BIOINFORMATICS Bioinformatics Advance Access published February 22, 2005

Real spherical harmonic expansion coefficients as 3D shape descriptors for protein binding pocket and ligand comparisons

*Richard J. Morris, Rafael J. Najmanovich, Abdullah Kahraman, Janet M. Thornton**

EMBL-EBI, Wellcome Trust Genome Campus, Hinxton, Cambridge CB10 1SD, UK

ABSTRACT

Motivation: An increasing number of protein structures are being determined for which no biochemical characterisation is available. The analysis of protein structure and function assignment is becoming an unexpected challenge and major bottleneck towards the goal of well-annotated genomes. As shape plays a crucial rôle in biomolecular recognition and function, the examination and development of shape description and comparison techniques is likely to be of prime importance for understanding protein structure-function relationships.

Results: A novel technique is presented for the comparison of protein binding pockets. The method uses the coefficients of a real spherical harmonics expansion to describe the shape of a protein's binding pocket. Shape similarity is computed as the L_2 distance in coefficient space. Several thousand such comparisons can be carried out per second on a standard linux PC. Other properties such as the electrostatic potential fit seamlessly into the same framework. The method can also be used directly for describing the shape of proteins and other molecules.

Availability: A limited version of the software for the real spherical harmonics expansion of a set of points in PDB format is freely available upon request from the authors. Binding pocket comparisons and ligand prediction will be made available through the protein structure annotation pipeline Profunc (written by Roman Laskowski) which will be accessible from the EBI website shortly.

Contact: rjmorris@ebi.ac.uk

1 INTRODUCTION

With the worldwide rise of Structural Genomics initiatives and the increasing number of protein structures that are being solved for which little or no biochemical characterisation exist, the challenge of understanding how protein structure is related to function is becoming much more than a mere academic interest. Structure provides an excellent interpretation aid for biochemical data and a good starting point for generating further hypotheses, potentially leading to drug design. However function assignment based on structural analyses is in most cases far from trivial. Structural Genomics may well provide many new folds which will aid and greatly enhance the number of sequences for which homology models can be built, but without functional annotation and sufficient understanding of structure it could be argued that these new structures are not being fully exploited and add little new biological information.

In this paper, a method for modelling shapes - especially binding pockets in protein structures - is presented. Shapes are considered as functions on the unit sphere by describing each surface point by its spherical coordinates (r, θ, ϕ) and setting $f(\theta, \phi) = r$. We describe an efficient approach to describe this shape function by the coefficients of a real spherical harmonics expansion. Any (square-integrable) function on the unit sphere can be expanded in this manner and therefore the same methodology can be applied to enhance the description by including other properties such as the electrostatic potential.

1.1 Function Prediction from Structure

There have recently been a flood of interesting approaches to extract structural information about protein active sites and potential binding partners. See for example Aloy *et al*., 2001; Exner *et al*., 2002; Cai *et al*., 2002; Schmitt *et al*., 2002; Laskowski *et al*., 2003; Bate and Warwicker, 2004; Shulman-Peleg *et al*., 2004, and references within. For reviews related to function prediction from structure, see Orengo *et al*., 1999; Teichmann *et al*., 2001; Wild and Saqi, 2004, and especially Whisstock and Lesk, 2003, for a clear description of the problems involved and many examples.

1.2 Shape

Shape has long being recognised as a key concept in chemistry (Mezey, 1993). Especially in molecular biology where shape is a major factor (largely through non-covalent interactions) in virtually all processes/interactions within the cell and plays an important rôle in molecular recognition. It

*to whom correspondence should be addressed

is perhaps surprising that precisely in this area, shape is somewhat ill-defined, although a few accepted operational definitions exist.

For rigid bodies and macroscopic objects, shape is often described by combinations of standard geometrical objects such as the Platonic solids, by sets of intersecting planes, or by algebraic equations defining the extent of the object. At the molecular level, the electron density ultimately defines shape (Bader, 1990; Mezey, 1993). either by isocontours or extrema and saddle-points. For some applications (e.g. medium resolution scattering techniques, molecular graphics, structural bioinformatics) this approach is simplified and atoms are commonly approximated by a solid sphere representation.

Biomolecules under physiological conditions are constantly in motion, often undergoing quite dramatic conformational changes induced by the thermal energy of the surroundings and enhanced by the low-lying multi-minima form of the free energy landscape (Onuchic *et al*., 1997). The concept of shape is rather more tricky to define in such cases. A proper description of shape in this context would incorporate some probabilistic method to include coordinate uncertainties and motion, similar to the Gaussian surface description (Duncan and Olson, 1993; Laskowski, 1995) or fuzzy surfaces (Agishtein, 1992). However, it is common to consider the shape of a molecule as being determined by its surface and to treat it as a static entity. Surfaces are frequently determined using van der Waals atomic radii and solvent/probe accessible regions (Lee and Richards, 1971; Greer and Bush, 1978; Connolly, 1983; Voorintholt *et al*., 1989; Gabdoulline and Wade, 1996).

Mathematically there are several ways of describing and representing shapes including triangulations, polygons, distance distributions and landmark theory. The focus here will be on functional forms. Functions can be either local (piecewise - such as splines) or global in which the whole shape is described by an often very complex expression. Global representations are often termed parametric as the whole shape can be reduced to a number of parameters and each parameter affects the entire shape. Functions can either be explicit, meaning that one coordinate is expressed in terms of the others, or implicit, meaning that the surface points satisfy a given equation (isosurfaces). For example, spherical harmonics can be used for explicit functions, whereas super and hyperquadrics are implicit representations.

2 APPROACH

The goal is to extract functional information from protein structure based on the analysis and comparison of binding pockets. Our approach makes the basic assumption that proteins that bind similar ligands have clefts of similar size, shape and chemistry.

The first step is therefore to determine what we think might be a binding pocket and to get the best guess of its spatial extent. Although we wish to elaborate here mainly on the shape comparison algorithm, a brief overview of the whole process may help to put things in context. The overall flow of our approach for binding pocket shape description can be summarised as follows:

- Compute the clefts and cavities for a given macromolecule (SURFNET, Laskowski, 1995).
- Use characteristics such as the residue conservation score to identify which clefts may be binding pockets (ConSurf, Armon *et al*., 2001; Glaser *et al*., 2003) and to reduce its volume to encompass the most likely site of ligand binding.
- Transform the binding pocket to a standard frame of reference and compute the real spherical harmonic expansion coefficients that best approximate the shape.
- Scan a database of precomputed expansion coefficients of protein binding pockets and ligands and choose the most significant matches.

The individual steps are explained in the following subsections with additional background information and implementation details given thereafter.

2.1 Determination of potential binding pockets

Protein surfaces contain clefts and indentations of varying sizes and depths. It has been shown that in enzymes, the active site is commonly found within the largest cleft (Laskowski *et al*., 1996). While this observation makes the detection of the active site relatively easy, the shapes of the clefts often extend significantly beyond the region occupied by the ligand (see Figure 2 in Laskowski *et al*., 1996) and as such are not well-suited for direct shape comparisons with ligands. These larger volumes may be of functional importance but need to be reduced in size for our purposes. A detailed description and analysis of a new algorithm for predicting potential binding pockets will be presented elsewhere (Glaser *et al.*, in preparation). In brief, the algorithm uses the SURFNET program (Laskowski, 1995) to generate 3D shapes of the protein's cavities and clefts. SURFNET does this by placing spheres between protein atoms such that the radius of these spheres does not penetrate the atomic (van der Waals) radii of these two atoms or any nearby atoms. The union of these spheres is used by SURFNET to describe the 3D cleft shape. These SURFNET spheres used for the definition of the clefts in the protein are then filtered by the residue conservation score of the nearest residue. Residue conservation is calculated using the ConSurf algorithm (Armon *et al*., 2001). The retained non-overlapping clusters of SURFNET spheres are ranked by volume and total assigned residue conservation. The cluster with the top ranking is taken to determine the potential binding pocket. The binding pocket shape is thus

determined by the union of SURFNET spheres that are near conserved residues.

2.2 Harmonics and Real Spherical Harmonics

Spherical harmonics have a well-established standing in the molecular sciences. They are perhaps best known as the orbital shape determining functions as solutions of the angular part of Schrödinger's equation for the hydrogen atom (eigenfunctions of the angular momentum operators L^2 and L_z), although they occur in a great variety of different physical problems such as electromagnetism, gravity, mechanics, or hydrodynamics. Their attractive properties when dealing with rotations, spherical averaging procedures, or smooth surface representations on the sphere has led to extensive use in protein crystallography for the rotation function in molecular replacement (Crowther, 1972), in the computation of radially-averaged normalised structure factor profiles (Morris and Bricogne, 2003), in molecular docking (Ritchie and Kemp, 1999), in small-angle scattering low resolution shape determination (Stuhrmann, 1970; Svergun, 1991), and protein surface display routines (Duncan & Olson, 1993). Recently, Cai *et al*., 2002 built on a very similar approach to that of Ritchie and Kemp (Ritchie and Kemp, 1999, 2000) and described an efficient virtual screening algorithm using spherical harmonic molecular surfaces.

Spherical harmonics, $Y_{lm}(\theta, \phi)$, are single-valued, smooth (infinitely differentiable), complex functions of two variables, θ and ϕ, indexed by two integers, l and m. In quantum physics terminology, l is the angular quantum number and m the azimuthal quantum number. Roughly speaking, l gives the number of local minima of the function and therefore represents a spatial frequency. See any quantum mechanics or functional analysis textbook for more definitions and properties (for example Cohen-Tannoudji *et al*., 1977; Edmonds, 1996).

Spherical harmonics form a complete set of orthonormal functions and thus form a vector space analogue to unit basis vectors. In the same way that vector projections onto each axis (scalar product between vectors) can be used to describe any vector in the familiar form $\mathbf{x} = (x, y, z)^T$, expansion coefficients (scalar product between functions) can be used to describe functions. Any (square-integrable) function of θ and ϕ can be expanded as follows

$$f(\theta, \phi) = \sum_{l=0}^{\infty} \sum_{m=-l}^{l} c_{lm} Y_{lm}(\theta, \phi). \qquad (1)$$

Note that this expansion is exact and not merely an approximation. Errors are introduced by limiting the series to a certain order of l. A second source of error arises from the fact that the surface to be modelled need not be a function of θ and ϕ. For a closed surface in 3D to be single-valued and therefore a function of θ and ϕ, requires that any ray leaving the expansion centre should only penetrate the surface once, i.e. a continuous mapping exists between the surface and the unit sphere S_2. Such figures are called single-valued surfaces or *star-shape*, Figure 1. Proteins and ligands at high resolution are often not star-shape, although their low resolution images often approximately fulfil this requirement (Svergun, 1991; Ritchie and Kemp, 1999). Even if a surface is not truly star-shaped, the approximation of using an outer shell can nevertheless give useful and discriminative information about the shape. However, for binding pockets we have found the star-shape requirement is often fulfilled (the star-shape requirement is not used at any point for the construction of the binding pockets).

The expansion coefficients, c_{lm}, can be obtained by multiplying the above equation by the complex spherical harmonics and integrating over the solid angle Ω,

$$c_{lm} = \int_{S_2} f(\theta, \phi) Y^*_{lm}(\theta, \phi) d\Omega. \qquad (2)$$

These coefficients are unique and can therefore be used directly as a feature vector for describing the shape.

In this manner, a spectral decomposition of any (square-integrable) function may be carried out. The lower l values correspond to the low frequencies and describe the overall low-resolution shape, whereas the higher values add finer, high-frequency detail to the picture. The termination of the series at a given l thus corresponds to a (spatial) frequency filtering method (low-pass filter).

The complex-valued spherical harmonics can be combined to give real valued functions that share the same orthonormal and completeness properties. Real spherical harmonics are better suited for describing real surface functions. *Surface harmonics* are often defined as any combination of real spherical harmonics for fixed l and commonly used in shape analysis and deformational studies. In general, a surface harmonic is simply a harmonic function whose domain is a surface and is not restricted to any coordinate system or specific family of functions.

3 ALGORITHMIC DETAILS

3.1 Orientation

To be able to compare expansion coefficients directly they must be put in a standard frame of reference. Each molecule or binding pocket is translated so that its centre of geometry coincides with the origin of the coordinate system. The system is then rotated such that the second moment about the mean, the variance-covariance matrix,

$$\mu^2_{ij} = \mathcal{E}\{(X_i - \mu_i)(X_j - \mu_j)\} \quad i, j = x, y, z \qquad (3)$$

becomes diagonal with maximal values in x, followed by y, followed by z. This is equivalent to ordering the moments of inertia as z, y, x. This orientation is, however, only unique up to rotations of π around each axis (axis flips) due

to the symmetry of the second moment tensor (and the fact that negative eigenvectors remain eigenvectors with the same eigenvalues). In a similar manner, one can compute the third moment around the mean, which is a measure related to the skewness of a distribution. We have chosen to define an orientation for which the two diagonal elements of the third moment with the largest absolute values are made positive (the third diagonal element is then determined by the requirement that the system remains right-handed) as our standard orientation. This can always be achieved by a rotation about any of the axes, x, y, z, by π. The final position is indistinguishable from the original one by the first and second moments.

As spherical harmonics enjoy mathematically convenient rotational properties - the coefficients can be rotated in the same way as vectors with so-called Wigner matrices (Edmonds, 1996; Chaichian and Hagedorn, 1997) - the orientation convention introduced above is merely to speed up the process by avoiding having to search for optimal rotations between coefficients. The registration of 3D shapes is, however, not a trivial task and can be severely hampered by errors in the original shapes (Besl and McKay, 1992; Lanzavecchia *et al.*, 2001; Dugan and Altman, 2004). This skewness method should therefore be seen as a heuristic. Another approach would be either to store and search for all axis flips or to fall back on the optimisation problem of finding the best alignment. An attractive alternative would be the use of rotationally invariant descriptors (Kazhdan *et al.*, 2003).

3.2 Legendre Polynomials

As solutions of the angular part of Laplace's equation in spherical coordinates, spherical harmonics are functions of θ and ϕ that can be separated further into a purely θ dependent term multiplied by a purely ϕ dependent term. The θ functions are the well-known Legendre polynomials with argument $\cos\theta$, and the ϕ functions are simply exponential functions of $im\phi$. The computation of spherical harmonics therefore requires the evaluation of Legendre polynomials. It is well-known - especially in areas such as astrophysics and geophysics - that routines for the Legendre polynomials based on standard recurrence relationships (as found in the Numerical Recipes, Press *et al.*, 1996) become unstable and lead to overflow problems for higher orders of l. We therefore employ a stable recursion formula using extended-range arithmetic (Smith *et al.*, 1981).

3.3 Integration on S_2 and spherical t-designs

As may be seen from Equation 2, the computation of the expansion coefficients requires that the function to be expanded is integrated over the whole unit sphere. For functions that are available in analytical form, this integration can be carried out analytically in favourable cases, otherwise one must resort to numerical integration. The determination of integration points and their correct weights in a summation approach to the integral is, however, far from trivial and is still an active area of research (Jetter *et al.*, 1998). Techniques exist for determining the integration weights given a set of sample points and also for creating such a layout. Such techniques are often demanding and would require intensive computation to first establish a good point layout and then to optimise the parameters. Instead, we have employed mathematical objects known as spherical t-designs (Goethals and Seidel, 1979). A spherical t-design is a set of points, $\{\mathbf{p_1}, \mathbf{p_2}, \ldots, \mathbf{p_N}\}$, for which the integral of any polynomial, $f(\mathbf{x})$, of degree at most t over the sphere is equal to the average value (with equal weights) of the polynomial over the set of points

$$\int_{S_{d-1}} f(\mathbf{x}) d\mu(\mathbf{x}) = \frac{1}{N} \sum_{i=1}^{N} f(\mathbf{p_i}). \qquad (4)$$

$\mu(\mathbf{x})$ is a uniform normalised measure on the sphere of dimension d. For modelling surfaces, the function $f(\mathbf{p_i})$ is the radius from the centre for a given θ and ϕ. For binding pockets, this radius is computed by rolling a 1.4 Å ball over the union of spheres obtained from the combination of SURFNET and conservation score filtering. The radius is the distance from the expansion centre to the closest surface point of the ball that is penetrated by a ray travelling outwards from the expansion at θ and ϕ. This gives a smoothed surface that is well-suited for the expansion in spherical harmonics. For molecules (ligands), the approach is similar but using the original molecule's atoms (and their van der Waal radii) instead of the SURFNET spheres.

Spherical t-designs are actually only proven to exist algebraically up to $t = 9$, but considerable numerical testing provides evidence to suggest that spherical t-designs have been found up to $t = 21$. These points are very uniformly spread and represent a highly efficient (optimal) sampling on the sphere for a given degree of function variability (order of the polynomial). We have tested a number of published sphere integration schemes and have consistently found the best and most stable results with the spherical t-designs. In particular, we have used the 240 point set that is suspected to be a spherical design of order 21 (Hardin and Sloane, 1996). For higher order expansions we use approximate equal-weight integration layouts of up to 900 points, published by Fliege and Maier, 1996. Triangulation methods typically represent a huge oversampling of surface (depending of the details one is interested in and the degree of surface subdivision). Our set of points is far smaller (and yet still sufficient) than that typically obtained from a surface triangulation algorithm and doesn't require the computation of such or the integration weights and is therefore a major factor in the efficiency of our method.

Although this direct coefficient computation has huge advantages in terms of speed, one can envision cases in

which a different approach may be more appropriate. Protein models can exhibit a great deal of variance with respect to the accuracy of individual atoms. To a very first approximation this is reflected in the spread of the crystallographic atomic displacement factors (temperature or B-factors). There are means of estimating the individual coordinate errors from Cruickshank's diffraction precision index (Cruickshank, 1999; Schneider, 2000) and this positional uncertainty (Ten Eyck, 2003) could be taken into account to construct a probability surface. It would then seem natural to fit the expansion coefficients so as to minimise the error-weighted squared differences between the probability surface values and the reconstructed radii. A similar method has been used for geophysical data and can be implemented to perform very efficiently due to the independency of the coefficients (Matheny and Goldgof, 1995).

3.4 Implementation

The above method was developed in LISP and then rewritten in C for speed and portability reasons. For the computation of the surface and the expansion coefficients to the order of $l = 20$, the compiled LISP code typically required execution times of under 1s on a standard linux box running Redhat 7.3 and using the LISP package CMUCL (CMUCL User's Manual, 2004). In C we made use of the GNU GSL (Galassi *et al*., 2003) libraries for matrix manipulations and the associated Legendre polynomials (stable to $l = 150$) for the computation of the spherical harmonics. The C code allowed for aggressive optimisation that pushed the execution time down to typically well under 0.01s per structure. A JAVA version of the method using the Colt Library takes approximately 2s for the same task. Note that this does not reflect in any way a fair comparison of the programming languages mentioned here, but merely indicates the efficiency of the method that runs in reasonable times even for byte-code.

4 RESULTS

Real spherical harmonic expansions have been computed for all ligands in the EBI-MSD (Boutselakis *et al*., 2003) ligand database, the HIC-Up database (Kleywegt and Jones, 1998), for a few selected enzyme families with well-known binding pockets, and for a large number of predicted binding pockets derived from surface cleft analyses and residue conservation scores. A detailed analysis of more biochemical focus will not be presented here. In this paper, we instead focus on the methodology but give examples that show the power of our approach and also the current limitations with suggestions for future improvements.

4.1 Accuracy

For functions on the sphere, spherical harmonic expansions can be constructed to any arbitrary error threshold (within the numerical accuracy of the method). However, it makes sense to truncate the expansion to match the type of feature detail one is looking at and thereby decrease the chances of fitting noise. To roughly capture the overall shape of a small molecule or binding pocket, we have found that terminating the expansion series at $l_{\max} = 6$ is usually sufficient (giving rise to $N = \sum_{l=0}^{l_{\max}=6} \sum_{m=-l}^{l} = 49$ coefficients). However, for the chosen spherical t-design integration layout the best results in terms of reproducing the original values were obtained for $l_{\max} = 14$ ($N = 225$ expansion coefficients). The accuracy was measured in root mean square deviations (rmsd) between the original sample points and reconstructed values. Taking different sets of points with which to compute the rmsd changed the picture remarkably little, showing that the spherical harmonics are very well-behaved between the sample points and represent well the overall shape. We found values of about 0.001-0.02 Å rmsd for most small molecules and binding pockets we tested (the values typically increase with the overall size). Figure 2 shows a few selected approximations for a predicted binding pocket. The comparison and clustering studies take place in 225-dimensional space ($l_{\max} = 14$). The difference in coefficient space between shape i and shape j is calculated as a standard L_2 distance,

$$d(i,j) = \sqrt{\sum_{l=0}^{l_{\max}} \sum_{m=-l}^{l} (c_{lm}^{i} - c_{lm}^{j})^2}. \qquad (5)$$

For visualisation purposes, these multi-dimensional spaces are displayed in 2D plots in which the coefficients are placed along the x-axis and their values are given on the y-axis. Figure 3 shows the first 49 coefficients (corresponding to a cut-off of $l_{\max} = 6$) for four ligands that fall into two clusters. One such cluster is shown in Figure 4 which contains all ligands within a sphere of radius of 2.5 in coefficient space.

4.2 Sensitivity

The combination of SURFNET with residue conservation scores often gives a good binding pocket prediction but unfortunately of varying accuracy when compared to the known location of the ligand (Glaser *et al*. in preparation). Small changes in the PDB coordinates (for instance due to side chain flexibility) can give rise to a different set of spheres from which the pocket is built. In Figure 5 we show the effect of randomising the SURFNET sphere centres for a predicted binding pocket. As can be seen this can potentially introduce quite large shifts in the expansion coefficient vectors, given significant binding pocket rearrangements (top curve). When comparing binding pockets with ligands using shape alone, it therefore doesn't make sense to distinguish between the different molecules shown in Figure 4 as the binding pocket error will be in the order of these differences or larger.

In Figure 6, the average deviations of not getting the reference frames exactly the same are shown. The rotations were selected by randomly sampling quaternion space (Kuipers, 2002). The plots show that deviations from the correct centre of geometry (the expansion centre) of about 0.5 Å and

rotational deviations of up to about 20° generate differences in the expansion coefficients comparable to the differences within the ligand cluster shown in Figure 4.

4.3 Conformational Flexibility

Structural flexibility is a major challenge in any comparison technique based on 3D objects. It is not immediately clear how to best handle the multiple conformations commonly observed in high resolution X-ray structures and especially in NMR models in structural comparisons, although there have been promising attempts (Schneider, 2002). For our approach, we are faced with conformational flexibility both on the side of the protein and ligand. On the protein side, the overall effect results in a plot very similar to Figure 5 for small changes. For larger changes such as domain movements, the binding pockets are often no longer detectable or are distorted beyond recogniton from the liganded protein structure. In Figure 7 the distances between expansion coefficients (computed with Equation 5) are displayed for some diverse conformations of current depositions of nicotinamide-adenine-dinucleotide (NAD) and adenosine-triphosphate (ATP) in the PDB (Bernstein *et al.*, 1977). For other flexible ligands the behaviour and spread of coefficient distances is very similar (data not shown).

4.4 Binding Pocket Comparisons

In order to show the potential of the present method for binding pocket comparisons, we generated a test set of forty proteins with low pairwise sequence identity. The test set contains ten examples each of three different ligands (Adenosine-5'-Triphosphate, ATP; Nicotinamide-Adenine-Dinucleotide, NAD; and Heme, HEM) with the remaining ten examples all bound to five distinct but chemically similar steroids (Estradiol, EST; Progesterone, STR; Equitinin, EQU; Testosterone, TES and Dihydrotestosterone, DHT). Binding pockets were determined using SURFNET (Laskowski *et al.*, 1996) as described in Section 2.1 with an additional filtering method based on the proximity of the SURFNET spheres to atoms belonging to residues known to interact with the ligand in question. The residue interaction information was obtained from PDBsum (Laskowski *et al.*, 2005). The binding pocket shapes were then expanded in real spherical harmonics and their coefficients were compared using Equation 5. Figure 8 shows a dendrogram obtained from hierarchical clustering based on these distances and indicates the extent to which predicted binding pockets can be matched. As can be seen, steroid binding pocket shapes are similar to each other and sufficiently dissimilar to the remaining three binding pocket types to provide a clear distinction. The other binding pockets do not show such a clear separation, especially the ATP binding pockets show large variability and do not cluster well. The predicted heme and NAD binding pockets exhibit pronounced tendencies to cluster into separate groups. We expect the performance to increase as more properties are included into the binding pocket feature vector.

5 DISCUSSION

A fast method has been presented for capturing the global shape of a protein's binding pocket or ligand. The method can also be applied directly to the protein itself. The surface is treated as a (single-valued) function of the spherical coordinates, θ and ϕ, that can then be expanded in a linear combination of real spherical harmonics. The use of spherical designs for the integration provides a robust, fast and elegant approach for the determination of the expansion coefficients. The computation time for the expansion typically takes well under 0.01s (C version) on a standard i686 linux PC. If two objects share a common orientation, these coefficients can be directly compared using the standard Euclidean distance (L_2) metric in N-dimensional space, where N is determined by the order of the spherical harmonics chosen to represent the shape. We have presented a robust heuristic that defines a standard frame of reference based on the first, second, and third moments of a 3D object.

The method presented here has, however, a number of inherent difficulties. A problematic hurdle is predicting the binding pocket correctly. The whole idea behind comparing binding pockets by shape is based on actually having something close to the functionally relevant shape to start with. Especially when comparing binding pocket shapes to ligands, it is important to get a good geometric model of where the ligand may bind. Our approach will therefore perform badly in cases where the ligand lies on a fairly flat surface without much indentation or the ligand sticks out into the solvent. In many instances, the binding pocket could not be determined well with current methods, thus hampering any further steps. Given that each expansion coefficient is an integral over the whole surface with the spherical harmonics (Equation 2), i.e. their support is the full S_2, it is not possible to relate the coefficients to local shape features. Each coefficient always acts globally. It is therefore not well-suited to find local matches (subsolutions).

Another problem is how to deal with flexibility within the protein and the ligand. The inclusion of such effects into our approach is not straightforward without dealing with ensembles of potential conformations, probability surfaces or averaged coefficients for similar conformations. Given the speed of our approach, we are currently circumventing this problem and obtaining satisfactory results by simply storing all coefficients of various conformations (based on PDB entries rather than exploring the whole conformational space) and comparing against these.

The registration of a 3D object is not trivial and our heuristic for determining the coordinate frame of reference is not faultless. In the field of Computer Vision this problem has led to the development of 3D shape retrieval systems based

on rotation invariant descriptors (Funkhouser *et al*., 2003; Kazhdan *et al*., 2003). This approach loses information in that the original shape cannot be reconstructed from its descriptors but this is over-compensated for by the avoidance of registration errors.

Shape alone does not determine when interactions occur. At least of equal importance is the electronic configuration of all interacting partners. The electrostatic potential describes the total effective interaction energy that would be exerted on a point charge placed in this field. As the electrostatic potential is governed by Poisson's or Laplace's equation (for zero charge density), spherical harmonics are again a good choice for describing its solutions (Tsirelson and Ozerov, 1996; Ritchie and Kemp, 2000). The integration of the electrostatic potential is currently being analysed and will be presented elsewhere.

6 CONCLUSION

In this paper, the method of using real spherical harmonics to describe surfaces has been explained in detail, including implementation issues. It has been shown how this approach can be employed to compare (star-like) shapes efficiently. For protein binding pockets this method offers a robust, compact and fast shape-driven description and comparison method. It is not well-suited for the location of sub-groups or sub-patterns for which alternative approaches are currently being tested.

ACKNOWLEDGEMENT

RJM is grateful for financial support from SPINE contract-no QLG2-CT-2002-00988. We thank Roman Laskowski for providing help with and modifications to SURFNET, Gareth Stockwell for the NAD and ATP datasets, Fabian Glaser for advice on ConSurf, and Jonathan Barker for help with Figure 4. RJM would like to thank Gèrard Bricogne for the exposure to some of the mathematical methods employed in this approach.

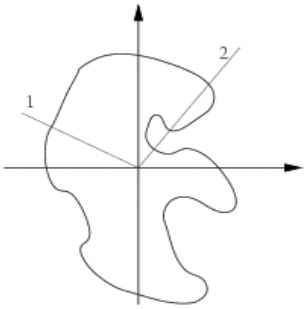

Fig. 1. For a figure to be *star-shaped* every ray leaving a defined centre (typically the centre of geometry) should penetrate the surface once and only once, as ray 1. Non-star-like surface features similar to those along ray 2 cannot be described as a unique function of their spherical coordinate angles. One can, however, always approximate any shape by taking the furthest penetration point of each ray. Despite losing internal pockets and similar features this approach often still provides valuable shape information.

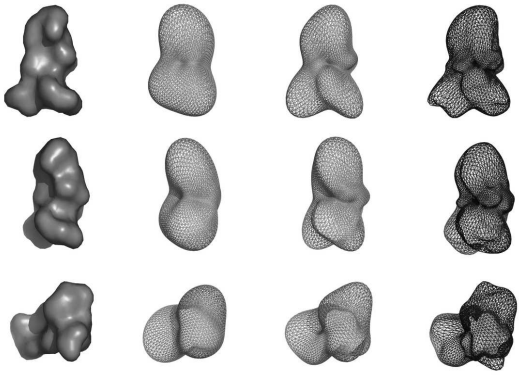

Fig. 2. The left column shows different views of a predicted binding pocket (SURFNET + residue conservation) in 1b14. The second column shows the reconstructed 3D shape from computed real spherical harmonic expansion coefficients up to an order of $l_{\max} = 4$. The third column shows the reconstruction up to $l_{\max} = 6$ which already captures rather well the overall shape. The next additional coefficients give only fairly small improvements. The fourth column shows the reconstructed model up to the fourteenth order.

REFERENCES

Agishtein, M. E. (1992) Fuzzy molecular surfaces, *J. Biomol. Struct. Dynam.*, **9**, 759-768.

Aloy, P., Querol, E., Aviles, F. X. and Sternberg, M. J. E. (2001) Automated Structure-based Prediction of Functional Sites in Proteins: Applications to Assessing the Validity of Inheriting Protein Function from Homology in Genome Annotation and to Protein Docking, *J. Mol. Biol.* **311**, 395-408.

Armon, A., Graur, D. and Ben-Tal, N. (2001) ConSurf: an algorithmic tool for the identification of functional regions in proteins by surface mapping of phylogenetic information, *J. Mol. Biol.*, **307**, 447-463.

Babbit, P. (2003) Definitions of enzyme function for the structural genomics era, *Curr. Op. Chem. Biol.*, **7**, 230-237.

Bader, R. F. W. (1990) Atoms in Molecules - A Quantum Theory. Oxford University Press. ISBN: 0-19855-865-1.

Bate, P. & Warwicker, J. (2004) Enzyme/Non-enzyme Discrimination and Prediction of Enzyme Active Site Location Using Charge-based Methods, *J. Mol. Biol.*, **340**, 263-276.

Bernstein, F. C., Koetzle, T. F., Williams, G. J. B., Meyer, Jr., E. F., Brice, M. D., Rodgers, J. R., Kennard, O., Shomanouchi, T., and Tasumi, M. (1977) The Protein Data Bank. A computer-based archival file for macromolecular structures. *J. Mol. Biol.* **112**, 535-542.

Besl, P. J. and McKay, N. D. (1992) A method for registration of 3D shapes, *IEEE Transactions on Pattern Analysis and Machine Intelligence*, **14**, 239-256.

Boutselakis, H., Dimitropoulos, D., Fillon, J., Golovin, A., Henrick, K., Hussain, A., Ionides, J., John, M., Keller, P. A.,

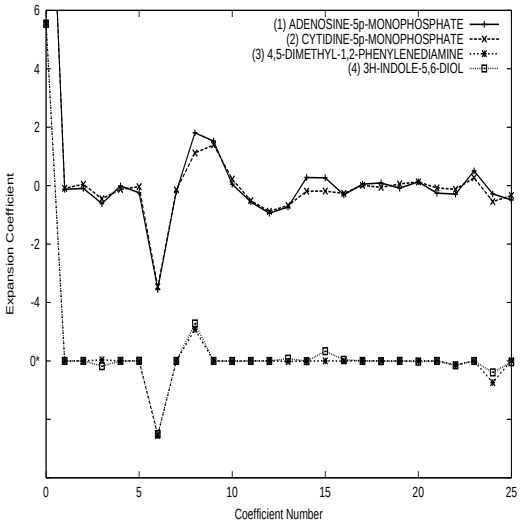

Fig. 3. The real spherical harmonics expansion coefficients for HIC-Up entries for adenosine-5p-monophosphate, cytidine-5p-monophosphate, 4,5-dimethly-1,2-phenylinediamine and 3h-indole-5,6-diol. Only the first 25 coefficients are shown. The curves (3,4) are offset vertically (for representation purposes only) compared to curves (1,2) as indicated by 0* in the plot. The distances in coefficient space (Equation 5) between curves (1) and (2) are 1.35, between (1) and (3) 4.65, between (1) and (4) 4.63, and between (3) and (4) 0.67. Although the curves may not differ visually too strongly, the Euclidean distances between their coefficients clearly clusters (1) with (2) and (3) with (4) into separate groups.

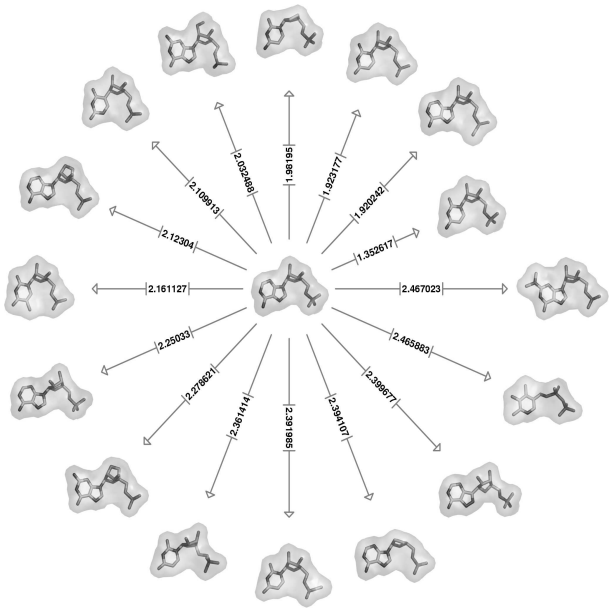

Fig. 4. The cluster of ligands associated with the centre molecule (adenosine-5p-monophosphate) for a tolerance distance in coefficient space of 2.5 (Equation 5).

Krissinel, E., McNeil, P., Naim, A., Newman, R., Oldfield, T., Pineda, J., Rachedi, A., Copeland, J., Sitnov, A., Sobhany, S., Suarez-Uruena, A., Swaminathan, J., Tagari, M., Tate, J., Tromm, S., Velankar, S. and Vranken, W. (2003) E-MSD: the European Bioinformatics Institute Macromolecular Structure Database, *Nucleic Acids Research*, **31**, 1, 458-462.

Cai, W., Shao, X., and Maigret, B. (2002) Protein-ligand recognition using spherical harmonic molecular surfaces: towards a fast efficient filter for large virtual throughput screening, *J. Mol. Graphics and Modelling*, **20**, 313-328.

Chaichian, M., Hagedorn, R. (1997) Symmetries in Quantum Mechanics: from Angular Momentum to Supersymmetry, IOP Institute of Physics, ISBN 0-750304073.

CMUCL User's Manual (2004). Editor Robert A. MacLachlan. Technical Report CMU-CS-92-161.

Cohen-Tannoudji, C., Dui, B., Laloê, F. (1977) Quantum Mechanics, Vol. 1 & 2, Wiley Interscience, ISBN 0-471-16432-1 & 0-471-16434-8.

Connolly, M. L. (1983) Analytical Molecular Surface Calculation, *J. Appl. Cryst.*, **16**, 548-558.

Crowther, R. A. (1972) The Fast Rotation Function, "The Molecular Replacement Method", Editors Rossmann, M. G., Gordon & Breach, New York, pp 173-178.

Cruickshank, D. W. J. (1999) Remarks about protein structure precision, *Acta Cryst.* **D**55, 583-601.

Dugan, J. M. and Altman, R. (2004) Using surface envelopes for discrimination of molecular models, *Protein Science*, **13**, 15-24.

Duncan, B. S. and Olson, A. J. (1993) Shape analysis of molecular surfaces, *Biopolymers*, **33**, 219-229.

Edmonds, A. R. Angular Momentum in Quantum Mechanics (1996), Princeton University Press, ISBN 0-691-02589-4.

Exner, T. E., Keil, M. and Brickmann, J. (2002) Pattern Recognition Strategies for Molecular Surfaces. I. Pattern Generation Using Fuzzy Set Theory, *J. Comput. Chem.*, **23**, 1176-1187.

Fliege, J. and Maier, U. (1996) A Two-Stage Approach for Computing Cubature Formulae for the Sphere. Technical Report. Ergebnisberichte Angewandte Mathematik 139T. Universität Dortmund, Fachbereich Mathematik.

Funkhouser, T., Min, P., Kazhdan, M., Chen, J., Halderman, A., Dobkin, D. and Jacobs, D.(2003) A search engine for 3D models, *ACM Transitions on Graphics*, **22**,1, 1-28.

Gabdoulline, R. R., Wade, R. C. (1996) Analytically defined surfaces to analyze molecular interaction properties, *J. Mol. Graph.*, **14**, 341-353.

Galassi, M., Davies, J., Theiler, J., Gough, B., Jungman, G., Booth, M. and Rossi, F. (2003) GNU Scientific Library Reference Manual (2nd), ISBN 0-9541617-3-4.

Glaser. F., Pupko, T., Paz, I., Bell, R. E., Bechor-Shental, D., Martz, E., and Ben-Tal, N. (2003) ConSurf: Identification of Functional Regions in Proteins by Surface-Mapping of Phylogenetic Information, *Bioinformatics*, **19**, 1, 163-164.

Goethals, J.-M. and Seidel, J. J. (1979) Spherical Designs, in D. K. Ray-Chaudhuri, ed. *Relations between Combinatorics and Other Parts of Mathematics, Proc. Symp. Pure Math.*, **34**, pp 255-272.

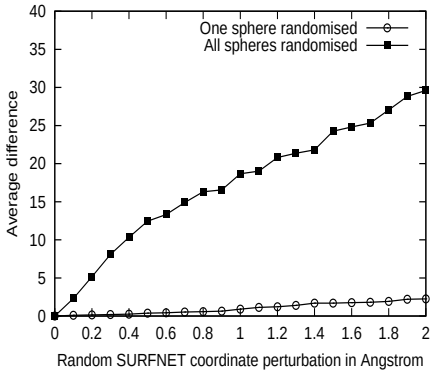

Fig. 5. This plot shows the differences in coefficient space (Equation 5) as a function of positional randomisations of SURFNET spheres up to a given threshold for each coordinate. Each point is the average over 1000 individual randomisations for a given maximum coordinate change. These plots give a very rough estimate of the order of total coefficient changes that can be expected for different magnitudes of shape deviation. The top curve shows a rough upper bound estimate (all SURFNET spheres were randomised), whereas the botton curve shows the effect of changing only one sphere (out of 168) at a time. The relative influence of atomic coordinate changes depends on the size of the binding pocket or ligand.

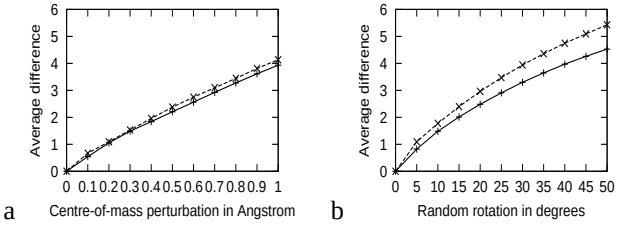

Fig. 6. The average differences in the expansion coefficients over 1000 geometry randomisations for adenosine-5p-monophosphate (lower curves) and adenosine-triphospate (top curves). Plot **a** shows the average deviations from the original expansion coefficient feature vector with respect to individual centre-of-geometry coordinate perturbations. Plot **b** depicts the coefficient vector distances for random rotations. For comparison, the nearest neighbour of adenosine-5p-monophosphate in coefficient space is 1.35 away. These plots merely show tendencies, the precise changes naturally depend on the individual ligand shape and size.

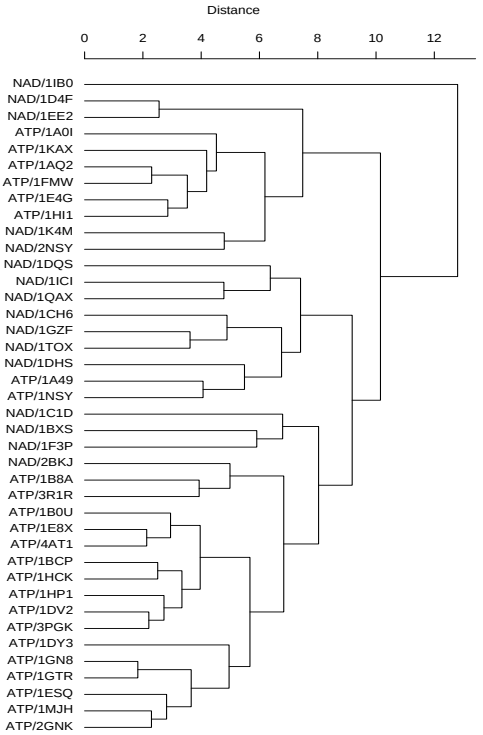

Fig. 7. Hierarchical clustering of the real spherical expansion coefficients for 16 different conformations of NAD and 25 different conformations of ATP taken from a set of low-sequence-similarity proteins with ligands. The separation is not perfect but overall the molecules are separated rather well by shape alone. The distances between the clusters are shown at the top.

Greer, J., Bush, B. (1978) Macromolecular Shape and Surface Maps by Solvent Exclusion, *Proc. Nat. Ac. Sc.*. **75**, 303-307.

Hardin, R.H and Sloane, J. A. (1996) McLaren's Improved Snub Cube and Other New Spherical Designs in Three Dimensions, *Discrete and Computational Geometry*, **15**, 429-441.

Jetter, K., Stöckler, J. and Ward, J. D. (1998) Norming sets and spherical cubature formulas, *Computational Mathematics*, pp 237-245, Eds. Z. Chen, Y. Li, C. A. Micchelli, Y. Xu, Marcel Dekker, Inc. New York, ISBN: 0-8247-1946-8.

Jones, S., Shanahan, H., Berman, H. M. and Thornton, J. M. (2003) Using electrostatic potentials to predict DNA-binding sites on DNA-binding proteins, *Nucl. Acids Res.*, **31**, 7189-7198.

Kazhdan, M., Funkhouser, T. and Rusinkiewicz, S. (2003) Rotation Invariant Spherical Harmonic Representation of 3D Shape Descriptors, in *Eurographics Symposium on Geometry Processing* editors Kobbelt, Schröder and Hoppe.

Kleywegt, G. J. and Jones, T. A. (1998) Databases in Protein Crystallography, *Acta Cryst.* D**54**, 1119-1131.

Kuipers, J. B. (2002) Quaternions and Rotations: a primer with applications to orbits, aerospace, and virtual reality, Princeton University Press, ISBN 0-691-10298-8.

Lanzavecchia, S., Cantele, F., Bellon, P. L. (2001) Alignment of 3D structures of macromolecular assemblies, *Bioinformatics*, **17**, 58-62.

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

Laskowski, R., Luscombe, N. M., Swindells, M. B. and Thornton, J. M. (1996) Protein clefts in molecular recognition and function, *Protein Sci.*, **5**. 2438-2452.

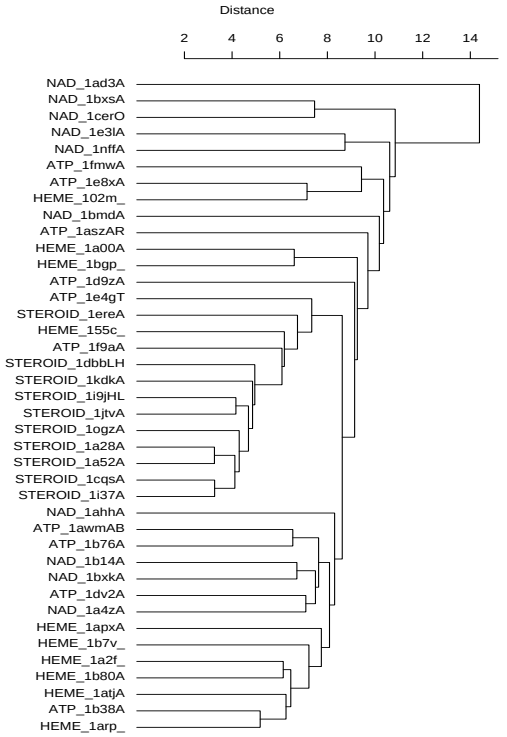

Fig. 8. Hierarchical clustering of 40 predicted binding pockets using their real spherical harmonic shape descriptors. The ligand type is given is capital letters (NAD, ATP, HEME, STEROID) followed by the PDB code and the chain IDs with which the ligand has contact. As can be seen, steroid binding pockets cluster quite well, heme and NAD binding pockets show some degree of separation, whereas ATP binding pockets cannot be well identified based only on their predicted shape.

Laskowski, R., Watson, J. D. and Thornton, J. M. (2003) From protein structure to biochemical function?, *J. Struct. Func. Genomics*, **4**, 163-177.

Laskowski R. A., Chistyakov V. V., Thornton J. M. (2005) PDBsum more: new summaries and analyses of the known 3D structures of proteins and nucleic acids. *Nucleic Acids Res.*, **33**, D266-D268.

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

Matheny, A. and Goldgof, D. B. (1995) The Use of Three- and Four-Dimensional Surface Harmonics for Rigid and Nonrigid Shape Recovery and Representation, *IEEE Trans. Pattern Analysis and Machine Intelligence*, **17**, 10, 967-981.

Mezey, P. G. (1993) Shape in Chemistry. An Introduction to Molecular Shape and Topology. VCH Publishers, Inc. ISBN: 0-89573-727-2

Morris, R. J. and Bricogne, G. (2003) Sheldrick's 1.2 Å Rule and beyond, *Acta Crystallogr.* D**59**, 615-617.

Onuchic, J. N., Luthey-Schulten, Z. and Wolynes, P. G. (1997) Theory of protein folding: The energy landscape perspective, *Annu. Rev. Phys. Chem.* **48**, 545-600.

Orengo, C. A., Todd, A. E. and Thornton, J. M. (1999) From protein structure to function, *Curr. Opin. Struct. Biol.*, **9**, 374-382.

Press, W. H., Flannery, B. P., Teukolsky, S. A. and Vetterling, W. T. (1996) Numerical Recipes in C: The Art of Scientific Computing, Cambridge University Press, ISBN: 0-521-43108-5.

Ritchie, D. W. and Kemp, G. J. L. (1999) Fast Computation, Rotation, and Comparison of Low Resolution Spherical Harmonic Molecular Surfaces, *J. Comp. Chem.*, **20**,4, 383-395.

Ritchie, D. W. and Kemp, G. J. L. (2000) Protein Docking Using Spherical Polar Fourier Correlations, *Proteins: Structure, Function, and Genetics*, **39**,4, 178-194.

Shulman-Peleg, A., Nussinov, R. and Wolfson, H. J. (2004) Recognition of Functional Sites in Protein Structures, *J. Mol. Biol.*, **339**, 607-633.

Schneider, T. R. (2000) Objective comparison of protein structures: error-scaled difference distance matrices, *Acta Cryst.* **D**56, 714-721.

Schneider, T. R. (2002) A genetic algorithm for the identification of conformationally invariant regions in protein molecules, *Acta Cryst.* **D**58, 196-298.

Schmitt, S., Kuhn, D. and Klebe, G. (2002) A New Method to Detect Related Function Amoung Proteins Independent of Sequence and Fold Homology, *J. Mol. Biol.*, **323**, 387-406.

Smith J. M., Olver F. W. J. and Lozier D. W. (1981) Extended-range arithmetic and normalised Legendre Polynomials, *ACM Transactions on Mathematical Software*, 93-105.

Stuhrmann, H. (1970) Interpretation of small-angle scattering functions of dilute solution and gases. A representation of the structures related to a one-particle scattering function, *Acta Crystallogr.* A**26**, 297-306.

Svergun, D. (1991) Mathematical methods in small-angle scattering data analysis, *J. Appl. Cryst.*, **24**, 485-492.

Teichmann, S. A., Murzin, A. G. and Chothia, C. (2001) Determination of protein function, evolution and interactions by structural genomics, *Curr. Op. Struct. Biol.*, **11**, 354-363.

Ten Eyck, L. F. (2003) Full Matrix Refinement as a Tool to Discover the Quality of a Refined Structure, *Meth. Enzymol.*, **374**, Macromolecular Crystallography, Part D, Eds. C. W. Carter and R. M. Sweet, pp 345-369.

Tsirelson, V. G. and Ozerov, R. P. (1996) Electron Density and bonding in crystals. ISBN 0 7503 0284 4. Institute of Physics Publishing, pp 147-167.

Yao, H., Kristensen, D. M., Mihalek, I., Sowa, M., Shaw, C., Kimmel, M., Kavraki, L. E. and Lichtarge, O. (2003) An accurate, scalable method to identify functional sites in protein structures, *J. Mol. Biol.*. **326**,1, 255-261.

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. Graph.*, **7**, 243-245.

Wei, L. and Altman, R. B. (2003) Recognizing Complex, Asymmetric Functional Sites in Protein Structures Using a Bayesian Scoring Function, *J. Bioinformatics and Comp. Biol.*, **1**, 1, 119-138.

Whisstock, J. C. and Lesk, A. (2003) Prediction of protein function from protein sequence and structure, *Quart. Rev. Biophysics*, **36**, 03, 307-340.

Wild, D. L. and Saqi, M. A. (2004) Structural Proteomics: Inferring Function from Protein Structure, *Current Proteomics*, **1**. 59-65.