Parameterization and Classification of the Protein Universe via Geometric Techniques

Ashish V. Tendulkar¹, Pramod P. Wangikar²*, Milind A. Sohoni³, Vivekanand V. Samant² and Chetan Y. Mone²

¹Kanwal Rekhi School of Information Technology, Indian Institute of Technology Bombay, Powai, Mumbai 400 076, India
²Department of Chemical Engineering, Indian Institute of Technology, Bombay, Powai Mumbai 400 076, India
³Department of Computer Science and Engineering, Indian Institute of Technology Bombay, Powai, Mumbai 400 076, India

We present a scheme for the classification of 3487 non-redundant protein structures into 1207 non-hierarchical clusters by using recurring structural patterns of three to six amino acids as keys of classification. This results in several signature patterns, which seem to decide membership of a protein in a functional category. The patterns provide clues to the key residues involved in functional sites as well as in protein-protein interaction. The discovered patterns include a “glutamate double bridge” of superoxide dismutase, the functional interface of the serine protease and inhibitor, interface of homo/hetero dimers, and functional sites of several enzyme families. We use geometric invariants to decide superimposability of structural patterns. This allows the parameterization of patterns and discovery of recurring patterns via clustering. The geometric invariant-based approach eliminates the computationally explosive step of pair-wise comparison of structures. The results provide a vast resource for the biologists for experimental validation of the proposed functional sites, and for the design of synthetic enzymes, inhibitors and drugs.

© 2003 Elsevier Ltd. All rights reserved.

Keywords: geometric invariants; clustering; protein structure comparison; functional site; protein-protein interface

Introduction

Unraveling of the evolutionary relationships between proteins has been of central interest to biologists for decades. This is achieved via either primary sequence alignment¹,² or overall structural alignment.³,⁴ The conserved amino acids in the sequence alignment led to construction of family-wise sequence signatures, which are now well documented.⁵,⁶ The conserved amino acids are deemed to be important for function.⁷ On the structural side, several methods are known for optimal pair-wise alignment of protein structures. Algorithms such as Protein Structure Alignment by Distance Matrices⁸ and Vector Alignment Search Tool⁹ are widely used, each optimizing a different measure of similarity. An extensive all-against-all structure comparison has led to hierarchical classification systems such as SCOP¹⁰ and CATH.¹¹ These are extremely useful in understanding structural, evolutionary and functional relationship in proteins of known structure. Much of the structure analysis hinges on the hypothesis that nearly all proteins have structural similarities and, in many cases share a common evolutionary origin.¹²,¹³ Furthermore, substructures of small number of amino acids are known to be conserved across several proteins. Methods have been developed to search for user-defined conformations, which are typically useful in searching for known active sites¹²,¹⁵ or peptide segment conformations.¹⁴ We have described an unbiased graph theoretic approach for detection of recurring side-chain patterns from protein families.¹⁶ It is now well accepted that a family of proteins with common function typically conserves a functional site made up of a small number of amino acids, which is not necessarily detectable from the sequence signatures.¹⁵,¹⁶ Classic examples include the Ser/His/Asp catalytic triad of serine proteases.¹⁷ Thus, our first objective was to establish

Supplementary data associated with this article can be found at doi: 10.1016/j.jmb.2003.09.021

Abbreviation used: PCA, principal component analysis.

E-mail address of the corresponding author: pramodw@iitb.ac.in

0022-2836/$ - see front matter © 2003 Elsevier Ltd. All rights reserved.
relationships between proteins based on recurring structural patterns, which could potentially be functional site patterns.

Many proteins naturally occur as complexes of multiple chains as homo/hetero-dimers, trimers, etc. Interactions between the amino acid residues at the protein–protein interface stabilize the complex. The understanding of the key residues involved in protein–protein interaction is of interest to biologists. Thus, in addition to the functional sites as mentioned above, we probe into the conserved structural patterns at the protein–protein interface.

In this paper, we report an approach to the detection of recurring structural patterns via geometric invariants. These invariants help in deciding the super-imposability of candidate patterns without the computationally expensive step of actually constructing and verifying the superimposing transformation. A geometric invariant is a quantity, which is unchanged under a group of geometric transformations, in this case, the group of translations and rotations in three-dimensional space. The simplest invariant is content, the labels of amino acids in the pattern. Examples of continuous invariants associated with a geometrical structure of amino acids, or pattern, are signed volumes, areas, lengths, etc. For our group of transformations, it has been shown that invariants suffice to decide superimposability of two structures. Thus, in effect, if two patterns $K_1$ and $K_2$ are not superimposable then there is an invariant $f$ such that $f(K_1)$ does not equal $f(K_2)$.

Using this as the starting point, we coordinate a pattern by its evaluations on a fixed suite of $N$ invariants (Figure 1). Using earlier work, we examine patterns from 3487 non-redundant protein structures. Each such possible pattern is mapped into the $N$-dimensional space as a vector. A single protein may thus generate thousands of pattern vectors of hundreds of contents. This procedure is carried out for all the proteins, one at a time, to obtain a collection of pattern-vectors and their contents. This collection may now be subjected to the usual tools of data-mining and extraction. A cluster is a small region in this $N$-space, which has a large number of pattern-vectors, and thus is visited by a large number of proteins. Such a cluster corresponds to a recurring pattern.

Since the method eliminates computationally explosive pair-wise comparisons, large datasets of protein structures may now be analyzed and classified. When this classification scheme is compared with the well accepted SCOP classification, we find that several of our clusters are visited by members of a single SCOP superfamily with the detected recurring pattern acting as a signature of that SCOP superfamily and not present in any other superfamily. In addition, we find several clusters visited by proteins from two to three SCOP superfamilies. This points to novel relationships between superfamilies and potentially common functional sites/mechanisms of action across superfamilies. We discuss some example patterns along with the reported biological implications of the pattern forming amino acid residues and indeed find many of these to be involved in functional/mechanistic role.

**Algorithm**

In this section, we describe our approach to pattern classification via geometric invariants (Figure 1). Our input dataset consists of non-redundant protein structures from the PDB database. No two entries of our dataset have identical domain as well as species as per SCOP database.

![Figure 1](image-url) The overall algorithm for classification of the protein universe via geometric techniques.
A protein, for us, is a sequence of colored points in three-dimensional space, the colors coming from the amino acid labels and the co-ordinates as locations. From each protein, all possible patterns of 3–6 amino acids are enumerated as described before. For each pattern, three geometric objects are represented by constructing three polyhedra, made up of the C-alpha atoms, C-beta atoms and functional atoms, separately. One functional atom or a pseudo atom from the side-chain is considered for each amino acid. We refer to any of these atom functional sites of proteins. At first, we describe these amino acids are unlikely to be part of hydrophobic amino acids, Gly, Ala, Val, Leu, Ile, Phe and Pro, are omitted from our analysis as Phe and Pro, are omitted from our analysis as the amino acid labels and the co-ordinates as locations. From each protein, all possible patterns of all 5 × 5-matrices of the type shown below:

\[
P' = \begin{bmatrix} P & 0 & 0 \\ 0 & 1 & 0 \\ 0 & 0 & 1 \end{bmatrix}
\]  

(3)

where \( P \) is a 3 × 3-permutation matrix such that there is exactly one 1 in each row and column. Thus the effect of multiplying \( K' \) by \( P' \) has the effect of re-ordering the \( \alpha \)-amino acids amongst themselves. In general, the perturbations act on the right and re-arrange amino acids of the same label. Thus the general transformation on a pattern \( K' \) is given by two matrices \( X', P' \) acting on \( K' \) to give

\[
L' = X'K'P'.
\]  

(4)

Given another pattern \( M' \) with the same content as \( K', M' \) is super-imposable with \( K' \) if there are matrices \( X', P' \) such that \( M' = X'K'P' \). We denote this as \( M \sim \sim K \). Note that \( \sim \sim \) is an equivalence relation. This means that (i) \( K \sim K \), (ii) if \( K \sim L \) then \( L \sim K \) and finally (iii) if \( K \sim L \) and \( L \sim M \) then \( K \sim M \).

### Invariant theory

Given the space \( C \) of all patterns of a fixed content, and the transformations as discussed above, a real valued function \( f : C \rightarrow \mathbb{R} \) is called an invariant if \( K \sim L \) implies that \( f(L) = f(M) \). Thus \( f \) is a function which is left invariant under the transformations on the pattern. The most obvious invariant is content, i.e. the number of D, E and N amino acids. A more non-trivial invariant is the volume of the convex hull of the 5 atoms in \( \mathbb{R}^3 \). Clearly, if \( K \) and \( L \) are super-imposable, then \( \text{vol}(K) = \text{vol}(L) \). Thus invariants may be used as quick tests to judge if \( M \sim L \): simply compare \( f(L) \) with \( f(M) \), where \( f \) is an invariant. This motivates two important questions: (i) Are the invariants robust to small variations? For example, if there is a small perturbation of the pattern \( M \) to say \( \overline{M} \), then is \( |f(M) - f(\overline{M})| \) small? (ii) If \( L \sim M \) then is there an invariant such that \( f(L) \neq f(M) \)?

These questions and many others have been investigated by mathematicians over the last 150 years and is now known as Geometric Invariant Theory. The seminal paper of D. Hilbert in 1893, addressed the questions of \( \sim \) and its relationship with polynomial invariants in the language of groups. Another landmark is the classic book of H. Weyl, which also gave recipes for the construction of invariants and relationships amongst them. Several modern books are available on this theme.
The crucial point in using invariants is that it is possible to determine the validity $K \sim L$ without actually computing the $X$ and $P$ which take $K$ to $L$. This is of considerable interest to us, since the evaluation and comparison on invariants will substantially reduce the computational effort of determining $K \sim L$. Secondly, we may select a suite of invariants $f = \{f_1, \ldots, f_k\}$ which we use to "coordinate" $C$, the space of configurations. Indeed, for any $K \in C$, we construct the $k$-vector $f(K) = (f_1(K), \ldots, f_k(K))$. In some norm on $\mathbb{R}^k$ if $\|f(K) - f(L)\| < \epsilon$, then we declare $K$ and $L$ to be super-imposable. We point out here that computing $K \sim L$ via invariants does not necessarily construct for us the transformation taking $K$ to $L$. However, the proof of existence of such a transformation rather than its exact parameters, is what we need.

Classically, this phenomenon of the ease of determination of $\sim$ is well known. Compare for example, the problem of determining if two diagonalizable matrices $A$ and $B$ have the same eigenvalues. This corresponds to declaring $A \sim B$ if there is an $X$ such that $XAX^{-1} = B$. Determining $A \sim B$ for diagonalizable $A, B$ is simple, just compare $\text{Trace}(A^i)$ with $\text{Trace}(B^i)$ for $i = 1, \ldots, n$. However, construction of the $X$ taking $A$ to $B$ is extremely intricate.

The invariants

We now come to the construction of the actual invariants that we may use for our configuration space and allowed transformations. For the case when all the atoms in the pattern are distinctly labeled and when reflections are also allowed, the basic set of invariants is well-known. It is precisely the set of inner products $K = \{q_{ij}^k = (p_i - p_{i+1}, p_j - p_{j+1})\}$ where $p_i$ denotes the position of the $i^{th}$ atom and $N$ is the number of atoms in the pattern. Note that this set includes $Q = \{d_{ij} = |p_i - p_j|\}$ for $1 \leq i < j \leq N$, the distance between $p_i$ and $p_j$, the position of the $i$-th and the $j$-th atoms.

If reflections are not allowed, then the basic set of invariants includes the determinantal terms $D_{ijk} = \text{det}(p_1 - p_i, p_j - p_i, p_k - p_j)$. The determinant of the matrix formed by the columns $p_i - p_a, p_i - p_b, p_i - p_c$ respectively. Note that the square of $D_{ijk}$ is already expressible in terms of $Q$ above, and thus, it is the all-important sign of the quantity which captures the orientation. Once this basic set of invariants $L$ is chosen, all polynomial invariants are generated from the set $L$.

Having stated the general result for the unit content case, we move to the typical case such as DDDEN, when the multiplicities of atom labels may be greater than unity. This involves the transformation $P^i$ above, which acts on a configuration from the right (Equation (4)). Since matrix multiplication is associative, the action of a typical $P^i$ commutes with that of a typical $X$. The upshot is that, we may take the set $L$, the invariants for the left action, and consider the (right) action of $P$ on it. The invariants for this action on $L$ will be invariants for the total action on $C$.

For our example DDDEN, $d_{ia}$, the distance between the first and the fourth atom is an invariant under translation and rotation, which act on the left. The right action of the permutation of the three D atoms "symmetrizes" $d_{ia}$ to $d_{i+1a} + d_{i+2a} + d_{i+3a}$, the sum of the distances between the D-E atoms. Another invariant is $|\text{det}(p_1 - p_5, p_2 - p_5, p_3 - p_5)|$, the unsigned volume of the tetrahedron formed by the three D and the N atom. If we define $p_0$ as $PD = p_1 + p_2 + p_3/3$, and define $q$ as $\text{det}(p_4 - p_0, p_5 - p_0, p_5 - p_0) + \text{det}(p_4 - p_0, p_5 - p_0, p_2 - p_0) + \text{det}(p_4 - p_0, p_5 - p_0, p_3 - p_0)$. We easily see that $q$ is an invariant and differentiates between stereo-isomers. This invariant or others above are easily obtained in a systematic manner from $L$. All the invariants above are intrinsically polynomials in the entries of the matrix $K$ which defines a configuration. Thus these are continuous and tolerant to observation errors. In other words, if $M$ and $L$ are nearby configurations with the same content, then $f(M) = f(L)$. However, their sensitivity to perturbations differs for different invariants and largely depends on the degree of the invariant. Whence, in the selection of the suite $f$, preference is usually given to low degree invariants, and/or equivalent evaluation schemes must be developed for higher degree invariants.

The overall clustering algorithm

We now outline our scheme of coordinatizing the so-called protein universe. There are four basic functions, which are used repeatedly. These are:

- $[\text{IList}, c] = \text{InvariantList}(c)$: This function takes as input a typical content $c$ such as DDDEN and constructs a collection of invariant functions. These functions are to be evaluated on patterns to parameterize them. Currently, for patterns of 4 amino acids, $\text{IList}$ has 45 elements, for those of 5 amino acids it has 60 elements, and of 6 amino acids, it has 75 elements.

- $\text{PatternList} = \text{Patterns}(P,c)$: This function locates patterns of content $c$ in a given protein $P$. Each element of $\text{PatternList}$ is of the form $[K,\text{label}]$, where the label comes from the protein $P$ and the residue positions of the pattern forming amino acids.

- $[v,\text{label}] = \text{Plot}(\text{IList},c,K,\text{label})$: This function is the core evaluator of our scheme. When a pattern $K$ of content $c$ is input, $\text{Plot}$ evaluates each invariant in $\text{IList}$ on $K$, produces the pattern vector $v$ and labels it with $\text{label}$.

- $\text{CList} = \text{Clusters}(V,\text{Windows})$: This takes as input a collection of pattern vectors $V$ and generates a list of clusters in $\text{CList}$. The clustering parameters are passed in the variable $\text{Windows}$. This procedure is described in more detail in the next section.
Algorithm MakeClusters

1. ContentList = EnumerateContents;
2. for c in ContentList do
3.    [IList,c] = InvariantList(c);
4.    Klist = empty;
5.    V = empty;
6.   for pattern [K,label] in Klist do
7.       v = Plot(IList,c,K,label);
8.       V = V + v;
   end;
9.   CList(c) = Clusters(V,Windows);
10.  end;

We now briefly describe the above algorithm. Step 1 produces ContentList, a list of possible patterns contents, which need to be examined. Step 2 is the main loop of the algorithm. The actions 3–7 must be done for each content c in ContentList. Step 3 produces the suite of invariants, which will be used to map patterns into vectors. Step 4 needs as input the list of proteins to be analysed. For each protein P in ProteinList, Step 4 produces a list of patterns which may be considered as potential active sites. Step 5 produces the set V of vectors for each pattern coming from every protein in ProteinList. Once V is produced, Clusters produces a list of clusters for the content c. Each cluster C may be regarded as a recurring geometric pattern in many proteins and of potential biological significance.

The Clusters algorithm of step 6 is currently a standard data mining subroutine. The input is a collection of points in N-dimensional space. Clusters searches for dense regions, i.e. regions that have a large number of points. A region is defined as the N-dimensional rectangle obtained by the conjunction of the tolerance windows in each dimension. This is illustrated in Figure 2, where the original collection of points is separated into sets with each set satisfying the tolerance window on the 1st co-ordinate of the N-coordinates. Further, each of the sets is passed on to the second co-ordinate and so on. There are three obvious parameters: (i) the dimensions of the candidate regions, (ii) the threshold number of points above which a region is declared "dense" and (iii) order in which the co-ordinates are deployed. These parameters are passed to Clusters as variables Windows.

The parameters Windows are extracted from a training regime as follows. We first select a collection of proteins with known superimposable patterns. For each invariant, a tolerance value is determined which exceeds the variance in the patterns in the training data. The second parameter, threshold number of points, essentially reflects the number of distinct proteins in a dense region/cluster. Based on our earlier work, we chose this threshold as 5 for a pattern of six amino acids. Finally, the order of co-ordinates, and in turn the invariants, was chosen after some trial and error but is not optimized either for performance or for tested results.

Results
Training of the classification scheme

Our objective was to estimate the variation in geometric invariants for functional site patterns acceptable as superimposable from the training data of four superfamilies (Table 2). To illustrate, we visualize the superimposable and
non-superimposable patterns on a two-dimensional plot using two of the geometric invariants (Figure 3). In each of the plots shown here, and as is generally observed, the superimposable patterns always appear as a dense cluster. We obtain the tolerance windows, \( W \), for each geometric invariant for the superimposable patterns by placing a box around a known cluster (Table 1). In some of the plots, we could observe additional clusters made up of non-superimposable patterns. It is clear from the two plots that just two invariants may not suffice to determine superimposability. Indeed, depending on the pattern size, we have used up to 75 invariants to co-ordinate a pattern.

**Performance of the clustering algorithm**

The speed of execution of the clustering program depends on (i) the number of invariants, \( N \), used to co-ordinate the patterns in an \( N \)-space, (ii) the dimensions of the candidate regions in the \( N \)-space, (iii) the threshold number of points above which a region is declared dense and (iv) the order in which the co-ordinates (invariants)

---

**Figure 3.** Training data: (a) Perimeter (sum of all edges) of functional atom pentahedra plotted against perimeter of C-alpha pentahedra for the super-imposable (■) and non-superimposable (△) DDDES patterns from the “Concanavalin A like lectins/glucanase” SCOP superfamily. (b) Volume of functional atom tetrahedron plotted versus C-alpha tetrahedron for CHHM patterns from cytochrome c superfamily. The tolerance windows, \( W_i \), are obtained for each geometric invariant by placing a box around a known cluster of superimposable patterns.

**Table 1.** List of geometric invariants used for describing the patterns of amino acids and their windows for the known similar patterns from the training data

<table>
<thead>
<tr>
<th>Geometric invariants</th>
<th>Pattern size</th>
<th>Window for known similar patterns from training data</th>
</tr>
</thead>
<tbody>
<tr>
<td>Edge between two atoms</td>
<td>—</td>
<td>1.75 Å</td>
</tr>
<tr>
<td>Variance (distance between centroid of the pattern and the atom)</td>
<td>—</td>
<td>1.2 Å</td>
</tr>
<tr>
<td>( \sqrt[3]{\text{volume of tetrahedron}} )</td>
<td>Tetrahedron</td>
<td>1.4 Å</td>
</tr>
<tr>
<td>Perimeter</td>
<td>Tetrahedron</td>
<td>4.0 Å</td>
</tr>
<tr>
<td></td>
<td>Pentahedron</td>
<td>5.0 Å</td>
</tr>
<tr>
<td></td>
<td>Hexahedron</td>
<td>7.0 Å</td>
</tr>
<tr>
<td>Sum of variances</td>
<td>Tetrahedron</td>
<td>2.5 Å</td>
</tr>
<tr>
<td></td>
<td>Pentahedron</td>
<td>3.5 Å</td>
</tr>
<tr>
<td></td>
<td>Hexahedron</td>
<td>5.0 Å</td>
</tr>
<tr>
<td>Sum of (edge)(^2)</td>
<td>Tetrahedron</td>
<td>50 Å(^2)</td>
</tr>
<tr>
<td></td>
<td>Pentahedron</td>
<td>120 Å(^2)</td>
</tr>
<tr>
<td></td>
<td>Hexahedron</td>
<td>200 Å(^2)</td>
</tr>
<tr>
<td>Sum of (variance)(^2)</td>
<td>Tetrahedron</td>
<td>60 Å(^2)</td>
</tr>
<tr>
<td></td>
<td>Pentahedron</td>
<td>120 Å(^2)</td>
</tr>
<tr>
<td></td>
<td>Hexahedron</td>
<td>150 Å(^2)</td>
</tr>
</tbody>
</table>
are deployed. Of these, the first three are based on the training regime and implicitly reflect the notion of geometric similarity and biological significance of patterns. Set generation, sorting and set intersection, the three intrinsic steps of the clustering algorithm, consume a major portion of the execution time. The number of sets initially increase with the invariant number and then decrease and stabilize to a small value, typically between 0 and 5 (Figure 4). Our first cut clustering program took approximately 9 days for execution on a 4-processor (Xeon® Intel processors) server.

Are the patterns in a cluster really superimposable

In the absence of the ability to visualize the clusters in the N-dimensional space, we visualize a representative cluster of content CHHMMN on a 3-dimensional plot where the number of dimensions have been reduced to 3 using principal component analysis (Figure 5). A dense cluster, shown in a box, is detected by our clustering algorithm by applying conjunctive criteria on each of the N invariants. When the patterns were superimposed using SwissPDB,21 the patterns of the cluster indeed superimposed to a single family of superimposable patterns. This was found to be generally true for a set of randomly selected clusters.

Clustering of the non-redundant SCOP representatives and Statistics of the results

Protein structures of 3487 non-redundant SCOP representatives were used as input. As a first step, between 2000 and 50,000 patterns were enumerated from each of the proteins as described before.15 A total of 26,900 different types of pattern contents emerged from the protein universe with each pattern content having between 50 and 10,000 patterns. As an example, the CCCCT pattern type had 1256 patterns. Each pattern type was subjected to the clustering algorithm, separately, using the criteria shown in Table 1, which resulted in 1207 clusters (Table 2). Each of the resultant clusters was examined for the number of proteins and the number of SCOP superfamilies contained in the cluster. We categorize the clusters as (Tables 3 and 4) (i) Signature Patterns of Functional Sites: Clusters of patterns defining the functional category and the functional site of its members and are typically visited by PDB structures of one to three SCOP superfamilies, (ii) Signature Patterns at Protein–Protein Interface: The pattern forming amino acids are drawn from two or more distinct chains of a homo-dimer, hetero-dimer/tetramer, or a protein–inhibitor complex.

The overall statistics of the representation of SCOP superfamilies in the clusters is shown in Figure 6. Of the 1090 SCOP superfamilies, 634

Table 2. Patterns15 and SCOP superfamilies10 used for training of geometric invariant based classification scheme

<table>
<thead>
<tr>
<th>SCOP superfamily</th>
<th>Pattern content</th>
<th>No. of known superimposable patterns</th>
</tr>
</thead>
<tbody>
<tr>
<td>Zinc finger</td>
<td>Cys/Cys/</td>
<td>29</td>
</tr>
<tr>
<td></td>
<td>His/His</td>
<td></td>
</tr>
<tr>
<td>Cytochrome c</td>
<td>Cys/His/</td>
<td>45</td>
</tr>
<tr>
<td></td>
<td>His/Met</td>
<td></td>
</tr>
<tr>
<td>Concanavalin A like lectins/glucanases</td>
<td>Asp/Asp/</td>
<td>19</td>
</tr>
<tr>
<td></td>
<td>Asp/Glu/Ser</td>
<td></td>
</tr>
<tr>
<td>Cupredoxins</td>
<td>Cys/Cys/</td>
<td>31</td>
</tr>
<tr>
<td></td>
<td>His/Met/</td>
<td></td>
</tr>
<tr>
<td></td>
<td>Asn</td>
<td></td>
</tr>
</tbody>
</table>
superfamilies are not represented in any of the detected clusters (Figure 6(a)). This indicates that these superfamilies are under-represented in the PDB database and the relevant PDB structures do not have substructure similarity with other PDB structures. The remaining superfamilies are represented in at least one pattern, some which are signature patterns. The overrepresented families such as Concanavalin A like lectins, trypsin-like serine proteases, etc. find presence in several patterns. Likewise, of the detected 1207 patterns, 806 patterns are signature patterns, which are visited by exactly one SCOP superfamily (Figure 6(b)). Several patterns are visited by multiple superfamilies and may provide clues to relationship between these families. We show a few representative patterns with the proteins that contribute to those patterns (Figure 7). In addition, residue positions of the pattern forming amino acids of representative proteins has been shown in Table 3. A summary of the additional signature patterns along with their SCOP superfamily affiliation is shown in Table 4. We present the potential biological implications of some of the detected patterns below.

### Table 3. Representative signature patterns with the pattern index of one representative protein that visits a pattern

<table>
<thead>
<tr>
<th>SCOP superfamily*</th>
<th>PDB code*</th>
<th>Pattern content*</th>
<th>Number of proteins*</th>
</tr>
</thead>
<tbody>
<tr>
<td>(Phosphotyrosine protein) phosphatases II(c.45.1)</td>
<td>1jn-a</td>
<td>C480 H479 N340 Q358 R519</td>
<td>7</td>
</tr>
<tr>
<td>(Trans)glycosidases(c.1.8)</td>
<td>1b30-a</td>
<td>D128 D174 E132 E238 H84 T237</td>
<td>5</td>
</tr>
<tr>
<td>2Fe-2S ferredoxin-like(d.15.4)</td>
<td>1a70-a</td>
<td>C39 C47 C77 R40 S38</td>
<td>7</td>
</tr>
<tr>
<td>Aldolase(c.1.10)</td>
<td>1ald-a</td>
<td>C134 C177 R133 W147 Y173</td>
<td>6</td>
</tr>
<tr>
<td>1ald-a</td>
<td>E187 K107 K146 R148 W147</td>
<td>6</td>
<td></td>
</tr>
<tr>
<td>Bacterial enterotoxins(b.40.2)</td>
<td>1b12-a</td>
<td>E144 K137 K148 Q143 T104 T141</td>
<td>5</td>
</tr>
<tr>
<td>Beta-Lactamase (e.3.1)</td>
<td>1bgs-a</td>
<td>D131 E166 K73 N132 S50 S70</td>
<td>5</td>
</tr>
<tr>
<td>Concanavalin A-like lectins/glucanases(b.29.1)</td>
<td>1fk-a</td>
<td>E123 H137 N129 S147 W131</td>
<td>13</td>
</tr>
<tr>
<td>1ak3-a</td>
<td>E184 H158 N174 R162 R186</td>
<td>6</td>
<td></td>
</tr>
<tr>
<td>1axk-a</td>
<td>E234 Q283 W227 Y225 V235 Y236</td>
<td>8</td>
<td></td>
</tr>
</tbody>
</table>

### Table 4. Representative signature patterns with the pattern index of one representative protein that visits a pattern

<table>
<thead>
<tr>
<th>SCOP superfamily*</th>
<th>PDB code*</th>
<th>Pattern content*</th>
<th>Number of proteins*</th>
</tr>
</thead>
<tbody>
<tr>
<td>Cyclophilin (peptidylprolyl isomerase)(b.62.1)</td>
<td>1a58-a</td>
<td>C25 H159 N175 Q19 W177</td>
<td>12</td>
</tr>
<tr>
<td>EF-hand(a.39.1)</td>
<td>1a75-a</td>
<td>D95 E50 Q51 S49 Y68 Y88</td>
<td>6</td>
</tr>
<tr>
<td>Lysozyme-like(d.2.1)</td>
<td>1dka-a</td>
<td>D48 D52 D66 N39 S50 Y53</td>
<td>5</td>
</tr>
<tr>
<td>Metalloproteases (&quot;zincins&quot;), (d.92.1)</td>
<td>1cx-a</td>
<td>E202 H201 H205 M219 S206 W124</td>
<td>8</td>
</tr>
<tr>
<td>NAD(P)-binding Rossmann-fold (c.2.1)</td>
<td>1gad-o</td>
<td>C149 C153 H176 T150 Y311</td>
<td>6</td>
</tr>
<tr>
<td>Periplasmic binding protein-like II(c.94.1)</td>
<td>1a8-e-a</td>
<td>D292 D297 E83 K206 Y188</td>
<td>8</td>
</tr>
<tr>
<td>Riboflavin synthase (b.43.4)</td>
<td>1fn-a</td>
<td>C114 C272 T172 Y134 Y95</td>
<td>6</td>
</tr>
<tr>
<td>RuBisCo, C-terminal domain(c.1.14)</td>
<td>1bur-a</td>
<td>E204 H292 H294 M266 N241</td>
<td>6</td>
</tr>
</tbody>
</table>

---

One row indicates a representative pattern from a cluster.

* The peptide chain ID is indicated after the PDB code when the all the pattern forming amino acids belong to a single peptide chain.

* One letter code of the amino acid followed by the peptide chain ID (where applicable) and residue position.

* Number of PDB structures that contribute to a given pattern.

---

### Signature patterns of functional sites

(a) CEEHKT pattern extends the active site of TIM beta/alpha barrel. Triosephosphate isomerase (TIM) is a homodimer catalyzing the interconversion of dihydroxyacetone phosphate and D-glyceraldehyde 3-phosphate in the glycolytic pathway and represents a bigger family characterized by a common folding motif called the β/α barrel. Each TIM monomer consists of an eightfold repeat of a β-strand-loop-α-helix-loop motif, which folds up to form a (β/α)8 barrel. The reported catalytic triad residues, Glu/His/Lys, belong to loops connecting the β-strands to the following α-helices. We find a structural pattern of CEEHKT conserved in the members of TIM family derived from diverse biological origins such as *Bacillus stearothermophilus*, *Archaean*, *Trypanosoma cruzi*, **Parameterization & Classification of Protein Universe**
Escherichia coli, Plasmodium falciparum, Chicken, Human, etc. Five residues lie on one chain of the homodimer while the Thr residue lies on the other chain. Of this pattern, the residue positions of Glu167/His95/Lys13 residues match with the reported active site residues of TIM. However, the residues, Cys126/Glu97/Thr75 (5tim numbering) have not yet been implicated in catalysis.

(b) An active site and a novel site in cysteine proteases. Members of this family include important therapeutic targets such as human cathepsin F, a target for development of immunomodulator and cruzain, a therapeutic target of Chagas’ disease. The family shares a common catalytic mechanism and active site despite diverse biological origins of its members. Cys25 and His159 form the core of the active site while the neighboring Asn179, Gln19 and Trp181 (1ppo numbering), play a role in stabilizing the transition state. Residue positions of several cysteine protease-specific patterns such as CHNQW, HNQWW, CQSSWY, match with the reported active site residues, Cys25, His159, Asn179, Gln19 and Trp181 (Figure 8(a)). The patterns show additional residues, Asn18 and Trp185, conserved in the vicinity of the active site and positioned to interact with the oxygen ion hole. In addition, we detect a distinct pattern DEQSY, which has not yet been implicated in either the active site or in any of the substrate binding clefts. DEQSY forms a cleft in the vicinity of the active site.

(c) Three mutually exclusive patterns of concanavalin A-like lectins/glucanases. This SCOP superfamily consists of several families such as legume lectins, galectins, glucanases, xylanases, etc. We find several patterns specific to this superfamily (Tables 3 and 4). Three representative patterns, EHNSTR, EHHNWT and EQWYYY, are shown (Figure 7(a)). These patterns separate the members of this superfamily into three mutually exclusive sets of PDB entries. The EHNSTR pattern is populated by legume lectins, the EHNRR by galectins while the EQWYYY pattern contains xylanases. The residue positions of amino acids, which make up the legume lectin specific patterns, match with the known metal ion coordinating ligands (Figure 8(b) and (c)). The conserved amino acids of the EQWYYY pattern, Glu79, Gln129, Tyr70 and Tyr81 (1bk1 numbering) are known to be part of the active site of xylanases (Figure 8(d)) and Glu79, which is thought to be the active nucleophile, is held in place by interactions with Gln129 and Tyr70. The substrate binding cleft is lined with several tyrosine residues and other aromatic amino acid residues, which are perfectly positioned to stabilise at least four xylose residues by stacking and hydrogen-bonding interactions. In galectins, the residue positions of the EHNRR pattern match with the reported carbohydrate binding site (Figure 8(e)). Thus, these three patterns not only classify the members of this superfamily

Table 4. Additional example signature patterns and their SCOP superfamily affiliations

<table>
<thead>
<tr>
<th>SCOP superfamily</th>
<th>Pattern content</th>
</tr>
</thead>
<tbody>
<tr>
<td>(Trans)glycosidases(c.1.8)</td>
<td>DEHHNQ, EEQWW, HKNNQT, NNTYYY</td>
</tr>
<tr>
<td>Acid proteases(b.50.1)</td>
<td>DDTWY</td>
</tr>
<tr>
<td>alpha/beta-Hydrolases(c.69.1)</td>
<td>EHHQS</td>
</tr>
<tr>
<td>alpha-Amylases(b.71.1)</td>
<td>DDDER, DDEHY, DDHQ</td>
</tr>
<tr>
<td>Beta-Lactamase/d-ala carboxypeptidase(e.3.1)</td>
<td>DEKNSS, DDDRRT, DKSS</td>
</tr>
<tr>
<td>Concanavalin A-like lectins/glucanases(b.29.1)</td>
<td>DDDEHS, DDEWY, EEOQRY, EEQFY, EWYYY</td>
</tr>
<tr>
<td>Cytochrome c(a.3.1)</td>
<td>CHHMNN, CHHMT</td>
</tr>
<tr>
<td>Heme-dependent peroxidases(a.93.1)</td>
<td>DDEES, DEEQS, DEQSS</td>
</tr>
<tr>
<td>Lysozyme-like(d.2.1)</td>
<td>DDDNSY, DDNNSY, DDSTY</td>
</tr>
<tr>
<td>Metalloproteases (zincins) (d.92.1)</td>
<td>DDDHY, DHHHMS, DHHHSS, EHHHMS</td>
</tr>
<tr>
<td>Multitheme cytochromes(a.138.1)</td>
<td>CCHHH, CCHM</td>
</tr>
<tr>
<td>NAD(P)-binding Rossmann-fold (c.2.1)</td>
<td>CCHST, CCMTT, CCNTY, CTTYY, NSTWY</td>
</tr>
<tr>
<td>Periplasmic binding protein-like II(c.94.1)</td>
<td>DDEFY, EHRY, KKSSYY</td>
</tr>
<tr>
<td>Phospholipase A2, PLA2(a.133.1)</td>
<td>DDNTYY, DDNYYY, DDYY</td>
</tr>
<tr>
<td>Protein kinase-like (PK-like)(d.114.1)</td>
<td>DDDHY, DDDHY, DDHY</td>
</tr>
<tr>
<td>Xylose isomerase-like(c.1.15)</td>
<td>CDERS, CEQWY, DDHMS, CRSWW, DDKNN</td>
</tr>
<tr>
<td>Zn2/Cys6 DNA-binding domain(g.38.1)</td>
<td>DDDTW, DDDHW, EEHKMW, EHHMNW</td>
</tr>
<tr>
<td>Metallothionein(g.46.1)</td>
<td>CCCC</td>
</tr>
<tr>
<td>Fe,Mn superoxide dismutase (SOD)(a.2.11)</td>
<td>CCCCC</td>
</tr>
<tr>
<td>Triosephosphate isomerase (TIM)(c.1.1)</td>
<td>DHHHT, DHMNST, HTTYTQ, QTQ</td>
</tr>
<tr>
<td>Subtilisin-like(c.41.1)</td>
<td>DDHMRR, DHQSSW, DHTSS, EKMT, NNNN</td>
</tr>
<tr>
<td>Trypsin-like serine proteases(b.47.1)</td>
<td>CDKNS, CHQNQ, EKRRA, MQQBSW</td>
</tr>
<tr>
<td>Viral coat and capsid proteins(b.10.1)</td>
<td>DEEEHS, DHHSY, EEHWY, EHWW</td>
</tr>
</tbody>
</table>

Parameterization & Classification of Protein Universe
into three functional categories, but also provide clues about the functional sites of the respective categories.

(d) Metal binding and catalytic sites of Xylose isomerase. Several patterns including DDDKN, DDEEH, and DDHKWW, characterize the xylose isomerase family. In all, the patterns of xylose isomerase result in Asp245, Asp257, Asp287, Glu217, His220, Lys183 and Trp137 as conserved residues (2xis numbering). In many xylose isomerases of known structure, seven residues (four Asp, two Glu, and one His) are thought to participate in binding metal ions and three residues (one His, one Asp, and one Lys) are thought to participate in substrate binding.28,29 The Trp137 is proposed to play a role of creating a hydrophobic environment in the vicinity of C1 and C2 of the substrate, where the hydride shift is thought to take place.30 Amino acid positions of the detected patterns match with the reported metal and substrate binding residues (Figure 8(f)).

(e) Coordination sites for structural and catalytic zinc ions in metalloproteases. Matrix metalloproteinases (MMPs) belong to a family of zinc dependent endopeptidases, which cleave proteins of the extracellular matrix. These enzymes are attractive therapeutic targets as they are involved in normal physiological processes such as angiogenesis, extracellular matrix remodeling, embryogenesis and differentiation,31,32 and also in pathological processes such as metastasis, arthritis, and tumor invasion.33,34 MMPs are characterized by two tetrahedrally coordinated zinc ions, one catalytic and one structural.31 The catalytic zinc ion is coordinated by the imidazole nitrogen atoms His201, His205 and His211 and a carboxyl oxygen of Glu202 (1cvx numbering). The structural zinc ion coordinates His151, His166, His179, and the carboxyl oxygen atom of Asp153. The metalloprotease specific patterns, DDHMY, DHHHMS and EHHMSW, consist of residues, which ligate with the catalytic zinc ion and the structural zinc ion...
However, the role of the Met219 and Trp124 residues, conserved in these patterns, has not been reported.

(f) **Active site of β-lactamases.** Pathogenic organisms acquire resistance to β-lactam antibiotics mainly through β-lactamases. The most common β-lactamases are the class A and class C enzymes that function *via* an active site serine residue. The reaction mechanism involves N-protonation of the substrate, catalyzed through a hydrogen bonded network...
cluster involving the 2-carboxylate group of the substrate, the side-chains of Ser130 and Lys234, and a solvent molecule. Nucleophilic attack on the carbonyl carbon is carried out by the side-chain of Ser70 with proton abstraction catalyzed by a water molecule hydrogen-bonded to the side-chain of Glu166. Of the β-lactamase specific pattern detected here, DEKNSS, the Glu, Lys Ser residues match with the reported active site residues whereas the Asp and Asn residue has not been implicated in catalysis (Figure 8(h)).

(g) Two types of haem coordination sites in Cytochromes. Cytochrome-c acts, besides other functions in the cell, as a mediator of electron transport in the respiratory chain of prokaryotic and eukaryotic organisms. Nitrite reductases as well as photosynthetic reaction centers are made up of cytochrome-like domains. Cytochrome domains are known to bind between one and sixteen haem prosthetic groups per peptide chain. The haem groups covalently bind to two cysteinines of the CXXCH-consensus motif. The histidine from this sequence and another methionine or histidine coordinate the iron atom of the porphyrin ring. We detect two distinct patterns, CCHH and CCHM which are specific for cytochromes. The CCHH pattern contained mainly the multihaem cytochromes whereas the CCHM pattern contained mainly the single haem cytochrome c proteins with the exception of 1ey and 2pqc, both of which are photosynthetic reaction centers. Interestingly, 1ey and 2pqc contain one CCHH pattern and three CCHM patterns each, per peptide chain. In both the patterns, the first three residues are derived from the CXXCH motif whereas the relative sequence position of the fourth residue, His or Met, varies significantly from pattern to pattern, even within a single polypeptide chain. The number of distinct CCHH patterns detected per peptide chain correlates well with the number of haem groups known to bind to that chain. To exemplify, for PDB code 19hc, the nine-haem cytochrome from Desulfovibrio desulfuricans, we detected nine distinct CCHH patterns. It may be noted that the geometry of the cytochrome-specific CCHH pattern detected here is similar to the zinc finger specific CCHH patterns.

(h) Novel sites in Protein Kinases. Certain Protein kinases have a membrane binding domain, a unique region, a Src homology 3 (SH3) domain, a Src homology 2 (SH2) domain, a kinase catalytic loop and an activation loop. The SH2 domain binds phosphotyrosine containing ligands, and the SH3 domain binds ligands containing proline rich regions. Activation loop is hydrogen-bonded to Asp-1132 and Arg-1136 in the catalytic loop (1ir3 numbering). Two of the detected patterns, DDHMS and DDHRS, contain Asp1132 (1ir3 numbering), which has been reported to be part of catalytic mechanism. Other residues of this pattern have not been implicated in the reaction mechanism. The pattern forming residues are derived from a helix adjoining the catalytic loop (helix-5) (Met 1120), the catalytic loop (His1130, Arg1131 and Asp1132) and helix-9 (Asp1191 and Ser1194). The third pattern, CRSWW, was away from the catalytic loop as well as the activation loop and the residues were drawn from Helix-11 (Cys1245), loop between helix 11 and 12 (Trp1246), Helix 12 (Arg1253) Helix-9 (Ser1194 and Trp1193). These residues have not yet been implicated in catalysis or binding and thus provide novel avenues to probe their role.

(j) Active site of Trans-glycosidase. We detected several patterns including DEEEHT, DKNNQY and EHHNRW, which are specific for this family. The residue positions match with those reported for the active site of the enzyme family (Figure 8(i)). Glu166 and Glu352 (1bg numbering) are positioned in the vicinity of the anomeric sugar C1 atom, with Glu352 acting as a nucleophile. Trp398, makes stacking interaction with the plane defined by the ring-like guanulate chain. Tyr296 is hydrogen bonded to Glu352 and is thought to contribute to stabilization of the oxocarbonium species. Arg77 is known to be conserved in the vicinity of active site; however the role of this residue has not been described.

Signature patterns at protein–protein interface

(a) Tetramer association in viral coat and capsid proteins. Members of this family include animal and human pathogens, such as rhinoviruses (etiological agent of common cold), foot-and-mouth disease virus, coxsackievirus, and polioviruses and are collectively known as picornaviruses. Picornaviruses are small (30 nm) icosahedral particles composed of 60 copies each of the four capsid proteins VP1-VP4, which enclose a single-stranded positive-sense RNA. Majority of the viral coat protein specific patterns were formed by amino acids at the protein–protein interface (Tables 3 and 4). As an example, the CERWYY pattern is formed by three residues from VP1 and three residues from VP2 (Figure 9(a)). These patterns provide a clue to the conserved nature of the hetero-tetramer, which may have significant implications for drug discovery.

(b) Patterns at the interface of Trypsin-like serine protease and inhibitor. We detected several patterns such as, DKRRWW, EEERRW and EEKRW, at the interface of the enzyme–inhibitor complex. Of these, the aspartate, two glutamate and one arginine residues belong to the inhibitor while one glutamate, one lysine, one arginine and two tryptophan residues belong to the serine protease (Figure 9(b)). We conjecture that the complementarily charged residues on the enzyme and the inhibitor play a role in stabilizing the complex. The role of the two Trp residues is not clear. The patterns were detected in several different enzymes complexed with inhibitors such as leech antihemostatic proteins, Ovomucoid/PCI-1 like, and BPTI-like inhibitors.

(c) A conserved Glutamate “double bridge” of Fe/Mn
**superoxide dismutase.** Superoxide dismutases (SODs) are ubiquitous metalloenzymes that catalyze the one-electron redox cycle of superoxide disproportionation. This radical-scavenging function defends biological systems against oxidative damage by intercepting reactive oxygen species. They have been shown to make an important contribution to the survival of several intracellular bacterial pathogens such as *M. tuberculosis* mycobacteria and therefore, are considered as an important therapeutic target. FeSODs and MnSODs share substantial sequence and three-dimensional structural similarity. In *E. coli* MnSOD, Glu170 (1vew numbering) of one monomer, which hydrogen-bonds to the Mn ligand His171 of the other monomer, has been proposed to form a double bridge at the dimer interface. Mutations at this Glu residue were shown to destroy activity. We detected several Fe/Mn-SOD specific patterns such as DEEHHY, HHHHHY, EESWWY and EEYYY. These are formed by one Asp, one Glu, four His, one Trp and two Tyr residues per chain. Of these, the Asp, Glu and His residues are the metal ligands where the Glu also acts as a double-bridge as discussed above (Figure 9(c)). The hydrophobic Trp and Tyr residues potentially provide stability to the homo-dimer by stacking interactions (Figure 9(d)).

**Discussion**

This work presents a scheme for classification of the protein universe based on recurring structural patterns. This provides a significant refinement over the conventional methods, which are based on overall sequence or structure alignment. The recurring patterns, which are used as keys of classification, provide clues about the functional sites in the proteins as is shown for several examples here. Superimposability of patterns is arrived at by comparing geometric invariants, which is a novel feature of this work. The results include signature patterns, which correspond to the functional sites of the respective protein families where the patterns have revealed additional residues compared to the known functional sites in several families. This includes the important therapeutic targets such as phospholipase A2, cruzain, superoxide dismutase, etc. Several of the patterns were detected at the protein–protein interface, which provide clues to the dimer interactions, protein–inhibitor interactions, etc. Likewise, all the members of a SCOP superfamily or family do not necessarily appear in a given pattern. For example, the Concanavalin A like lectins/clusanases superfamily contains 63 PDB structures. Of this, lectins, galectins and xylanases get separated into three mutually exclusive clusters. In this manner, the clustering algorithm is able to sort proteins by their functional sites and in turn by functional category.

Recently, a purely geometrical approach has

---

**Figure 9.** Conserved patterns at the protein–protein interface. (a) Interactions between VP1 (red) and VP2 (blue) viral coat proteins in the hetero-tetrameric complex: PDB code 1aym. (b) Interactions between the trypsin-like serine protease (purple) and inhibitor (blue): PDB code 1tom. (c) The conserved ligands of the Mn$^{2+}$ ion and the glutamate double-bridge of Fe,Mn-superoxide dismutase. (d) The amino acids at protein–protein interface which could potentially be involved in homo-dimer stabilization of Fe,Mn-superoxide dismutase. For (c) and (d), PDB code 1vew, chain C: blue; chain D: red. (c) and (d) are generated by considering amino acids resulting from all the signature patterns of that superfamily and then separating them into patterns of hydrophilic residues (His, Glu, and Asp) and those of hydrophobic residues (Tyr and Trp) in (c) and (d), respectively. The Figures are generated using SwissPDB viewer.
been described, which uses a combination of interatomic distances based comparison of patterns followed by superposition of the candidate patterns. In fact, several other distance based pattern comparison methods are known. Indeed, distance-related measures are present in the suite of geometric invariants chosen by us. These can be considered as the first order geometric invariants. Additionally, we use second order geometric invariants such as surface area and summation of products of edges and third order invariants such as the signed volume where the sign can distinguish between non-superimposable mirror images. The different types of invariants have different degree of sensitivity toward perturbation in the conformation. A discussion on the sensitivity analysis of invariants is beyond the scope of the current manuscript. A carefully chosen suite of geometric invariants guarantees superposition of the patterns/structures/substructures with similar values of invariants. We believe that our method attempts to overcome the limitations of the current methods, which are based on comparison of distance matrices.

The patterns at the protein–protein interface reveal the conserved nature of the dimeric/tetrameric assembly as well as the key conserved residues involved in the protein–protein interaction. When the pattern forming amino acids residues were compared with literature, a majority of them were either in the vicinity of the reported functional site or were at the protein–protein interface. It may be noted that the number of patterns shown here (Tables 3 and 4) is a small fraction of the total number of patterns discovered. Thus, the results presented here provide a vast resource for the biologists for experimental verification of the proposed functional sites in the thousands of PDB structures.

The clustering approach uses several adjustable parameters based on (i) the tolerant windows of the geometric invariants for declaring the patterns as superimposable and (ii) the minimum recurrence frequency of a pattern, to consider only those patterns which have a certain level of representation in the non-redundant database. Both type of parameters are tunable. The current results are based on one set of values for these adjustable parameters, values chosen primarily from data management point of view as we describe below. The SCOP database contains 329 non-redundant PDB structure entries for the immunoglobulin superfamily. Thus, when all the non-redundant PDB entries were used as input, a large fraction of the resultant patterns were found to be signature patterns for immunoglobulins. We conjecture that this has resulted due to overrepresentation of this family in the database. Thus, in order to manage and analyze our results in a better manner, we decided to omit the 329 PDB structures belonging to the immunoglobulin superfamily. Furthermore, a refined analysis of results may be possible for extraction of the underrepresented patterns, by using a subset of the current input data, or by lowering the threshold number of patterns for consideration of a cluster to be dense. Alternatively, it is possible to allow a greater degree of geometric dissimilarity among the patterns keeping in view the extent of uncertainty involved in experimental structure determination. For example, in the current results, 634 of the 1090 SCOP superfamilies do not appear in any of the patterns. This may be a result of low levels of structure/sub-structure similarity of these protein structures with other structures available in the PDB database. It may be important to solve crystal structures of those proteins, which belong to these “under-represented” superfamilies.

Finally, it might be appropriate to provide a rough idea about the improvement in the computational speed with the current approach. Computation of the superimposing transformation is an optimization problem, which is intrinsically computationally intensive and needs to be performed on a pair of structures at a time. Likewise, the distance matrix based structure comparison methodologies typically operate on a pair of structures. With millions of structural patterns enumerated from the protein universe, a majority of the trillions of comparisons would be redundant as only a small fraction of structure-pairs are superimposable. The current method provides a substantial improvement in computational speed by providing unilateral descriptive features in the form of geometric invariants for each pattern, which makes the patterns amenable to the conventional clustering programs. The patterns, which form a cluster in an N-dimensional “geometric invariants” space, are almost guaranteed to superimpose with each other. Computation of the transformation for each pattern pair in the cluster could be an additional step, although it can be omitted in our application once we are convinced of the “superimposability” of the patterns in a cluster. We emphasize that it might be feasible to employ any of the currently available probabilistic clustering algorithms, hierarchical or non-hierarchical, such as K-means, self organizing maps, etc. In such a case, one may have to devise an appropriate strategy to evaluate each cluster for its compactness. We chose to employ a deterministic clustering approach whereby the tolerance for the deviation on each geometric invariant was pre-calculated using a training data of superimposable patterns.

This is the first report of systematic application of the geometric invariant theory for analysis of protein structures. The strategy developed here could find applications in several domains of structural biology and chemistry including protein folding, drug design, etc. From the computational biology/chemistry perspective, this work opens new opportunities for geometric invariant-based extraction of relationships based on either local or global structural similarities.
Acknowledgements

The work was partially funded by a grant from the Council of Scientific and Industrial Research, Government of India.

References


Edited by J. Thornton

(Received 25 April 2003; received in revised form 5 July 2003; accepted 11 September 2003)

www.sciencedirect.com

Supplementary Material consisting of the recurring patterns with all pattern forming amino acids on a single peptide chain (Table S1), the recurring patterns at the protein–protein interface (Table S2) and a list of PDB codes used in our analysis (Table S3) are available on Science Direct