Available online at www.sciencedirect.com
ScienceDirect

Journal of Molecular Graphics and Modelling 26 (2008) 1035–1045

Journal of Molecular Graphics and Modelling

www.elsevier.com/locate/JMGM

An extension of spherical harmonics to region-based rotationally invariant descriptors for molecular shape description and comparison

Lora Mak, Scott Grandison, Richard J. Morris *

John Innes Centre, Norwich Research Park, Colney Lane, NR4 7UH Norwich, UK

Received 27 May 2007; received in revised form 23 August 2007; accepted 25 August 2007
Available online 31 August 2007

Abstract

The use of spherical harmonics in the molecular sciences is widespread. They have been employed with success in, for instance, the crystallographic fast rotation function, small-angle scattering particle reconstruction, molecular surface visualisation, protein–protein docking, active site analysis and protein function prediction. An extension of the spherical harmonic expansion method is presented here that enables regions (bodies) rather than contours (surfaces) to be described and which lends itself favourably to the construction of rotationally invariant shape descriptors. This method introduces a radial term that extends the spherical harmonics to 3D polynomials. These polynomials maintain the advantages of the spherical harmonics (orthonormality, completeness, uniqueness and fast computation) but correct the drawbacks (contour based shape description and star-shape objects) and give rise to powerful invariant descriptors. We provide proof-of-principle examples illustrating the potential of this method for accurate object representation, an analysis of the descriptor classification power, and comparisons to other methods.

Keywords: Spherical harmonics; Zernike polynomials; Invariant descriptors; Molecules; Shape comparison

1. Introduction

The shape and the 3D distribution of physico-chemical properties of biological entities play a major rôle in biochemical processes, especially molecular recognition. The importance of shape, especially global shape, and shape metrics in molecular research has been argued convincingly in an elegant paper by Gramada and Bourne [1] that describes their own new approach to this problem. Hawkins et al. [2] show in a recent evaluation that shape-matching can be superior to docking for virtual screening. We refer to these papers [1,2] for a general introduction and background.

The main disadvantage of using RMSD (root mean square deviation) as a metric for structural comparisons is that it requires an atom-to-atom correspondence. This restriction is rather limiting. There are many examples for which this one-to-one mapping is not feasible but shape comparisons are still of interest. Especially for binding pockets but also for comparing protein and ligand shapes, a more global descriptor has advantages. As many before us, we have chosen to use a spherical harmonics based approach. Spherical harmonics have been applied with success in many areas of molecular biology. For instance, in the visualisation of molecular surfaces [3,4], crystallographic structure solution [5,6], molecular docking [7,8], virtual screening [9,10], in small-angle scattering shape determination [11,12], in the computation of radially averaged normalised structure factor profiles [13,14] and protein structure representation and comparison [10,1].

The main advantages of spherical harmonics expansions are the completeness and orthonormality, meaning that any function of the spherical coordinate angles θ and ϕ can be described uniquely to any required level of detail. In addition, it allows for physico-chemical properties to be described within the same mathematical framework. The main limitations are the restriction to representing only star-shape surfaces and the alignment problems associated with the fast comparison of objects.

Here, we present a new moment method for describing and comparing molecular shapes in 3D. This method may be viewed as an extension for the spherical harmonics expansion that employs Zernike radial functions [15] to sample objects over regions rather than surfaces. This extension results in a full

* Corresponding author.
E-mail address: Richard.Morris@bbsrc.ac.uk (R.J. Morris).

1093-3263/$ – see front matter © 2007 Elsevier Inc. All rights reserved.
doi:10.1016/j.jmgm.2007.08.009

3D density modeling method and provides a number of advantages whilst maintaining and enhancing the features of the pure spherical harmonics series representation.

2. From spherical harmonics to Zernike polynomials

2.1. The basis set

For shape representation and comparison there are a number of desirable properties that a well-suited method should possess such as completeness, uniqueness, transformation invariance, and speed of computation. Moment-based methods fulfill many of these desirable properties. Especially in 2D image analysis moment-based methods have become very popular due to their compression performance, efficient indexing and retrieval capabilities. 2D Zernike moments are amongst the most powerful moment-based methods. See Zhang and Lu [16] for a review on shape description and Celebi and Aslandogan [17] for a recent comparison of popular moment-based methods.

In Ref. [10] we approximated the surface of a molecule or binding pocket by a function of the spherical coordinates θ and ϕ. As any square-integrable function of θ and ϕ can be expanded as follows:

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

we then chose to describe the shape of a molecule or binding pocket by its expansion coefficients of a similar function series (for convenience we employed real spherical harmonics). Spherical harmonics representations have many of the desired properties.

We start our present developments with complex spherical harmonics. The spherical harmonics form a complete orthonormal basis which warrants that the expansion is unique and that the coefficients are independent (zero redundancy). However, the orthonormality and completeness are valid only on the unit sphere, making this method a contour-based method (can only represent surfaces) and in addition requiring a one-to-one mapping to the unit sphere (star-shape surface requirement). Although any square-integrable function of θ and ϕ can be represented to any arbitrary degree of accuracy, the function is itself only an approximation to the real surface and is only exact when a continuous mapping exists between the surface and the unit sphere. To address these limitations it becomes necessary to extend the spherical harmonics with a radial function, $R(r)$, thus enabling a full sampling of 3D space. Care must, however, be taken in not to lose the advantageous properties of the spherical harmonics in doing so. Different choices of radial function are possible. The main requirements being a good sampling of space (not too highly localised) and the orthogonality. Gram-Schmidt orthogonalisation can always be used to generate polynomials that are orthogonal.

As Canterakis [18] and Novotni and Klein [19], we have employed the radial part of Zernike polynomials. An alternative to the presented Zernike polynomial approach is the expansion into so-called regular and irregular solid harmonics [20] for which an efficient fast-multipole algorithm has been developed [21]. Other radial functions may also be employed as in the work of Ritchie [8]. Interesting new approaches include developments by the groups of Otwinowski and coworkers [22,23], Navaza [24] and Trapani and Ritchie [25]. Zernike polynomials have good sampling properties that have made them a popular choice for image recognition problems. Zernike polynomials were introduced to model the propagation of wavefronts in circular optical systems [15]. The Zernike polynomials are related to optical lens properties and different orders describe spherical aberration, focus, coma, astigmatism, and field curvature [26].

The radial part of the 2D Zernike polynomials is given by

$$R_{nl}(r)=\sum_{k=0}^{(n-l)/2}N_{nlk}r^{n-2k} \tag{2}$$

for $n-l$ even, otherwise zero. These functions can also be expressed in terms of Jacobi polynomials and are related to Bessel functions of the first kind. We previously explored the use of zeroth order Bessel functions of the first kind to describe rotationally invariant Fourier transforms of scattering density [14,10]. Bessel functions and spherical harmonics arise naturally from a separation of variables from Laplace's equation or by decomposing the 3D Fourier transform in radial and spherical components.

The normalisation constant N_{nlk} is typically chosen to ensure that $R_{nl}(1)=1$, leading to

$$N_{nlk}=\frac{(-1)^k(n-k)!}{k!\,[(1/2)(n+l)-k]![(1/2)(n-l)-k]!}. \tag{3}$$

With this normalisation, the radial Zernike functions obey the following orthogonality relationship in [0,1]

$$\int_0^1 R_{nl}(r)R_{n'l}(r)r\,\mathrm{d}r=\frac{1}{2n+2}\delta_{nn'}. \tag{4}$$

In 3D, we wish to build up an orthonormal basis over the full unit ball. We denote these new basis functions (3D Zernike polynomials) as

$$Z_{nlm}(\mathbf{r})=R_{nl}(r)Y_{lm}(\theta,\phi). \tag{5}$$

The normalisation requirement and the use of spherical harmonics naturally extends the 2D Zernike polynomials, which have been used with success in 2D image retrieval problems, to 3D [18,19]. Whereas 2D Zernike polynomials live on the unit disc, the 3D polynomials cover the unit ball. The radial functions of 2D Zernike polynomials must be modified to warrant orthonormality, as

$$\int_0^1\int_0^{2\pi}\int_0^{\pi}R_{nl}(r)Y_{lm}(\theta,\phi)R_{n'l'}(r)Y^*_{l'm'}(\theta,\phi)r^2\sin\theta\,\mathrm{d}\theta\,\mathrm{d}\phi\,\mathrm{d}r = \delta_{nn'}\delta_{ll'}\delta_{mm'} \tag{6}$$

results in the following condition for R_{nl}

$$\int_0^1 R_{nl}(r)R_{n'l}(r)r^2\,\mathrm{d}r=\delta_{nn'}. \tag{7}$$

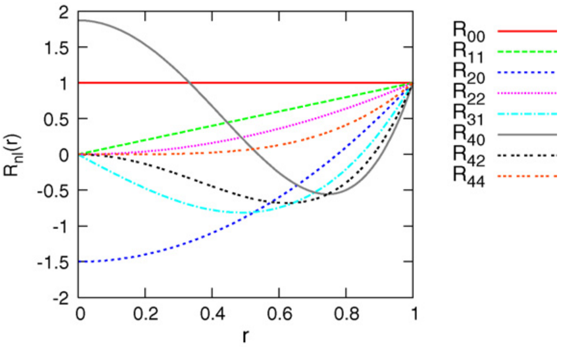

Fig. 1. Radial functions of the Zernike polynomials. The first few non-zero radial functions of the 3D Zernike polynomials, Eqs. (2) and (8), are shown.

Only the case $l = l'$ need be considered as the angular part of the Zernike polynomials (the spherical harmonics) contains $\delta_{ll'}$ in its orthogonality relationship. As may be seen by insertion or the use of Zernike recurrence relationships, the radial functions from the 2D Zernike polynomials result in a break of orthogonality. The radial functions above therefore require a re-orthogonalisation and normalisation due to the r^2 weighting. This can be achieved by setting

$$N_{nlk} = (-1)^k 2^{l-n}\sqrt{2n+3} \times \frac{(2n-2k+1)![(1/2)(n+l)-k]!}{[(1/2)(n-l)-k]!(n+l-2k+1)!(n-k)!k!} \qquad (8)$$

A sample of the first few radial functions (scaled by $\sqrt{2n+3}$) is shown in Fig. 1.

Any function within the unit ball can be represented as

$$f(\mathbf{r}) = \sum_{nlm} c_{nlm} Z_{nlm}(\mathbf{r}). \qquad (9)$$

The expansion coefficients (complex numbers), c_{nlm}, are often referred to as moments. The determination of these 3D moments requires that the object of interest (any distribution over $r\theta\phi$) multiplied by the Zernike polynomials be integrated over the unit ball.

2.2. Rotational properties of 3D Zernike polynomials and moments

The spherical harmonics themselves are not rotationally invariant and thus an initial alignment or a coefficient distance minimisation must be performed to compare objects. Although heuristics have been developed for the alignment of 3D objects, without additional optimisation these approaches are error-prone. The spherical harmonics, however, enjoy convenient rotational properties that allow them to be transformed easily with the use of so-called Wigner matrices (D matrices, Clebsch–Gordan coefficients, [27–29]). It can be shown that a rotation (change of θ and ϕ) introduces a mixing of the m states but leaves l untouched. This is a consequence of the fact that the rotation operator commutes with the total squared angular momentum operator (see [29] for an excellent and very didactic exposition of rotational symmetry in quantum mechanics and spherical harmonics). The Wigner matrices thus have a diagonal block structure, each block (submatrix) corresponding to l values and defining a subspace of the spherical harmonics. Each such submatrix is unitary, meaning that the transformation of this subspace is distance preserving. See Fig. 2 for a pictorial explanation of state-mixing. Thus, spherical harmonics are not transformation invariant, however, rotations can be performed easily using Wigner matrices and rotationally invariant subspaces can be found [30].

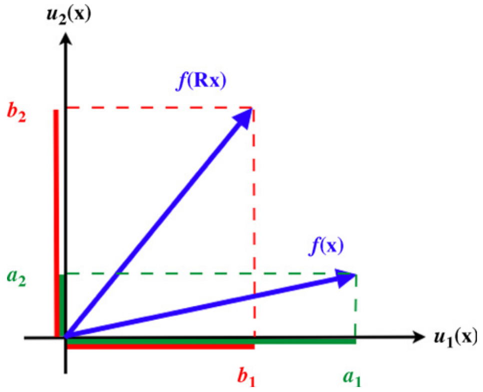

Fig. 2. Mixing of states. The projection of a function onto basis functions and the grouping of the expansion coefficients into rotationally invariant subspaces can be understood with the example of vectors in 2D. Projecting the function $f(\mathbf{x})$ onto the basis functions $u_1(\mathbf{x})$ and $u_2(\mathbf{x})$ gives expansion coefficients, a_1 and a_2, that are not rotationally invariant. A rotation, $\mathbf{R}$, of the object introduces a different distribution between these projections (state mixing). The magnitude of the vector in the space defined by a_1 and b_2 or b_1 and b_2 is, however, rotationally invariant. The amplitude of the vector built from spherical harmonics coefficients is not rotationally invariant, however, they can be broken down into subspaces of $2l+1$ dimensions similar to the 2D case shown here.

Taking the norm of the $(2l+1)$-dimensional vectors (with components c_{lm} from a spherical harmonics expansion) in these subspaces produces rotationally invariant features F_l

$$F_l = \left\| \begin{matrix} c_{l,-l} \\ c_{l,-l+1} \\ \vdots \\ c_{l,l} \end{matrix} \right\|. \qquad (10)$$

In this way, coefficients may be formed that indeed are rotationally invariant [30]. This work, in combination with expansions into shells, has led to high performance shape descriptors and powerful shape search engines [31]. This approach also enables the star-shape requirement to be removed by performing the expansion in concentric shells and not just the outer surface. A disadvantage is the loss of directional information, thus making a reconstruction impossible.

As a consequence of the spherical harmonic properties, the Zernike moments are also not invariant under rotation. However, the radial term naturally is invariant, so the same approach of defining $2l+1$ vectors is applicable and yields rotationally invariant descriptors. These rotationally invariant descriptors, F_{nl}, can be defined as norms of $(2l+1)$-dimensional vectors (with components c_{nlm} from a Zernike

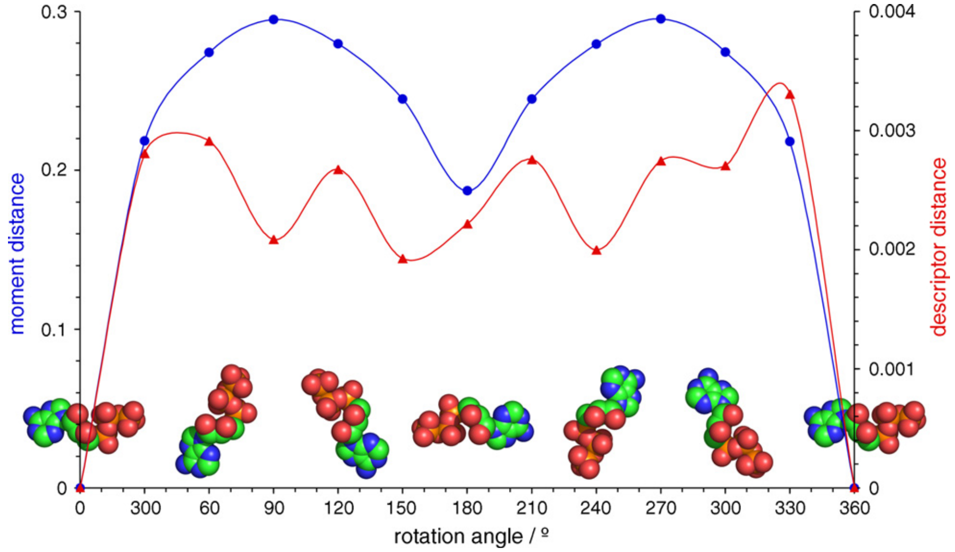

Fig. 3. Rotational behaviour of the Zernike moments and descriptors. This plot depicts the change in Zernike coefficient distance rotating an adenosine triphosphate (ATP) molecule through 360° around a chosen axis. For comparison the rotationally invariant descriptors are also shown. Although the values are not truly rotation invariant due to numerical inaccuracies, these errors are two orders of magnitude below the moment difference.

polynomial expansion)

$$F_{nl} = \left\| \begin{matrix} c_{nl,-l} \\ c_{nl,-l+1} \\ \vdots \\ c_{nl,l} \end{matrix} \right\|. \tag{11}$$

As for spherical harmonics, the coefficients obey the symmetry relation $c_{nl,-m} = (-1)^m c^*_{nlm}$.

To show the rotational invariance numerically, we selected molecules from our test dataset, systematically applied different rotations to them around an arbitrary axis of rotation, and compared the resulting Zernike moments and descriptors between the different starting orientations. In Fig. 3, the coefficient distance

$$d = \sqrt{\sum_{n=0}^{N_{\max}} \sum_{l=0}^{n} (F_{nl}^{\text{shape1}} - F_{nl}^{\text{shape2}})^2}, \tag{12}$$

relative to the object's initial orientation of the Zernike moments and descriptors is shown for adenosine triphosphate (ATP) over 360°. The moments vary significantly (differences up to 0.3), whereas the descriptors reproduce the rotational invariance to within numerical errors of the current implementation (differences less 0.0034).

2.3. The algorithm

To progress from a molecular structure (in, for instance, PDB format) to a set of Zernike expansion coefficients (moments) requires placing the molecule or the molecular properties one wishes to describe in a datastructure that is well-suited to sampling for the integration procedure. Algorithms for computing spherical harmonics are well-known, see Morris et al. [10] and references within. For the computation of the Zernike expansion coefficients, the integration must be carried out over the unit ball. We have experimented with extending the fast integration approach of Morris [32] for molecular-like objects by taking different spherical designs for sampling radii to ensure a near-uniform volume sampling within the unit sphere. Although this approach initially seemed appropriate and gave decent approximations for low order expansions, the method of Novotni and Klein [19] offers a substantial improvement in both ease of implementation and mathematical elegance. Novotni and Klein [19] present a 1D description of their approach and show how to extend to 3D. To summarise, their method expands the integrated Zernike polynomials as summations of geometric moments that are computed on an orthogonal grid which samples the shape to be described. This approach thereby provides a simple shape sampling strategy and an elegant integration scheme that avoids the need to compute the spherical harmonics or the radial terms directly. Interested readers should consult the papers of Canterakis [18] and Novotni and Klein [19] for further details.

The choice of integration scheme has a profound influence on the overall algorithm design. Following the cartesian integration method mentioned above leads to the following steps as depicted in Fig. 4.

(1) Read in the PDB file (or other format).
(2) Place the coordinate origin in the molecule's centre of mass. Scale such that the molecule fits within the unit ball.
(3) Place an orthogonal grid around the object and sample the molecule's properties at these grid points.
(4) Use the sampled values to compute an approximation to the integral for the determination of the Zernike moments via the geometric moments.
(5) Compute the Zernike moments and descriptors.

To compute the Zernike moments, an integration of the object over the full domain of the 3D Zernike polynomials, i.e. the unit ball, must be performed. Although the previous

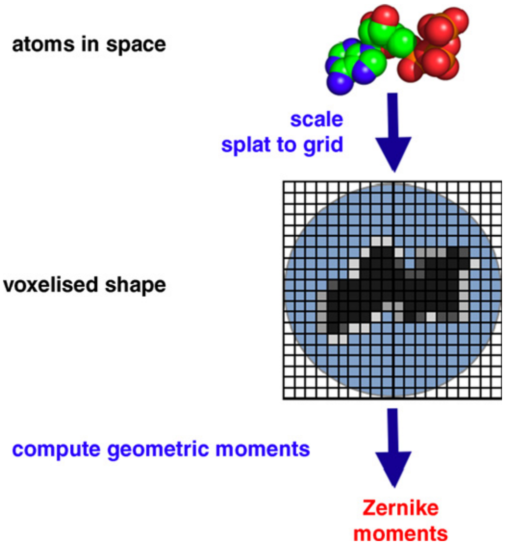

Fig. 4. Schematic representation of the Zernike moment computation. A molecule is scaled to fit well within the unit ball and converted to voxels. The integration region of the grid is shown is blue. The integration over the Zernike polynomials and the object is performed via the computation of geometric moments on the grid.

derivations have used only spherical coordinates, $\mathbf{r} = (r, \theta, \phi)$, which are indeed a very natural choice for this kind of problem, there are advantages of moving to cartesian coordinates, $\mathbf{x} = (x, y, z)$. In this basis, the Zernike polynomials can be expressed as linear combinations of monomials up to order n

$$\sum_{\alpha+\beta+\gamma \le n} x^\alpha y^\beta z^\gamma. \tag{13}$$

Neglecting normalisation constants and the complex linear combination terms, the computation of the expansion coefficients thus boils down to the calculation of geometric moments

$$\mu_{\alpha\beta\gamma} = \int_{|\mathbf{x}|\le 1} f(\mathbf{x}) x^\alpha y^\beta z^\gamma \, \mathrm{d}\mathbf{x}. \tag{14}$$

If one places a cubic grid around the object and samples the function one wishes to describe at $N \times N \times N$ grid points, then the above integral collapses into a summation over approximately $4\pi[(N-1)/2]^3/3$ boxes bounded by the sample points,

$$\mu_{\alpha\beta\gamma} \approx \sum_{\substack{ijk\\ |\mathbf{x}|\le 1}}^{N-1} \int_{x_i,y_j,z_k}^{x_{i+1},y_{j+1},z_{k+1}} f(\mathbf{x}) x^\alpha y^\beta z^\gamma \, \mathrm{d}\mathbf{x} \approx \sum_{\substack{ijk\\ |\mathbf{x}|\le 1}}^{N-1} \frac{x_{i+1}^{\alpha+1} - x_i^{\alpha+1}}{\alpha+1} \times \frac{y_{j+1}^{\beta+1} - y_j^{\beta+1}}{\beta+1} \frac{z_{k+1}^{\gamma+1} - z_k^{\gamma+1}}{\gamma+1} f_{ijk},$$

in which f_{ijk} is the voxel function value within the box defined by the sample points between i, j, k and $i+1, j+1, k+1$. f_{ijk} should be assigned such that the integral over each cube can be accurately reproduced, or alternatively the grid should be chosen such that $f(\mathbf{x})$ is approximately constant within each box. For modeling shapes that do not vary too wildly locally, this does not represent a problem and good results may be obtained by averaging the surrounding eight sample points (or just sampling the box mid-points).

3. Results and discussion

We first tested the reconstruction quality to ensure that we are indeed able to represent 3D shapes accurately with Zernike moments. Examples that highlight the improvements of this region-based method over our previous surface-based approach were chosen. The left column of Fig. 5 shows glutamine synthetase (1FPY) in two orthogonal views. The second column shows the voxelised (iso-contoured on a 64^3 grid) complex which was then used for the integration for the determination of the Zernike expansion coefficients. The next columns show reconstructions for the expansion orders 5, 10, 15, and 20. The improvement with expansion order can be observed, leading to a good reproduction quality of a highly non-starshape object. All molecular images in this paper were produced with PyMol [33] and Chimera [34].

As the Zernike moment computation takes place over the unit ball, this is the relevant domain for which the original and the reconstructed object should be compared. As a reconstruction metric we employ the root mean square deviation between the original and reconstructed object, $[\sum_{i=1}^{N} (f_i^{\text{original}} - f_i^{\text{reconstructed}})^2 / N^{\text{ball}}]^{1/2}$, where N^{ball} is the number of voxels within the unit ball. This has currently been

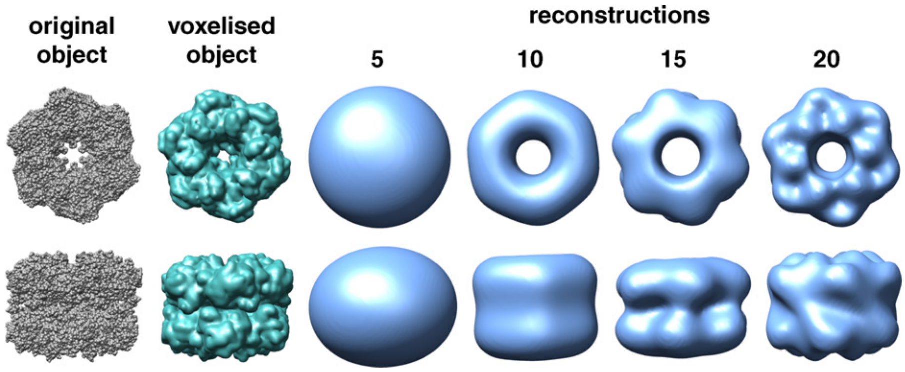

Fig. 5. The reconstruction of glutamine synthetase. This figure shows a spacefill, voxel description, and four Zernike moment reconstructions of expansion order 5, 10, 15, and 20 of the dimer of hexamers of glutamine synthetase (PDB code 1FPY).

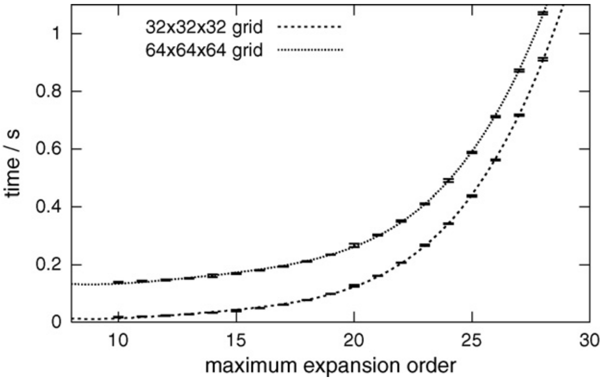

Fig. 6. The computation time dependence for Zernike descriptors as a function of expansion order. The curves represent the average values of 100 molecules with their standard deviations.

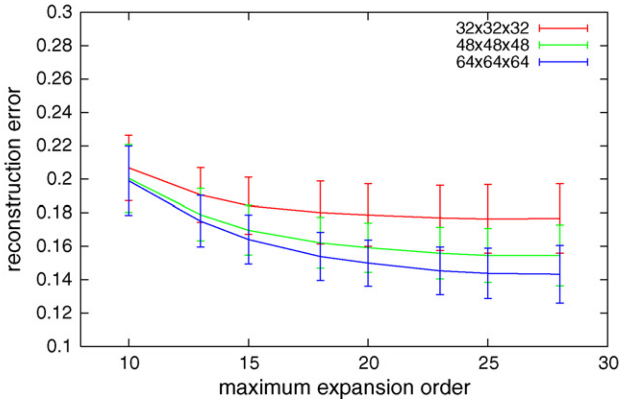

Fig. 8. The reconstruction error as a function of expansion order. The average values and standard deviations of the root mean square deviations are shown for a small test set consisting of ten molecules.

computed only for actual shapes (binary distributions within the ball) for which this metric becomes proportional to the number of incorrectly filled voxels over the total number of voxels, if one thresholds the reconstructed object. This thresholding gets rid of some of the noise in the reconstruction and can reduce the root mean square error down to in the order of 0.2 and less. The threshold for the images in this manuscript was chosen such as to minimise the reconstruction error. The threshold value is only of significance for reconstruction purposes to tidy the resulting image and is not used in any way for shape comparisons.

Figs. 6 and 7 show an analysis of the computation time as a function of expansion order and grid size over the ligand dataset. Figs. 8 and 9 show the reconstruction error. Higher expansion orders and larger grids do perform better but at a higher computational cost. In Fig. 10 the molecules flavin-adenine dinucleotide (FAD), nogalamycin (NGM), zanamivir (ZMR), and heme (HEM) are shown (first column) with their voxelised representations (second column), the Zernike moment reconstructions (third column), and their spherical harmonics reconstructions (fourth column). All these pictures show surfaces but the Zernike method reconstructs the full density. In Fig. 11 a slice through a reconstructed model of a heme is shown, displaying the voxel values also in the interior of the molecule. Many of these ligand structures have convoluted surfaces that deviate significantly from the star-shape property. A spherical harmonics reconstruction does very well at reproducing the outer shape but misses important features. As can be seen, the Zernike moments are capable of reproducing the original shape to a high degree of accuracy.

The total number of moments describing an object can be calculated from the moment order n as $(n+1)(n+2)(n+3)/6$. This can be significantly higher than the number of spherical harmonic expansion coefficients that would be needed to describe the surface to an equivalent level of detail, $(2l+1)^2$. Thus, for reconstruction purposes the Zernike polynomial expansion requires many more coefficients. For shape description, however, the rotationally invariant descriptors may be employed. The total number of Zernike descriptors is $(n/2+1)^2$ when n is even and $(n+1)(n+3)/4$ for n is odd, see Fig. 12. This leads to a significant saving and also eliminates any problems related to structural alignments prior to shape comparison. Despite some numerical shortcuts, the extension to 3D space comes at a cost and the current version takes over an order of magnitude longer to compute than our

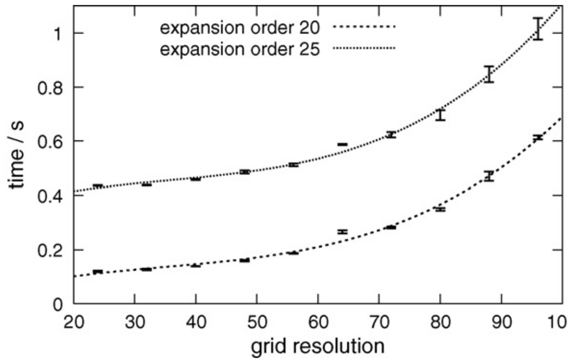

Fig. 7. The computation time dependence for Zernike descriptors as a function of grid resolution. The curves represent the average values of 100 molecules with their standard deviations.

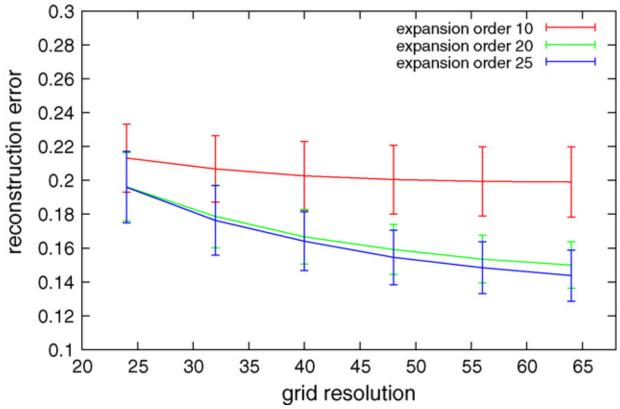

Fig. 9. The reconstruction error as a function of grid size. The average values and standard deviations of the root mean square deviations are shown for a small test set consisting of ten molecules.

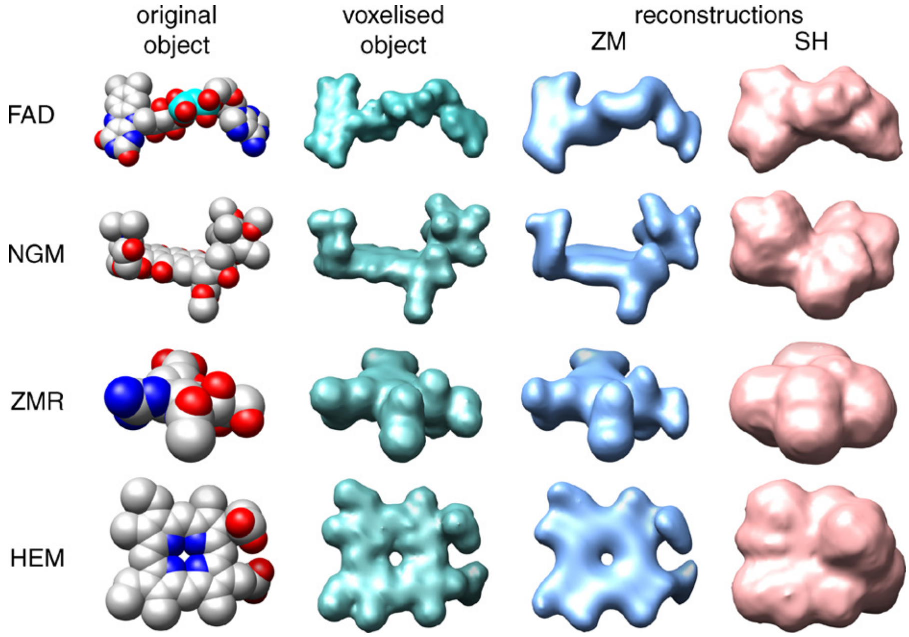

Fig. 10. Comparison of spherical harmonic and Zernike moment representations. The first row contains images of flavin-adenine dinucleotide (FAD), the second nogalamycin (NGM), the third zanamivir (ZMR), and fourth heme (HEM). The first column shows spacefill representations of the original molecules, the second the input shape, the third the Zernike reconstruction to order $n_{\max} = 20$, and the fourth column shows the spherical harmonic reconstruction to $l_{\max} = 14$.

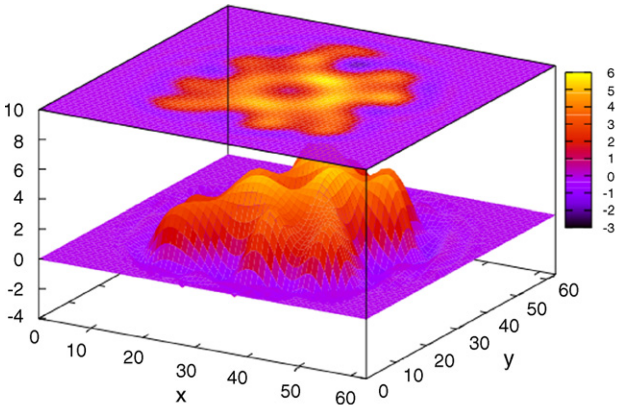

Fig. 11. A slice through a heme displaying the reconstructed density.

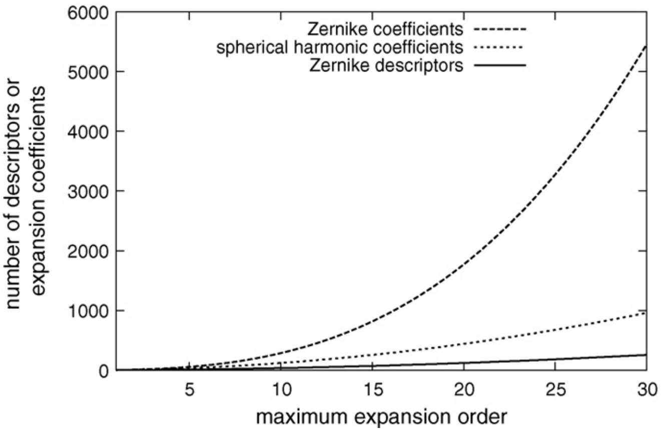

Fig. 12. The number of Zernike moments and descriptors as a function of the expansion order. The number of Zernike moments increases with the third power of the expansion order, spherical harmonics with the square, and Zernike descriptors also with the second power.

spherical harmonics implementation. This is, however, still under a second on a 2 GHz dual processor G5 with 4 GB RAM. Comparison times are fast, allowing for thousands to be carried out per second. We have computed Zernike moments and descriptors for a large set of ligands and protein structures. We present first our initial analysis on a test set of 100 ligands. This dataset consists of cognate ligands from non-homologous proteins (different CATH H-levels). The dataset is described in detail in Kahraman et al. [35]. For this 100 ligand dataset, we tested how well the Zernike descriptors retrieve the correct ligands for each query ligand. The performance was evaluated by computing precision-recall plots and ROC curves. The results are summarised as area under the ROC curve values, AUC, in Table 1. Although the reconstruction is significantly better than for spherical harmonics, the classification power is only marginally improved for this data set, producing an increase of only around 3%, or 7% if one excludes the size of

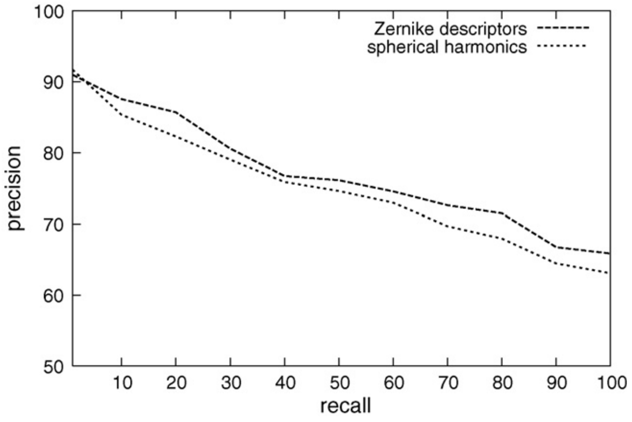

Fig. 13. Precision-recall plot over the 100 ligand dataset.

the molecule. This is reflected in the averaged precision-recall curves shown in Fig. 13. Precision is defined as TP/(TP + FP) and recall as TP/(TP + FN) in which TP stands for true positives, FP for false positives, and FN for false negatives. In this figure, the retrieval performance of the Zernike descriptors and the spherical harmonic coefficients is shown for the 100 ligand dataset. The Zernike descriptors perform consistently better, however, the improvement is less than we initially expected from the reconstruction power of the Zernike moments. Nevertheless, this classification power can be achieved with far fewer descriptors. Indeed, already order 10 gives good results which are only marginally worse than higher orders. Fig. 14 shows the classification by shape descriptors ($n_{\max} = 10$) for the ligand data set. Overall, the ligand types cluster very well. Two examples where this is not the case are highlighted, showing that although the ligand type is misclassified, the task of shape matching was preformed correctly. This illustrates the problem of using the ligand type to evaluate shape comparisons. In Fig. 15 a heat map of the Zernike descriptor distance matrix ($n_{\max} = 10$) is depicted showing distinct groupings of shapes (dark blue in block structures along the diagonal). It must be pointed out that ligand type matching is not the optimal test for evaluating shape comparison performance, although it is the first obvious test to carry out. Other metrics for the evaluation are being investigated.

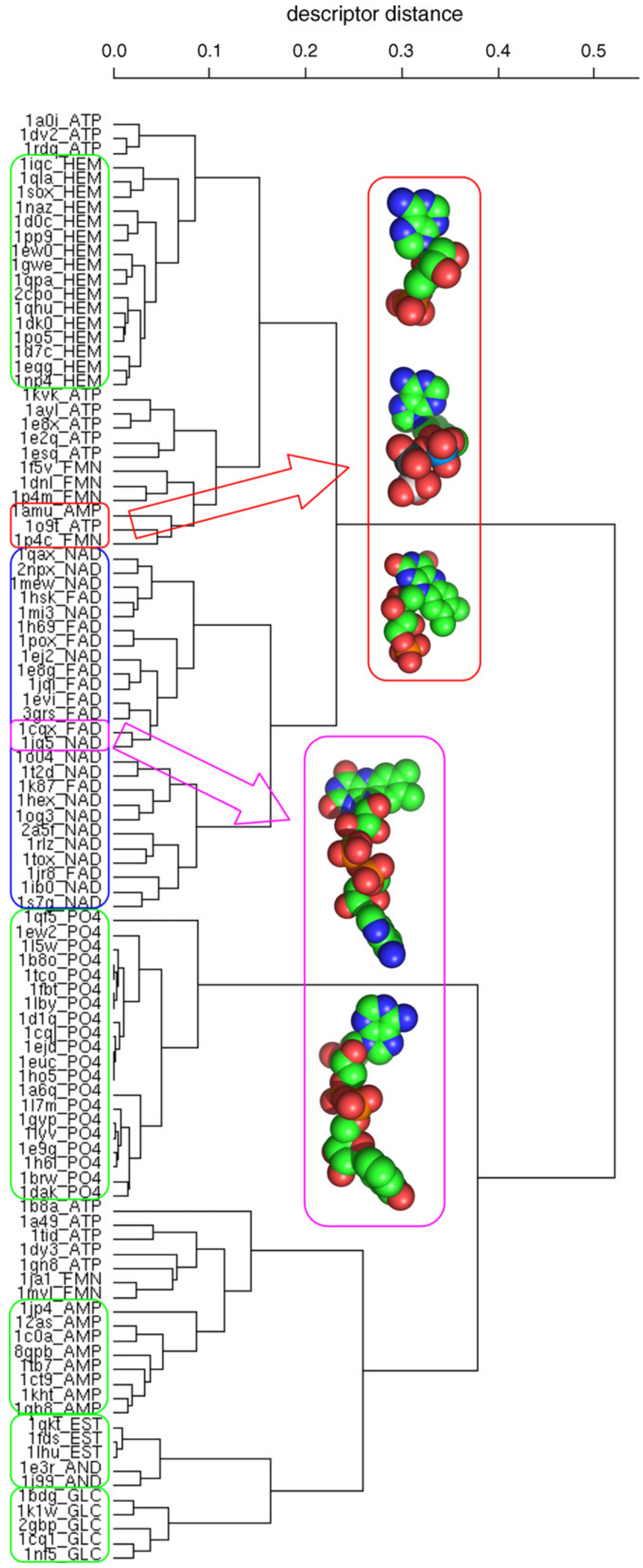

Fig. 14. Clustering of the 100 ligand dataset based on their Zernike descriptors.

Table 1
Zernike descriptor retrieval performance

Expansion order	Number of descriptors	AUC
10	36	0.943
15	72	0.944
20	121	0.945
25	182	0.946

The expansion order, the number of Zernike descriptors, and the area under the ROC curves (AUC) are shown. For comparison, our spherical harmonics approach using 225 coefficients achieved an AUC value of 0.87 for shape only and 0.92 for shape and size.

The second application is on a large set of proteins taken from Daras et al. [36]. This set of proteins was used by the authors to evaluate their Fourier-based structure comparison and classification software. Although not specifically designed for protein comparisons, we used our method on this set to test

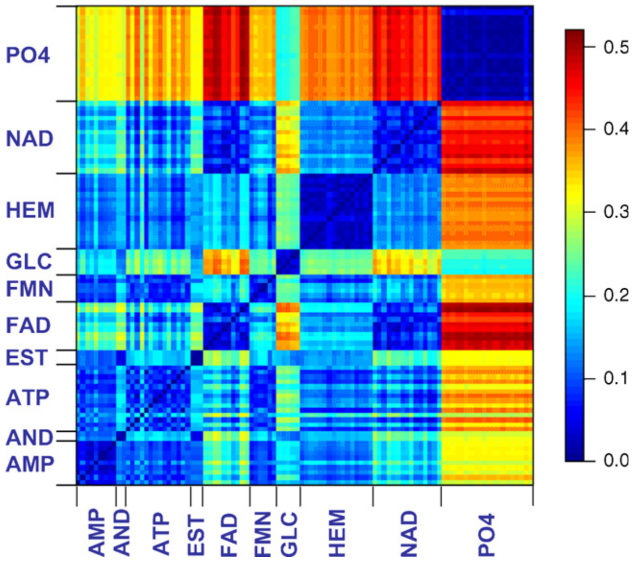

Fig. 15. Zernike descriptor distance heat map.

the performance of the Zernike descriptors for this purpose. The area under the ROC curve is 0.939 indicating that the global shape descriptors can perform this classification task extremely well. Misclassified proteins are shown in Fig. 16. As can be seen, the method picks up similar shapes, which in these cases do not, however, correspond to an evolutionary-based classification. A more detailed analysis will be presented elsewhere.

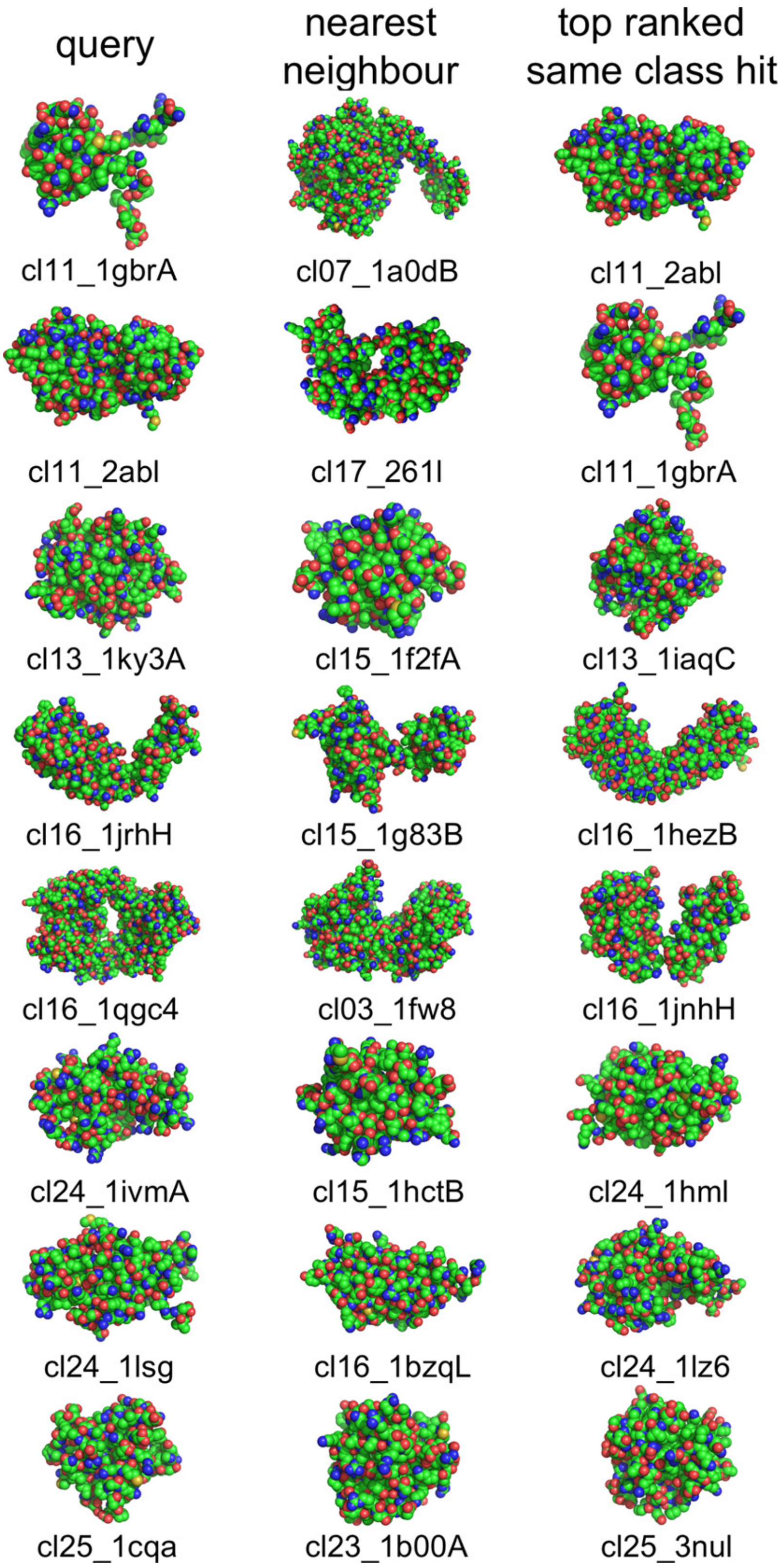

Fig. 16. Incorrectly classified proteins based on their Zernike descriptors. The proteins are denoted by ‘cl’ followed by the class number and their PDB code.

4. Conclusions

Powerful shape and 3D function matching techniques may have broad applications in computational biology, aside from numerous applications related to proteins and ligands, other interesting examples may include indexing cell imaging databases or describing the shape of biological phenotypes. The extension of spherical harmonics to incorporate radial sampling, whilst taking care to maintain the desirable orthonormality and completeness relationships, has led to the construction of functions equivalent to 3D Zernike functions. We have shown that these functions are well-suited to present molecular shapes and can successfully overcome some of the limitations of surface harmonics. This extra power comes at an additional computational cost per coefficient and also in that many more coefficients are required for the reconstruction. For shape matching, however, rotationally invariant descriptors may be employed thus reducing the number of coefficients greatly. For retrieval, we achieve a slightly higher performance over our previous work and a significantly higher information compression rate. Aside from this reduction in the number of coefficients and the rotational invariance, the advantages of this new method for shape matching were less impressive than we expected. Although, the reconstruction quality is superior compared to the pure spherical harmonics approach, the improvement in terms of classification and shape matching is only marginal. This confirms our previous analyses with spherical harmonics [10,35], in which we found that (1) the star-shape approximation is often sufficient for molecular shapes, (2) the surface often provides a sufficiently good description of a body (albeit ignoring internal structure), and (3) that higher moment heuristics for the orientation of objects are capable of providing a sufficiently accurate frame of reference. For significant deviations from star-shape objects, methods such as the one presented here provide the added detail required for a better reconstruction and shape matching. The reduced number of coefficients and rotational invariance of the Zernike descriptors, make this approach well-suited for indexing molecular databases based on their shape (or other spatial features). Comparing only shape cannot compete with more sophisticated sequence and secondary structure 3D alignment methods for classifying proteins, but it can avoid problems related to RMSD alignments and is very efficient. We therefore see the application of this approach in fast database scanning for finding similar shaped molecules, i.e. a fast low-resolution filtering method to be used prior to more detailed atom-based comparisons or docking studies. We have computed Zernike descriptors and spherical harmonic coefficients for the NCI ligand diversity set and have set up a shape-based virtual screening approach that is currently undergoing testing and evaluation.

Acknowledgements

We thank Dr. Dave Ritchie for helpful comments on an earlier draft of this manuscript. We would like to thank the authors of software packages who generously make their tools freely available (LATEX, gnuplot, emacs, gcc, and many many others) without which this work would have not been possible. LM is grateful for the funding provided by the John Innes Foundation and the ORSAS award. Research in RJMs group is supported by the UK Biotechnology and Biological Sciences Research Council (BBSRC).

References

[1] A. Gramada, P.E. Bourne, Multipolar representation of protein structure, BMC Bioinform. 7 (2006) 242–255.

[2] P.C.D. Hawkins, A.G. Skillman, A. Nicholls, Comparison of shape-matching and docking as virtual screening tools, J. Med. Chem. 50 (1) (2006) 74–82.

[3] N. Max, E.D. Getzoff, Spherical harmonic molecular surfaces, IEEE Comput. Graphics Appl. 8 (1988) 42–50.

[4] B.S. Duncan, A.J. Olson, Shape analysis of molecular surfaces, Biopolymers 33 (1993) 219–229.

[5] R.A. Crowther, The fast rotation function, in: M.G. Rossmann (Ed.), The Molecular Replacement Method, Gordon & Breach, New York, 1972, pp. 173–178.

[6] J. Navaza, AMoRe: an automated package for molecular replacement, Acta Crystallogr. A 50 (1994) 157–163.

[7] D.W. Ritchie, G.J.L. Kemp, Fast computation, rotation, and comparison of low resolution spherical harmonic molecular surfaces, J. Comput. Chem. 20 (4) (1999) 383–395.

[8] D.W. Ritchie, G.J.L. Kemp, Protein docking using spherical polar fourier correlations, Proteins: Structure Funct. Genet. 39 (4) (2000) 178–194.

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

[10] R.J. Morris, R.J. Najmanovich, A. Kahraman, J.M. Thornton, Real spherical harmonic expansion coefficients as 3D shape descriptors for protein binding pocket and ligand comparisons, Bioinformatics 21 (2005) 2347–2355.

[11] H. Stuhrmann, 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 (1970) 297–306.

[12] D. Svergun, Mathematical methods in small-angle scattering data analysis, J. Appl. Crystallogr. 24 (1991) 485–492.

[13] R.J. Morris, G. Bricogne, Sheldrick's 1.2 Årule and beyond, Acta Crystallogr. D 59 (2003) 615–617.

[14] R.J. Morris, E. Blanc, G. Bricogne, On the interpretation and use of $\langle |E|^2 \rangle (d^*)$profiles, Acta Crystallogr. D 60 (2004) 227–240.

[15] F. Zernike, Diffraction theory of the cut procedure and its improved form, the phase contrast method, Physica 1 (1934) 689–704.

[16] D. Zhang, G. Lu, Review of shape representation and description techniques, Pattern Recog. 37 (2004) 1–19.

[17] M.E. Celebi, Y.A. Aslandogan, A comparative study of three moment-based shape descriptors, in: Proceedings of the International Conference on Information Technology: Coding and Computing (ITCC'05), vol. 1, 2005, pp. 788–793.

[18] N. Canterakis, 3D Zernike moments and Zernike affine invariants for 3D image analysis and recognition, in: Proceedings of the 11th Scandinavian Conference on Image Analysis, 1999.

[19] M. Novotni, R. Klein, Shape retrieval using 3D Zernike descriptors, Comput. Aided Des. 36 (2004) 1047–1062.

[20] E.O. Steinborn, K. Ruedenberg, Molecular integrals between real and between complex spherical harmonics, in: P.-O. Löwdin (Ed.), Advances in Quantum Chemistry, vol. 7, Academic Press, New York, 1973, pp. 1–81.

[21] H.Y. Wang, R. LeSar, An efficient fast-multiple algorithm based on an expansion in the solid harmonics, J. Chem. Phys. 104 (11) (1996) 4173–4179.

[22] F. Pavelcik, J. Zelinka, Z. Otwinowski, Methodology and applications of automatic electron-density map interpretation by six-dimensional rotational and translational search for molecular fragments, Acta Crystallogr. D 58 (2002) 275–283.

[23] A. Kudlicki, M. Rowicka, M. Gilski, Z. Otwinowski, An efficient routine for computing symmetric real spherical harmonics for high orders of expansion, J. Appl. Crystallogr. 38 (2005) 501–504.

[24] S. Trapani, J. Navaza, Calculation of spherical harmonics and Wigner d functions by FFT. Applications to fast rotational matching in molecular replacement and implementation into AMoRe, Acta Crystallogr. A 62 (2006) 226–269.

[25] D.W. Ritchie, High-order analytic translation matrix elements for real-space six-dimensional polar Fourier correlations, J. Appl. Crystallogr. 38 (2005) 808–818.

[26] J.C. Wyant, K. Creath, Basic wavefront aberration theory for optical metrology, in: R. Shannon, J. Wyant (Eds.), Applied Optics and Optical Engineering, vol. XI, Academic Press, New York, 1992, pp. 28–39.

[27] C. Cohen-Tannoudji, B. Dui, F. Laloê, Quantum Mechanics, vols. 1 & 2, Wiley–Interscience, 1977 ISBN 0-471-16432-1 & 0-471-16434-8.

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

[29] M. Chaichian, R. Hagedorn, Symmetries in Quantum Mechanics: From Angular Momentum to Supersymmetry, IOP Institute of Physics, 1997 ISBN 0-750304073.

[30] M. Kazhdan, T. Funkhouser, S. Rusinkiewicz, Rotation invariant spherical harmonic representation of 3D shape descriptors, in: Kobbelt, Schröder, Hoppe (Eds.), Eurographics Symposium on Geometry Processing, 2003.

[31] T. Funkhouser, P. Min, M. Kazhdan, J. Chen, A. Halderman, D. Dobkin, D. Jacobs, A search engine for 3D models, ACM Trans. Graphics 22 (1) (2003) 1–28.

[32] R.J. Morris, An evaluation of spherical designs for molecular-like surfaces, J. Mol. Graphics Model. 24 (5) (2006) 356–361.

[33] W.L. DeLano, The PyMOL Molecular Graphics System, 2002, World Wide Web http://www.pymol.org.

[34] E.F. Pettersen, T.D. Goddard, C.C. Huang, G.S. Couch, D.M. Greenblatt, E.C. Meng, T.E. Ferrin, UCSF Chimera—a visualization system for exploratory research and analysis, J. Comput. Chem. 25 (13) (2004) 1605–1612.

[35] A. Kahraman, R.J. Morris, R.A. Laskowski, J.M. Thornton, Shape variation in protein binding pockets and their ligands, J. Mol. Biol. 368 (1) (2007) 283–301.

[36] P. Daras, D. Zarpalas, D. Tzovaras, M.G. Strintzis, 3D shape-based techniques for protien classification, in: IEEE International Conference on Image Processing, 2005, 1257–1260.