18

Relating Protein Structure and Function Through a Bijection and Its Implications on Protein Structure Prediction

Marco Ambriz-Rivas[1], Nina Pastor[2] and Gabriel del Rio[1]
[1]*Universidad Nacional Autónoma de México, Instituto de Fisiología Celular*
[2]*Universidad Autónoma del Estado de Morelos, Facultad de Ciencias*
México

1. Introduction

Proteins are studied by measuring different properties, typically the chemical structure and biochemical activity. Given that these measurements are done on the same protein molecule, they must be related. Despite the fact that this relationship exists, the mathematical nature of this relationship has remained elusive to our understanding, and is not commonly considered in the so called "structure-function relationship problem of proteins" (Punta & Ofran, 2008). While this is a fundamental problem in biochemistry and biology, that is, to establish a procedure that allows scientists to reliably relate protein structure and protein activity, the likelihood to succeed in this enterprise depends on our ability to understand the mere nature of this relationship. The possibility to effectively relate structure and activity has motivated years of research in different areas in biology, including biophysics, molecular biology, biochemistry, bioinformatics, and computational biology, among others. Although great advances have been achieved from these different areas of expertise, the question remains unsolved. That is, there is no general procedure that may have proven to effectively relate protein structure and activity. However, recent results in the prediction of protein three-dimensional structure (from now on referred simply as 3D structure) are addressing this problem with a fresh look, revealing a new aspect of this relationship that may explain why this particular problem has remained elusive. The present work reviews the general concepts being used to predict protein 3D structure with emphasis on the contribution of these methods to unravel the structure-activity relationship of proteins.

We divide this review in three sections. In the first section, we will present a mathematical view on the evolution of the concept about the 3D structure-activity relationship in proteins. The second section presents the general concepts behind template-based modelling and *ab initio* methods for the prediction of protein 3D structure. There, we will describe how these approaches have contributed to our current understanding of the 3D structure-activity relationship of proteins. Finally, we will review new methods for protein 3D structure prediction and how these may contribute to unravel the 3D structure-activity relationship of proteins.

2. Evolution of the 3D structure-activity paradigm from a mathematical perspective

Back in 1936 Mirsky and Pauling (Mirsky & Pauling, 1936) proposed that protein activity, or its function within a biological context, should be determined by its 3D structure. Considering that the characterization of protein activity has frequently been cumbersome, the possibility to determine it by simply looking at the 3D structure of proteins could be considered an impulse to establish this relationship. Yet, determination of protein 3D structure has not been an easy treat either. Perhaps the main motivation to establish this relationship consists in the possibility to design new devices capable of reproducing the highly efficient capabilities of proteins (Drexler, 1994; Robson, 1999; Balzani *et al*., 2000) or to simply engineer proteins in order to adapt these for industrial use (Zaks, 2001; Huisman & Gray, 2002; Straathof *et al*., 2002; Luetz *et al*., 2008). Ultimately, establishing the 3D structure-activity relationship of proteins may serve to test our level of understanding of these molecules.

Hitherto, the approximation most frequently used to solve this relationship is to consider knowledge-based classification schemes. Such schemes are based on the existence of a given set of proteins with known activity; from that knowledge, it has been possible to identify new proteins sharing similar activity, from protein sequence comparisons. Although quite useful to classify the ever-increasing number of new protein sequences generated nowadays, this type of approaches has a limited ability to assist researchers in the design of protein activity (see next section). Alternatively, the activity of a protein is commonly analyzed from the knowledge of its 3D structure using biophysical methods (Neet & Lee, 2002; Chollet & Turcatti, 1999). In either case, previous knowledge of both protein 3D structure and activity is required to establish this relationship, indicating our current limitation in understanding this problem from basic principles. Even when new enzymatic activities have been designed "from scratch" (Siegel *et al.*, 2010) the active site residues are nestled within previously known protein folds. It has been possible to design completely novel folds, such as Top7, from scratch (Kuhlman *et al*., 2003), but this refers to the sequence-3D structure relationship, which is not the main focus of this review.

We propose that one of the reasons for this limited understanding of the 3D structure-activity relationship of proteins is the absence of knowledge as to what type of mathematical relationship this one is. As we will show, determining the nature of this relationship may lead researchers to analyze this relationship with a new perspective and may accelerate the full understanding of it.

To explain this, let us first formally describe the 3D structure-activity relationship of proteins as a postulate:

P1: Protein activity depends on its 3D structure.

That is, protein activity may be represented as a mathematical relation of the protein 3D structure. Since both activity and 3D structure can always be measured on a given protein, that is they come in pairs, we postulate that this relation may be represented by a mathematical function. To further describe this postulate, let us define:

D1: Protein activity is defined as the capacity of proteins to interact with other molecules resulting in a change (on the interacting molecule or the environment) that is measurable (e.g., the chemical transformation of glucose to glucose 6-phosphate).

D2: Protein 3D structure is defined by two sets: the set of amino acid residues included in the protein and the set of physical interactions between these residues in the 3D space.

D3. A mathematical function is a particular class of relation between sets and it describes the dependence between the elements of these sets: an independent variable (an element in one of the sets) and the dependent variable (another element in the other set). In other words, for a given value of the independent variable there is one value of the dependent variable.

Postulate **P1** then refers to a mathematical function between two features of proteins: the activity and the 3D structure. The activity is usually expressed as a quantity (kinetic constants such as the Michaellis-Menten constant Km) and the structure may be represented by a quantity also, for instance the fold classification; yet, such quantities have not been easily related, so a new set of measurements is needed to evaluate **P1** (see below for a further discussion on this aspect). To do so, the question we want to address first is: what type of mathematical function is this? Basically, there are three types of mathematical functions:

D4: Injections. In mathematics, this refers to one-to-one relations: given two sets S (3D Structure) and A (protein Activity), there is at least one element in S related with one element in A (see Figure 1A and 1B). Therefore, there can be elements of the set A that do not have a matching partner in set S (Figure 1B).

D5: Surjections. This is defined as a mathematical function where given two sets S and A, there is an association of at least one element in S with an element in A (see Figure 1A and 1C). Therefore, there can be elements of the set A that have one or more relations with elements in set S.

D6: Bijections. These are defined as mathematical functions where for every element in set S there is exactly one element in set A associated to it. They occur when both an injection and a surjection relation exist (see Figure 1A).

In all these cases (injections, surjections and bijections), the mathematical function f might be reversible: given f: S → A, then it is possible to find a function g such that g: A → S. However, only in the case of bijections the reversibility of the association is a necessary condition of the function.

Expressing these concepts in terms of the 3D structure-activity relationship of proteins, we may say that this relationship presents the properties of injections. For a long time biochemists have characterized the activities of proteins; however, for some time many activities were known but no protein 3D structures were associated to them; more recently, with the advent of DNA sequencing, many protein sequences and 3D structures are known for which no activity has being assigned yet (Norin & Sundström, 2002). However, given postulate P1, we must expect that for each protein there have to be both an activity and a 3D structure associated to it; consequently, the currently unknown 3D structures or activities of proteins will be measured eventually.

Alternatively, most of the current approaches to study the 3D structure-activity relationship of proteins treat this as a surjection: the evolution theory postulates that protein activity or 3D structure has been conserved in different species (orthologous proteins); thus this is a case of a one-to-many (one function-many structures) relation. Additionally, in protein

evolution the term "convergence" refers to the cases where different 3D structures of proteins have evolved to share a similar activity; conversely, an alternative example are single-domain moonlighting proteins, where one 3D structure is associated with multiple activities, albeit, using different molecular surfaces (Jeffery, 1999, 2003, 2009; Copley, 2003). In any case though, the one-to-many association prevails as much as we group together 3D structures or activities that are not identical. That is, to the best of our knowledge, there are no two proteins with identical activities reported so far with perfectly different 3D structures, nor are there two proteins with identical 3D structures with perfectly different activities. Take for instance the triose-phosphate isomerase proteins; these are proteins with a high degree of sequence-3D structure similarity, sharing similar but not identical activities (see Table 1). In the case of moonlighting proteins, there is no evidence that the two different activities may be performed in the same protein having exactly the same 3D structure, yet the structure may be slightly altered to accomplish different activities (Bateman *et al*., 2003; Krojer *et al*., 2002).

The need to move from considering similar to identical activity or 3D structure in the structure-activity relationship of proteins is important to improve our understanding of this relationship. On the one hand, it is convenient to assume similarity in 3D structure or activity of proteins in the discovery phase of biology (i.e., accelerated discovery of new proteins) because this assumption allows for the classification of new proteins into known families of proteins with known activity. Alternatively, provided the existence of an activity assay, it is possible to identify new proteins with such activity and presumably related in their 3D structure. However, after this initial phase of discovery, full understanding of the activity or 3D structure of a protein requires more detailed analysis both experimentally and theoretically. For the theoretical part, here we claim that in order to gain a better understanding of the 3D structure-activity relationship of proteins it is necessary to be precise in the terms used to relate these properties.

From this analysis we noted that since the 3D structure-activity relationship of proteins presents features of both injections and surjections, thus it may be best represented by a bijection. Furthermore, assuming that the injective feature is only a temporal one, and the surjective feature exists if and only if the definition of activity or 3D structure is not precise, we may conclude that the best way to analyze the 3D structure-activity relationship of proteins is as a bijection, where we postulate that for any given protein there is always one activity related to a given 3D structure. This approach necessarily implies that one has to come up with a rigorous and precise definition for both 3D structure and activity. Herein lies the challenge.

This conclusion leads us to the following scenario: let us assume that there is a set S with every possible 3D structure of proteins, and a set A with every possible measurable activity of proteins; then, for a given protein 3D structure in set S there is exactly one protein activity in set A; conversely, for a given protein activity in set A, there is exactly one protein 3D structure in set S. In this scenario, there are no identical activities in set A, neither there are identical structures in set S. To formally express this:

$$A = f(S) \tag{1}$$

Now, in order to express this relation in numerical terms, let us define the 3D structure as a matrix (e.g., adjacency matrix) and activity as a vector (e.g., list of critical residues for

protein activity). Choosing this set of critical residues is a convenient pick since it has been reported that proteins sharing high 3D structural similarity do not share the same set of critical residues (Cota E *et al.*, 2000; Rivera MH *et al.*, 2003), yet some critical residues are indeed shared between homologue proteins (Zhang Z & Palzkill T, 2003). Thus, representing 3D structure as a matrix (M) and activity as a vector of critical residues (C) provides us with a way to express this relation formally and look for mathematical tools to define the mathematical function inherent to these quantities. Thus:

$$C = f(M) \tag{2}$$

In other words, given a set of contacts between the residues of a protein (3D structure), our problem is to find a mathematical transformation of this matrix into a vector containing the critical residues for the protein function. If the mathematical function relating M and C is a bijection, then it must be possible to transform the vector C back into the matrix M. In order to find the mathematical function involved in this transformation, having access to multiple 3D structures and multiple sets of critical residues for several proteins is required.

Our analysis has several implications for the analysis of the 3D structure-activity relationship. In the present review, we will discuss only those relevant for the prediction of protein 3D structure. That is encouraged by the emergence of new approaches for the prediction of protein 3D structure that are based on the notion that the 3D structure-activity relationship is a bijection. However, these approaches have been developed in the absence of the current mathematical context, as we will describe below; embracing this bijection may provide the basis to improve the current methods of protein 3D structure prediction.

3. Current methods for protein 3D structure prediction

In this section we will summarize the ideas behind them and the kind of relationship that they assume between 3D structure and activity. This review does not attempt to cover in detail these methodologies, but to present the basic aspects of them in the context of postulate P1. For detailed descriptions of these methodologies, there are other reviews published elsewhere (Jones & Thornton, 1993; Martí-Renom *et al.*, 2000; Osguthorpe, 2000; Hardin *et al.*, 2002; Koretke *et al.*, 2002; Zhang, 2002; Godzik, 2003).

3.1 General considerations

Despite of the diversity of approaches to perform structural predictions, they all share a common design. The two key components of any method are the model generator and a quality evaluator (Figure 2).

1. Model generators refer to algorithms that create native-like protein 3D structures. There are two ways to generate such structures: knowledge-based strategies that depend on the available structures in databases and *ab initio* strategies (also known as physics-based), which consider physics principles to generate structures. Typically, model generators produce many alternative 3D structures that are potential solutions to the native structure of the protein.

2. Quality evaluators. These algorithms aim to evaluate the quality of the models produced by the model generators, in order to select the best models; i.e., those resembling the known native-like structure of proteins. Like the model generators, quality evaluators can be knowledge-based or *ab intio*.

It is important to keep in mind that these methodologies have limitations, especially if they are used to gain insights into the relation between the 3D structure and activity of poorly characterized proteins. Knowledge-based model generators and evaluators assume surjective relations between structure and activity, since the common idea of modellers of protein 3D structures is to assist in the grouping of protein structures based on similar attributes (Gerstein & Hegyi, 1998; Domingues *et al*., 2000; Skolnick *et al*., 2000). Therefore, in these cases knowledge of the protein 3D structure may provide inaccurate information about the activity (Martin *et al*., 1998). On the other hand, *ab initio* methods do not take into account the 3D structure-activity relation to perform predictions. With this kind of predictions, it is unlikely to get precise information about the activity of the protein from its 3D structure (Baker & Sali, 2001, and the results from CASP9).

Often the 3D structure is used to interpret the activity and rarely the other way around (Gherardini & Helmer-Citterich, 2008), thus it is not surprising that the current methods of protein 3D structure prediction do not address the prediction of 3D structure from the activity of the protein. In spite of this limitation, current methodologies for protein 3D structure prediction have been important in the development of the ideas about protein 3D structure determinants and their relationship with activity. Consequently, in the next two sections we will describe briefly the current methods for protein structure predictions, their features and limitations to elucidate protein activity.

3.2 Template-based modeling

This kind of predictions uses a protein of known 3D structure as a template to build the model of a protein whose 3D structure is unknown (target). The most critical part of this methodology is to identify adequate template(s) for the target. Accordingly, template-based modelling is classified in two main areas: homology modelling and fold recognition.

The idea behind homology modelling is that similar sequences have similar 3D structures (Doolittle, 1981, 1986; Chothia & Lesk, 1986). In this regard, the quality of a 3D model for a target protein depends strongly on the percentage of sequence identity between the target and template; the greater the identity, the more accurate the model will be. Likewise, below 30% of identity between the target and template proteins (sometimes referred as the "twilight zone"; Doolittle, 1986), several false templates may be identified for the target protein (Sander & Schneider, 1991; Rost, 1999). In that case, templates should be searched with fold recognition algorithms (Rost, 1999; see below). Templates can be found by searching databases of proteins with known 3D structure (e.g. the Protein Data Bank) with sequence alignment tools like BLAST (Altschul *et al*., 1990, 1997) or FASTA (Pearson & Lipman, 1988). Then, models of the target protein are built from the templates, taking into account changes that must be introduced like insertions and deletions in the template (indels), side chain conformations of non-conserved residues, possible rearrangements in the backbone, among others (Jones & Thirup, 1986; Bruccoleri & Karplus, 1987; Vásquez, 1996).

Afterwards, the quality of the resulting models is evaluated (Laskowski *et al*., 1993; Hooft *et al*., 1996; Wallner & Elofsson, 2003; Ginalski *et al*., 2003).

On the other hand, fold recognition methodologies identify proteins sharing similar 3D structures even if they do not have any obvious sequence similarity (Jones & Thornton, 1993; Godzik, 2003). Fold recognition can be performed in two ways. The first involves the enhancement of homology detection (Fischer & Eisenberg, 1996; Jaroszewski *et al*., 1998; Rychlewski *et al*., 2000), by using sequence profiles compiled from protein sequences that are compatible with the target. Two examples of this approach are PSI-BLAST (Altschul, 1997) and hidden Markov models (Durbin *et al*., 1998). Accuracy of prediction is increased further if structural information (e.g. secondary structure) is incorporated in the profiles (Di Francesco *et al*., 1997a, 1997b). The second approach is termed "threading" (Jones *et al*. 1992; Godzik & Skolnick, 1992). Here, the target sequence is forced to adopt the 3D structure of a potential target. Then the quality of the model is evaluated with a structure-based score. If the model has a high score, there is confidence that the target adopt a similar 3D structure as the template, otherwise the model is discarded. Once the template(s) is (are) found, a 3D-structural model of the target protein is built following the steps described in homology modelling after the initial template identification.

Template-based modelling has been recognized as the most accurate approach for protein structure prediction, especially if the identity between target and template is high (Chothia & Lesk, 1986; Sali *et al*., 1995; Cozzetto *et al*., 2009). However, as any model, these need to be tested in their ability to reproduce a biologically relevant feature, such as the activity. Since these methods assume a surjection for the structure-activity relationship, there are limitations imposed by such assumption, which are more notorious in the cases of low sequence similarity between the target and template proteins. One example of the limitation induced by the surjection conjecture in the structure-activity relationship of proteins is the TIM barrel fold, a common 3D-structure present in enzymes with very different activities such as oxidoreductases, hydrolases, lyases and isomerases (Greene *et al*., 2007). Likewise, the opposite situation is common: proteins with very similar activities and structurally unrelated. For instance, both chymotrypsin and subtilisin are serine-proteases with the same catalytic triad in the active site even thought they have completely different 3D-structures (Wallace *et al*., 1996).

Furthermore, even when there is a clear similarity between target and template sequences, there can be measurable structural differences. The most common example is loop structure. Precise prediction of loop regions is usually hard to accomplish since they tend to exhibit higher sequence variability and often have insertions and deletions relative to templates (Martí-Renom *et al*., 2000). Loops though, play an important role conferring specificity to the protein activity. Another less frequent situation is when there are visible differences in active sites of related proteins. This can lead to inaccurate modelling of the structure of target proteins (Moult, 2005). One way to improve the modelling of loops would be to evaluate the predicted activity of the model.

The information summarized above provides a general notion about the relationships that template-based modelling assumes. One-to-many relations between protein structure and activity are quite common with this kind of predictions. Thus, it is frequent to misrelate the activity of a protein from the knowledge of its fold alone (Martin *et al*., 1998). It is often

necessary to use other resources to predict the activity more accurately, as the use of local structural features of proteins in active sites (see Gherardini & Helmer-Citterich, 2008 for more details). Such tools work with the traditional approaches for predictions: knowledge-based like the 3D-templates (Wallace *et al*., 1996); or physics based, for example the identification of clefts and pockets in protein structures (Laskowski *et al*., 1996; Binkowski *et al*., 2003). These methods provide a theoretic framework to understand the 3D structure-activity relation in a one-way path: the prediction of activity from structure.

3.3 *Ab initio* modeling

Template based modelling can provide insights into the 3D structure and activity of poorly characterized proteins. In terms of generating reliable models it has an intrinsic limitation: it requires a protein of known 3D structure in order to produce a model. This may not be a problem in many situations, but there are proteins without any detectable template (more than half of the sequenced proteins in known genomes, see Yura *et al*., 2006). In such cases the alternative is *ab initio* modelling, also known as template-free modelling (Osguthorpe, 2000; Hardin *et al*., 2002; Koretke *et al*., 2002). The premise of these modelling methods is that the protein sequence determines the native structure, which has the global minimum potential energy among all the alternative conformations. In other words, *ab initio* methods assume that sequence alone would be sufficient to model the structure of proteins. For this reason, *ab inito* methods are adjured to predict the structure folds that were previously unknown.

Ab initio methods carry out a large-scale search for protein structures that have a particularly low energy for a given amino acid sequence. The two critical parts of these predictors are the conformational search strategy and the energy evaluation method (known as energy potential). To perform a fast and efficient search of the conformational space, *ab initio* methods use sophisticated algorithms suited to solve combinatorial problems since it is impossible to systematically explore all the conformations of a polypeptide chain. Monte Carlo algorithms (Simons *et al*., 1999; Ortiz *et al*., 1999), genetic algorithms (Pedersen & Moult, 1997a, 1997b), zipping and assembly (Ozcan *et al*., 2007) and molecular dynamics (Duan & Kollman, 1998; Shaw DE *et al.* 2010) are among the most frequently used methods to explore the conformational space of protein structures. Likewise, the energy potential is crucial to evaluate and select models of the target protein. Energy potentials can be of two kinds: molecular mechanics potentials, that are derived from physical-chemical calculations (Brooks *et al*., 1983; Pearlman *et al*., 1995) and knowledge-based potentials are constructed from the statistical analysis of the available structures in databases (Sippl, 1990; Koretke *et al*., 1998; Kuhlman and Baker, 2000).

Ab initio predictions usually consume a great deal of time and computer power. Recent methods make simplifications on the protein 3D structure in order to keep an acceptable speed (Helles, 2007). One of the solutions is to reduce the number of atoms that represent the protein 3D structure in order to simplify the model generation process (Kolinski, 2004; Lee *et al*., 1999). An alternative to speed up calculations is to consider fragment assembly strategies (Simons *et al*., 1999; Jones & Thirup, 1986). The idea with this approach is to split the structure into smaller fragments composed by many residues. Fragments are selected from a knowledge-based database on the basis of structural compatibility with the target

sequence and secondary structure propensities. The assembly of such substructures is determined by the energy potential and the conformation searching strategy. There are also multi-scale methods, like those of Cecilia Clementi, which change the resolution of the model depending on the questions that want to be asked of the protein (Shehu *et al*., 2009).

Template-free modelling has experienced much progress since the first blind prediction experiment known as "Critical Assessment of Techniques for Protein Structure Prediction" (CASP) took place in the early 90's (Bourne, 2003; Moult, 2005). However, despite of the considerable efforts the accuracy of *ab initio* predictions is still very low, compared to template-based modelling. That is, models generated with *ab initio* methods may have very large deviations from the experimental structures. In other cases, the 3D structure of the model can be completely wrong (this is actually a common situation). These limitations have hindered the practical use of *ab intio* modeling for the inference of the 3D structure-activity relationship on the target proteins (Baker & Sali, 2001; the results from CASP9).

Finally, *ab initio* predictions do not take into account the relation between 3D structure and activity explicitly, therefore they provide little reliable information about this relationship. On the other hand, they assume that proteins fold autonomously to the 3D structure with the minimum free energy (this is the case for most globular proteins), but there are cases where this assumption may be unjustified, as in the case of protein folding under kinetic control. For example, it has long been recognized that transmembrane proteins do not adopt their final, functional 3D-structure unassisted, but they need a translocation machinery to insert into the membrane and fold (Elofsson & von Heijne, 2007). Hence, the use of these strategies is inadequate for transmembrane proteins. Nonetheless, the ROSETTA method (originally developed for globular proteins) has been adapted to predict transmembrane proteins, with limited success (Yarov-Yarovoy *et al*., 2006). Additionally, *ab initio* predictions are unsuited for natively unstructured proteins (proteins that do not have a defined, unique structure), because they perform their activities as many alternative, rapidly interchanging conformations that correspond to multiple energy minima (Radivojac *et al*., 2007).

Despite of these disadvantages, *ab initio* predictions sometimes provide insights about protein activity. For example, in the fourth CASP experiment, the ROSETTA method was able to predict the structure of a couple target proteins that are structurally related to proteins of known 3D structure that were missed by fold recognition methods (Bonneau *et al*., 2001; Baker & Sali, 2001). Interestingly, the activities of the target proteins were similar, even thought there was no significant sequence identity between the proteins. A second example is the signalling protein Frizzled, whose critical residues for activity (previously characterized) were clustered together in the predicted structure in a surface patch likely to be involved in key protein-protein interactions (Baker & Sali, 2001). From these examples, it can be concluded that *ab initio* methods are more effective to gain information about the activity if they are combined with knowledge-based approaches (carrying on their limitations).

3.4 Concluding remarks about the current methods for protein 3D structure prediction

The available methodologies for the 3D structure prediction of proteins have provided useful insights about the relation between 3D structure and activity, and helped to construct the current paradigm. However, further refinement of these methods may assist

to fully relate protein 3D structure with activity. In this review we propose that such refinement may come from the recognition of the bijective nature of the 3D structure-activity relationship. For instance, knowledge-based methods imply a surjective relationship between activity and 3D structure. Consequently, predicting details on the activity of a modelled 3D structure of a protein can be hard, since there are examples of folds associated with many activities and *vice versa*. Furthermore, *ab initio* methods do not consider the structure-activity relationship, therefore the information they provide about the activity is commonly inaccurate. Additionally, template-free methods assume that proteins fold autonomously into a stable, minimum energy conformation, limiting their applicability in proteins that do not have these features because they fold under kinetic control.

In summary, it is necessary to develop methods that take into account the bijective nature of the 3D structure-activity relationship, in order to improve the usefulness and reliability perhaps, of protein 3D structure prediction methods. In the following section we will describe the available methodologies that take into account this bijection.

4. Emerging methods for protein structure prediction based on the bijective nature of protein 3D structure and activity

The previous section outlined the current status in the protein 3D structure prediction field, its strengths and weaknesses with regard to activity inference. It is evident that current methodologies still have limitations to exploit the usefulness of the 3D structure-activity relationship. Fortunately, new methodologies have been developed that take into account the bijective nature between the 3D structure and activity of proteins. This section will discuss the principles behind these methods and their capabilities.

4.1 Relevance of critical residues in the 3D structure-activity relation

These methods are based on the concept of critical residues, which are defined as those residues that upon mutation abolish the activity of a protein. Such definition depends on the experimental procedure used to measure the activity of the protein, but generally speaking, residues are considered critical if they tolerate few if any mutations (Loeb *et al*., 1989; Rennell *et al*., 1991; Terwilliger *et al*., 1994; Huang *et al*., 1996; Axe *et al*., 1998). Therefore, an experimentally determined critical residue may be either important to maintain the 3D structure of a protein or critical for the interaction with another molecule, or both. Thus, these residues constitute a key piece of knowledge that can be exploited to relate activity and 3D structure. Not surprisingly, methods have been developed to predict critical residues from protein sequence and/or 3D structure (Elcock, 2001; del Sol Mesa *et al*., 2003; Glaser *et al*., 2003; Thibert *et al*., 2005; Cusack *et al*., 2007).

Additionally, critical residues may provide a useful way to quantify structural features of proteins and relate them with the activity of a protein. As we mentioned earlier, there are no reports of two proteins with identical 3D structures with perfectly different activities and *vice versa* (please note that correctly representing both 3D structure and activity is one of the biggest challenges, and therefore, a Cartesian representation of the protein may not be the best to distinguish identical 3D structures). Hence, it is expected that proteins with similar,

yet strictly different 3D structures, will have different sets of critical residues. If that is the case, the set of critical residues for a given protein should reflect its unique 3D structural and activity properties. Such assumption provides the framework for methodologies that are based in the bijective relation between 3D structure and activity.

In the next two sections, we will describe the available bijective approaches. To simplify, they are classified in two categories: phylogeny and structure-based methods. The usefulness of these methodologies to relate 3D structure and activity will also be discussed.

4.2 Phylogeny-based approaches

The idea behind phylogenetic methods is to exploit the evolutionary information that can be extracted from the analysis of the sequences of related proteins. To do so, it is necessary to identify a group of similar protein sequences and to construct a multiple sequence alignment with them. There are two types of information that can be extracted from the alignments: sequence conservation and sequence correlation.

The first property refers to the frequency of a specific amino acid at a given position in the alignment; residues occurring at high frequencies at particular positions are considered conserved residues. Sequence conservation is related to the direct evolutionary pressure to maintain the physical-chemical characteristics of some positions in order to retain the activity and/or 3D structure of a family of homologous proteins. Therefore, highly conserved residues are regarded as critical to retain the 3D structure and activity of the protein. In the literature, there are many reports of methods to calculate conservation (see Valdar *et al.*, 2002 and Sadowski & Jones, 2009 for comprehensive reviews).

Residue correlation (also known as co-evolution or co-variation) is defined as concerted patterns of variation between two or more different positions in a multiple sequence alignment of homologous proteins (Altschuh *et al.*, 1987). Such co-variating residues are proposed to correspond to compensatory substitutions that maintain the structural stability or functional properties of proteins throughout their evolutionary history. It has been observed that correlated residues tend to be in physical contact (Altschuh *et al.*, 1988); thus, this feature was proposed to be useful in residue contact predictions (Göbel *et al.*, 1994; Pazos *et al.*, 1997; Olmea & Valencia, 1997).

Critical residues predicted with phylogenetic approaches can be exploited to improve structural predictions. For example, the method reported by the group of Valencia (Olmea *et al.*, 1999) uses sequence conservation and correlation as part of a structure quality evaluator for a fold recognition structure predictor. The authors of this work report that the method is capable to distinguish correct models from incorrect models generated by the TOPITS threading algorithm (Rost, 1995). However, the accuracy of the algorithm decreases for large proteins, thus restricting its applicability.

Another exciting application of sequence correlation and co-variation is the design of new proteins (a field that strongly depends on 3D structure prediction tools). An illustrative example of protein engineering is the use of the Statistical Coupling Analysis method (SCA; Lockless & Ranganathan, 1999), which was used to design a novel artificial protein sequence with the same 3D structure and activity as natural WW domain proteins (Socolich *et al.*,

2005; Russ *et al*., 2005). In order to design the protein, the method took into account the critical residues of the protein as well as their patterns of conservation and correlation (Socolich *et al*., 2005). Furthermore, the methodology has been used recently to design a light-modulated chimerical enzyme (Lee *et al*., 2008).

Ultimately, conserved residues will only capture the common critical residues for a set of homologous proteins, and will most likely miss the critical residues specific for the activity and 3D structure of each protein in that set. In that sense, conserved residues may be useful to score common structural features of proteins but may not be useful to evaluate the different 3D structure and biological activity of each homologous protein. To do so, a new method has recently being described that is now reviewed.

4.3 Methods based on structural information

A complementary approach to identify critical residues is to consider only 3D structural properties of proteins. One of the most recent approaches to study the 3D structure of proteins is graph theory (Vendruscolo *et al*., 2002; Greene & Higman, 2003; Thibert *et al*., 2005; Cusack *et al*., 2007; Montiel Molina *et al*., 2008), a theoretical approximation that has been used to characterize other biological systems, such as metabolism, genetic regulation and protein-protein interaction networks (Jeong *et al*., 2000, 2001; Del Rio *et al*., 2009). Under this view, protein 3D structure is modelled as a graph (network), which is defined by one set of nodes that represent the amino acid residues in a protein, and a set of edges that can be considered as molecular interactions between any two residues (nodes). The criterion to link two residues by an edge is based on maximum distances among the atoms of residues (Vendruscolo *et al*., 2002; Greene & Higman, 2003; Thibert *et al*., 2005; Cusack *et al*., 2007; Milenković *et al*., 2009).

Graph theory provides the mathematical basis to study the topological properties of networks derived from the protein structure. One useful concept of this field to characterize networks is network centrality, which measures the relative importance of nodes in the network. Thus, centrality can be used to predict critical residues (Thibert *et al*., 2005; Cusack *et al*., 2007) or to study topological features of protein structures (Vendruscolo *et al*., 2002; Greene & Higman, 2003). Some of the most common centralities used to study networks derived from protein structures are betweenness and closeness, which relate nodes through the shortest paths among all the nodes in the graph (Freeman, 1977).

Centrality is reliable when it comes to predict critical residues (Chea & Livesay, 2007), but how can these be used to predict 3D structural and functional features? We have recently reported a tool named "JAMMING" to facilitate this task. The method predicts critical residues using betweenness or closeness centrality (Cusack *et al*., 2007). We have shown that JAMMING may be used to identify protein structures involved in ligand binding by screening thousands of conformations generated from protein 3D structures in the unbound form; such functional conformers were found by a scoring system that matches critical residues with central residues (Montiel Molina *et al*., 2008). Our results show that critical residues for a molecular interaction are preferentially found as central residues of protein structures in complex with a ligand. Therefore, the tool helps to relate the activity of the protein (binding to a molecule) with its structural properties (the conformers).

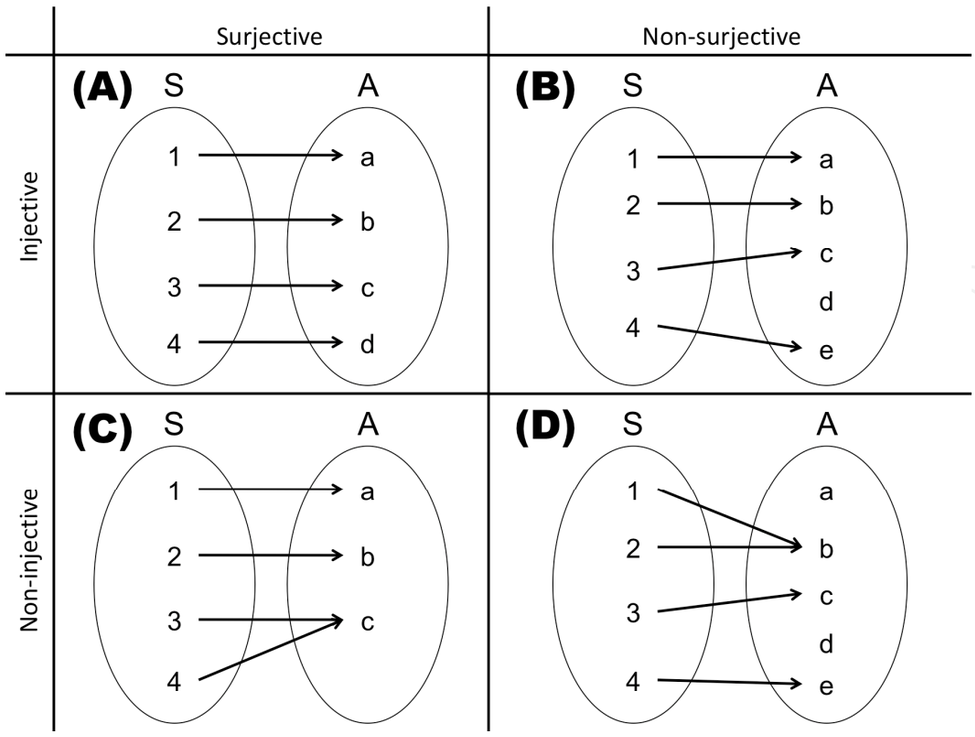


Fig. 1. Examples of injective and surjective functions
A) Injective and surjective (bijection). B) Injective and non-surjective. C) Non-injective and surjective. D) Non-injective and non-surjective.

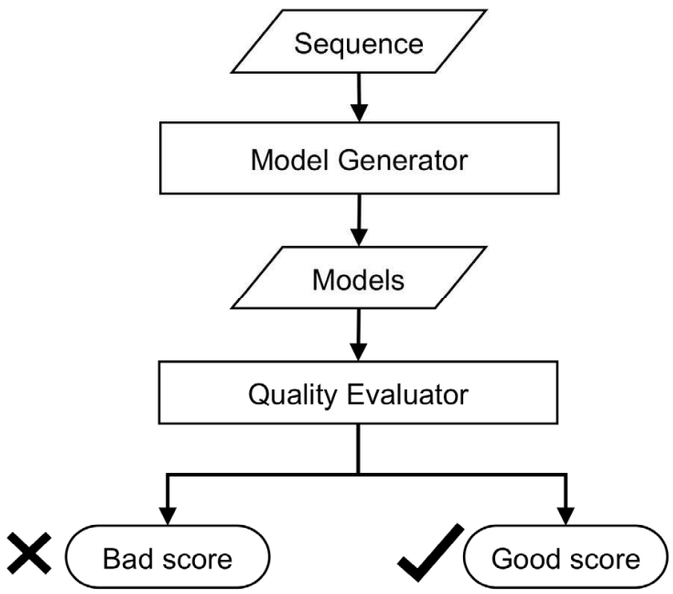


Fig. 2. Flowchart of structural prediction methods. The protein sequence is the input of the model generator algorithm. As a result, the generator produces multiple models that are assessed by the quality evaluator. Finally, the best scoring models are selected, whereas the models with bad scores are discarded.

Species	PDB	Identity[1] [%]	RMSD[1] [A]	K_m [mM]	K_{cat} [1/s]	References[2]
Homo sapiens	1wyi	100	0.0	0.34	16320	Gracy, 1975
Oryctolagus cuniculus	1r2t	98	0.4	0.42	8670	Krietsch, 1975a
Gallus gallus	8tim	89	0.8	0.47	4300	Xiang *et al.*, 2004
Saccharomyces cerevisiae	1ypi	53	1.0	1.27	16700	Krietsch, 1975b
Trypanosoma brucei	1tpf	53	1.1	0.19	6000	Kursula *et al.*, 2002
Leishmania mexicana	1amk	50	1.6	0.30	4170	Kohl *et al.*, 1994
Escherichia coli	1tre	46	1.4	1.03	9000	Alvarez *et al.*, 1998
Vibrio marinus	1aw2	42	1.4	1.90	7000	Alvarez *et al.*, 1998

[1] Sequence identities and RMSDs were calculated with the program DaliLite (Holm & Park, 2000) using 1wyi as the first molecule in all comparisons.
[2] References originally reporting the values for Km and Kcat.

Table 1. Structural and functional features of triose-phosphate isomerases from different species.

5. Conclusions

The structure-activity paradigm has travelled a long way since the first efforts to characterize the 3D structure and biological activity of proteins were performed back in the 1930's. Traditionally, the relationship between 3D structure and activity has been considered as a surjection to assist in the classification of the known proteins. Consequently, knowledge-based classification schemes, although useful to give sense to an ever-increasing list of known protein sequences, may not provide the basis to understand the subtleness of protein activity and structure in nature. In a similar fashion, most of the current methods for protein 3D structure prediction are unable to provide better insights about the activity of a protein of unknown structure (especially if it does not have a close homologue).

In this review, we propose that the relation between structure and activity may be modelled by a bijection. Critical residues provide a way to relate the structure and the activity of proteins, especially in the situation where structure and activity are represented by a bijection. Current methodologies based on the bijective 3D structure-activity relationship unnoticeably provided novel tools to explore the subtle determinants of protein activity, structure and their interaction. We claim that the incorporation of these methods into the traditional tools for protein structure prediction will improve the usefulness of the structural predictions to understand the details on the evolution of protein activity.

6. Acknowledgment

We want to acknowledge the technical assistance of Dra. Maria Teresa Lara Ortiz and the IT core facility of the Instituto de Fisiología Celular; Dr. Alejandro Fernández for his fruitful discussions on this subject and reading of the manuscript. This work was supported in part by one grant from CONACyT (82308) and two grants from the Universidad Nacional Autónoma de México to GDR, including the Macroproyecto: Tecnologías para la Universidad de la Información y la Computación and PAPIIT IN205911, and CONACyT (102182 and 133294) to NP.

7. References

Altschuh D, Lesk AM, Bloomer AC, Klug A. (1987). Correlation of co-ordinated amino acid substitutions with function in viruses related to tobacco mosaic virus. J Mol Biol. 193(4):693-707.

Altschuh D, Vernet T, Berti P, Moras D, Nagai K. (1988). Coordinated amino acid changes in homologous protein families. Protein Eng. 2(3):193-199.

Altschul SF, Gish W, Miller W, Myers EW, Lipman DJ. (1990). Basic local alignment search tool. J Mol Biol. 215(3):403-410.

Altschul SF, Madden TL, Schäffer AA, Zhang J, Zhang Z, Miller W, Lipman DJ. (1997). Gapped BLAST and PSI-BLAST: a new generation of protein database search programs. Nucleic Acids Res. 25(17):3389-402.

Alvarez M, Zeelen JP, Mainfroid V, Rentier-Delrue F, Martial JA, Wyns L, Wierenga RK, Maes D. (1998). Triose-phosphate isomerase (TIM) of the psychrophilic bacterium Vibrio marinus. Kinetic and structural properties. J Biol Chem. 273(4):2199-2206.

Axe DD, Foster NW, Fersht AR. (1998).A search for single substitutions that eliminate enzymatic function in a bacterial ribonuclease. Biochemistry. 37(20):7157-7166.

Baker D, Sali A (2001). Protein structure prediction and structural genomics. Science. 2001. 294(5540):93-6.

Balzani V V, Credi A, Raymo FM, Stoddart JF. (2000). Artificial Molecular Machines. Angew Chem Int Ed Engl39(19):3348-3391.

Bateman OA, Purkiss AG, van Montfort R, Slingsby C, Graham C, Wistow G. (2003). Crystal structure of eta-crystallin: adaptation of a class 1 aldehyde dehydrogenase for a new role in the eye lens. Biochemistry. 42(15):4349-56.

Binkowski TA, Adamian L, Liang J. (2003). Inferring functional relationships of proteins from local sequence and spatial surface patterns. J Mol Biol. 332(2):505-26.

Bonneau R, Tsai J, Ruczinski I, Baker D. (2001). Functional inferences from blind *ab initio* protein structure predictions. J Struct Biol. 134(2-3):186-90.

Bourne PE. (2003).CASP and CAFASP experiments and their findings. Methods Biochem Anal. 44:501-7.

Brooks BR, Bruccoleri RE, Olafson BD, States DJ, Swaminathan S, Karplus M. (1983). CHARMM: A program for macromolecular energy, minimization, and dynamics calculations. J Comp Chem. 4(2):187–217.

Bruccoleri RE, Karplus M. (1987). Prediction of the folding of short polypeptide segments by uniform conformational sampling. Biopolymers. 26(1):137-168.

Chea E, Livesay DR. (2007). How accurate and statistically robust are catalytic site predictions based on closeness centrality? BMC Bioinformatics. 8:153.

Chollet A, Turcatti G. (1999). Biophysical approaches to G protein-coupled receptors: structure, function and dynamics. J Comput Aided Mol Des. 13(3):209-219

Chothia C, Lesk AM. (1986). The relation between the divergence of sequence and structure in proteins. EMBO J. 5(4):823-6.

Copley SD. (2003). Enzymes with extra talents: moonlighting functions and catalytic promiscuity. Curr Opin Chem Biol. 7(2):265-72.

Cota E, Hamill SJ, Fowler SB, Clarke J. (2000). Two proteins with the same structure respond very differently to mutation: the role of plasticity in protein stability. J Mol Biol. 302(3):713-25.

Cozzetto D, Kryshtafovych A, Fidelis K, Moult J, Rost B, Tramontano A. (2009) Evaluation of template-based models in CASP8 with standard measures. Proteins 77 Suppl 9:18-28.

Cusack MP, Thibert B, Bredesen DE, del Rio G (2007) Efficient identification of critical residues based only on protein structure by network analysis. PLoS ONE 2(5):e421.

Del Rio G, Koschützki D, Coello G (2009) How to identify essential genes from molecular networks? BMC Syst. Biol. 3:102.

del Sol Mesa A, Pazos F, Valencia A (2003) Automatic methods for predicting functionally important residues J Mol Biol. 326(4):1289-1302.

Di Francesco V, Garnier J, Munson PJ. (1997a). Protein topology recognition from secondary structure sequences: application of the hidden Markov models to the alpha class proteins. J Mol Biol. 267(2):446-463.

Di Francesco V, Geetha V, Garnier J, Munson PJ.(1997b). Fold recognition using predicted secondary structure sequences and hidden Markov models of protein folds. Proteins. Supplement 1:123-128.

Domingues FS, Koppensteiner WA, Sippl MJ. (2000). The role of protein structure in genomics. FEBS Lett. 476(1-2):98-102.

Doolittle RF. (1981). Protein Evolution. Science. 214(4525):1123-1124.

Doolittle RF. (1986). Of URFs and ORFs: a primer on how to analyze derived amino acid sequences. University Science Books, Mill Valley, CA, USA.

Drexler KE. (1994). Molecular nanomachines: physical principles and implementation strategies. Annu Rev Biophys Biomol Struct. 23:377-405.

Duan Y, Kollman PA. (1998). Pathways to a protein folding intermediate observed in a 1-microsecond simulation in aqueous solution. Science. 282(5389):740-744.

Durbin R, Eddy S, Krogh A and Mitchison G. (1998). Biological Sequence Analysis: Probabilistic Models of Proteins and Nucleic Acids. Cambridge: Cambridge University Press.

Elcock AH (2001) Prediction of functionally important residues based solely on the computed energetics of protein structure J Mol Biol. 312(4):885-896.

Elofsson A, von Heijne G. (2007). Membrane protein structure: prediction versus reality. Annu Rev Biochem. 76:125-140.

Fischer D, Eisenberg D. (1996). Protein fold recognition using sequence-derived predictions. Protein Sci. 5(5):947-55.

Freeman LC. (1977). A set of measures of centrality based on betweenness. Sociometry 40:35-41.

Gerstein M, Hegyi H. (1998). Comparing genomes in terms of protein structure: surveys of a finite parts list. FEMS Microbiol Rev. 22(4):277-304.

Gherardini PF, Helmer-Citterich M. (2008). Structure-based function prediction: approaches and applications. Brief Funct Genomic Proteomic. 7(4):291-302.

Ginalski K, Elofsson A, Fischer D, Rychlewski L. (2003). 3D-Jury: a simple approach to improve protein structure predictions. Bioinformatics. 19(8):1015-8

Glaser F, Pupko T, Paz I, Bell RE, Bechor-Shental D, Martz E, Ben-Tal N. (2003). ConSurf: identification of functional regions in proteins by surface-mapping of phylogenetic information. Bioinformatics. 19(1):163-4.

Göbel U, Sander C, Schneider R, Valencia A. (1994). Correlated mutations and residue contacts in proteins. Proteins. 18(4):309-317.

Godzik A, Skolnick J. (1992). Sequence-structure matching in globular proteins: application to supersecondary and tertiary structure determination. Proc Natl Acad Sci USA. 89(24):12098-12102.

Godzik A. (2003). Fold recognition methods. Methods Biochem Anal. 44:525-546.

Gracy RW. (1975).Triosephosphate isomerase from human erythrocytes. Methods Enzymol. 41:442-447.

Greene LH, Higman VA (2003). Uncovering network systems within protein structures. J Mol Biol. 334(4):781-91.

Greene LH, Lewis TE, Addou S, Cuff A, Dallman T, Dibley M, Redfern O, Pearl F, Nambudiry R, Reid A, Sillitoe I, Yeats C, Thornton JM, Orengo CA. (2007). The CATH domain structure database: new protocols and classification levels give a more comprehensive resource for exploring evolution. Nucleic Acids Res. 35:D291-D297.

Hardin C, Pogorelov TV, Luthey-Schulten Z. (2002). *Ab initio* protein structure prediction. Curr Opin Struct Biol. 12(2):176-181.

Holm L, Park J. (2000). DaliLite workbench for protein structure comparison. Bioinformatics. 16(6):566-567.

Hooft RW, Vriend G, Sander C, Abola EE. (1996). Errors in protein structures. Nature. 381(6580):272.

Huang W, Petrosino J, Hirsch M, Shenkin PS, Palzkill T. (1996).Amino acid sequence determinants of beta-lactamase structure and activity. J Mol Biol. 258(4):688-703.

Huisman GW, Gray D. (2002). Towards novel processes for the fine-chemical and pharmaceutical industries. Curr Opin Biotechnol. 13(4):352-358.

Jaroszewski L, Rychlewski L, Zhang B, Godzik A. (1998). Fold prediction by a hierarchy of sequence, threading, and modeling methods. Protein Sci. 7(6):1431-40.

Jeffery CJ. (1999). Moonlighting proteins. Trends Biochem Sci. 24(1):8-11.

Jeffery CJ. (2003). Moonlighting proteins: old proteins learning new tricks. Trends Genet. 19(8):415-7.

Jeffery CJ. (2009). Moonlighting proteins – an update. Mol. BioSyst. 5:345-350.

Jeong H, Mason SP, Barabasi AL, Oltvai ZN (2001). Lethality and centrality in protein networks. Nature. 411 (6833), 41-42.

Jeong H, Tombor B, Albert R, Oltvai ZN, Barabasi AL (2000). The large-scale organization of metabolic networks. Nature. 407 (6804), 651-654.

Jones D, Thornton J. (1993). Protein fold recognition. J Comput Aided Mol Des. 7(4):439-456.

Jones DT, Taylor WR, Thornton JM. (1992). A new approach to protein fold recognition. Nature. 358(6381):86-89.

Jones TA, Thirup S.(1986). Using known substructures in protein model building and crystallography. EMBO J. 5(4):819-822.

Kohl L, Callens M, Wierenga RK, Opperdoes FR, Michels PA. (1994).Triose-phosphate isomerase of Leishmania mexicana mexicana. Cloning and characterization of the gene, overexpression in Escherichia coli and analysis of the protein. Eur J Biochem. 220(2):331-338

Kolinski A. (2004). Protein modeling and structure prediction with a reduced representation. Acta Biochim Pol. 51(2):349-71.

Koretke KK, Luthey-Schulten Z, Wolynes PG. (1998). Self-consistently optimized energy functions for protein structure prediction by molecular dynamics. Proc Natl Acad Sci U S A. 95(6):2932-2937.

Koretke KK, Luthey-Schulten Z, Wolynes PG. (2002). Ab initio protein structure prediction. Curr Opin Struct Biol. 12(2):176-181.

Krietsch WK. (1975a). Triosephosphate isomerase from rabbit liver. Methods Enzymol. 41:438-442.

Krietsch WK. (1975b). Triosephosphate isomerase from yeast. Methods Enzymol. 41:434-438.

Krojer T, Garrido-Franco M, Huber R, Ehrmann M, Clausen T. (2002). Crystal structure of DegP (HtrA) reveals a new protease-chaperone machine. Nature. 416(6879):455-459.

Kuhlman B, Baker D (2000) Native protein sequences are close to optimal for their structures. Proc. Natl. Acad. Sci. USA 97:10383-10388.

Kursula I, Partanen S, Lambeir AM, Wierenga RK. (2002).The importance of the conserved Arg191-Asp227 salt bridge of triosephosphate isomerase for folding, stability, and catalysis. FEBS Lett. 518(1-3):39-42.

Kuhlman B, Dantas, G, Ireton GC, Varani G, Stoddard BL, Baker D. (2003). Design of a novel globular protein fold with atomic-level accuracy. Science 302:1364-1368.

Laskowski RA, Luscombe NM, Swindells MB, Thornton JM. (1996). Protein clefts in molecular recognition and function. Protein Sci. 5(12):2438-2452.

Laskowski RA, Moss DS, Thornton JM. (1993). Main-chain bond lengths and bond angles in protein structures. J Mol Biol. 231(4):1049-1067.

Lee J, Liwo A, Scheraga HA. (1999).Energy-based de novo protein folding by conformational space annealing and an off-lattice united-residue force field: application to the 10-55 fragment of staphylococcal protein A and to apo calbindin D9K. Proc Natl Acad Sci USA. 96(5):2025-30.

Lee J, Natarajan M, Nashine VC, Socolich M, Vo T, Russ WP, Benkovic SJ, Ranganathan R. (2008). Surface sites for engineering allosteric control in proteins. Science. 322(5900):438-442.

Lockless SW, Ranganathan R. (1999). Evolutionarily conserved pathways of energetic connectivity in protein families. Science. 286(5438):295-299.

Loeb DD, Swanstrom R, Everitt L, Manchester M, Stamper SE, Hutchison CA 3rd. (1989). Complete mutagenesis of the HIV-1 protease. Nature. 340(6232):397-400.

Luetz S, Giver L, Lalonde J. (2008). Engineered enzymes for chemical production. Biotechnol Bioeng. 101(4):647-653.

Martí-Renom MA, Stuart AC, Fiser A, Sánchez R, Melo F, Sali A. (2000). Comparative protein structure modeling of genes and genomes. Annu Rev Biophys Biomol Struct. 29:291-325.

Martin AC, Orengo CA, Hutchinson EG, Jones S, Karmirantzou M, Laskowski RA, Mitchell JB, Taroni C, Thornton JM. (1998). Protein folds and functions. Structure. 6(7):875-884.

Milenković T, Filippis I, Lappe M, Pržulj N, (2009) Optimized Null Model for Protein Structure Networks. PLoS ONE 4: e5967.

Mirsky AE, Pauling L. (1936). On the Structure of Native, Denatured, and Coagulated Proteins. Proc Natl Acad Sci USA. 22(7):439-447.

Montiel Molina HM, Millan-Pacheco C, Pastor N, del Rio G (2008) Computer-Based Screening of Functional Conformers of Proteins. PLoS Comput Biol. 4(2):e1000009.

Moult J. (2005). A decade of CASP: progress, bottlenecks and prognosis in protein structure prediction. Curr Opin Struct Biol. 15(3):285-289.

Neet KE, Lee JC. (2002).Biophysical characterization of proteins in the post-genomic era of proteomics. Mol Cell Proteomics. 1(6):415-420

Norin M, Sundström M. (2002). Structural proteomics: developments in structure-to-function predictions. Trends Biotechnol. 20(2):79-84.

Olmea O, Rost B, Valencia A. (1999). Effective use of sequence correlation and conservation in fold recognition. J Mol Biol. 293(5):1221-1239.

Olmea O, Valencia A. (1997). Improving contact predictions by the combination of correlated mutations and other sources of sequence information. Fold Des. 2(3):S25-S32.

Ortiz AR, Kolinski A, Rotkiewicz P, Ilkowski B, Skolnick J. (1999). Ab initio folding of proteins using restraints derived from evolutionary information. Proteins. Supplement 3:177-185

Osguthorpe DJ. (2000). Ab initio protein folding. Curr Opin Struct Biol. 10(2):146-152.

Ozkan SB, Wu GA, Chodera JD, Dill KA (2007) Protein folding by zipping and assembly. Proc. Natl. Acad. Sci. USA 104: 11987-11992.

Pazos F, Helmer-Citterich M, Ausiello G, Valencia A. (1997).Correlated mutations contain information about protein-protein interaction. J Mol Biol. 271(4):511-523.

Pearlman DA, Case DA, Caldwell JW, Ross WS, Cheatham TE III, DeBolt S, Ferguson D, Seibel G, Kollman P. (1995) AMBER, a package of computer programs for applying molecular mechanics, normal mode analysis, molecular dynamics and free energy calculations to simulate the structural and energetic properties of molecules. Comp Phys Commun. 91:1-41.

Pearson WR, Lipman DJ.(1988). Improved tools for biological sequence comparison. Proc Natl Acad Sci USA. 85(8):2444-2448.

Pedersen JT, Moult J. (1997a). Ab initio protein folding simulations with genetic algorithms: simulations on the complete sequence of small proteins. Proteins. Supplement 1:179-184.

Pedersen JT, Moult J. (1997b). Protein folding simulations with genetic algorithms and a detailed molecular description. J Mol Biol. 269(2):240-259.

Punta M, Ofran Y. (2008). The rough guide to in silico function prediction, or how to use sequence and structure information to predict protein function. PLoS Comput Biol. 4(10):e1000160.

Radivojac P, Iakoucheva LM, Oldfield CJ, Obradovic Z, Uversky VN, Dunker AK. (2007). Intrinsic disorder and functional proteomics. Biophys J. 92(5):1439-56.

Rennell D, Bouvier SE, Hardy LW, Poteete AR. (1991). Systematic mutation of bacteriophage T4 lysozyme. J Mol Biol. 222(1):67-88.

Rivera MH, López-Munguía A, Soberón X, Saab-Rincón G. (2003). Alpha-amylase from Bacillus licheniformis mutants near to the catalytic site: effects on hydrolytic and transglycosylation activity. Protein Eng. 16(7):505-514.

Robson B. (1999). Beyond proteins. Trends Biotechnol. 17(8):311-315.

Rost B. (1995). TOPITS: threading one-dimensional predictions into three-dimensional structures. Proc Int Conf Intell Syst Mol Biol. 3:314-421.

Rost B. (1999). Twilight zone of protein sequence alignments. Protein Eng. 12(2):85-94.

Russ WP, Lowery DM, Mishra P, Yaffe MB, Ranganathan R. (2005). Natural-like function in artificial WW domains. Nature. 437(7058):579-83.

Rychlewski L, Jaroszewski L, Li W, Godzik A. (2000). Comparison of sequence profiles. Strategies for structural predictions using sequence information. Protein Sci. 9(2):232-41.

Sadowski MI, Jones DT. (2009) The sequence-structure relationship and protein function prediction. Curr. Opin. Struct. Biol. 19:357-362.

Sali A, Potterton L, Yuan F, van Vlijmen H, Karplus M. (1995). Evaluation of comparative protein modeling by MODELLER. Proteins. 23(3):318-326.

Sander C, Schneider R. (1991). Database of homology-derived protein structures and the structural meaning of sequence alignment. Proteins. 9(1):56-68.

Shaw DE, Maragakis P, Lindorff-Larsen K, Piana S, Dror RO, Eastwood MP, Bank JA, Jumper JM, Salmon JK, Shan Y, Wriggers W. (2010) Atomic-level characterization of the structural dynamics of proteins. Science 330: 341-346.

Shehu A, Kavraki LE, Clementi C. (2009) Multiscale characterization of protein conformational ensembles. Proteins 76:837-851.

Simons KT, Bonneau R, Ruczinski I, Baker D. (1999). Ab initio protein structure prediction of CASP III targets using ROSETTA. Proteins. Supplement 3:171-176.

Sippl MJ. (1990). Calculation of conformational ensembles from potentials of mean force. An approach to the knowledge-based prediction of local structures in globular proteins. J Mol Biol. 213(4):859-83.

Skolnick J, Fetrow JS, Kolinski A. (2000). Structural genomics and its importance for gene function analysis. Nat Biotechnol. 2000 Mar;18(3):283-287.

Socolich M, Lockless SW, Russ WP, Lee H, Gardner KH, Ranganathan R. (2005). Evolutionary information for specifying a protein fold. Nature. 437(7058):512-518.

Straathof AJ, Panke S, Schmid A. (2002). The production of fine chemicals by biotransformations. Curr Opin Biotechnol. 13(6):548-556.

Terwilliger TC, Zabin HB, Horvath MP, Sandberg WS, Schlunk PM. (1994). In vivo characterization of mutants of the bacteriophage f1 gene V protein isolated by saturation mutagenesis. J Mol Biol. 236(2):556-71.

Thibert B, Bredesen DE, del Rio G (2005). Improved prediction of critical residues for protein function based on network and phylogenetic analyses. BMC Bioinformatics. 6:213.

Valdar WS. (2002). Scoring residue conservation. Proteins. 48(2):227-241.

Vásquez M. (1996). Modeling side-chain conformation. Curr Opin Struct Biol. 6(2):217-21.

Vendruscolo M, Dokholyan NV, Paci E, Karplus M (2002). Small-world view of the amino acids that play a key role in protein folding. Phys Rev E. 65(6 Pt 1):061910.

Wallace AC, Laskowski RA, Thornton JM. (1996). Derivation of 3D coordinate templates for searching structural databases: application to Ser-His-Asp catalytic triads in the serine proteinases and lipases. Protein Sci. 5(6):1001-1013.

Wallner B, Elofsson A. (2003).Can correct protein models be identified? Protein Sci. 12(5):1073-1086.

Xiang J, Jung JY, Sampson NS. (2004).Entropy effects on protein hinges: the reaction catalyzed by triosephosphate isomerase. Biochemistry. 43(36):11436-11445.

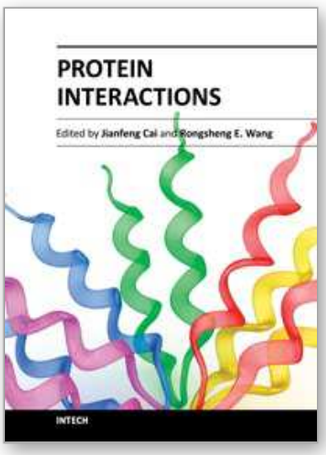
Yarov-Yarovoy V, Schonbrun J, Baker D. (2006).Multipass membrane protein structure prediction using Rosetta. Proteins. 62(4):1010-1025.

Yura K, Yamaguchi A, Go M. (2006) .Coverage of whole proteome by structural genomics observed through protein homology modeling database. J Struct Funct Genomics. 7(2):65-76.

Zaks A. (2001). Industrial biocatalysis. Curr Opin Chem Biol. 5(2):130-6.

Zhang H (2002). Protein Tertiary Structures: Prediction from Amino Acid Sequences. In: ENCYCLOPEDIA OF LIFE SCIENCES. John Wiley & Sons, Ltd: Chichester http://www.els.net/ [doi:10.1038/npg.els.0006101].

Zhang Z, Palzkill T. (2003). Determinants of binding affinity and specificity for the interaction of TEM-1 and SME-1 beta-lactamase with beta-lactamase inhibitory protein. J Biol Chem. 278(46):45706-45712.



Protein Interactions

Edited by Dr. Jianfeng Cai

ISBN 978-953-51-0244-1
Hard cover, 464 pages
Publisher InTech
Published online 16, March, 2012
Published in print edition March, 2012

Protein interactions, which include interactions between proteins and other biomolecules, are essential to all aspects of biological processes, such as cell growth, differentiation, and apoptosis. Therefore, investigation and modulation of protein interactions are of significance as it not only reveals the mechanism governing cellular activity, but also leads to potential agents for the treatment of various diseases. The objective of this book is to highlight some of the latest approaches in the study of protein interactions, including modulation of protein interactions, development of analytical techniques, etc. Collectively they demonstrate the importance and the possibility for the further investigation and modulation of protein interactions as technology is evolving.

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

IntechOpen