World Research Journal of Peptide and Protein
ISSN: 2278-4586 & E-ISSN: 2278-4608, Volume 1, Issue 1, 2012, pp.-33-40.
Available online at http://www.bioinfo.in/contents.php?id=131

ARABESQUE: A TOOL FOR PROTEIN STRUCTURAL COMPARISON USING DIFFERENTIAL GEOMETRY AND KNOT THEORY

HOI TIK ALVIN LEUNG[1], BERNARDO OCHOA MONTAÑO[2], TOM BLUNDELL[2], MICHELE VENDRUSCOLO[1] AND RINALDO WANDER MONTALVÃO[1]*

[1]Department of Chemistry, University of Cambridge, Lensfield Road, CB2 1EW, Cambridge, UK.
[2]Department of Biochemistry, University of Cambridge, 80 Tennis Court Road, CB2 1GA, Cambridge, UK.
*Corresponding Author: Email- rwm35@cam.ac.uk

Received: January 11, 2012; Accepted: February 14, 2012

Abstract- We present ARABESQUE, a new tool for protein structure analysis, which includes structure comparison, generation of annotated structural alignments, and annotated superposition of structures. By combining differential geometry and knot theory, this method produces an accurate analysis of structural conservation in a family of proteins. The annotated alignment and superposed structures are used to characterise the local and global structural information content, to refine the sequence alignment and to produce fragments and 3D probability density functions for comparative modelling.
Keywords- Protein structure comparison, Protein structure alignment, Differential geometry, Knot theory.

Citation: Hoi Tik Alvin Leung, et al (2012) ARABESQUE: A tool for protein structural comparison using differential geometry and knot theory. World Research Journal of Peptide and Protein, ISSN: 2278-4586 & E-ISSN: 2278-4608, Volume 1, Issue 1, pp.-33-40.

Introduction

The comparison of the structures of proteins enables one to determine close and distant relationships amongst them [1]. This type of analysis can be used to classify proteins into families with similar folds and properties [2]. Such a classification is very useful since ensembles of related proteins within a family contain enough information to allow patterns in both sequences and structures to be identified. These patterns play a vital role in the understanding of a variety of aspects of protein behaviour, including their structural stability, biological activity, molecular evolution and structural conservation. While close relationships can easily be identified by using sequence similarity alone, distant relationships can often be determined only through a comparison of three-dimensional structures, since two proteins with low sequence identity can share similar folds, biological function and physicochemical properties. These aspects follow as a direct consequence of the fact that the tertiary structure of proteins is more conserved than their sequences due to the action of selective pressures on the protein function(s) [3-4].

The systematic organisation of proteins into families can be used to predict the fold of proteins through homology modeling [5]. In this approach, the structure of a protein is predicted from its sequence using information derived from homologous (i.e. divergently evolved) structures, together with additional rules inferred from general structural data [5-8]. Homology modelling programs can predict the structure for proteins from their amino acid sequences by extrapolating their structural features from the structures in their families. MODELLER [5] and ORCHESTRAR [6-8] are examples of homology modelling packages used for building new structures from currently available structures.

It is not always easy to identify structurally conserved regions based on sequence alone, especially when the average percentage of identity (PID) for a given protein family is low. Consequently the development of methods to identify such regions is essential not only for protein comparison but also for homology modelling. The successful identification of structurally conserved regions demands a measure for structural divergence between two protein fragments that satisfies the triangle inequality rule [9]. However, most of the current measures, such as RMSD, violate this rule and are unable to judge dissimilarity [9], thus creating substantial difficulties for using clustering algorithms to identify structurally conserved regions in protein families with large structural divergences.

Other geometrical measures have been employed in order to

overcome such limitations. Differential geometry was employed by Louie and Somorjai to study structural and dynamical representation of patterns in proteins [10-11] and for creating models for enzymes [12]. They suggested that differential geometry is the adequate framework for a "unifying and natural description of the three-dimensional conformation of proteins". Their main argument is that regular secondary structure elements are related to geodesics on minimal surfaces. They found that α-helices lie on the conjugate minimal surfaces of the helicoid, β-barrels lie on the catenoid and, additionally, the intermediate states of the isometric transformation between both surfaces model a variety of β-twisted sheets commonly found in proteins. Rackovsky and Scheraga have also developed a differential geometry representation of protein backbone structures and demonstrated a number of applications. They showed that differential geometry representations can be used for comparing local folding of backbone structures in a quantitative manner [13-14] or to inspect the initial stages of protein folding and to predict which structures are likely to be formed [15]. These results are possible due to the fact that their differential geometry representations operate on a four-Cα length scale, therefore highlighting structural features that are not clearly emphasised by the φ/ψ dihedral angles as they operate on a single-residue length scale. One of the practical uses of differential geometry in protein structure comparison is employed in a method for protein structure similarity search called CTSS [16]. This method is based on the differential geometry theory applied to continuous 3D space curves created by a spline interpolation of the Cα atoms. Their method is able to extract compact, robust and localized shape signatures that can be used for pairwise alignment of protein structures.

In addition to differential geometry methods, knot theory has also been employed for analysing protein conformations on a global [9,17-20] or a local scale [21]. Peter Røgen and collaborators have employed numbers inspired by Vassiliev knot invariants to construct a dissimilarity measure, named Scale Gauss Metric (SGM), to study the automatic classification of protein folds. This measure was applied in an automatic procedure for the CATH2.4 database, which resulted in the classification of 20,937 connected domains from the CATH2.4 with a success rate of more than 95%. In another study Dewey et al. [21] described a method based on a local geometric property for pairwise (TLOCAL) and multiple structure alignments (TCLUSTALW). The local geometric measure in this method is a quantity derived from Vassiliev integral formulas for knot invariants called the writhing number. This geometric measure is calculated in a sliding window, typically five residues long, and represented by a 20-letter code. By converting the continuous writhing number into a discrete 20-letter code the structural alignment problem is mapped into a sequence alignment problem similar to the traditional one, thus allowing the use of the same algorithms designed for sequence alignment. One of the greatest advantages of this type of geometric alignment is that it is able to deal with structural intricacies present in highly divergent protein families.

The main objective of ARABESQUE is to produce not only a structural alignment but also an annotation to highlight the sets of fragments that are structurally conserved across the members of a specific protein family. It combines measures originated from differential geometry and knot theory in order to create an approach for determining structural conservation on the local and global scales. This approach extends our previous one (CHORAL), which applies differential geometry, but not knot theory, for modelling the conserved cores of proteins [6] by inferring their structure from the sets of conserved fragments.

One problem in finding the ensemble of conserved fragments is that a Structurally Conserved Region (SCR) is usually defined as a region where all proteins in the same family show the same conformation for the main chain atoms independent of their structural classification as secondary structure elements or loops. This aspect implies that the length of the SCR tends to be proportional to the family percentage of identity and inversely proportional to the number of its members and, additionally the superposition of Cα atoms often leads to the equivalency of regions displaying quite different conformations. This problem creates a paradoxical situation where the more structures are known for a given low PID family the less SCRs can be inferred from it.

A clear example of this problem is presented in the program SCORE [22], created to model the conserved core of proteins, which defines an SCR as a continuous stretch of three or more aligned residues conserved across all aligned structures. The residues are said to be conserved when the distance amongst all Cα atoms for each aligned position is less than 3.8 Å and the difference in backbone torsion angles lies within a threshold of 150◦. As just one residue from a protein falling outside the thresholds is enough to disallow a three residue long region from being considered conserved the addition of more divergent structures will reduce the number of SCRs. This definition is also clearly a necessary condition for any structural conservation of the main chain but it is not sufficient to insure it as the relaxed threshold for the backbone torsion angles, used to improve on the numbers of SCRs found, can enforce regions with incompatible geometry to be defined as conserved.

The method employed by CHORAL was designed to improve the performance of a fragment-based program for modelling families or super-families showing low sequence identity. It introduces the concept of Structurally Conserved Clusters (SCCs), which uses as much information available as possible for a protein family and also introduces new strong geometric requirements towards a sufficient condition for structural conservation. In CHORAL, an SCC is defined as a geometric pattern (shape) that may be unique to any single member of the structural alignment or common to any combination of structures in the family. This approach allows several SCCs to span a single region of the structural alignment where no single SCR would be defined.

We created ARABESQUE in order to extend and improve the definition of SCC used by CHORAL and output the internal representation of the conserved sets of fragments as an annotated sequence alignment and superposed structures. Both the sequence alignment and the superposed structures are colour coded in a way to reflect the structural conservation of the regions and sub-regions. The information can be used not only for understanding the patterns of conservation in a protein family but also for correcting any mistake in user defined sequence alignments. The final alignment can be optimally used not only by ORCHESTRAR (that uses CHORAL) but by any modern homology modelling program as it enforces a better geometric alignment.

Methods

Structural alignment

The initial step for the ARABESQUE algorithm is the production of superposed structures and a sequence alignment. The method used for this purpose is called BATON, which is based on COMPARER [23]. In this method, proteins are treated as an array of elements that inhabit different layers of the protein structural hierarchy. The first layer is the sequence similarity that acts as the basic indicator for equivalence between parts of the structures. Additional information about features derived from the higher hierarchical levels (secondary structure, super-secondary structure, motif and domain) is added to each residue. These features can describe either a property of that residue in particular or relationships between residues. The residue properties are classified into two distinct groups:

- Inherent: charge, side-chain size, hydrophobicity, etc.
- Structure dependent: φ and ψ dihedral angles, side-chain and main-chain orientation in relation to the centre of mass, solvent accessibility, etc.

The property and relationship scores are incorporated into a residue-by-residue weight matrix resembling the similarity matrix used in sequence/sequence alignment where each property contributes its weight to the similarity between two residues [24]. A dynamic programming algorithm is used in order to find the optimal pairs of residues and it outputs the result as a sequence alignment and superposed structures. Additionally, the sequence alignment is annotated with three-dimensional structural features using JOY [24]. This annotation helps the understanding of the amino-acids conservation in their specific local environments (table 1).

Table 1- Codes representing the JOY annotation

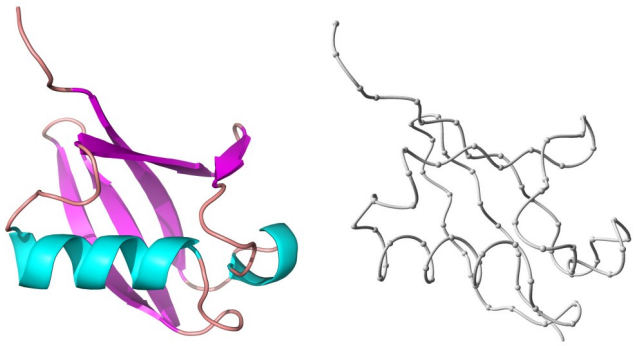
Amino Acid Conversation	Annotation	
Solvent inaccessible	UPPER CASE	X
Solvent accessible	lower case	x
α-helix	Red	x
B-strand	Blue	x
3_{10}-helix	Maroon	x
Hydrogen bond to main-chain amide	Bold	x
Hydrogen bond to main-chain carbonyl	Underlinex	
Disulphide bond	Cedilla	ç
Positive φ	Italic	x

Differential geometry of proteins

Protein structures are complex geometric objects that can be described in a wide variety of ways. The application of differential geometry to protein structure analysis is based on the expression of its geometry as 3D space curve through a parametric function [25]. As here we are interested in formulating a very sensitive method for comparing the backbone geometry of proteins we need to use a representation that retains the maximum amount of features that characterises such geometry. The most natural approach to produce parametric curves that preserves most of the backbone structural information is to apply spline interpolation to fit the Cα atoms of proteins, which creates smooth continuous curves [6,16]. We fit the coordinates of Cα atoms individually generating three different parametric equations, one for each coordinate, with the residue number as the parameter. It results in a parametric vector equation (1) where each parametric equation for the individual coordinates is a spline function instead of an analytical function.

$$\vec{r}(t) = x(t)\hat{x} + y(t)\hat{y} + z(t)\hat{z} \tag{1}$$

Different from the method used by Can and Wang [16] that employs smoothing quintic splines, cubic splines are used in the current work. By computing appropriate control knots we ensure that the 3D curve is smooth, crosses all the Cα positions and does not oscillate wildly between points, which sometimes happen with spline fitting of space curves. Figure 1 shows an example of the spline parametric fitting, for the Cα atoms, applied to the ubiquitin model (1UBQ). It is easy to see that the spline representation preserves all the important structural features shown in the cartoon representation as α-helices and β-strands can be clearly observed. The spline fitting has been tested against 800 protein models and it does not produce any distortion of the structural representation and therefore smoothing is not necessary.



(a) **(b)**

Fig. 1(a)- Cartoon representation of the native state of the protein ubiquitin (1UBQ) showing coloured secondary structures elements. **(b)-** Cubic spline representation of the same protein structure. The spheres represent the positions of the Cα atoms and the tube is the spline reconstruction for the curve that connects all the Cα atoms.

The curvature (κ) of a regular curve for any value of the arbitrary parameter *t* is related to its parametric representation (equation 1) and can be expressed as

$$\kappa = \frac{\left|\frac{d\vec{r}}{dt} \times \frac{d^2\vec{r}}{dr^2}\right|}{\left|\frac{d\vec{r}}{dt}\right|^3} = \frac{|\dot{r} \times \ddot{r}|}{|\dot{r}|^3} \tag{2}$$

This value is essentially the rate of change of the direction of the unit tangent vector with respect to the arc length, which means that it is a measure of the deviation from a straight line. In addition, torsion can be seen as the rate that a regular curve deviates from a plane and is related to the curve parametric equation by [25]

$$\tau = \frac{\left[\dot{r}\ \ddot{r}\ \dddot{r}\right]}{|\dot{r} \times \ddot{r}|^2} \tag{3}$$

Although the parametric representation using cubic splines curves is very effective in preserving the visual characteristics of the

structural elements, it is not adequate for the numerical evaluation of the derivatives used by equations 2 and 3. In contrast to CHORAL, our new approach employs cubic spline parametric functions just as a framework for enforcing the characteristics we want into the shape. The derivatives are evaluated by means of a Chebyshev approximation around the residue [26]. At residue *i* the parametric function is approximate by 50 points from *i-1* to *i+1* by the Chebyshev function and derivatives can be evaluated easily from its coefficients [26].

As curvature and torsion fully characterise a regular curve in the three-dimensional Euclidian space, one can anticipate that proteins, fitted with spline curves, which are regular curves, should possess similar geometry. One can intuitively understand this complete characterisation of a curve through curvature and torsion by observing that curves in three-dimensional space that bend (change of curvature) and twist (change of torsion) at the same points in the same way are similar because a curve possesses only two degrees of freedom.

One example of such behaviour is the elicitin HOMSTRAD family [27-29]. The elicitin is a two-member all-alpha class protein family, with a PID of 87% and a superposed RMSD of 0.88Å. The superposed structures of its family members (Fig. 2 and 3) show a high degree of structural conservation. As expected for proteins with such high degree of sequence similarity and low RMSD, the geometric signatures derived from both structures are very similar (Fig. 4). A significant point to be observed by comparing the curvature (Fig. 4a) with the alignment is that the valleys correspond to the helices. Helices are also easy to spot, as their curvature is small compared to the curvatures of β-strands or loops.

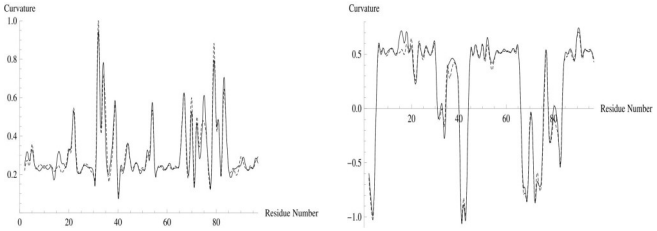
Fig. 2- Superposed structures of the elicitin HOMSTRAD family. The 1lira structure is shown in green and 1jpa structure in blue. The family has a PID of 87% and a superposed RMSD of 0.88Å.

```
                10        20        30        40        50
1lria  ( 1) taÇt asq̃q̃taAỹktLvsIlsdasfnq̃Çstd̃SgỸsM̃ltakalPt̃taq̃ỸklM
1ljpa  ( 1) taÇt atq̃qtaAyktlvsIlsessfsq̃Çskd̃SgỸs̃mlt̃atalPtñaqykl M
                αααααααααα        αααααααααα        ααααααα

                60        70        80        90
1lria  ( 51) çaš taÇntMikkIvtlnPpñÇdLtVptsglvlÑVysyAngFsnkçssl
1ljpa  ( 51) çaš taÇntMIkkivalnPpdçdLtvptsglvldvytyAngfssk̃çasl
                αα ααααααααααα        ββ        ββ ααααα ααααααα
```

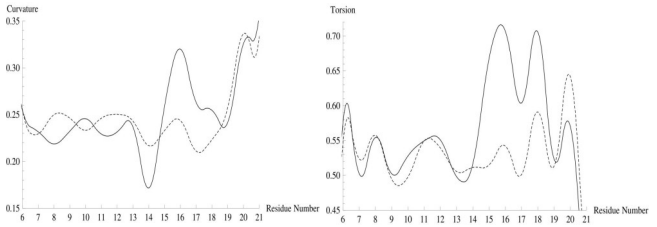
Fig. 3- JOY annotated sequence alignment of the elicitin HOMSTRAD family.

Given the simplification of the protein geometry imposed by the regular curve representation, the sensitivity of curvature and torsion signatures to the secondary structure element is surprising. In the region between aligned positions 14 and 20 in the elicitin alignment the secondary structure of 1lria protein changes from α-helix to 3_{10}-helix. Starting from residue number 13, the measures for the structures start to diverge and small changes can be observed clearly in the graphs for curvature and torsion for that region (Fig. 5). This is a very good example of the sensitivity of curvature and torsion to the morphological aspects that are relevant for studying the geometric similarities of proteins ensembles. Although a few small differences can be seen, the proteins possess a very similar geometry. We found very small values for the Euclidian distance for the curvature and torsion pairs of both proteins, as expected, but such values still highlight the changes in the helical conformation (Fig. 6).



(a) **(b)**

Fig. 4- Normalised values for (a) curvature and (b) torsion for the elicitin HOMSTRAD family of proteins. The values are normalised against the maximum value obtained from all the structures in the HOMSTRAD database. (1lria – solid, 1ljpa – dashed)



(a) **(b)**

Fig. 5- Normalised values for (a) curvature and (b) torsion for elicitin HOMSTRAD family around the first α-helix. (1lria – solid, 1ljpa – dashed)

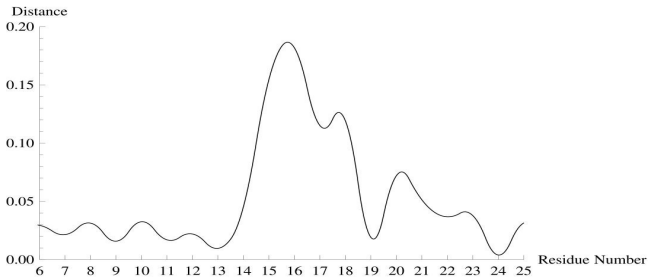


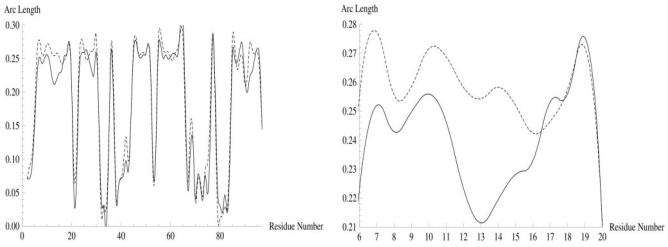
Fig. 6- Euclidean distance between the normalised curvature and torsion pairs. This region shows the difference between the 310-helix of 1lria protein and the α-helix of 1ljpa.

Curvature and torsion are local characteristics of the curve representing the protein backbone and, despite the fact that they show the local changes in geometry very well, they do not contain enough information about the geometry of the region. Although two proteins may share the same geometrical property locally, the topology of the fragment in neighbourhood of the residue can be very different. The protein backbone arc length is a natural choice

for a fragment dissimilarity measure due to its sensitivity to the compactness of regions of the spline representation and its easy geometrical interpretation. In the current method, the region for fragment centred at residue *i* is defined as the window encompassing the residues from *i-2* to *i+2* and is relate to the parametric functions though the equation

$$a(i) = \int_{i-2}^{i+2} \sqrt{\left(\frac{dx}{dt}\right)^2 + \left(\frac{dy}{dt}\right)^2 + \left(\frac{dz}{dt}\right)^2}\, dt \qquad (4)$$

It is straightforward to evaluate this expression by means of a simple numeric integration over the cubic spline parametric function. Fig. 7 shows an example of such integration for the elicitin family.



(a) **(b)**

Fig. 7- Normalised values for arc length for elicitin HOMSTRAD family for the whole alignment (a) and around the first α-helix (b). (1lria – solid, 1ljpa – dashed)

Knot theory and proteins

The writhing number, a number originated from knot theory, is used by ARABESQUE to characterise a five-residue long region centred on the central residue. This number is used as a local geometric measure that describes the degree of curvature of the protein backbone formed from the vectors connecting all the Cα atoms. The writhing number can be easily calculated by means of

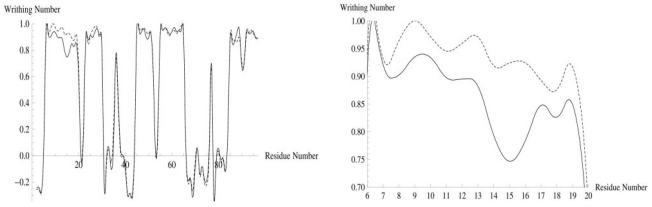
$$W_r = 2 \sum_{i=1}^{N-3} \sum_{j=i+2}^{N-1} \frac{\Omega_{i,j}}{4\pi} \qquad (5)$$

$$\Omega_{i,j} = \left[\sin^{-1}\left(\vec{a}_{i,j} \cdot \vec{b}_{i,j}\right) + \sin^{-1}\left(\vec{b}_{i,j} \cdot \vec{c}_{i,j}\right) + \sin^{-1}\left(\vec{c}_{i,j} \cdot \vec{d}_{i,j}\right) + \sin^{-1}\left(\vec{d}_{i,j} \cdot \vec{a}_{i,j}\right)\right] \cdot \mathrm{sign}\left(\vec{r}_{j,j+1} \times \vec{r}_{i,i+1} \cdot \vec{r}_{i,j+1}\right)$$

$$\vec{a}_{i,j} = \frac{\vec{r}_{i,j} \times \vec{r}_{i,j+1}}{\left|\vec{r}_{i,j} \times \vec{r}_{i,j+1}\right|}, \qquad \vec{b}_{i,j} = \frac{\vec{r}_{i,j+1} \times \vec{r}_{i+1,j+1}}{\left|\vec{r}_{i,j+1} \times \vec{r}_{i+1,j+1}\right|},$$

$$\vec{c}_{i,j} = \frac{\vec{r}_{i+1,j+1} \times \vec{r}_{i+1,j}}{\left|\vec{r}_{i+1,j+1} \times \vec{r}_{i+1,j}\right|}, \vec{d}_{i,j} = \frac{\vec{r}_{i,j} \times \vec{r}_{i,j+1}}{\left|\vec{r}_{i,j} \times \vec{r}_{i,j+1}\right|}$$

where $\vec{r}_{i,j}$ denotes the vector between the Cα *i* and *j* [21]. Fig. 8 shows an example of this number for the elicitin family. Althought this measure describes a type of curvature for the protein backbone, it is very different in nature from differential geometry curvature and they are not correlated. For the differential geometry case, the curvature describes just the local degree of curvature in space and is always positive. On the other hand, the writhing number describes a longer region and can be either positive or negative. The writhing number shows negative values for certain regions of the proteins (Fig. 8a). Figure 9 illustrates the differences between the differential geometry measure of curvature and the knot theory writhing number. The figure shows that their natures are different as the numbers describe curvature for different representations and lengths.



(a) **(b)**

Fig. 8- Normalised values for the writhing number for elicitin HOMSTRAD family for the whole alignment (a) and around the first α-helix (b). (1lria – solid, 1ljpa – dashed)

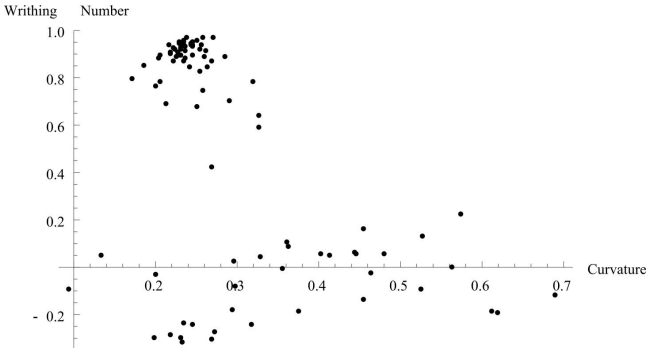


Fig. 9- Correlation between curvature and writhing number; for the elicitin family, the value for R2 is 0.25.

Structural Conservation Analysis

The determination of the conserved residues and regions is done by applying clustering algorithms to the values describing the geometry of the protein. For CHORAL, a Modified Basic Sequential Algorithmic Scheme (MBSAS) [30] was employed for simultaneously clustering both the Cα atoms Euclidian distance and the curvature and torsion pairs, as it can be done with just one threshold. The results produced by the MBSAS are strongly dependent on the order in which the vectors are analysed and on the value of the threshold parameter. In some situations such sensitivity to the clustering parameter will cause the construction of non-optimal clusters.

In order to overcome this problem, ARABESQUE employs a Two-Threshold Sequential Algorithmic Scheme (TTSAS) [30], which defines a "grey" area between two distinct threshold. Values above the second threshold are considered non-conserved, values below the first threshold are considered conserved and anything between the two thresholds are left to be analysed later. After the first pass, any non-assigned coordinate is reassessed and assigned either to an existent set or to a new set.

For ARABESQUE we decided to use a "sieve" procedure where clusters are created in the following order:

1. Cα-Cα distances.
2. Arc length and writhing number.
3. Curvature and torsion.

The first step consists of clustering the aligned and superposed Cα atoms into sets where the distances between the elements are below 3.8Å, and thresholds of 2.0Å and 3.8Å are used. For the dissimilarity measure, the best choice for our problem is the *max proximity function*

$$D_{\substack{ps \\ \max(\vec{p},C)}} = max_{\vec{q}\in C} D(\vec{p},\vec{q})$$

as it ensures the formation of compact clusters. In this function $D(\vec{p},\vec{q})$ is the Euclidian distance between coordinate vectors $\vec{p}$ and $\vec{q}$ belonging to the cluster *C*.

During the second step, each set is then individually considered and their elements are split into sub-sets in case they are deemed not conserved by the arc length and writhing number criteria. The two thresholds used are 0.20 and 0.35 respectively and the dissimilarity measure is the *average proximity function*

$$D_{\substack{ps \\ \max(\vec{p},C)}} = \frac{1}{n_C}\sum_{\vec{q}\in C} D(\vec{p},\vec{q})$$

,

where n_C is the cardinality of *C*. As both the arc length and the writhing number describe the local structure at a medium range they can discern differences between β-strands and loops that are difficult to recognise using only curvature and torsion alone, as is done in CHORAL. The inherent flexibility of loops causes them to display a vast range of curvature and torsion pairs that often overlaps the ones displayed by β-strands. The addition the long-range information about the local structures helps to avoid the false positives during curvature and torsion clustering frequently observed in CHORAL.

For the third and final step, the sub-sets are analysed using curvature and torsion pairs and they can be split into newer sub-sets. The two thresholds for this case are 0.20 and 0.35 and the dissimilarity measure is also an average proximity function. This final set of sets contains all the information needed to describe the structural conservation across the proteins. The sets for each aligned position are analysed and collected into groups containing the same structures, called of Structurally Conserved Clusters (SCCs), as described in CHORAL [6].

The Structurally Conserved Clusters are displayed in the alignment using the same colour (Fig. 10); it means that residues shown with the same colour possess the similar conformations. As mentioned before, a JOY colour annotated alignment is also provided by ARABESQUE. In addition to the colour coded alignments, ARABESQUE also produces a 3D kinemage file (Fig. 11) of the Cα trace, coloured as in the annotated alignment, that can be displayed by KiNG program [31] and all the fragments are also outputted individually. This 3D model allows the user to explore the structural conservation at the fragment level.

The structural superposition and the Euclidian Cα-Cα distance requirements during the first clustering step enforce structural similarity dependence in the 3D space. If bypassed, the algorithm can detect similarity even when segments are not superposed, thus allowing the algorithm to detect similarity even in hinged proteins. The restricted Euclidian Cα-Cα distance is important for producing fragments for homology modelling, as in CHORAL, or for generating Probability Density Functions (PDFs) for restraint-base conformational sampling for RAPPER [35].

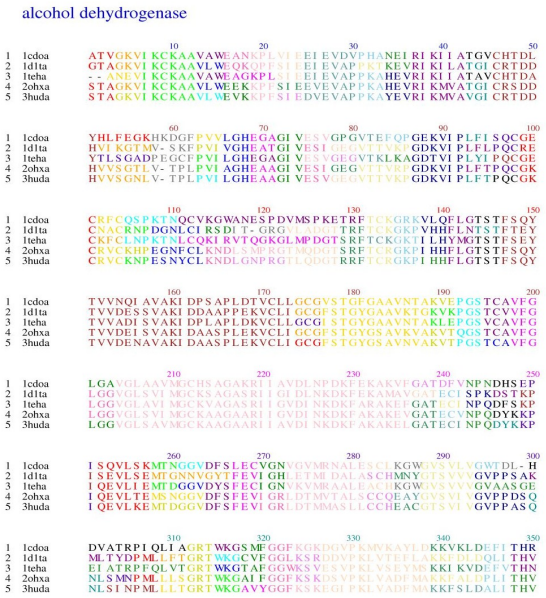


Fig. 10- Colour annotated alignment for the HOMSTRAD alcohol dehydrogenase family. Each sequence segment with the same colour has similar structure.

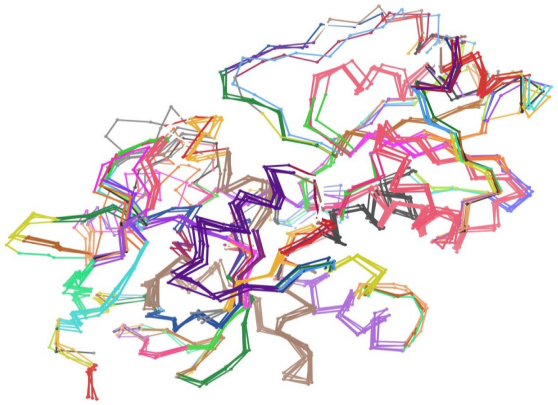


Fig. 11- Colour annotation for characterising the structural conservation in the alcohol dehydrogenase family. As in the sequence alignment, residues with similar structure are shown in the same colour. The original kinemage file can be rotated and zoomed using the KiNG program.

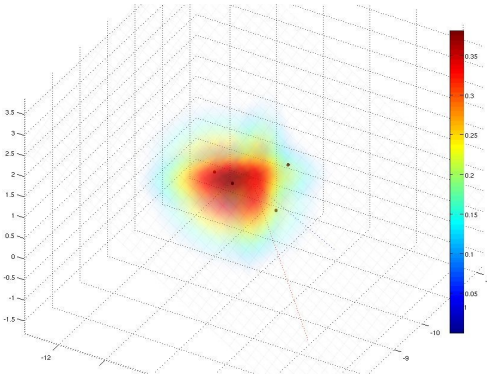


Fig. 12- Example of a Cα Probability Density Function using SCCs

There are various conceivable ways of incorporating the propensity score information and multiple templates. A straightforward and intuitive one is to combine PDFs additively while weighting them according to their relative propensities

$$p_c(x,y,z) = \frac{1}{N}\sum_{i=1}^{N} \frac{\eta_i}{\sqrt{2\pi\sigma}} e^{-\left[(x-\mu_{i,x})^2+(y-\mu_{i,y})^2+(z-\mu_{i,z})^2\right]/2\sigma^2}$$

$$\eta_i = \frac{\omega_i}{\sum_{j=1}^{N}\omega_j}$$

$$\omega_i = e^{S_i / \max_{1\le j\le N} S_j}$$

where N is the number of templates in the SCC, η is the normalised weight, ω is the weight score and *S* is the propensity score of the fragment, if the propensity score of that particular SCC is positive, or of the entire sequence, if the score is negative. The scoring system for SCCs, necessary for the implementing the Probability Density Function (PDF), has already been described elsewhere [6, 22]. Figure 12 shows an example of a Cα PDF that can be generated by data outputted by ARABESQUE when the user provides a sequence for a target protein in addition to the template structures.

Conclusion

We have described ARABESQUE, a new methodology for protein structure comparison. This approach extends and enhances the method designed for CHORAL [6]. The development of ARABESQUE was motivated by the fact that although CHORAL is very successful in many cases, it could become problematic in analysing the complex geometries often present in protein families with very low sequence similarity. The multiple layers of analysis employed by ARABESQUE are able to address the most demanding clustering situations arising from such geometric complexities. This new method provides us with several useful tools for analysing structural conservation and homology modelling.

ARABESQUE is very useful not only for analysing the structural conservation of a given protein family, but also as a tool for checking the quality of structural and sequence alignments. The curvature, torsion, arc length and writhing number graphics, in conjunction with the 3D kinemage, can be used for locating misaligned residues and fragments. ARABESQUE also allows users to make corrections to the alignment and to produce a new conformational analysis. By comparing the new SCCs with the old ones, the user is able to verify if the modification has decreased the total fragmentation of the alignment and, therefore, reduced its information entropy.

We anticipate that it will be possible to use ARABESQUE as the basis for a new hybrid modelling program that combines information derived from the probability density functions with experimental data such as NMR chemical shifts [32], residual dipolar couplings [33] and low-resolution electron density maps [34]. We intend to combine restrained comparative modelling, as done for RAPPER ([35-37]), and NMR restrained molecular dynamics simulations [38] into a system able to deal with the problem of computing the entropic changes in free energy during flexible protein-ligand docking.

Acknowledgment

We would like to thank Simon Lovell and Kenju Mizuguchi for the useful discussion during the early work on this project. HTAL would like to thank Cambridge Overseas Trusts and St John's College, Cambridge for financial support.

References

[1] Goldsmith-Fischman S., Honig B. (2003) *Protein Science* 12, 1813-1821.
[2] Murzin A.G., Brenner S.E., Hubbard T., Chothia C. (1995) *Journal of Molecular Biology* 247(4), 536-540
[3] Bajaj M. and Blundell T. (1984) *Annual Review of Biophysics and Bioengineering* 13, 453-492.
[4] Illergard K., Ardell D.H., Elofison A. (2009) *Proteins-Structure Function and Bioinformatics* 77(3), 499-508.
[5] Fiser A. and Sali A. (2003) *Macromolecular Crystallography*, Pt D 374, 461-+.
[6] Montalvao R.W., Smith R.E., Lovell S.C., Blundell T.L. (2005) *Bioinformatics* 21(19), 3719-3725.
[7] Smith R.E., Lovell S.C., Burke D.F., Montalvao R.W., Blundell T.L.(2007) *Bioinformatics* 23(9), 1099-1105.
[8] Dolan M.A., Keil M., Baker D.S. (2008) *Proteins-Structure Function and Bioinformatics* 72(4), 1243-1258.
[9] Fain B. and Rogen P. (2003) *National Academy of Sciences of the United States of America* 100(1), 119-124.
[10] Louie A.H. and Somorjai R.L. (1982) *Journal of Theoretical Biology* 98(2), 189-209.
[11] Louie A.H. and Somorjai R.L. (1983) *Journal of Molecular Biology* 168(1), 143-162.
[12] Louie A.H. and Somorjai R.L. (1984) *Bulletin of Mathematical Biology* 46(5-6), 745-764.
[13] Rackovsky S. and Scheraga H.A. (1978) *Macromolecules* 11 (6), 1168-1174.
[14] Rackovsky S. and Scheraga H.A. (1980) *Macromolecules* 13 (6), 1440-1453.
[15] Rackovsky S. and Scheraga H.A. (1981) *Macromolecules* 14 (5), 1259-1269.
[16] Can T. and Wang Y.F. (2003) *IEEE Bioinformatics Conference*, 169-179.
[17] Rogen P. and Bohr H. (2003) *Mathematical Biosciences,* 182 (2), 167-181.
[18] Rogen P. and Sinclair R. (2003) *Journal of Chemical Information and Computer Sciences* 43(6), 1740-1747.
[19] Rogen P. (2005) *Journal of Physics-Condensed Matter* 17 (18), S1523-S1538.
[20] Nielsen B.G., Rogen P., Bohr H.G. (2006) *Mathematical and Computer Modelling* 43(3-4), 401-412.
[21] Chang P.L., Rinne A.W., Dewey T.G. (2006) *BMC Bioinformatics* 7.346
[22] Deane C.M., Kaas Q., Blundell T.L. (2001) *Bioinformatics,* 17 (6), 541-550.
[23] Sali A. and Blundell T.L. (1990) *Journal of Molecular Biology* 212(2), 403-428.
[24] Mizuguchi K., Deane C.M., Blundell T.L., Johnson M.S., Overington J.P. (1998) *Bioinformatics* 14(7), 617-623.
[25] do Carmo M. (1976) Differential Geometry of Curves and Surfaces.
[26] Broucke R. (1973) *Communications of the Acm.* 16(4), 254-

256.
[27]Mizuguchi K., Deane C.M., Blundell T.L., Overington J.P. (1998) *Protein Science* 7(11), 2469-2471.
[28]de Bakker P.I.W., Bateman A., Burke D.F., Miguel R.N., Mizuguchi K., Shi J., Shirai H., Blundell T.L. (2001) *Bioinformatics* 17(8), 748-749.
[29]Stebbings L.A. and Mizuguchi K. (2004) *Nucleic Acids Research* 32, D203-D207.
[30]Theodoridis S. and Koutroumbas K. (2006) Pattern recognition.
[31]Chen V.B., Davis I.W., Richardson D.C. (2009) *Protein science* 18(11), 2403-2409.
[32]Kohlhoff K.J., Robustelli P., Cavalli A., Salvatella X., Vendruscolo M. (2009) *Journal of the American Chemical Society*, 131 (39), 13894-95
[33]Zweckstetter M. (2008) *Nature Protocols*, 3(4), 679-690.
[34]DePristo M.A., de Bakker P.I.W., Johnson, R.J.K., Blundell, T.L. (2005) *Structure*, 13(9), 1311-1319
[35]Furnham N., de Bakker P.I., Gore S., Burke D.F., Blundell T.L. (2008) *BMC Structural Biology* 8, 7.
[36]Furnham N., Dore A.S., Chirgadze D.Y. Bakker P.I.W., DePristo M.A., Blundell T.L. (2006) *Structure* 14(8), 1313-1320.
[37]Karmali A.M., Blundell T.L., Furnham N. (2009) *Acta crystallographica. Section D, Biological crystallography* 65(Pt 2), 121-127.
[38]De Simone A., Montalvao R.W., Vendruscolo M. (2011) *Journal of Chemical Theory and Computation*.