UPTEC X 02 018 ISSN 1401-2138
APR 2002

MARIA STRÖMGREN

EXCAVAT: A program for the calculation of solvent accessible surface area and the identification of cavities in proteins

Master's degree project

Molecular Biotechnology Programme
Uppsala University School of Engineering

UPTEC X 02 018 | **Date of issue 2002-04**

Author

Maria Strömgren

Title (English)

EXCAVAT: A program for the calculation of solvent accessible surface area and the identification of cavities in proteins

Title (Swedish)

Abstract

A program for the calculation of solvent accessible surface area (SASA) and the identification of putative binding sites in proteins has been developed based on the classification of grid points depending upon their shortest distance to the protein surface and their accessibility. Simple descriptors related to geometrical properties of these cavities are also calculated. Input to the program are the atomic coordinates in PDB format. The cavity finding algorithm was validated using a test set of 31 proteins from different functional classes. It was possible to identify the experimentally observed binding site for all test cases. An analysis of the ability of the descriptors to cluster these groups showed that different descriptors might be relevant for different groups of proteins.

Keywords

Solvent accessible surface area, cavities, binding sites, cavity identification, geometrical descriptors, adsorption chromatography

Supervisors

Dr. Enrique Carredano
Separations R&D, Amersham Biosciences

Examiner

Dr. Gerard Kleywegt
Dept. of Cell and Molecular Biology, Uppsala University

Project name	Sponsors
Language **English**	Security
ISSN 1401-2138	Classification
Supplementary bibliographical information	Pages **26**

Biology Education Centre Biomedical Center Husargatan 3 Uppsala
Box 592 S-75124 Uppsala Tel +46 (0)18 4710000 Fax +46 (0)18 555217

EXCAVAT: A program for the calculation of solvent accessible surface area and the identification of cavities in proteins

Maria Strömgren

Sammanfattning

I denna studie har EXCAVAT, ett program för beräkning av tillgänglig ytarea och identifiering av håligheter i proteiner, utvecklats. Genom beräkningen av den tillgängliga ytan av ett protein får man viktig information om vilka delar av proteinet som kan nås av lösningsmolekyler (t.ex. vatten). Bindning av små molekyler sker ofta i håligheter i proteiner, och därför är det intressant att hitta och undersöka håligheter. Metoden som använts i programmet bygger på att punkter, placerade runt hela proteinet, får olika värden beroende på deras tillgänglighet för vatten och andra molekyler. Dessa punkter utnyttjas sedan för att beräkna den tillgängliga ytarean för varje atom och för att identifiera håligheterna i proteinet. EXCAVAT beräknar också några geometriska värden som beskriver varje hålighet. Indata till programmet är filer som innehåller information om varje atoms position i proteinerna.

Examensarbete 20 p i Molekylär bioteknikprogrammet

Uppsala universitet april 2002

1. Introduction

1.1. Surface accessible surface area

A trend in biology today is the accumulation of structural data, increasing the need for efficient bioinformatics tools, that concentrate the information into few, but descriptive parameters. An example of this is the availability of an increasing number of protein structures in the Protein Data Bank (PDB) (Berman *et al.*, 2000) which has given great opportunities of studying geometrical aspects of proteins. For instance, several ways of describing the surface of a protein using the atomic coordinates have been developed. One of the most useful numerical descriptions of the exposure of the atoms of the protein is the solvent accessible surface area (SASA).

The SASA of an atom is defined (Lee and Richards, 1971) as the area of the accessible part of the surface at a distance of *R* (from the atom center), where *R* is the sum of the van der Waals radius of the atom (usually implicitly accounting for hydrogen atoms) and the radius of the probe (Figure 1). In other words, SASA describes the area over which contact between protein and solvent can occur.

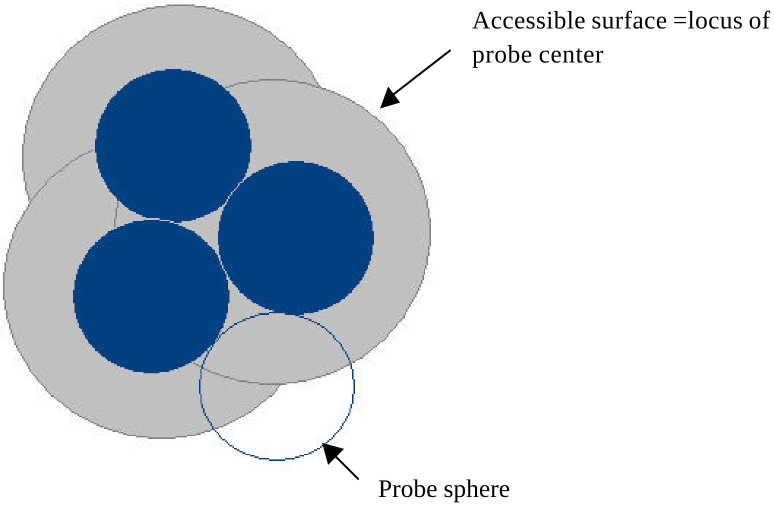


Figure 1. Definition of SASA. The solvent accessible surface is defined as the locus of the center of the desired solvent probe, if the latter is rolled over the surface. In the figure, the blue spheres symbolize the van der Waals volume of atoms. The border of the gray area symbolizes the solvent accessible surface.

Besides its value in the analysis of the structures and interactions of proteins and other molecules, SASA has been used in the study of the protein-folding problem and hydrophobicity studies (Connolly, 1996). The free energy of solvation is linearly related to SASA (Fraczkiewicz & Braun, 1998; Eisenberg & McLachlan, 1986).

1.1.1. Potential value of surface accessibility computations in ligand design

Here, the term cavity is used for all the following concepts in a protein context; holes, invaginations, tunnels, channels, depressions, indentations, voids, pockets and clefts.

Computation of surface accessibility also has importance in ligand design. Most binding sites for small ligands in proteins are cavities (Levitt & Banaszak, 1992; Ho & Marshall, 1990) and there is a connection between cavities and SASA since cavities are accessible only to small molecules, giving them a specific accessibility by imposing an upper limit of the probes (Figure 2). The specific accessibility could be used as an identification tag in the search of cavities and also probably in the search for potential ligands from small molecule databases.

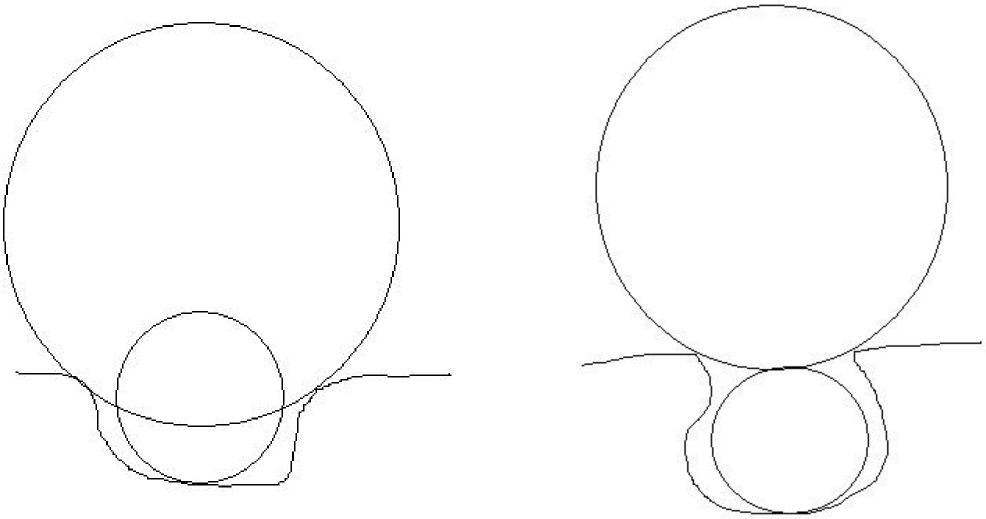


Figure 2. Cavities are characterized by a specific accessibility. Both cavities are partly accessible to a small solvent probe, but inaccessible to a large solvent probe.

The characteristics of a binding site, such as size, form, accessibility and the nature of surrounding amino acid residues are of greatest importance for the binding specificity. Using different numbers, related to attributes of this kind, might open the possibilities for a classification of cavities, where members of a class accommodate similar ligands. Visual inspection is probably the most complete tool for cavity analysis. However in data-mining contexts where it is necessary to search for matches from hundreds of proteins or protein complexes and hundreds of thousands or probably millions of compounds in small molecule databases visual inspection is of little use. Computational methods for analyzing cavities are therefore desirable.

Evolution tends to conserve structural features that are of importance to biological function, activity and specificity. Thus, binding sites in related proteins normally have related structures. Related proteins can be assumed to have similarities in the cavities at their binding sites. Using cavity descriptors, based on the geometrical properties of the cavity, it should be possible to find resemblance within a protein family. Therefore the concept of surface accessibility could also find applications in the field of proteomics.

1.1.2. Potential value of SASA in chromatography

The concept of SASA might also prove useful in the area of protein purification. Separation and purification of proteins is frequently accomplished using different adsorption chromatography techniques including ion-exchange chromatography, hydrophobic interaction chromatography (HIC), reversed phase chromatography (RPC) and affinity chromatography (http://www.chromatography.apbiotech.com, 27 Mar. 2002). All these methods achieve a separation of biomolecules by different types of interactions between ligands immobilized on a solid phase matrix and solute molecules during their travel through the column. In order to be able to predict the chromatographic behavior of proteins, a deeper theoretical understanding of the interactions of proteins in adsorption chromatography systems is necessary.

In ion-exchange chromatography, the separation of proteins depends mainly on electrostatic interactions (Amersham Pharmacia Biotech, 1999b). A theory, modeling effects of the retention factor of proteins in ion-exchange chromatography, has been proposed based on solutions of the Poisson-Boltzmann equation for oppositely charged surfaces (Ståhlberg *et al.* 1991; Jonsson & Ståhlberg, 1997). Solutions to the linear approximation of the Poisson-Boltzmann equation taking into account the charge distribution at the atomic level can be calculated (Honig & Nicholls, 1995), but are not easily obtained. Alternative approaches are therefore of interest. The number of charged groups and their positions in the proteins are probably important. Many charged groups are located on the surface, but some are inside the protein. As the electrostatic forces decrease with distance, charges at the surface probably have more influence than buried ones. Thus, the surface charge characteristics of the protein could be an important factor for the protein behavior in ion-exchange chromatography.

In HIC, hydrophobic groups of the solute molecules interact with the hydrophobic ligands coupled to the matrix, lowering the total entropy in accordance to the hydrophobic effect (Amersham Pharmacia Biotech, 1999a). In their biological environment, most proteins exist in hydrophilic surroundings, and in accordance with this, most hydrophobic groups are located inside the protein. The exceptions to this rule, the exposed hydrophobic groups, are important in HIC. The size of the hydrophobic effect is correlated with the amount of hydrophobic surface area buried by the interaction (Melander *et al.* 1989; Bartlett *et al.* in press), indicating a separation of proteins based directly on the hydrophobicity of the surface. Conformational changes of the structure due to denaturation must be considered, as this could change the exposure of different groups. With a stable protein structure, knowing the surface area of hydrophobic groups would give a good estimate of the protein behavior in HIC.

RPC also relies on hydrophobic interactions (Amersham Pharmacia Biotech, 1999c). However RPC adsorbents are more highly substituted with hydrophobic ligands than HIC adsorbents. Thus, the protein binding to RPC is usually stronger, requiring non-polar solvents for elution and increasing the risk of denaturation effects (McNay & Fernandez 2001; El Rassi *et al.* 1990). As a consequence of this, it is somewhat more complex to predict surface properties. However, to a good approximation the denaturation of many proteins may often be described by a simple two-state mechanism model (Shirley, 1992).

The behavior of the proteins in ion-exchange chromatography, HIC, and RPC might then be dependent on characteristics of the protein surface. In these kinds of adsorption chromatography, the processes are mainly of a stochastic nature and the entire surface is likely to be important. There is a need for a description of the surface that includes information on the accessibility of different groups. This is the kind of information that can be obtained directly from SASA calculations.

Affinity chromatography differs from the techniques described, as the interactions are of a more specific nature involving a particular patch of the protein. The protein is separated by reversible absorption to a ligand bound to the chromatographic media (http://www.chromatography.apbiotech.com, 27 Mar. 2002) under favorable binding conditions. Recovery of the molecules of interest can be accomplished by changing the conditions. A straightforward approach in affinity chromatography is to make use of the biological function of the protein, taking advantage of a known binding site for the separation with a ligand similar to the biologically active ligand. An example of this is purification of fusion proteins with MBP using its maltose-binding region with amylose (Hearn & Acosta, 2001). In the past, peptidic ligands have been the most commonly used in affinity chromatography. However, the drawbacks of peptides and proteins in terms of stability (especially during column cleaning procedures), have led to a search for small organic compounds as affinity ligands and in order to identify putative binding sites it is necessary to probe the protein surface for cavities. The concept of cavity descriptors mentioned above may prove useful in focusing the screening efforts towards a reduced number of candidates in the search for new affinity ligands.

1.1.3. The history of SASA

During the last decades, much work has been done in this area of SASA. In the calculations, atoms are commonly treated as hard spheres with fixed coordinates and van der Waals radii, which are frequently augmented to account for attached hydrogen atoms (Shrake & Rupley, 1973; Richards, 1974).

Lee & Richards (1971) developed a method for the calculation of SASA based on computing areas by multiplying arc lengths by the spacing between the planes. Shrake & Rupley (1973) developed a method where points were placed on expanded atomic spheres and the accessibility of each point was determined. Richards (1977) defined the concept of molecular surface (MS), which is closely related to the accessible surface. The MS is considered as the sum of two components, one of them being the molecular van der Waals surface that can be in contact with the surface of the probe, and the other component being those patches of the probe surface facing the interior. Connolly (1983) proposed an algorithm for calculating the MS, where the solvent molecule, modeled by a sphere, is used to generate a smooth outer-surface contour.

Voorintholt *et al.* (1989) described a method for visualizing protein surfaces, channels and cavities, using a grid in which every point contains a value that depends on the distance to the nearest atom. A faster version of the Shrake-Rupley algorithm was developed (Wang & Levinthal, 1991). A method for analytical calculation of SASA (Fraczkiewicz & Braun, 1998), by finding solvent-exposed vertices of intersecting atoms has been implemented in the web-based program GETAREA (http://www.scsb.utmb.edu/cgi-bin/get_a_form.tcl, 19 Apr. 2002).

1.2. Cavity identification

A number of different methods for detecting protein cavities, tunnels and potential sites for internal waters have been described (Connolly, 1996). Richards (1979) identified possible channels from the surface to the interior by studying connectivity of protein packing defects. With the method of Voorintholt (1989), cavities are shown as contours around volumes large enough to hold a probe with a certain radius. The method implemented in POCKET (Levitt & Banaszak, 1992) defines indentation, cavities and holes in a protein as points where a probe can fit, and identifies the surrounding amino acid residues. Delaney (1992) has described a method using cellular logic operations to find concave regions of the protein. Kleywegt & Jones (1994) developed VOIDOO, which finds, measures and displays cavities by the procedure of increasing atomic radii until a volume is closed off inside the increased protein. Thus, this method will not find all kinds of cavities. SURFNET (Laskowski, 1995) visualizes cavities, intermolecular interactions as well as molecular surfaces. In a study of packing defects (Hubbard & Argos 1995), an analysis of internal cavities of different protein groups highlighted problems in cavity detection. Stahl *et al.* (2000) have described a cavity mapping method as well as the prediction of enzyme class by a self-organizing neural network using SASA. In that study it was possible to classify and predict active site cavities among a set of proteins with zinc in the active site.

1.3. Goal of this work

The aim of this study has been to develop a program for the calculation of SASA, the identification of putative binding sites in the form of cavities and the extraction of simple descriptors related to geometrical properties of these cavities, the input to the program being the atomic coordinates in PDB format.

The method of choice is the use of a grid and the classification of the grid points depending upon their shortest distance to the protein surface. In this respect the approach resembles the method of Voorintholt and coworkers (1989). However a novel ingredient is the simultaneous classification of the grid points into maximal-accessibility classes which in combination with the shortest-distance classes provides a straightforward, plausible and practical definition of any cavity with relevance for the recognition of small molecules by proteins.

2. Materials and methods

2.1. Classification of grid points

A three dimensional grid covering the protein and its surroundings is created and every grid point is assigned two integers according to its maximal accessibility and its shortest distance to the protein surface respectively. This assignment corresponds to two different partitionings consisting of

i) *maximal-accessibility* classes and
ii) *shortest-distance* classes

The class assignment is carried out in two loops.

In the first loop, grid points within the van der Waals radius of every atom are identified. These grid points are assigned maximal-accessibility and shortest-distance zero.

For all other grid points, the maximal-accessibility integer n_{acc} is closely related to the maximum radius of a sphere, which may include the grid point without intersecting the protein volume. The shortest-distance integer n_{dist} is closely related to the maximum radius of a sphere centered on the grid point and not intersecting the protein volume.

The second loop (Figure 3) excludes the grid points identified in the first round. Before looping, all maximal-accessibility integers are set to zero. At every grid point *i*, the shortest distance R_{i} to the protein surface is calculated by first identifying the closest atom. From this the shortest-distance integer is calculated according to

$$n_{\mathrm{dist,i}} = \mathrm{Int}(R_{\mathrm{i}} / R_{\mathrm{P}}) + 1 \qquad \text{Eq. 1}$$

and

$$n_{\mathrm{dist,i}} = \mathrm{Max}(n_{\mathrm{dist,i}}, n_{\mathrm{max}}) \qquad \text{Eq. 2}$$

where R_{P}, the radius of the probe, is usually set to 1.4 Å (Richards, 1977) and n_{max}, the maximal size considered, is introduced for practical purposes. Before moving to the next grid point, all other grid points j within a sphere of radius R_{i} from grid point *i* are identified and assigned a maximal-accessibility integer according to

$$n_{\mathrm{acc,j}} = \mathrm{Max}(n_{\mathrm{dist,i}}, n_{\mathrm{acc,j}}) \qquad \text{Eq. 3}$$

2.1.1. Speed improvements

Assigning all grid points shortest-distance integers is time consuming because proteins consist of many atoms and the resolution of the grid has to be high (0.5-1.0 Å). In the most basic approach, every grid point has to be compared with every atom to ensure that the shortest distance to the protein surface is found. To accelerate the procedure, previous to the looping, the grid space is divided into cubic regions and the atoms located inside each region are identified. In the search for the closest atom, only atoms of the region of the grid point and neighboring regions are considered. To increase the speed even more, a coarser and a finer division into regions are prepared. When looping from one grid point to a neighboring one, the shortest distance to the protein surface of the previous step is stored and used to decide upon which division to take advantage of in the current step. In this way, the coarser division is used when the grid point is far away from the protein and the finer when close to the protein, reducing the number of atoms to be considered.

The speed of assigning the maximal-accessibility integers is improved in the case when the previous (neighbor) grid point belongs to the shortest-distance integer corresponding to $n_{\max}$. Only about half of the grid points within the sphere of radius R_i have to be looped through then as the other half have already been assigned a larger or equal maximal-accessibility integer.

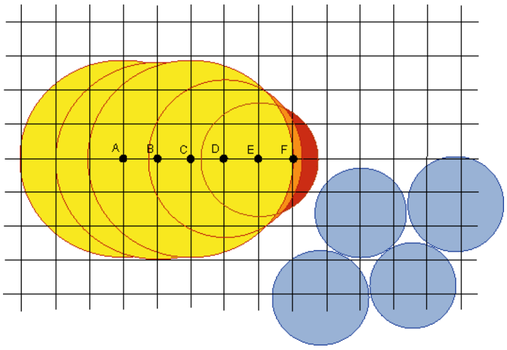


Figure 3. Looping through all grid points. When looping from grid point A to grid point F, the distance to the closest atom is calculated, and the grid point is classified to the appropriate shortest-distance integer. This corresponds to the size of the largest sphere that the point could be the centre of. All grid points within each sphere are assigned a maximal-accessibility integer from 4 to 2, in accordance to the size of the sphere. Blue circles represent atoms.

2.2. Solvent accessible surface area

When all grid points have been assigned a maximal-accessibility integer, the accessible area of each atom is easily estimated, by defining neighbor grid points to the atom as the grid points within a shell with a certain thickness outside the van der Waals surface (Figure 4). The SASA of an atom is calculated as:

$$\mathrm{SASA} = 4\pi R^2 * (\mathrm{N}_{\mathrm{acc}} / \mathrm{N}_{\mathrm{tot}}), \qquad \text{Eq. 4}$$

where $\mathrm{N}_{\mathrm{acc}}$ and $\mathrm{N}_{\mathrm{tot}}$ are the number of accessible and total number of points in the shell respectively and where R equals the sum of the van der Waals radius of the atom and the probe radius R_{P}. The SASA of an amino acid residue is calculated as the sum of the SASA of the atoms in the residue. The integer $\mathrm{N}_{\mathrm{acc}}$ is calculated as the number of grid points with maximal-accessibility integer greater than or equal to two.

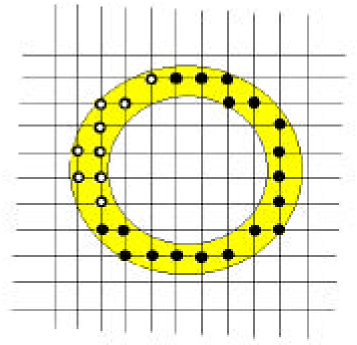


Figure 4. Calculation of SASA. The solvent accessible area of an atom is calculated as: SASA=$4\pi R^2$ *(number of accessible points in shell /total number of points in shell), where R equals the sum of the van der Waals radius and the probe radius. The central circle represents the atomic van der Waals volume and the yellow area represents the shell used for the calculation. In the figure, the white points might be thought of as accessible and the black inaccessible.

2.2.1. Possibility of modeling the accessibility of missing atoms

An often-encountered problem in the calculation of SASA is that PDB structures sometimes lack information about atomic coordinates of some amino acid residues. The missing residues are frequently exposed, and their accessibility might therefore be crucial for different applications. However, the accessibility of these absent atoms can easily be modeled.

A 100-residue peptide consisting of five copies of each natural amino acid type in a random order was built using BIOPOLYMER (Tripos, 2002). After energy minimization SASA was calculated for every atom and average values for each residue type were recorded. These values can then be used as estimates for the SASA of atoms for which coordinate information is absent.

2.3. Cavity identification

As a result of visual inspection of different combinations, cavities are identified using the grid points with shortest-distance integer equal to two or three but with a maximal accessibility integer not larger than three *i.e.*

$$(n_{\text{dist}} = 2 \vee n_{\text{dist}} = 3) \wedge (n_{\text{acc}} \leq 3) \qquad \text{Cond.1}$$

These grid points are then clustered in the following way. All grid points satisfying Cond.1 are put into a list and the first grid point is compared with every other grid point in the list. The first grid point and all grid points within a certain distance D from the first grid point are regarded as cluster 1, and removed from the list. This procedure is repeated generating clusters 2,3,4 etc until the list is empty.

In the next step, the clusters are re-clustered. For every cluster, the average of the coordinates of the cluster points is computed. These central coordinates are used to identify the closest neighbor cluster. If there is at least one point in the cluster, at a distance smaller than *D* to a point in the neighbor cluster, the two clusters are merged. Then, all clusters with fewer than 4 grid points are removed and the re-clustering procedure is repeated. All clusters with fewer than 25 grid points are removed and re-clustering is carried out once more.

Finally each one of the final clusters is considered to be a cavity. Cavities are then sorted according to the number of grid points.

The development of the cavity clustering method was performed with default parameters. As parameters (such as number of grid points) are grid dependent, modifications will be necessary for the use of other grid spacings.

2.4. Calculation of geometrical cavity descriptors

In order to get a relevant description of the surroundings of a cavity, all amino acid residues containing at least one atom center within a distance of 4 Å from a cavity point are identified as the residues forming the cavity.

A number of geometrical cavity descriptors were defined (Table 1). The descriptors have been developed and analyzed using default values of parameters. As the NRGR descriptor is grid-dependent, this descriptor has to be modified to get comparable values for other grid spacings. The calculation of these descriptors from their definition is straightforward with one exception. The mean of the estimated distance to closest atom (MDCA) is calculated by assuming that the grid-points in the cavity with shortest-distance integer $n_{\mathrm{dist,i}}$ are on average a distance of ($n_{\mathrm{dist,i}}$ + 0.5) probe radii from the surface.

Cavity descriptor	**Definition**
MDCA	Mean of estimated distance to the protein surface for all grid points in the cavity
MACC	Mean of maximal-accessibility integers of grid points in the cavity
NRGR	Number of grid points in the cavity
ACCA	Sum of SASA of amino acids forming the cavity
MACCA	Mean of ACCA

Table 1. Cavity descriptors and their definition. A more detailed explanation of how the estimated distance to the protein surface is calculated is given in the running text.

2.5. Validation methods

2.5.1. Validation of class assignment and SASA

The web-based program GETAREA was used to compare SASA calculations in terms of correlation coefficients with several structures, among those carboxylic ester hydrolase (PDB id: 1une) and phospholipase A2 (PDB id: 1vip). In order to test the method for orientation-dependency, SASA values from a set of ten rotated structures of the structure 1vip were compared with the values from GETAREA.

SYBYL (Tripos, 2002) was used to visualize results from the class assignment of grid points (n_{dist} and n_{acc}) and to verify the identification of cavities. For this purpose the grid points were written out as oxygen atoms in PDB format. The appropriate identification of surrounding amino acid residues was verified as well.

2.5.2. Validation of cavity identification

SYBYL and MOLCAD (Tripos, 2002) were used to visualize identified cavities. The structures of two complexes, one of lysozyme in complex with NAG (PDB id: 1lzb) and one of streptavidin in complex with biotin (PDB id: 1stp) were used to see if the position of the ligands coincide with the positions of the cavity grid points.

Further validations of the cavity identification algorithm and cavity were performed using 31 protein structures, belonging to different protein functional classes (Table 2) including serine proteases, glycosyl hydrolases, phospholipase A2, retinol-binding and biotin-binding proteins, alkaline phosphatases, galactose-binding proteins and dihydrofolate reductase. Within each class, proteins from different organisms were chosen. Some protein structures are complexes with ligands or inhibitors. Structures with mutations were present among the chosen structures.

The observed binding site of all protein structures was compared with the calculated cavities in order to identify the cavity, if any, that corresponded to the binding site. Then descriptors were calculated for this cavity and the relevance of the descriptors for the clustering of the proteins into groups was analyzed.

2.5.3. Validation of relevance of cavity descriptors

The 31 structures were used to study the relevance of the cavity descriptors as parameters related to clustering of different protein groups. Ideally this study should be done with groups containing a large number of members since the number of data points affects the spread of the data. If this number is small then even random numbers may give the appearance of clustering. To illustrate this, consider a set of *N* points uniformly distributed along a line and randomly selected subsets. Under these conditions, subsets of two points will most probably have a range smaller than the total range. Two points have a larger probability of being close to each other by chance than three points etc. In the extreme case of one point one obtains the spread of zero. To compensate for the small numbers a corrected relative range (CRR) was calculated as the relative range normalized by the relative range expected from random numbers. CRR was calculated as:

$$\mathrm{CRR} = \text{relative range} / \text{Random relative range} \qquad \text{Eq. 5}$$

where

Relative range = Range of descriptor within protein group / Total range of descriptor for all proteins Eq. 6

and

Random relative range = 1-(1/*n*) Eq. 7

where *n* is the number of investigated structures within the protein group. A hand-waving derivation of Eq 7 is as follows (see Figure 5). Given the event of choosing *n* real numbers between 0 and 1 and assuming a uniform distribution of probabilities then the expected value of the distance between two of these numbers is 1/*n*. Then the expected value of the range of the sample (largest minus smallest value) is

Expected sample range = (No. of distances)*(length of each distance)
=(n-1)*(1/n)
=(1-(1/n)) Eq. 8

and the random relative range should be this divided by the total range that is this case is one leading to Eq. 7.

A principal component analysis (PCA) using all descriptors was attempted using SYBYL's molecular spreadsheet, retaining the largest components.

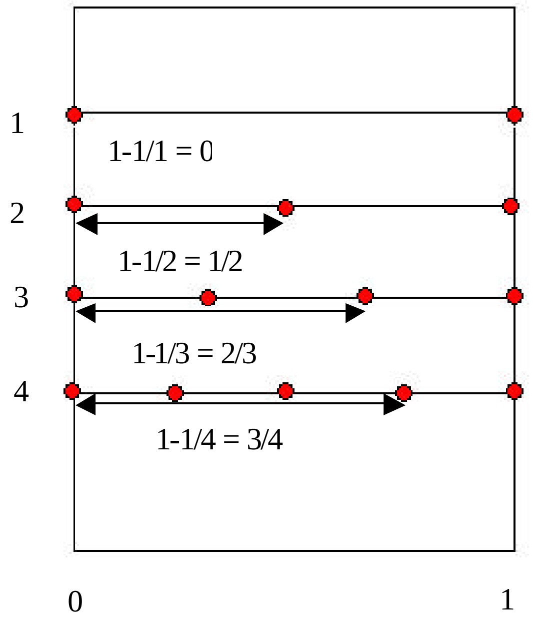


Figure 5. Expected range of a sample from a uniform distribution. To the left the sample density.

PDB id	Functional class	Source	Special notes
1acb	Chymotrypsin	Bovine	Alphachymotrysin complex with inhibitor
1afq	Chymotrypsin	Bovine	Gammachymotrypsin complex with synthetic inhibitor
1bp4	Papain	Papaya	Complex with aldehyde inhibitor
1amh	Trypsin	Rat	
1aq7	Trypsin	Bovine	Complex with inhibitor
1ezm	Elastase	Pseudomonas aeruginosa	
1amy	Alpha amylase	Barley seeds	
1bli	Alpha amylase	Bacillus licheniformis	Mutant N190F, Q264S, N265Y
1jfh	Alpha amylase	Pig	Complex with methyl 4,4'-dithio -alpha-maltotrioside
1lzb	Lysozyme	Hen	Complex with NAG
153l	Lysozyme	Goose	Lacks a catalytic aspartate
1cu6	Lysozyme	Bacteriophage T4	Mutant L91A, complex with HED
1cel	Cellobiohydrolase	Trichoderma reesei	Complex with O-iodo-benzyl-1-thio-B-D-glucose
1bvw	Cellobiohydrolase	Humicola insolens	Complex with NAG and mannose
1bp2	Phospholipase A2	Bovine	Complex with MPD
1clp	Phospholipase A2	Terciopelo	Mutant D49K
1db4	Phospholipase A2	Human	Complex with indole 8
1vip	Phospholipase A2	Vipera ruselli	
1crb	Retinol binding protein	Rat	Complex with retinol
1ggl	Retinol binding protein	Human	
1hbp	Retinol binding protein	Bovine	Holo form, complex with retinol
1brp	Retinol binding protein	Human	Holo form, complex with retinol
1stp	Biotin binding protein	Streptomyces	Streptavidin complex with biotin
1avd	Biotin binding protein	Hen	Avidin complex with biotin
1b8j	Alkaline phosphatase	E. coli	Complex with vanadate
1ew2	Alkaline phosphatase	Human	Complex with NAG
1gof	Galactose oxidase	Dactylium dendroides	
1eut	Sialidase	Micromonospora viridifaciens	Complex with galactose
1ai9	Dihydrofolate reductase	Yeast	Complex with NADPH
1cz3	Dihydrofolate reductase	Thermotoga maritima	
1ddr	Dihydrofolate reductase	E. coli	Complex with methotrexate,urea

Table 2. Protein structures chosen for the validation of cavity identification and relevance of cavity descriptors.

3. Results and Discussion

3.1. Validation of grid point accessibility

Careful visual inspection and several manual measurement of distances revealed no error in the classification of grid point accessibility and distance to the surface. All grid points, accessible to probes with the radius of water or larger and with shortest-distance integer equal to two, cover as might be expected the surface of the protein (Figure 6).

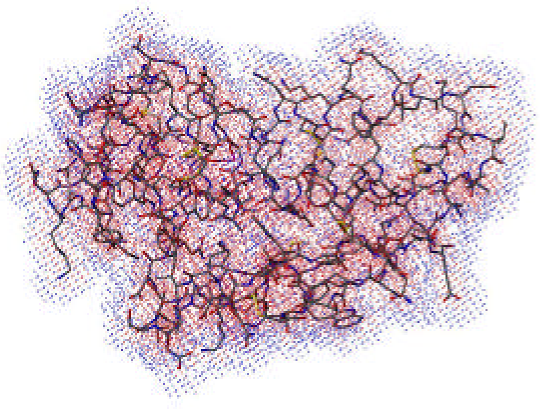


Figure 6. Visual validation of classification of grid points. Grid points (visualised as pseudo-atoms in SYBYL) defining the surface of the protein (blue) and inside atomic volumes (red) in carboxylic ester hydrolase (PDB id: 1une).

3.2. Validation of SASA

Comparing the results of SASA calculations with those from the program GETAREA yields correlation coefficients of 0.86 and 0.94 when the calculations were carried out per atom and amino acid residue respectively for the structures with PDB id 1une and 1vip Figure 7). Results from other structures were in agreement with these values (data not shown). From linear regression analysis it was found that the calculated SASA per atom from EXCAVAT was about 1.5 times larger than that from GETAREA. A possible reason for this overestimation is the finite thickness of the shell sampling grid points and their accessibility for the definition of SASA (Figure 8). In the limit of infinite grid point density and infinitely thin shell the program would probably converge to more precise and also more accurate SASA values. However time requirements prohibit a denser grid than about 0.5 Å and consequently the sampling shell must be thick enough to give good statistics. As it is the program gives a reasonable estimation.

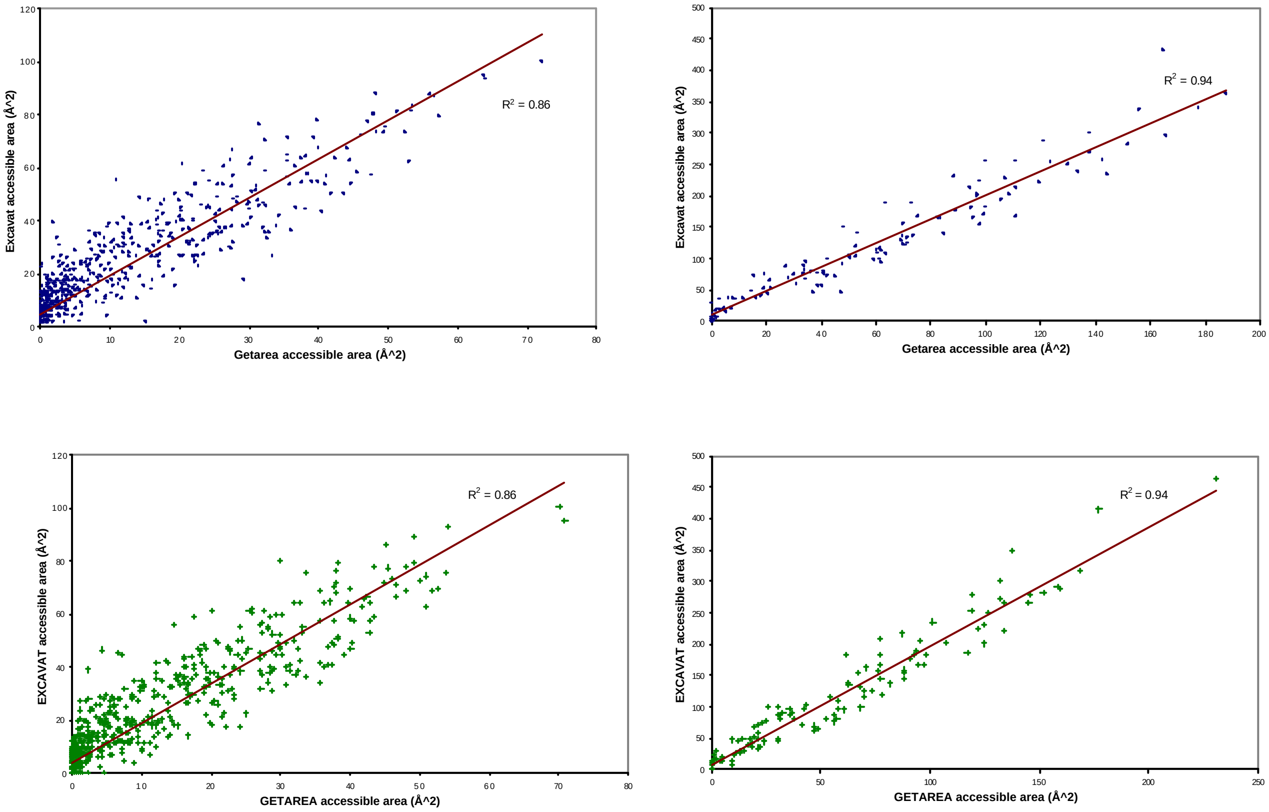


Figure 7. Comparison with SASA values from the program GETAREA. Left: SASA per atom. Right: SASA per residue. The upper figures show values for 1une and the lower show values for 1vip.

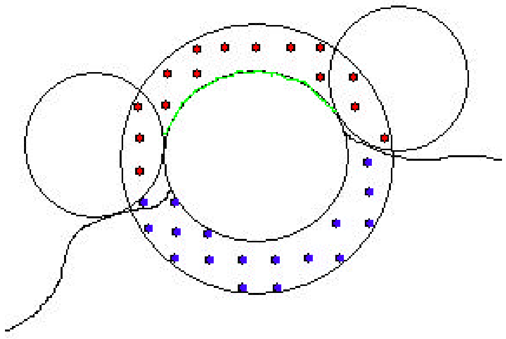


Figure 8. Shell effect possible cause for large SASA. The green part of the atom corresponds to the accessible part of the van der Waals surface. However, by using a shell with a thickness, a larger proportion of the atom is regarded accessible as illustrated. Blue dots represent inaccessible and red accessible grid points.

A comparison of ten different rotations of phospholipase A2 (PDB id: 1vip) show fluctuations in single atomic values and the correlation coefficients vary between 0.85 and 0.87. Trend lines, based on linear regression analyses, from each set show similar behavior (Figure 9). Comparing the average SASA values of each atom for the set of rotated structures with the GETAREA values yield slightly higher correlation coefficients than correlation coefficients from individual structures (Figure 10).

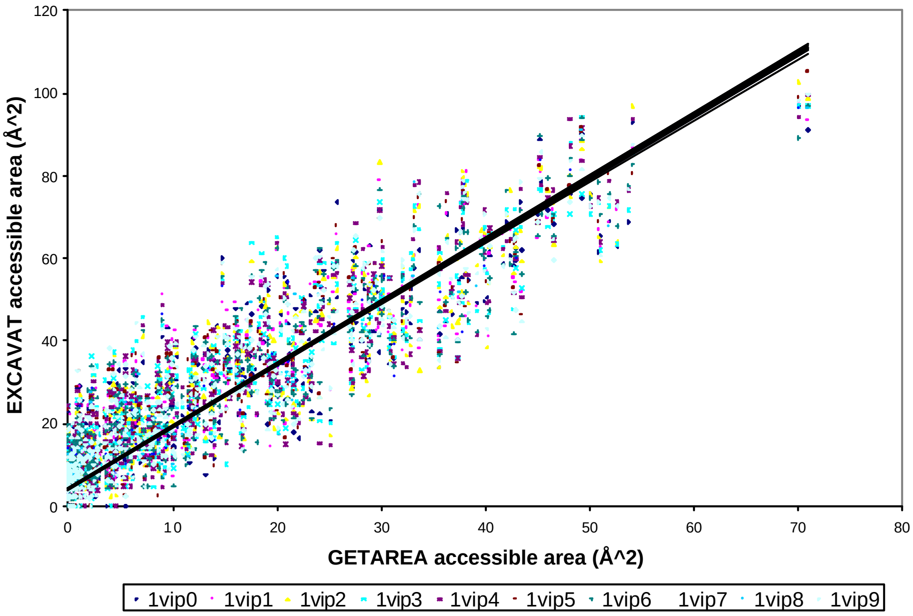


Figure 9. Results from rotated structures of phospholipase A2. The results from the ten differently rotated structures of 1vip show large spread for individual atoms. However, the trendlines almost coincide and correlation coefficients (not shown) are similar.

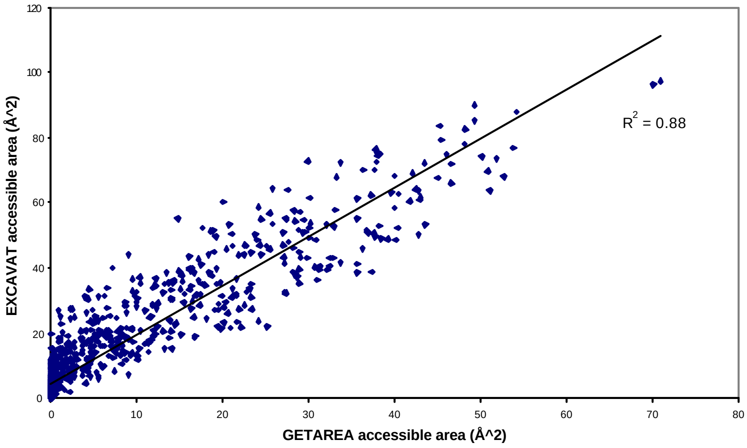


Figure 10. Results from average of rotated structures. The average of the rotated structures of 1vip show a slightly higher correlation than the individual structures.

3.3. Cavity identification

The position of the grid points corresponding to the largest cavities in lysozyme in complex with NAG coincides roughly with the position of the ligand (Figure 11). The position of the grid points corresponding to the only cavity in streptavidin coincides very well with the position of biotin (Figure 12).

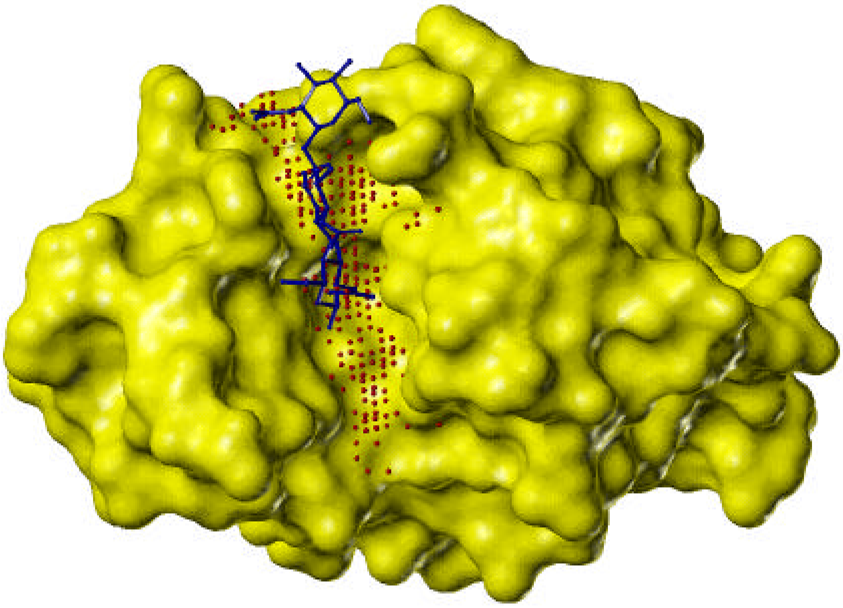


Figure 11. Cleft in lysozyme. The cleft in lysozyme (PDB id: 1lzb) is filled with computed cavity points (red). The ligands (NAG) in the complex bind in this site.

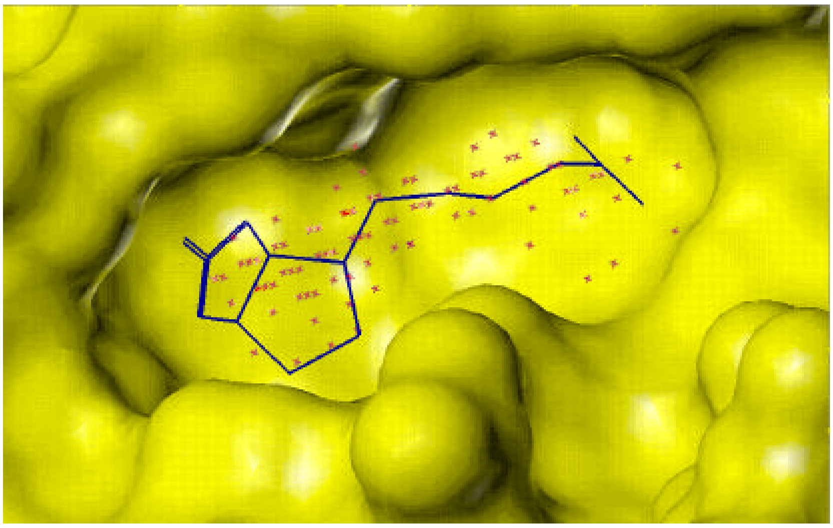


Figure 12. Cavity in streptavidin. The binding-site of biotin in streptavidin (PDB id: 1stp) is identified with the computed cavity points (red).

In all 31 investigated protein structures, the observed binding site was identified as a cavity. In most of the cases it corresponded to the largest identified cavity (Table 3).

PDB id	Family	Cavity rank in terms of size	Total number of cavities
1acb	Chymotrypsin	1	10
1afq	Chymotrypsin	1	13
1bp4	Papain	2	8
1amh	Trypsin	1	10
1aq7	Trypsin	4	8
1ezm	Elastase	1	11
1amy	Alpha amylase	2	16
1bli	Alpha amylase	1	16
1jfh	Alpha amylase	1	17
1lzb	Lysozyme	1	4
153l	Lysozyme	1	5
1cu6	Lysozyme	1	4
1cel	Cellobiohydrolase	2	19
1bvw	Cellobiohydrolase	1	11
1bp2	Phospholipase A2	1	5
1clp	Phospholipase A2	1	12
1db4	Phospholipase A2	1	4
1vip	Phospholipase A2	1	4
1crb	Retinol binding protein	1	7
1ggl	Retinol binding protein	1	19
1hbp	Retinol binding protein	1	12
1brp	Retinol binding protein	1	9
1stp	Biotin binding protein	1	1
1avd	Biotin binding protein	3 (one chain)	11
1b8j	Alkaline phosphatase	22 (one chain)	37
1ew2	Alkaline phosphatase	2	8
1gof	Galactose oxidase	16	25
1eut	Sialidase	9	18
1ai9	Dihydrofolate reductase	1	13
1cz3	Dihydrofolate reductase	1	9
1ddr	Dihydrofolate reductase	1 (one chain)	12

Table 3. Cavity identification. In all the protein structures, a cavity corresponding to the binding site was found. In most cases, the largest cavity (containing most grid points) corresponded to the observed binding site.

3.4. Cavity descriptor relevance

Diagrams based on values obtained from the descriptor calculation for all 31 proteins are given in Appendix B. A table with the average and corrected random range for each group is given in Appendix C. The corrected random range CRR for all descriptors and protein groups is depicted (Figure 13) and the following trends can be observed. Some descriptors are apparently relevant for clustering certain functional classes of proteins. Considering "CRR<0.2" as the relevance criterium then MDCA (related to distance to surface) is relevant for chymotrypsin, trypsin, cellobiohydrolase and phospholipase A2; MACC (related to accessibility) is relevant only for chymotrypsin. NRGR (related to cavity volume) is relevant for phospholipase A2, biotin binding proteins and alkaline phosphatase. ACCA (related to cavity surface) is relevant for lysozyme, phospholipase A2 and alkaline phosphatase and MACCA (related to surface per residue) is relevant for lysozyme and dihydrofolate reductase.

The group of proteins, which cluster with respect to the largest number of descriptors (three), is phospholipase A2. The structural similarity among the active sites of the enzymes belonging to this group has been highlighted (Carredano *et al*., 1998). However, according to a sequence alignment using CLUSTALW (Thompson, 1994; http://www.ebi.ac.uk/clustalw, 19 Apr 2002), the sequences of the chosen phospholipases are only about 25% identical. Serine proteases in general, alpha amylases, as well as retinol and galactose binding proteins do not show any clustering behavior in the space defined by these descriptors and set of proteins.

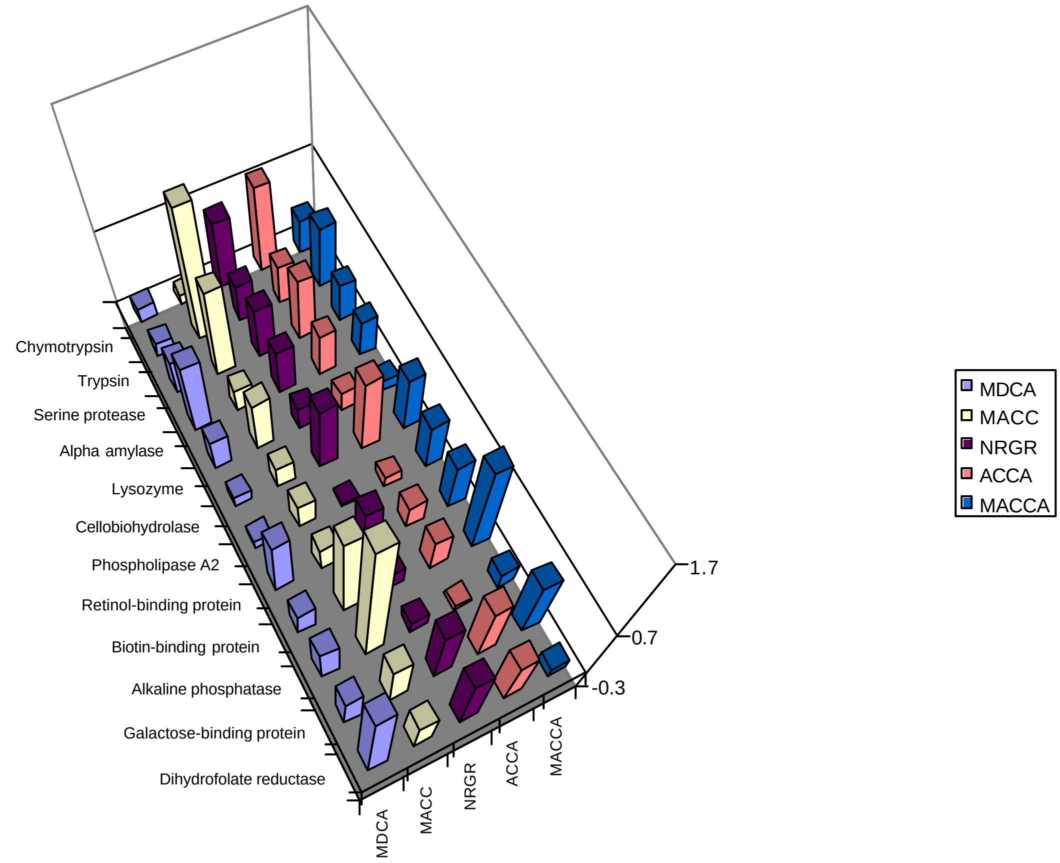


Figure 13. CRR of the cavity descriptors. For each protein group, the CRR is shown as a bar.

Relations between descriptors are also seen, most of which show a non-linear character. MDCA and MACC are related, both holding information on the accessibility in the cavity. ACCA and NRGR are different measures of the size of the cavity, and thus related. MACCA appears to be related to ACCA and NRGR.

Results from the PCA analysis are shown as diagrams in Appendix D. A weak tendency to clustering behavior is observed, in which phospholipase A2 cluster particularly well.

3.5. Performance

One of the disadvantages of any grid method is the dependence on the grid spacing and the grid location, limited by time and memory requirements. A medium-sized protein requires a couple of minutes with a grid spacing of 1 Å. As an attempt to increase the resolution without large time losses, the calculations were performed twice, with a shift in the grid placement, but no improvement was seen.

3.6. Concluding remarks

The calculations of SASA are carried out at a reasonable precision (rather high correlation coefficients) and speed. However, the SASA obtained by EXCAVAT is about 1.5 times too large (as compared to results from GETAREA). In future versions this overestimation could be corrected for. One can also think of this method as providing a slightly different measure of accessibility than SASA, which does not necessarily have to be less useful. Also, this method can very easily be modified to estimate the area of the molecular surface, as defined by Richards (1977).

It is also possible to define SASA in a more refined way by using the information stored in the maximal-accessibility integer of the grid points in the definition shell (section 2.2). Different definitions could be of use for different applications where knowledge of the size (and shape) of the physical “probe” exists. Further investigations of the possible benefits of different SASA definitions are needed.

A weighted accessibility definition might find use in the area of chromatography. It is necessary to try to correlate chromatographic data with SASA obtained using different weighting schemes to try to identify the most appropriate one.

The effects from surface character might be due to the smoothness of the surface. Using this method, one can easily think of and calculate different descriptors describing the smoothness, to give three examples, the number of cavities, the number of grid points in cavities and the number of grid points in cavities per total surface area. An alternative measure could be the accessibility in surface points. Such descriptors could be of particular importance to adsorption chromatography where the complementarity of surfaces is important, for instance, in the case of HIC and RPC, where the burial of hydrophobic areas influences the chromatographic behavior.

The methods developed in this study are a response to the need for bioinformatics tools to extract information from the growing amount of 3D structural information as both the SASA and the cavity descriptors are numerical extracts from the whole protein structure. Interestingly, this type of methods might be fruitfully combined with other bioinformatics tools such as homology modeling to produce descriptors for the even larger set of proteins where only primary structure informatio n is available.

The method for cavity identification has enabled the identification of the binding sites in the entire set of protein structures chosen. As pointed out (Kleywegt & Jones) the extent of any cavity is a subjective matter. Here the definition was not based on the identification of "holes" or "pockets" as such but rather on grid points around and inside the protein allowing the volume of a probe with radius in a specific range. In the case of external cavities this translates into focusing on groups of grid points of certain accessibility. In a molecular recognition context, specific accessibility is likely to be a motive for the location of binding sites in cavities, opening possibilities for an objective definition of a binding site.

One good approach for immediate comparison of cavity structures would be mapping different cavities onto each other. This kind of super-positioning only makes sense if there is a unique direction, which can be related to all cavities. One possibility would be to define such direction in terms of gradient of accessibility. Using the geometrical center of the cavities as origin, a unique transformation of the coordinates of the cavity grid points could be defined, enabling the desired comparison.

In order to get sufficient material for a complete statistical analysis, a comparison of more proteins from different structural classes would be needed. A comparison among several structures of the same protein and very different structures with identical function could yield a deeper understanding of the relationship between structure and the cavity descriptors as well as function and the cavity descriptors. Care must be taken when including mutants and inhibitor complexes that no significant changes in conformation have taken place.

The matter of the design of relevant cavity descriptors is complex. Thus further studies in this respect are needed. Using the rather simple descriptors defined, protein structures with high structural similarity in their binding site, such as phospholipase A2, showed good clustering behavior, whereas other groups of proteins like cellobiohydrolases did not cluster well. However, the descriptors developed here offer a starting point in the search for information rich parameters describing cavities. Other descriptors, holding information not only of the geometrical properties of the cavity but also of chemical properties such as charge and hydrophobicity of surrounding amino acid residues may also be included. The combination of these chemical descriptors with the geometrical descriptors may prove to be extremely useful.

4. Acknowledgements

I would like to thank my supervisor Dr Enrique Carredano for his support and enthusiasm in the work with this project.

I would also like to thank all the people in the sections Biochemistry and Chromatographic characterization at Amersham Biosciences for making my time here pleasant!

Thank you, Dan, for listening to all my questions and thoughts during this work.

5. References

Amersham Pharmacia Biotech (1999a) Hydrophobic interaction chromatography. Principles and Methods, 18-1020-90

Amersham Pharmacia Biotech (1999b) Ion exchange chromatography. Principles and Methods, 18-1114-21

Amersham Pharmacia Biotech (1999c) Reversed phase chromatography. Principles and Methods, 18-1134-16

Bartlett, P.A., Yusuff, N., Rico, A.C. & Lindvall, M.K. (2002) Antihydrophobic solvent effects: an experimental probe for the hydrophobic contribution to enzyme-inhibitor binding. *J Am Chem Soc*, in press

Berman H.M., Westbrook, J., Feng, Z., Gilliland, G., Bhat, T.N., Weissig, H., Shindyalov, I.N. & Bourne, P.E. (2000) The Protein Data Bank. *Nucleic Acids Research*, 28, 235-242

Carredano, E., Westerlund, B., Persson, B., Saarinen, M., Ramaswamy, S., Eaker, D. & Eklund, H. (1998) The three-dimensional structures of two toxins from snake venom throw light on the anticoagulant and neurotoxic sites of phospholipase A2. *Toxicon*, 36, 75-92

Connolly, M.L. (1983) Analytical molecular surface calculation. *J Appl Cryst*, 16, 548-558

Connolly, M.L. (1996) Molecular surfaces: a review. http://www.netsci.prg/Science/Compchem/feature14.html, 15 Mar 2002

Delaney, J.S. (1992) Finding and filling protein cavities using cellular logic operations. *J Mol Graphics*, 10, 174-177

Eisenberg D. & McLachlan, A.D. (1986) Solvation energy in protein folding and binding. *Nature*, 316, 199-203

El Rassi, Z., Lee, A.L. & Horvath, C. (1990) Reversed-phase and hydrophobic interaction chromatography of peptides and proteins. *Bioprocess Technol*, 9, 447-94

Fraczkiewicz, R. & Braun W. (1998) Exact and efficient analytical calculation of the accessible surface areas and their gradients for macromolecules. *J Comp Chem*, 19, 319-333

Hearn, M.T.W. & Acosta, D. (2001) Applications of novel affinity cassette methods: use of peptide fusion handles for the purification of recombinant proteins. *J Mol Recognit*, 14, 323-369

Ho, C.M.W. & Marshall, G.R. (1990) Cavity search: an algorithm for the isolation and display of cavity-like binding regions. *J Comp-Aided Molecular Design*, 4, 337-354

Honig B. & Nicholls A. (1995) Classical electrostatics in biology and chemistry. *Science*, 268, 1144-9

Hubbard, S.J. & Argos P. (1995) Detection of internal cavities in globular proteins. *Prot Eng*, 8, 1011-1015

Jonsson, B. & Ståhlberg, J. (1999) The electrostatic interaction between a charged sphere and an oppositely charged planar surface and its application to protein adsorption. *Colloids Surf, B* 14, 67-75.

Kleywegt, G.J. & Jones, T.A. (1994) Detection, delineation, measurement and display of cavities in macromolecular structures. *Acta Cryst*, D50, 178-185

Laskowski, R.A. (1995) SURFNET: A program for visualizing molecular surfaces, cavities, and intermolecular interactions. *J Mol Graphics*, 13, 323-330

Lee, B. & Richards, F.M. (1971) The interpretation of protein structures: estimation of static accessibility. *J Mol Biol*, 5, 379-400

Levitt, D.G. & Banaszak, L.J. (1992) POCKET: A computer graphics method for identifying and displaying protein cavities and their surrounding amino acids. *J Mol Graphics*, 10, 229-234

McNay, J.L. & Fernandez, E.J. (2001) Protein unfolding during reversed-phase chromatography: I. Effect of surface properties and duration of adsorption. *Biotechnol Bioeng*, 76, 224-232

Melander, W.R., el Rassi Z. & Horvath C. (1989) Interplay of hydrophobic and electrostatic interactions in biopolymer chromatography. Effect of salts on the retention of proteins. *J Chromatography*, 469, 3-27

Richards, F.M. (1974) The interpretation of protein structures: total volume, group volume distributions and packing density. *J Mol Biol*, 82, 1-14

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

Richards, F. M. (1979). Packing defects, cavities, volume fluctuations, and access to the interior of proteins. *Carlsberg Res. Commun.*, 44, 47-63.

Shirley, B. A. (1992) Ch. 6 *in* Tim J. Ahren and Mark C. Manning [eds]. Stability of Protein Pharmaceuticals. Part A: Chemical and Physical Pathways of Protein Degradation. Plenum Press, New York.

Shrake A. & Rupley J.A. (1973) Environment and exposure to solvent of protein atoms. Lysozyme and Insulin. *J Mol Biol*, 79, 351-371

Stahl, M., Taroni, C. & Schneider, G. (2000) Mapping of protein surface cavities and prediction of enzyme class by a self-organizing neural network. *Prot Eng*, 13, 83-88

Ståhlberg J., Jonsson, B. & Horvath, C. (1991) Theory for electrostatic interaction chromatography of proteins. *Anal Chem*, 63, 1867-74

Thompson JD, Higgins DG, Gibson TJ. (1994) "CLUSTAL W: improving the sensitivity of progressive multiple sequence alignment through sequence weighting, position-specific gap penalties and weight matrix choice."; Nucleic Acids Res. 22, 4673-4680

Tripos Associates, Inc. (2002), SYBYL molecular modelling software. St. Louis, MO

Voorintholt, R., Kosters, M.T., Vegter, G., Vriend, G. & Hol, W.G.J. (1989) A very fast program for visualizing protein surfaces, channels, and cavities. *J Mol Graphics*, 7, 243-245

Wang, H. & Levinthal C. (1991) A vectorized algorithm for calculating the accessible surface of macromolecules. *J Comp Chem*, 12, 868-871

6. Appendix A: Overview of the program

6.1. Outline of the program

EXCAVAT was written in C++, compiled with Borland C++ Builder, and run on a 333 MHz Pentium II with 256 MB RAM. The standard template library (STL) was used for implementing lists and vectors.

In principle, the program follows a one-way path (Figure 14). However, there is a choice of performing the calculation with a single structure at a time or as a batch job with structure file names provided in an input file. Then, the (first) PDB file is read, grid points are created and assigned maximal-accessibility integers as well as shortest-distance integers. In the next step, the SASA per atom and amino acid are calculated and cavities are identified and characterized. Finally, output files are written. For the batch version, the procedure is repeated.

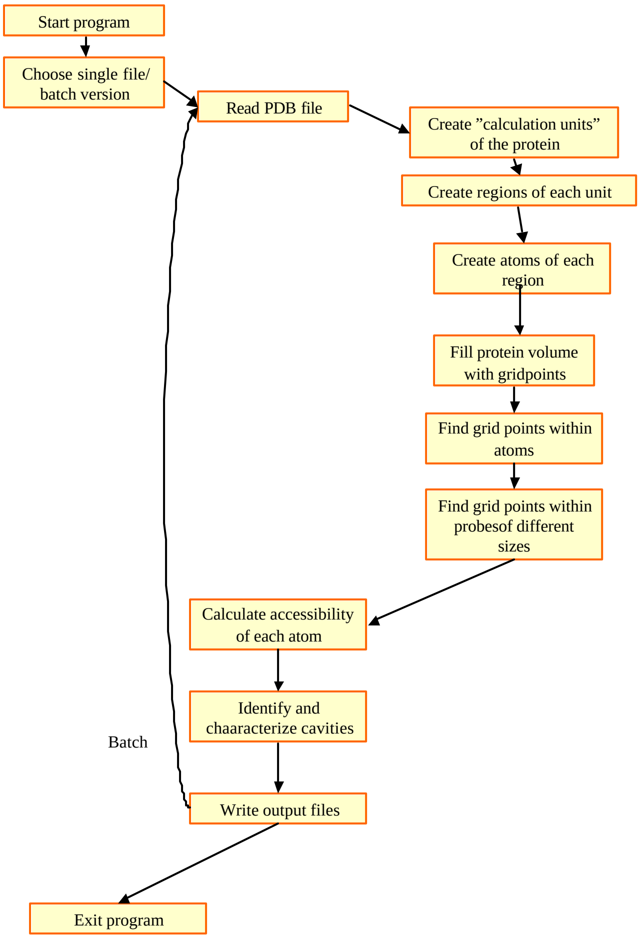

Start program
Choose single file/ batch version
Read PDB file
Create ”calculation units” of the protein
Create regions of each unit
Create atoms of each region
Fill protein volume with gridpoints
Find grid points within atoms
Find grid points within probesof different sizes
Calculate accessibility of each atom
Identify and chaaracterize cavities
Write output files
Batch
Exit program

Figure 14. Flow chart of the program.

6.2. Program input and output

The input of the program consists of a parameter file, one or several PDB files, and in the batch case, a file with PDB file names and a file containing estimated SASA values to be used for missing amino acids. The parameter file contains information for the advanced user to change if desired (default values to be seen in Table 4). During the development, parameters have been altered in order to examine the effects in terms of accuracy and speed. The shell thickness for the calculation of SASA has been varied and optimized to 0.95 Å for a grid of 1.0 Å.

Grid spacing	1.0 Å
SASA shell thickness	0.95 Å
n_{max} (number of probes)	8
R_P (smallest probe radius)	1.4 Å

Table 4. Default values of parameters.

From PDB files, sequence information (SEQRES lines) is recorded, if present. Of the PDB information on atoms (ATOM lines), atom type, amino acid type and number as well as x, y and z coordinates are saved. All heteroatoms (HETATM lines) are ignored by the program.

A number of output files are generated. The SASA values per atom and per amino acid are given in separate files. The descriptors for each identified cavity are given in a file. As an option, all grid points can be written as a pseudo-PDB file, which can be displayed in SYBYL or similar programs. In the same way, the coordinates of all grid points in every cavity of the protein can be written to a file.

6.3. Treatment of atoms

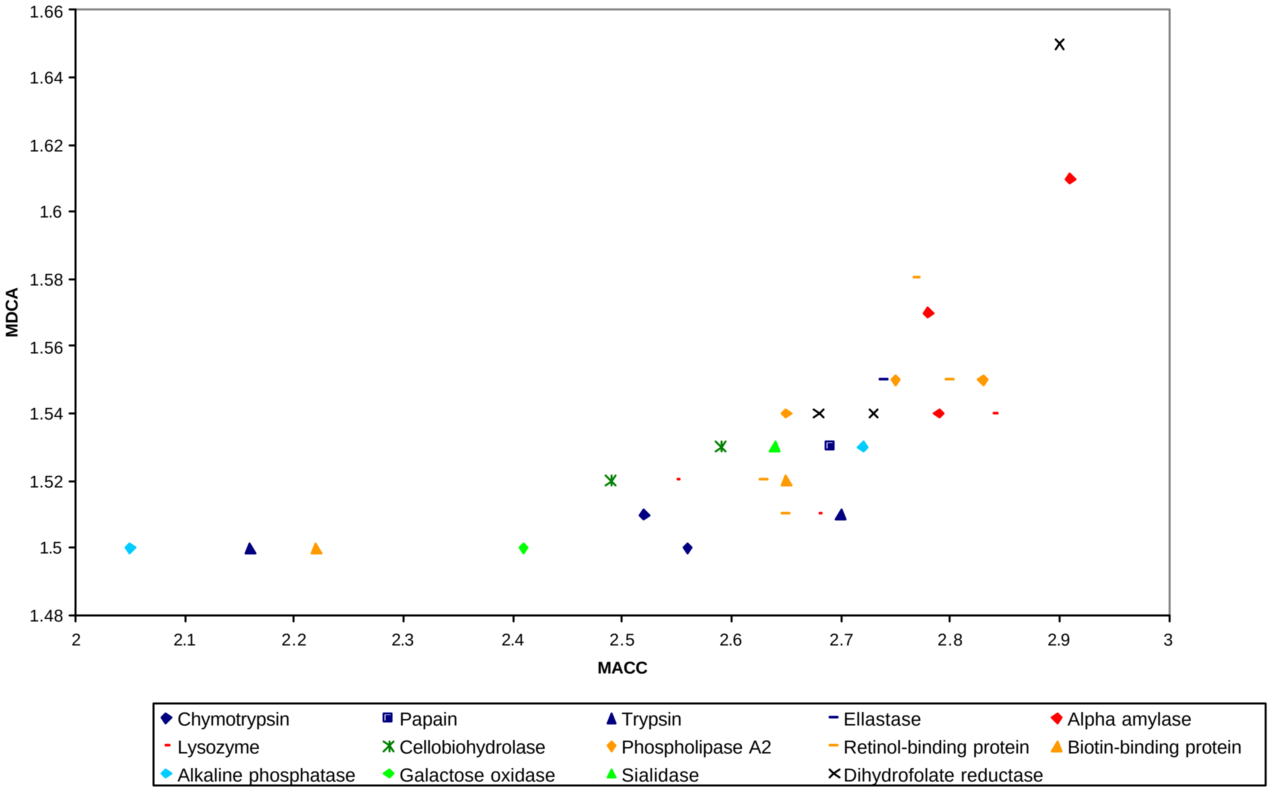
In the case of separate chains or models of the PDB file, the appropriate treatment of the protein is not obvious. If the protein exists as a monomer, SASA and cavities are calculated most properly by regarding the chains individually (separate calculation units). Otherwise, all chains should be treated as one calculation unit. The user may choose how to treat the protein. Each calculation unit is considered separately throughout the calculations. For PDB files containing different models (from NMR data), each model is treated separately throughout the calculations, and then an average is calculated (of SASA).

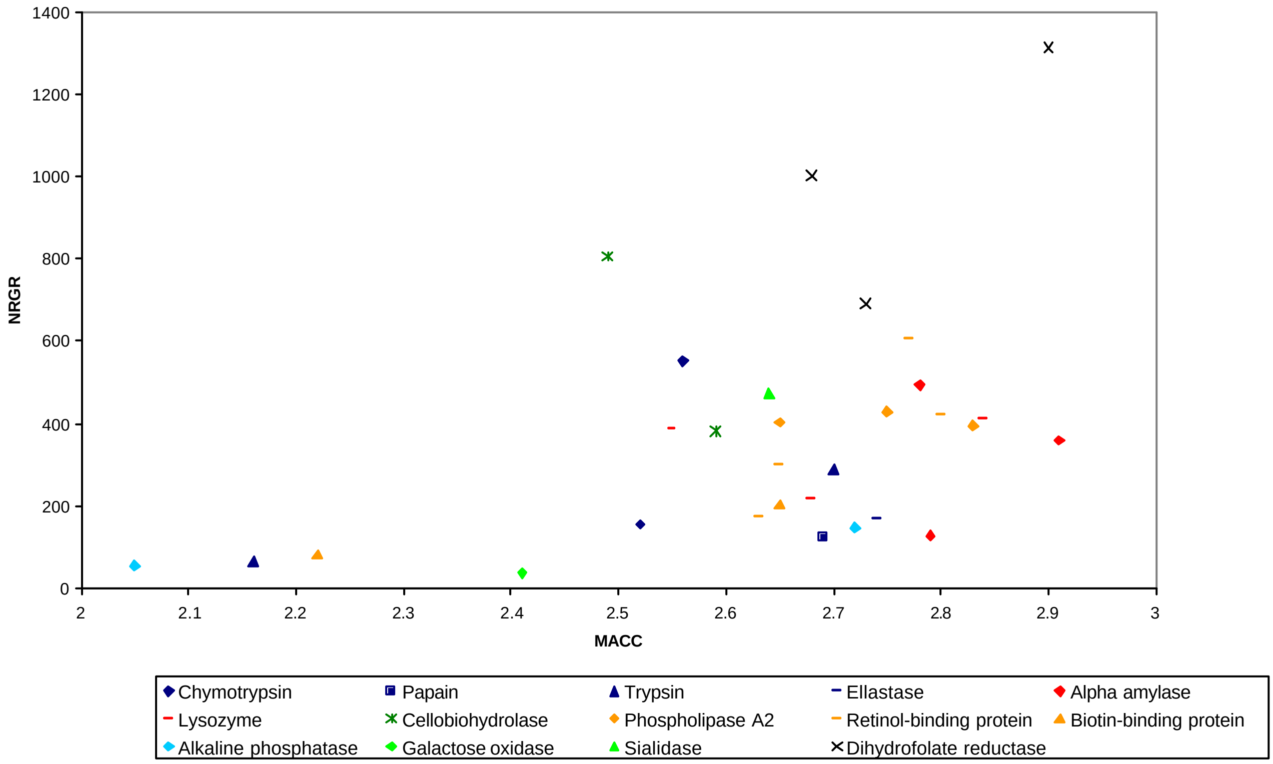
Before storing the atoms, the geometrical borders of each calculation unit are defined and regions are created for each calculation unit, by splitting the volume needed to cover the protein into several parts. The atoms of the PDB file are then stored as objects in the fitting region, with a radius (Table 5) given by the atom type (Shrake and Rupley). As the hydrogen atoms are implicitly included in the other atoms, all hydrogen atoms are given a radius of 0.

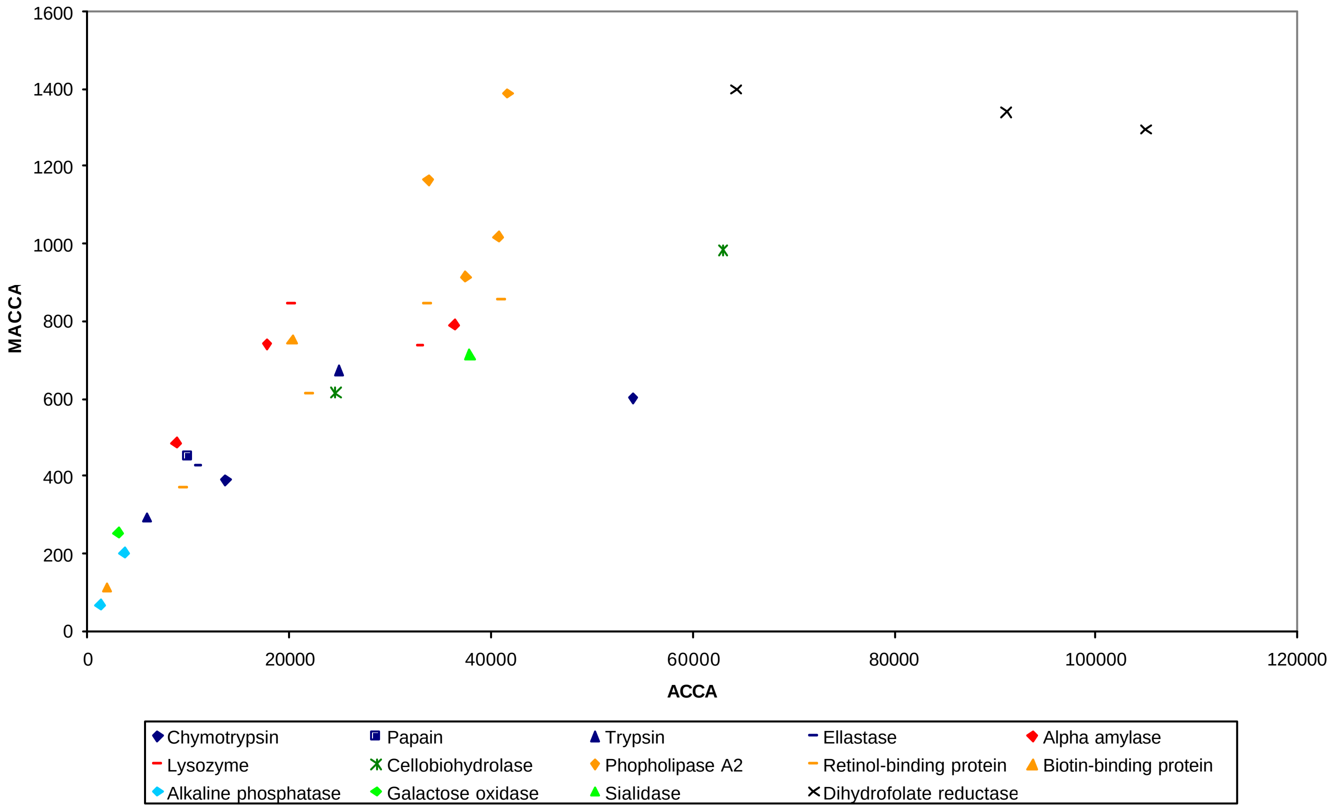
Atom	**Van der Waals radii (Å)**
All nitrogen	1.5
All oxygen	1.4
All sulfur	1.85
Non-aromatic carbon	2.0
Aromatic carbon	1.85
Carbonyl and all other carbon	1.5

Table 5. Van der Waals radii of different atoms.

7. Appendix B: Results with cavity descriptors




1400
1200
1000
800
600
400
200
0
NRGR
2
2.1
2.2
2.3
2.4
2.5
2.6
2.7
2.8
2.9
3
MACC
Chymotrypsin
Papain
Trypsin
Ellastase
Alpha amylase
Lysozyme
Cellobiohydrolase
Phopholipase A2
Retinol-binding protein
Biotin-binding protein
Alkaline phosphatase
Galactose oxidase
Sialidase
Dihydrofolate reductase


1600
1400
1200
1000
800
600
400
200
0
MACCA
0
20000
40000
60000
80000
100000
120000
ACCA
Chymotrypsin
Papain
Trypsin
Ellastase
Alpha amylase
Lysozyme
Cellobiohydrolase
Phopholipase A2
Retinol-binding protein
Biotin-binding protein
Alkaline phosphatase
Galactose oxidase
Sialidase
Dihydrofolate reductase

8. Appendix C: CRR and average values for cavity descriptors

MDCA	**CRR**	**Average**
Chymotrypsin	0.13	1.51
Trypsin	0.13	1.51
Serine protease	0.33	1.52
Alpha amylase	0.70	1.57
Lysozyme	0.30	1.52
Cellobiohydrolase	0.13	1.53
Phospholipase A2	0.09	1.55
Retinol-binding protein	0.62	1.54
Biotin-binding protein	0.27	1.51
Alkaline phosphatase	0.40	1.52
Galactose-binding protein	0.40	1.52
Dihydrofolate reductase	1.10	1.58
MACC		
Chymotrypsin	0.09	2.54
Trypsin	1.26	2.43
Serine protease	0.84	2.56
Alpha amylase	0.23	2.83
Lysozyme	0.51	2.69
Cellobiohydrolase	0.23	2.54
Phospholipase A2	0.28	2.75
Retinol-binding protein	0.26	2.71
Biotin-binding protein	1.00	2.44
Alkaline phosphatase	1.56	2.39
Galactose-binding protein	0.53	2.53
Dihydrofolate reductase	0.38	2.77
NRGR		
Chymotrypsin	0.62	354
Trypsin	0.35	176
Serine protease	0.48	225
Alpha amylase	0.43	327
Lysozyme	0.23	337
Cellobiohydrolase	0.67	594
Phospholipase A2	0.04	413
Retinol-binding protein	0.45	374
Biotin-binding protein	0.19	143
Alkaline phosphatase	0.15	102
Galactose-binding protein	0.68	255
Dihydrofolate reductase	0.73	1001

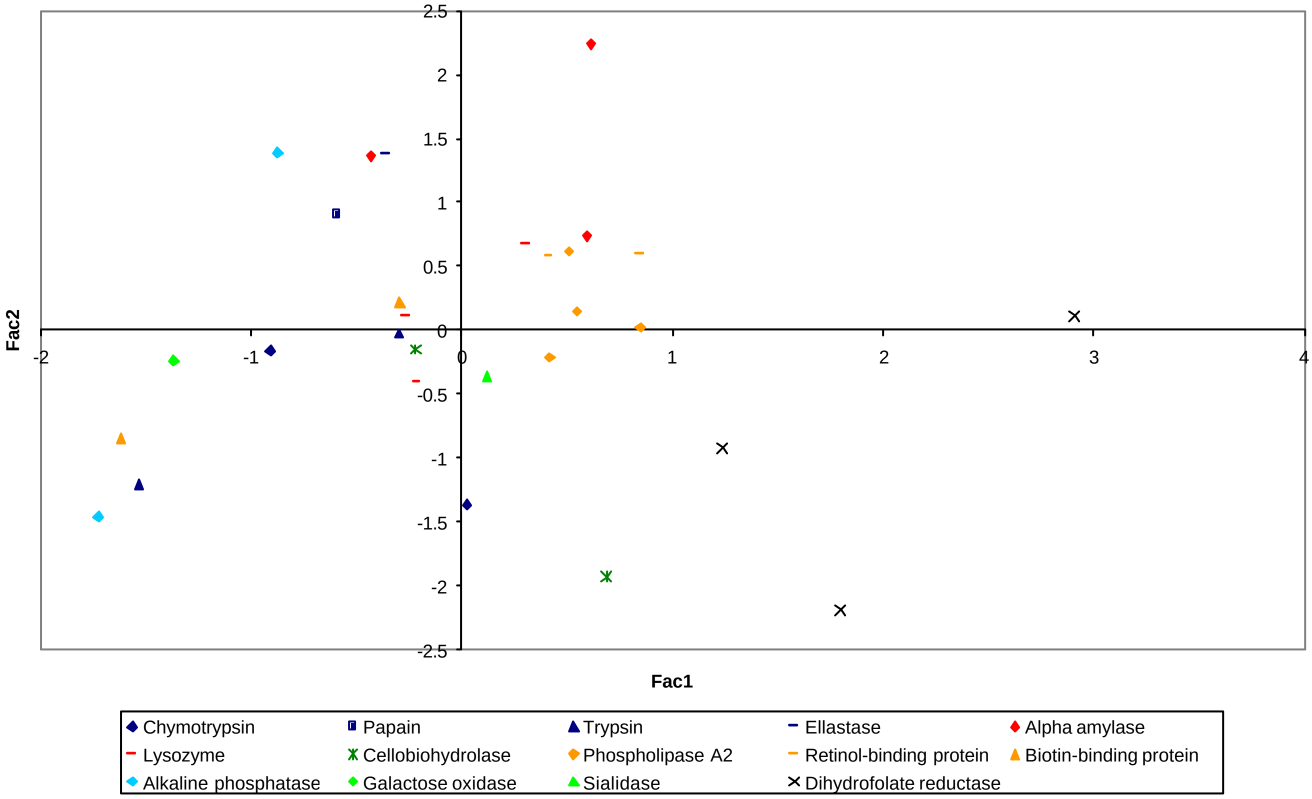
ACCA	**CRR**	**Average**
Chymotrypsin	0.78	3.39E+04
Trypsin	0.37	1.54E+04
Serine protease	0.58	1.99E+04
Alpha amylase	0.40	2.09E+04
Lysozyme	0.18	2.55E+04
Cellobiohydrolase	0.74	4.38E+04
Phospholipase A2	0.10	3.84E+04
Retinol-binding protein	0.24	2.66E+04
Biotin-binding protein	0.36	1.11E+04
Alkaline phosphatase	0.05	2.42E+03
Galactose-binding protein	0.67	2.05E+04
Dihydrofolate reductase	0.59	8.68E+04
MACCA		
Chymotrypsin	0.32	496
Trypsin	0.57	483
Serine protease	0.36	472
Alpha amylase	0.34	672
Lysozyme	0.11	775
Cellobiohydrolase	0.56	800
Phospholipase A2	0.47	1121
Retinol-binding protein	0.49	669
Biotin-binding protein	0.96	432
Alkaline phosphatase	0.20	134
Galactose-binding protein	0.69	484
Dihydrofolate reductase	0.12	1344

Protein groups in yellow marked rows have a CRR value of 0.20 or less.

9. Appendix D: Principal component analysis for cavity descriptors

The principal components corresponding to the two largest eigenvalues were extracted, and follow the equations:

```
Fac1 = 0.790*MDCA+0.700*MACC+0.924*NRGR+0.891*ACCA+0.892*MACCA
Fac2 = 0.456*MDCA+0.631*MACC-0.295*NRGR-0.422*ACCA-0.171*MACCA
```


Fac2
Fac1
Chymotrypsin
Papain
Trypsin
Ellastase
Alpha amylase
Lysozyme
Cellobiohydrolase
Phospholipase A2
Retinol-binding protein
Biotin-binding protein
Alkaline phosphatase
Galactose oxidase
Sialidase
Dihydrofolate reductase