gradient on protonation probabilities of bacteriorhodopsin: the structural basis of the back-pressure effect. J Mol Bio 339(3):571–589
26. Koepke J, Krammer E-M, Klingen AR, Sebban P, Ullmann GM, Fritzsch G (2007) pH modulates the quinone position in the photosynthetic reaction center from rhodobacter sphaeroides in the neutral and charge separated states. J Mol Biol 371:396–409
27. Klingen AR, Palsdottir H, Hunte C, Ullmann GM (2007) Redox-linked protonation state changes in cytochrome bc1 identified by Poisson-Bolt zmann electrostatics calculations. Biochem Biophys Acta 1767:204–221
28. Brooks BR, Bruccoleri RE, Olafson BD, States DJ, Swaminatha S, Karplus M (1983) CHARMM - A programm for macromolecular energy, minimization, and dynamics calculations. J Com Chem 4:187–217
29. Bondi A (1964) Van der Waals volumes and radii. J Phys Chem 68:441
30. Hiromoto T, Fujiwara S, Hosokawa K, Yamaguchi H (2006) Crystal structure of 3- hydroxybenzoate hydroxylase from comamonas testosteroni has a large tunnel for substrate and oxygen access to the active site. J Mol Biol 364:878–896

31. Borodich AI, Ullmann GM (2004) Internal hydration of protein cavities: studies on BPTI. Phys Chem 6:1906–1911
32. Takano K, Funahashi J, Yamagata Y, Fujii S, Yutani K (1997) Contribution of water molecules in the interior of a protein to the conformational stability. J Mol Biol 274:132–142
33. Otting G, Liepinsh E, Halle B, Frey U (1997) NMR identification of hydrophobic cavities with low water occupancies in protein structures using small gas molecules. Nat Struct Biol 4:396–404
34. Kodanadapani R, Suresh CG, Vijayan M (1990) Crystal-structure of low humidity tetragonal lysozyme at 2.1A resolution. J Biol Chem 265:16126–16131
35. Lanyi JK (2006) Proton transfers in the bacteriorhodopsin photocycle. Bioch et Biophys Acta - Bioener 1757:1012–1018
36. Heberle J (2000) Proton transfer reactions across bacteriorhodopsin and along the membrane. Bioch et Biophys Acta - Bioener 1458:135–147
37. Okamura MY, Paddock ML, Graige MS, Feher G (2000) Proton and electron transfer in bacterial reaction centers. Biochim Biophys Acta 1458:148–163
38. Stowell MHB, McPhillips TM, Rees DC, Soltis SM, Abresch E, Feher G (1997) Light-induced structural changes in photosynthetic reaction center: implications for mechanism of electron-proton transfer. Science 276:812–816
39. Paddock ML, Feher G, Okamura MY (2003) Proton transfer pathways and mechanism in bacterial reaction centers. FEBS Lett 555:45–50
40. Parkin S, Rupp B, Hope H (1996) Structure of bovine pancreatic trypsin inhibitor at 125K: definition of carboxyl-terminal residues Gly57 and Ala58. Acta Crystallogr, Sect.D 52:18–29
41. Jung A, Domratcheva T, Tarutina M, Wu Q, Ko WH, Shoeman RL, Gomelsky M, Gardner KH, Schlichting L (2005) Structure of a bacterial BLUF photoreceptor: insights into blue light-mediated signal transduction. Proc Natl Acad Sci USA 102:12350–12355
42. Brownlow S, Cabral JHM, Cooper R, Flower DR, Yewdall SJ, Polikarpov I, North ACT, Sawyer L (1997) Bovine Beta-lactoglobulin at 1.8 angstrom resolution - still an enigmatic lipocalin. Structure 5:481–495
43. Nogues I, Perez-Dorado I, Frago S, Bittel C, Mayhew SG, Gomez-Moreno C, Hermoso JA, Medina M, Cortez N, Carrillo N (2005) The Ferredoxin-NADP(H) reductase from rhodobacter capsulatus: molecular structure and catalytic mechanism. Biochemistry 44:11730–11740
44. Luecke H, Schobert B, Richter HT, Cartailler JP, Lanyi JK (1999) Structure of bacteriorhodopsin at 1.55 angstrom resolution. J Mol Biol 291:899–911
45. Retailleau P, Colloc'h N, Vivares D, Bonnete F, Castro B, El Hajji M, Mornon JP, Monard G, Prange T (2004) Complexed and ligand-free high-resolution structures of Urate Oxidase (Uox) from aspergillus flavus: a reassignment of the active-site binding. Acta Crystallogr, Sect.D 60:453–462
46. Andrade SLA, Dickmanns A, Ficner R, Einsle O (2005) Crystal structure of the archaeal ammonium transporter amt-1 from archaeoglobus fulgidus. Proc Natl Acad Sci USA 102:14994–14999
47. Fujimoto Z, Takase K, Doui N, Momma M, Matsumoto T, Mizuno H (1998) Crystal structure of a catalytic-site mutant alpha-amylase from bacillus subtilis complexed with maltopentaose. J Mol Biol 277:393–407
48. Brudler R, Hitomi K, Daiyasu H, Toh H, Kucho K, Ishiura M, Kanehisa M, Roberts VA, Todo T, Tainer JA, Getzoff ED (2003) Identification of a new cryptochrome class: structure, function, and evolution. Mol Cell 11:59–67
49. Wohlfahrt G, Witt S, Hendle J, Schomburg D, Kalisz HM, Hecht HJ (1999) 1.8 and 1.9 angstrom resolution structures of the penicillium amagasakiense and aspergillus niger glucose oxidases as a basis for modelling substrate complexes. Acta Crystallogr, Sect.D 55:969–977
50. Hohenester E, Maurer P, Timpl R (1997) Crystal structure of a pair of follistatin-like and EF-hand calcium-binding domains in BM-40. EMBO J 16:3778–3786
51. Seiffert GB, Ullmann GM, Messerschmidt A, Schink B, Kroneck PMH, Einsle O (2007) Structure of the non-redox-active tungsten/[4Fe : 4S] enzyme acetylene hydratase. Proc Natl Acad Sci USA 104:3073–3077