5

An Exhaustive Shape-Based Approach for Proteins' Secondary, Tertiary and Quaternary Structures Indexing, Retrieval and Docking

Eric Paquet and Herna L. Viktor
National Research Council & University of Ottawa, Canada

1. Introduction

Over the past ten years, the number of three-dimensional protein structures has grown exponentially (Holm, 2008). This is due, mainly, to the advent of high throughput systems. Consequently, molecular biologists need systems to enable them to effectively store, manage and explore these vast repositories of three-dimensional structures. They want to determine if an unknown structure is in fact a new one, if it has been subjected to a mutation, and/or to which family it possibly belongs. Furthermore, they require the ability to find similar proteins in terms of functionalities. Importantly, they aim to find docking sites. That is, they aim to determine the possible sites for the binding of two proteins, namely the ligand and the receptor, in order to form a stable complex. This similarity in functionality, and specifically the task to find docking sites, are related to outer the shape of the protein (Binkowski & Joachimiak., 2008). The outer shape (or envelope), in part, determines whether two proteins may have similar functionalities and may thus aid us to determine the location of such protein binding sites. The previously introduced docking problem may be better understood from the perspective of drug design. Most diseases and drugs work on the same basic principle. When we become ill, a foreign protein docks itself on a healthy protein and modified its functionality. Such a docking is possible if the two proteins have two sub-regions that are compatible in terms of three-dimensional shape, a bit like two pieces of a puzzle. Drugs are designed to act in a similar way. Namely, a drug docks on the same active site and prevents the docking of foreign proteins which can potentially cause illness (Paquet and Viktor, 2010).

Consequently, one of the main objectives of macromolecular docking, also known as protein-protein docking, is to find compatible regions in between two proteins from a geometrical point of view; the better the fit, the better the efficiency of the docking. There are currently many approaches that have been developed in order to simulate in silico, i.e. with a computer and with algorithms, this docking of two macromolecules. We shall briefly review them in order to enhance their most salient features, their potential and their main weaknesses. We shall limit ourselves to the most recent works, which at any rate, capture most previous advances.

This chapter is organized as follow. In Section 2, approaches based on correlation techniques are presented, while in Section 3 various methods based on spherical harmonics and orthogonal polynomial are reviewed. Section 4 briefly describes shape signatures based on curvature invariants and Section 5 provides an overview of distance-based techniques. In Section 6, techniques based on alignment and Monte Carlo simulations are explored while the specific contribution of computer graphics is outlined in Section 7. Section 8 completes this chapter by presenting some high throughput methods. The main conclusions are presented in Section 9.

2. Correlation-based techniques

One of the earliest approaches for macromolecule docking is based on finding correlations. Such methods aim to determine the best alignment, in terms of translations, in between two macromolecules. In order to evaluate the correlation, a representation or shape must be associated with each macromolecule. Such a representation may take various forms, amongst which the binary and the volumetric representations are the most common. Then, given a representation, the product of their shapes is evaluated for each relative translation in order to determine their relative overlapping. Similar regions, that is, regions that present a substantial overlap, are usually characterized by a high value of their correlation. The position of the correlation peak determines the relative translation in between the two macromolecules. Because the relative translation is unknown, the translation space must be searched exhaustively. In order to be more efficient from a computational point of view, the correlation is implemented with the Fourier transform. To further reduce the complexity of the calculations, the Fourier transform is evaluated with the fast Fourier transform algorithm.

Many methods distinguish themselves by the representation they use in order to describe the proteins. For instance, (Katchalskip-Katzir et al., 1992) use a binary representation for the macromolecule. Here, the points inside the boundary and on the boundary have a value of one, while external points have a value of zero. Such a representation tends to be oversensitive and only allows for exact marching. In an attempt to further speed up the calculations, (Nukuda et al., 2007) implement the correlation on an IBM Blue Gene supercomputer. By also taking the rotations into account, they sample the rotation space and compute a separate correlation for each sampled rotation. Such an exhaustive approach does not scale well, even with a supercomputer. A similar approach is used by (Kasakov et al., 2006), with the difference being that the binary representation is replaced by a continuous one, which is obtained from the molecular potentials associated with the constituent atoms. Their method provides a representation both in terms of shape and physicochemical properties and offers more robustness, since the potentials are real valued. Nevertheless, the scalability problem remains unsolved. Sukhwani et al. (Sukhwani et al., 2008) also use more sophisticated molecular potentials and parallelize the calculation of the correlation. However, the problems associated with an exhaustive search in the rotation space are still untargeted. In order to address this problem, (Vadja & Kozakov, 2009) and (Gray, 2006) propose to replace the exhaustive sampling of the rotation space by a Monte Carlo optimization of a cost function associated with the three-dimensional rotation.

Consequently, a transition probability is associated with the process. Rotations are assimilated to a random walk and are accepted, or rejected, according to a Metropolis

criterion based on the transition probability. Gray, in particular, uses the simulated annealing algorithm in order to optimize the cost function. Such Monte Carlo approaches allow one to explore efficiently, and parsimoniously, the rotation space and are consequently much more scalable. The rotation may also be obtained directly in a non iterative way. For instance, (Katebi et al., 2009) determine the rotation in between two macromolecules, with the singular value decomposition (SVD) based on the relative position of the constituent carbon atoms. This approach leads to an oversimplification of the shape, which is not necessarily compatible with docking. Here, SVD has been selected against principal component analysis (PCA) because of its statistical robustness.

Finally, (Bonvin et al., 2006) present general considerations, applicable to all these methods, about the flexibility of proteins. It is well known that most macromolecules are flexible. This flexibility must be taken into account when simulating docking. To this end, Bonvin et al. propose to use various conformations of the same protein, as obtained from nuclear magnetic resonance (NMR), and to repeat the calculations for each one of them. This approach requires a large amount of experimentally determined structures in order to provide realistic results. This, in turn, involves an exhaustive search which eventually does not scale well. Furthermore, NMR is currently limited to relatively small macromolecules. Bonvin et al., also propose to parameterize the deformations of the macromolecules. However, their parameterization appears to be relatively arbitrary and their proposal thus would need further evaluation to determine the feasibility and suitability thereof.

3. Angular basis, spherical harmonics and orthogonal polynomials

In order to address the above-mentioned problems associated with the correlation-based techniques, methods based on spherical harmonics and orthogonal polynomials have been developed. The aim, here, is to provide a representation more adapted to rotation. One of the seminal works was presented by Canterakis (Canterakis, 1994).

Sael & Kihara (Sael & Kihara, 2010a) use a coordinate system centred on the barycentre of the macromolecules. They then project each one of them on an orthogonal basis, formed by the product of a spherical harmonic and an orthogonal polynomial which is, in this particular case, the Zernike polynomial. The resulting basis is known as the 3D Zernike polynomial. Then, rotation invariants are calculated from this representation. The invariants associated with each macromolecule are compared to each other using the Euclidian distance. Although relatively efficient, this approach is limited to global comparison and is not suitable for docking in its present form. Based on their similar work, (Sael & Kihara, 2010b) have extended their approach to docking. Pockets are extracted from the macromolecules and are individually describes in terms of the 3D Zernike polynomials, from which rotation invariants are calculated. In order to determine a docking region, each pair of invariants, that is, each “pair of pockets”, must be independently compared. This limits the scalability of the method. Furthermore, it has been pointed out that pockets are far from being the only candidates for docking. This means that many potentially interesting docking regions are potentially ignored by the algorithm. Consequently, (Sael & Kihara, 2010d) expanded their work by randomly select an ensemble of small regions on the macromolecules. Each region is represented in terms of 3D Zernike polynomials, from which a set of rotation invariants are calculated. In order to alleviate the scalability problem which occurs when each pair of invariants is compared, they cluster the invariants with a

self organizing map, also known as the Kohonen map. Here, the number of clusters are not predetermined but determined by the algorithm itself. In order to avoid the problems associated with the border effect, they use a boundless Kohonen map with either a spherical or a toroidal topology. Being unsupervised, the Kohonen map may lead to nonsensical results. Furthermore, the 3D Zernike polynomials are more adapted to closed shapes like macromolecules than to an open regions or patches. Finally, (Sael et al., 2008) study the impact of the metric when comparing the rotation invariants. They compare the Euclidian distance, the Manhattan distance and the correlation coefficient, which all seem to provide comparable results in this particular case.

Venkatraman et al. (Venkatraman et al., 2009) also use a global approach, based on 3D Zernike polynomials. The shape is represented by a mixture of Gaussians centred on the constituent carbon atoms. This analytical representation seems relatively adapted to represent the shape of a macromolecule. Nevertheless, the limitations associated with the global approach of Sael et al. remain. Mak et al. (Mak et al., 2008) also employ a global approach similar to Sael et al. However, in contrast with Venkatraman et al., they use a binary voxelised representation for the shape of the macromolecule, which presents a lower expressive power.

Sovic et al. (Sovic et al., 2010) propose a different approach based on the 3D Zernike polynomials. Each macromolecule is represented with a subset of 3D Zernike polynomials. As pointed out earlier, these polynomials are particularly suited for closed shapes presenting a certain degree of spherical symmetry, as many macromolecules do. Then, each pair of macromolecules is compared in that particular representation. Translation invariance is obtained from a standard correlation while rotation invariance is obtained by exhaustively searching the rotation space. Sovic et al. claim that it is computationally less expensive to perform the alignment in the 3D Zernike polynomials representation than it is in the standard Euclidian representation. However, they do not provide evidence that their approach is scalable.

Scalability is a particularly acute issue for docking. There are currently more than 76,000 (28 October 2011) macromolecular structures that have been experimentally determined and which may be found in the Protein Data Bank (PDB) (Dutta et al., 2009). Let us assume, for instance, that we partition each structure in a thousand of structural regions or patches. Then, evaluating any potential match in between two patches would involve 2,812,499,962,500,000 comparisons, if one is to consider all the combinations of two patches selected from a set of 75,000,000 patches. Of course, this is an oversimplification of the problem, but it shows that the importance of the scalability should not be underestimated.

Furthermore, (Ritchie, 2005; Mavridis & Ritchie, 2010) propose a new approach which follows Canterakis' scheme. Here, the shape is represented in terms of spherical harmonics for the angular part and in terms of orthogonal polynomials for the radial part. However, they replace the Zernike polynomials by either the spherical Bessel Gaussian-type orbitals (GTO) which is formed by the product of a Gaussian and a Laguerre polynomial, the exponential type orbitals (ETO) which is formed by the product of an exponential and of a Jacobi polynomial or the Bessel type orbitals. They use various representations for the shape of the macromolecule, including the electronic density. These orthogonal polynomials seem to be more adapted to the geometry of macromolecules than the Zernike polynomial, although no definitive conclusions may be drawn at this stage.

4. Curvature invariants

Geppert et al. (Geppert et al., 2010) extract local curvature invariants, in terms of rigid transformations, from the envelope. They use geometric hashing in order to structure the invariants and in order to compare them efficiently. Nevertheless, geometric hashing is notoriously known for its lack of scalability and robustness. Furthermore, the expressive power of local curvature invariants is relatively low. Geppert et al. claim that the local curvature invariants are oblivious against a change of conformation (robustness against the macromolecule flexibility). This is certainly true for the regions where there is a limited bending or deformation, but it does not hold in general. Ranganath et al. (Ranganath et al., 2007) also extract local curvature invariants, in terms of rigid transformations, from the envelopes of the macromolecules. The calculation of the curvature invariant is based on the wavelet transform. Although the calculation of the curvature invariant is more robust when performed with the wavelet transform, the limitations inherent to these invariants remain.

5. Distance-based and graph-based methods

Liu et al. (Liu et al., 2009) propose to sample the envelope of each macromolecule with a certain number of landmarks. The landmarks are first obtained by randomly sampling points on the surface of the macromolecule and then by clustering them in order to obtain an informative sample. Then, the distances in between each pair of landmarks are calculated and their probability distribution is determined. Here, the distance is defined has the shortest path in between two landmark points within the macromolecular shape. This distance is also known as the geodesic distance which can be obtained, for instance, with the Dijkstra algorithm (Dijkstra, 1959). This probability constitutes a unique geometrical signature associated with the global shape of the macromolecule. The distance, as previously defined, is invariant under rigid transformations, invariant to isometries and relatively oblivious to a small set of general transformations.

Chi (Chi, 2004) follows an approach inspired by image indexing. He computes the distance in between the constituent carbon atoms associated with each macromolecule and creates a distance matrix. Such a matrix is invariant under rigid transformations. Then, the texture associated with the distance matrix is analyzed in terms of uniformity, entropy, homogeneity, contrast, correlation and cluster tendency. These measures constitute the geometrical signature of the macromolecule. It should be noted that the carbon atoms provide a very low resolution representation of the shape of the macromolecule. This approach has been extended by (Chen et al., 2010). They propose to analyze the texture associated with the distance matrix in terms of principal component analysis (PCA). Although more efficient from an information theory point of view, this approach still provide a very low resolution representation of the global shape of the macromolecule.

Furthermore, (Zhang et al. 2009), (Peng & Tsay, 2010), (Novosad et al., 2010) and (Reddy et al., 2011) also use the relative position of the constituent carbon atoms of the macromolecules. The distance matrix is replaced by a graph associated with the topology of the carbon atoms. It should be noted that the graph is highly sensitive to the original topology of the carbon atoms. Such a graph is invariant under rigid transformation. By using partial matching of graphs, they evaluate possible docking configurations. Nevertheless, the expressive power of a graph based on the topology of the carbon atoms is relatively low. Furthermore, extensive partial graph matching may lead to an excessive computational complexity.

Borgwardt et al. (Borgwardt et al., 2005) follow the same approach then in (Zhang et al., 2009; Peng & Tsay, 2010; Novosad et al., 2010; Reddy et al., 2005) for the construction of the graphs associated with the macromolecules. The graphs are compared with a graph kernel which probes the graphs with a random walk. Although the limitations previously outlined remain, it is interesting to note that there is a strong fundamental connection in between the shape and the random walk, which should be further explored. Such a connexion might lead to better metrics for shapes comparison.

6. Alignment and Monte Carlo based methods

In their respective papers, (Wang et al., 2007) and (Hoffmann et al., 2010) propose to find the best global alignment, in terms of rigid transformations, in between two macromolecules. Wang et al. define a cost function which evaluates the local discrepancy in between the shape of two macromolecules. Then, they minimize the cost function in terms of translation and rotation in order to find the best alignment. The rotations are either represented in terms of matrices or in terms of quaternions. Algorithms based on the later are often computationally more efficient. In order to avoid local minima, a Monte Carlo approach is used for the optimization. A transition probability function is defined which allows to explore efficiently and parsimoniously the transformation space and to escape local minima as required during the exploration stage. Wang et al.'s optimization approach is based on the simulated annealing technique, which mimics the physical optimization of the crystalline structure associated with a solid during his cooling phase. Hoffmann et al. propose to represent the shape of the macromolecule with a mixture of Gaussians centred on the constituent carbon atoms and to perform a pre-alignment of each pair of macromolecules with principal component analysis (PCA). The pre-alignment facilitates the convergence of the simulated annealing algorithm toward the global minimum and reduces the computational complexity, since PCA is a linear algorithm. Both methods are relevant for precise global comparison and have some degree of relevance for docking.

7. Computer graphics based methods

In addition to the previous methods based on invariant patterns recognition, graph theory and computer vision, methods originating from computer graphics have also been suggested. For instance, (Zauhar, 2003) proposes to randomly sample a set of points on the envelope of a macromolecule. Next, for each one of these points, he propagates a virtual optical ray inside the envelope up to a predetermined number of reflexions. The probability distribution associated with the length of the rays constitutes a geometrical signature for the global shape of the macromolecule. Such a representation is invariant under rigid transformations. It is unlikely that such a method may be applied for docking, because the patches involve in docking are not closed shapes and consequently are not suitable for virtual light ray propagation: most light rays would escape the patch without any reflexion.

Shapira et al. (Shapira et al., 2008) propose a robust definition of the diameter for a closed object called the shape diameter function (SDF). Given a point on the surface of a closed object, a cone is constructed. The apex of the cone is the given point and the axis of the cone is the inner normal associated with the given point. The opening angle for the cone is typically 120 degrees. Then, the SDF is defined as the weighted average of the lengths of the optical rays that propagate from the apex to the closest point on the surface of the shape

under the constraint that they remain inside the cone. The weight is defined as the inverse of the angle in between the ray and the axis of the cone.

The SDF may be defined for each point of the object or for a given subset. In order to describe the shape, a histogram of its associated SDFs is constructed which constitutes its geometric signature. With the aim of having an analytical form for the histogram, a mixture of Gaussians is constructed from the later. Such a signature is invariant under rigid transformations and oblivious under more general transformations. The degree of obliviousness is far from clear, though. Furthermore, the signature seems to be deficient from an expressiveness point of view. With the intention of addressing the expressiveness problem, (Gal et al., 2007) define two bidimensional histograms which encapsulate additional information about the shape. This is done so as to obtain a better discrimination in between shapes. Given a set of points on the surface of the object, the first histogram is constructed from the SDF at each point and from the distance in between each point and the barycentre. The second histogram is constructed from the SDF and from the centricity function calculated at each point. Given a point on the surface of the object, the centricity function (CF) is defined as the average geodesic distance from that point to all the other points in a given neighbourhood. It should be noted that the CF is invariant under isometry. Gal et al.'s approach seems to provide better discrimination. In addition, the SDF appears to provide a natural partition or segmentation of simple shapes, in the sense that points belonging to the same part tend to have a comparable value for their SDF.

Fang et al. (Fang et al., 2009) have applied the SDF to the description of macromolecular shapes. The points forming the envelope of the protein are sampled by clustering them with the K-mean algorithm, with the purpose of obtaining a more informative sample as opposed to random sampling. For each point belonging to the sample, they compute the SDF. Finally, they construct a histogram of the later. This histogram constitutes the global geometry signature for the envelope. Signatures and consequently shapes are compared with the Euclidian distance.

Lo et al. (Lo et al., 2010) propose to describe the envelope of macromolecules with solid angles. The solid angles are computed for each point of the envelope. Then, the surface is partitioned, or segmented, by clustering the solid angles based on the observation that points belonging to a given pocket tend to have a comparable solid angle. In order to speed up the calculation of the solid angles, Lo et al. take advantage of the graphical processing unit (GPU) and of the Compute Unified Device Architecture (CUDA) parallel technology. Docking candidates are found by exhaustively matching the various patches. The expressiveness power of the solid angle is very low. Also, the algorithm, in its present form, may only detect certain types of pockets, while ignoring a large proportion of potentially interesting docking sites.

8. High throughput methods

Shibberu & Holder (Shibberu & Holder, 2011) present an approach for proteins alignment. They compute the distance in between each pair of constituent carbon atoms and construct a distance matrix. A continuous cut-off function is applied to the distance matrix, in order to obtain a contact matrix. The quadratic form constructed from this contact matrix defines a generalized inner product. With the help of the Eigen decomposition of the contact matrix, it is possible to associate a unique Eigen value to each residue or carbon atom. A scoring matrix, which measures the distance in between each pair of Eigen values associated with

two distinct proteins, is constructed in order to evaluate the quality of the alignment. The carbon atoms provide only a low resolution description of the surface of the protein. The assignment of an Eigen value to a specific residue might be subject to a certain level of ambiguity and might be computationally unstable, for instance, when some Eigen values are very similar.

Hue et al. (Hue et al., 2010) propose a method based on support vector machine (SVM) which aims to address the scalability problem. Instead of predicting the docking regions, they try to predict the outcome of the docking in between two macromolecules. The SVM algorithm determines the optimal partition, in terms of a hyperplane, in between two classes: the interacting macromolecules and the non-interacting ones. The SVM is trained with a training set of macromolecules for which the presence or the absence of interaction is known experimentally. For each pair of macromolecules in the training set, an alignment is performed and a feature vector, here a similarity matrix, is associated with the quality of the alignment. These feature vectors span a vector space which is partitioned by the SVD with a hyperplane into two classes, namely interacting and non-interacting macromolecules. Once the training is completed, the presence of the absence of interaction in between two proteins may be determined by aligning them. This is accomplished by computing the feature vector associated with the quality of the alignment and by determining the region, or class, to which the feature vector belongs. In order to improve the efficiency of the method, the metric used to quantify the alignment, the Euclidian inner product, is replaced by a nonlinear kernel. This allows one to improve the classification accuracy by introducing a non linearity in what is, otherwise, a linear partition of the classification space. Further, a singular value decomposition (SVD) of the similarity matrix is performed, in order to reduce the complexity associated with the evaluation of the nonlinear kernel. The method may contribute to reduce the computational complexity. However, the precision-recall curves tend to indicate that there is a very high proportion of misclassification involved in the process. In addition, a global alignment is far from being the most efficient method to characterize the degree of interaction in between two macromolecules. Nevertheless, a similar approach could be followed if the alignment is replaced by a more efficient geometrical signature. The limitations, associated with a hyperplane, may be overcome by replacing the SVP with a Bayesian technique (Rodriguez & Schmidler, 2009).

Finally, (Paquet & Viktor, 2010; Paquet & Viktor, 2011) perform a virtual fragmentation of the macromolecules. For each fragment, they estimate a translation and rotation invariant intrinsic reference frame which is statistically optimal in terms of shape. Then, they evaluate various probability distributions associated with the spatial distribution of the elementary surface elements forming the fragments. A geometrical signature is constructed from these probability distributions. This signature is invariant under rigid transformation and oblivious under more general transformations. The signatures are either exhaustively compared pair-wise, with a Euclidian metric, or clustered with a K-means algorithm in order to reduce the computational complexity. The method may be applied to both local and global analysis. It is compatible with various representations including the envelope, the constituent carbon atoms and the van der Waal representation, among others. The method is scalable and a search may be performed in a database of a few hundred thousand of fragments or macromolecules, in approximately a second on a portable computer. A typical query against the 1tyv protein is shown in Fig. 1.

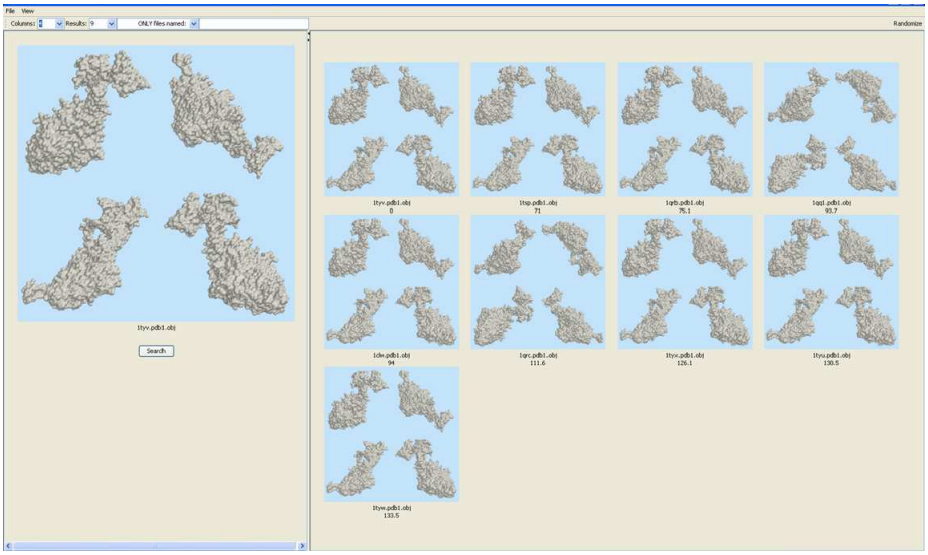


Fig. 1. Retrieval of all members from the P22 tailspike protein from Salmonella phage P22 conformation, using the 1tyv structure as a query, with a precision 100% and recall 100% from a database of more than 35,000 structures.

9. Conclusions

In this section, we summarize our main conclusions about the various algorithms. Further, we identify some potentially important research directions and challenges. Docking is by no means a trivial problem. That is, one has to identify two regions on two complex shapes which may present intricate deformations, while possessing very little a priori knowledge about the location, shape and deformations associated with these regions. The situation is further complicated by the combinatorics of the search. In order to find candidates for docking, one first has to search through large databases of macromolecular structures. Subsequently, for each potentially interesting pair of proteins, a large combination of regions, or docking sites, has to be considered.

We may only conclude that, unless one has an important a priori knowledge about the interaction, it is important to develop approaches that may rapidly determine if two proteins might, or not, interact independently of the details of their interaction. Such a binary classification can be obtained, for instance, with machine learning methods based on support vector machine (SVM) as well as their Bayesian generalization, relevance vector machine (RVM), which provides more robust probabilistic classification (Damoulas et al., 2008).

Because of the large number of regions that are compared, it is important to develop efficient methods to do so. Large scale clustering and constrained clustering (Basu et al., 2008) methods might be able to target such a problem. Constrained clustering, for instance, is particularly adapted to enforce physicochemical constraints.

Macromolecules are not rigid, but deformable entities. As we saw earlier, some methods, up to a certain extent, take this fact into account. It may come as a surprise, that so few methods

do, indeed, take these deformations into account. Such a situation might originate from the fact that most structures are obtained from X-ray crystallography which implicitly implies rigidity: this is indeed the case for the crystal, but it is not the case for the protein, in vivo. The present models are not entirely adequate, in order to describe the deformations associated with proteins. Instead, one should be able to distinguish in between two types of transformations: the deformations on the surface of a protein and the flexibility associated with its backbone or carbon atoms chain(s). The later is far from arbitrary, since the bending occurs at a small and specific number of sites or hinges. Consequently, a protein behaves like an articulated object. Following this line of thought, (Rodriguez & Schmidler, 2009) have developed a Bayesian approach which takes this phenomenon into account and which allows to align the backbone of two proteins. They determine a prior for both the number of hinges and their location while, for the likelihood, they associate either a rigid or an affine transformation to each articulated segment of the backbone.

Finally, (Raviv et al., 2010) have shown that, even in the case where an invariance has been obtained against a large group of deformations, as in their case for isometries, the lack of constraints might lead to nonsensical results. Indeed, although possible, from an invariance point of view, some deformations should not be taken into account. This is because of their nonexistence in nature. For example, a protein cannot be flattened, although such a deformation may be allowed under certain groups of transformations. The authors of this chapter foreseen that a volumetric approach might resolve the problem. For macromolecules, a volumetric representation might be obtained, for instance, from the electrostatic, or from the van der Waal potential, associated with each constituent atom.

10. References

Binkowski, T. A. & Joachimiak, A. (2008). Protein Functional Surfaces: Global Shape Matching and Local Spatial Alignments of Ligand Binding Sites. *BMC Structural Biology*, Vol. 8, 23 pp.

Holm, L. et al. (2008). Searching protein structure databases with DaliLite v.3. Bioinformatics, Vol. 24, pp. 2780–2781

Vajda, S. & Kozakov, D. (2009). Convergence and combination of methods in protein-protein docking. *Current Opinion in Structural Biology*, Vol. 19, pp. 164-170

Katchalskip-Katzir, E. et al. (1992). Molecular surface recognition: Determination of geometric fit between proteins and their ligands by correlation techniques. *Proc. Natl. Acad. Sci.*, Vol. 89, pp. 2195-2199

Kazakov D. et al. (2006). PIPER: An FFT-Based Protein Docking Program with Pairwise Potential. *Proteins: Structure, Function, and Bioinformatics*, Vol. 65, 392-406

Nukuda, A. et al. (2007). High Performance 3D Convolution for Protein Docking on IBM Blue Gene. *The Fifth International Symposium on Parallel and Distributed Processing and Applications (ISPA)*, Springer LNCS 4742, pp. 958-969, Niagara Falls, ON, Canada, August 29-31, 2007

Katebi A. R. et al. (2009). Computational Testing of Proteins-Protein Interactions. *Bioinformatics and Biomedicine Workshop, IEEE Int. Conf. on Bioinformatics and Biomedicine*, pp. 144-151, Washington, D. C., USA, November 1-4, 2009

Sukhwani, B. & Herbordt, M. C. (2008). Acceleration of a Production Rigid Molecule Docking Code. *Int. Conf. on Field Programmable Logic and Applications*, pp. 341-346, Heidelberg, Germany, September 8-10, 2008

Gray, J. J. (2006). High-resolution protein-protein docking. *Current Opinion in Structural Biology*, Vol. 16, pp. 183-193

Bonvin, A. M. J. J (2006). Flexible protein-protein docking, *Current Opinion in Structural Biology*, Vol. 16, pp. 194-200

Sael, L. & Kihara, D. (2010a). Improved protein surface comparison and application to low-resolution protein structure data. *Bioinformatics*, Vol. 11 (Suppl. 11): S2, 12 pp.

Sael, L. & Kihara, D. (2010b). Binding Ligand Prediction for Proteins Using Partial Matching of Local Surface Patches. *Int. J. Mol. Sci.*, Vol. 11, pp. 5009-5026

Sael, L. & Kihara, D. (2010c). Characterization and Classification of Local Protein Surfaces Using Self-organizing Map. *International Journal of Knowledge Discovery in Bioinformatics*, Vol. 1, pp. 32-47

Venkatraman, V. et al. (2009). Protein-protein docking using region-based 3D Zernike descriptors. *BMC Bioinformatics*, Vol. 10, 21 pp.

Sael et al., L. (2008). Fast protein tertiary structure retrieval based on global surface shape similarity. *Proteins*, pp. 1259-1273

Canterakis, N. (1994). 3D Zernike Moments and Zernike Affine Invariants for 3D Image Analysis and Recognition, *ESPRIT Basic Research Workshop on Visual Invariances*, 8 pp.

Mak, L. et al. (2008). An extension of spherical harmonics to region-based rotationally invariant descriptors for molecular shape description and comparison. *Journal of Molecular Graphics and Modelling*, Vol. 26, pp. 1035-1045

Sovic, I et al. (2010). Parallel Protein Docking Tool. *33rd International Convention on Information and Communication Technology, Electronics and Microelectronics (MIPRO)*, pp. 1333-1338, May 24-28, Opatija, Croatia, 2010

Ritchie, D. W. (2005). High-order analytic translation matrix elements for real-space six-dimensional polar Fourier correlations. *J. Appl. Cryst.*, Vol. 34, pp. 808-818

Mavridis L. & Ritchie, D. W. (2010). 3D-BLAST: 3D Protein Structure Alignment, Comparison, and Classification using Spherical Polar Fourier Correlations. *Pacific Symposium on Biocomputing*, Vol. 15, pp. 281-292, January 4-8, Kamuela, Hawaii, USA, 2010

Geppert, T. et al. (2010). Protein-Protein Docking by Shape-Complementarity and Property Matching. *Journal of Computational Chemistry*, Vol. 31, pp. 1919-1928

Liu, Y.-S. et al. (2009). IDSS: deformation invariant signatures for molecular shape comparison. *BMC Bioinformatics*, Vol. 10, 14 pp.

Chi, P.-H. (2004). A Fast Protein Structure Retrieval System Using Image-Based Distance Matrices and Multidimensional Index. *Fourth IEEE Symposium on Bioinformatics and Bioengineering (BIBE)*, pp. 522-529, May 19-21, Taichung, Taiwan, 2004

Chen, Y. et al. (2010). 2nd PCA on 3D Protein Structure Similarity. *IEEE Fifth International Conference on Bio-Inspired Computing: Theories and Applications (BIC-TA)*, pp. 253-257, September 8-10, Liverpool, United Kingdom, 2010

Zhang, T. et al. (2009). A Graph-Based Approach for Protein-Protein Docking, *2nd International Conference on Biomedical Engineering and Informatics (BMEI)*, pp. 1-8, October 17-19, Tianjin, China, 2009

Peng, S.-L. & Tsay, Y.-W. (2010). Measuring Protein Structural Similarity by Maximum Common Edge Subgraphs. *Sixth International Conference on Intelligent Computing (ICIC)*, Springer LNAI 6216, pp. 100-107, August 18-21, Changsha, China, 2010

Novosad, T. et al. (2010). Searching Protein 3-D Structures for Optimal Structure Alignment Using Intelligent Algorithms and Data Structures. *IEEE Transactions on Information Technology in Biomedicine*, Vol. 14, pp. 1378-1386

Borgwardt, K. M. et al. (2005). Protein function prediction via graph kernels. *Bioinformatics*, Vol. 21, pp. i47-i56

Reddy, A. S. et al. (2011). Analysis of HIV Protease Binding Pockets Based on 3D Shape and Electrostatic Potential Descriptors. *Chem. Biol. Drug Des.*, Vol. 77, pp. 137-151

Wang, C. et al. (2007). Protein-Protein Docking with Backbone Flexibility. *J. Mol. Biol.,* Vol. 373, pp. 503-519

Hoffmann, B. et al. (2010). A new protein binding pocket similarity measure based on comparison of clouds of atoms in 3D: application to ligand prediction. *BMC Bioinformatics*, Vol. 11, 16 pp.

Zauhar, R. J. (2003). Shape Signatures: A New Approach to Computer-Aided Ligand- and Receptor-Based Drug Design. *J. Med. Chem.*, Vol. 46, pp. 5674-5690

Shapira, L. et al. (2008). Consistent mesh partitioning and skeletonisation using the shape diameter function. *Visual Comput.*, Vol. 24, pp. 249-259

Gal, R. et al. (2007), Pose-Oblivious Shape Signature. *IEEE Transactions on Visualization and Computer Graphics*, Vol. 13, pp. 261-271

Fang, Y. et al. (2009). Three dimensional shape comparison of flexible proteins using the local-diameter descriptor. *BMC Structural Biology*, Vol. 9, 15 pp.

Ranganath, A. et al. (2007). Efficient Shape Descriptors for Feature Extraction in 3D Protein Structures. *In Silico Biology*, Vol. 7, 169-174 (2007).

Lo, Y.-T. et al. (2010). Using Solid Angles to Detect Protein Docking Regions by CUDA Parallel Algorithms. *International Symposium on Parallel and Distributed Processing with Applications*, pp. 536-541, April 19-23, Atlanta, GA, USA, 2010

Shibberu, Y. & Holder, A. (2011). A Spectral Approach to Protein Structure Alignment. *IEEE/ACM Transactions on Computational Biology and Bioinformatics*, Vol. 8, pp. 867-875

Hue, M. et al. (2010). Large-scale prediction of protein-protein interactions from structures. *BMC Bioinformatics*, Vol. 11, 9 pp.

Paquet, E. & Viktor, H. L. (2010). Addressing the Docking Problem: Finding Similar 3-D Protein Envelopes for Computer-aided Drug Design, *Advances in Computational Biology*, Advances in Experimental Medicine and Biology 680, Springer, ISBN 978-1-4419-5912-6, pp. 447-454

Paquet, E. & Viktor, H. L. (2011). Multimodal Representations, Indexing, Unexpectedness and Proteins. *Twenty-fourth International Conference on Industrial, Engineering and Other Applications of Applied Intelligent Systems (IEA/AIE 2011)*, Springer Lecture Notes in Artificial Intelligence (LNAI), pp. 85-94, June 28 - July 1, Syracuse, NY, USA, 2011

Dijkstra, E. W. (1959). A note on two problems in connexion with graphs. *Numerische Mathematik*, Vol. 1, pp. 269–271

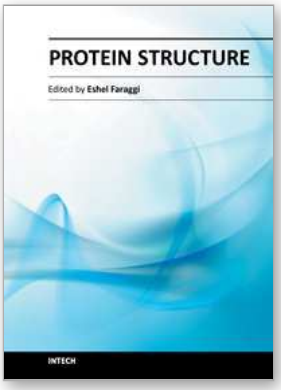
Dutta, S. et al. (2009). Data Deposition and Annotation at the Worldwide Protein Data Bank. Molecular Biotechnology, Vol. 42, pp. 1-13

Rodriguez, A. & Schmidler, S. C. (2009). Bayesian Protein Structure Alignment. Submitted to *Annals of Applied Statistics*

Basu, S. (Ed.) et al. (2008). *Constrained Clustering: Advances in Algorithms, Theory, and Applications*. Chapman and Hall/CRC, ISBB: 1584889969, Boca Raton, USA

Raviv, D. et al. (2010). Volumetric Heat Kernel Signatures. *ACM Workshop on 3D Object Retrieval (3DOR)*, pp. 39-44, October 25, Firenze, Italy, 2010

Damoulas, T. et al. (2008). Inferring Sparse Kernel Combinations and Relevance Vectors: an Application to Subcellular Localization of Proteins. *Seventh International Conference on Machine Learning and Applications(ICMLA)*, pp. 577-582, December 11-13, San Diego, USA, 2008



Protein Structure
Edited by Dr. Eshel Faraggi

ISBN 978-953-51-0555-8
Hard cover, 396 pages
Publisher InTech
Published online 20, April, 2012
Published in print edition April, 2012

Since the dawn of recorded history, and probably even before, men and women have been grasping at the mechanisms by which they themselves exist. Only relatively recently, did this grasp yield anything of substance, and only within the last several decades did the proteins play a pivotal role in this existence. In this expose on the topic of protein structure some of the current issues in this scientific field are discussed. The aim is that a non-expert can gain some appreciation for the intricacies involved, and in the current state of affairs. The expert meanwhile, we hope, can gain a deeper understanding of the topic.

How to reference
In order to correctly reference this scholarly work, feel free to copy and paste the following:

Eric Paquet and Herna L. Viktor (2012). An Exhaustive Shape-Based Approach for Proteins' Secondary, Tertiary and Quaternary Structures Indexing, Retrieval and Docking, Protein Structure, Dr. Eshel Faraggi (Ed.), ISBN: 978-953-51-0555-8, InTech, Available from: http://www.intechopen.com/books/protein-structure/an-exhaustive-shape-based-approach-for-proteins-secondary-tertiary-and-quaternary-structures-indexin

INTECH
open science | open minds

InTech Europe
University Campus STeP Ri
Slavka Krautzeka 83/A
51000 Rijeka, Croatia
Phone: +385 (51) 770 447
Fax: +385 (51) 686 166
www.intechopen.com

InTech China
Unit 405, Office Block, Hotel Equatorial Shanghai
No.65, Yan An Road (West), Shanghai, 200040, China
中国上海市延安西路65号上海国际贵都大饭店办公楼405单元
Phone: +86-21-62489820
Fax: +86-21-62489821