

# Methods for the Prediction of Protein-Ligand Binding Sites for Structure-Based Drug Design and Virtual Ligand Screening

Alasdair T.R. Laurie and Richard M. Jackson\*

*Institute of Molecular and Cellular Biology, Faculty of Biological Sciences, University of Leeds, Leeds, LS2 9JT, UK*

**Abstract:** Structure Based Drug Design (SBDD) is a computational approach to lead discovery that uses the three-dimensional structure of a protein to fit drug-like molecules into a ligand binding site to modulate function. Identifying the location of the binding site is therefore a vital first step in this process, restricting the search space for SBDD or virtual screening studies. The detection and characterisation of functional sites on proteins has increasingly become an area of interest. Structural genomics projects are increasingly yielding protein structures with unknown functions and binding sites. Binding site prediction was pioneered by pocket detection, since the binding site is often found in the largest pocket. More recent methods involve phylogenetic analysis, identifying structural similarity with proteins of known function and identifying regions on the protein surface with a potential for high binding affinity. Binding site prediction has been used in several SBDD projects and has been incorporated into several docking tools. We discuss different methods of ligand binding site prediction, their strengths and weaknesses, and how they have been used in SBDD.

## INTRODUCTION

Drug design is a time consuming and expensive process. The first stages of this process are lead discovery and lead optimisation. Traditionally, lead compounds have been discovered serendipitously, by chemically modifying and improving existing drugs (the so-called “me-too” approach) or by isolating the active ingredients in herbal remedies. More recently, pharmaceutical companies have focussed on high-throughput screening (HTS). This involves screening a large chemical library against a protein target. There are around 3 million chemicals publicly available for purchase to be used in HTS, and many more in proprietary commercial databases. However, large scale HTS is expensive and it is beneficial to restrict the size of a chemical library to compounds that are most likely to be successful. Screening of a virtual library is one way in which potentially successful compounds can be identified. HTS and virtual screening are limited by the size of the library they use. *De novo* drug design attempts to overcome this limitation by increasing the exploration of chemical search space. Both virtual screening and *de novo* drug design require a three-dimensional representation of the protein target and are therefore referred to as “structure-based” drug design (SBDD) methods. Structure-based drug design has already yielded several drugs currently on the market. These include the HIV protease inhibitors Viracept [1] and Agenerase [2].

There are two prerequisites for SBDD. Firstly, a three-dimensional representation of the protein target must be available. Preferably, this structure should be derived from X-ray crystallography or NMR. Alternatively, a comparative (homology) model of the protein structure may be created if there is sufficient sequence similarity between the target and

a protein of known structure. The second prerequisite is knowledge of the location of the ligand binding site. Computational methods for the detection and characterisation of functional sites on proteins have increasingly become an area of interest [3]. Binding sites can be identified by co-crystallisation of a protein with a ligand, by identifying structural or sequence similarity with a known binding site or by using a binding site prediction tool.

Several types of algorithms have been developed to predict ligand binding sites. Some analyse the protein surface for pockets. Many studies have suggested that the binding site is usually in the largest pocket [4-7]. Another type of algorithm analyses the binding energies of probes placed on a grid around the protein. Probe clustering [8, 9] and energy contour analysis [10] can be used to predict ligand binding sites. Alternatively, more complex simulation methods can also be used to predict binding sites *e.g.* Bhinge *et al.* [11] used molecular dynamics simulations to identify ligand binding sites. Elcock [12] used the assertion that functionally important residues are often in electrostatically unfavourable positions.

A series of functional site comparison tools also exist to identify binding sites, and have recently been reviewed by Jones and Thornton [13]. These tools can be used to assign function to newly resolved protein structures with unknown function. Such tools include 3D templates [14, 15], graph theory [16, 17], ‘fuzzy pattern matching’ [18] and evolutionary trace methods [19]. Such tools are not normally used for binding site prediction in SBDD studies. They are more often used to allocate function to newly resolved protein structures from structural genomics projects. Other approaches include the expectation that amino acids within a binding site mutate simultaneously during evolution (correlated mutations), which has been applied to protein-protein binding site prediction [20]. It has also been observed that brackets of proline residues are often found in protein-protein binding sites [21], although the same has not been

\*Address correspondence to this author at the Institute of Molecular and Cellular Biology, Faculty of Biological Sciences, University of Leeds, Leeds, LS2 9JT, UK; Tel: +44 (0)113 343 2592; Fax: +44 (0)113 343 3167; E-mail: r.m.jackson@leeds.ac.uk

noted for protein-ligand binding sites. It should be noted that the prediction of protein-protein binding sites (see Szilagyi *et al.* [22] for a recent review) usually requires a different computational approach to protein-ligand binding site prediction and is beyond the scope of this review.

There are many problems involved in predicting ligand binding sites. One major problem is induced fit. The binding site can change shape significantly upon binding a ligand. Another problem is where a ligand binding site occurs at an inter-subunit interface. Some algorithms have only been tested on single subunits, and some have been shown to perform less well on complexes. A third problem is the sheer variety of ligands that exist, and the corresponding variety of binding sites. It is difficult to design an algorithm that accounts for all conformationally and physicochemically different ligand binding sites. There remains the problem of how to validate a binding site prediction tool. Often a successful prediction is defined as covering a certain number of ligand atoms. However, if the predicted sites are very big (for example, covering the whole protein) then the prediction can be still counted as a success even though it is not very precise. In general, SBDD requires a precise definition of the ligand binding site to restrict the search space to relevant areas of the protein, and reduce false-positive results. In this review, we explore some of the different methods used to predict ligand binding sites on proteins as a first step in locating sites for the SBDD process. Pocket-detection, both geometry and energy-based methods are of principal importance in defining the binding site in SBDD, we therefore concentrate on describing these methods. However, increasingly functional site prediction and “blind docking” methods will play a role in SBDD, therefore, we have covered albeit more briefly recent advances and applications in these areas.

### POCKET DETECTION: GEOMETRY-BASED METHODS

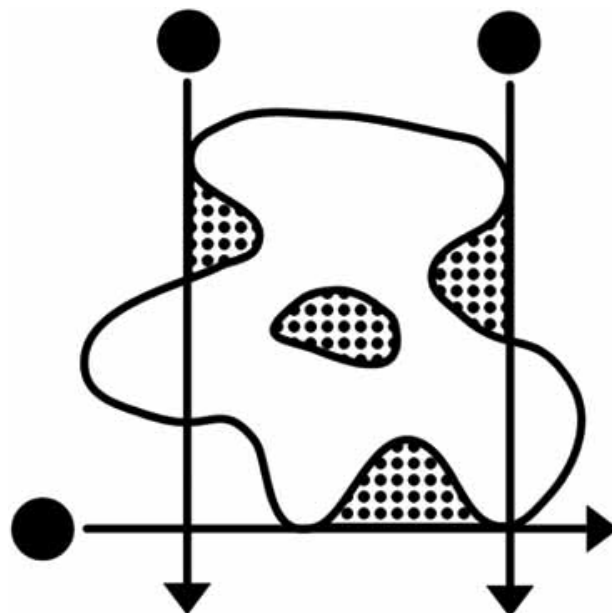
Protein pocket detection is a widely used technique to identify potential ligand binding sites. It uses geometric considerations to define pockets and studies have shown that the binding site is commonly found in the largest pocket. For example, SurfNet [23] was used to analyse 67 protein structures, and the ligand binding site was found to be in the largest pocket in 83% of cases [7]. Another method, APROPOS [24] looks for characteristic patterns of small “caves” into which molecular groups can fit into, and has a high reported success rate. Other pocket detection algorithms include Cavity Search [25], POCKET [26], VOIDOO [27], LIGSITE [5], CAST [28, 29], PASS [30], LigandFit [31] and algorithms developed by Delaney [32], Del Carpio *et al.* [33] and Masuya & Doi [34].

Pocket detection algorithms frequently employ a three-dimensional grid surrounding the protein or a definition of the molecular surface. The molecular surface can be defined purely using a grid, by finding the interface at which grid points no longer coincide with protein atoms. This technique is employed by LIGSITE [5], POCKET [26] and the method of Delaney [32]. Molecular surface algorithms can also be used. These have the advantage of not being dependent on grid resolution. Molecular surface algorithms are generally dependent on the radius of a “solvent” probe that rolls across

the surface (this is generally taken to be water, with a radius of 1.4Å). The Solvent Accessible Surface of Lee & Richards [35] is the surface defined by the centre of the probe, whilst the Molecular Surface or Connolly surface [36] is defined by the protein-solvent interface i.e. the surface completely excluded from solvent volume, and therefore defines the point of contact between the solvent probe and van der Waals surface of the protein atoms. Several pocket detection algorithms are discussed in more detail below.

### POCKET [26]

A probe sphere of radius 3Å is passed across the protein along each line of a Cartesian three-dimensional grid in the x, y and z directions. An interaction between the protein and probe sphere occurs if the centre of a protein atom is found to be within the probe sphere. A pocket is identified if an interaction occurs followed by a period of no interaction, followed by another interaction. Pockets are shown by dotted areas in Fig. (1). The main disadvantage of this algorithm is that the precise nature of the pockets is dependent on the angle of rotation of the protein relative to the coordinate reference frame.

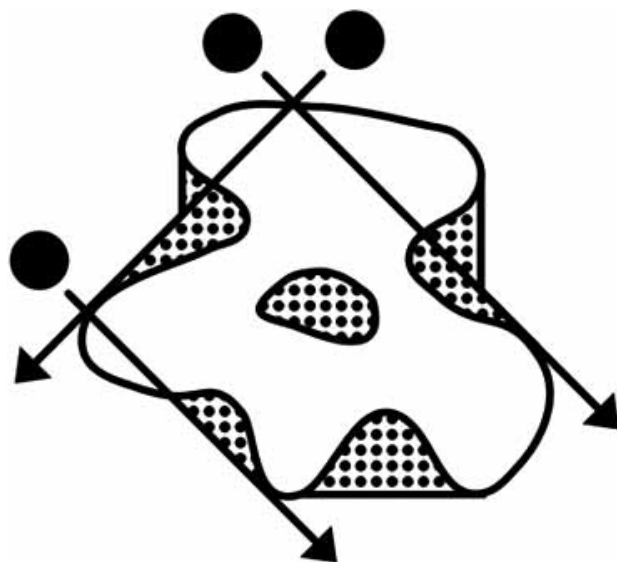


**Fig. (1).** The POCKET algorithm. Probe spheres (black circles) scan a protein. Dotted areas indicate pockets and cavities identified by the algorithm.

### LIGSITE [5] and Pocket-Finder [8]

LIGSITE is very similar to POCKET (described above). However, LIGSITE also scans with probes along the Cartesian cubic diagonals as well as the x, y and z axes, i.e. seven scan directions as opposed to three. This makes identification of protein pockets much less dependent on the orientation of the protein in the three-dimensional grid (compare Fig. 1 and Fig. 2). LIGSITE has a variable known as the MINPSP (minimum protein-site-protein) threshold. A single grid point has seven probe lines passing through it (x, y, z and the four cubic diagonals). The grid point can be defined to be a pocket (PSP event) up to seven times. The MINPSP thresh-

old defines how many PSP events must occur for a grid point to be defined as being part of a pocket. By setting the threshold higher, shallow pockets are excluded. LIGSITE was verified on ten protein structures, and was shown to give good results, with seven of the proteins having the binding site in the largest pocket. The accuracy, speed and simplicity of this type of algorithm has made it ideal for use in several subsequent studies, including CavBase [37] and SuperStar [38]. Recently, we have implemented our own version of the LIGSITE method, called Pocket-Finder [8]. This was done in order to make direct comparison with the energy-based method, Q-SiteFinder (see below).



**Fig. (2).** LIGSITE scans cubic diagonals in addition to the x, y and z axes.

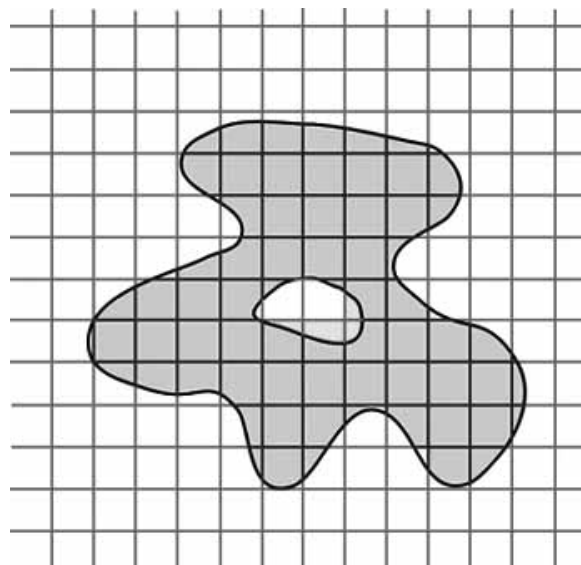
#### Delaney [32]

The protein is placed within a three-dimensional grid. Where grid points intersect the protein, they are set to 'true', otherwise they are set to 'false' (Fig. 3A). The protein surface (and cavity boundaries) are defined to be grid points set to 'true' that are adjacent to grid points set to 'false'. A monolayer of particles is then added to the protein surface (a surface expansion) and the true/false representation is recalculated to redefine the surface (Fig. 3B). A surface contraction then takes place, where a monolayer of particles is removed (Fig. 3C). Some of the particles added to pockets survive the surface contraction. This is because expansion into pockets can add particles that are not subsequently defined to be part of the protein surface. After repeated expansions and contractions (usually five to ten), the protein cavities are filled with particles (Fig. 3D).

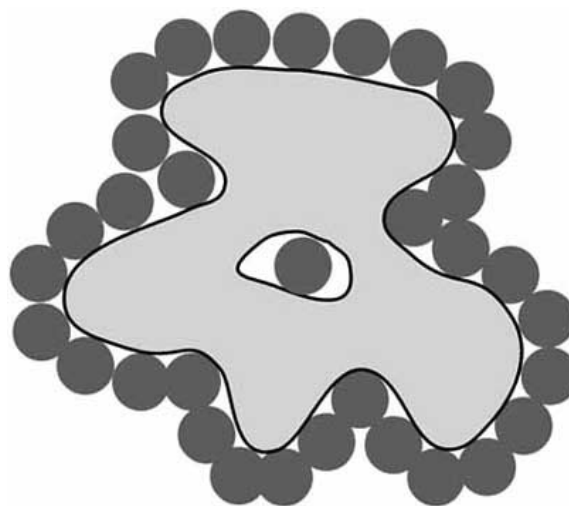
#### PASS [30]

PASS (Putative Active Site with Spheres) uses a similar concept to Delaney [32], described above. However, a different type of analysis is used to achieve a similar effect. The algorithm looks at all possible combinations of three protein atoms. If the three atoms are close enough together, the algorithm calculates the two possible positions for a probe sphere

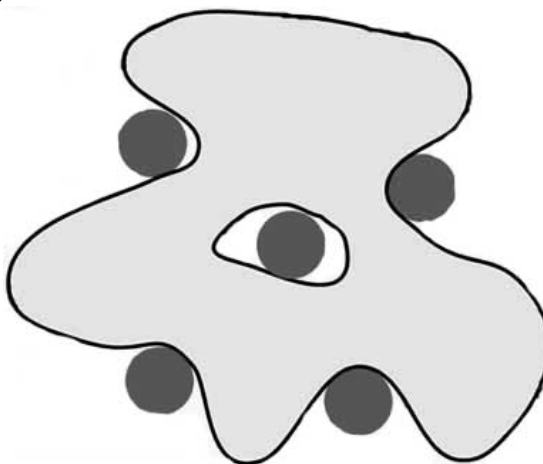
A)



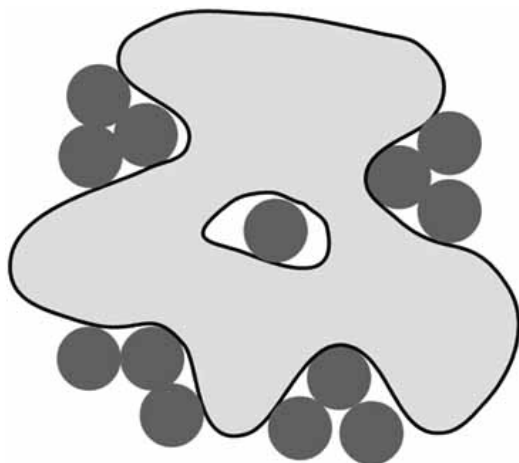
B)



C)

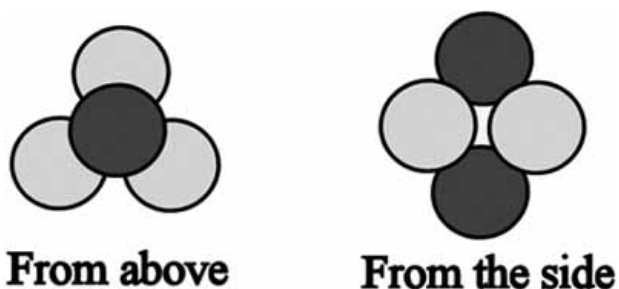


(Fig. 3) contd....

**D)**

**Fig. (3).** The algorithm of Delaney [32]. **A:** The protein is placed in a three dimensional grid. **B:** A surface expansion takes place. **C:** A surface contraction takes place. **D:** After repeated surface expansions and contractions, particles accumulate in the pockets and cavities.

to just touch the surface of all three protein atoms (Fig. 4). The probes are rejected if they clash with protein atoms. This results in a covering of the protein surface with probes, similar to the surface expansion of Delaney [32], shown in Fig. (3B).



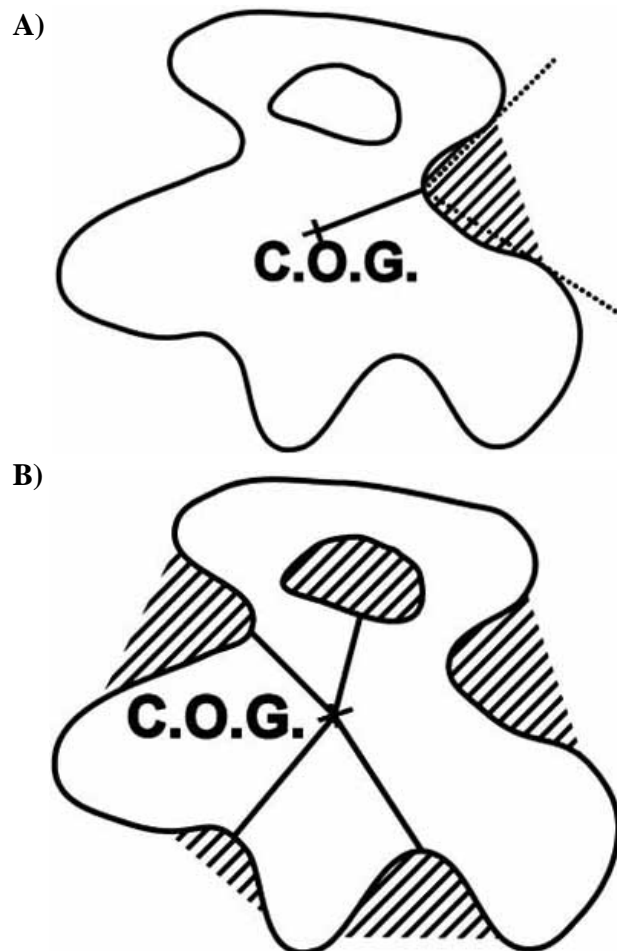
**Fig. (4).** The PASS algorithm. The positions of probes (dark grey) are calculated from the positions of three protein atoms (light grey). There are two possible probe positions, and each is tangential to all three protein atoms.

A filtration step then occurs, which has a similar effect to the surface contraction of Delaney [32]. The burial count of each probe is measured by calculating the number of protein atoms found within an 8Å radius of the probe. Probes in pockets have a higher burial count than those outside pockets. A burial count threshold is applied to remove probes outside pockets. This filtration step leaves probes in pockets and cavities as shown in Fig. (3C). Repeated cycles of addition of probe spheres followed by filtration to remove spheres not found in protein pockets causes accumulation of spheres in a similar fashion to that shown in Fig. (3D). When these cycles of addition and filtration no longer lead to a change in the number of probes bound to the protein, the end

point is reached. PASS then assigns probe weights using a method related to free surface volume analysis of Stouten *et al.* [39]. These weights are used to determine a single point to represent the binding site, referred to as “Active Site Points”. The algorithm was tested on 30 complexes and most of the ligand binding sites were identified by one or more of the three largest pockets [30].

#### Del Carpio *et al.* [33]

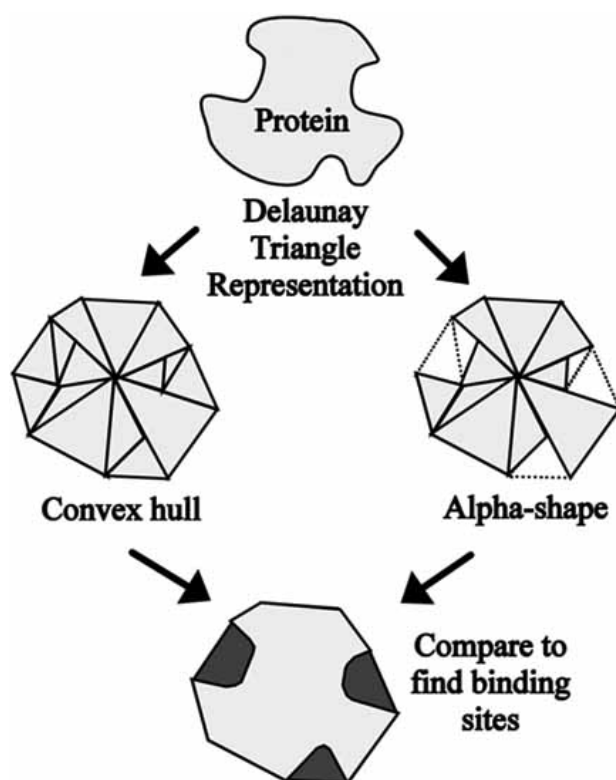
This algorithm uses a surface “growing” process to identify cavities and pockets. The molecular surface is first identified using the method of Lee & Richards [35]. The centre of gravity of the protein is identified along with the closest surface atom (Fig. 5A). The surrounding surface atoms are then flagged such that a concave pocket is defined by atoms within line of sight of the first atom. The algorithm then searches for the next closest unflagged atom to the centre of gravity and repeats the process. The algorithm continues until no more concave regions on the surface can be identified (Fig. 5B).



**Fig. (5).** The algorithm of Del Carpio *et al.* **A:** The nearest surface point to the centre of gravity (C.O.G.) is taken as the first starting point, and the resulting binding site is indicated by the shaded area. **B:** Other starting points are identified in order of their proximity to the centre of gravity.

## APROPOS [24]

The Automatic PROtein POcket Search (APROPOS) algorithm is based upon creating an  $\alpha$ -shape representation of the protein. The algorithm used to generate the  $\alpha$ -shape creates a Delaunay [40] representation of the protein. The nature of the  $\alpha$ -shape is dependent on a parameter " $\alpha$ ". This can be thought to be the radius of a probe that is rolled over the surface of the protein. The probe can erase the sides and edges of the triangles, but not the vertices (atomic centres). When  $\alpha$  approaches infinity, the convex hull is formed (Fig. 6). In practice, an experimental value of around 20 Å was used, otherwise false positive pockets were identified. The alpha-shape (Fig. 6) is created by using values of  $\alpha$  between 2.8 Å (oxygen atom radius) and 4.5 Å (methyl group radius) to find pockets that could bind to ligand groups. The pockets are identified by comparing the structures of the alpha-shape and convex hull. Protein pockets are revealed where the structures of the two representations differ significantly.

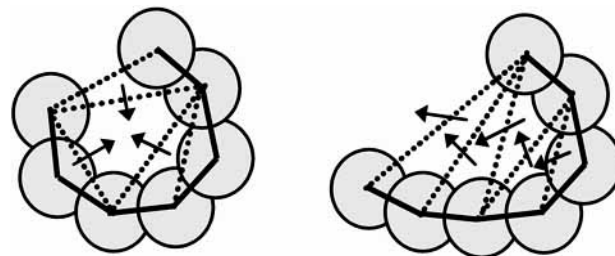


**Fig. (6).** The APROPOS algorithm. Pockets are identified by comparing the convex hull with the  $\alpha$ -shape (see text).

APROPOS also includes a method to predict which pocket(s) are ligand binding sites. It has been noted that ligand groups tend to fit into small "caves" in protein molecules. APROPOS searches for these characteristic "caves". The algorithm was shown to have a 95% success rate on a dataset of proteins consisting of one subunit, although the accuracy was much lower when protein complexes were tested.

## CAST [28, 29]

CAST uses a similar method to APROPOS (described above) to detect protein pockets. Delaunay representations of the proteins are created and discrete flow theory is used to determine which pockets to consider (Fig. 7). The algorithm was tested on the data set of 67 protein structures used by Laskowski *et al.* [7]. When using CAST, 74% of ligand binding sites were identified in the largest pocket as opposed to 83% found using SurfNet. However, the authors concluded that differences between the size and nature of the pockets produced by the two methods make direct comparison difficult. CAST has been made available online as CASTp (Table 1).



**Fig. (7).** The CAST algorithm. A diagram demonstrating discrete flow theory (adapted from [29]). **A:** One Delaunay triangle acts as a sink for the flow. CAST considers this a true pocket. **B:** The triangles flow to infinity. CAST does not consider these types of pockets.

## SurfNet [23]

SurfNet takes pairs of relevant atoms within a protein and forms a test sphere between them. If this sphere overlaps any protein atoms, the radius of the test sphere is reduced until there is no further overlap (Fig. 8A). The test spheres therefore accumulate in pockets and cavities as shown in (Fig. 8B). Test spheres are only retained if they have a radius between 1 Å and 4 Å. SurfNet has been made available for download (Table 2).

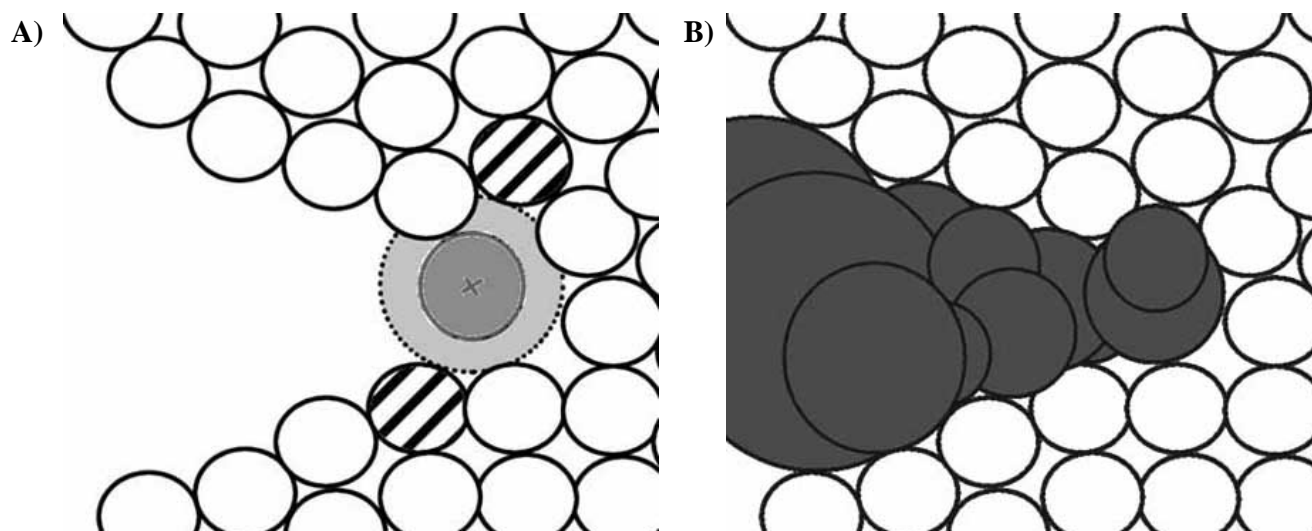
In summary, the main issue when designing a pocket detection algorithm is defining where the boundary between protein and pocket occur. Many algorithms include a method to estimate pocket volume, often by counting the number of grid points contained within the predicted binding site. We have shown that pocket volume calculated with Pocket-Finder [8] (our implementation of LIGSITE) increases linearly with protein size, but the average volumes of bound ligands are independent of protein size and show little change. Sometimes, pockets defined by geometric means can therefore be very large relative to the size of the volume occupied by the ligand.

## Pocket Modelling by Docking Tools

Docking tools require a three dimensional representation of the binding pocket. However, it is not always necessary to use a pocket-detection approach. Sometimes the scoring function and 'bumps checking' (for protein-ligand overlap) are sufficient to representing the binding pocket. For example, Q-fit creates a series of three-dimensional energy grids

**Table 1. Website Addresses for Online Servers that can Identify Ligand Binding Sites**

Type	Method	Address
Pocket Detection	CASTp	<a href="http://cast.engr.uic.edu/cast">http://cast.engr.uic.edu/cast</a>
	Pocket-Finder	<a href="http://www.bioinformatics.leeds.ac.uk/pocketfinder">http://www.bioinformatics.leeds.ac.uk/pocketfinder</a>
Energy-based site detection	Q-SiteFinder	<a href="http://www.bioinformatics.leeds.ac.uk/qsitefinder">http://www.bioinformatics.leeds.ac.uk/qsitefinder</a>
Phylogenetic Analysis	Consurf	<a href="http://consurf.tau.ac.il">http://consurf.tau.ac.il</a>
Binding Site databases and functional site comparison	SitesBase	<a href="http://www.bioinformatics.leeds.ac.uk/sb">http://www.bioinformatics.leeds.ac.uk/sb</a>
	ProFunc	<a href="http://www.ebi.ac.uk/thornton-srv/databases/ProFunc">http://www.ebi.ac.uk/thornton-srv/databases/ProFunc</a>
	eF-site	<a href="http://ef-site.hgc.jp/eF-site">http://ef-site.hgc.jp/eF-site</a>
	SiteEngine	<a href="http://bioinfo3d.cs.tau.ac.il/SiteEngine">http://bioinfo3d.cs.tau.ac.il/SiteEngine</a>
	PINTS	<a href="http://www.russell.embl.de/pints">http://www.russell.embl.de/pints</a>



**Fig. (8).** The SurfNet algorithm. **A:** A protein pocket is shown. Protein atoms are represented by white circles. For each pair of atoms (indicated by stripes) a test sphere (light grey circle with dotted outline) is created between them. If the test sphere overlaps with protein atoms, the radius is reduced until they no longer overlap (dark grey circle). If the radius falls below an arbitrary value (for example,  $1.0\text{\AA}$ ), no test sphere is placed at this location. **B:** The process continues, testing all relevant pairs of atoms, until the pockets are filled with spheres.

**Table 2. Web Addresses for Downloadable Tools that can Identify Ligand Binding Sites**

Type	Method	Address
Pocket Detection	SurfNet	<a href="http://www.biochem.ucl.ac.uk/~roman/surfnet/surfnet.html">http://www.biochem.ucl.ac.uk/~roman/surfnet/surfnet.html</a>
	VOIDOO	<a href="http://xray.bmc.uu.se/usf/voidoo.html">http://xray.bmc.uu.se/usf/voidoo.html</a>
	PASS	<a href="http://www.ccl.net/cca/software/UNIX/pass/overview.shtml">http://www.ccl.net/cca/software/UNIX/pass/overview.shtml</a>
Phylogenetic Analysis	Rate4Site	<a href="http://www.tau.ac.il/~itaymay/cp/rate4site.html">http://www.tau.ac.il/~itaymay/cp/rate4site.html</a>

around the binding pocket for a variety of probe types [41]. The potential energy maps are used to define the binding pocket. Algorithms such as DOCK define the binding pocket more explicitly. The “sphgen” program forms part of the DOCK suite and generates a series of overlapping spheres to describe the three dimensional shape of a binding pocket. The Connolly algorithm [36] is used to generate a molecular surface. Spheres are placed on the molecular surface such that each sphere just touches the surface at two points and the radius of each sphere passes through the surface normal of one of the points. The spheres are then filtered so that only the largest sphere touching each protein atom is retained.

## POCKET DETECTION: ENERGY-BASED METHODS

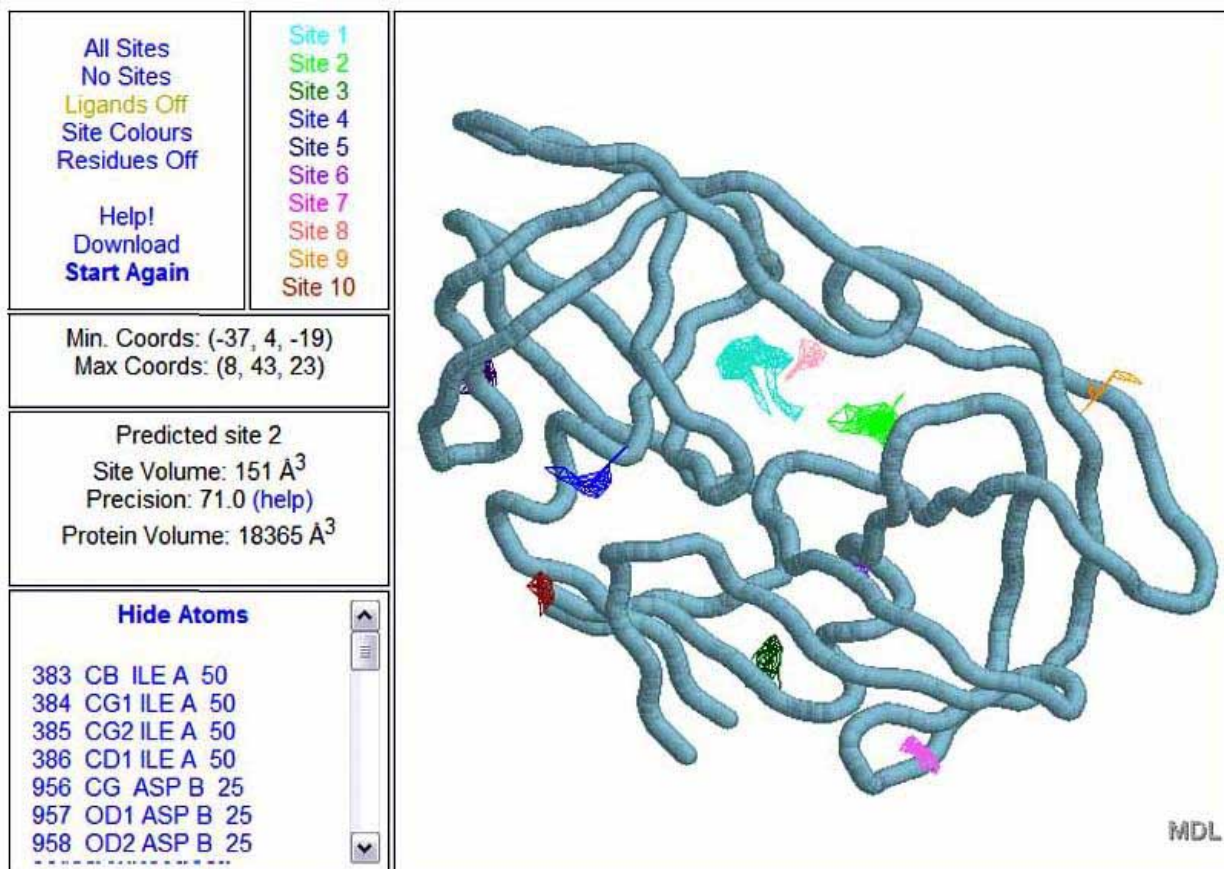
### Goodford [10]

Several techniques have been developed for estimating the interaction energy between a probe molecule (*e.g.* a methyl, hydroxyl, or amine groups) at a given point and a protein. Probably the first time this was introduced and certainly one of the most widely established methods is GRID developed by Goodford [10]. It identifies sites of favourable interaction with specific probe types. This is particularly useful for viewing energy contours to find favourable sites on the protein surface and has been widely used in structure based drug design, since it identifies which parts of the protein are likely to interact favourably with functional groups

on a drug-like molecule. For example, studies have been carried out to identify the hydrogen bonding potential of drug-like molecules using GRID [42, 43]. The Multiple copy simultaneous search (MCSS) method of Miranker [44] has also been used to detect favourable binding sites for different functional groups. However, neither method has been used to locate the ligand binding sites on a protein directly.

### Ruppert *et al.* [9]

The method of Ruppert *et al.* has been developed for estimating the interaction energies between a probe at a given point and a protein. They use the scoring function developed by Jain [45] to optimise interaction energies of three different probe types (hydrophobic, hydrogen atom; hydrogen bond donor, NH; hydrogen bond acceptor, C=O). They retain probes with the most favourable interaction energies. They then identify “sticky spots”, which are regions that have the highest density of probe interaction energy. Next a pocket is grown, by defining protein-free spheres in the protein void around the sticky spot. Lastly, a process of accretion takes place, which enlarges the sticky spots into larger pockets, by adding nearby accessible probes defined by the pocket. Thus, both energetic and geometric criteria are used to define a ligand binding site. Their algorithm was shown to give good results on nine ligand-bound and two proteins in the unbound state.



**Fig. (9).** Q-SiteFinder web page showing the ligand binding site prediction for HIV protease (1aaq). The protein is shown in grey tubing with ten colour-coded predicted binding sites, each represented by a cluster of methyl probes. The location of the ligand binding site is identified by the 1<sup>st</sup> (turquoise), 2<sup>nd</sup> (light green) and 8<sup>th</sup> (pink) predicted binding sites.



### Q-SiteFinder [8]

Q-SiteFinder locates ligand binding sites by clustering favourable regions for van der Waals (CH<sub>3</sub>) probes on the protein surface (Fig. 9). It uses the GRID forcefield parameters [41] to estimate the interaction energies of probes placed at all points on a three dimensional grid that encompasses the entire protein. Probes with favourable interaction energies are retained and are clustered according to their spatial proximity. The clusters are ranked according to their total interaction energy.

The algorithm was shown to have a 90% success rate in the top three predicted sites when tested on 134 protein-ligand complexes corresponding to the GOLD docking test set described by Nissink *et al.* [46]. The success rate showed a small decrease (to 86%) when tested on proteins in the unbound state, possibly because of the effect of induced fit.

### Comparison of Q-SiteFinder and Pocket-Finder [8]

Q-SiteFinder was validated using a precision-based threshold for success. Precision is defined as the percentage of probes in a single cluster that are within 1.6Å of ligand atoms. A precision threshold of 25% was used to define a successful prediction. Q-SiteFinder obtained an average precision of 68% in the first predicted site. Q-SiteFinder was compared with a pocket detection algorithm, Pocket-Finder, (our implementation of LIGSITE [5]), which was optimised and tested on the same data set as Q-SiteFinder. Pocket-Finder obtained a similar success rate to Q-SiteFinder but only when the precision threshold was dropped to zero. Pocket-Finder had an average precision of 29% for the largest pocket. We think that the high precision and success rate of Q-SiteFinder will be of benefit in SBDD studies and functional site analysis. Q-SiteFinder and Pocket-Finder have been made available online (Table 1).

### Pocketome [47]

The pocketome algorithm is similar to that of Q-SiteFinder. It creates a three-dimensional grid around the protein and calculates van-der Waals potentials at each point. The potential map is then smoothed, and envelopes of favourable binding energy were identified. Only envelopes with volume exceeding 100Å<sup>3</sup> were retained. A site coverage threshold is used rather than a precision threshold [8] to define success. 85.7% of 5616 protein-ligand binding sites were correctly identified with coverage greater than 80%. The vast majority of these were identified in the largest predicted site. Without a coverage threshold, the success rate increased to 96.8%.

## KNOWLEDGE-BASED FUNCTIONAL SITE PREDICTION

Knowledge of binding sites can often be obtained using biochemical data, if not directly in the form of high resolution structure determination methods (X-ray, NMR), then indirectly by NMR relaxation studies or site-directed mutagenesis and related experimental techniques. Comparative (homology) modelling studies can sometimes identify ligand binding sites, since they are often highly conserved. This knowledge is employed by Pupko *et al.* [48]

(Rate4Site), de Rinaldis *et al.* [49] and Armon *et al.* [50] (ConSurf). Databases such as PROSITE [51] can also be used to identify sequence similarity with known ligand binding sites for which structures are available. Alternatively, structural similarity with a known ligand-protein complex can indicate the presence of a binding site. Structural alignments of binding sites are held in the LigBase database [52]. Structural similarity studies are often particularly successful with enzymes. The enzyme active site is often the primary ligand binding site, and usually shows very high levels of structural similarity between enzymes that catalyse similar reactions [14]. One example is the serine protease family, whose active sites consist of a serine, a histidine and an aspartate residue which each have highly conserved orientations. An algorithm designed to identify them was nearly 100% successful [53]. The same research group also used a support vector machine to identify enzyme serine hydrolase active sites and achieved an 85% success rate on a data set of 139 structures [54].

Several databases contain information about binding sites and/or allow comparison of binding sites to allow the recognition of shared protein function. CavBase [37], Patterns In Non-homologous Tertiary Structures (PINTS) [55], SiteEngine [56], eF-site [57], ProFunc [15] and SitesBase [58] hold three dimensional structural information about protein pockets and allow structural comparisons between them. This is also useful for analysing newly resolved protein structures with unknown function to assign function from structural similarity with characterised sites. This is also useful for analysing newly resolved protein structures with unknown function to assign function from structural similarity with sites of known function. These databases also have potential use in identifying potential drug interactions with proteins other than the intended drug target. SitesBase, ProFunc, eF-site, SiteEngine, and PINTS are available online (Table 1). SitesBase uses geometric hashing to pre-calculate results for an all-against-all binding site comparison at the atomic level for the Protein DataBank (PDB) [59]. PINTS, uses a different method which calculates results on the fly and uses a depth-first search. CavBase stores property based surface patches which can be compared against each other and is commercially available as part of Relibase+ [60]. SiteEngine [56] uses a low then high-resolution geometric search method to either search a known functional site against a set of complete protein structures, or against other known binding sites. eF-site [61] uses a method based on graph theory [57] to detect similarity between molecular surfaces. They used a training set of 10 pairs of proteins to determine what level of similarity calculated by their algorithm indicates a likely match. They applied their method to 18 newly resolved hypothetical proteins and found potential matches for each of them. Using this information, they were able to propose possible functions for the proteins. Laskowski *et al.* [14] derive 3D templates from the PDB which represent ligand and DNA binding sites as part of the ProFunc [15] server for predicting protein function from 3D structure. They calculate the similarity between any two templates by performing a rotation/translation to align the two templates. They apply the Dayhoff mutation data matrix to take into account mutations between physiochemically similar amino acids. They also calculate an expectation



value, which indicates the likelihood that two templates would match through random chance when screening the database. The algorithm was tested by searching a random subset of 100 CATH domains against 1337 CATH domains and was shown to distinguish between true positives and false positives. The best match was a true positive in 74% of cases.

## STATISTICAL AND MACHINE LEARNING APPROACHES

Statistical analysis of protein-ligand contacts and orientations can be used to predict ligand binding sites, for example PATCH [62] was developed to detect carbohydrate binding sites with a 65% success rate when tested on 40 proteins. Neural networks have also been used to classify enzymes based upon an identification of their active sites similarity [63, 64]. Similarly, surface property based approaches have also been used for predicting protein-protein interactions, including the use of support vector machines [65]. Stahl *et al.* [64] found the solvent accessible surface area using the Connolly algorithm [36] and surface points were allocated an interaction type ('aliphatic', 'hydrogen-bond donor', 'hydrogen-bond acceptor', 'aromatic-face' and 'aromatic-edge'). They located the five largest protein pockets, and classified them according to whether or not they bound to a ligand. A neural network was trained on 176 proteins and tested on 18 zinc-containing enzymes. The pockets were correctly classified for 16 of these structures. The neural network can be used for binding site classification and could be applied to the identification of protein function from structure.

## BLIND DOCKING

Blind docking is the process by which a standard docking tool is applied to a whole protein. It implicitly includes binding site prediction but also aims to give information about the correct ligand binding orientation. However, blind docking requires that the structure of the ligand is known. Other binding-site prediction tools do not have this prerequisite. Blind docking is very slow, especially when screening a large number of ligands against a protein. In summary, blind docking is most useful where the two binding partners are already known and the user is trying to identify a biologically relevant binding mode.

Blind docking has been conducted by Hetényi and Van Der Spoel [66] using AutoDock [67]. They successfully replicated the protein-ligand orientation in eight complexes. Blind docking was also conducted in the CASP2 docking challenge [68]. The challenge posed in CASP2 was: given the three dimensional structure of a ligand and protein, determine where a ligand will bind. Nine groups submitted predictions for seven protein-ligand complexes and one protein-protein complex. The overall results were good, with nearly all of the 77 predictions being within 3Å of the actual orientation, and over half were within 2Å. Therefore, despite the slow speed of such docking simulations, the results appear to be useful. There are many examples of studies where docking has been used to characterise a putative binding site *e.g.* Kurowski *et al.* [69] used AutoDock to characterise the

cofactor-binding site of methyltransferase by docking *S*-adenosylmethionine.

Blind docking has also been carried out by Bliznyuk and Gready [70], using a technique called van der Waals – fast Fourier transform (vdW-FFT). This involves using the OPLS or AMBER forcefield to estimate van der Waals energy terms at each point on a grid surrounding the protein. Fast Fourier transform is used to evaluate the binding affinity of different ligand orientations. The most favourable orientations should identify the binding site. The method was shown to be successful on a small test set.

## DOCKING INTO PREDICTED BINDING SITES

Virtual screening studies allow large libraries of compounds to be analysed to find out which ones are likely to bind to a protein target with high affinity. Such studies are much faster and cheaper than their experimental high-throughput screening counterparts, and frequently generate good results. Two review papers [71, 72] give many examples of virtual screening successfully leading to a drug being developed and ultimately released onto the market. There are many examples of binding site prediction being used in virtual screening studies, some of which are described below.

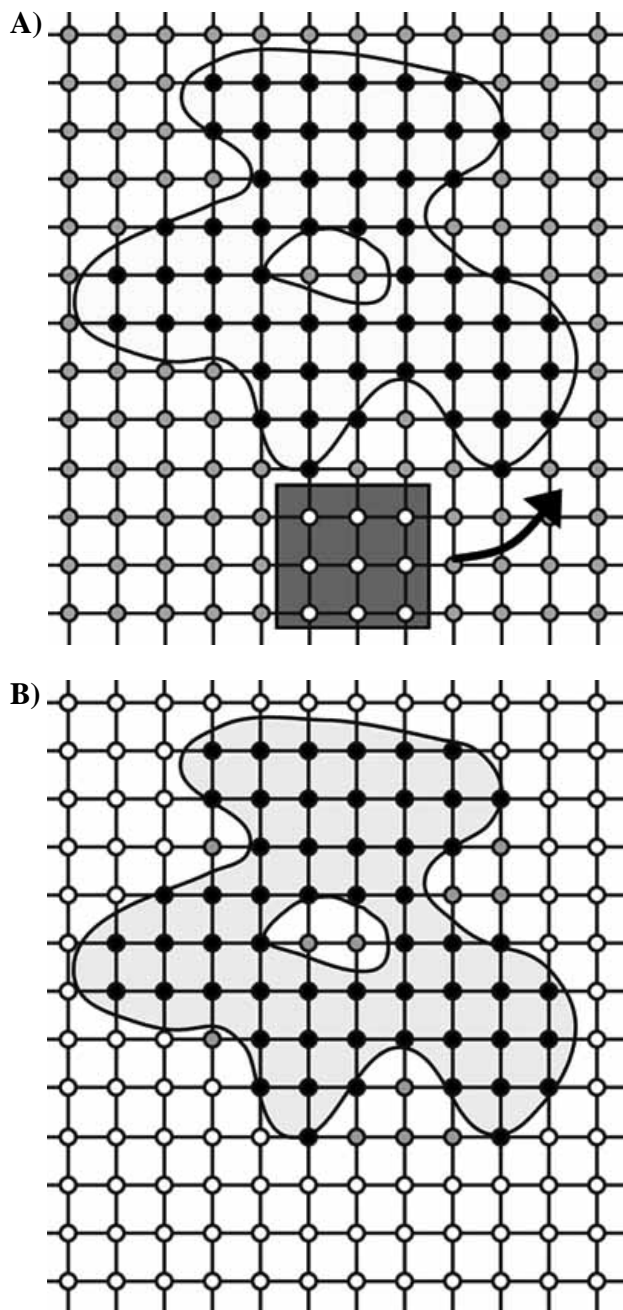
Li *et al.* [73] used APROPOS and DOCK [6, 74] to identify a potential CD4 binding site on a major histocompatibility complex (MHC) class II protein. This is an unusual use of APROPOS, since the CD4 binding site is a protein-protein rather than a protein-ligand binding site. However, the aim of the study was to find small molecule (non-peptide) inhibitors, so APROPOS was used to locate a potential ligand binding site within the protein-protein interface. They then performed a virtual screening study with 150,000 chemicals from the ACD database [75]. 41 compounds were selected for laboratory binding affinity testing, of which eight were found to have some activity (<100µM). The four most potent inhibitors were tested further to ensure that the inhibition was specific to MHC-II. These inhibitors are probably too weak to be used as lead compounds but could form a starting point for the design of other potential immunosuppressive drugs.

LIGSITE [5] is a popular method for pocket detection. It has been used to define binding sites in several applications, including *de novo* drug design [38], docking [76], functional site comparison [77] and CavBase [37]. For example, SuperStar [38] has been developed to generate propensity maps of basic molecular probe types on the surface of the protein, with the intention that it could be used in SBDD studies as a pharmacophore descriptor. It incorporates an implementation of the LIGSITE algorithm to analyse the protein surface for pockets.

Automated docking tools are those capable of docking a library of compounds to the potential sites, scoring them and returning the output to the user without requiring any user-interactivity. Sometimes they incorporate a binding site prediction algorithm. Hammerhead is a completely automated flexible docking tool [78]. It uses a method of binding site prediction related to that of Ruppert *et al.* [9] described above and uses the same scoring function [45]. It was tested by performing a virtual screening experiment on streptavidin

using 80,000 molecules. The highest scoring molecule was biotin, and it was docked successfully in the experimentally derived conformation.

LigandFit [31] incorporates a binding site prediction tool. Bindewald & Skolnick [79] also implemented the same cavity search routine in their docking algorithm. The algorithm is summarised in Fig. (10).



**Fig. (10).** The LigandFit pocket detection algorithm. The protein is placed in a three dimensional grid. Each grid point is tested to see if it is within the protein (black circles). **A:** A cubically shaped "eraser" is passed over the protein. It is obstructed by grid points within the protein. **B:** The cube "erases" free grid points (white circles). The remaining grey circles represent protein pockets and cavities.

LigandFit incorporates a protein pocket detection algorithm and a Monte Carlo stochastic ligand docking routine. Pocket detection is summarised in Fig. (10A) and Fig. (10B). It involves the creation of a grid representation around the protein, with each grid point being defined as free or occupied. A cubically shaped "eraser" then removes all accessible free grid points. The remaining free grid points define the pockets and cavities. The size of the eraser had a significant bearing on the calculated pocket volumes (estimated from the number of grid points that form a pocket). The size of the eraser had to be specified manually to obtain the best results. If an appropriate eraser size was used, the binding site was found in the largest identified pocket in 53 out of the 75 proteins tested. If a fixed eraser size of 5Å was used, 45 out of 75 proteins were found to have the ligand binding site in the largest pocket. LigandFit has been used in several virtual screening studies [80, 81] and has also been incorporated into a screensaver to allow the processing of large virtual screening tasks to be distributed across several computers (see, [www.grid.org](http://www.grid.org)).

## SUMMARY

Ligand binding site prediction is a broad and active field of research, and several different approaches have been adopted to address the problem. No method is 100% successful, and each have their own advantages and disadvantages. The first computational method to determine binding sites was pocket detection, pioneered by Cavity Search [25] and POCKET [26]. Subsequent advances in pocket detection yielded algorithms such as LIGSITE [5], APROPOS [24], PASS [30] and SurfNet [23]. Such algorithms show very good coverage of the ligand binding site. Pocket detection algorithms report success rates in the 70-90% range. However, the ligand binding site can be much smaller than the pocket in which it is found. Therefore, algorithms that rely on a geometric approach do not always define the precise location of the ligand binding site.

Energy-based methods of pocket detection include those of Ruppert *et al.* [9] and Q-SiteFinder [8]. The reported success rates are similar to the pocket detection algorithms, although we observed that Q-SiteFinder predicted sites with a higher average precision than pocket detection, which makes this type of algorithm suited to restricting the search space for SBDD. It is possible that a combination of the two types of algorithm may prove even more successful, *i.e.* detect pockets first with a pocket detection algorithm, and use Q-SiteFinder to further restrict the search space.

Identifying sequence and structural similarity with proteins with known functional sites is an emerging area for identifying ligand binding sites. This type of analysis is most often used to identify function from structure. The main drawback of this method is that it cannot detect binding sites that have no similarity with proteins of known function, whereas pocket detection and energy-based methods are independent of this information.

Identification of binding sites is of fundamental importance in SBDD and virtual ligand screening. It restricts the search space to the relevant parts of a protein complex, accelerating the process and reducing false-positive results. Functional site location is also extremely important for pre-

dicting function from structure [13]. We have provided an overview of the different methods of ligand binding site prediction and an insight into their use in SBDD. If little information is available about the ligand binding site or protein function it is strongly recommended that several different types of tool are used simultaneously to predict ligand binding sites before starting a virtual screening project. Otherwise, pocket detection and energy-based methods may be appropriate for defining the search space appropriate for SBDD.

## REFERENCES

- [1] Kaldor, S.W., Kalish, V.J., Davies 2nd, J.F., Shetty, B.V., Fritz, J.E., Appelt, K., Burgess, J.A., Campanale, K.M., Chirgadze, N.Y., Clawson, D.K., Dressman, B.A., Hatch, S.D., Khalil, D.A., Kosa, M.B., Lubbehusen, P.P., Muesing, M.A., A. K. Patick, S. H. Reich, K. S. Su. and J. H. Tatlock (1997) *J. Med. Chem.*, 40, 3979-85.
- [2] Kim, E. E., Baker, C. T., Dwyer, M. D., Murcko, M. A., Rao, B. G., Tung, R. D. and Navia M. A. (1995) *J. Am. Chem. Soc.*, 117, 1181-1182.
- [3] Campbell, S. J., Gold, N. D., Jackson R. M. and Westhead D. R (2003) *Curr. Opin. Struct. Biol.*, 13, 389-95.
- [4] DesJarlais, R. L., Sheridan, R. P., Seibel, G. L., Dixon, J. S., Kuntz, I. D. and Venkataraghavan R. (1988) *J. Med. Chem.*, 31, 722-9.
- [5] Hendlich, M., Rippmann, F. and Barnickel, G. (1997) *J. Mol. Graph. Model.*, 15, 359-63, 389.
- [6] Kuntz, I. D., Blaney, J. M., Oatley, S. J., R. Langridge and T. E. Ferrin (1982) *J. Mol. Biol.*, 161, 269-88.
- [7] Laskowski, R. A., Luscombe, N. M., Swindells, M. B. and Thornton, J. M. (1996) *Protein Sci.*, 5, 2438-52.
- [8] Laurie, A. T. and Jackson, R. M. (2005) *Bioinformatics*, 21, 1908-16.
- [9] Ruppert, J., Welch, W. and Jain, A. N. (1997) *Protein Sci.*, 6, 524-33.
- [10] Goodford, P. J. (1985) *J. Med. Chem.*, 28, 849-57.
- [11] Bhinge, A., Chakrabarti, P., Uthanumallian, K., Bajaj, K., Chakraborty, K. and Varadarajan, R. (2004) *Structure, (Camb.)*, 12, 1989-99.
- [12] Elcock, A. H. (2001) *J. Mol. Biol.*, 312, 885-96.
- [13] Jones, S. and Thornton, J. M. (2004) *Curr. Opin. Chem. Biol.*, 8, 3-7.
- [14] Laskowski, R. A., Watson, J. D. and Thornton, J. M. (2005) *J. Mol. Biol.*, 351, 614-26.
- [15] Laskowski, R. A., Watson, J. D. and Thornton, J. M. (2005) *Nucleic. Acids Res.*, 33, W89-93.
- [16] Artymiuk, P. J., Poirrette, A. R., Grindley, H. M., Rice, D. W. and Willett, P. (1994) *J. Mol. Biol.*, 243, 327-44.
- [17] Spriggs, R. V., Artymiuk, P. J. and Willett, P. (2003) *J. Chem. Inf. Comput. Sci.*, 43, 412-21.
- [18] Kleywegt, G. J. (1999) *J. Mol. Biol.*, 285, 1887-97.
- [19] Lichtarge, O., Bourne, H. R. and Cohen, F. E. (1996) *J. Mol. Biol.*, 257, 342-58.
- [20] Pazos, F., Helmer-Citterich, M., Ausiello, G. and Valencia, A. (1997) *J. Mol. Biol.*, 271, 511-23.
- [21] Kini, R. M. and Evans, H. J. (1995) *Biochem. Biophys. Res. Commun.*, 212, 1115-24.
- [22] Szilagy, A., Grimm, V., Arakaki, A. K. and Skolnick, J. (2005) *Phys. Biol.*, 2, S1-S16.
- [23] Laskowski, R. A. (1995) *J. Mol. Graph.*, 13, 323-30, 307-8.
- [24] Peters, K. P., Fauck, J. and Frommel, C. (1996) *J. Mol. Biol.*, 256, 201-13.
- [25] Ho, C. M. and Marshall, G. R. (1990) *J. Comput. Aid. Mol. Des.*, 4, 337-54.
- [26] Levitt, D. G. and Banaszak, L. J. (1992) *J. Mol. Graph.*, 10, 229-34.
- [27] Kleywegt, G. J. and Jones, T. A. (1994) *Acta. Crystallogr. D. Biol. Crystallogr.*, 50, 178-85.
- [28] Binkowski, T. A., Naghibzadeh, S. and Liang, J. (2003) *Nucleic Acids. Res.*, 31, 3352-5.
- [29] Liang, J., Edelsbrunner, H. and Woodward, C. (1998) *Protein Sci.*, 7, 1884-97.
- [30] Brady Jr, G. P. and Stouten, P. F. (2000) *J. Comput. Aid. Mol. Des.*, 14, 383-401.
- [31] Venkatachalam, C. M., Jiang, X., Oldfield, T. and Waldman, M. (2003) *J. Mol. Graph. Model.*, 21, 289-307.
- [32] Delaney, J. S. (1992) *J. Mol. Graph.*, 10, 174-7, 163.
- [33] Del Carpio, C. A., Takahashi, Y. and Sasaki, S. (1993) *J. Mol. Graph.*, 11, 23-9, 42.
- [34] Masuya, M. and Doi, J. (1995) *J. Mol. Graph.*, 13, 331-6.
- [35] Lee, B. and Richards, F. M. (1971) *J. Mol. Biol.*, 55, 379-400.
- [36] Connolly, M. L. (1983) *Science*, 221, 709-13.
- [37] Hendlich, M. (1998) *Acta Crystallogr. D. Biol. Crystallogr.*, 54, 1178-82.
- [38] Verdonk, M. L., Cole, J. C., Watson, P., Gillet, V. and Willett, P. (2001) *J. Mol. Biol.*, 307, 841-59.
- [39] Stouten, P. W. F., Frommel, C., Nakamura, H. and Sander, C. (1993) *Mol. Simul.*, 10, 97-120.
- [40] Delaunay, B. (1934) *Otdelenie Matematicheskikh i Estestvennykh Nauk*, 7, 793-800.
- [41] Jackson, R. M. (2002) *J. Comput. Aid. Mol. Des.*, 16, 43-57.
- [42] Wade, R. C., Clark, K. J. and Goodford, P. J. (1993) *J. Med. Chem.*, 36, 140-7.
- [43] Wade, R. C. and Goodford, P. J. (1989) *Prog. Clin. Biol. Res.*, 289, 433-44.
- [44] Miranker, A. and Karplus, M. (1991) *Proteins*, 11, 29-34.
- [45] Jain, A. N. (1996) *J. Comput. Aided. Mol. Des.*, 10, 427-40.
- [46] Nissink, J. W., Murray, C., Hartshorn, M., Verdonk, M. L., Cole, J. C. and Taylor, R. (2002) *Proteins*, 49, 457-71.
- [47] An, J., Totrov, M. and Abagyan, R. (2005) *Mol. Cell Proteom.*, 4, 752-61.
- [48] Pupko, T., Bell, R. E., Mayrose, I., Glaser, F. and Ben-Tal, N. (2002) *Bioinforma.*, 18, Suppl. 1, S71-7.
- [49] de Rinaldis, M., Ausiello, G., Cesareni, G. and Helmer-Citterich, M. (1998) *J. Mol. Biol.*, 284, 1211-21.
- [50] Armon, D. Graur and N. Ben-Tal (2001) *J. Mol. Biol.*, 307, 447-63.
- [51] Appel, R. D., Bairoch, A. and Hochstrasser, D. F. (1994) *Trends Biochem. Sci.*, 19, 258-60.
- [52] Stuart, A. C., Ilyin, V. A. and Sali, A. (2002) *Bioinforma.*, 18, 200-1.
- [53] Chou, K. C. and Cai, Y. D. (2004) *Proteins*, 55, 77-82.
- [54] Cai, Y. D., Zhou, G. P., Jen, C. H., Lin, S. L. and Chou, K. C. (2004) *J. Theor. Biol.*, 228, 551-7.
- [55] Stark, A., Sunyaev, S. and Russell, R. B. (2003) *J. Mol. Biol.*, 326, 1307-16.
- [56] Shulman-Peleg, A., Nussinov, R. and Wolfson, H. J. (2004) *J. Mol. Biol.*, 339, 607-33.
- [57] Kinoshita, K. Furui, J. and Nakamura, H. (2002) *J. Struct. Funct. Genomics*, 2, 9-22.
- [58] Gold, N. D. and Jackson, R. M. (2006) *Nucleic. Acids. Res.*, in press.
- [59] Berman, H. M., Westbrook, J., Feng, G. Gilliland, T. N. Bhat, H. Weissig, I. N. Shindyalov and P. E. Bourne (2000) *Nucleic. Acids. Res.*, 28, 235-42.
- [60] Hendlich, M., Bergner, A., Gunther, J. and Klebe, G. (2003) *J. Mol. Biol.*, 326, 607-20.
- [61] Kinoshita, K. and Nakamura, H. (2005) *Protein Sci.*, 14, 711-8.
- [62] Taroni, C., Jones, S. and Thornton, J. M. (2000) *Protein Eng.*, 13, 89-98.
- [63] Gutteridge, A., Bartlett, G. J. and Thornton, J. M. (2003) *J. Mol. Biol.*, 330, 719-34.
- [64] Stahl, M., Taroni, C. and Schneider, G. (2000) *Protein Eng.*, 13, 83-8.
- [65] Bradford, J. R. and Westhead, D. R. (2005) *Bioinformatics*, 21, 1487-94.
- [66] Hetényi, C. and Van Der Spoel, D. (2002) *Protein Sci.*, 11, 1729 - 1737.
- [67] Morris, G., Goodsell, D., Halliday, R., Huey, R., Hart, W., Belew, R. and Olson, A. (1998) *J. Comp. Chem.*, 19, 1639 - 1662.
- [68] Dixon, J. S. (1997) *Protein, (Suppl. 1)*, 198-204.
- [69] Kurowski, M. A., Sasin, J. M., Feder, M., Debski, J. and Bujnicki, J. M. (2003) *BMC. Bioinformatics*, 4, 9.
- [70] Bliznyuk, A. and Gready, J. (1999) *J. Computa. Chem.*, 20, 983 - 988.
- [71] Alvarez, J. C. (2004) *Curr. Opin. Chem. Biol.*, 8, 365-70.
- [72] Hardy, L. and Malikayil, A. (2003) *Curr. Drug Discov.*, Dec., 16-20.

- [73] Li, S., Gao, J., Satoh, T., Friedman, T. M., Edling, A. E., Koch, U., Choksi, S., Han, X., Korngold, R. and Huang, Z. (1997) *Proc. Natl. Acad. Sci. USA*, 94, 73-8.
- [74] Ewing, T. J., Makino, S., Skillman, A. G. and Kuntz, I. D. (2001) *J. Comput. Aid. Mol. Des.*, 15, 411-28.
- [75] Voigt, J. H., Bienfait, B., Wang, S. and Nicklaus, M. C. (2001) *J. Chem. Inf. Comput. Sci.*, 41, 702-12.
- [76] Rarey, M., Kramer, B. and Lengauer, T. (1995) *Proc. Int. Conf. Intell. Syst. Mol. Biol.*, 3, 300-8.
- [77] Schmitt, S., Kuhn, D. and Klebe, G. (2002) *J. Mol. Biol.*, 323, 387-406.
- [78] Welch, W., Ruppert, J. and Jain, A. N. (1996) *Chem. Biol.*, 3, 449-62.
- [79] Bindewald, E. and Skolnick, J. (2005) *J. Comput. Chem.*, 26, 374-83.
- [80] Aparna, V., Rambabu, G., Panigrahi, S. K., Sarma, J. A. and Desiraju, G. R. (2005) *J. Chem. Inf. Model.*, 45, 725-38.
- [81] Mpamhanga, C. P., Chen, B., McLay, I. M., Ormsby, D. L. and Lindvall, M. K. (2005) *J. Chem. Inf. Model.*, 45, 1061-74.

Copyright of Current Protein & Peptide Science is the property of Bentham Science Publishers Ltd. and its content may not be copied or emailed to multiple sites or posted to a listserv without the copyright holder's express written permission. However, users may print, download, or email articles for individual use.