# UNIVERSIDAD NACIONAL AUTÓNOMA DE MÉXICO

## Maestría y Doctorado en Ciencias Bioquímicas

ESTUDIO DE LAS ESTRUCTURAS Y CONFÓRMEROS FUNCIONALES DE LAS PROTEÍNAS POR MEDIO DE LA TEORÍA DE GRÁFICAS

TESIS

QUE PARA OPTAR POR EL GRADO DE:

Doctor en Ciencias

PRESENTA:
HÉCTOR MARLOSTI MONTIEL MOLINA

TUTOR PRINCIPAL
DR. GABRIEL DEL RIO GUERRA
Instituto de Fisiología Celular

MIEMBROS DEL COMITÉ TUTOR
DR. LORENZO SEGOVIA FORCELLA
Instituto de Biotecnología
DR. ARTURO ROJO DOMÍNGUEZ
Departamento de Ciencias Naturales, UAM Cuajimalpa

Ciudad Universitaria, CDMX. Octubre, 2022

# STUDY OF PROTEIN STRUCTURE AND FUNCTIONAL CONFORMERS BY A GRAPH-THEORETICAL APPROACH

## Abstract

A long-standing goal in biology is to understand the relationship between function, structure, and dynamics of proteins. Protein structure prediction for proteins without recognized templates requires a substantial improvement in conformational search and accurate model selection that nowadays is missing. Likewise, considering that protein function at the molecular level is understood by the ability of proteins to dynamically bind and transform other molecules, the limited data on protein dynamics and on protein structures in association with their ligands represents a major hurdle to our understanding of protein function at the structural level. Recent reports show that protein function can be linked to protein structure and dynamics through graph centrality analysis, suggesting that the structures of functional and basal-state conformers may be inferred computationally. In the present work, we study protein structure form a graph-theoretical approach in order to find connectivity features that facilitate the sampling of protein functional model selection. We found that graphs derived from proteins have non-trivial distributions of degrees, clustering coefficients, contact orders and motifs. We successfully used the connectivity information for predicting structural domains and measured the effect of local connections on the global connectivity. Also, a new method is described to discriminate protein conformations relevant to the specific recognition of a ligand. The method relies on a scoring system that matches critical residues with central residues for function in different structures of a given protein. Central residues are the most traversed residues with the same frequency in graphs derived from protein structures. We tested this method in a set of 24 different proteins and more than 260,000 conformers in bounded and unbounded states. To illustrate the usefulness of our method in the study of the structure-dynamics-function relationship of proteins, we were able to predict those residues of the TATA-box binding protein whose mutation impairs DNA binding. Our results indicate that critical residues for an interaction are preferentially found as central residues of protein structures in complex with a ligand. Thus, our method effectively distinguishes protein conformations relevant to the function of interest.

# CONTENTS

# ACKNOWLEDGEMENTS

I would like to acknowledge the direction, support and excellent exchange of ideas of my advisor Gabriel del Rio. His admirable impetus and range of interests were fundamental to the completion of this work.

Thanks to Nina Pastor and César Millán for their technical advice and for providing the top-class molecular dynamics simulations used in this work. They were an invaluable body of data where we could make a better testing of our ideas.

My sincere gratitude to Arturo Rojo and Lorenzo Segovia for all their ideas and their recommendations for improvement during all phases of this research. Thanks to Lorenzo Segovia for allowing us to employ his computing cluster for the motifs study.

My appreciation for the technical assistance received from the Information Technology core of the Instituto de Fisiología Celular-UNAM and the important help of Alondra Solares in the compilation of experimental reports of TBP mutants.

This work was funded by two grants from the Universidad Nacional Autónoma de México (UNAM) to Gabriel del Rio: PAPIIT-IN210705 and Macroproyecto UNAM: Tecnologías para la Universidad de la Información y la Computación; grants J33190-E from CONACyT, and the program "Cómputo Científico" (SEP-FOMES 2000) to Nina Pastor, which that provided computer access to the IBM-4 Regatta at the Universidad Autónoma del Estado de Morelos.

*A mis padres y hermanos.*

# INTRODUCTORY REVIEW

## I. The Molecular Study of Life

The study of Biology from the molecular perspective began during the 19th century, when advances in the construction of microscopes permitted to visualize for the first-time macromolecular structures inside cells, like the centrosome, which sudden apparition during mitosis was seen with great interest (Farmer 1898).

In 1926 Sumner working on the enzyme urease established that proteins are the actual molecules responsible for the chemical transformations performed by cells (Sumner 1926). Eighteen years later, in a classical experiment performed by Avery-MacLeod-McCarty, DNA was recognized as the molecule storing genetic information (Avery 1944). And from that epoch and onwards, the molecular description of the cell has advanced at a frenetic pace.

New discoveries that followed helped unraveling an intricated biomolecular world, full of complexity, orchestrated by specific interactions between molecules that seem to be taking place in a very-crowded environment[1]. Inside this milieu, proteins move and find their relevant targets, with enough frequency to produce significant effects on the cell. Once these effects are produced, the cell's internal environment change and is conductive to the

[1] Estimates of the total number of protein molecules that coexist in a mammalian cell have a lower bound of around one billion molecules (Lodish, et. al. 2000). We can link this to properties like the viscosity of molecules in liquid mediums. For example, for a small molecule, the viscosity of the cytoplasm of a cell type like fibroblasts, is around 2 centipoises (cP) (1 cP is the viscosity of liquid water at 20°C), but for a protein-size dextran (a carbohydrate that has minimal interaction with molecules inside a cell) the viscosity goes up to 4-5 cP. And for proteins like bovine albumin, which do have interactions with other cytoplasm proteins, the viscosity of the cytoplasm of a fibroblast increases to around 70 centipoises. For reference: mercury has 1.5 cP, whereas milk has 3 cP and olive oil has 81 cP. So just considering size, a molecule as big as a protein can diffuse through the cytoplasm with a little less freedom than an inert nanoparticle in milk, but when the interactions of a protein are considered, we can compare their movement to that of an inert nanoparticle in olive oil. (Brazma, et. al. 2001) (Mastro 1984)

subsequent molecular interactions. These chains of events occur accordingly to a blueprint codified in the DNA that is executed through the lifetime of every organism. A blueprint that has been constructed through selection, recombination and randomness.

Scientists have been collecting an immense amount of information about biomolecules. They have deposited this information in public databases expecting that they will provide a powerful toolbox to tackle some of the problems that arise when trying to understand the complexity of the cellular machinery[2]. Still, the functional annotation for the hundreds of complete genomes from diverse phyla remains relatively low, even for model organisms.

This thesis belongs to the field of Structural Bioinformatics. This field could be defined as the study of the data of positions and movements of the atoms that constitute the biomolecules by computational means. Nowadays, large teams of structural biologists solve hundreds of structures with considerable speed (including some of those considered as very difficult such as the ribosome, ion channels or the $F_0/F_1$ ATPase gained special notoriety), providing the raw data to which computational methods can be applied. In Structural Bioinformatics, much effort goes to the study of Protein structure data. During the following pages, I will describe different aspects of Protein Structural Bioinformatics. First, I will argue why analyzing three-dimensional structures by computational means provides insights about protein function and evolution. Then I will write about Protein Structure Prediction and Structural Genomics and I will discuss the relationship between Protein Structural Dynamics and protein function.

# II. Structure as a tool to elucidate Protein Function

Protein function is a concept that encompasses different levels and aspects of the actions of a protein. For example, in physiology one can say that the function of a given protein is to

[2] This complexity can be overwhelming. The protein p53, for example, was first discovered in 1979, and despite initially being misjudged as a cancer promoter, it soon gained notoriety as a tumor suppressor — a 'guardian of the genome' that stifles cancer growth by condemning genetically damaged cells to death. Few proteins have been studied more than p53, it even commands its own meetings. Yet the p53 story has turned out to be immensely more complex than it seemed at first. In 1990, several labs found that p53 binds directly to DNA to control transcription, supporting the traditional Jacob–Monod model of gene regulation. But as researchers broadened their understanding of gene regulation, they found more facets to p53. In 2009, Japanese researchers reported (Susuki, et. al. 2009) that p53 helps to process several varieties of small RNA that keep cell growth in check, revealing a mechanism by which the protein exerts its tumor-suppressing power. Even before that, it was clear that p53 sat at the center of a dynamic network of protein, chemical and genetic interactions. Researchers now know that p53 binds to thousands of sites in DNA, and some of these sites are thousands of base pairs away from any genes. It influences cell growth and death, and DNA structure and repair. It also binds to numerous other proteins, which can modify its activity, and these protein–protein interactions can be tuned by the addition of chemical modifiers, such as phosphates and methyl groups. Through a process known as alternative splicing, p53 can take nine different known forms, each of which has its own activities and chemical modifiers. Biologists are now realizing that p53 is also involved in processes beyond cancer, such as fertility and very early embryonic development.

induce cell proliferation, but in biochemistry one can say that the function of that same protein is to phosphorylate proteins at their tyrosine residues. It seems however, that it's not necessary to come up with a precise definition of the term function, but most likely with a framework that allows us to organize the related knowledge. The Gene Ontology (GO) consortium has adopted such framework. With it we can categorize the different aspects and levels of the activity of proteins in a broadly accepted manner and that is useful for understanding the cells and their evolution. In GO, three different aspects are considered and defined separately: the cellular localization (e.g. nucleus or ribosome); the biological process or pathway in which the protein is involved (e.g. metabolism, cell cycle); and the molecular function, defined as the ensemble of specific activities it can undertake (e.g. binding, transport). A hierarchy in GO that allows for further description represents these aspects.

The catalogues of sequenced protein-coding genes are filled with uncertainties in the annotation of the GO molecular function aspect. Having the three-dimensional structure of a protein can help to describe its function at different levels and it has proven to be a key element in the elucidation of important details of their molecular functions[3]. One example of this is the protein that forms the potassium channels in cell membranes. This class of proteins shows a seemingly counterintuitive activity: they permit the passage of potassium ions, whereas they block the passage of the equally charged but much smaller sodium ions. Before obtaining the three-dimensional structure, the detailed molecular architecture of such channels and the exact means by which they convey ions remained speculative. In 1998 however, and despite a barrier to the structural study of integral membrane proteins that had thwarted most attempts for decades, MacKinnon and colleagues determined the detailed structure of a potassium channel from a bacterium (Doyle 1998). With this structure in hand and other biochemical experiments, they could propose a mechanism by which potassium channel selectivity occurs: it appears that the filter -which is held in a very precise conformation- is more tuned for the larger potassium ion: when these ions enters the channel, water flows away, but for this to be energetically feasible, the pore must offer a surrogate for water. In this case, the surrogate role is carried out by oxygen atoms from the protein filter, which surround in a more coupled way this particular size of ion, making it transiently more stable.

Structural comparative studies are an important source of function assignment; it is assumed that when two proteins have significant structural resemblance their respective

[3] The structure of the DNA is perhaps the most famous example but that was also the case for haemoglobin - the second protein structure to be solved -. After the initial structure of the haemoglobin was solved by Max Perutz and his colleagues in 1959, the high-resolution 2.8 Å structures of horse oxyhaemoglobin in 1968 and deoxy haemoglobin in 1970 finally provided atomic models which Perutz could use to explain the mechanism of the cooperative binding of oxygen. However, Perutz's theory depended on small changes in the displacement of the iron atoms from the haem plane when oxygen was bound, which was then conveyed through the proximal histidine to the subunit surface where salt bridges between the subunits were broken or weakened. In the absence of oxygen, stronger salt bridges stabilized the deoxy state where the iron atoms were pulled out of the haem plane by the proximal histidine bond. The key concepts were tension at the haem, salt bridges at the subunit interface, and their coupling through internal Van der Waals interactions inside each subunit. The overall change in energy has remained too small to calculate theoretically, but, by the mid 1980s, with improved higher-resolution structures using synchrotron radiation data, the movements of the iron atoms were measured beyond doubt. Perutz's proposed mechanism was essentially correct (Perutz 1990).

functions may be also similar. This hypothesis has an impact in current drug discovery and the understanding of cellular processes but should be applied carefully since there are several examples where this is not the case[4]. Nevertheless, it is common practice to assign the known molecular function of a protein to all the other proteins to which its structure is similar. Even if the global structure similarity is not high, it is possible to find common 3D motifs that can point to a putative function; this is especially true for the active sites of certain enzymes (Lee 2011). In other cases, the functional assignment based on structure can be extended to the fold level. For example, if a given fold contains just families grouped into one functional superfamily in the SCOP database, or if all functionally annotated PFam families from one-fold were grouped into one PFam clan (a group of proteins families with evidence of common ancestry) linked to a single functional category, then it can be assumed that any protein in question that adopts this fold is likely to have a similar function to that associated with that fold. A greater challenge is presented in the functional annotation of multi-domain proteins since even if their individual domains have been functionally characterized, the actual molecular function of the complete protein may be difficult to infer. In this case knowing the 3D arrangement between domains can help to understand how is that they interplay in a global multi-domain molecular function (Shuman 2004).

Structure studies have also helped to discover that proteins are not the only biomolecules that carry on interesting functions. In another very famous example of structure-assisted function elucidation, the structure of the ribosome was solved in the year 2000. One of the oldest and biggest macromolecular complexes, the ribosome was intensively studied for its central role in the cell, but the determination of its structure presented big challenges because of its size. However, the groups of Ramakrishnan (Wimberly B. 2000), Steintz (Ban N. 2000), Yonath (Schluenzen F. 2000) succeeded in solving the complete structure of its two subunits. This helped them to conclude that the main catalytic task of the ribosome is to provide a template for the precise positioning of tRNA molecules, rather than to participate in the actual chemical reaction, and that this activity is performed exclusively by rRNAs since only their atoms are in contact with the substrates during the different positions that they take in the ribosome sites (in fact, the ribosome crystals used for determining the structure were able to catalyze the reaction and allowed to visualize some of the involved steps), thus providing affirmative support for the existence of a pre-protein RNA world. Also, the surprising identification (Agmon I. 2005) of a two-fold rotation axis in the peptidyl transferase center of all known ribosomal structures led to a proposal of how this machinery might work for peptide bond formation, translocation and nascent protein progression. This symmetry is totally absent at the sequence level, and thus, is a good example of the kind of insight that can be gained by analyzing three-dimensional structures.

[4] Counterexamples are the pairs of homologous enzyme/non-enzyme proteins that have maintained highly similar structures. Commonly, the non-enzyme protein descends from an enzyme by loss or blockade of catalytic residues, but interestingly, there are cases where the enzymes are descendants of non-enzyme proteins (Todd 2002).

# III. Protein Structures as Guides for Discovering Evolutionary Relationships

The genetic code is constructed in such a way that has allowed the cells to produce an enormous number of different proteins with a low diversity in the chemical groups of the molecule that stores that information, the DNA. The potential number of proteins that can be produced with this code is truly astronomical. In contrast, the actual number of protein folds that exist today is several orders of magnitude below this number[5]. Some estimates put it in the order of a few thousands (Grant A. 2004). That means that the genetic code has been used in a very redundant way in terms of the number of different folds implicitly coded by it[6].

Since structure similarity is an indicative of common ancestry, the magnitude of the expected number of different folds has fueled the hope of obtaining soon a structural classification that describes the evolutionary relationships of almost all known proteins. In general, protein pairs with a sequence identity higher than 30% are close homologous and very likely to be structurally similar. When the sequence similarity falls in the range 20-30% (referred as the "twilight zone") the structural similarity is considerably less common. And it has been estimated that around 5% of proteins pairs with sequence identity below 20% have similar structures[7]. These are general numbers, so it can be expected that the relationship between sequence and structure vary depending on the particular fold or super-family in

[5] Various authors have investigated the original source of the current fold diversity. Some reports point out that before the branching of eukarya, bacteria and archaea, fold diversity was achieved not by gene duplication but primarily by shuffling small exon units whose sequences were originated randomly (Dorit 1990). A team has reported that the structures in the α/β class were the first to appear but that have been steadily superseded (in terms of abundance) by structures in the α+β class, which they conclude was the following to appear. The all- and all-β classes originated later maybe as the result of less aggressive conditions that allowed the stability of not so rigid structures (Caetano-Anollés 2003). Other authors propose that the ancestral proteins of the α+β class were formed at different times from previous pieces of proteins from the all- and all-β classes (Alva 2010).

[6] In nature, there are variations of the genetic code, but they are similar enough for this argument to remain valid (Santos M 2004). It is very difficult to estimate how big the number of different functional folds would it be if all the possible single domain medium-sized proteins were produced, especially since it is not known to what extent the fold universe is constrained by folding/stability requirements and functional adequacies.

[7] Molecular evolution may eliminate easily recognizable sequence similarity among protein genes that diverged a long time ago, but still it may leave behind traces of statistically significant patterns of conserved residues that are apparent only when multiple, related sequences are aligned. To reflect the concept of different degrees of divergence between proteins genes, proteins are often subjected to a multilevel classification, with the term *family* reserved for groups of proteins related by short evolutionary distances so that any two proteins inside the family retain identifiable traces of similarity in their primary sequences. After this, it is possible to bring together two or more families into one *superfamily* (*clan* is the term used in PFam) if there are some members of the different families that retain identifiable traces of similarity between them, e.g. there is a transitivity property in the common ancestry relationship, the key difference with respect to the family definition is that inside a superfamily no any two members retain identifiable traces of sequence similarity between them. Sometimes the sequence based superfamilies are overlapped/complemented by structural based superfamilies that bring together families with enough common structural features that it is inferred they come from a common ancestor. In structural classifications, there exists a further level up: the *fold,* which groups structurally similar superfamilies but with different characteristic features so there is no certainty that two superfamilies inside a fold come from a common ancestor. The fold level is more an organizational need than a proved evolutionary reality. We can expect that further development of even more sensitive algorithms for recognition of distant homologues would expand the list of superfamilies groupings.

question or in the algorithm for sequence similarity and the measure of structure similarity chosen.

When the sequence similarity between two proteins is inside or below the twilight zone, there is no straightforward sequence-based procedure for detecting evolutionary relationships. Protein structure represents a powerful means of discovering distant common ancestry since structure similarity is conserved to a greater extent through evolutionary time as sequences diverge from each another. There are celebrated cases of homology inferred from structure, including the unexpected similarity between actin and the 70-kDa heat-shock cognate protein (Flaherty 1991), or the Top Rim domain shared between some topo-isomerases, primases and nucleases (Aravind 1998). Current evolutionary driven classifications (CATH, SCOP, Superfamily, Gene3D, PFam) establish that two proteins are homologous only if there's enough sequence, structural or functional evidence.

As for the detection of very-far evolutionary relationships, the development of novel comparative approaches (Alva 2010) has permitted to find signals of common ancestry for certain superfamilies/folds. Also, in recent studies (Cuff 2009), it has been recognized that the clustering of proteins into distinct superfamilies related by common structural features may be somewhat artificial, since for some parts of the structural space the divisions are not clear-cut and resemble more of a continuum. Around 14% of the structures in CATH contribute to this overlapping between superfamilies. In addition, some of the more populated superfamilies (accounting for 4% of the total number of superfamilies in CATH) have significant structural divergence and together comprise one quarter of the structural diversity if an RMSD threshold of 5Å groups all the structures. These findings reflect some of the difficulty in identifying and defining protein homology.

There exist, however, a typical range where structure and sequence similarities can be equaled, and this have permitted the transference of structural knowledge between proteins with enough sequence similarity. Even when a protein with a given sequence can adopt different conformations, the odds that two close sequences will fold into significantly different structures are so small[8] that are often neglected in practice.

[8] There are some examples where a few point mutations change the overall fold of a protein (Alexander 2009), normally these mutations are substitutions of cysteine residues that form disulfide bonds. Several proteins - most notably lysozymes from humans, hen egg-white and phage T4 (over 900 structures in all) - have been systematically mutated, crystallized and structurally determined to explore the effects of single amino-acid changes on protein structure and stability. The net conclusion from these examples is that most changes have little effect on the fold, although they may modify the protein's stability, function or rate of folding to varying degrees (Matthews 1993) (Sinha 2001). In most cases the modified residue is accommodated by slight shifts in nearby side chains and adjustments to the protein backbone. Proteins also tolerate certain insertions that are engineered artificially, as demonstrated by a study using linker insertion mutagenesis. Here many 5-residue insertion mutants of the α-complementing domain of Escherichia coli β-galactosidase were generated. The insertions were made, essentially at random, along the length of the sequence using bacteriophage μ *in vitro* DNA transposition (Poussu 2004). Most insertions were tolerated - that is, they did not prevent the protein folding - even those that disrupted secondary structure elements and those that ended up within the protein's interior. As many as half of the mutants showed β-galactosidase activity at least equivalent to that of the wild type, and with activities up to two-fold higher in some mutants. A small but growing number of "metamorphic" proteins adopt different folded conformations for the same amino acid sequence under native conditions. Unlike prions, they undergo reversible conformational changes. And unlike allosteric modulation or changes upon binding, metamorphic proteins are capable of independent interconversion that may result in

So far, I have described the relevance of having the three-dimensional structure of a protein for establishing its function and evolutionary relationships. I will end this section by briefly mentioning that, in cases where the protein in question has a recognized therapeutic or technological use, obtaining its three-dimensional structure is an important step in the process of finding effective drugs that alter its function or in its redesign for further applications.

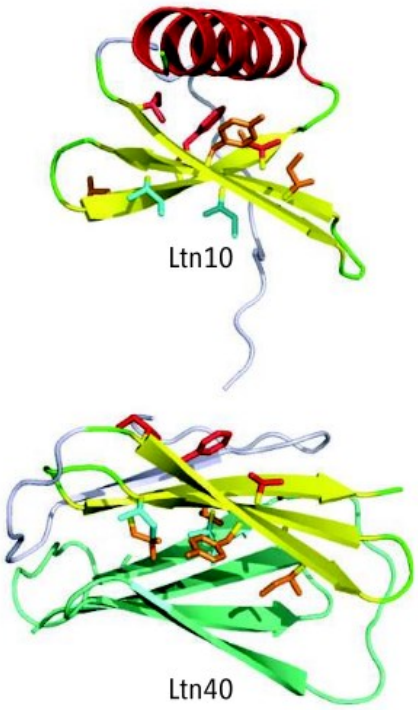


FIGURE 1. Metamorphic proteins. The chemokine lymphotactin (Ltn) adopts two distinct folds at equilibrium in physiological conditions, and interconversion between the conformers involves almost complete restructuring of its hydrogen bond network and other stabilizing interactions. One conformer, Ltn10, adopts the canonical chemokine fold and binds natural Ltn receptors. The other, Ltn40, forms a dimeric β-sheet sandwich and binds to heparin, a polysaccharide component of the extracellular matrix. These mutually exclusive activities of the two conformers are both essential for full Ltn function in vivo. Structurally equivalent residues are few and contribute either to the Ltn10 core (red) or to the dimeric interface of Ltn40 (cyan). Other non-polar residues (orange) change sides, such that the formation of the dimeric interface on one side of the β sheet destabilizes the hydrophobic core on the other side and vice versa (Tuinstra 2008).

---

an abrupt fold change (FIGURE 1). The existence of multiple folded conformations is not prohibited by the principles of physics and chemistry. However, in vivo, a protein must quickly form its biologically active conformation, and stable alternative folds would act as kinetic traps that slow the rate of protein folding. To avoid this complication, different folded states should be able to interconvert without going through a fully unfolded state. Also, there have been recent discoveries of protein families containing members with distinct folds. For example, the protein RfaH of *E. coli* is composed of two domains. The N-terminal domain displays high similarity to that of its paralog NusG, a general transcription factor. In contrast, the α-helical coiled-coil C domain, while retaining sequence similarity, is strikingly different from the β barrel of NusG. Such an all-β to all-α transition of the entire domain is an extreme example of protein family divergence (Belogurov 2007).

# IV. Protein Structure Prediction

The experiments of Anfinsen (Haber 1961) in the sixties showed that for some globular proteins, the protein structure is determined by their sequence and the surrounding aqueous environment alone. Since then, scientists have been developing an enormous array of approaches to predict protein structures. Nowadays, Computational Protein Structure Prediction can help to assign a three-dimensional structure to a vast number of proteins, and, as progress is being made, it may aid to the modeling of domain families that don't have a representative structure. There are two broad categories in which most structure prediction approaches fall: comparative structure modeling (CM) and Ab Initio structure prediction (or Free Modeling FM). Comparative structure modeling methods are based on the availability of known structures of conserved homologues for the target protein. Once the templates are identified by sequence or profiles comparisons, using the framework of the template structure and refining the details, like side chain conformation and spatial restraints, can generate a high-resolution model. Ab Initio methods by contrast, are focused on predicting structures for proteins that don't have conserved homologues with proteins with known three-dimensional structures.

In the last years, the boundaries between both categories have become increasingly blurred. Much of the state-of-the-art Ab Initio modeling algorithms use evolutionary or knowledge-based information for collecting spatial restraints or for identifying local structural building blocks or fragments. Recent community-wide critical assessments of protein structure prediction (CASP) experiments have shown the advantages of this class of composite approaches.

Solving a new protein structure by experimental means remains a difficult task. However, at the turn of this century, improvements in cloning, protein expression in heterologous systems and protein purification by affinity chromatography has significantly increased the ability to obtain microgram to milligram quantities of protein needed for structure determination. Recently, technological advances in cryo-electron microscopy have significantly increased the resolution, speed and quality of the data collection, expanding the class of proteins available to such kind of structure determination. Likewise, developments in improved crystallization screening methods that enhance the ability to obtain and optimize protein crystals, the invention of cryo-cooling techniques to obtain better quality data from single crystals, the increased brightness, stability, availability and "user-friendliness" of synchrotrons and the genetic engineering methods that facilitate heavy-atom integration into the target of interest, have contributed at continuing the expansion of the range of proteins that can be subjected to X-ray crystallography, extending it to smaller and smaller crystals. In parallel, advances in the "dry lab", such as continued development and improvement of the corresponding software, have increased the speed, reliability and quality of structure determination.

Structural Genomics (SG) is the given name to a recent series of high-throughput structure-solving efforts carried on by specialized centers around the globe and coordinated by consortiums that have been succeeding in providing thousands[9] of new structures in the past ten years. One of the goals of the SG centers is to obtain the 3D structures of at least one representative of as much single-domain protein families as possible, and then generate models for similar proteins by CM building[10]. This goal has received a boost from the observation that, even when the addition of new sequences maintains its pace, the rate of newly discovered single-domain families appears to be diminishing (contrary to the case of newly discovered multiple-domain architectures, whose rate seems to be increasing). Many of the advances at synchrotron beam-lines have occurred in partnership with Structural Genomic centers and likewise many of the available crystallization robots and protein expression and purification devices and platforms, as well as software packages and tools, have their origin in the Protein Structure Initiative (PSI) and other SG centers.

In SG, target selection is a key aspect when the intention is to solve as much structures as possible using a CM criterion[11]. One natural set of targets is the Domains of unknown function, or DUF. Interestingly, a number of these families are present in all kingdoms of life. In Pfam release 24.0 there are 3,067 DUFs and the fraction of DUF families had increased with every release to about 25% of all families. As expected, the number of DUFs is increasing mainly because of the large number of new genomic and metagenomic sequences that have many new clade-specific families. SG centers have solved the structures of around 250 of these DUF families. And in some cases, this has helped to narrow down the possible function of some of them[12]. For example, some of the structures were co-

[9] However, the most of the almost 5,000 protein structures determined by these centers have yet to be described in the peer-reviewed literature. In a high-throughput structural genomics environment, the process of structure determination occurs independently of any associated experimental characterization of function, which creates a challenge for the annotation and analysis of structures and the publication of these results. Developments like TOPSAN (The Open Protein Structure Annotation Network), enables the generation of knowledge via collaborations among globally distributed contributors supported by automated amalgamation of available information.

[10] For example, based on 53 newly solved proteins from SG projects, Sali and coworkers (Pieper 2006) built reliable models for domains in 24,113 sequences from the UniProtKB database with their CM tool MODELLER. These models have been deposited in a CM model database, MODBase (http://salilab.org/modbase). MODBase contains around 18 million models or fold assignments for domains from 3.3 million sequences. In this study, the structure assignments were based on an all-against-all search of the amino acid sequences in UniProtKB using the solved structures of PDB. Structural genomics can also benefit from improvements in high-resolution structure prediction algorithms. A study estimated that a 10% decrease in the threshold needed for accurate modeling, from 30 to 20% sequence identity, would reduce the number of experimental structures required by more than a factor of two (Vitkup 2001).

[11] In 2001 Vitkup and collaborators estimated that at least 16,000 new structures needed to be determined by experiments to ensure that CM could generate good structures for 90% of single-domain protein families. For their calculation, they supposed that CM technology generates a good model when there is a sequence identity of at least 30% with 80% alignment coverage (Vitkup 2001).

[12] Jaroszewski and collaborators analyzed the structures of 248 of these families and found that 67 had new folds. One question related to those DUF structures that adopt a previously known fold, is whether they diverged from already known families and, therefore, can be classified into already known clans or super-families, or whether they are examples of convergent evolution. While rigorous proof of homology is often difficult, usually the combination of several arguments enables to arrive at a satisfactory answer. After closer structural and sequence-based comparisons, they classified the DUFs in (i) ''recognizable homologues'', when both sequence and structure similarities are significant, (ii) ''putative homologues'', when there is structural similarity but only marginal sequence similarity, (iii) ''putative analogs'', when there is structural similarity but cannot be connected to previously known proteins of the same fold with current day, state-of-the-art, sequence-based remote homology recognition methods, and (iv) "new folds", when there is no structural nor sequence similarity to previously known folds. Since the profiles of structural similarities in the first three groups are very similar,

crystallized with a putative natural ligand, others had a distant evolutionary relationship with members of a characterized family thus forming part of the same superfamily, and others were multiple-domain architecture families that contained already characterized domains.

Before the first X-ray protein structures appeared, protein structure was visualized in terms of analogies based on organic chemistry and symmetry. In light of these reasonable expectations, the low-resolution X-ray structure of myoglobin came as a considerable surprise. Kendrew in describing his model said: "Perhaps the most remarkable features of the molecule are its complexity and its lack of symmetry, the arrangement seems to be almost totally lacking in the kind of regularities which one instinctively anticipates". Perutz was even clearer about his initial disappointment "Could the search for ultimate truth really have revealed so hideous and visceral-looking an object". In the last fifty years, we have learned to appreciate the aesthetic merits of protein structure. Helical models are indeed kind of visceral and electron-density maps are like intricate branched corals, but still there are elegant patterns and arrangements between secondary structure elements that were selected based on their own idiosyncratic function: "propellers", "barrels", "knots", "zippers", "sandwiches", all of them coexist with more irregular features to provide an enormous spectrum of diversity.

The notion of protein structure classification emerged from several studies conducted during the late 1970s and early 1980s that aimed to elucidate the basic principles of protein folding and protein structure evolution. The early work of Chothia and coworkers pioneered the division of protein structures into four major classes based on their secondary-structure composition and showed that simple geometrical features of secondary-structural elements give rise to their mutual arrangement in distinct architectures (Chothia 1977). Later, a more detailed classification deduced from the topological details of less than 200 structures was proposed (Richardson 1977). By the end of the 1980s, the term "fold" was already established and it was intended to outline three major aspects of protein three-dimensional structure: the secondary structures of which the protein is composed of, their relative arrangement and the path taken through the structure by the polypeptide

---

they suggest that most proteins from the group of putative analogs may be, in fact, distant, but not readily recognizable, homologues of previously characterized protein families. At the same time, some of the putative analogs, especially those that consist of a small number of secondary structure elements, such as a-helical hairpins, probably arose from convergent evolution. Over a third of the new folds contained fragments with significant structural similarity to fragments of known proteins that adopt different overall folds. The presence of some structural similarity among sections of different folds has been recognized for some time, and some authors suggest that, in most cases, it has its origin in the general evolution of protein structure (Friedberg 2005). Upon closer examination, they found that this is not only the case for new folds from DUF families but also holds for many recently solved proteins that were identified to have new folds. The percentage of new folds with some structural similarity to another fold has grown to almost 30% in the last 2 years. These sub-fold similarities have a discrete distribution, so the finding does not necessarily argue for (or necessarily disprove) a continuum in protein fold space. Their explanation of this phenomenon is that, with an increasing number of known protein structures, there is a saturation of the available fold space at the level of micro-domains that represent shorter, usually compact, structures that become component pieces of different folds. This concept of structural pieces or fragments that are treated as sorts of structural alphabets have been incorporated to the most successful free modeling protein structure prediction methods. But, as words in human speech, no common ancestry is assumed between proteins sharing some number of letters in this structural alphabet. (Jaroszewski 2009).

chain. Thus, the fold of a protein was defined through its composition, architecture and topology (Chothia 1990).

It was thought at that time that the number of architectural types was limited. Moreover, although some structural variations were observed among evolutionarily related proteins, none of these affected the common structural core. Therefore, it was assumed that the protein fold is evolutionarily stable in that it retains its features, although some structural variations could be anticipated. Similarly, it was thought that in general every protein folds into a single three-dimensional structure and that its structural core is insensitive to large conformational changes related to function or formation of quaternary structure. These themes have been increasingly challenged as more and more structural variations are observed in protein families and in certain individual metamorphic proteins.

Before the advent of structural genomics and the Protein Structure Initiative, analysis on the trends of newly discovered folds seemed to indicate that much of the protein fold space had been explored. More than one structural representative was solved for most of the characterized families (Andreeva 2008) and the growth in the number of new folds in SCOP had almost stalled. But after the SG and PSI launch, the number of new folds, super-families and families rose again, mainly because the PSI SG targeted proteins with no significant sequence similarity to known structures. Recent analysis of the distribution of protein families characterized by structural genomics has confirmed the dominant role of the largest known super-families, which have grown further in their number of constituent families (Andreeva 2008). In addition, other super-families have grown large rather unexpectedly. The evolutionary success of these "new rich" super-families is probably a consequence of the presence of unusual conserved and presumably functionally important features in their folds. One of these "new rich" super-families, is the dimeric $\alpha+\beta$ barrel super-family in SCOP, several new members of which have come from the first structures of meta-genomic sequences.

Initially, it was anticipated that a large number of new folds would be discovered owing to the breath of coverage of fold space targeted by the PSI. Interestingly, this has not turned out to be the case as a substantial portion of the structures expected to have novel folds revealed significant structural similarities to already known folds and in fact represent variations of known protein architectures and topologies. However, there were several unexpected findings of previously unseen topologies and architectures[13]. PSI also greatly increased the number of protein topologies with high contact order (3D contacts between amino acids that sit far away from each other in the sequence), which is known to limit the success of current Ab Initio structure-prediction methods, thus providing invaluable high-

[13] The structural data delivered over the past decade by SG and independent groups has revealed examples of atypical structural features and structural variations that have challenged many longstanding tenets in protein structure. Amongst these, for instance, is the discovery of the deep trefoil knot (Nureki 2002). SG has determined the structures of several knotted proteins, which in turn helped to dispel one of the oldest dogmas in molecular biology, since it was believed that the process of protein folding could not efficiently produce deep knots in protein backbones.

resolution templates for modeling. Without previous preconceptions, comparisons of some SG structures revealed dramatic structural variations in related proteins that go beyond the expectations based on their sequence similarity. These provided examples of how protein folds can evolve without compromising the integrity of the structure of the functional site.

The biological usefulness of a predicted protein model relies on the accuracy of the structure prediction. For example, high-resolution models with root mean square deviation (RMSD) values in the range of 1–2 Å, typically generated by CM using close homologous templates, can be suitable for computational ligand-binding studies and virtual compound screening. Medium-resolution models, roughly in the RMSD range of 2–5 Å and typically generated by threading and CM from distantly homologous templates, can be used for identifying the spatial locations of functionally important residues such as active or binding sites. However, many of the functionally important sites are located on the loop regions that show more structural variability than helices and sheets. Thus, accurate modeling of loop regions is still an important, yet unsolved problem in template-based modeling. Finally, even models with the lowest resolution, from an otherwise meaningful prediction, i.e., models with an approximately correct topology, predicted using either Ab Initio approaches or based on weak hits from threading, have several uses including protein domain boundary identification, topology recognition and family/super-family assignment.

There are two critical problems in the field of protein structure prediction. The first problem is related to the template-based modeling: How to identify the most suitable templates from known protein structures in the PDB library? Furthermore, following template structure identification, how can the template structures be refined to better approximate the native structure? The second major problem is related to free modeling for the target sequences without appropriate templates: How can a correct topology for the target proteins be constructed from scratch? Progress made in these areas has been assessed in the CASP experiments under the categories of template based modeling (TBM) and free modeling (FM), respectively.

In the following sections, current protein structure prediction methods will be reviewed for both template-based modeling and free modeling. The basic ideas and advances of these directions will be discussed.

## IV.1 Template-Based Predictions

For a given target sequence, template-based prediction methods build 3D structures based on a set of solved 3D protein structures, termed the template library. The canonical procedure of template-based modeling consists of four steps: (1) finding known structures (templates) related to the sequence to be modeled (target); (2) aligning the target sequence on the template structures; (3) building the structural framework by copying the aligned regions, or by satisfying spatial restraints from templates; (4) constructing the unaligned

loop regions and adding side-chain atoms. The first two steps are usually performed as a single procedure because the correct selection of templates relies on their accurate alignment with the target. Similarly, the last two steps are also performed simultaneously since the atoms of the core and loop regions interact closely.

Historically, template-based methods can be categorized in two types: (1) comparative modeling (CM) and (2) threading. CM builds models based on evolutionary information between target and template sequences, while threading is designed to match target sequences directly onto 3D structures of templates with the goal to detect target-template pairs even without evolutionary relationships. The schematic overview of CM and threading is depicted in the upper part of FIGURE 5. In recent years, as a general trend in the field, the borders between CM and threading are becoming increasingly blurred since both comparative modeling and threading methods rely on evolutionary relationships, e.g. both use sequence profile-based alignments. In CASP experiments they are placed in the same category of template-based modeling without explicitly distinguishing them.

### IV.1.1 Template Structure Identification

Since its first application in the early 1990s (Bowie 1991) (D. T. Jones 1992), threading has become one of the most active areas in proteins structure prediction. Numerous algorithms have been developed during the previous twenty years for identifying structure templates from the PDB. Threading techniques include sequence profile–profile alignments (Ginalski 2003) (Zhou 2005), structural profile alignments (Shi 2001), hidden Markov models (HMM) (Karplus 1998) (Soding 2005), and machine learning (D. Jones 1999) (Cheng 2006) among others.

The sequence profile–profile alignment (PPA) is probably the most often-used and robust threading approach. Instead of matching the single sequences of target and template, PPA aligns a target multiple sequence alignment (MSA) with a template MSA. The alignment score in the PPA is usually calculated as a product of the amino acid frequency at each position of the target MSA with the log-odds of the matching amino acid in the template MSA, though there are also alternative methods for calculating the profile–profile alignment scores (Sadreyev 2003). Profile–profile alignment-based methods demonstrated advantages in several recent blind tests where several sequence profile-based methods were ranked at the top of single threading servers (Battey 2007).

HHsearch (Soding 2005), a HMM–HMM alignment method, is distinguished as one of the bests threading servers. The principles of the HMM–HMM alignments and the profile–profile alignments are similar in that both attempt pair-wise alignments of the target MSA with the template MSA. Instead of representing the MSAs by sequence profiles, HHsearch uses profile HMMs that can generate the sequences with certain probabilities (determined

by the product of the amino acid emission and insertion/deletion probabilities). HHsearch aligns the target and template HMMs by maximizing the probability that two models co-emit the same amino acid sequence. In this way, amino acid frequencies and insertions/deletions of both HMMs are matched in a better way. HHsearch also deals with non-homologous sequence stretches at the ends of correctly aligned homologous regions. These stretches can recruit many more unrelated segments in further search iterations, thus HHsearch simply prunes away the ends of sequences to be included in the alignment if their score per column is below a specified threshold.

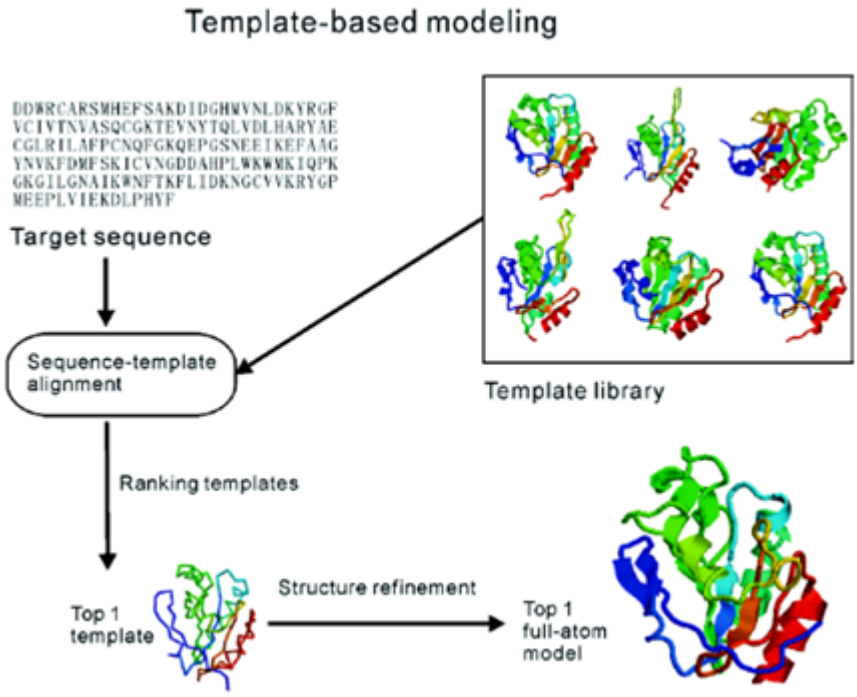


FIGURE 5. Schematic overview of the methodologies employed in template-based modeling.

In addition to sequence profiles, top-performing structure prediction methods compare the predicted secondary structure of the query protein with the actual secondary structure of the candidate template proteins. Such 1D properties defined for each position have a big advantage: their similarity scores can be combined with the similarity score between profile columns in the dynamic programming algorithms that calculate the optimal alignment. Hence 1D similarity scores may improve both the sensitivity of fold recognition and the alignment quality. Although secondary structure has had the largest impact, many other 1D scores have been proposed. Other scores that have become widely used recently are predicted solvent accessibility, predicted number of tertiary residue-residue contacts ("coordination number") and 1D environmental fitness scores, which evaluate how well the amino acid distribution at each query position would fit into the structural environment at

each template position. Profile column scores ignore correlations between columns. In contrast, 1D predictions are done on context windows. Comparing 1D predictions therefore amounts to scoring the similarity of local amino acid patterns, which may contain strong inter-column correlations.

### IV.1.2 Meta-Servers

Although average performance differs among threading algorithms, there is no single threading program which outperforms all others on every target. This motivated the use of the meta-server (Fischer, Rychlewski L, Fischer D 2003), which collects and combines results from a set of existing threading programs. There are two ways to generate predictions in meta-servers. One is to build a hybrid model by cutting and pasting the selected structure fragments from the templates identified by threading programs. The combined model has on average larger coverage and better topology than any single template. One defect is that the hybrid models often have non-physical local clashes. The second approach is to select the best models based on a variety of scoring functions or machine-learning techniques. This approach has emerged as a new research area called Model Quality Assessment Programs (MQAP) (Fischer 2006). Despite considerable efforts in developing various MQAP scores, the most robust score turns out to be the one based on the structure consensus, i.e. the best models are those simultaneously hit by different threading algorithms. The idea behind the consensus approach is simple: there are more ways for a threading program to select a wrong template than a right one. Therefore, the chance for multiple threading programs working collectively to make a commonly wrong selection is higher than the chance to make a commonly correct selection.

In later experiments, the Zhang-Server - an automated server based on profile/profile threading and I-TASSER structure refinement (Y. Zhang 2008) - outperformed the meta-servers, which included it as an input. This highlighted the challenge of the MQAP methods in correctly ranking and selecting the best models since the performance of a consensus method depends on the performance of individual servers and also their correlation. In principle if the individual servers are highly correlated or the best individual server is significantly better than the others, it is possible that a simple clustering/consensus method may not perform better than the best individual server. However, when there are a few very good and independent individual servers, a simple clustering/consensus method may perform better than the best individual server.

## IV.1.3 Template Structure Assembly/Refinement

The goal of protein structure assembly/refinement is to draw the templates closer to the native structure. This has proven to be a non-trivial task. Until only a few years ago, most of the TBM procedures either kept the templates unchanged or drove the templates away from the native structures. Early efforts on template structure refinement have relied on molecular dynamics (MD) based atomic-level simulations; these attempts to refine low-resolution models using classic MD programs such as AMBER and CHARMM. However, except for some isolated instances, this approach has not achieved systematic improvements.

Good template refinements have been achieved by combining the knowledge and physics-based potentials with spatial restraints from templates (Misura 2006) (Y. Zhang 2008). The group of Baker first built low-resolution models with ROSETTA (Simons 1997) using a fragment library enriched by the query-template alignment. The Cβ-contact restraints are used to guide the assembly procedure, and the low-resolution models are then refined by a physics-based atomic potential.

The Zhang server for example, incorporates composite knowledge-based energy terms that have been optimized using large-scale structure decoys (Zhang. 2007). This approach helps to coordinate the complicated correlations of different interaction terms. Another feature of this server is that the force field includes multiple sources of knowledge-based potentials and consensus tertiary restraints from multiple templates. The consensus spatial information usually has higher accuracy/confidence than that of individual templates. In the CASP7 experiment for example, this server managed to achieve a higher GDT-TS[14] score than the best possible structural template (or "virtual predictor group") in more than half the assessment units and a higher GDT-HA[15] score in approximately one-third of cases (Kopp 2007). This comparison may not entirely reflect the template refinement ability of the algorithms because the predictors start from threading templates rather than the best structural alignments; the latter requests the information of the native structures, which were not available when the predictions were made. However, a global GDT score comparison may favor the full-length model because the template alignment has a shorter length than the model. If the best possible template is provided (from a template library with a maximum of 35% pairwise sequence identity), this server can generate models of excellent quality for

[14] The GDT-TS (global distance test - total score) score is intended as a more accurate measurement than the more common RMSD metric, which is sensitive to outlier regions created by poor modeling of individual loop regions in a structure that is otherwise reasonably accurate. For example, the RMSD of two protein structures can be high if the tails or some loops have a different orientation even though the global topology of the core part is the same; this cannot be distinguishable, based on the RMSD value alone, from the case where two structures have completely different topologies. The GDT score is calculated as the largest set of amino acid alpha carbon atoms in the model structure that fall within a defined distance cutoff of their position in the experimental structure. It is typical to calculate the GDT score under several cutoff distances, and scores generally increase with increasing cutoff. A plateau in this increase may indicate an extreme divergence between the experimental and predicted structures, such that no additional atoms are included in any cutoff of a reasonable distance.

[15] The high accuracy version of the GDT measure is called GDT-HA. It uses smaller cut off distances (half the size of GDT_TS) and thus is more rigorous.

most proteins and models of good quality for the most divergent sequences, thus reassuring the notion that the crucial step in protein structure prediction is finding the best possible templates.

Sometimes homology models are not sufficiently reliable to accurately predict small but important features. For example, when no 3D structure of a splice variant is known, it is common to build a homology model to see what the structural difference might be. An example of this is the neuronal protein aczonin (FIGURE 6). A homology model of a splice variant of its C2A domain - containing a nine-residue insertion relative to the normal isoform - had suggested that the structure should be mostly unaffected by the insertion. The additional nine residues were predicted to add a surface loop some distance from the crucial Ca2+ binding site of the protein. Solving the structure of this longer isoform by NMR (PDB code 1rh8) showed that the structural change is in fact dramatic. The inserted sequence became part of a β-sheet in the core of the protein, displacing the segment that had previously been there and causing the displaced sequence to adopt a helical conformation on the protein's surface that severely disrupted the Ca2+ binding site. So, rather than the minor effect that was predicted from the homology model, the splice variant had a marked effect on the structure and consequently reduced its Ca2+ binding affinity. This shows the possible pitfalls of just relying on CM to infer the effects of even fairly small structural changes. Nevertheless, because of the scarcity of structures of variants, it is necessary to improve to predictive methods to try to understand the role of alternative splicing in eukaryotes.

## IV.2 Free Modeling

When structural analogs do not exist in the PDB library or could not be detected by threading (which is more often the case), the structure prediction must be generated from scratch. This type of prediction has been termed Ab Initio or de novo modeling, a term that may be easily understood as modeling "from first principles". Since CASP7, it is termed free modeling, which more appropriately reflects the status of the field, since the most efficient methods in this category still consider hybrid approaches including both knowledge-based and physics-based potentials. Evolutionary information is often used in generating sparse spatial restraints or identifying local structural building blocks.

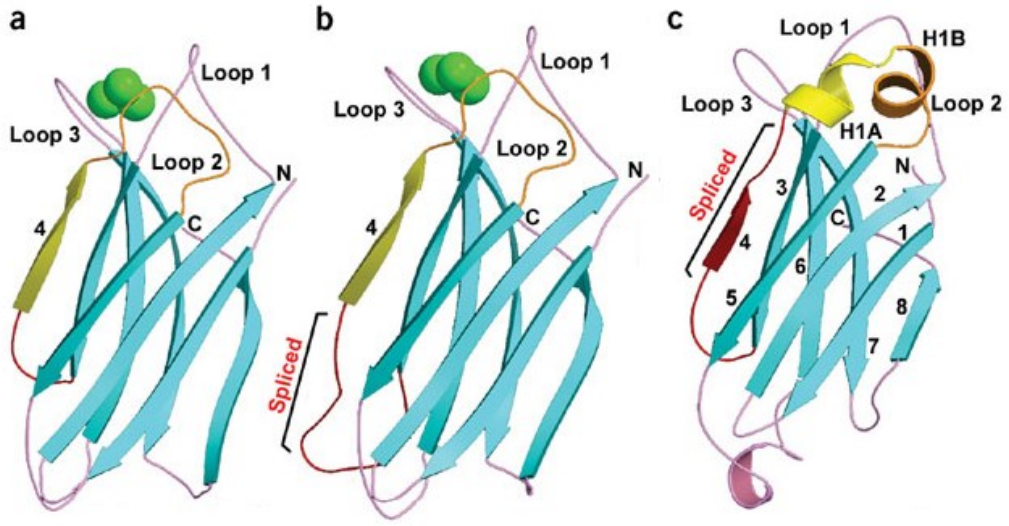


FIGURE 6. Pitfalls associated to homology-built models. In this example, the C2A domain of aczonin has an alternative splice variant, which has a nine-residue insertion relative to the normal form. Homology modeling suggested this would have little or no effect on function. However, the experimentally determined structure showed this to be far from the truth. (a) The structure of the synaptotagmin-1 $C_2A$ domain served as template for the CM of the insertion-less isoform. (b) Homology model of the insertion variant of aczonin. The model suggested that the insertion should merely insert a loop (shown here in red) on the protein's surface, far from the Ca2+ binding site (defined by the loops 1-3 at the top of the structure, calcium ions are shown in green). Hence it should have little effect on the protein's function. (c) Actual structure solved by nuclear magnetic resonance. The inserted sequence, again shown in red, becomes a β-strand, displacing the sequence that makes this strand in the short form. The displaced sequence becomes a helical region that seriously interferes with the binding site (Garcia 2004).

### IV.2.1 Physics-Based Free Modeling

From a physics point of view, interactions between atoms should be based on quantum mechanics, the coulomb potential and only a few fundamental parameters such as the electron charge and the Planck constant; their atom types should describe all atoms where only the number of electrons is relevant. However, few attempts have been made to start from quantum mechanics to predict structures of even small proteins, simply because the computational resources required for such calculations are far beyond the computer capabilities available now. Without quantum mechanical treatments, a practical starting point for Ab Initio protein modeling is to use a compromised force field with a large number of selected atom types; in each atom type, the chemical and physical properties of the atoms are alike with the parameters calculated from crystal packing or computational quantum mechanics. Well-known examples of such all-atom physics-based force fields include AMBER (Weiner 1984) and CHARMM (Brooks 1983). These potentials contain parameters associated with bond lengths, angles, torsion angles, Van der Waals, and electrostatics interactions. The

major difference between them lies in the selection of atom types and the interaction parameters.

Compared to template-based approaches, the purely physics-based Ab Initio methods – all-atom potential functions, like AMBER or CHARMM, combined with molecular dynamics (MD) conformational sampling – have been less successful in protein structure prediction. Significant efforts have been made on the purely physics-based protein folding. The first widely recognized milestone of successful Ab Initio protein folding is the 1997 work of Duan and Kollman, who folded the villin headpiece (a 36-mer). This work used MD simulations in explicit solvent for 2 months on parallel supercomputers with models up to 4.5 Å (Duan 1998). This small protein was recently folded by Pande and coworkers (Zagrovic 2002) to 1.7 Å, with a total simulation time of 300 ms or approximately 1,000 CPU.

In a more recent article, Shaw and collaborators correctly folded two proteins: BPTI, a serine-protease inhibitor, and the villin headpiece using the AMBER force field with explicit solvent and running in a special-purpose machine[16]. They observed not just the folding but the subsequent protein dynamics near the native state that agree to a good extent with the experimental data, showing that the current force fields can be accurate enough. The simulations of 1 ms are the longest to date and took about three months for each protein (Shaw 2010). Despite this remarkable effort, physics-based folding is far from routine for general protein structure prediction of normal size proteins, mainly because of the prohibitive computing demand and the general problem of finding a global minimum in the potential energy landscape.

Another niche for physics-based simulation is protein-structure refinement. This approach starts from low-resolution structures with the goal to draw the initial models closer to the native structure. Because the starting models are usually not far away from the native state, the conformational change is relatively small and the simulation time is much less than in Ab Initio folding. One of the earliest MD-based protein structure refinements was for the GCN4 leucine zipper (a 33-residue dimer) (Vieth 1994). In that work, a low-resolution coiled-coil dimmer structure (2 ~ 3 Å) was first assembled using Monte Carlo simulation. Recently, another team used CHARMM22 to refine five CASP6 CM targets with lengths in the 70–144 residue range. In four cases, considerable refinements with up to 1 Å RMSD reduction were achieved (Chen 2007). One of the major differences of this work is that an implicit solvent force field based on the generalized Born approximation was exploited, which significantly speeds up the MD simulations, while the spatial restraints extracted from the initial models are used to guide the refinement procedure.

Yet another use of the physics-based potential is in the discrimination of the native/near-native structures from structure decoys. For example, (Karplus 1998) exploited CHARMM19 and an implicit solvation potential to discriminate the native structure from the

[16] This special-purpose machine consists of a substantial number of regular cpu. They are interconnected by a specialized high-speed three-dimensional torus network that allows to massive parallelization. This approach to simulate folding competes with the Blue Gene/L machine which is a general purpose parallel supercomputer and the Folding@home distributed computing project which uses the cpu of regular computers connected to the internet.

decoys generated by threading the native sequences on other protein structures. They found the energy of the native states is lower than that of the decoys in most cases. Various authors obtained similar results, i.e. the native structure can be distinguished from non-native decoys by the physics-based potentials. Recently, however, (Wroblewska 2007) showed that the AMBER plus generalized Born approximation potential can only discriminate the native structure from roughly minimized TASSER decoys. But after a 2-ns MD simulation to further minimize the decoys, none of the native structures had lower energy than the refined decoys. This result partially explains the reported discrepancy between the decoy-discrimination ability of the physics-based potentials and less-successful folding/refinement results.

### IV.2.2 Knowledge-Based Free Modeling

Bowie and Eisenberg pioneered the following approach for free modeling: assembling of new tertiary structures using small fragments (mainly 9-mers) cut from other PDB proteins (FIGURE 6B) (Bowie 1991). Based on this idea, Baker and coworkers later developed ROSETTA (Simons 1997), which worked very well for free modeling of small/medium proteins in the CASP experiments, and popularized the fragment assembly approach in the field. In new developments with ROSETTA (Das 2007), the method first assembles structures from the fragments in a reduced knowledge-based model with conformations specified by the heavy backbone atoms and Cβ. In the second stage, Monte Carlo simulations with an all-atom physics-based potential are performed to refine the details of the low-resolution models. A notable achievement was demonstrated in CASP6 by generating a model for a small hard target T0281 (70 residues) that is 1.6 Å away from the crystal structure. In CASP7, a very extensive sampling was carried out using the distributed computing network of Rosetta@home that used about 500,000 CPU hours for each target domain. With this computer power the protocol of ROSETTA built a model for T0283 (a template modeling target) with RMSD = 1.8 Å over 92 residues out of the 112 (FIGURE 8). And in CASP9 (CASP9 2010), a ROSETTA model was significantly better than the best template available for a target of the α/β class (FIGURE 7). However, despite significant success, the computer cost of the procedure (~150 CPU days for a small protein <100 residues) was still too expensive for routine use[17].

[17] In a subsequent approach to circumvent the large conformational sampling needed, the ROSETTA team developed FoldIt (Cooper 2010), an online multiplayer game where players fold proteins starting from their extended-chain state and accumulate points by moving the structure to a lower energy state. This approach is based in the hypothesis that human search strategies using 3D and visual cues are better than Monte Carlo methods for sampling highly-rugged energy landscapes. In a ten-structure blind test, the players performed better than the ROSETTA protocol at five of them (two models had a RMSD < 2 Å to the native structure) and had the same score at other three.

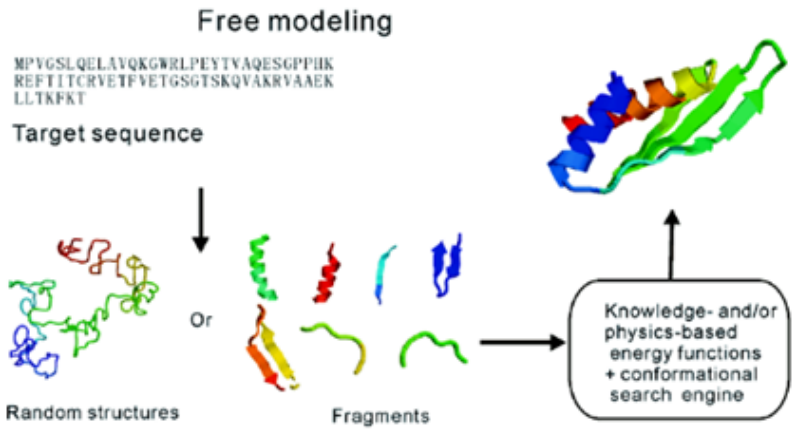


FIGURE 6B. Schematic overview of a Free Modeling pipeline

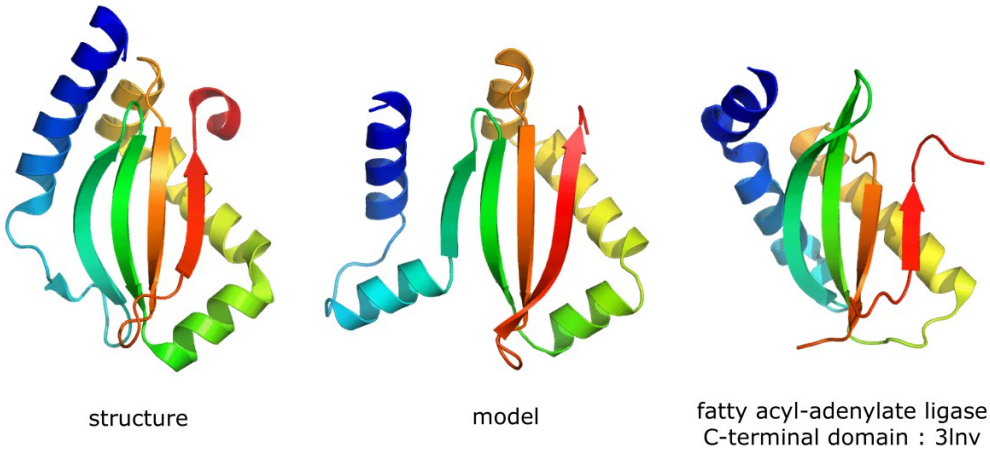


FIGURE 7. In CASP9 a ROSSETA model for Target 581 was the largest improvement over the closest template (fatty acyl-adenylate ligase C-terminal domain; PDB 3lnv). It wasn´t the best scoring model of the whole FM category but for the correct assignments of the secondary structure elements achieved a GDT score 44% higher than the template, even when the secondary structure prediction was markedly incorrect (CASP9 2010).

Another knowledge-based free modeling approach is called TASSER (Threading Assembly Refinement) and was first developed by Skolnick and Zhang (Skolnick 2004). In its I-TASSER version (iterative-TASSER, Y. Zhang 2008) the protocol has three stages: i) Threading. A meta-threading server combining seven top threading programs (including HHSearch and Prospect) finds templates that are sorted by its alignment quality. ii) Structural Assembly. Continuous fragments with various sizes are excised from threading alignments and used to reassemble protein structures in a reduced model of just Cα and side-chain center of mass. The regions not aligned by threading are modeled on a lattice system and those that were aligned are kept off this lattice. The reassembly process of the fragments is conducted by parallel Monte Carlo simulation and cluster centroids are obtained by

averaging the 3D coordinates of all the clustered structural decoys. The energy terms of the potential include information about predicted secondary structure propensities, backbone hydrogen bonds, a variety of short- and long-range correlations and hydrophobic energy based on the structural statistics from the PDB library. Weights of knowledge-based energy terms are optimized using a large-scale structure decoy set, which coordinates the complicated correlations between various interaction terms. iii) Model Selection and Refinement. The fragment assembly simulation is performed again starting from the selected cluster centroids. Although the inherent I-TASSER potential remains unchanged in the second run, external constraints are pooled from the original threading alignments and the PDB structures that are structurally closest to the cluster centroids according to their structural alignment tool (Zhang 2005). The purpose of the second iteration is to remove steric clashes and to refine the global topology of the cluster centroids. The decoys generated during the second round of simulations are clustered again, and the lowest energy structures are selected as input for the generation of the final structural models by building all-atom models from Cα traces through the optimization of hydrogen bonding networks. Although the whole procedure uses structural fragments and spatial restraints from threading templates, it often constructs models of correct topology even when the topologies of individual templates are incorrect. In CASP7 (Zhang. 2007), among 19 FM and FM/TBM targets, I-TASSER built the correct topology (~3–5 Å) for 7 cases with sequences up to 155 residues long (FIGURE 8).

In the CASP9 experiment (CASP9 2010), I-TASSER was tied with ROSSETA and the then newly developed QUARK as the best predictors in the Free Modeling category. QUARK (an Ab Initio server also from the Zhang group) was the best performing method when no templates were available.

Although a purely physics-based Ab Initio simulation has the advantage in revealing the pathway of protein folding, most free modeling methods combine both knowledge-based and physics-based approaches. There were consistent successes in building correct topology (3–6 A˚) for small proteins, but the more exciting high-resolution free modeling (< ˚2 A) was not frequent and computationally expensive. Prediction of structures with high contact order[18] and/or novel folds continued to be poor in general. There was evidence that the atomic potentials do give the lowest energy near the native state and that the bottleneck of high-resolution folding seems to be the insufficient conformational sampling (Cooper 2010). The artificial golf-hole-like energy landscapes without middle-range funnels could be a contributor to the shortcomings of conformational search, especially for proteins of larger sizes.

[18] A high average contact order is the result of a complex topology. It has been observed that the higher the contact order, the slowest a protein folds, and that the folding rate doesn´t depend too much on the sequence (Alm 1999), this impacts the capabilities of MD-assisted folding simulations for proteins with a complex topology, since they would require much more computational power to simulate longer time scales.

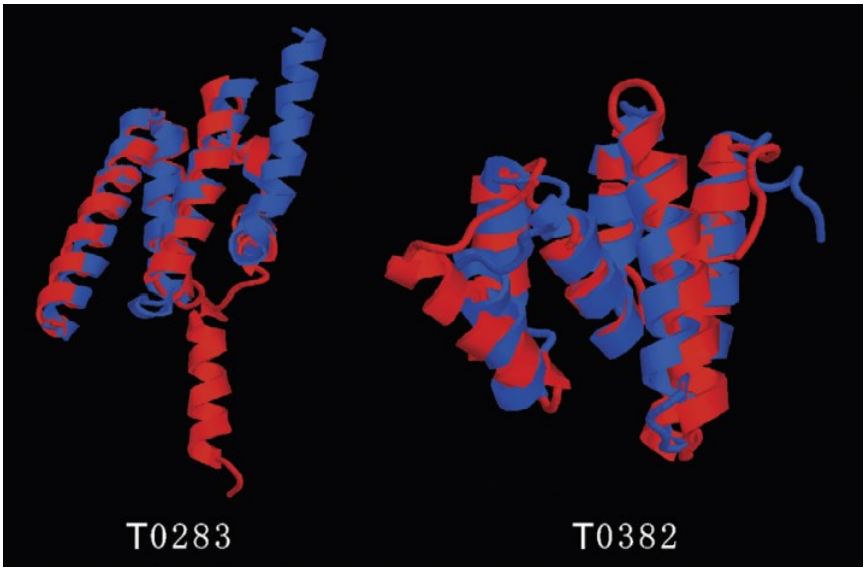


FIGURE 8. Examples of free modeling in CASP7 generated by the two top methodologies (Zhang. 2007). T0283 (left) is a TBM target of 112 residues; but the model is generated by all-atom ROSETTA (a hybrid knowledge-based and physics-based approach), which gives a TM-score[19] 0.74 and an rmsd 1.8 A˚ over the first 92 residues. T0382 (right) is a FM/TBM target of 123 residues and all initial templates had incorrect topologies (>9 Å); the model is generated by I-TASSER (a purely knowledge-based approach) with a TM-score 0.66 and an rmsd 3.6 A˚. Blue and red represent the model and the crystal structures, respectively. It is common that these two methods excel when modeling all-α proteins.

### IV.2.3 Coarse-grained and Deep Learning Free-Modeling

By the time ROSETTA achieved mainstream use in the molecular biology community, the ideas that lead to rapid progress in free-modeling had exhausted their potential and for a time the focus turned to refining them. A lack of collaboration between teams and the scarcity of significant progress in the understanding of the folding process[20] may have accentuated this. Beta-proteins, large size and high contact order continued to pose the biggest challenges for structure prediction, overestimation of helix sizes and underestimation of strand sizes were common.

The set of shapes that a protein might take can be likened to a landscape: different locations in the landscape correspond to different shapes, with nearby locations having similar shapes. The height of a location corresponds to how energetically favorable the associated shape is, with the lowest point being the most favored. Natural proteins evolved

[19] The TM-score (Zhang.. 2004) was developed as an alternative to some of the drawbacks of the GDT score like its dependence to protein size and its subjective distance cutoffs. This score uses weights developed by (Levitt 1998), where shorts distances are weighted stronger than shorter distances, thus making the score more sensible to global topology than to local similarity. By associating a statistical significance test to this score, the authors found that a TM-score above 0.5 is an indicative of the same SCOP or CATH fold (Xu 2010).

[20] The successful idea of fragment assembly (Bowie 1991) emerged from folding theory, which hasn´t had any breakthroughs since the folding funnel hypothesis.

to have funnel-shaped landscapes that enable newly synthesized proteins, powered by the thermal fluctuations of the cell, to cross the landscape and find their way to a favored conformation in physiologically relevant timescales (milliseconds to minutes). Algorithms can search the landscape to find favored conformations by following the landscape's inclination, but the ruggedness of the terrain causes them to get stuck in local valleys far from the lowest basin. At some point, coarse-grained representations of protein structures started to be relevant due to their ability to significantly extending the conformational sampling (Kolinski 2005). In particular, the use of graph representations/networks of proteins (Amitai 2004) (Thibert 2005), where vertices represent residues (or atoms) and edges represent 3D contacts between them is an interesting approach because combines computational speed in their handling with the vast theoretical edifice of graph theory results and their applications (Jacobs 2001). An important fact is that if a graph represents all the native contacts, it is possible to reconstruct the 3D structure from it (Havel 1979). One of the purposes of this thesis is to try to understand the potential benefits of this approach towards a free modeling protein structure prediction method. Graph representations of 3D contacts can also provide helpful information for model selection (Latek 2008).

The course of the structure-prediction field started to change with the implementation of an old idea: that the evolutionary record contains clues about how proteins fold (Altschuh 1987). If two amino-acid residues in a protein are close together in 3D space, then a mutation that replaces one of them with a different residue will probably induce a mutation that alters the other residue in a compensatory direction to maintain energetically favorable interactions, thus residues in spatial proximity may co-evolve across a protein family and the set of co-evolving residues can encode valuable spatial information, especially if they are far apart in the sequence. With the advent of the big sequence databases that permitted to construct useful multiple sequence alignments and by transforming this co-evolutionary information into 3D contacts in a graph/matrix representation of the contact map, the set of conformations that merit consideration by algorithmic searches can be greatly restricted (Göbel 1994). At the beginning of last decade, several groups started to identify a number of biases that had stymied prior attempts to augment the coevolutionary-3D contacts signal, and developed powerful statistical machinery to correct them. There was some consistent progress for several years, with direct coupled analysis (DCA) methods based on the Ising model being the more successful (Weigt 2009). And then, in CASP13 (CASP 2018), several groups demonstrated that there was actually no need for robust statistics, it was sufficient to train deep residual neural networks.

The initial injection of deep learning (a type of machine learning) into co-evolutionary analyses improved matters by incorporating richer inputs. By 2018, the modelers were often scoring in the mid-70s of the performance scale (scores above 90 were considered on par with experimentally solved structures). AlphaFold, developed by Google's DeepMind, had a median score close to 80 in CASP13. Instead of binary contact data, AlphaFold predicts the probabilities of residues being separated by different distances. Because probabilities and energies are interconvertible, AlphaFold predicts an energy landscape -one that overlaps in its lowest basin with the true landscape, but is much

smoother. In fact, AlphaFold's landscape smoothness nearly eliminates the need for searching. This makes it possible to use a simple procedure to find the most favorable conformation, rather than the complex search algorithms employed by other methods. The resulting algorithm outperformed all entrants at CASP13, generating the best structure for 25 out of 43 proteins, compared with 3 out of 43 for the next-best method. AlphaFold's predictions had a median accuracy of 6.6 ångströms on this set of proteins – that is, for the middle-ranked protein in this set, the atoms in the proposed structures were on average 6.6 Å away from their actual positions.

At CASP 14 in 2020, its successor, AlphaFold 2 (Jumper 2021), had a median score of 92.4—on par with experimental technique and as such, it can be considered one of the major science's achievements of this century. The first section of the gigantic AlphaFold 2 neural network, the Evoformer, has the task of maximizing the extraction of coevolutionary/3D contacts information out of the multiple sequence alignment and a structure model made of templates. The central idea behind the Evoformer is that the information flows back and forth the network. Before AlphaFold 2, most deep learning models would take a multiple sequence alignment and output some inference about geometric proximity. Geometric information was therefore a product of the network. In the Evoformer, instead, the structure contacts representation is both a product and an intermediate layer. At every cycle, the model leverages the current structural hypothesis to improve the assessment of the multiple sequence alignment, which in turns leads to a new structural hypothesis. Both representations, sequence and structure, exchange information until the network reaches a solid inference. Evoformer is a transformer neural network, the transformer architecture was introduced in 2017 by a team at Google Brain, the key ingredient is a mechanism called *attention*. The objective of attention is to identify which parts of the input are more important for the objective of the neural network. In other words, to identify which parts of the input it should pay attention to. The main reason why transformers have not been widely implemented is that the construction of the attention matrix leads to a quadratic memory cost. The Evoformer architecture uses two transformers, with one communication channel between the two. Each head is specialized for the particular type of data it is looking at, either a multiple sequence alignment, or a representation of pairwise interactions between amino acids. They also incorporate the information of the contiguous representation, allowing for regular exchange of information and iterative refinement. The unparalleled performance of the AlphaFold 2 network seems down to DeepMind's engineering. It seems that the ideas in the model don't provide new insights on protein folding or about protein structure, it is their access to compute resources, and their engineering capabilities that turned them into the successful neural network it became. Many of the performance-increasing details are probably due to intensive experimentation.

In July 2022 researchers announced they have used AlphaFold 2 to predict the structures of more than 200 million proteins from some 1 million species, covering almost every protein sequence of Uniprot. The data is freely available on a database set up by DeepMind and the EMBL–EBI. According to EMBL–EBI, around 35% of the more than 214 million predictions are deemed to be highly accurate, which means they are as good as

experimentally determined structures. Another 45% are considered to be accurate enough for many applications. Many AlphaFold 2 structures are good enough to replace experimental structures for some applications. In other cases, researchers use AlphaFold predictions to validate and make sense of experimental data. Poor predictions are often caused by intrinsic disorder in the protein itself that means it has no defined shape — at least, not without other molecules present.

# V. Protein Functional Conformers

Proteins present conformational fluctuations (dynamics) (Hvidt 1954), and they are assumed to be important in several biological processes such as molecular recognition (Frederick 2007), catalysis (Henzler-Wildman 2007), and allosteric regulation (Popovych 2006). The consequence of dynamics as a prerequisite to function suggests that in addition to structural requirements, such as shape and chemical affinity, function imposes requirements for flexibility as well. The biological functions of proteins can be viewed as arising from interplay between protein structure and dynamics, and new experimental and computational strategies are needed to understand better these two contributions.

In the case of enzymes, catalytic requirements seem to arise case-by-case, but precise spatial arrangement of catalytic residues is clearly of central importance. As for the how sequences and structures are designed to facilitate conformational change, a clearer picture is emerging: to enable conformational changes, proteins need to make relatively large-amplitude fluctuations toward specific directions. In native-basin dynamics, it has been established that the quasi-harmonic fast fluctuations are encoded largely in the three-dimensional architecture, as illustrated by the accurate reproducibility of the RMSD fluctuations around the native state by the Gaussian network model (Haliloglu 1997), a type of structure-based elastic network model (ENM) in which vertices represents atoms (or amino acids) that are joined by strings according to interacting or distance criteria. In the case of allosteric changes, the direction of conformational change from the open state to the closed state can usually be well represented by a few low-frequency modes of the anisotropic network model (Atilgan 2001), another type of ENM.

Protein motions modeled by simple elastic network models are harmonic or quasi-harmonic, but the large-amplitude fluctuations required for functional relevant conformational changes are inharmonic (Miyashita 2003) and subject to a nonlinear

potential derived from the chemical identities of the residues. Evidence suggests that protein architecture alone is not sufficient, and that sequence specificities also play crucial roles in encoding motions. Some mutations that do not modify catalysis per se or the native structure can alter functional behavior substantially. In a recent study with Adenylate kinase (Schrank 2009), the authors found that certain mutations on residues localized away from the active site, increased the probability of a locally unfolded state that correlated with a change in the binding affinity. In addition, the interactions within a protein can be locally frustrated due to the restraints imposed by functions, a survey of allosteric proteins shows that hinge regions are located near regions of high frustration according to a residue-based simplified energy function (Ferreiro 2011), suggesting that functional requirements on the sequence may conflict with the optimal conditions for folding or packing.

Another observation pointing to the role played by dynamics in enzyme function is that time constants for enzyme catalysis and product release, protein folding, and allosteric transition are all between microseconds and milliseconds. Since enzymatic reaction is essentially a change in the electronic state of the active site, one expects that the time scale must be quite short, between femto-seconds and pico-seconds. In addition, the diffusion-controlled reaction detected by fluorescence quenching also occurs quite rapidly, around a time scale of pico-seconds. However, the enzymatic reactions are slower and its time scales are comparable to that of protein folding (Henzler-Wildman 2007). This suggest that, in order to prepare the special nuclear coordinate for the transition state of the bound active site, a protein must rearrange its nuclear coordinates substantially, and this process may generally take a period of time almost as long as that of protein folding.

However, the discussions over the exact nature of this phenomenon will likely occupy the field for more years to come. Once a structural rearrangement linked to transition state formation is identified, the question is whether the detected movement contributes to catalysis by lowering the transition state barrier. In a dramatic substrate-to-product rearrangement such as proline cis-trans isomerization carried out by cyclophilin A (P. Agarwal 2006), the rearranging substrate may simply push certain side chains out of the way to reach the transition state. If these side chains are flexible and it costs negligible energy to brush them aside, then the accompanying movement will have little effect on catalysis. At the other extreme, rearrangements of the enzyme structure could be required to properly align the catalytic residues around the developing transition state, such that the protein dynamics that occur during transition state formation are essential to transition state stabilization and thus catalysis.

Recent results suggest that conformational fluctuations resulting from the concerted motions of many atoms can push the unbound states of enzymes into conformations closely resembling the bound states, thereby priming them to form complexes with specific ligands (Boehr 2006). Thus, although the unbound state of a protein is inherently flexible, fluctuations are not random. Rather, they take place preferentially in a way that prepares the protein to bind to its cofactors and substrates. The free-energy landscapes of the free and the bound states may differ just enough to cause changes in the relative populations of their

principal states. After binding, the free-energy landscape is plastically deformed just enough to make a slightly different state of the protein become the most populated. An example of this is the five successive transitions that carry out the enzyme dihydrofolate reductase (DHF) (FIGURE 9), where it has been shown that a binding event can cause a protein molecule to occupy a new free-energy minimum, stabilized by a ligand and geared to fluctuate toward another state that binds the next ligand in the catalytic cycle (Boehr 2006). This view may point out that the energy landscape is more than a mere funnel, and resembling rather a *battered sombrero*. Such picture reflects how evolution could have coordinated the thermal motions of hundreds of atoms to perform biological functions.

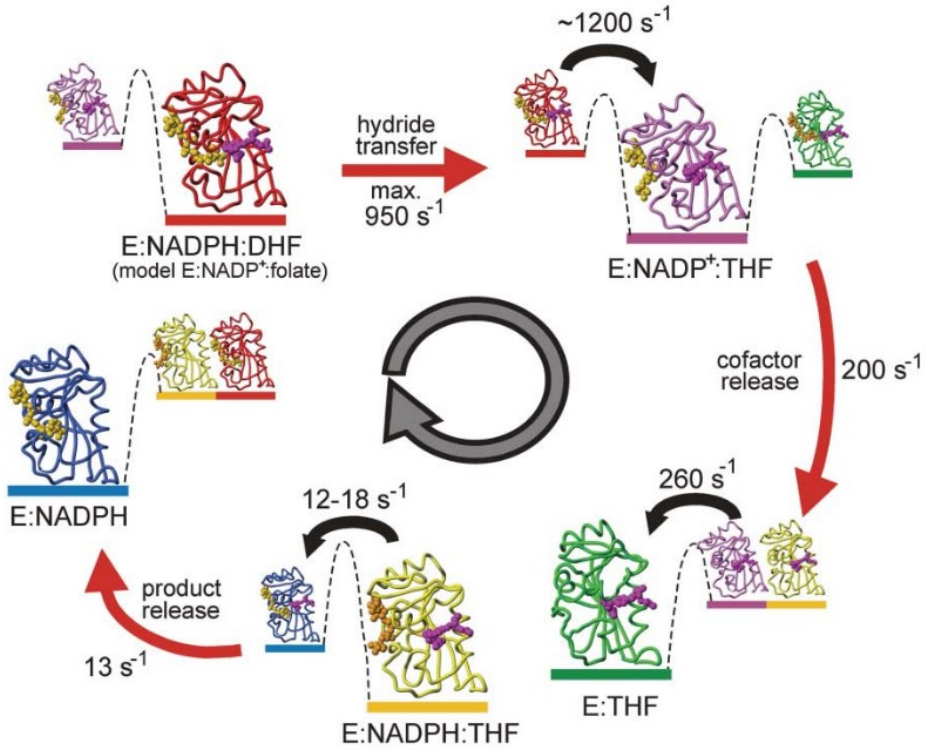


FIGURE 9. The dynamic energy landscape of DHF catalysis (Boehr 2006). The ground state (larger) and higher energy (smaller) structures for each intermediate in the cycle are shown. For each intermediate in the catalytic cycle, the higher energy conformations detected resemble the ground-state conformations of adjacent intermediates.

In a series of recent millisecond MD simulations carried out for two small proteins (Shaw 2010), the authors could observe the particularities of the short and large amplitude fluctuations. An identified region of reduced dynamical activity in the interval between those time regimes could be a common feature of proteins. One section of one of the proteins changed its conformation during hops between basins and the side chains of the protein moved slowly compared to their movement in the short amplitude fluctuations, where backbone movement is minimal. The transition time between conformational transitions was at least several hundred nanoseconds, a time that might tend to increase with protein size. Also, binding and escape events of a water molecule showed to be considerably faster than

the lifetime of the bounded state. A 1-millisecond simulation of the folded protein BPTI revealed a small number of structurally distinct conformational states whose reversible inter-conversion is slower than local fluctuations within those states by a factor of 1000 or more. Interestingly, it has been proposed that the conformational diversity of proteins allows them to be functionally evolvable (Tokuriki 2009). Minor conformers may mediate alternate functions, and mutations could shift the conformational equilibrium to favor these conformers and thus increase the level of the alternate function.

There are several approaches for predicting functional residues, most of them based on conserved positions or solvent-accessibility surface (SAS) (Ashkenazy 2010). These predictors normally point to the active or binding site residues. But as allosteric regulation shows, important residues for function may not necessarily reside there. A report on the enzyme dihydrofolate reductase showed that a network of coupled systematic motions in distant residues of the protein is associated to the reaction trajectory from the reactant to the transition state (P. B.-S. Agarwal 2002). Thus, if as discussed above, certain functionally relevant fluctuations appear to depend not only in the position but also in the chemical identity of residues positioned away from the active site, it is desirable to predict this class of residues too.

# VI. Graph Representation of Protein Structures

Depending on the problem studied, protein structures can be chosen to be represented in a number of different ways: as atomic coordinates, as secondary structure elements, etc. The employ of graph/network approaches to model protein structures, where amino acids residues are represented as the vertices of a graph (in the graph theory sense), and the contact or proximity in space between them as the edges[21] (FIGURE 10), has been chosen for problems like fold recognition (Mirny 1996) and the study of transition states of folding (Vendruscolo 2001). In this work we will call this approach Protein Graphs.

[21] From this definition we can see that a graph or network derived from a 3D protein structure is equivalent to the notion of a contact map of amino acid residues, which is the term preferred in the context of protein structure solving and prediction. In this work we treat graph, network or contact map as indistinguishable terms. Therefore, the terms vertex and edge of a graph are equivalent, respectively, to the terms node and connection of a network.

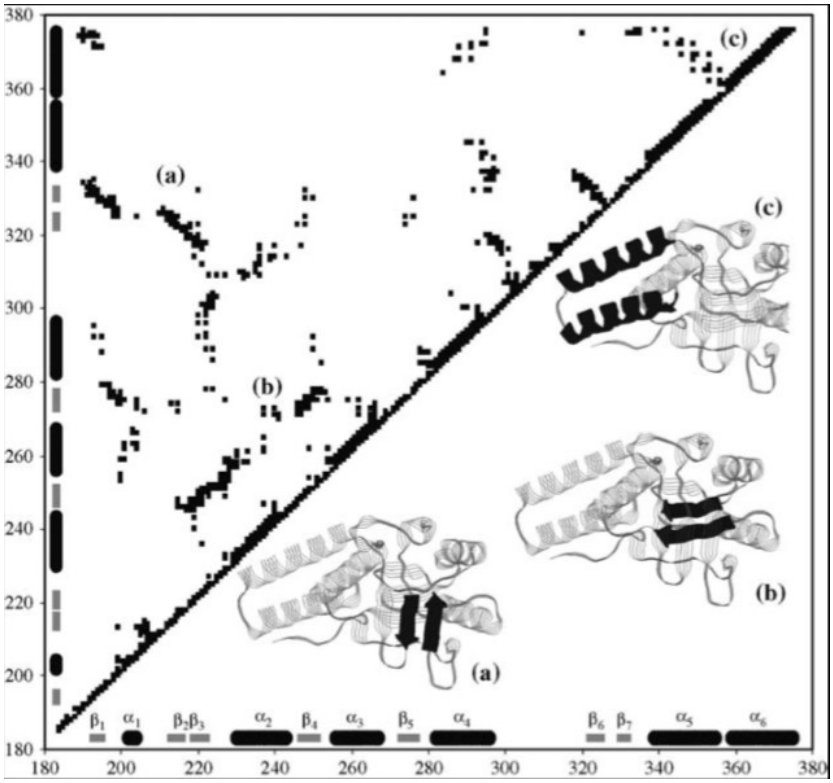


FIGURE 10. Matrix representation for the graph corresponding to the structure of the HSP-60 protein (PDB: 1KID). Each x or y-value corresponds to one vertex (from residue 180 to 380). On the upper left triangle each dot indicates an edge that is present between the corresponding pair (x,y) of vertices in the graph. In this example an edge is present whenever the CB atoms of two residues are at most 8.0 Å apart. The secondary structure elements are indicated along the x and y axis and on the lower right triangle some structural features are highlighted in dark: (a) β2_ β3 β6_ β7 anti-parallel sheet, (b) β4 β5 parallel sheet and (c) α5 α6 contacting helical regions.

Protein Graphs have been fruitful in identifying functional residues beyond those identified by evolutionary or SAS methods. Measures like closeness centrality (CC) and transitivity, that captures the load of flux that could pass through a given vertex have proved especially suited for this task (Amitai 2004) (Thibert 2005) (Cusack 2007).

Thermal motions act as a molecular lubricant during conformational changes, directing the protein through the energy landscape. Yet, in a given macromolecule only a subset of possible motions is important for biological function. It is a challenge to identify these functionally relevant conformations. If, as previously noted, Protein Graphs permit that certain functionally important residues can be identified by structure alone, it is reasonable to propose that those conformers that are directly involved in function are those that harbor the functional residues as central residues in the corresponding graphs. One of the objectives of this thesis is to test this hypothesis.

# OBJECTIVES AND HYPOTHESIS

There were two objectives in the development of the present work:

I. To find patterns in the graph representations of protein structures that could be used in the conformational space search step of a free modeling protein structure prediction protocol. Our **hypothesis** in this case is that the graph representations of native protein structures have a particular class of topological properties that distinguishes them and that we look to characterize. We think that these features can be exploited to accelerate the search of near-native structures in the conformational space landscape.
II. To predict functional conformers by mapping critical residues with graph-derived central residues. In this case our **hypothesis** is that graph-derived central residues are a good predictor of the critical residues for the function of a protein and that this becomes evident only when the structure of the protein adopts the particular conformers that carry on the function. Thus, we hypothesize that the functional conformers are those whose critical residues coincide the most with their central residues.

# METHODS

**Graph Representations of Protein Structures (PGs).** Coarse-grained models are seen as a reasonable approximation to full-atom models of protein structures in order to speed computational methods of structure prediction. We used the graph representation (a set of vertices connected by edges) of protein structures to achieve fast calculations but more importantly due to previous work (Thibert 2005) that found that certain measures derived from this type of representation are useful for identifying critical residues (residues essential for protein function). Graph representations (FIGURE 10B) where constructed as follows: **i)** We used the 3D atom coordinates of PDB format files. **ii)** Each amino acid residue is represented as a vertex. **iii)** Whenever any two atomic centers belonging to different amino acid residues lie inside a sphere of 5 Å of diameter, we connect those two amino acids by an undirected edge[22]. This particular way to build the PGs was chosen due a previous study

[22] We calculated the distance $d(a_1,a_2)$ between two residues $a_1$, $a_2$ as:

that found its effectiveness to map graph-central vertices with critical residues for function (Thibert 2005).

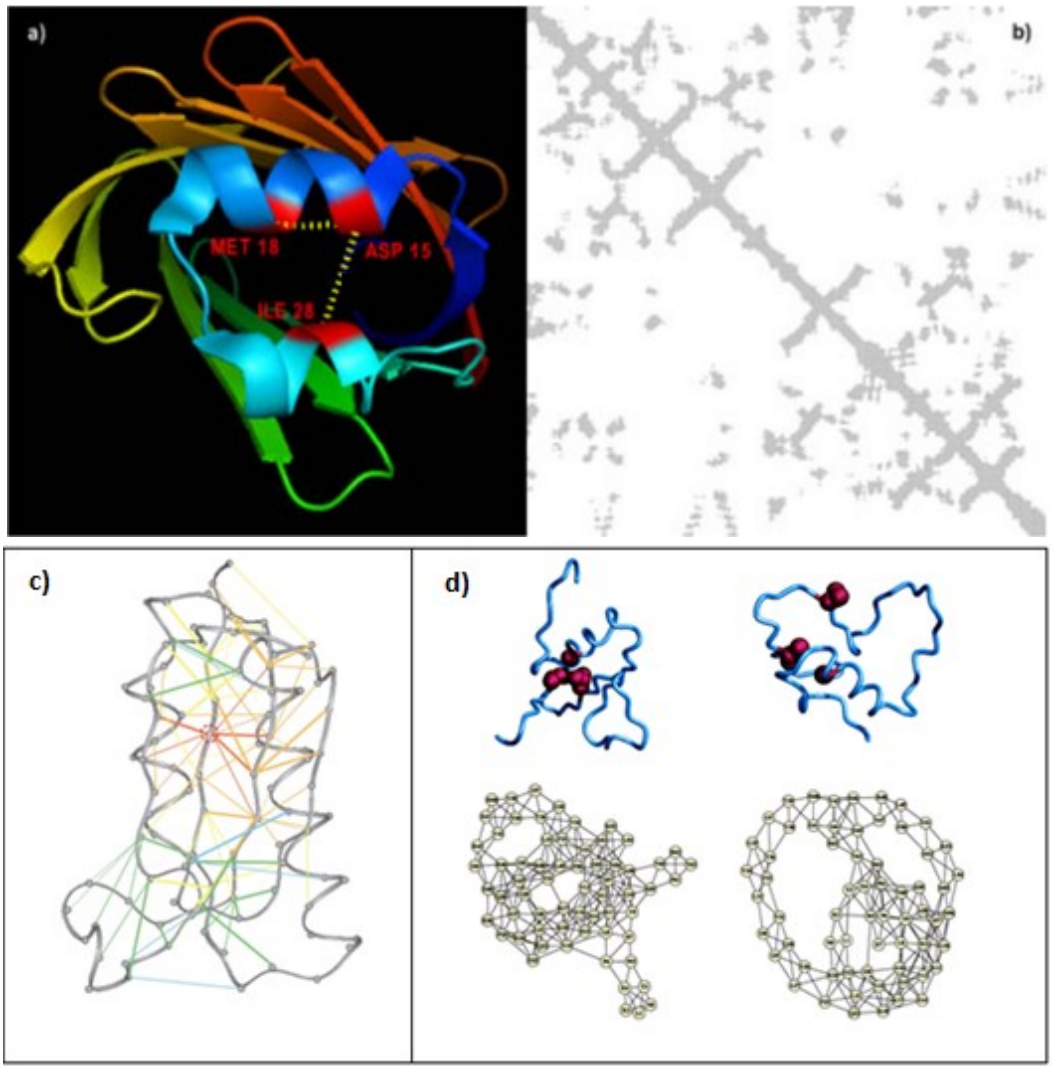


FIGURE 10B. Schematic view of graph representations of protein structures (PGs). (a) Since certain atoms of the residues in red are within a distance of 5 Å of each other, the vertices corresponding to those residues become connected in the PG. (b) Matrix representation of a PG. Called adjacency matrices, they have as much columns (and rows) as the number of residues in the protein. Every pair of entries $a_{i,j}$ / $a_{j,i}$ are colored in grey if the *i* residue is connected to the *j* residue. (c) Cartoon of a backbone showing the edges of its PG. (d) Depiction of the corresponding PGs for two proteins.

**Non-redundant Structural Representatives (NRR).** To find general patterns of graph representations of proteins, we selected 1,899 significantly different protein structures from

,

where $a_{k,i}$ denotes the positions of all the atomic centers of the *i* atoms of the residue $a_k$. Then we assigned an edge to all pairs whose $d(a_1,a_2) < 5$ Å.

the FSSP database[23]. Each of these proteins comes from a different first-level structural class in the FSSP classification and the 1,899 structures covered the entire first level of the FSSP database at the time of the study. By being from different first-level classes we make sure that for every pair of these structures their DALI structural similarity has a z-score < 2 (a DALI similarity between two structures with z-score ≥ 2 means that the DALI similarity is two standard deviations away from the average DALI similarity obtained from an all-against-all PDB-wide structural comparison, or in other words, a z-score ≥ 2 means that the similarity of these two structures is statistically significant).

**Measures derived from PGs.** We derived several measures and looked at their distribution across the 1,899 structures to find characteristic features of this kind of graphs. **i)** Degree (k): In an undirected graph, the degree of a vertex corresponds to its edges. **ii)** Clustering coefficient (CC). The clustering coefficient estimates how close the neighbors of a vertex (those vertices directly connected to the vertex in question) are to form a clique (a subgraph where all vertices are connected), thus CC = [Observed No. edges between the N-neighbors of a vertex]/[Expected No. edges in a clique of size N]. The CC gives an idea of how strongly connected are the neighbors of a given vertex. **iii)** Contact order (CO). This is a common measure in the structure prediction methods. It is defined as the number of residues in the primary sequence that lie between two residues that are in contact in the 3D structure. In the graph representation, the CO is just the number of edges that has the primary sequence-derived path (a path is any set of consecutive edges) between two vertices connected by a non-primary sequence-derived edge. **iv)** Matrices of Normalized amino acid type contact preferences (MNCP). We wanted to know if there are preferential connections between different amino acid types. A simple way to look at this is to count the number of edges between any two different amino acid types in a structure or set of structures and then divide this number by the total number of edges, we end up with symmetrical matrices of 20x20 NCPs.

**Identifying Central Residues from PGs.** Central residues were defined as those residues that correspond to the vertices with the largest and less frequent transitivity values. The transitivity value of a vertex was obtained by counting the number of times this vertex was in the shortest paths connecting every pair of vertices in the graph. The frequency of a transitivity value is the number of vertices presenting that transitivity value in a graph. Thus, each vertex will have a transitivity value and a value-frequency in the graph; only those having transitivity values immediately close to the largest transitivity value in the graph and with the same transitivity value frequency as those with the largest transitivity values are considered central. Using this strategy, we observe that about 20% or less of the vertices were central in the PGs. For these calculations, we used our software available

[23] The FSSP database is a continuously updated structural classification of three-dimensional protein folds. It is derived using a structure comparison program (DALI) for the all-against-all comparison of three-dimensional coordinate sets in the PDB. Hierarchical clustering based on structural similarities yields a fold tree that defines 1,899 fold classes. For each representative protein chain, there is a database entry containing structure-structure alignments with its structural neighbors in the PDB. DALI stands for Distance matrix ALIgnment and is a server that does a 3D comparison of protein structures that assigns a similarity score by finding an alignment that minimize the distance matrix between the Cα of two proteins.

at http://bis.ifc.unam.mx/jamming/ (Cusack 2007). Transitivity, T, is related to betweeness B (Brandes 2008), as follows: $B_i = T_i/SP_i$; where $B_i$ is the betweeness value calculated for the *i*-vertex, $T_i$ is the Transitivity value of the *i*-vertex, and $SP_i$ is the number of shortest paths connecting the *i*-vertex to the rest of the vertices in the graph.

**Overrepresented Subgraphs (Motifs).** For discovering the frequency of different types of subgraphs inside the PGs that are overrepresented with respect to a random expected frequency (so-called motifs) we used mfinder, a free tool from the lab of Uri Alon (Milo 2002) that count the frequency of any type of subgraph of less than 9-vertices. We were interested in which subgraphs were overrepresented in the PGs with respect to a null random model. For this we generated, for each PG to analyze, different sets of random graphs that conserved the exact sequence of degrees of each vertex in the PG. For evaluating the significance of the frequency of 3- and 4-vertices subgraphs we compared the relative frequencies of these subgraphs in 10,000 random graphs (RGs), for the 5-vertices subgraphs we used a random sample taken from 50,000 RGs and for the 6-vertices subgraphs we used a random sample taken from 100,000 RGs. The need for random sampling is due to the computational load that is expected from the exponential growth in the number of different subgraphs as a function of the number of vertices that form them (Harary 1973).

**Connectivity measures for PGs of Backbone-less Proteins.** The particular distance criterion used for building the PGs[24] implies that every two consecutive residues in the polypeptide chain are always connected by an edge. To evaluate the connectedness of the PGs irrespective of the connectivity directly provided by the polypeptide chain we removed all the edges that linked every vertex with its first, or both its first and second neighbors along the polypeptide chain, we called them 1-simplified and 1/2-simplified PGs, respectively. We then proceeded to calculate some connectivity measures on these "backbone-less" PGs. **i)** The Eccentricity ($\varepsilon_v$) of a vertex v is the greatest geodesic distance (number of edges in a shortest path connecting two vertices) between v and any other vertex. It measures how far a vertex is from its most distant vertex in the graph. **ii)** The Radius of a graph is the minimum of all the $\varepsilon_v$ of the graph. **iii)** The Diameter of a graph is the maximum $\varepsilon_v$ of the graph. That is, it is the greatest distance between any pair of vertices. To find the diameter of a graph, first we find the shortest path between each pair of vertices. The greatest length of any of these paths is the diameter of the graph. **iv)** Connectedness. We can verify if the graph was left disconnected (tested for the presence of vertices for which there is no path connecting them) by the removal of the backbone-derived edges. For this we simply check if the eccentricity of any vertex is zero.

**Structural Data used for Analyzing Central Residues and Functional Conformers.** To study the relationship between functional residues and central residues in multiple protein structures, two proteins were used: HIV protease and the T4 lysozyme. For the HIV protease, 73 experimentally determined crystal structures were used: 1a30, 1a8g, 1a9m, 1aaq, 1ajv, 1ajx, 1axa, 1bdr, 1bv7, 1bv9, 1bwa, 1bwb, 1cpi, 1dif, 1dmp, 1gnm, 1gnn, 1gno, 1hbv, 1hih,

[24] If any two atomic centers belonging to different amino acid residues lie inside a sphere of 5 Å of diameter then we link with an edge the vertices corresponding to those amino acids residues.

1hiv, 1hos, 1hps, 1hpv, 1hpx, 1hsg, 1hte, 1htf, 1htg, 1hvc, 1hvi, 1hvj, 1hvk, 1hvl, 1hvr, 1hvs, 1hwr, 1hxb, 1hxw, 1mer, 1mes, 1met, 1meu, 1mtr, 1odw, 1odx, 1ody, 1ohr, 1pro, 1qbr, 1qbs, 1qbt, 1qbu, 1sbg, 1tcx, 1vij, 1vik, 1ytg, 1yth, 2aid, 2bpv, 2bpw, 2bpx, 2bpy, 2bpz, 2upj, 3aid, 4hvp, 4phv, 5hvp, 7hvp, 8hvp, 9hvp. For the T4 lysozyme 23 experimentally determined crystal structures were used: 1ctw, 1cu0, 1cu2, 1cu3, 1cu5, 1cu6, 1cup, 1cuq, 1cv0, 1cv1, 1cv3, 1cv4, 1cv5, 1cv6, 1cvk, 1cx7, 1d2w, 1d2y, 1d3f, 1d3j, 1d3m, 1d3n, 1qsq.

To identify functional conformers, three sets of protein structures were used: HIV protease, the yeast TATA-Binding Protein (TBP) and the MolMov set of proteins. For the HIV protease, the same protein structures described above were used. The PDB code of those structures in complex with a substrate analogue are: 1aaq, 1cpi, 1dmp, 1hbv, 1hih, 1hiv, 1hos, 1hps, 1hpv, 1hte, 1htf, 1htg, 1hvi, 1hvj, 1hvk, 1hvl, 1hvr, 1hvs, 1ohr, 1sbg, 2bpv, 2bpw, 2bpx, 2bpy, 2bpz, 4hvp, 4phv, 5hvp, 7hvp, 8hvp, 9hvp. For TBP, the crystal structures used had the PDB codes: 1tbp for TBP without DNA, and 1ytb for the TBP complex with a TATA box (TATATAAA).

In the case of the MolMov set, we used the proteins reported at the database of macromolecular movements (Flores 2006). This database includes structures of proteins motions and we have analyzed only those including an interaction with a ligand. Thus, this set includes proteins solved in the absence of a ligand (MolMov subset U) and the same proteins solved in the presence of a ligand (MolMov subset I). The PDB codes in the MolMov subset U includes: 1bjz, 1beb, 1dqz, 1tre, 1pin, 1dv7, 4crx, 1ex6, 1fto, 1omp, 1rkm, 1oib, 1nyl, 1urp, 1akz, 1d6m, 1gp2, 2pfk and 1pjr. The PDB codes in the MolMov subset I include: 1bjy, 1b0o, 1dqy, 6tim, 1f8a, 1dvj, 1crx, 1ex7, 1ftm, 3mbp, 1qai, 2rkm, 1quk, 1gtr, 2dri, 1ssp, 1i7d, 1cip, 1pfk and 3pjr. The MolMov set includes very diverse types of ligands and protein architectures (TABLE 4) and the number of amino acids per protein ranked from 156 to 647. Finally, for each structure in these subsets, 26 normal modes of vibration were calculated using ElNèmo (Suhre 2004) and 11 protein conformations derived for each. Thus, the MolMov set includes a total of 5,720 protein structures, with 2,860 protein structures in each subset. TE HACE FALTA HABLAR DE LOS DATOS DE LA TIM EN DONDE COMPARAS LOS CONSERVADOS CON LOS CENTRALES

**Molecular Dynamics Simulations (MD).** The group of Nina Pastor carried out the MD simulations used for the study of functional conformers. The initial structure for the simulation of free TBP was derived from 1TBP (Kim 1993). For the bound TBP the initial structure was 1YTB (Chasman 1993) (chains B and D), which is the carboxyl terminal domain of TBP from *Saccharomyces cerevisiae* bound to a TATA box hairpin (5′ TATATAAA 3′, CYC1); the bases in the hairpin were removed, and only 10 base pairs were kept (the TATA box and one-base pair at the 5′ and 3′ end). The complex of TBP bound to sequence 5′ GCGCGCGCGC 3′ (CG) was constructed introducing the necessary modifications to the 1YTB structure using the Biopolymer module of Insight II program. The structures were solvated placing the solute molecules on a cubic TIP3 water box and removing all the waters within 2.5 Å of the solute. The cubic water box was trimmed to a hexagonal box employing the Simulaid program (Mezei 1997). Initially, the water molecules and sodium atoms were

submitted to an energy minimization using 4 stages of 500 Steepest Descent (SD) steps and 2 stages of 1000 Adopted Basis Newton-Raphson (ABNR) steps. After solvent minimization, periodic boundary conditions (PBC) were turned on employing the CRYSTAL module of the CHARMM (Brooks 1983) program version 28 using CHARMM27 parameters (Foloppe 2000) (MacKerell 1998). The solvent was again minimized with 500 ABNR steps keeping the solute molecule fixed. Two final minimization stages were applied to the whole system with 250 SD steps and 250 ABNR steps. The solvent was equilibrated with 150 ps of molecular dynamics using a 1.5 fs step in the NPT ensemble at 300 K with the Leap-Frog integrator. Later, the whole system was equilibrated using the same protocol for the solvent. The Berendsen algorithm was used. A value of 600.0 atomic mass units (amu) was used for the mass of the pressure piston. The reference pressure was set to 1 atm. The Langevin piston collision frequency was set to 10.0 $ps^{-1}$. The Langevin piston bath temperature was set to 300 K. The Hoover constant temperature was used. The Hoover reference temperature was set to 300.0 K. The mass of the thermal piston was set at 1000 $kcal*ps^{-2}$. The target temperature was 300 K. The image and neighbor list update were done when necessary (heuristic test), with a distance cut-off set to 14 Å; electrostatic interactions were shifted, and van der Waals interactions were switched, to ensure smooth forces at the cutoff distance. All calculations were performed using SHAKE algorithm and an integration time step of 1.5 fs was used. All the systems were simulated for 10.65 ns using PBC with the CRYSTAL module of CHARMM in the NPT ensemble at 300 K with the Leap-Frog integrator saving coordinates every 100 steps. The last 9 ns were used for analysis.

**Solvent Accessible Surface Area of Residues (SASA).** To measure the movements of individual residues during the MD simulations, we calculated the SASA of each residue in each conformer (AQUI DEBES DECIR AL MENOS CON QUE SOFTWARE LO CALCULASTE).

**Estimating the Reliability of the Predictions.** Two measurements were used to account for this: sensitivity and specificity. Sensitivity, Se, is defined as Se = (TP+FN)/AP, where TP: true positives, FN: false negatives and AP: all positives. In our case, AP are all the critical residues determined experimentally, TP are the critical residues correctly predicted and FN the critical residues not predicted as critical. Specificity, Sp, is defined as Sp = (AN−FP)/AN; where AN: all negatives and FP: false positives. In our case, AN are the non-critical residues determined experimentally and FP are the residues predicted as critical, which are not critical. Additionally, to compare the sensitivity of the predictions in paired comparisons, we defined the Combined Sensitivity parameter as:

//

Where C1 refers to the observed central residues in protein 1 and, C2 refers to the observed central residues in protein 2. M is the number of central residues that are truly critical residues for either protein 1 or protein 2. Thus, 2< = CS> = 0 to distinguish it from Sensitivity.

**Prediction of Critical Residues as Conserved Residues.** The ConSurf server (Berezin 2004) was used for this. The parameters used to run the ConSurf server were: Maximum likelihood method used to calculate the conservation scores, PSI-BLAST E-value = 0.001, maximum

number of homologous sequences = 50 and the number of PSI-BLAST iterations = 1. Conserved residues were those with the most negative score (color code of 9).

**Identification of Structural Domains.** We wanted to know if a natural division of the vertices of a PG into non-overlapping groups leads to the identification of structural domains in the corresponding protein structure. A good division of a graph into groups is not merely one in which there are few edges between the proposed groups; it is one in which there are fewer than expected edges between groups. If the number of edges between groups is significantly less than we expect by chance, or equivalent if the number within groups is significantly more, then it is reasonable to conclude that the division is meaningful. This idea can be quantified by using the measure known as modularity. The modularity is, up to a multiplicative constant, the number of edges falling within groups minus the expected number in an equivalent network with edges placed at random. Thus, one can search for grouping structure by looking for the divisions of a graph that have positive and large values of modularity. For calculating which partition maximizes the modularity we made a program implementing the spectral method of Newman (Newman 2006) involving adjacency matrices (FIGURE 10b) and applied it to the PG derived from the structure of the ε-subunit of the ATP synthase (PDB code: 1AQT).

# RESULTS

**The Graphs derived from Protein Structures (PGs) show Distinctive Features.** Our first hypothesis is that the graph representations of native protein structures have properties that distinguishes them and can be used to speed the discrimination of near-native protein structures in a conformational space search. We started searching for these properties in a set of graphs built for 1,899 non-redundant protein structural representatives (see Methods). We plotted the most frequently used graph-derived measures (see Methods) since in this way we could compare our results with the literature in order to find a model that could reproduce our findings. The first plot (FIGURE 11) shows the probability P(k) of finding a vertex with degree k in a PG. The most probable vertex degree is 12 (P=0.113). We could fit a Poisson distribution to the data (with a confidence level of 98%). It is important to note that all but the two residues at each end of the protein sequence always have four edges derived from the primary structure just because of the way we built the graphs. We see from the data that it is possible to find vertices that have > 20 edges, which tell us about the compactness that certain regions of a protein could achieve. Next we plotted the average clustering coefficient as a function of the degree CC(k) of a vertex (FIGURE 12), we found that this data could be fitted (with a confidence level of 98%) by a power law distribution with exponent $\alpha$=0.53. We also plotted the average clustering coefficient as a function of the number of vertices CC(n) of a PG (FIGURE 13) and we could fit the data to exponential decay distribution with rate $\beta$=0.64 (confidence level $\alpha$ = 0.907).

The resulting distributions were compared with those corresponding to well-studied graph models: random graphs, scale-free graphs and hierarchical graphs (Barabasi 2004) (Ravasz 2003). The principal result is that none of these models could be adjusted to our observed distributions. The observed distribution of P(k) can be replicated with a random model but that models fails in the rest of the distributions. Likewise, the observed CC(k) distribution can be modeled with a hierarchical network, but this kind of network can't generate the distributions of the other measures. Finally, for the CC(n) distribution we found that can be modeled in a better fashion by a scale-free model (fit with 85% confidence). Recently, a new model of graphs that seems to be more appropriate has been developed. These models, called geometric graphs (GG)[25] (Dall 2002), are constructed on the basis of distance and spatial relationships between objects. We observed that the GG model with a random spatial arrangement of vertices can replicate the observed P(k) distribution.

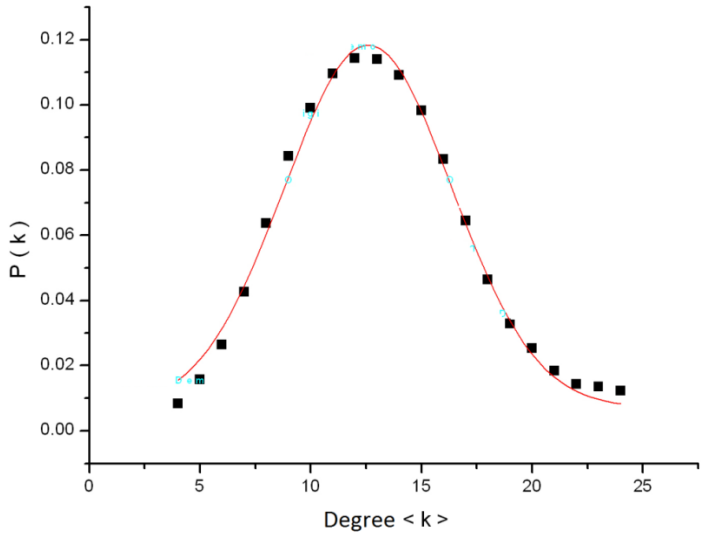


FIGURE 11. The probability P(k) of finding a vertex with degree k in a PG. The most probable vertex degree is 12 (P=0.113). We could fit a Poisson distribution (red line) to the data with a confidence level of 98%.

[25] The model of a GG is as follows: In a d-dimensional space, with a defined distance function and a distance-threshold value d, all vertices that reside within a distance d of each other are connected. The spatial localization of every vertex can be derived from any distribution (random, power-law, exponential).

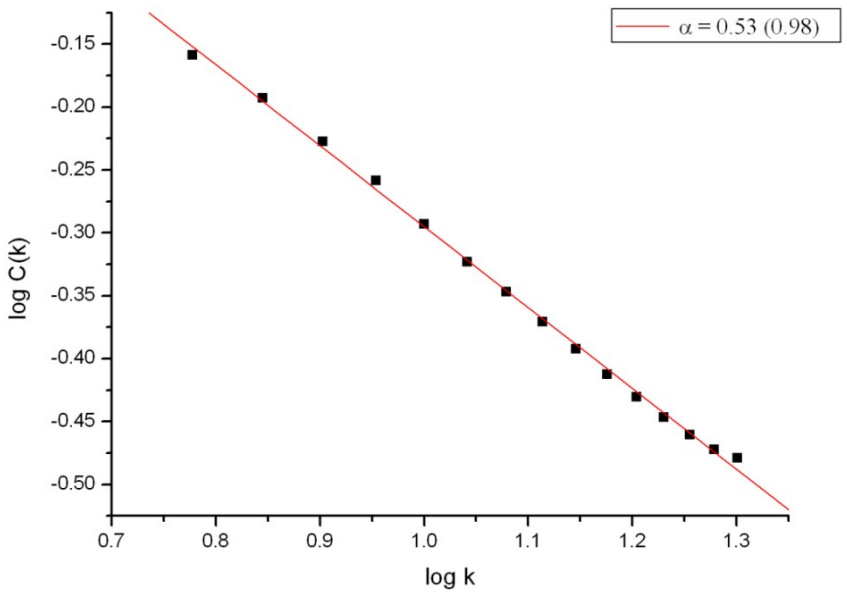


FIGURE 12. Average clustering coefficient as a function of the degree CC(k) of a vertex. The data could be fitted with a confidence level of 98% by a power law distribution (red line) with exponent α=0.53.

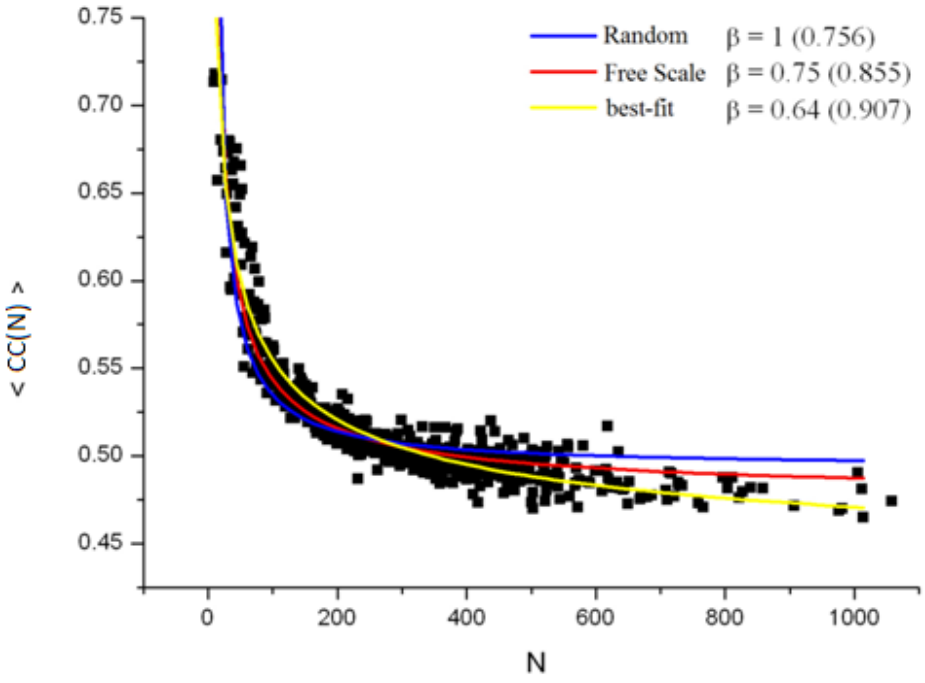


FIGURE 13. Average clustering coefficient as a function of the number of vertices CC(n) of each PG. We could fit the data to exponential decay distribution with rate β=0.64 at a confidence level α = 0.907 (yellow line). Alternative models like random (blue line) and scale-free (red lines) graphs achieved lower fittings.

**Number of Edges Distribution.** A very basic feature that we could look at is the number of edges as a function of the number of vertices that a PG has. FIGURE 14 shows the plot for this. We observe a strong linear relationship that could be the consequence of a homogeneous atomic density across most proteins. There are some PGs that appear to be more compact as observed for the slightly more pronounced slope of the line they seem to conform. We then evaluate if this particular linear relationship is a good predictor of when the PG is derived from a native structure versus when it is derived from an incorrectly folded decoy. For this we downloaded a set of incorrect decoys from http://dd.stanford.edu and compared their number of edges with the number of edges of the corresponding native structures. We found that there is not a significant difference between them (TABLE 1), but we believe that this criterion is necessary even if clearly not sufficient when searching for near-native derived PGs

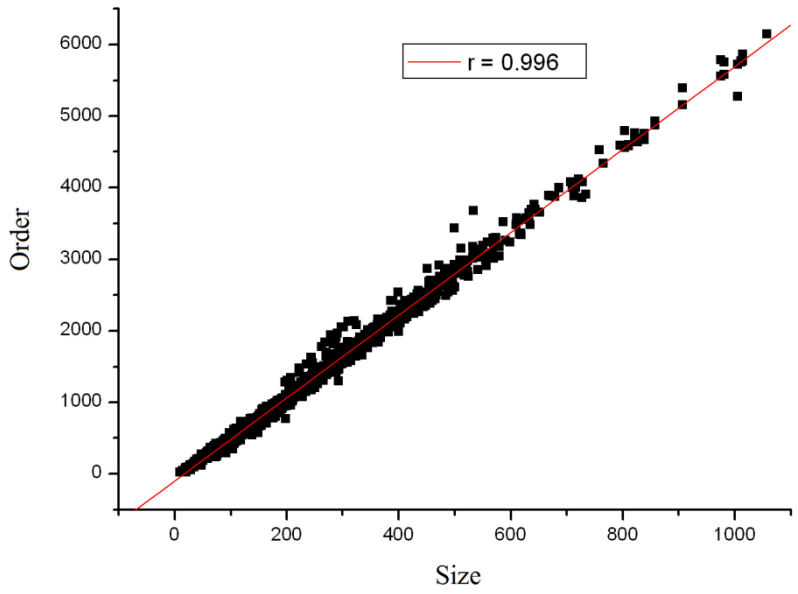


FIGURE 14. Number of edges (order) as a function of the number of vertices (size) for each NRR PG. We could fit the data to a linear correlation (red line) with coefficient r=0.996.

| NATIVE STRUCTURE | EDGES | INCORRECT DECOYS | EDGES |
|---|---|---|---|
| 1bp2 | 610 | 1bp2on2paz | 598 |
| 1cbh | 164 | 1cbhon1ppt | 136 |
| 1fdx | 227 | 1fdxon5rxn | 214 |
| 1hip | 393 | 1hipon2b5c | 359 |
| 1lh1 | 787 | 1lh1on2i1b | 740 |
| 1p2p | 613 | 1p2pon1rn3 | 593 |
| 1ppt | 138 | 1ppton1cbh | 147 |
| 1rei | 1,117 | 1reion5pad | 1,078 |
| 1rhd | 1,533 | 1rhdon2cyp | 1,510 |
| 1rn3 | 639 | 1rn3on1p2p | 591 |
| 1sn3 | 325 | 1sn3on2ci2 | 273 |
| | | 1sn3on2cro | 294 |
| 2b5c | 403 | 2b5con1hip | 373 |
| 2cdv | 444 | 2cdvon2ssi | 487 |
| 2ci2 | 301 | 2ci2on1sn3 | 307 |
| | | 2ci2on2cro | 301 |
| 2cro | 318 | 2croon1sn3 | 316 |
| | | 2croon2ci2 | 274 |
| 2cyp | 1,596 | 2cypon1rhd | 1,488 |
| 2i1b | 769 | 2i1bon1lh1 | 749 |
| 2paz | 639 | 2pazon1bp2 | 572 |
| 2ssi | 504 | 2ssion2cdv | 420 |
| 2tmn | 1,780 | 2tmnon2ts1 | 1,522 |
| 2ts1 | 1,721 | 2ts1on2tmn | 1,722 |
| 5pad | 1,155 | 5padon1rei | 1,038 |
| 5rxn | 247 | 5rxnon1fdx | 221 |

TABLE 1. Number of edges of PGs from 23 native structures and their incorrect decoys from http://dd.stanford.edu.

**A small exercise for predicting native contacts.** We wanted to implement a small routine for predicting native contacts of a selected protein. We choose the structure of the gene V protein from F1 phage (PDB code: 1GVP) since there exists a considerable amount of mutagenesis data regarding their critical residues. Our routine makes use of the structural data so is not an Ab Initio protocol. Still we thought the exercise would be helpful to start benchmarking futures developments. We worked with the matrix representation of the graph, this is called the Adjacency Matrix of the graph **A**(G), and is constructed by filling with 1's the $A_{i,j}$ entries of the matrix corresponding to 3D contacts between pairs of vertices *i* and *j* and with 0's the rest of the entries. We also calculated three MNCPs (See Methods) **1)** One from the complete set of NRR PGs (MNCP1). **2)** One from the 1GVP PG (MNCP2). **3)** A uniform MNCP that we used as null model (MNCP3).

The protocol is the following: **i)** Choose randomly a vertex *i* of the new graph G. **ii)** Verify that the number of edges of this vertex don't be greater that the degree ($d_{i(1GVP)}$) of the corresponding vertex in the 1GVP PG, if not, go back to the first step. **iii)** If its number of

edges is < than $d_{i(1GVP)}$ proceed to assign a new edge with that vertex *j* whose corresponding value in the MNCP is the highest and only if its corresponding $d_{j(G)}$ is less than $d_{j(1GVP)}$, otherwise repeat this step and try to connect it with the vertex with the next highest value in the MNCP. If there is more than one vertex satisfying these conditions proceed to choose one randomly. **iv)** Go back to the second step and apply the procedure to the newly connected vertex *j*. **v)** Stop when all vertices converge to its maximum allowed number of edges or when it had passed 1000 cycles without converging.

We obtained 1,000 graphs when using MNCP2 and MNCP3 in step iii). Interestingly, when using MNCP1 the routine could not converge to a single graph. We then calculated the percentages of native contacts that were formed in the graphs built using the MNCP2 and compared them with the percentages produced by using MNCP3. FIGURE 15 shows the results of this comparison. A few of the graphs built using the uniform-valued MNCP3 can achieve ~8% of the native contacts on top of the 29% derived by the proximity in the primary sequence, ~8% is a maximum achieved by random contacts between different amino acid types and the restraint that we keep the same sequence of degrees that the 1GVP PG. When using the 1GVP-derived MNCP2 we see that the minimum percentage of native contacts is very near the maximum obtained when using MNCP3 and that the use of the 1GVP-derived MNCP2 can contribute with a further ~ 5%.

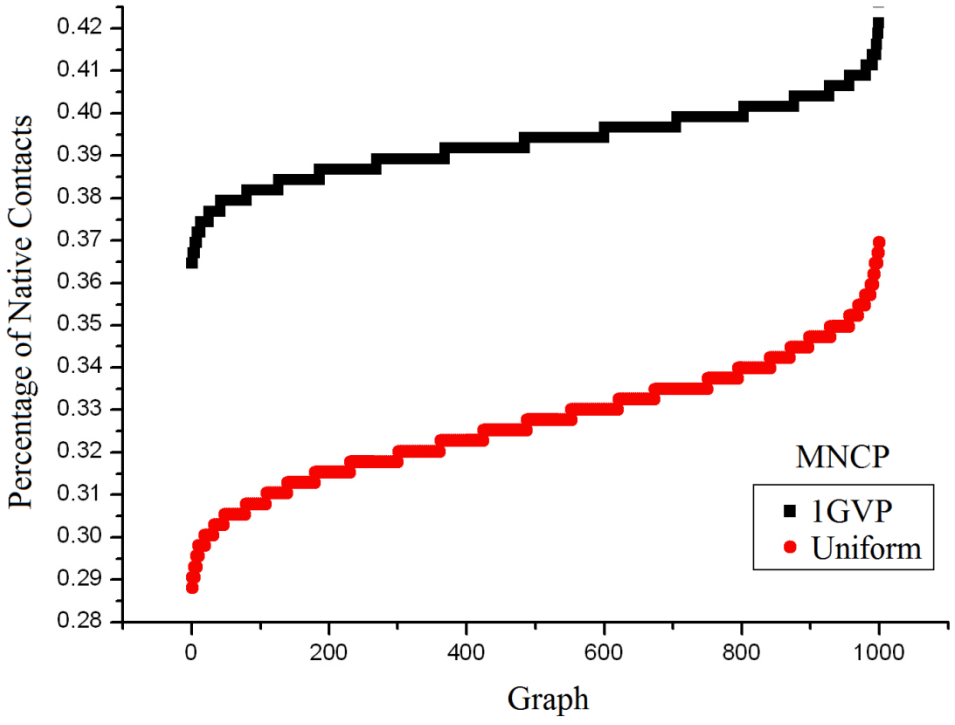


FIGURE 15. Percentages of native contacts that were formed in each of the 1,000 graphs built using the MNCP2 (red line) and each of the 1,000 graphs produced by using MNCP3 (black line).

**The Distributions of Contact Orders.** Another property that can be particularly helpful is the contact order (CO) since this is a natural feature of the PGs; we say that the CO between two vertices that are connected by an edge is equal to the distance at which they sit apart in the protein sequence. We looked at different distributions of the COs calculated from the NRR set. FIGURE 16 is the distribution of frequencies for each value of CO that we found, behind the distribution there is a protein-size effect since the larger values of CO are only possible in sufficiently large proteins. FIGURE 17 is the distribution of frequencies for each value of normalized CO (CO/protein length), thus we are removing the protein-size effect of the previous plot. We observe a fast decay with a long tail that reflects the degree of complex folding that can achieve certain proteins. Then we calculated the average CO and the normalized average CO of each PG (FIGURE 18). Again the non-normalized values bear a protein-size effect that favors the small values, but in the normalized average CO (FIGURE 19) we can see a distribution that reflects the heterogeneity of the fold universe, with the β-sheet proteins likely contributing to the larger CO values. Both curves can be fitted by a log-normal distribution. The FIGURE 20 shows the frequency of each normalized CO value found in each of the PGs of the set of NRR (in this case for each PG the sum of their normalized COs adds one). The diagonal that goes from x-axis = 1 to y-axis = 1 defines the corresponding theoretical maximum frequency that any normalized CO value can have. For example, a maximum of 20% of the total number of residues can have a normalized CO of 0.8 (i.e. 3D contacts between residues separated in the sequence by 80% of the length of the protein). We observe a large region below this diagonal that is scarcely populated, a feature that can be exploited when searching for near-native PGs. Again, we see that most of the contacts are local and that there is a slow decreasing in the frequency of increasingly larger contact orders. Thus, the distribution shows that large contact orders are just a little less probable than medium contact orders.

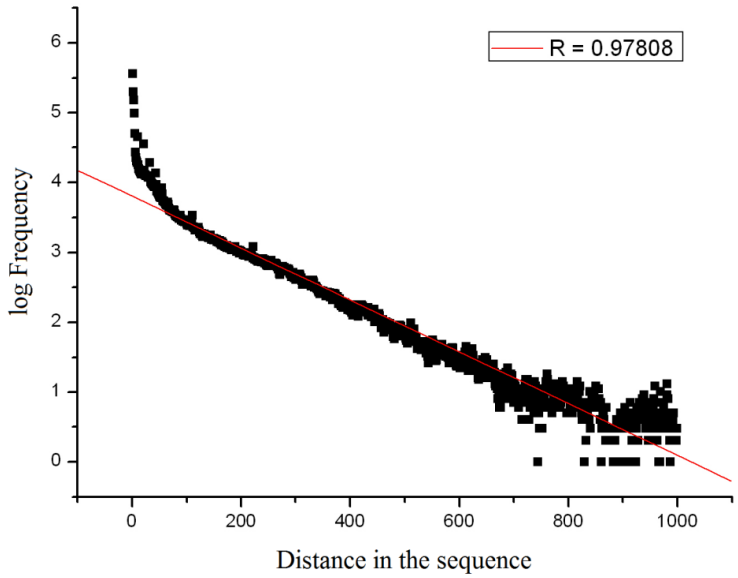


FIGURE 16. Distribution of frequencies for each value of Contact Order (distance in the sequence) found in the NRR PGs set. The major part of the distribution can be fitted by a exponential model (red line) (correlation coefficient r=0-978).

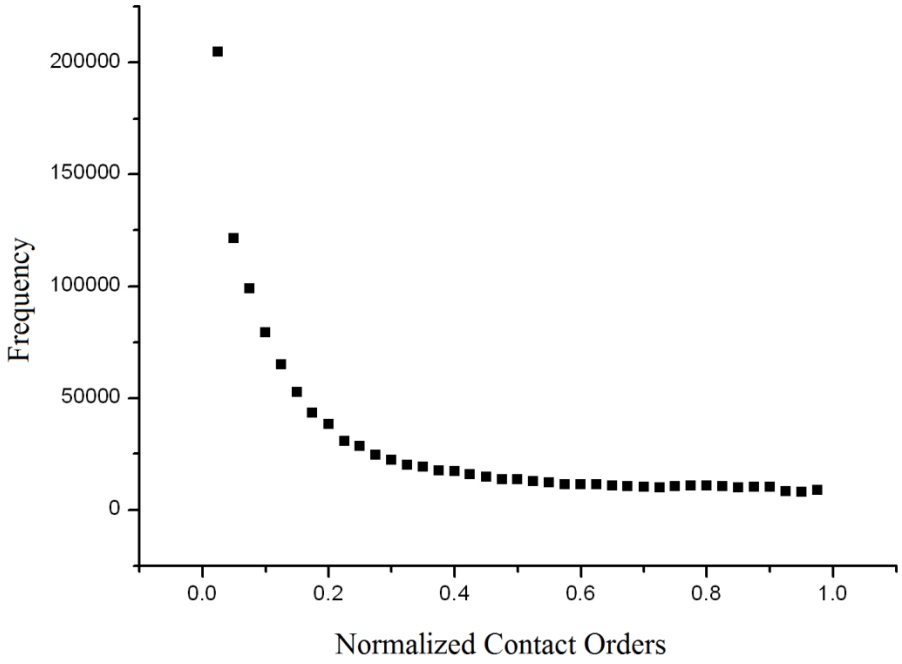


FIGURE 17. Distribution of frequencies for each value of normalized Contact Order (CO/protein length).

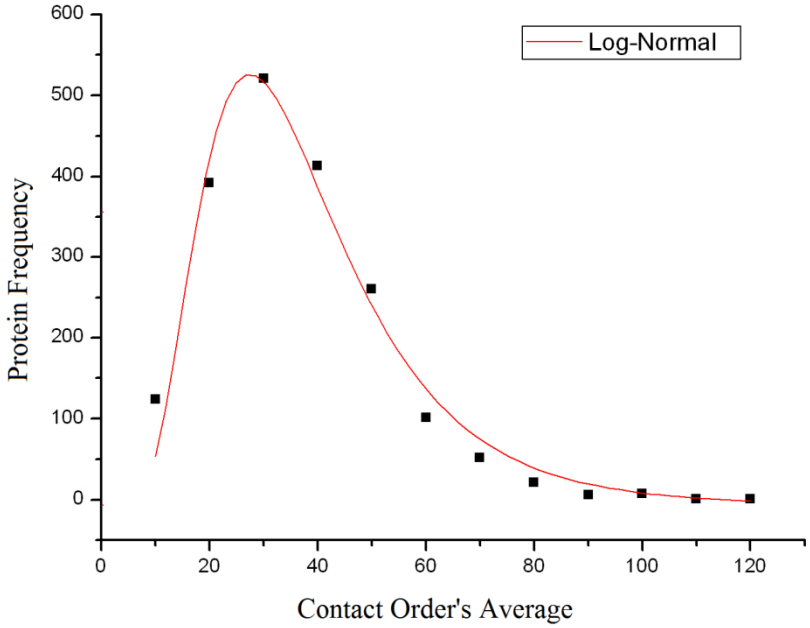


FIGURE 18. Frequency of PGs according to its Average CO. The distribution can be fitted by a log-normal model (red line).

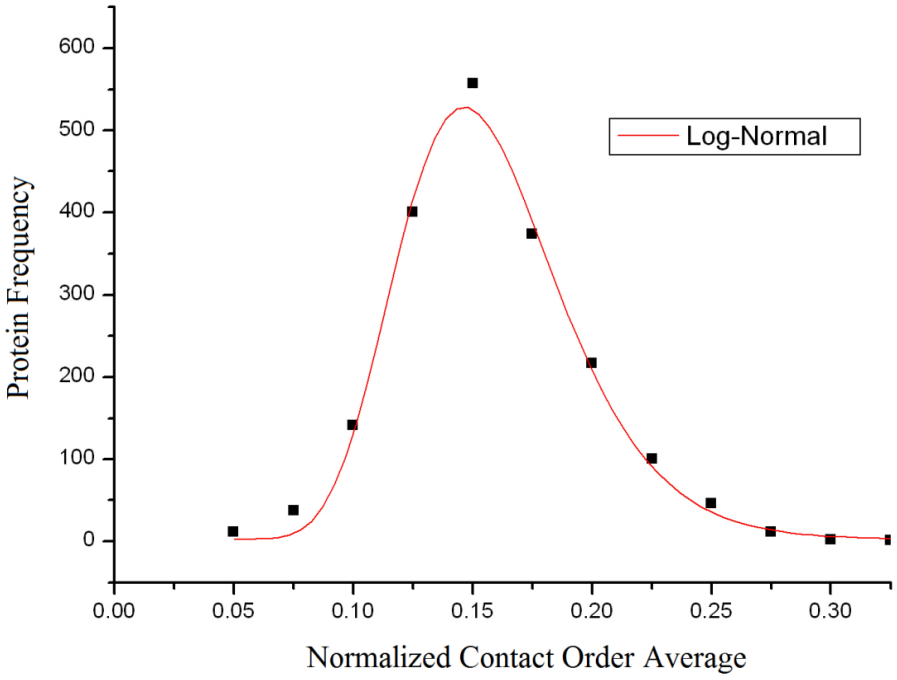


FIGURE 19. Frequency of PGs per Normalized CO Average. The distribution can be fitted by a log-normal model (red line).

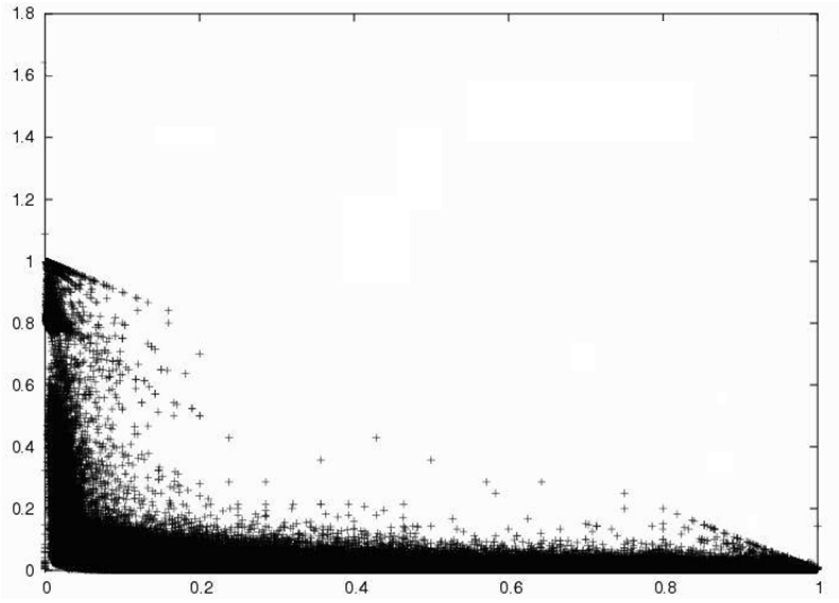


FIGURE 20. Frequency of each normalized CO value found in each of the PGs of the set of NRR (in this case for each PG the sum of their normalized COs adds one). The diagonal that goes from x-axis = 1 to y-axis = 1 defines the corresponding theoretical maximum frequency that any normalized CO value can have.

**Motifs.** When a protein starts folding it forms well-defined local substructures that eventually give rise to secondary structure elements. We wanted to know if these substructures can be identified as graph motifs, that is, particular types of overrepresented subgraphs with respect to a random expectancy. For any given number of vertices n, there exists a finite number of possible connected graphs, and this number grows exponentially with n (Harary 1973). We searched for subgraphs of up to six vertices (those were the largest subgraphs we could look for due to computational restraints since the search time grows approximately as a power of 2 on the size of the subgraphs), that are overrepresented in each of the NRR PGs with respect to a null model of random graphs that shared the same sequence of degrees of the PG in question (see Methods). FIGURE 21 shows the 3-, 4- and 5-motifs found. We observe that in general, the subgraphs that are motifs don't contain the square nor the pentagon subgraphs, by contrast, the triangle subgraph is contained in almost all motifs which could be an indication of steric constrains inside proteins. We then filtered and pooled the NRR PGs in those belonging to the α-class proteins and those belonging to the β-class. We wanted to know if any of the five and six vertices-subgraphs are indicators of α-helices or β-structures. FIGURES 22 and 23 shows that this is not the case, the frequency of each 5- and 6-subgraph is practically the same in an α-class protein that in a β-class protein.

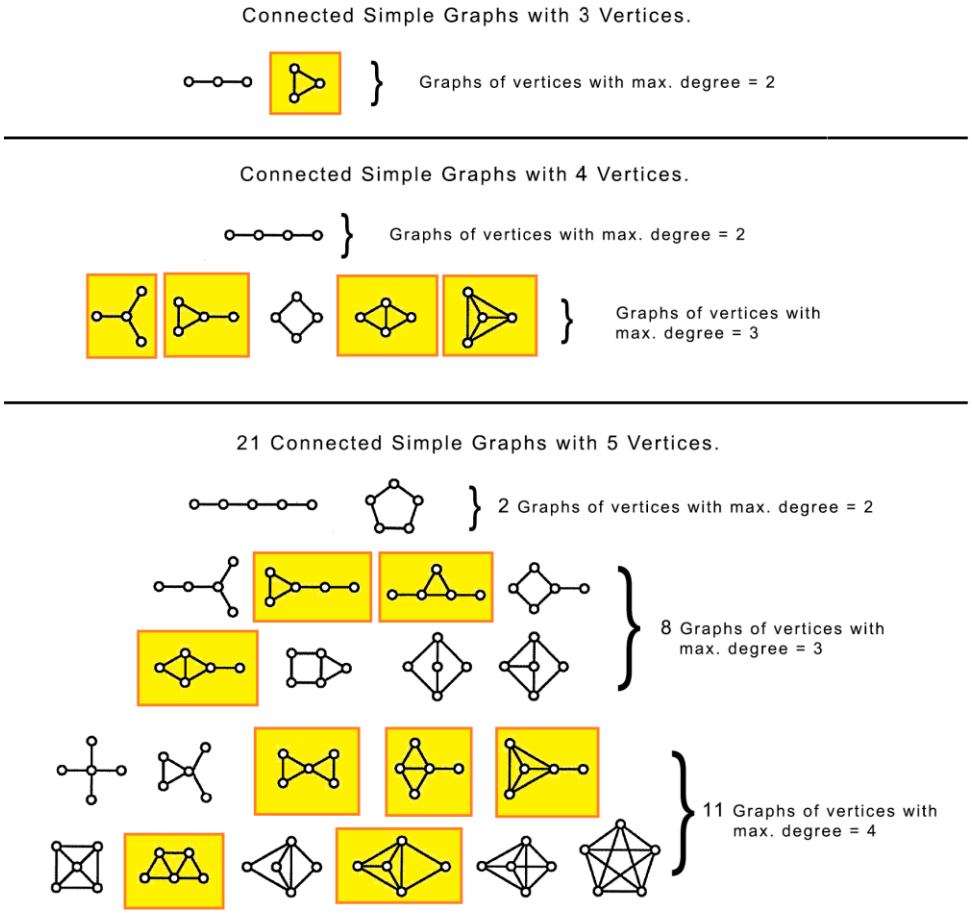


FIGURE 21. Diagrams of all possible connected simple graphs up to five vertices. Yellow squares indicate the Motifs found in the NRR PGs set.

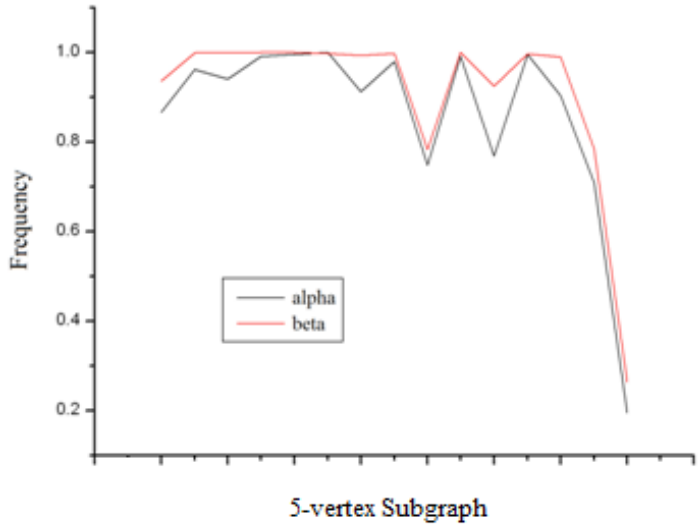


FIGURE 22. Frequency of each 5-subgraph in α-class and β-class proteins. The x-axis runs for the 21 possible 5-subgraphs in increasing number of edges

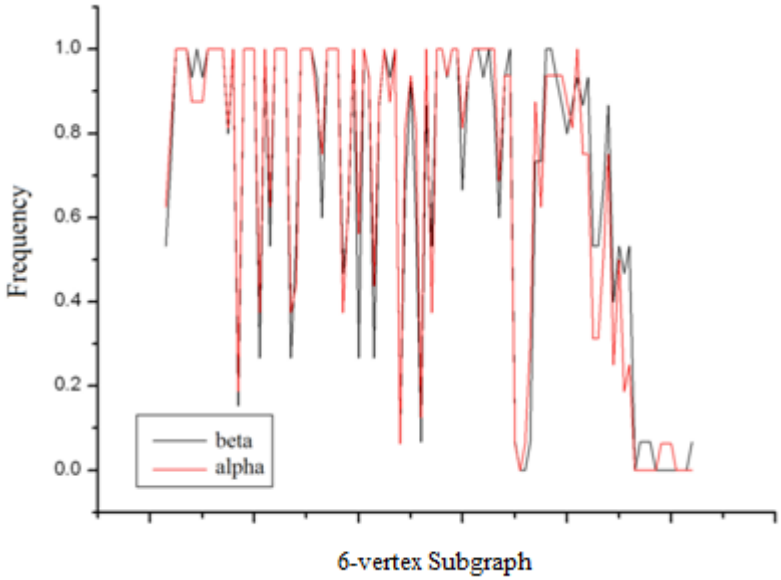


FIGURE 23. Frequency of each 6-subgraph in α-class and β-class proteins. The x-axis runs for the 112 possible 6-subgraphs in increasing number of edges

We also looked for motifs in geometric graphs GG due to their capacity to replicate the PGs degree distribution. One hundred 3D-GGs of 100 vertices each were generated. For this we used an arbitrary Euclidian distance and random spatial distribution. Then we looked for 6-motifs and compared their frequencies with those corresponding of the 6-motifs of the PGs. FIGURE 24 shows the results of the comparison. As observed in the leftmost bar, almost half of the PG's 6-motifs can be also found in 3D-GGs at similar frequencies.

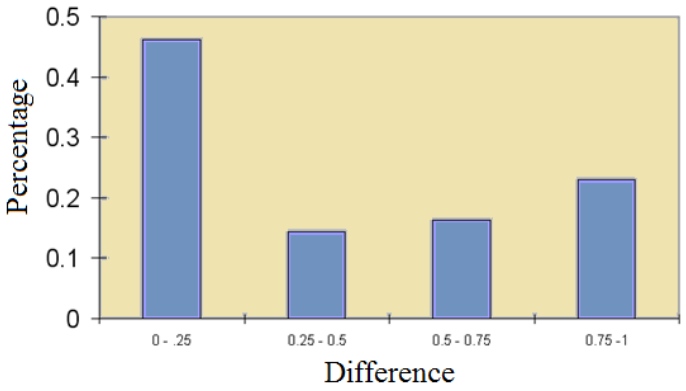


FIGURE 24. Percentages of four ranges of differences in the frequencies of 6-vertex motifs between PGs and geometric random graphs.

**PGs without the contacts derived from the backbone proximity.** We were also interested in studying how much of the extent of the connectedness[26] of the PGs relied on the contacts derived from the proximity that the backbone brings, since our way of constructing the PGs implies that every two consecutive residues in the polypeptide chain are always connected by an edge. We derived two sets of modified PGs from the 1,899 NRR PGs by removing those contacts derived directly from the backbone (1-simplified PGs) and also those contacts between amino acids two residues apart in the backbone (1/2-simplified PGs) (see Methods). To evaluate the connectedness of these two sets of graphs we measured their eccentricity and diameter (see Methods for these definitions). TABLE 2 shows the results. The effect of the 1-removal is minimal in terms of the connectedness. We expected this since the maximum theoretical separation between two amino acids separated by one residue in the backbone is around 4.88 Å, below our cut when defining the edges, thus the reason of studying 1/2-simplified PGs. We see that 7.7% of the 1/2-simplified PGs got disconnected and that there is a significant change in the diameter of the 1/2-simplified PGs versus the PGs, a result that suggest that the backbone-derived edges are in charge of some of the short-circuiting. The change in the average eccentricity is more gradual.

| Graph Type | % of Disconnected Graphs | Avg Eccentricity | e |
|---|---|---|---|
| PGs | 0 | 6.41 | 9.44 |
| 1-Simplified | 0.21% | 6.55 | 9.64 |
| 1/2-Simplified | 7.70% | 6.74 | 10.41 |

TABLE 2. Measures of connectedness in three sets of graphs: PGs, 1-Simplified PGs and 1/2-Simplified PGs.

We explored further the change in the diameter from the PGs to the 1/2-Simplified PGs (FIGURE 25). As expected the diameter increases in a broad range of values for different PGs. In some cases, the diameter diminishes, most probable as a result of the disconnection

[26] The connectedness measures if the graph is formed by one or more unconnected/separate pieces. A graph is simply connected if there exist a path between any two vertices, thus because of the backbone, all our PGs were simply connected.

of certain PGs, those with hinge regions are more susceptible to this. Similar results were obtained for the eccentricity (FIGURE 26). It is not clear what structural correlate is responsible of the pronounced increase in the diameter of some PGs. It could be that those proteins are less compact or have large loops that got cut out.

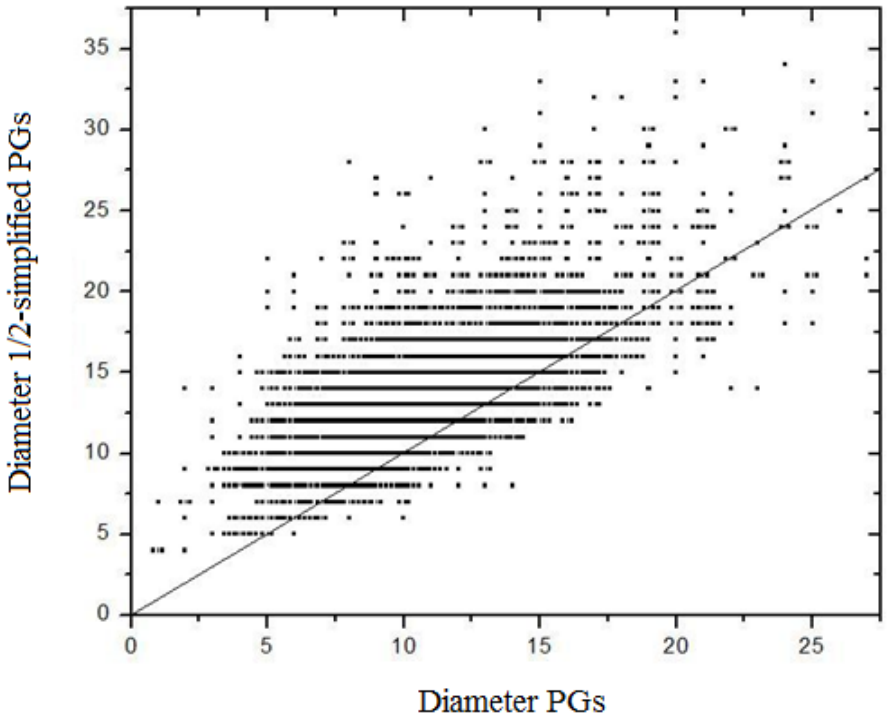


FIGURE 25. Changes in the diameter after removing edges from the original protein graphs (PGs). The removed edges correspond to those contacts between amino acids one and two residues apart in the backbone (1/2-Simplified PGs).

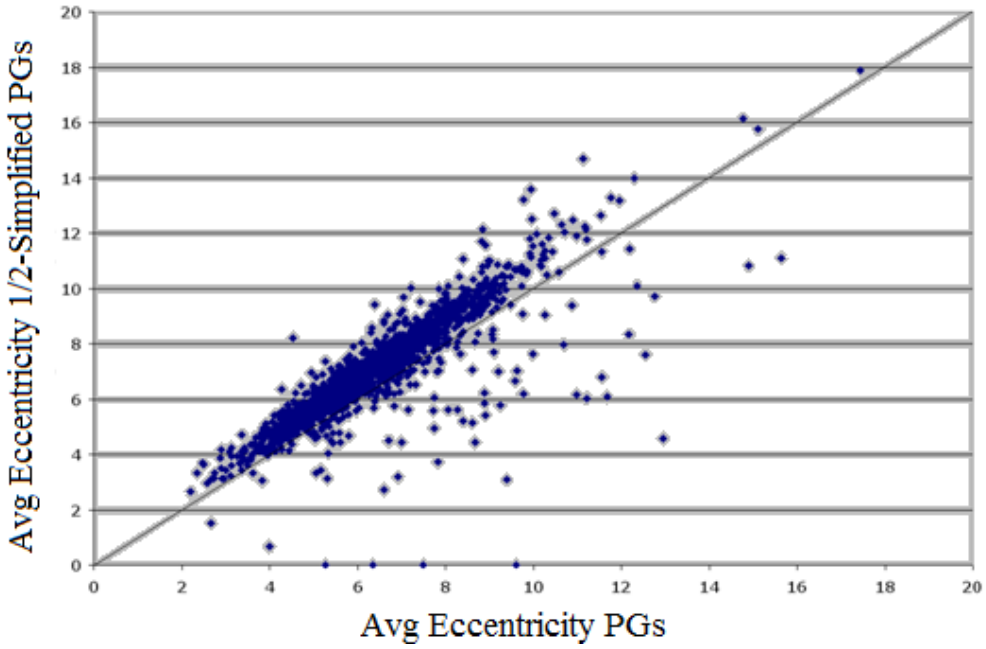


FIGURE 26. Changes in the average eccentricity after removing edges from the original protein graphs (PGs). The removed edges correspond to contacts between amino acids one and two residues apart in the backbone (1/2-Simplified PGs).

**Identification of Structural Domains.** Traditionally, a structural domain is defined as that region of the protein that can acquire its 3D native conformation in an independent manner. This view can sometimes conflict with the evolutionary definition of a domain. Nevertheless, this structural definition is of practical importance in the protein structure prediction field. It is customary to divide every protein in structural domains and then try to predict the structure of each domain separately. As mentioned at the end of the protein structure prediction section in the introductory review, one of the advantages of using a graph representation of protein structures is that we can make use of a vast ensemble of techniques already developed in graph theory for different problems. In this case, we are interested in testing the effectiveness of a graph theory algorithm that improves the graph partition in order to identify the structural domains.

This algorithm (Newman 2006) resolves effectively the problem of dividing a graph in natural groupings. This concept involves the notion of modularity: inside a graph there is a set of vertices that are more connected between them that with the rest of the graph. The algorithm uses spectral theory to divide randomly the graph in two groups of vertices and then evaluate if the number of intra-group connections are significantly more (with respect to a random model) than the number of intergroup connections. This process is carried out for all possible partitions for selecting the one with a maximum score of significance. We applied this method to the structure of the epsilon subunit of the proton-translocating ATP synthase (PDB code: 1AQT), the results are showed in FIGURE 27. We found a strikingly good distinction of the two manually recognized structural domains of this protein. Just a few of

the residues in the N-terminal domain (red region in the figure) where not correctly assigned.

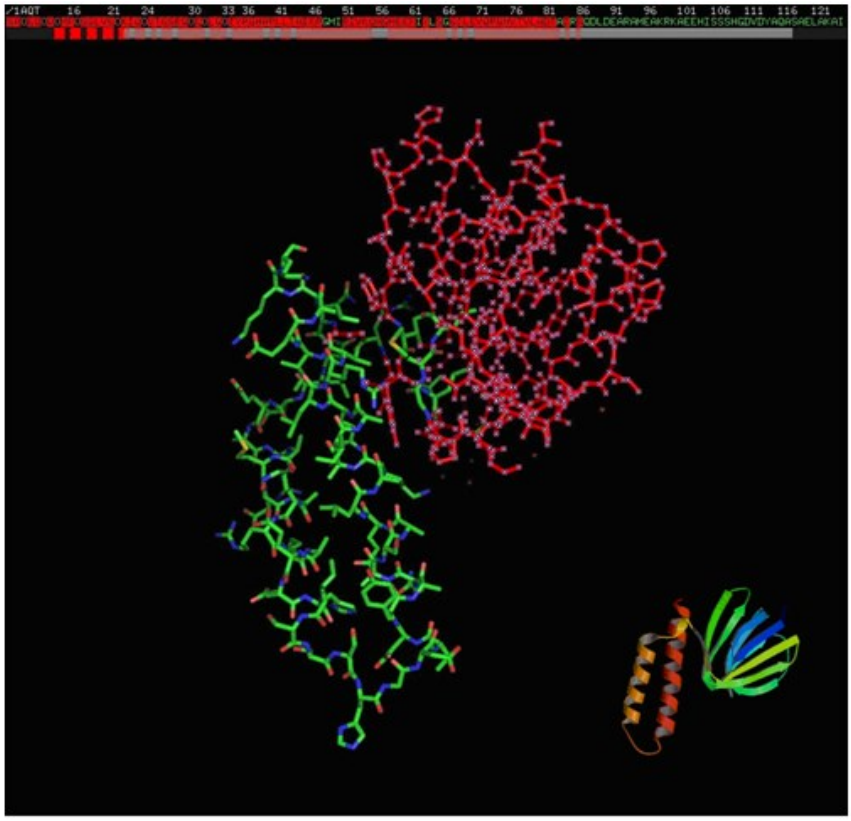

FIGURE 27. Structural domain discrimination of 1AQT by modularity maximization (MM). Each residue is colored in green or red according to the domain assignation given by the MM method. (lower right) A ribbon picture of 1AQT is depicted.

**Mapping Critical Residues for Protein Function onto Protein Conformers.** As pointed out in the introductory review, transitivity is a good predictor of functional residues (Thibert 2005) (Cusack 2007). However, in those previous studies the analysis was limited to a few structures per protein. We think that if function is carried out by conformational changes we can improve the identification of functional residues by considering the PGs of an ensemble of conformers. First, we tested this by constructing PGs for a large number of experimentally determined conformers of the HIV protease and the T4 lysozyme. We identified their central vertices (see Methods) and measured their effectiveness as predictors of functional residues by two parameters: sensitivity and specificity. A high sensitivity means that the method identified a large fraction of true positives and high specificity means that the method identified a large fraction of true negatives (see Methods). We observed that enlarging the number of conformers improves the prediction of functional residues (FIGURE 28)

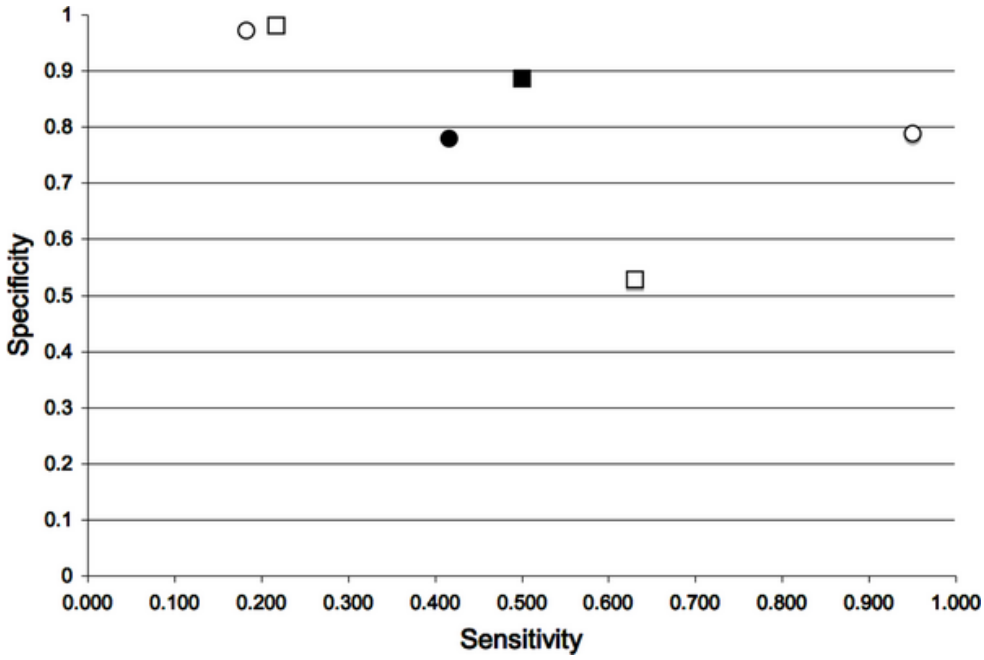


FIGURE 28. Residue centrality as a marker for protein conformational diversity. The sensitivity and specificity for predicting critical residues are plotted for 2 well-characterized proteins: HIV-protease (squares) and the T4 lysozyme (circles). The empty symbols correspond to the values obtained with a single protein conformer and the shadowed symbols correspond to those obtained with multiple conformers. For comparison, the filled symbols correspond to the values obtained with conserved residues predicted as critical residues (see Methods).

In addition, we looked at the triose-phosphate isomerases (TIMs) family. We included 16 protein orthologs of this family with known three-dimensional structures. We observed that central residues shared by most TIM structures, correspond to the most conserved residues (FIGURE 29). These results suggest that central vertices are indeed a good predictor of functional residues as long as the functional residues reside in their functional 3D positioning, so that by screening more conformers the chances of identifying them are increased.

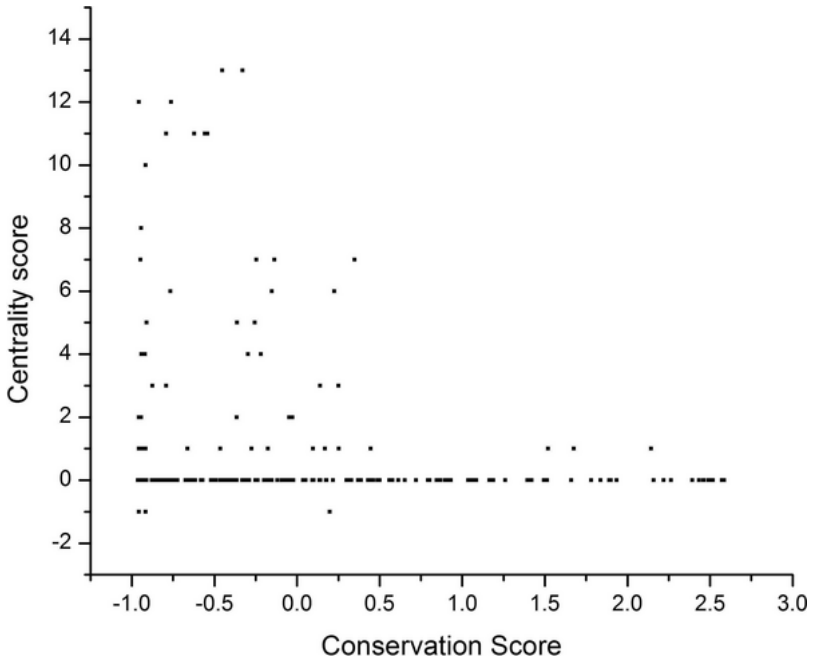


FIGURE 29. Reconstructing functional and phylogenetic relationships from central residues. For 16 structures of the SCOP structural family 51351 (Triose Phosphate Isomerase family, including: 1TIM, 1AMK, 1CI1, 1HG3, 1M6J, 1B9B, 1TCD, 1TRE, 1YYA, 1HTI, 1R2R, 1MO0, 1YDV, 1YPI, 1WYI, 8TIM), we calculated their central residues. Using a multiple sequence alignment, we mapped each central residue into the 1TIM structure. Then, we counted the frequency that each position of 1TIM was found as a central residue in all the family (centrality score). Here, we show the relationship of this frequency with a conservation score for each position of 1TIM derived using the Bayesian ConSeq procedure [50]. In this Bayesian approach, the highly-conserved positions are those with negative scores.

**Different Sets of Protein Conformers Have Different Sets of Central and Critical Residues.** The previous results suggest that different sets of protein conformers harbor different sets of central and critical residues. If this were correct, then it would be possible to find the set of protein conformers harboring in their functional residues in their functional positions: the functional conformers. In FIGURE 30, the fraction of identical central residues shared by every pair of protein conformers (y-axis) was calculated and normalized to 1; thus FIGURE 30 shows that even when two conformers are similar (e.g., some HIV-1 protease conformers share less than 1 Å RMSD values; see FIGURE 31 for the RMSD values), their central residues are not the same (no value of 1 was found between any protein conformer compared). To determine if there is a relationship between centrality and the structural differences between the conformers, we plotted the RMSD against the fraction of central residues shared by every conformer; we found that there is no such relationship (FIGURE 31).

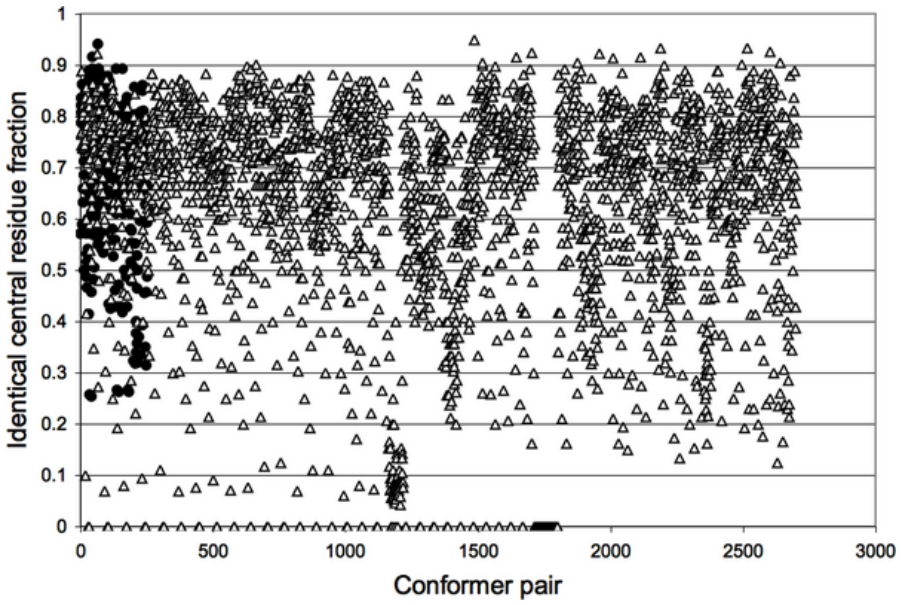


FIGURE 30. Paired comparison of central residues in protein conformers. The fraction of identical central residues shared by every pair of conformers (y-axis) is plotted against every pair of conformer analyzed (x-axis). The results are shown for every pair between the 23 T4 Lysozyme structures analyzed (filled circles) and the 31 complexed HIV-1 protease structures compared against all the 42 non-complexed HIV-1 protease structures (empty triangles). See the Methods section for the PDB codes of the structures used in this comparison.

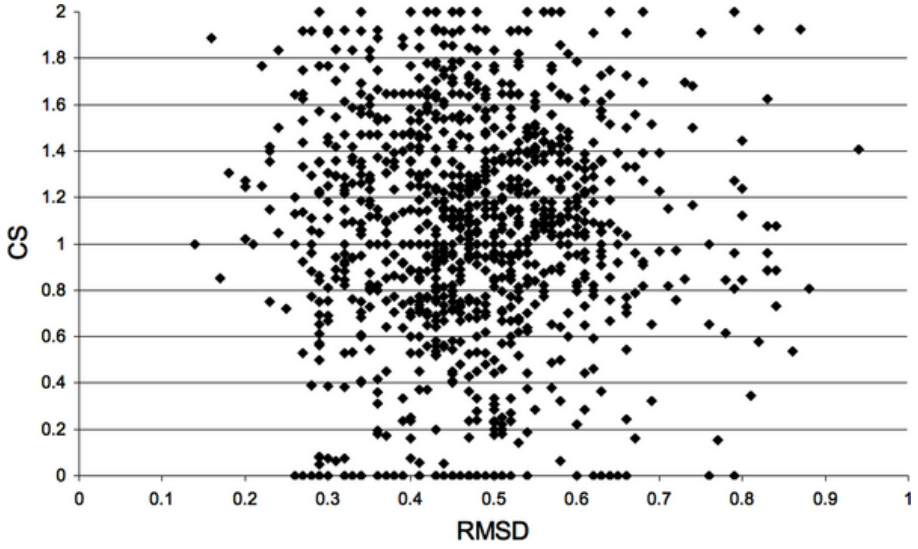


FIGURE 31. Mapping the relationship between RMSD and centrality in crystallographic conformers. Combined Sensitivity (CS) is plotted against the Root Mean Square Deviation (RMSD) values observed for every pair of structures compared. 31 HIV-1 protease structures in complex with a substrate were compared against 42 HIV-1 protease structures without a substrate. See the Methods section for the PDB codes of the structures used in this comparison.

**Screening for Protein Functional Conformers.** We propose that if a protein conformer participates in a given protein function, it must harbor as central residues those that are critical for that function. For instance, protein conformers of an enzyme solved in the presence of its substrate may show as central residues the critical residues involved in binding the substrate. To evaluate this, the sensitivity values reported in the following sections will use as functional residues those critical for ligand binding only, thus differing from the previous results shown so far. We looked at the HIV protease for which there are multiple protein complexes solved with a substrate or an inhibitor. From crystallographic (Zoete 2002) and mutagenesis studies (Loeb 1989) it has been shown that the residues Asp25, Gly27, Asp29, Asp30, Lys46 and Ile50 are critical for substrate binding and/or catalysis. For comparison, we analyzed 42 and 31 HIV protease structures solved in the absence or presence of a substrate analogue, respectively (see Methods). By looking at the fraction of critical residues harbored by these sets of conformers as central residues (expressed as the sensitivity value), we observed that the HIV protease conformers bound to a substrate analogue predominantly show as central residues those that are known to be involved in catalysis (FIGURE 32).

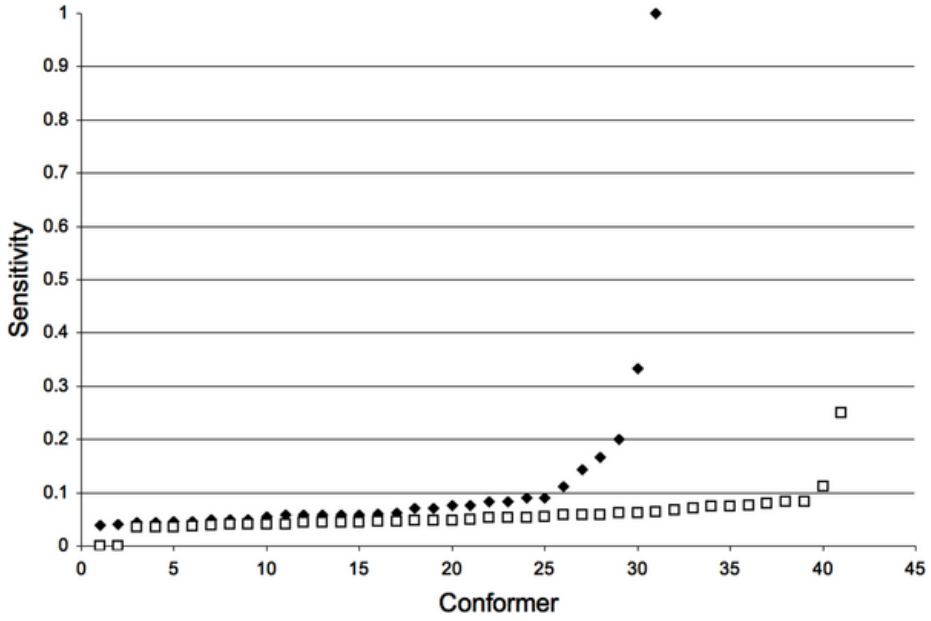


FIGURE 32. Mapping functional conformers in the HIV-protease by centrality measurements. The overall and average sensitivity for predicting critical residues of the HIV-protease was significantly higher when we used crystallographic structures of the HIV-protease associated with a substrate (black dots) than when the crystallographic structures did not include the substrate (white dots). To facilitate visual analysis, the points of each group were sorted in ascending order according to their sensitivity value.

We also analyzed multiple computationally generated protein conformers. In these studies, we used the yeast TATA binding protein (TBP), which has been solved both in the presence (Kim 1993) and in absence (Chasman 1993) of its ligand: the DNA TATA box. It has been previously shown by mutagenesis that at least 53 residues in yeast TBP are involved in DNA binding. We ran four molecular dynamics simulations, and for each of them 63,000 structures were generated. The four simulations included: (a) TBP+WtDNA, TBP in the presence of a high affinity substrate (the TATA sequence), using PDB file 1YTB (Kim 1993) as the starting structure, (b) TBP-WtDNA, TBP that was solved in the presence of the TATA sequence (that is 1YTB), but the DNA was not included in the simulation, (c) TBP-GCDNA, TBP in the presence of a low affinity substrate (GC sequence) generated by *in silico* substitution of the TATA sequence present in 1YTB by the GCGCGCGCGC DNA duplex and (d) TBP solved without substrate, using PDB file 1TBP (Chasman 1993) as a starting structure. The abundance of critical residues for DNA binding found as central residues in these conformers follows the order: (a)>(b)>(c)>(d) (TABLE 3 and FIGURE 33). Also, there is no correlation between the RMSD differences of the conformers and the critical residues for DNA binding harbored by these conformers (FIGURE 34).

| Group | N | Mean | SD |
|---|---|---|---|
| TBP+WtDNA | 63000 | 0.254 | 0.122 |
| TBP-WtDNA | 63000 | 0.249 | 0.116 |
| TBP-GCDNA | 63000 | 0.234 | 0.117 |
| TBP | 63000 | 0.23 | 0.109 |

| Compared Groups | ModelDF | Model MS | Error DF | Error MS | F | $\alpha$ |
|---|---|---|---|---|---|---|
| TBP+WtDNA–TBP-WtDNA | 1 | 0.671 | 125998 | 0.014 | 46.709 | 0.05 |
| TBP-WtDNA–TBP-GCDNA | 1 | 7.319 | 125998 | 0.013 | 533.58 | 0.05 |
| TBP-GCDNA–TBP | 1 | 0.564 | 125998 | 0.012 | 43.873 | 0.05 |

TABLE 3. (Upper part) Each row shows the statistical parameters for each group of TBP conformers derived from molecular dynamics simulations. TBP+WtDNA: TBP with the TATA sequence. TBP-GCDNA: TBP with a GCGC sequence. TBP: TBP originally resolved without DNA and simulated without DNA. (N:Number of conformers, SD:Standard deviation). (Lower part) Each row summarizes the results for a one-way ANOVA (Null hypothesis: mean ($1^{st}$ group) = mean ($2^{nd}$ group)) for the pairs of groups indicated in the first column. In each case the null hypothesis is rejected at the 0.05 significance level (DF:Degrees of freedom, MS:Mean square, F:Calculated F-value, α:Significance level).

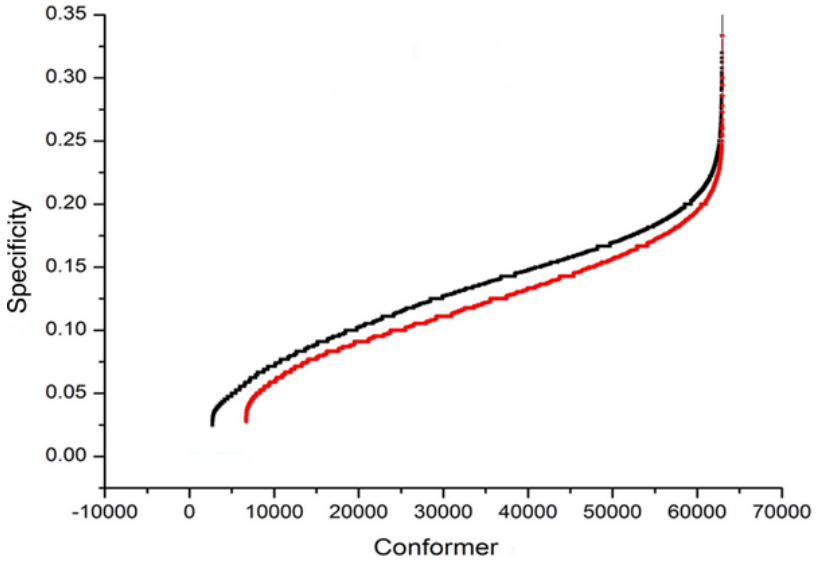


FIGURE 33. Mapping functional conformers in the TBP by centrality measurements. The overall and average sensitivity for predicting critical residues for the binding of the TBP to the TATA sequence was significantly higher when we used structures derived from a molecular dynamics simulation of the TBP associated with the TATA sequence, (labeled TBP+WtDNA, black dots) than when the simulated structures were without DNA, (labeled TBP, red dots). To facilitate visual analysis, the points of each group (63,000 structures each) were sorted in ascending order according to their sensitivity value. See TABLE 3 for a statistical analysis of these data.

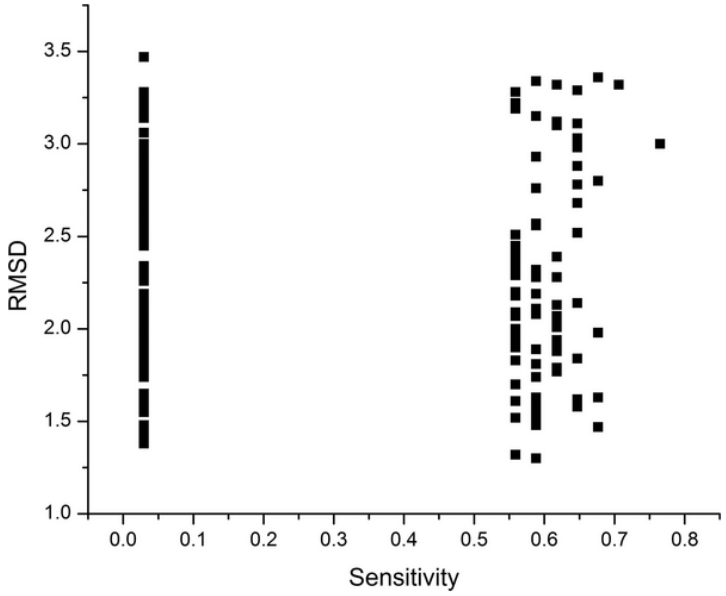


FIGURE 34. Mapping the relationship between RMSD and centrality in molecular dynamics. TBP conformers with the highest and lowest values of both sensitivity and specificity in the four molecular dynamic simulations of TBP were used to show the relationship between the sensitivity value and the RMSD of the conformer with respect to the 1YTB structure.

To analyze the reliability of our results in a larger data set of proteins, we employed the MolMov set that includes a total of 20 different proteins (see Methods and TABLE 4). This set includes a subset of protein structures solved in the absence of a ligand (subset U) and a subset of protein structures interacting with a ligand (subset I). A total of 286 alternative conformations were generated for every protein structure in each subset, providing a total of 2,860 protein structures in each subset, as derived from the normal modes of vibration (see Methods). The critical residues for ligand binding for each protein were assumed to be those conserved residues on the protein surface (see Methods). This assumption includes some degree of uncertainty (conserved residues not necessarily are functionally critical) and provides an additional way to evaluate our procedure. We observed that on average, the proportion of truly predicted critical residues (expressed as sensitivity) in the MolMov subset U is smaller than for the subset I (FIGURE 35a) but not in all cases (FIGURE 35b). We noticed that the MolMov set included 10 proteins for which the predicted critical residues were closer to the ligand (3 Å on average per protein, data not shown) in the crystal structure (FIGURE 35c for an example) than for the other 10 proteins in the MolMov set (FIGURE 35d for an example). Thus, only when the critical residues are truly related to the function of interest, our approach can identify the associated conformations to that function. These results are independent of the nature of either the ligand or the protein analyzed (see TABLE 4).

| PDB Code | Protein Name | Ligand Name | SCOP Classification |
|---|---|---|---|
| 1BJY | Tetracyclin repressor | Tetracycline | All alpha |
| 1DQY | Antigen85c (mycolyltransferase) | Diethyl phosphate inhibitor | Alpha/beta |
| 1CRX | CRE recombinase | DNA | All alpha |
| 1EX7 | Guanylate kinase | Guanosine 5-monophosphate | All beta |
| 1QUK | Phosphate-binding protein | Phosphate | Alpha/beta |
| 1GTR | Glutaminyl-tRNA synthase | ATP | Alpha/beta |
| 2DRI | Ribose-binding protein | Ribose | Alpha/beta |
| 1SSP | Uracyl-DNA glycosylase | Uracyl-DNA | Alpha/beta |
| 1CIP | Guanine nucleotide-bindign protein | Phosphoaminophosphonic acid-guanylate ester | Alpha/beta |
| 3PJR | Helicase | DNA | Alpha/beta |
| 1B0O | Beta-Lactoglobulin | Palmitate | All beta |
| 6TIM | TriosePhosphate Isomerase | 3-Phosphoglycerol | Alpha/beta |
| 1F8A | Peptidyl-prolyl cis-trans isomerase | Phosphoserine-proline peptide | Alpha + beta |
| 1DVJ | Orotinide monophosphate dehydrogenase | 6-AZA Uridine MonoPhosphate | Alpha/beta |
| 1FTM | Glutamate receptor | AMPA | Alpha/beta |
| 3MBP | Maltose-binding protein | Maltotriose | Alpha/beta |
| 1QAI | Reverse transcriptase | Nucleic acid | Multidomain protein (alpha and beta) |
| 2RKM | Oligopeptide binding protein | Lys-Lys peptide | Alpha/beta |
| 1I7D | DNA toposiomerase II | 8-bases single-stranded DNA | Multidomain protein (alpha and beta) |
| 1PFK | Phosphofructokinase | Fructose diphosphate | Alpha/beta |

TABLE 4. The MolMov set. The proteins solved in complex with a ligand in the MolMov set are listed with their ligands. The first ten rows correspond to the protein whose predicted critical residues where close to the ligand; the last ten rows are the proteins whose predicted critical residues were not close to the ligand. The last column indicates the structural classification as indicated in the SCOP database.

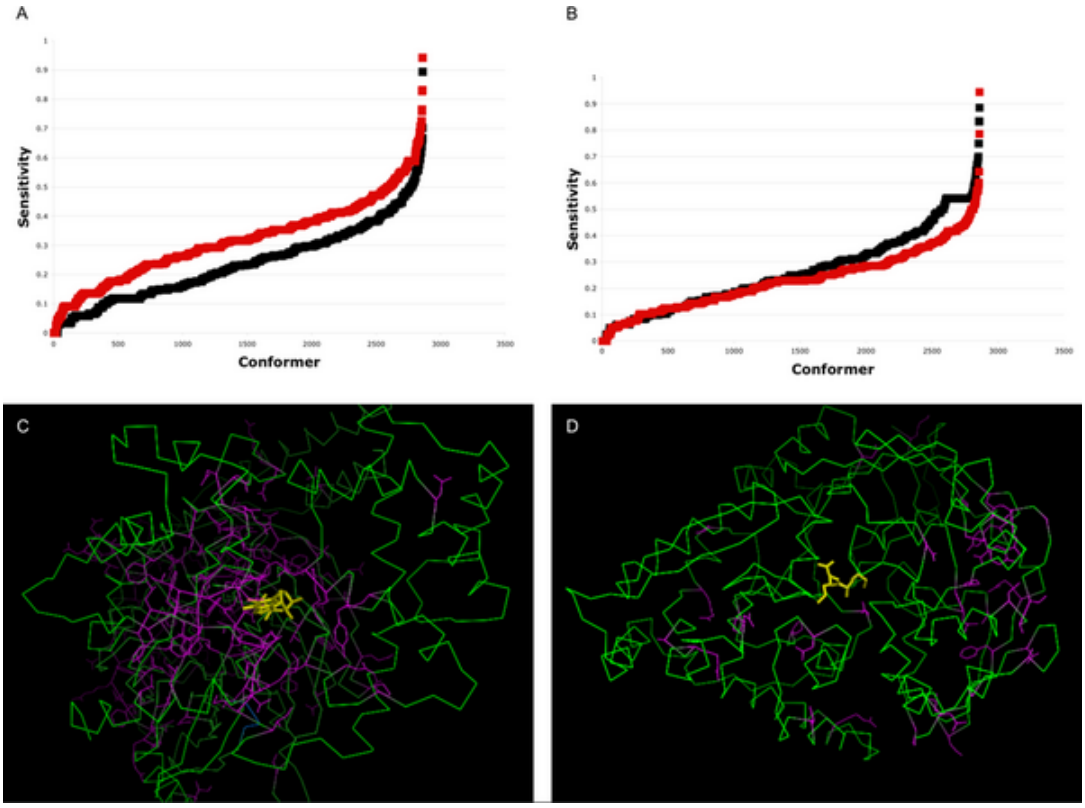


FIGURE 35. Mapping functional conformers in the MolMov set by centrality measurements. The sensitivity value for predicting critical residues in the MolMov set (see Methods) is plotted against each conformer evaluated. (A) The sensitivity values for 10 proteins with predicted critical residues close to the ligand showed significantly higher values when the protein was associated to a ligand (red squares) than the corresponding protein structures without the ligand (black squares). (B) As in (A), but here 10 proteins are shown for which the predicted critical residues were not close to the ligand. To facilitate visual analysis in (A) and (B), the points of each group were sorted in ascending order according to their sensitivity value. (C) 1CIP, Guanine nucleotide-binding protein in complex with a GTP analogue, is an example of a protein where the predicted critical residues were close to the ligand. (D) 2RKM, Oligopeptide-binding protein in complex with Lys-Lys peptide, is an example of a protein where the predicted critical residues were not close to the ligand. In (C) and (D) the ligand is in yellow, the protein in green, and the critical residues in purple.

**Linking Mutagenesis Data to Protein Structure and Dynamics.** TBP mutants that were identified with TBP-DNA binding gel-shift assays does not distinguish between folding-defective mutants and mutants directly involved in DNA binding. In contrast to the HIV protease, there are not numerous structures of the yeast TBP bound to the TATA DNA, thus limiting our ability to establish the structure-dynamics-function relationship of these mutants. For instance, the assumption that only residues less than 5 Å from DNA are directly involved in binding eliminates residues that are at a longer distance from DNA; yet, these distant residues may be at 5 Å or closer to the DNA in some alternative conformations of TBP bound to DNA. If multiple protein structures are computationally generated to determine which residues always fall within a cut-off distance from DNA, there is no a priori

knowledge to determine if all possible conformations were explored. Thus, simply measuring the distance between the ligand and the protein does not provide a comprehensive method to link structure to biological function. Similar reasoning may be applied to energy calculations, since there is no a priori energy value that may be used to specify the relevant residues for binding. In this context, our method does not measure the distance between the ligand and protein, thus is complementary to the criteria based on the distance between the ligand and a protein and could be used to improve our ability to identify critical residues for protein-ligand interactions.

All 53 critical residues in TBP involved in DNA binding qualified as central residues in the structures generated during the simulations (TABLE 5). This indicates that the simulations sampled relevant conformations of TBP associated to the function of the 53 DNA-binding null mutants. However, the centrality criteria used to map critical residues onto protein structures does not distinguish between critical residues for structure and binding. Thus, we examined if there are differences in the presence of these critical residues in the simulations. We would expect that critical residues found exclusively in simulations of TBP in the presence of DNA are more likely to be involved in binding, while those residues prevalently found in all the simulations (frequency> = 0.50) are more likely to be involved in maintaining TBP structure. From TABLE 5, we identified Lys97, Ser118, Pro191, Lys211, Val213 and Thr215 (yeast TBP numbering) as residues critical for binding, whereas critical residues for TBP structure would be Leu67, Leu76, Leu80, Val122, Leu172 and Leu175. In agreement with the yeast TBP-DNA structure, all residues that were predicted to be involved in DNA binding are oriented towards it, while those predicted to be involved in TBP structure are in the protein's core, except for Val122, which faces DNA. Moreover, Leu67, Leu76, Leu80, Leu172 and Leu175 were shown to produce misfolded proteins upon mutation to Lysine (Kim 1993).

| WT Residue | TBP+WtDNA | TBP-WtDNA | TBP | TBP-GCDNA |
|---|---|---|---|---|
| Pro65 | 0.579 | 0.277 | 0.247 | 0.429 |
| Leu67 | 0.786 | 0.621 | 0.552 | 0.57 |
| Asn69 | 0.237 | 0.051 | 0.045 | 0.026 |
| Val71 | 0.268 | 0.016 | 0.015 | 0.015 |
| Leu76 | 0.842 | 0.911 | 0.81 | 0.689 |
| Leu80 | 0.897 | 0.969 | 0.861 | 0.728 |
| Leu82 | 0.435 | 0.286 | 0.255 | 0.161 |
| Lys97 | 0.001 | 0 | 0 | 0 |
| Arg98 | 0.026 | 0.034 | 0.03 | 0 |
| Phe99 | 0.233 | 0.294 | 0.261 | 0.098 |
| Ala100 | 0 | 0.044 | 0.039 | 0.003 |
| Ile103 | 0.045 | 0.022 | 0.019 | 0.037 |
| Arg105 | 0.102 | 0.013 | 0.011 | 0.02 |
| Pro109 | 0.075 | 0.044 | 0.04 | 0.057 |
| Lys110 | 0.029 | 0.017 | 0.015 | 0.034 |
| Thr111 | 0.279 | 0.124 | 0.11 | 0.208 |
| Thr112 | 0.433 | 0.551 | 0.49 | 0.287 |
| Ala113 | 0.049 | 0.029 | 0.025 | 0.098 |
| Leu114 | 0.578 | 0.652 | 0.579 | 0.21 |
| Ile115 | 0.35 | 0.551 | 0.49 | 0.422 |
| Phe116 | 0.184 | 0.244 | 0.217 | 0.389 |
| Ser118 | 0.001 | 0 | 0 | 0 |
| Lys120 | 0.344 | 0.496 | 0.441 | 0.421 |
| Met121 | 0.326 | 0.089 | 0.079 | 0.128 |
| Val122 | 0.665 | 0.961 | 0.854 | 0.569 |
| Thr124 | 0.11 | 0.042 | 0.037 | 0 |
| Gly125 | 0.025 | 0.032 | 0.029 | 0.01 |
| Lys127 | 0.338 | 0.273 | 0.243 | 0.274 |
| Ser136 | 0.221 | 0.122 | 0.109 | 0.103 |
| Arg141 | 0.014 | 0.001 | 0.001 | 0 |
| Ile143 | 0.168 | 0.132 | 0.117 | 0.075 |
| Phe148 | 0.164 | 0.071 | 0.063 | 0.078 |
| Lys156 | 0.213 | 0.156 | 0.139 | 0.223 |
| Asn159 | 0.323 | 0.343 | 0.305 | 0.287 |
| Val161 | 0.247 | 0.221 | 0.197 | 0.141 |
| Leu172 | 0.897 | 0.998 | 0.887 | 0.758 |
| Leu175 | 0.958 | 1.007 | 0.895 | 0.757 |
| Leu189 | 0 | 0.033 | 0.029 | 0 |
| Phe190 | 0.046 | 0.03 | 0.027 | 0 |
| Pro191 | 0.001 | 0 | 0 | 0 |
| Leu193 | 0.479 | 0.497 | 0.442 | 0.336 |
| Ile194 | 0.016 | 0.016 | 0.015 | 0.002 |
| Arg196 | 0.045 | 0.035 | 0.031 | 0.031 |
| Lys201 | 0.003 | 0.003 | 0.002 | 0.082 |
| Val203 | 0.063 | 0.023 | 0.02 | 0.048 |
| Leu204 | 0.087 | 0.053 | 0.048 | 0.061 |
| Leu205 | 0.071 | 0.033 | 0.03 | 0.023 |
| Phe207 | 0.054 | 0.001 | 0.001 | 0.002 |
| Lys211 | 0.049 | 0 | 0 | 0 |
| Val213 | 0.093 | 0 | 0 | 0 |
| Leu214 | 0.453 | 0.336 | 0.299 | 0.245 |
| Thr215 | 0.013 | 0 | 0 | 0 |
| Lys218 | 0.572 | 0.644 | 0.572 | 0.477 |

TABLE 5. The observed Frequencies of DNA-binding null mutant positions (WT residue) for each of the 4 molecular simulations: (a) TBP+WtDNA, (b) TBP-WtDNA, (c) TBP and (d) TBP-GCDNA. The frequencies were obtained by normalizing the number of times any of the residues in this table was detected as central in all the 63,0000 conformers analyzed.

# DISCUSSION AND CONCLUSIONS

Two drivers of research in structural protein science are the ability to predict the 3D structure of proteins when no homology-based template can be found and the identification of the structure of functional conformers. Here we used a graph theoretical approach to attack both problems. Graphs that represent the three-dimensional contacts between residues in proteins (Protein Graphs) have proved to be a valuable tool for uncovering biological-relevant features like functional residues (Amitai 2004) (Thibert 2005) or aiding in model selection for structure prediction (Kolinski 2005).

All-atom free modeling methods are unable to sample extensive regions of the energy landscape. By simplifying protein molecules as graphs we keep enough information to reconstruct the structures, namely, the three-dimensional residue contacts. At the same time, the graphs are simple enough to carry on extensive conformational sampling. A large sampling is of no use if we can't score fast all the models to find the predicted structure. Building on the ability to identify functional residues as central vertices in the graph representations (Amitai 2004) (Thibert 2005), we propose that the protein graphs in the conformational sampling can be rapidly scored by the overlap between their functional residues and central vertices, with the protein graph with most overlapping corresponding to the native structure of the protein.

To make this kind of sampling more efficient and increase the chances of finding true native conformations, we searched for characteristic features that separate the graphs that are derived of protein structures from other types of graphs. In this way, the sampling can be restricted to only the subset of graphs that possess those features. We started by investigating basic graph measures like degree and clustering coefficient distributions. We found out that none of the most studied graph types derived from complex systems (random, hierarchical, scale-free) display the degree and clustering coefficient distributions of Protein Graphs, for this reason, we conclude that any efficient sampling must discard these big subsets of the space of graphs. The geometrical model (Dall 2002) that is based on a distance criterion seems more appropriate and, as expected, we found that the observed degree distribution can be generated with this model. We looked then at motifs in Protein Graphs and noted that most of them don't contain the square nor the pentagon subgraphs, by contrast, the triangle subgraph is contained in almost all motifs which could be an indication of steric constrains inside proteins. We also showed that $\alpha$-helices and $\beta$-sheets cannot be distinguished at the 6-subgraph level. Interestingly, almost half of the 6-motifs of Protein Graphs are shared with the geometrical model.

PGs exhibit a well-preserved linear relationship between number of vertices and number of edges, probably this relationship just reflects a general structural feature related to the compactness of proteins. Another result showed us that a non-negligible portion of 3D native contacts can be recreated pseudo-randomly by just keeping the sequence of degrees of each vertex and that other fraction can be effectively obtained by using

knowledge-based preferential contact usage between different amino acid types. By studying contact orders, we found a clear pattern in their distribution that showed that large contact orders are just a little less probable than medium contact orders. Thus, we got a series of results that showed that PGs do have prominent characteristic features, which is in agreement with one of our hypothesis. This information can be of great value in a carefully developed algorithm for constructing protein-like graphs to improve conformational sampling. Another question that we had was to know to what extent the connectedness of the PGs depends on the backbone derived contacts. We found that less than ten percent of the PGs relied entirely on these contacts to remain simply connected. We also found that a significant portion of the most central residues relied in the backbone contacts for their centrality and that the same could be said of the shortest paths.

A pleasant application of graph theory techniques applied to protein structure problems was our implementation of an algorithm that finds the graph partition with larger modularity, we found that the partition has a good agreement with the natural division of a protein in its structural domains. We expect that this method could add to structural domain identification algorithms.

Under the current view that proteins accomplish their function through dynamics, we hypothesize that the critical residues for function play their roles in those conformations directly involved in function. In such a case, having a method that identifies critical residues in particular protein structures may be capable to select the protein conformations associated to function. In previous reports, it has been shown that central residues to protein structure are related to residues critical for protein function (e.g., folding, catalysis) (Amitai 2004) (Thibert 2005). In these previous studies, central residues have been detected in a single protein structure. However, protein function comprises an ensemble of protein structures and presumably, each protein conformer may harbor a different subset of central and critical residues according to their role in function. Supporting this notion, a report (Vendruscolo 2002) showed that central residues in the folding transition state of 6 proteins map only to critical residues for folding. Here, we show evidence that including multiple conformers of a given protein improves the relationship observed between central residues and critical residues for protein function in three well-studied proteins.

In line with our hypothesis, we found that different protein conformers harbor different sets of central residues, despite their structural similarities (<1 Å) as measured by RMSD, indicating that centrality may depend on subtle geometrical differences between protein structures. This data indicates that central residues seem to be fingerprints of protein conformations. Understanding this correspondence between centrality and protein structure may lead to generate protein structures hosting specific sets of critical and central residues. This will require a more in-depth characterization of the topological features of protein structures represented as graphs. Despite current limitation to generate protein structures as PGs harboring a specific set of central residues, we were able to test the applicability of our hypothesis in identifying functional conformers of proteins through the screening of collections of protein structures. We determined the central residues for 73 experimentally determined conformers of the HIV protease and for 252,000 computationally generated conformers of TBP. For these two proteins, the critical residues for binding the substrate or other ligand have been identified. It is important to note, that it may be possible to have more than a single protein conformer binding a substrate/ligand, provided also that the substrate/ligand exists in several conformations. Given this condition, it is not surprising to find several conformers of these two proteins harboring as central residues those matching

the critical residues for binding the substrate/ligand. As expected, the protein conformers harboring most of the central residues corresponding to the critical residues for binding the substrate/ligand, are the experimentally determined conformers bound to the substrate/. We observed a similar trend for a larger data set of 20 different proteins. However, the protein structures in complex with a ligand cannot be identified if the critical residues provided are not related to the binding of such ligand (in our case, derived from a conservation index of exposed residues). These results are independent of the nature of either the ligand or the protein analyzed. This is evidence that if critical residues for ligand binding are preferentially in their functional position in the protein conformers bound to the ligand they can be identified as central, which is in line with our hypothesis.

We noticed that some conformers derived from the protein structure in the absence of a ligand present large sensitivity values. Understanding these results will require further studies, but a possible explanation could be the recent observations that suggest that conformational exploration by fluctuations of the unbound state may prime the protein to receive the ligand in a more appropriate manner (Boehr 2006) (Schrank 2009). Thus, even when a given conformation in the absence of a ligand is less-likely to harbor central residues matching critical residues, if there is enough conformational sampling we may find this primed structures. That could be the case with the yeast TBP dynamics that we studied. Our results show a large number of conformers in the absence of a ligand with a high proportion of central residues matching critical residues for binding. By contrast, if a protein uses the proposed induced-fit mechanism for ligand binding it may be less likely to found critical residues as central in unbound conformations. That could be the case for the HIV-1 protease, where we found a solid correspondence between critical and central residues just in the bound states.

We examined previously reported mutants of the yeast TBP that have been identified as critical for DNA binding. Since binding to DNA is a dynamic process, a single structure of TBP in complex with DNA may not be sufficient to determine which of the residues have a role in binding or in keeping the structure. We explored the use of our method for distinguishing these residues. Our results show that residues Lys97, Ser118, Pro191, Lys211, Val213 and Thr215 are more likely involved in binding, while residues Leu67, Leu76, Leu80, Val122, Leu172 and Leu175 appeared to be involved in the preservation of the structure of yeast TBP. Even when our method does not use a criterion based on the distance of the protein to the ligand, our results are in consonance with the distance and orientation of the critical residues observed in the structure of yeast TBP in complex with the TATA-box DNA.

Our results support the notion that protein function is achieved through fluctuations between specific of protein conformations. The method shown here may be applied to any other protein of interest to identify its potential functional conformers. We have made available the software to identify central residues. The identification of functional conformers of a target protein is indeed useful in many different areas of research, such as drug design, protein function design and protein-protein interaction predictions, among others. Likewise, the ability to differentially map critical residues onto a spectrum of conformations may increase our capacity to understand the role of specific residues. For instance, in many mutagenesis studies of proteins, especially those that test the in vivo function of the mutants, it is not obvious if the defects in function are related to a folding, processing, binding or catalytic effect. Our method may aid in the interpretation of such data.

# REFERENCES

Agarwal, P. "Enzymes: An integrated view of structure, dynamics and function." *Microbial Cell Factories* 5 (2006).

Agarwal, P., Billeter, S., Rajagopalan, R., Benkovic, S., Hammes-Schiffer, S. "Network of coupled promoting motions in enzyme catalysis." *PNAS* 99 (2002): 2794–2799.

Agmon I., Bashan A., Zavirach R., Yonath A. "Symmetry at the active site of the ribosome: structural and functional implications." *Biol. Chem.*, 2005: 833–844.

Alexander, PA, He Y, Chen Y, Orban J, Bryan PN. "A minimal sequence code for switching protein structure and function." *PNAS* 106 (2009): 21149-21154.

Alm, E., Baker, D. "Matching theory and experiment in protein folding." *Current Opinion in Structural Biology* 9 (1999): 189-196.

Altschuh D, Lesk AM, Bloomer AC, Klug A. Correlation of co-ordinated amino acid substitutions with function in viruses related to tobacco mosaic virus. J Mol Biol. 1987 Feb 20;193(4):693-707. doi: 10.1016/0022-2836(87)90352-4. PMID: 3612789.

Alva, V., Michael Remmert, Andreas Biegert, Andrei N. Lupas, Johannes Söding. "A galaxy of folds." *Protein Science* 19 (2010): 124-130.

Amitai, G., Shemash, A., Sitbon, E., Shklar, M., Metanely, D., et al. "Network analysis of protein structures identifies functional residues." *J Mol Biol* 344 (2004): 1135–1146.

Andreeva, A., Howorth, D., Chandonia, J. M., Brenner, S. E., Hubbard, T. J., Chothia, C. & Murzin, A. G. "Data growth and its impact on the SCOP database: new developments." *Nucleic Acids Res.* 36 (2008): D419–D425.

Aravind, L., Leipe, D. D. & Koonin, E. V. "Toprim — a conserved catalytic domain in type IA and II topoisomerases, DnaG-type primases, OLD family nucleases and RecR proteins. ." *Nucleic Acids Res. 26,*, 1998: 4205–4213.

Ashkenazy, H., Erez, E., Martz, E., Pupko, T., Ben-Tal, N. "ConSurf 2010: calculating evolutionary conservation in sequence and structure of proteins and nucleic acids." *Nucleic Acids Research* 38 (2010): W529–W533.

Atilgan, A., et al. "Anisotropy of fluctuation dynamics of proteins with an elastic." *Biophys J* 80 (2001): 505-515.

Avery, O., MacLeod, C., McCarty, M. "Studies on the Chemical Nature of the Substance Inducing Transformation of Pneumococcal Types." *J Exp Med*, 1944: 137-158.

Ban N., Nissen P., Hansen J., Moore P, Steitz T. "The Complete Atomic Structure of the Large Ribosomal Subunit at 2.4 Å Resolution." *Science* 289, no. 5481 (2000): 905-920.

Barabasi, A., Oltvai, Z. "Network Biology: Understanding the Cells's Functional Organization." *Nature Reviews Genetics* 5 (2004): 101-113.

Battey, JN, Kopp J, Bordoli L, Read RJ, Clarke ND, Schwede T. "Automated server prediction in CASP7." *Proteins* 69 (S8) (2007): 68–82.

Belogurov, Vassylyeva, Svetlov. "Structural Basis for Converting a General Transcription Factor into an Operon-Specific Virulence Regulator." *Molecular Cell* 26 (2007): 117-129.

Berezin, C., Glaser F., Rosenberg J., Paz I., Pupko T., Fariselli P., Casadio R. and Ben-Tal N. "ConSeq: The Identification of Functionally and Structurally Important Residues in Protein Sequences." *Bioinformatics* 20 (2004): 1322-1324.

Boehr, D., McElheny, D., Dyson, J., Wright, P. "The Dynamic Energy Landscape of Dihydrofolate Reductase Catalysis." *Science* 313 (2006): 1638-1342.

Bowie, JU., Luthy R, Eisenberg D. "A method to identify protein sequences that fold into a known three-dimensional structure." *Science* 253 (1991): 164–170.

Brandes, U. "On Variants of Shortest-Path Betweenness Centrality and their Generic Computation." *Social Networks* 2 (2008): 136-145.

Brazma, et. al. *A quick introduction to elements of biology - cells, molecules, genes, functional genomics, microarrays.* 2001. http://www.ebi.ac.uk/microarray/biology_intro.html.

Brooks, B.R, Bruccoleri RE, Olafson BD, States DJ, Swaminathan S, Karplus M. "CHARMM: a program for macromolecular energy, minimization, and dynamics calculations." *J Comput Chem* 3 (1983): 187–217.

Caetano-Anollés, G., Caetano-Anollés D. "An Evolutionarily Structured Universe of Protein Architecture." *Genome Research* 13 (2003): 1563-1571.

CASP9. *9th Community Wide Experiment on the Critical Assessment of Techniques for Protein Structure Prediction.* 2010. http://predictioncenter.org/casp9/docs.cgi?view=presentations.

Chasman, D., Flaherty KM, Sharp PA, Kornberg RD. "Crystal structure of yeast TATA-binding protein and model for interaction with DNA." *PNAS* 90 (1993): 8174–8178.

Chen, J., Brooks CL III. "Can molecular dynamics simulations provide high-resolution refinement of protein structure?" *Proteins* 67 (2007): 922–930.

Cheng, J, Baldi P. "A machine learning information retrieval approach to protein fold recognition." *Bioinformatics* 22 (2006): 456–1463.

Chothia. "Structure of proteins: packing of alpha-helices and pleated sheets." *Proc. Natl Acad. Sci. USA* 74 (1977): 4130-4134.

Chothia. "The Classification and Origins of Protein Folding Patterns." *Annu. Rev. Biochem.* 59 (1990): 1007–1039.

Chothia, Levitt, Richardson. "Structure of proteins: packing of alpha-helices and pleated sheets." *PNAS* 74 (1977): 4130-4.

Cooper, S., Khatib Firas, Treuille Adrien, Barbero Janos, Lee Jeehyung, Beenen Michael, Leaver-Fay Andrew, Baker David, Popović Zoran, and Players Foldit. "Predicting protein structures with a multiplayer online game." *Nature* 466 (2010): 756-760.

Cuff, A., et. al. "The CATH Hierarchy Revisited—Structural Divergence in Domain Superfamilies and the Continuity of Fold Space." *Structure*, 2009: 1051-1062.

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

Dall, J., Christensen, M. "Random Geometric Graphs." *Phys Rev E* 66 (2002): 016121.

Das, R., Qian B, Raman S, Vernon R, Thompson J, Bradley P. "Structure prediction for CASP7 targets using extensive all-atom refinement with Rosetta@home." *Proteins* 69 (S8) (2007): 118–128.

Dorit, R., Schoenbach L, Gilbert W. "How big is the universe of exons?" *Science* 250, no. 4986 (1990): 1377-1382.

Doyle, D. A. Cabral, J. M. Pfuetzner, R. A. Kuo, A. Gulbis, J. M. Cohen, S. L. Chait, B. T. Mackinnon, R. "The Structure of the Potassium Channel: Molecular Basis of K+ Conduction and Selectivity." *Science*, no. 5360 (1998): 69-76.

Duan, Y., Kollman, P. "Pathways to a Protein Folding Intermediate Observed in a 1-Microsecond Simulation in Aqueous Solution." *Science* 740-744 (1998): 282.

Farmer. "The Present Position of Some Cell Problems." *Nature* 58, no. 1490 (1898): 63-67.

Ferreiro, D., Hegler, J., Komives, E., Wolynes, P. "On the role of frustration in the energy landscapes of allosteric proteins." *PNAS* 108 (2011): 3499–3503.

Fischer, D. "Rychlewski L, Fischer D." *Proteins* 51 (2003): 434-441.

Fischer, D. "Servers for protein structure prediction." *Curr Opin Struct Biol* 16 (2006): 178–182.

Flaherty, K. M., McKay, D. B., Kabsch, W. & Holmes, K. C. "Similarity of the three-dimensional structures of actin and the ATPase fragment of a 70-kDa heat shock cognate protein." *Proc. Natl Acad. Sci. USA*, 1991: 5041–5045 .

Flores, S., Echols N, Milburn D, Hespenheide B, Keating K, et al. "The Database of Macromolecular Motions: new features added at the decade mark." *Nucleic Acids Res* 34 (2006): D296–D301.

Foloppe, N., MacKerell AD. "All-atom empirical force field for nucleic acids: I. Parameter optimization based on small molecule and condensed phase macromolecular target data." *J Comp Chem* 21 (2000): 86–104.

Forster A., Church G. "Towards synthesis of a minimal cell." *Mol Syst Biol*, 2006: 45.

Frederick, K., Marlow, M., Valentine, K., Wand, A. "Conformational entropy in molecular recognition by proteins." *Nature* 448 (2007): 325-329.

Friedberg, I., Godzik, A. "Connecting the protein structure universe by using sparse recurring fragments." *Structure* 13 (2005): 1213-1224.

Garcia, J., Gerber, S., Sugita, S., Südhof, T., Rizo, J. "A conformational switch in the Piccolo C2A domain regulated by alternative splicing." *Nat. Struct. Mol. Biol.* 11 (2004): 45-53.

Ginalski, K, Pas J, Wyrwicz LS, von Grotthuss M, Bujnicki JM, Rychlewski L. "ORFeus: Detection of distant homology using sequence profiles and predicted secondary structure." *Nucleic Acids Res 3* 31 (2003): 3804–3807.

Göbel, U., Sander, C., Schneider, R. and Valencia, A. (1994), Correlated mutations and residue contacts in proteins. Proteins, 18: 309-317. https://doi.org/10.1002/prot.340180402

Grant A., Lee D. and Orengo C. "Progress towards mapping **tmotifhe** universe of protein folds." *Genome Biology*, 2004: 107.

Haber, E., Anfinsen, C. "Regeneration of Enzyme Activity by Air Oxidation of Reduced Subtilisin-Modified Ribonuclease." *J. Biol. Chem.*, 1961: 422-424.

Haliloglu, T., Bahar, I., Erman, B. "Gaussian dynamics of folded proteins." *Phys Rev Lett* 79 (1997): 3090–3093.

Harary, F., Palmer, E.M., Graphical Enumeration, Academic Press, NY, 1973, page 90

Havel, T., Crippen, G., Kuntz, I. "Effects of distance constraints on macromolecular conformation. II. Simulation of experimental results and theoretical predictions." *Biopolymers* 18 (1979): 73–81.

Henzler-Wildman, K., et al. "A hierarchy of timescales in protein dynamics is linked to enzyme catalysis." *Nature* 450 (2007): 913-916.

Hume, Douglas. *Bechamp Or Pasteur: A Lost Chapter in the History of Biology.* Kessinger Publishing, 1996 .

Hvidt, A., Linderstrom–Lang, K. "Exchange of hydrogen atoms in insulin with deuterium atoms in aqueous solutions." *Biochim Biophys Acta* 14 (1954): 574–575.

Jacobs, D., Rader, A., Kuhn, L., Thorpe, M. "Protein Flexibility Predictions Using Graph Theory." *PROTEINS: Structure, Function, and Genetics* 44 (2001): 150–165.

Jaroszewski, L., Li Z, Krishna SS, Bakolitsa C, Wooley J, Deacon AM, Wilson IA, Godzik A. "Exploration of uncharted regions of the protein universe." *PLoS Biol.*, no. 7(9) (2009): e1000205.

Jewett M., Forster A. "Update on designing and building minimal cells." *Current Opinion in Biotechnology*, 2010: 697–703.

Jones, DT. "GenTHREADER: an efficient and reliable protein fold recognition method for." *J Mol Biol* 287 (1999): 797–815.

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

Jumper, J., Evans, R., Pritzel, A. et al. Highly accurate protein structure prediction with AlphaFold. *Nature* 596, 583–589 (2021). https://doi.org/10.1038/s41586-021-03819-2

Karplus, K, Barrett C, Hughey R. "Hidden Markov models for detecting remote protein homologies." *Bioinformatics* 14 (1998): 846–856.

Kim, Y., Geiger JH, Hahn S, Sigler PB. "Crystal structure of a yeast TBP/TATA-box complex." *Nature* 365 (1993): 512–520.

Kolinski, A., Bujnicki, J. "Generalized protein structure prediction based on combination of fold-recognition with de novo folding and evaluation of models." *Proteins* 61 Suppl 7 (2005): 84-90.

Kopp, J., Bordoli L, Battey JN, Kiefer F, Schwede T. "Assessment of CASP7 predictions for template-based modeling targets." *Proteins* 6(S8): (2007): 38–56.

Kuruma Y., Stano P., Ueda .T, Luisi P. "A synthetic biology approach to the construction of membrane proteins in semi-synthetic minimal cells." *Biochim Biophys Acta*, 2009: 567-574.

Latek, D., Kolinski, A. "Contact prediction in protein modeling: Scoring, folding and refinement of coarse-grained models." *BMC Structural Biology* 8 (2008): 36.

Lee, D., de Beer TA, Laskowski RA, Thornton JM, Orengo CA. "1,000 structures and more from the MCSG." *BMC Struct Biol.* 11 (2011): 2.

Levitt, M. and Gerstein,M. "A unified statistical framework for sequence comparison and structure comparison." *PNAS* 95 (1998): 5913–5920.

Lodish, et. al. *Molecular Cell Biology » The Dynamic Cell » 1.2 The Molecules of Life.* 2000. http://www.ncbi.nlm.nih.gov/bookshelf/br.fcgi?book=mcb&part=A199.

Loeb, D., Swanstrom, R., Everitt, L., Manchester, M., Stamper, S., et al. "Complete mutagenesis of the HIV-1 protease." *Nature* 340 (1989): 397–400.

MacKerell, A. Jr, Bashford D, Bellott M, Dunbrack RL Jr, Evanseck JD, et al. "All-atom empirical potential for molecular modeling and dynamics studies of proteins." *J Phys Chem B* 102 (1998): 3586–3616.

Mastro, Babich, Taylor, Keith. "Diffusion of a small molecule in the cytoplasm of mammalian cells." *PNAS* 81, no. 11 (1984): 3414-3418.

Matthews, B. W. "Structural and genetic analysis of protein stability." *Annu. Rev. Biochem.*, 1993: 139-160.

Mezei, M. "Optimal Position of the Solute for Simulations." *J Comp Chem* 18 (1997): 812–815.

Milo, R., Shen-Orr, S., Itzkovitz, S., Kashtan, N., Chklovskii, D., Alon, U. "Network Motifs: Simple Building Blocks of Complex Networks." *Science* 298 (2002): 824-827.

Mirny, L., Domany, E. Protein fold recognition and dynamics in the space of contact maps. Proteins. 1996;26(4):391-410.

Misura, K., Chivian D, Rohl CA, Kim DE, Baker D. "Physically realistic homology models built with ROSETTA can be more accurate than their templates." *PNAS* 103 (2006): 5361–5366.

Miyashita, O., Onuchic, J., Wolynes, P. "Nonlinear elasticity, proteinquakes, and the energy landscapes of functional transitions in proteins." *PNAS* 100 (2003): 12570–12575.

Newman, M. "Modularity and community structure in networks." *PNAS* 103 (2006): 8577-8582.

Noireaux V., Libchaber A. "A vesicle bioreactor as a step toward an artificial cell assembly." *roc. Natl Acad. Sci. USA*, 2004: 17669–17674.

Nureki, O., Shirouzu, M., Hashimoto, K., Ishitani, R., Terada, T., Tamakoshi,M., Oshima, T., Chijimatsu, M., Takio, K., Vassylyev, D. G., Shibata, T.,Inoue, Y., Kuramitsu, S. & Yokoyama, S. "An enzyme with a deep trefoil knot for the active-site architecture." *Acta Cryst.* D58 (2002): 1129-1137.

Perutz, Max. "Mechanisms regulating the reactions of human hemoglobin with oxygen and carbon monoxide." *Annual review of physiology*, 1990: 1-25.

Pieper, Eswar N, Davis FP, Braberg H, Madhusudhan MS, Rossi A. "MODBASE: a database of annotated comparative protein structure models and associated resources." *Nucleic Acids Res*, no. 34 (Database issue) (2006): D291–D295.

Popovych, N., Sun, S., Ebright, R., Kalodimos, C. "Dynamically driven protein allostery." *Nat Struct Mol Biol* 13 (2006): 831-838.

Poussu, E., Vihinen, M., Paulin, L. & Savilahti, H. "Probing the α-complementing domain of E. coli β-galactosidase with use of an insertional pentapeptide mutagenesis strategy based on Mu in vitro DNA transposition." *Proteins*, 2004: 681–692.

Ravasz, E., Barabási, A. "Hierarchical organization in complex networks." *Physical Review E* 67 (2003): 026112.

Richardson, Jane. "β-Sheet topology and the relatedness of proteins." *Nature* 268 (1977): 495-500.

Sadreyev, R, Grishin N. "COMPASS: a tool for comparison of multiple protein alignments with assessment of statistical significance." *J Mol Biol* 326 (2003): 317–336.

Santos M, Moura G, Massey S, Tuite M. "Driving change: the evolution of alternative genetic codes." *Trends in Genetics* 20, no. 2 (2004): 95.102.

Schluenzen F., Tocilj A., Zarivach R., Harms J., Gluehmann M., Janell D., Bashan A., Bartels H., Agmon I., Franceschi F., Yonath A. "Structure of Functionally Activated Small Ribosomal Subunit at 3.3 Å Resolution." *Cell* 102, no. 5 (2000): 615-623.

Schrank, T., Bolen, W., Hilser, V. "Rational modulation of conformational fluctuations in adenylate kinase reveals a local unfolding mechanism for allostery and functional adaptation in proteins." *PNAS* 106 (2009): 16984 –16989.

Shaw, D., et. al. "Atomic-Level Characterization of the Structural Dynamics of Proteins." *Science* 330 (2010): 341-346.

Shi, J, Blundell TL, Mizuguchi K. "FUGUE: sequence-structure homology recognition using environment-specific substitution tables and structure-dependent gap penalties." *J Mol Biol* 310 (2001): 243–257.

Shuman, S., Lima, C. "The polynucleotide ligase and RNA capping enzyme superfamily of covalent nucleotidyltransferases." *Current Opinion in Structural Biology* 14 (2004): 757–764.

Simons, K., Kooperberg C, Huang E, Baker D. "Assembly of protein tertiary structures from fragments with similar local sequences using simulated annealing and Bayesian scoring functions." *J Mol Biol* 268 (1997): 209–225.

Sinha, N. & Nussinov, R. "Point mutations and sequence variability in proteins: redistributions of preexisting populations." *Proc. Natl Acad. Sci. USA*, 2001: 3139–3144.

Skolnick. "Automated structure prediction of weakly homologous proteins on a genomic scale." *PNAS* 101 (2004): 7594-7599.

Soding, J. "Protein homology detection by HMM-HMM comparison." *Bioinformatics* 21 (2005): 951–960.

Suhre, K., Sanejouand Y-H. "ElNémo: a normal mode web server for protein movement analysis and the generation of templates for molecular replacement." *Nucleic Acids Res* 32 (2004): W610–W614.

Sumner, James. "The isolation and crystallization of the enzyme urease." *Journal of Biological Chemistry* 69 (1926): 435-441.

Susuki, et. al. "Modulation of microRNA processing by p53." *Nature*, 2009: 529-533.

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

Todd, A., Orengo C., Thornton J. "Sequence and Structural Differences between Enzyme and Nonenzyme Homologs." *Structure*, 2002: 1435–1451.

Tokuriki, N., Tawfik, D. "Protein Dynamism and Evolvability." *Science* 324 (2009): 203-207.

Tuinstra, R., Francis C. Peterson, Snjezana Kutlesa, Sonay Elgin, Michael A. Kron, Brian F. Volkman. "Interconversion between two unrelated protein folds in the lymphotactin native state." *PNAS* 105 (2008): 5057–5062.

Unger, R., Uriel S, Havlin S. "Scaling law in sizes of protein sequence families: From super-families to orphan genes." *Proteins* 51 (2003): 569–576.

Vendruscolo, M., Paci, E., Dobson, C.M., Karplus, M., *Nature* 409 (2001) : 641.

Vendruscolo, M., Dokholyan, N., Paci, E., Karplus, M. "Small-world view of the amino acids that play a key role in protein folding." *Phys Rev E Stat Nonlin Soft Matter Phys* 65 (2002): 061910.

Vieth, M., Kolinski A, Brooks CL III, Skolnick J. "Prediction of the folding pathways and structure of the GCN4 leucine zipper." *J Mol Biol* 237 (1994): 361–367.

Vitkup, Melamud, Moult, Sander. "Completeness in structural genomics." *Nature Structural Biology* 8 (2001): 559 - 566.

Weigt M, White RA, Szurmant H, Hoch JA, Hwa T. Identification of direct residue contacts in protein-protein interaction by message passing. Proc Natl Acad Sci U S A. 2009 Jan 6;106(1):67-72. doi: 10.1073/pnas.0805923106. Epub 2008 Dec 30. PMID: 19116270; PMCID: PMC2629192.

Weiner, S., Kollman PA, Case DA, Singh UC, Ghio C, Alagona G. "A new force field for molecular mechanical simulation of nucleic acids and proteins." *J Am Chem Soc* 106 (1984): 765–784.

Wimberly B., Brodersen D., Clemons W., Morgan-Warren R., Carter A., Vonrhein C., Hartsch T., Ramakrishnan V. "Structure of the 30S ribosomal subunit." *Nature* 407 (2000): 327-339.

Wroblewska, L., Skolnick J. "Can a physics-based, all-atom potential find a protein's native structure among misfolded structures? I. Large scale AMBER benchmarking." *J Comput Chem* 28 (2007): 2059–2066.

Xu, J., Zhang, Y. "How significant is a protein structure similarity with TM-score=0.5?" *Bioinformatics* 26 (2010): 89–895.

Zagrovic, B., Snow CD, Shirts MR, Pande VS. "Simulation of Folding of a Small Alpha-helical Protein in Atomistic Detail using Worldwide-distributed Computing." *J Mol Biol* 323 (2002): 927–937.

Zhang. "TM-align: a protein structure alignment algorithm based on the TM-score." *Nucleic Acids Res* 33 (2005): 2302–2309.

Zhang, Y. "I-TASSER server for protein 3D structure prediction." *BMC Bioinformatics* 9:40 (2008).

Zhang. "Template-based modeling and free modeling by I-TASSER in CASP7." *Proteins* 69(Suppl 8) (2007): 108–117.

Zhang. Y.., J, Skolnick. "Automated structure prediction of weakly homologous proteins on a genomic scale." *PNAS* 101 (2004): 7594-7599.

Zhang.., Y., Skolnick,J. "Scoring function for automated assessment of protein structure template quality." *Proteins* 57 (2004): 702–710.

Zhang..., Y., Skolnick J. "TM-align: a protein structure alignment algorithm based on the TM-score." *Nucleic Acids Res* 33 (2005): 2302–2309.

Zhou, H, Zhou Y. "Fold recognition by combining sequence profiles derived from evolution and from depth-dependent structural alignment of fragments." *Proteins* 58 (2005): 321–328.

Zoete, V., Michielin, O., Karplus, M. "Relation between sequence and structure of HIV-1 protease inhibitor complexes: a model system for the analysis of protein flexibility." *J Mol Biol* 315 (2002): 21–52.

OPEN ACCESS Freely available online



# Computer-Based Screening of Functional Conformers of Proteins

**Héctor Marlosti Montiel Molina[1], César Millán-Pacheco[2], Nina Pastor[2], Gabriel del Rio[1]***

**1** Departamento de Bioquímica, Instituto de Fisiologia Celular, Universidad Nacional Autonoma de Mexico, Mexico City, Mexico, **2** Departamento de Bioquimica y Biologia Molecular, Facultad de Ciencias, Universidad Autonoma del Estado de Morelos, Morelos, Mexico

## Abstract

A long-standing goal in biology is to establish the link between function, structure, and dynamics of proteins. Considering that protein function at the molecular level is understood by the ability of proteins to bind to other molecules, the limited structural data of proteins in association with other bio-molecules represents a major hurdle to understanding protein function at the structural level. Recent reports show that protein function can be linked to protein structure and dynamics through network centrality analysis, suggesting that the structures of proteins bound to natural ligands may be inferred computationally. In the present work, a new method is described to discriminate protein conformations relevant to the specific recognition of a ligand. The method relies on a scoring system that matches critical residues with central residues in different structures of a given protein. Central residues are the most traversed residues with the same frequency in networks derived from protein structures. We tested our method in a set of 24 different proteins and more than 260,000 structures of these in the absence of a ligand or bound to it. To illustrate the usefulness of our method in the study of the structure/dynamics/function relationship of proteins, we analyzed mutants of the yeast TATA-binding protein with impaired DNA binding. Our results indicate that critical residues for an interaction are preferentially found as central residues of protein structures in complex with a ligand. Thus, our scoring system effectively distinguishes protein conformations relevant to the function of interest.

**Citation:** Montiel Molina HM, Millán-Pacheco C, Pastor N, del Rio G (2008) Computer-Based Screening of Functional Conformers of Proteins. PLoS Comput Biol 4(2): e1000009. doi:10.1371/journal.pcbi.1000009

**Editor:** James M. Briggs, University of Houston, United States of America

**Received** October 3, 2007; **Accepted** January 24, 2008; **Published** February 29, 2008



**Funding:** This work was funded by two grants from the Universidad Nacional Autónoma de México (UNAM) to GDR: PAPIIT-IN210705 and Macroproyecto UNAM: Tecnologias para la Universidad de la Información y la Computación; grants J33190-E from CONACyT, and the program ''Cómputo Científico'' (SEP-FOMES 2000) to NP, which gave us unlimited computer access to the IBM-4 Regatta at the Universidad Autónoma del Estado de Morelos.

**Competing Interests:** The authors have declared that no competing interests exist.

* E-mail: gdelrio@ifc.unam.mx