BIOINFORMATICS ORIGINAL PAPER Vol. 23 no. 23 2007, pages 3139–3146
doi:10.1093/bioinformatics/btm503

Structural bioinformatics

Moment invariants as shape recognition technique for comparing protein binding sites

Ingolf Sommer[1,*], Oliver Müller[1], Francisco S. Domingues[1], Oliver Sander[1], Joachim Weickert[2] and Thomas Lengauer[1]

[1]Max-Planck-Institute for Informatics, Stuhlsatzenhausweg 85, 66123 Saarbrücken and [2]Faculty of Mathematics and Computer Science, Saarland University, Building E1.1, 66041 Saarbrücken, Germany

Received on July 7, 2007; revised on September 13, 2007; accepted on October 2, 2007

Advance Access publication October 31, 2007

Associate Editor: Burkhard Rost

ABSTRACT

Motivation: An approach for identifying similarities of protein–protein binding sites is presented. The geometric shape of a binding site is described by computing a feature vector based on moment invariants. In order to search for similarities, feature vectors of binding sites are compared. Similar feature vectors indicate binding sites with similar shapes.

Results: The approach is validated on a representative set of protein–protein binding sites, extracted from the SCOPPI database. When querying binding sites from a representative set, we search for known similarities among 2819 binding sites. A median area under the ROC curve of 0.98 is observed. For half of the queries, a similar binding site is identified among the first two of 2819 when sorting all binding sites according the proposed similarity measure. Typical examples identified by this method are analyzed and discussed. The nitrogenase iron protein-like SCOP family is clustered hierarchically according to the proposed similarity measure as a case study.

Availability: Python code is available on request from the authors.

Contact: sommer@mpi-inf.mpg.de

Supplementary information: Supplementary data are available at *Bioinformatics* online.

1 INTRODUCTION

Geometric similarity plays an important role in structural biology, e.g. in the detection of the similarity of protein structures (Sierk and Kleywegt, 2004), or in the analysis of protein–protein binding interactions (Via *et al.*, 2000). To support the analysis of protein interactions, there are methods available for backbone structure comparison (Holm and Sander, 1993; Shindyalov and Bourne, 1998), for local spatial comparison of patterns in ligand binding sites (Artymiuk *et al.*, 2005; Kleywegt, 1999; Stark *et al.*, 2003), for comparing surfaces of ligand binding sites (Hofbauer *et al.*, 2004; Morris *et al.*, 2005), for searching ligand binding sites (Jambon *et al.*, 2005; Shulman-Peleg *et al.*, 2004b), for comparing protein binding sites (Hofbauer and Aszodi, 2005) or protein–protein interfaces (Bock *et al.*, 2005, 2007; Shulman-Peleg *et al.*, 2004a; 2005), and for analyzing protein–protein interactions (Cazals *et al.*, 2006). There are datasets of protein–protein interactions (Keskin *et al.*, 2004) and databases emerging for the classification of protein–protein interfaces, such as SCOPPI (Kim and Ison, 2005; Winter *et al.*, 2006) and SNAPPI (Jefferson *et al.*, 2007).

Recently, the structural analysis of protein–protein interactions has received a lot of attention, but still there is a need for methods assisting in the structural comparison of protein–protein binding sites. Here, we address the problem of efficiently identifying similarities in defined binding sites employing shape recognition techniques.

While the concept of shape is mathematically difficult to formulate (Edelsbrunner and Mücke, 1994), shape recognition techniques have been developed and employed very successfully. Among others, there are techniques based on extended Gaussian images (Horn, 1984), geometric hashing (Wolfson and Rigoutsos, 1997), spherical harmonics (Kazhdan *et al.*, 2003), shape distributions (Osada *et al.*, 2002), barcode shape descriptors (Collins *et al.*, 2004) and moment invariants (Hu, 1962) used for pattern recognition. Some of these, like geometric hashing (Shulman-Peleg *et al.*, 2005), spherical harmonics (Cai *et al.*, 2002; Morris *et al.*, 2005) and shape histograms (Ankerst *et al.*, 1999) have been successfully employed in structural bioinformatics. Also for identifying similarities between small-molecule ligands, shape comparison techniques have been successfully employed (Ballester and Richards, 2007; Grant *et al.*, 1996).

Moment invariants are well established in 2D image analysis (Hu, 1962), have been extended for 3D pattern only much later (Flusser *et al.*, 2003; Mamistvalov, 1998) and have not been applied to problems in structural bioinformatics yet.

We propose a novel, efficient approach for characterizing protein–protein binding sites such that similarities among them can be identified and protein binding sites can be classified structurally. The approach is based on a descriptor, capturing the essentials of the shape of each binding site in a vector of 3D moment invariants. This allows for comparing the shape independently from the sequence similarities between the binding sites. The approach is thus alignment-free, furthermore it does not need spatial superposition, and it is very runtime efficient.

*To whom correspondence should be addressed.

In the rest of the article, we will briefly review the theory of 3D moment invariants, describe how to represent parts of macromolecules such that moment invariants can be computed, and present experiments on a representative set of structurally known binding sites to demonstrate the power of the proposed descriptor.

2 METHODS

2.1 Geometric and central moments in 3D

The raw moments of order p of a 3D density function $\rho(\boldsymbol{x}) = \rho(x_1, x_2, x_3)$ are defined in terms of the integral

$$m_{p_1p_2p_3} = \int_{-\infty}^{\infty}\int_{-\infty}^{\infty}\int_{-\infty}^{\infty} x_1^{p_1} x_2^{p_2} x_3^{p_3}\,\rho(\boldsymbol{x})\,dx_1dx_2dx_3,$$

where $p = p_1 + p_2 + p_3$. The central moments are defined as

$$\mu_{p_1p_2p_3} = \int\int\int (x_1 - \bar{x}_1)^{p_1}(x_2 - \bar{x}_2)^{p_2}(x_3 - \bar{x}_3)^{p_3}\,\rho(\boldsymbol{x})\,dx_1dx_2dx_3,$$

where $\bar{x}_1 = m_{100}/m_{000}$, $\bar{x}_2 = m_{010}/m_{000}$ and $\bar{x}_3 = m_{001}/m_{000}$. Trivially, when the center of mass $(\bar{x}_1, \bar{x}_2, \bar{x}_3)$ is at the origin, the raw moments become the central moments.

2.2 Moment invariants

Moment invariants are rational functions of the moments that remain constant in value when the density ρ is subjected to transformation. For example, the following functions O_3, O_4, O_5, FL remain invariant, when ρ is subjected to orthogonal coordinate transformations (Flusser *et al.*, 2003; Mamistvalov, 1998); i.e. the values of O_3, O_4, O_5, FL remain constant when ρ is rotated in 3D space:

$$\begin{aligned}
O_3 &= (\mu_{200} + \mu_{020} + \mu_{002})/\mu_{000}\\
O_4 &= (\mu_{200}\mu_{020} - \mu_{110}^2 + \mu_{200}\mu_{002} - \mu_{101}^2 + \mu_{020}\mu_{002} - \mu_{011}^2)/\mu_{000}^2\\
O_5 &= (\mu_{200}\mu_{020}\mu_{002} + 2\mu_{110}\mu_{101}\mu_{011}\\
&\quad - \mu_{200}\mu_{011}^2 - \mu_{020}\mu_{101}^2 - \mu_{002}\mu_{110}^2)/\mu_{000}^3\\
FL &= (\mu_{003}^2 + 6\mu_{012}^2 + 6\mu_{021}^2 + \mu_{030}^2 + 6\mu_{102}^2 + 15\mu_{111}^2 - 3\mu_{102}\mu_{120}\\
&\quad + 6\mu_{120}^2 - 3\mu_{021}\mu_{201} + 6\mu_{201}^2 - 3\mu_{003}(\mu_{021} + \mu_{201})\\
&\quad - 3\mu_{030}\mu_{210} + 6\mu_{210}^2 - 3\mu_{012}(\mu_{030} + \mu_{210})\\
&\quad - 3\mu_{102}\mu_{300} - 3\mu_{120}\mu_{300} + \mu_{300}^2)/\mu_{000}^2
\end{aligned}$$

These moment invariants characterize the density of an object independently from the object's position or orientation. The particular functions are not invariant to scale. Since moments are continuous, the employed invariant functions of the moments are continuous as well. Slight changes in the density correspond to slight changes in the moment invariants. Similar density functions can be identified by identifying similar moment invariants. Thus, a feature vector $\boldsymbol{v} = (O_3, O_4, O_5, FL)$ of moment invariants can serve to describe densities independently from their position and orientation in 3D space.

Comparing different density functions now corresponds to comparing their feature vectors. In order to use a Euclidean metric, we normalize the feature vectors. Given N feature vectors, following the `scale` normalization routine in R (Ihaka and Gentleman, 1996), each feature is divided by its sample variance, i.e.

$$v'_{i,j} = v_{i,j}/\mathrm{rms}(v_{*,j})$$

where the features of vector $\boldsymbol{v}_i$ are indexed with j, and

$$\mathrm{rms}(v_{*,j}) = \sqrt{\frac{1}{N-1}\sum_{i=1}^{N} {v_{i,j}}^2}\,.$$

Other normalization schemes were suggested, e.g. (Mukundan and Ramakrishnan, 1998); in our experiments the specific kind of normalization affects the overall result only to a limited extent (data not shown).

Normalized feature vectors $\boldsymbol{v}'$ of density functions can be stored and a database or data structure of feature vectors can be queried for similarities. Search for similarities can be performed efficiently, e.g. by storing feature vectors in a space efficient k-d-tree or in a more runtime-efficient and less space-efficient range tree (Mehlhorn, 1984).

2.3 Describing molecules with moment invariants

For many molecules, coordinates of atoms are available from X-ray crystallography or NMR experiments. The coordinates of proteins, e.g. are deposited in the Protein Data Bank (PDB). With coordinates available, fitting a 3D Gaussian onto the center of each atom and summing over the Gaussians, the density of a molecule can be approximated (Grant *et al.*, 1996). The density can be interpreted as the molecule's shape. A feature vector characterizing the density thus serves to describe the molecule's shape.

When all atoms are employed in the above summation of Gaussians, the complete molecule's shape is characterized. Alternatively, the summation can be performed for only selected parts of the molecule. For example, the shape of the surface of a protein molecule can be approximated by summing over its surface atoms. Similarly, by selecting atoms comprising the binding site of a protein, the shape of that binding site is described.

2.4 Computing 3D moments of sums of Gaussians

With a molecule's density represented as sum of Gaussians, one can easily compute its moments and moment invariants. For a 3D Gaussian g_k, with standard deviation σ, centered at position $\boldsymbol{a}_k$

$$g_k(\boldsymbol{x}) = \frac{1}{\sigma^3\sqrt{(2\pi)^3}}\; e^{-\sum_{j=1}^{3}\left(\frac{x_j - a_{k,j}}{2\sigma}\right)^2},$$

the first few raw moments are as follows (Rinne, 2003, for example):

$$\begin{aligned}
m_{k,100} &= a_{k,1}, \quad m_{k,200} = a_{k,1}^2 + \sigma^2, \quad m_{k,300} = a_{k,1}^3 + 3a_{k,1}\sigma^2, \ldots\\
m_{k,010} &= a_{k,2}, \quad m_{k,020} = a_{k,2}^2 + \sigma^2, \quad m_{k,030} = a_{k,2}^3 + 3a_{k,2}\sigma^2, \ldots\\
m_{k,001} &= a_{k,3}, \quad m_{k,002} = a_{k,3}^2 + \sigma^2, \quad m_{k,003} = a_{k,3}^3 + 3a_{k,3}\sigma^2, \ldots
\end{aligned}$$

And since the distributions are independent along the coordinate axes:

$$m_{k,110} = a_{k,1}a_{k,2}, \quad m_{k,120} = a_{k,1}(a_{k,2}^2 + \sigma^2), \ldots$$

The raw moments of a density function ρ that is a sum of Gaussians $\rho = \sum_k g_k$, are readily computed as the sum of the moments of the individual Gaussians. In order to compute the central moments of the function ρ, its center of mass can be shifted to the origin such that:

$$\mu_{100} = \sum_k (a_{k,1} - \bar{x}_1), \quad \mu_{200} = \sum_k ((a_{k,1} - \bar{x}_1)^2 + \sigma^2), \ldots$$

and

$$\begin{aligned}
\mu_{110} &= \sum_k (a_{k,1} - \bar{x}_1)\,(a_{k,2} - \bar{x}_2),\\
\mu_{120} &= \sum_k (a_{k,1} - \bar{x}_1)\,((a_{k,2} - \bar{x}_2)^2 + \sigma^2), \ldots
\end{aligned}$$

These values are used to compute the moment invariants O_3, O_4, O_5, FL introduced in Section 2.2.

2.5 Experiments

We assess the moment invariant method on a representative set of protein–protein binding sites. We test how well the method identifies similar binding sites among a large representative set of protein binding sites.

2.5.1 Definition of the binding sites For the experiments, we use a set of protein–protein binding sites from the SCOPPI database (Kim and Ison, 2005; Kim *et al.*, 2006; Winter *et al.*, 2006). The database provides an evolutionary and structural classification of protein–protein interactions. Based on SCOP (Murzin *et al.*, 1995), SCOPPI superposes the domains of each family and classifies inter-domain interactions. One classification criterion, in particular, is the relative position of a binding site with respect to the structurally superposed domains of the family. Distinct positions are classified as distinct face types. The SCOPPI face type thus defines the position of the binding site with respect to the protein domain. Two binding sites with the same face type can be geometrically similar, but this is not always the case. SCOPPI classifies each binding site uniquely into one group defined by its family and face type.

SCOPPI provides removal of redundancy according to percent sequence identity of the protein domains. Here, we use a subset of SCOPPI whose domains have pairwise $< 80\%$ sequence identity and use only those SCOP families with at least 15 members. For the SCOP 1.69 database, this leaves us with 2819 binding sites from 96 SCOP families, falling into 501 groups of family and face type. We will refer to this set as PBSALL (for all protein binding sites).

For each binding site, the residues involved in the interaction are computed according to a criterion defined in Jones and Thornton (1995). Namely, binding site residues are defined as those residues whose side chains have an accessible surface area that decreases by more than 1Å^2 on dimerization. A visual catalog of the binding sites and binding site residues employed is provided in the Supplementary Material.

2.5.2 Guaranteeing geometric similarity within binding sites While binding sites within the same group of family and face type bind at similar positions, some display large differences (see Supplementary Material for some examples). Starting from the set PBSALL, a set of binding sites that falls into groups that are known to be geometrically similar is constructed by employing two restriction criteria.

(1) Binding sites that differ substantially in the number of residues involved in the binding cannot be geometrically similar. Consequently, groups for which the minimal and maximal number of binding site residues that occur within a group differ by at least 50 % are discarded.

(2) Furthermore, we confirm structural relatedness of the binding sites using the TM-align program (Zhang and Skolnick, 2005). TM-align is intended to compare predicted protein structure models against native structures. For the comparison, TM-align relies on the sequence order of the models to be compared; it still can align non-contiguous residues in sites. Independently from the sequence length, it scores model versus native structure on a scale of 0.0 (unrelated) to 1.0 (highly similar). Here, TM-align is employed for discarding binding sites, which are geometrically not similar. For evolutionary-related binding sites, the structure of binding site residues, corresponding to a subset of the atoms defined in the PDB files, can be directly compared with TM-align. Comparing only binding sites from the same group of family and face type, TM-align score values range from 0.05 to 0.8, mostly. The values for arbitrary unrelated binding sites range from 0.0 to 0.45, mostly (see Supplementary Material for details). Thus, all groups which contain a binding site that, compared to another binding site within the same group, displays a TM-align score of less than or equal to 0.45 are discarded.

Applying these two criteria leaves only groups of geometrically similar binding sites. We call such a group a SIMGROUP. Selecting SIMGROUPs containing at least three binding sites one obtains 53 SIMGROUPs from 32 SCOP families, containing a total of 224 binding sites. We will refer to this set of binding sites as PBSGEOM (for protein binding sites, restricted to geometrical similarity).

2.5.3 Precomputation of moment invariants The residues involved in the interaction of each binding site are computed as described above. For each of these residues, all atoms are modeled as Gaussian distributions. As an approximation, we choose the standard deviation $\sigma = 0.523$ such that for each Gaussian 99% of the density is within the Van-der-Waals radius 1.76 of a Carbon atom [VDW radius as in (Chothia, 1975), for the derivation of σ, see Supplementary Material]. Varying σ to some extent ($0.1 \leq \sigma \leq 1$; data not shown) does not have a large impact on the results.

For each binding site, from the sum of Gaussians the moment invariants are computed as described in Section 2.4. For the 2819 binding sites in PBSALL, this yields a 2819×4 dimensional matrix. This matrix is normalized columnwise as described in Section 2.2.

2.5.4 Experimental protocol We test the method on all binding sites b from PBSGEOM. By construction of PBSGEOM, each b belongs to a SIMGROUP of geometrically similar binding sites and we measure how well we can distinguish these members from the other unrelated binding sites. For b, based on the moment invariants we compute the similarity to each binding site in PBSALL and sort those decreasingly. In a perfect scenario, the binding sites from b's SIMGROUP would be on the first few ranks with all other binding sites following.

When querying for an individual b, and searching for members from its SIMGROUP, we report the best hit that corresponds to the top rank of one of the SIMGROUP members. Furthermore, we analyze the number of interspersed false positives when identifying all SIMGROUP members. This number is computed as maximum rank of a SIMGROUP member minus the size of the SIMGROUP. In a perfect scenario the best hit would be on rank one, the number of interspersed false positives would be zero.

As an additional summarizing quality measure we use the AUC-value, the 'area under the curve' in the ROC plot (Fawcett, 2006; Sing *et al.*, 2005). The AUC-value summarizes the distribution of the ranks of the other binding sites from b's SIMGROUP for one query. An AUC-value of 1.0 corresponds to perfect classification, whereas for random assignments an AUC-value of 0.5 can be expected.

3 RESULTS AND DISCUSSION

3.1 All against all

As described in the previous Section 2.5.4, for each binding site from PBSGEOM we report the ranks of its SIMGROUP members when searching within PBSALL. For 224 queries we report best hit, interspersed false positives and AUC-value. Tables 1 and 2 provide summary statistics (minima, 25%, 50%, 75% quantiles and maxima) of these values, Figure 1 provides histograms for the best hit ranks and AUC values; the Supplementary Material provides the associated P-values.

3.1.1 Identifying any similar binding site The median of the best hit ranks is 2, thus for at least half of the binding sites we find a geometrically similar binding site among the first two (0.1 %) of 2819 inspected binding sites. The maximum of the best hit ranks is 1902 (67.5%), thus for this 'worst case'

instance, geometrically similar binding sites are not detected. For three quarters of the query binding sites, we detect a geometrically similar one among the first 19 (0.7%) rank positions. For comparison, we also used the difference in the number of binding site residues for measuring similarity of binding sites (histogram provided in the Supplementary Material). This simple measure yields a median of the best hit ranks of 52; it is significantly outperformed by the moment invariant method (with a median of 2).

3.1.2 Identifying all similar binding sites Similarly, the median of the interspersed false positives is 149 (5.3%), the 75% quantile of the number of interspersed is 432.3 (15.3%). For 25% of the queries, we identify all SIMGROUP members with less than 27 (1%) false positives. For 5% of the queries, we identify all SIMGROUP members without any false positives.

Table 1. Summary statistics for the best hit ranks and interspersed false positives for the search results of 224 queries

	Min	25% q	Median	75% q	Max
Best hit rank	1	1	2	18.3	1902
Interspersed FP	0	26.8	149	432.3	2463

Table 2. Summary statistics for the AUC values for the search results of 224 queries

	Max	75% q	Median	25% q	Min
AUC	1	0.9949	0.9801	0.9402	0.3004

In summary it is fair to say, that we find similar binding sites for most queries, but for many we do not identify all related binding sites.

3.2 Inspecting individual cases

3.2.1 Nomenclature In order to discuss individual binding sites (as defined in Section 2.5.1) we will refer them by the PDB identifier of the structure, the SCOP domain on which the binding site resides and the SCOP domain with which it is interacting. Using the sunids from SCOP, we concatenate this information to obtain an identifier that is unique for each binding site: <PDB – ID>_<SCOP-domain of binding site>_<Chain – ID of binding site>_<SCOP-domain of binding partner>_<Chain – ID of binding partner>.

3.2.2 One case that went wrong We face 15 out of 224 cases for which the ranks of the best hits are above 200. Here, we analyze the worst case. When querying for 1bgy_75810_O_43691_Q, we identify the members of its group 1bcc_75804_C_43699_E, and 1kyo_75907_N_73272_P on ranks 1902 and 2041, respectively. Instead, we falsely identify 1d3a_30144_D_42110_C on rank 1. The situation is depicted in Figure 2A. The binding sites within the similarity group of 1bgy_75810_O_43691_Q have disconnected residues. For 1bgy_75810_O_43691_Q, the disconnected residues are shifted quite a bit with respect to the two other structures. This affects the center of the mass (yellow spheres in the graphics), which is used as reference point for computing the moment invariants. Therefore, the feature vectors are not similar anymore. Furthermore, this example clearly points out that the method is unable to identify partial matches between structures. For other binding sites, we observe a tendency that they are not identified when larger appendices are present in one but not in the other structure.

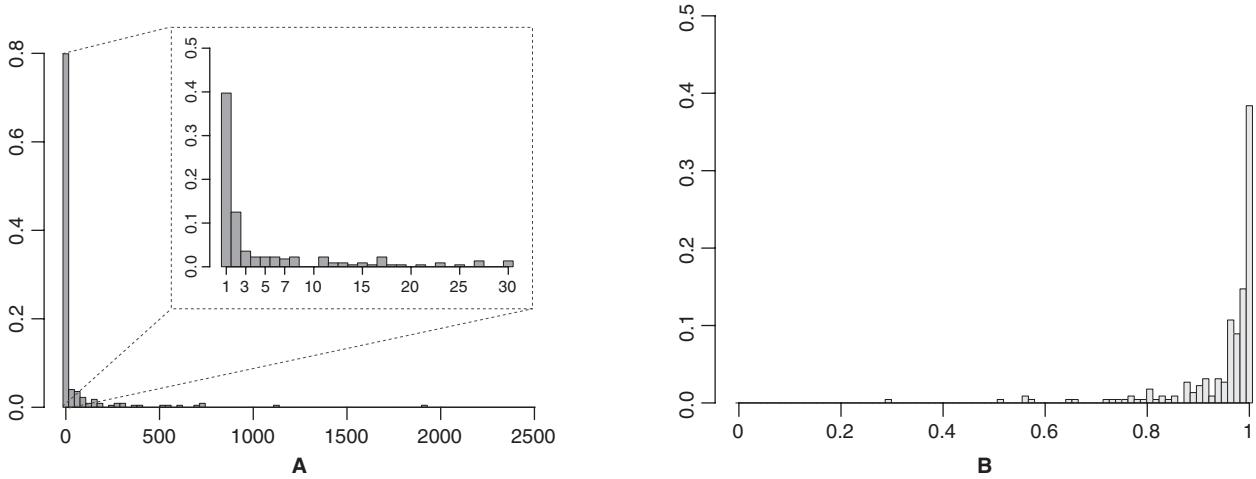

Fig. 1. (**A**) Histogram of the best hit ranks (i.e. the closest SIMGROUP members identified, c.f. Section 3.1). The possible scale of ranks is 1–2818 with an observed maximum of 1902 (x axis); for the histogram we use a bin width of 30. The figure summarizes 224 queries; the relative frequency of 0.8 corresponds to 179 cases (y axis). The insert details the observed counts within the first bin; 40%, i.e. 89 of the cases identify their closest SIMGROUP member on rank one, which is optimal. (**B**) Histogram of the AUC values observed for 224 queries. The x axis is split into bins of width 0.012. The relative frequency of observations is normalized with 224 (y axis).

Fig. 2. Visualization of three sample cases (**A, B** and **C**). In each case we have a query binding site (leftmost box) and geometric similarities as identified by the moment invariants (adjacent boxes from left to right). Each box contains a front view (above) and side view (below) of the protein domain (in gray) and of the binding sites atoms (red spheres). Moment invariants were computed for the binding sites atoms (red spheres) around the center of mass (yellow spheres, depicted in A only). For each query, we have members of the SIMGROUP in green boxes. (A) One case that went badly wrong (see Section 3.2.2). With query 1bgy_75810_O_43691_Q, we identify 1d3a_30144_D_42110_C (orange box) as geometrically similar instead of the two other binding sites from the SIMGROUP 1bcc_75804_C_43699_E, 1kyo_75907_N_73272_P. (B) With query 1hbm_60890_B_60888_A, we identify the SIMGROUP members 1e6y_18536_E_18530_D and 1e6v_18534_E_18528_D on rank one and three, respectively; binding site 1pv8_95155_E_95156_F is identified on rank two. (C) Query 1eys_43516_M_43515_L identifies both its SIMGROUP members 3prc_43438_M_43437_L and 1pcr_43495_M_43494_L on the first two ranks.

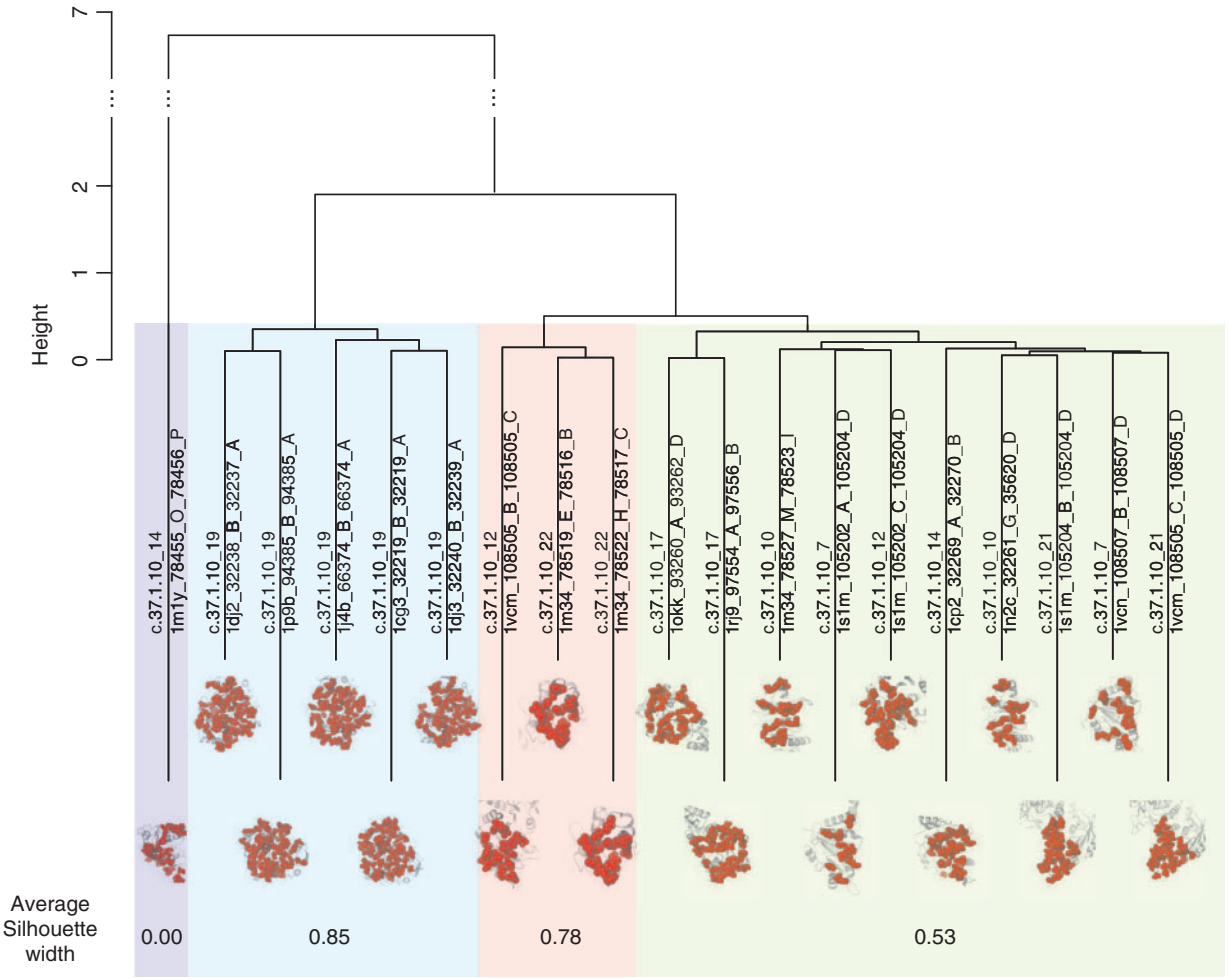

Fig. 3. The binding sites of SCOP family (c.37.1.10) of nitrogenase iron protein-like proteins are clustered according to moment invariant geometric similarity. All binding sites (within PBSALL) of that family are clustered irrespective of their face type. The height refers to distances when clustering hierarchically using average linkage. The average silhouette width is computed for the four clusters (shown in pastel colors), which result from cutting the hierarchical tree. The protein structures are oriented such that the binding sites face the camera, they have not been structurally superposed.

3.2.3 Cases that work well When querying for binding site 1hbm_60890_B_60888_A (see Fig. 2B), we correctly identify its SIMGROUP members 1e6y_18536_E_18530_D_48911 and 1e6v_18534_E_18528_D on ranks one and three, respectively; 1pv8_95155_E_95156_F that is no member of the SIMGROUP is incorrectly identified on rank two.

Figure 2C depicts query 1eys_43516_M_43515_L, which identifies its two SIMGROUP members 3prc_43438_M_43437_L and 1pcr_43495_M_43494_L on the first two ranks.

3.3 Clustering a family of binding sites

Figure 3 shows an example for applying clustering with moment invariant similarities to binding sites within a SCOP family. The SCOP family c.37.1.10 of nitrogenase iron protein-like proteins are clustered according to similarities of their binding sites. All members (within PBSALL) of that family, irrespective of their face type, are clustered hierarchically using average linkage. Cutting the hierarchical tree, four distinct clusters are identified in the example and the average silhouette values are computed for them. Generally, silhouette values lie in the range $[-1, 1]$. Entries with a silhouette value close to 1.0 are well clustered in the sense that distances to other entries in the same cluster are small compared to distances to entries in the closest different cluster. We observe two clusters with average silhouette values of 0.85 and 0.78, one cluster joins a variety of the binding sites with a silhouette value of 0.53, and one binding site is clearly marked as a geometric outlier with a silhouette value of 0.0. Low silhouette values correspond to inhomogenity regarding the external SCOPPI labeling, high silhouette values are found in clusters that reproduce the face types well.

This example serves to visualize the method's capabilities as well as to illustrate a concrete application. Comparing all members within one family, the SCOPPI face types are reproduced to a certain extent. Another application (data not shown) would be, to cluster according to geometrical similarities within the SCOPI face types.

4 CONCLUSIONS

Moment invariants are presented here as shape recognition technique for classifying protein–protein binding sites. Compared to other shape recognition techniques, moment invariants have some drawbacks and some advantages. Like some other shape recognition techniques, e.g. spherical harmonic-based concepts, the computation heavily depends on a reference point. The moment invariant feature vector continuously varies when affinely transforming the density with respect to the reference point. With their global, coarse-grained, view on density, moment invariant methods complement the geometric hashing methods popular in bioinformatics, which successfully detect similarities of features conserved in detail.

Computing moment invariants is extremely fast; in preprocessing, parsing the PDB files is the limiting factor for the descriptor construction, and with given descriptors a binding site can be compared to several thousand others within a second on a standard PC.

The current technique affords a global comparison of densities. For comparisons on a finer level, detail can be added in various ways: geometrically, the density can be partitioned—e.g. into a number of concentric shells within a sphere—to yield a larger and more specific feature vector. Similarly, chemical types of atoms can be used to partition the density, resulting in a more specific feature vector.

The problem of detecting similarities in protein–protein binding sites is highly relevant; there are no established 'gold standards' yet to which new methods can be compared to. Thus we assess the method proposed here, based on the SCOPPI database of protein–protein interfaces. From this database, groups of binding sites are extracted and the geometric similarity within the groups is verified using the sequence-based TM-align tool for structure superposition. While the verified similarity groups are used to demonstrate the performance of the method proposed here, our method does not rely on sequence alignment and can detect geometric similarity when there is no sequence similarity.

The method currently compares complete binding sites, but does not find partial matches. The main application will be to group predefined binding sites according to geometrical similarities. In our view, complete matching is also a necessary step in the direction of partial matching, as partial matching can be handled by subdivision of the objects to be compared. We are in the process of refining the method in order to also handle partial matches and to be less dependent on reference points.

In conclusion, moment invariants are a fast and robust technique for comparing densities on a global level, which we believe to be useful also for other applications in structural bioinformatics.

ACKNOWLEDGEMENTS

We thank Adrian Alexa, Holger Theisel, Christof Winter and Hongbo Zhu for helpful comments on the manuscript. We thank Christof Winter for providing SCOPPI data. This work forms part of the BioSapiens project, which is funded by the European Commission within its FP6 Programme under the thematic area 'Life sciences, genomics and biotechnology for health', contract number LSHG-CT-2003-503265.

Conflict of Interest: none declared.

REFERENCES

Ankerst,M. *et al*. (1999) Nearest neighbor classification in 3D protein data bases. In *Proceedings 7th International Conference on Intelligent Systems for Molecular Biology*. pp. 34–43.

Artymiuk,P.J. *et al*. (2005) Graph theoretic methods for the analysis of structural relationships in biological macromolecules. *J. Am. Soc. Inf. Sci. Technol.*, **56**, 518–528.

Ballester,P.J. and Richards,W.G. (2007) Ultrafast shape recognition to search compound databases for similar molecular shapes. *J. Comput. Chem.*, **28**, 1711–1723.

Bock,M.E. *et al*. (2005) Identifying similar surface patches on proteins using a spin-image surface representation. In *Combinatorial Pattern Matching: 16th Annual Symposium, CPM 2005, Lecture Notes in Computer Science 3537*. Springer, Berlin, Heidelberg, pp. 417–428.

Bock,M.E. *et al*. (2007) Discovery of similar regions on protein surfaces. *J. Comput. Biol.*, **14**, 285–299.

Cai,W. *et al*. (2002) Protein-ligand recognition using spherical harmonic molecular surfaces: towards a fast and efficient filter for large virtual throughput screening. *J. Mol. Graph. Model.*, **20**, 313–328.

Cazals,F. *et al*. (2006) Revisiting the voronoi description of protein–protein interfaces. *Protein Sci.*, **15**, 2082–2092.

Chothia,C. (1975) Structural invariants in protein folding. *Nature*, **254**, 304–308.

Collins,A. *et al*. (2004) A barcode shape descriptor for curve point cloud data. In Alexa,M. and Rusinkiewicz,S. (eds.) *Eurographics Symposium on Point-Based Graphics* http://graphics.ethz.ch/events/pbg/.

Edelsbrunner,H. and Mücke,E.P. (1994) Three-dimensional alpha shapes. *ACM Trans. Graph.*, **13**, 43–72.

Fawcett,T. (2006) An introduction to ROC analysis. *Pattern Recognit. Lett.*, **27**, 861–874.

Flusser,J. *et al*. (2003) Moment forms invariant to rotation and blur in arbitrary number of dimensions. *IEEE Trans. Pattern Anal. Mach. Intell.*, **25**, 234–245.

Grant,J.A. *et al*. (1996) A fast method of molecular shape comparison: a simple application of a gaussian description of molecular shape. *J. Comput. Chem.*, **17**, 1653–1666.

Hofbauer,C. and Aszodi,A. (2005) Sh2 binding site comparison: a new application of the surfcomp method. *J. Chem. Inf. Model.*, **45**, 414–421.

Hofbauer,C. *et al*. (2004) Surfcomp: a novel graph-based approach to molecular surface comparison. *J. Chem. Inf. Comput. Sci*, **44**, 837–847.

Holm,L. and Sander,C. (1993) Protein structure comparison by alignment of distance matrices. *J. Mol. Biol.*, **233**, 123–138.

Horn,B. (1984) Extended Gaussian images. *Proc. IEEE.*, **72**, 1671–1686.

Hu,M.-K. (1962) Visual pattern recognition by moment invatiants. *IRE. Tran. Inf. Theory*, **8**, 179–187.

Ihaka,R. and Gentleman,R. (1996) R: a language for data analysis and graphics. *J. Comput. Graph. Stat.*, **5**, 299–314.

Jambon,M. *et al*. (2005) The SuMo server: 3D search for protein functional sites. *Bioinformatics*, **21**, 3929–3930.

Jefferson,E.R. *et al*. (2007) SNAPPI-DB: a database and api of structures, interfaces and alignments for protein–protein interactions. *Nucleic Acids Res.*, **35**, D580–D589.

Jones,S. and Thornton,J.M. (1995) Protein-protein interactions: a review of protein dimer structures. *Prog. Biophys. Mol. Biol.*, **63**, 31–65.

Kazhdan,M. *et al*. (2003) Rotation invariant spherical harmonic representation of 3D shape descriptors. In Kobbelt,L., Schröder,P. and Hoppe,H. (eds.) Eurographics Symposium on Geometry Processing, http://www-i8.informatik.rwth-aachen.de/old-site/SGP/sgp03/.

Keskin,O. *et al*. (2004) A new, structurally nonredundant, diverse data set of protein–protein interfaces and its implications. *Protein Sci.*, **13**, 1043–1055.

Kim,W.K. and Ison,J.C. (2005) Survey of the geometric association of domain-domain interfaces. *Proteins*, **61**, 1075–1088.

Kim,W.K. *et al*. (2006) The many faces of protein–protein interactions: a compendium of interface geometry. *PLoS. Comput. Biol.*, **2**, 1151–1164.

Kleywegt,G. (1999) Recognition of spatial motifs in protein structures. *J. Mol. Biol.*, **285**, 1887–1997.

Mamistvalov,A. (1998) n-dimensional moment invariants and conceptual mathematical theory of recognition of n-dimensional data sets. *IEEE. Trans. Pattern Anal. Mach. Intell.*, **20**, 819–831.

Mehlhorn,K. (1984) *Data Structures and Efficient Algorithms. Volume 3: Multi-dimensional Searching and Computational Geometry,* volume 3 of *EATCS Monographs on Theoretical Computer Science*. Springer, Berlin, Germany.

Morris,R.J. *et al.* (2005) Real spherical harmonic expansion coefficients as 3D shape descriptors for protein binding pocket and ligand comparisons. *Bioinformatics*, **21**, 2347–2355.

Mukundan,R. and Ramakrishnan,K. (1998) *Moment Functions in Image Analysis.* World Scientific Publishing, Singapore, New Jersey, London, Hong Kong.

Murzin,A.G. *et al.* (1995) SCOP: a structural classification of proteins database for the investigation of sequences and structures. *J. Mol. Biol.*, **247**, 536–540.

Osada,R., Funkhouser,T., Chazelle,B. and Dobkin,D. (2002) Shape distributions. *ACM Trans. Graph*, **21**, 807–832.

Rinne,H. (2003) *Taschenbuch der Statistik*. Harri Deutsch, Frankfurt a.M., Germany.

Shindyalov,I.N. and Bourne,P.E. (1998) Protein structure alignment by incremental combinatorial extension (CE) of the optimal path. *Protein Eng.*, **11**, 739–747.

Shulman-Peleg,A. *et al.* (2004a) Protein-protein interfaces: recognition of similar spatial and chemical organizations. In Jonassen,I. and Kim,J. (eds.) *Algorithms in Bioinformatics: 4th International Workshop, WABI 2004, Bergen, Norway, 2004,* Volume 3240 of *Lecture Notes in Computer Science*. Springer Verlag, Berlin, Heidelberg, pp. 194–205.

Shulman-Peleg,A. *et al.* (2004b) Recognition of functional sites in protein structures. *J. Mol. Biol.*, **339**, 607–633.

Shulman-Peleg,A. *et al.* (2005) SiteEngines: recognition and comparison of binding sites and protein–protein interfaces. *Nucleic Acids Res.*, **33** (Web Server issue), W337–W341.

Sierk,M.L. and Kleywegt,G.J. (2004) Déjá vu all over again: finding and analyzing protein structure similarities. *Structure*, **12** (12), 2103–2111.

Sing,T. *et al.* (2005) ROCR: visualizing classifier performance in R. *Bioinformatics*, **21** (20), 3940–3941.

Stark,A. *et al.* (2003) A model for statistical significance of local similarities in structure. *J. Mol. Biol.*, **326**, 1307–1316.

Via,A. *et al.* (2000) Protein surface similarities: a survey of methods to describe and compare protein surfaces. *Cell. Mol. Life Sci.*, **57**, 1970–1977.

Winter,C. *et al.* (2006) Scoppi: a structural classification of protein–protein interfaces. *Nucleic Acids Res.*, **34**, D310–D314.

Wolfson,H.J. and Rigoutsos,I. (1997) Geometric hashing: an overview. *IEEE Comput. Sci. Eng.*, **97**, 10–21.

Zhang,Y. and Skolnick,J. (2005) The protein structure prediction problem could be solved using the current PDB library. *Proc. Natl. Acad. Sci. USA*, **102**, 1029–1034.