From Structure to Function: the Convergence of Structure Based Models and Co-evolutionary Information

Journal:	*Physical Chemistry Chemical Physics*
Manuscript ID:	CP-ART-12-2013-055275.R1
Article Type:	Paper
Date Submitted by the Author:	08-Feb-2014
Complete List of Authors:	Jana, Biman; Rice University, Center for Theoretical Biological Physics Morcos, Faruck; Rice University, Center for Theoretical Biological Physics Onuchic, Jose; Rice University, Center for Theoretical Biological Physics

From Structure to Function: the Convergence of Structure Based Models and Co-evolutionary Information

Biman Jana[a,b] , Faruck Morcos[a], José N. Onuchic[a]

[a]Center for Theoretical Biological Physics, Rice University, Houston, TX 77005-1827

[b]Department of Physical Chemistry, Indian Association for the Cultivation of Science, Jadavpur, Kolkata-700032, India

Corresponding authors email: José N. Onuchic, jonuchic@rice.edu

Significance

A combination of physical models and coevolutionary information helps to improve our understanding of biomolecular structure and function.

Abstract

Understanding protein folding and function is one of the most important problems in biological research. Energy landscape theory and the folding funnel concept have provided a framework to investigate the mechanisms associated to these processes. Since protein energy landscapes are in most cases minimally frustrated, structure based models (SMB) have successfully determined the geometrical features associated to folding and functional transitions. Structural information, however, is limited particularly about all different functional configurations. This is a major limitation for SBM's. Alternatively, statistical methods to study amino acid coevolution provide information on residue-residue interactions useful to study structure and function. Here, we show how the combination of these two methods gives rise to a novel way to investigate the mechanisms associated to folding and function. We use this methodology to explore the mechanistic aspects of protein translocation in the integral membrane protease FtsH. Dual basin-SBM simulations using the open and closed state of this hexameric motor reveals a functionally important paddling motion in the catalytic cycle. We also find that Direct Coupling Analysis (DCA) predicts physical contacts between AAA and peptidase domains of the motor, which are crucial for the open to close transition. Our combined method, which uses structural information from the open state experimental structure and coevolutionary couplings, suggests that this methodology can explore the functional landscape of complex biological macromolecules previously inaccessible to methods dependent on experimental structural information. This efficient way to sample the functional landscape of large systems creates a theoretical/computational framework able to better characterize the functional landscape in large biomolecular assemblies.

Introduction

The protein structure-function relationship is a fascinating area of biological research. The discovery of protein native structures through crystallography catalyzed many interesting questions in biological physics (1). For instance, the question of how proteins manage to minimize the search time to achieve its native structure over numerous other possibilities (2). The understanting of the mechanisms governing the challenging protein folding problem (3) relies on the energy landscape theory and the folding funnel (4-12). Supported by this theoretical framework, many different models of protein folding have been proposed and structure based models (SBMs) are found to be successful in understanding this complex process (13-16). The next natural step is to explore how this theoretical framework can also be extended towards protein function. This quest to understand function leads to several different modifications of SBMs needed to obtain insights on protein's functional mechanisms (17-20). While SBMs provide a wealth of information about biological process, they are limited by the availability of structural information, particularly about all different functional configurations. Alternatively, the analysis of large amounts of genomic data to understand similar problems has progressed independently. Statistical analyses of genomic data revealed interesting correlations in the mutational patterns of a particular protein family and provides information not only about folding but also function.(21-26). Direct coupling analysis (DCA) is one of these statistical genomic data analysis frameworks that has the power of predicting real physical interactions/contacts in proteins (21). These additional contacts supplement our SBMs when structural information is not available. This integration of the two approaches combining physical models with DCA evolutionary couplings can now be used for both protein folding and most importantly their functional landscape (27, 28).

In this manuscript, we summarize the current development of SBMs and discuss how direct coupling analysis of sequence information uncovers physical contacts in several protein families. We describe how the combination of these two methodologies helps to understand folding and function. Furthermore, we present our recent findings on extending this combined methodology towards the exploration of how complex multimeric molecular motors accomplish their function. We have chosen the FtsH AAA peptidase motor protein that forms a ring structure with six monomer units to translocate proteins into its pore for subsequent degradation.

Energy landscape theory and structure based models

Anfinsen's famous experiment suggests that a protein always folds to a unique three dimensional native structure (1) in timescales varying from sub-milliseconds to hours. This time range is too small for a random search process that in principle has an astronomically large number of conformations (2). Thus, the process of folding should be a biased exploration of the conformational space (3). Energy landscape theory proposes that the interactions present in the native state are energetically stabilized over other possible non-native interactions. This not only ensures a global minimum in the landscape for native state but also facilitates the search process. Conformations having a large fraction of native interactions are therefore partially stabilized. In addition, non-native states or traps are destabilized limiting the energetic frustration by smoothening the landscape to prevent deep entrapment in local minima during the folding event (principle of minimal frustration) (4, 13). Such a construction is called a folding funnel in which the stabilization energy increases towards the native tip and the periphery or diameter of the funnel at a particular value of stability is a measure of configurational entropy. The unfolded proteins that reside on the upper part of the funnel will diffuse down the funnel via an ensemble of converging pathways to the native state (5, 29, 30). In the folding process, for temperature below the folding temperature, proteins gradually become more energetically stable and native. A folding barrier generally occurs and it is a consequence of the imperfect cancellation between the entropic and energetic contributions (31).

Since the landscape is sufficiently smooth, the concept of a folding funnel is implemented in physical models of proteins using information solely from the protein native structure. In its most idealized version, only the native interactions are included and the non-native roughness is completely neglected. The Hamiltonian of these models generally incorporates two contributions from local and non-local interactions. The local part contains harmonic bonding interactions between adjacent atoms, harmonic angle potentials between three consecutive atoms, and dihedral potentials among three consecutive bonds which allow the rotation of bonds to explore conformational space. The non-local part contains native and non-native contributions. Native contacts are represented by an attractive potential between a pair of atoms that are within a cutoff distance in the native structure. The two dimensional representation of these interactions is called the native contact map. The non-native part is represented by a steric potential between pairs of atoms that are not within this cutoff distance. All the parameters in this Hamiltonian are derived from the native structure, hence called a Structure Based Model (SBM). SBMs exist in different coarse-graining versions, for instance, amino acids can be represented by a single bead at the position of its C-α atom (C-α SBM) (15). Alternatively, amino acids can be represented in full atomistic details (all-atom SBM) (16) as well as several other intermediate variations (32). Parameters to build these models can be computed using a web service called SMOG (33).

Overview of folding using SBM

As the SBMs are built using the basic ideas of the folding funnel and the principle of minimal frustration, these models are used widely to simulate folding mechanisms and kinetics. Depending on the temperature, the protein occupies preferentially the unfolded or folded basin. At high temperatures (above the folding temperature), the unfolded basin is the one mostly populated. Conversely, at lower temperatures the folded basin is dominant. Around the folding temperate, proteins interconvert between the folded and unfolded states. The fraction of native contacts, Q, has shown to be an excellent coordinate for folding and to measure the degree of nativeness of configurations during the folding event. Kinetic runs following Q and thermodynamic estimates based on the heat capacity provide vey similar values for the folding temperature and additional support for this proposed approach. These models have shown to be a powerful tool to observe geometrical features of the different stages towards folding. Geometry/topology gives rise to most of the structural heterogeneity of the different pathways towards the folding minimum. SBMs alone have shown to be sufficient to explain and predict the protein folding transition state ensemble determined by several techniques but prominently by φ-values analysis (34). These models are also found to be successful in identifying long-lived structural intermediates in the folding pathway of proteins that are not simple two-state folders. In addition to proteins with simple topologies, proteins with complex topologies, e.g. knotted proteins, are also found in nature. SBMs have also been successful for this more complex situation (35). Interesting extensions of SBM applications include the binding mechanism of two proteins (homo or hetero dimers (36, 37)). Given a dimeric structure, a SBM model is built using inter and intra protein contacts (36, 38, 39). Two types of scenarios are observed for homo-dimers. In one case, monomers fold independently and then bind to form a dimer. During folding, three states are encountered: unfolded, independently folded monomers and dimers. For other dimers, however, monomers cannot fold independently and binding-folding occurs simultaneously. For all these different situations, agreement have been observed between theory and experiments providing strong validation to energy landscape theory and the funnel landscape concept for folding and binding.

Extending the SBM to understand function

A current challenge is to show how this energy landscape that has been mostly used for folding can also be utilized to understand biological function. To perform their function, many proteins, however, need more than a single native structure and involve multiple functional conformations. Thus, it is apparent that building a SBM based on a single native state structure is not sufficient to completely explore its conformational landscape. This observation raises a fundamental question: is the functional landscape of proteins different from their folding landscape? To answer this challenging question, several extenstions have been made to SBMs as we elaborate below.

Multiple basin SBM

As the name suggests, multiple basin SBMs are built with more than one conformational state of a protein while still keeping the fundamental concept of SBM intact. A dual basin-SBM was used to solve “the mystery of the Rop dimer”, a dimer of two helix bundles that switched from a cis arrangement to a trans arrangement upon optimization of the hydrophobic core (40). These models were also used to understand folding mechanism of proteins that are capable of getting crystallized into more than one structure (41). A classical example of using multiple-basin SBM to understand function of an enzyme is adenylate kinase (AKE) (17, 18, 42). AKE is an enzyme that combines ATP and AMP to generate two ADP molecules as well as its reverse reaction to maintain the energy content in cells. AKE has two domains, LID binds to ATP and NMP binds to AMP. Upon binding both the domains undergo large conformational changes to reach its closed state. In its closed state, the reaction takes place and the two domains now open up to release two ADP molecules. Two crystal structures, one for the open state with no ligand and one for the closed one with an inhibitor, have been experimentally determined. Using these two structures, a dual-basin SBM Hamiltonian was derived. The parameters can be calibrated to achieve experimental relative populations between the open/closed states for a given concentration of the ligand. The most important outcome of these simulations was the determination of the sequence of events during closing and opening that shows different routes during the catalytic cycle. Similar methods have been used to understand the functional mechanism of kinase A (43). Therefore, by knowing the important structures in the functional path of a protein, it is possible to explore its functional landscape and kinetic mechanims.

Function as a perturbation of the native state

Another branch of research is dedicated towards understanding function of proteins as a perturbation of its native state structure by different kinds of interactions. These interactions can be ligand binding, binding to another protein or macromolecules, and mutational changes just to name a few. One representative example of this kind of systems is the kinesin-1 molecular motor

(20, 44). These are proteins which use the chemical energy from the ATP hydrolysis to generate mechanical work needed to walk on a microtubule (MT) track, which has structural polarity (minus and plus ends). While walking, these proteins carry important cellular material across the cell. Kinesin-1 is a molecular motor that walks on the MT towards the plus direction in a processive mechanism. The motor is a homo-dimer and uses its two catalytic motor heads to walk on the MT in a hand-over-hand fashion. An interesting question is how it acquires directionality and coordinates between two motor heads to generate a processive walk? Kinesin-1 is a homo-dimer with two structurally symmetric motor heads, a dimerization region and two connecting neck-linker regions (See Figure 1A). A SBM model was built from this symmetric crystal structure. A perturbation occurs when it binds to two consecutive binding sites in the MT at the same time, which is required for processive movement. The interaction sites to the MT are determined from a crystal structure of a single head kinesin bound to the MT. The interacting residue pairs extracted from this structure are used for both heads in such a way that kinesin-1 homo-dimer can bind with the two heads simultaneously. In order to bind the two heads of the dimer to two consecutive binding sites in MT, one head has to rotate to keep the binding orientation correct and to sufficiently extend to fit the MT profile. The neck-linker of the leading head (+ve end) experiences a strain as a consequence of a distortion in its native structure which simultaneously also distorts the nucleotide (NT) binding site of this head (See Figure 1B-C). This asymmetry ensures that the leading head cannot uptake ATP before the trailing head is released from the MT. This coordination is found to be important for functionality and processivity. Once the trailing head is detached, it releases strain from the leading head allowing it to bind ATP. Binding of ATP is introduced in this model as another perturbation by introducing some key contacts between the motor head and the neck-linker. These contacts induce a power stroke mechanism that takes the trailing head towards the +ve direction of the MT, becoming the leading head for the next cycle. In the entire cycle of stepping, the neck-linker goes though an order-disorder transition depending on the NT state of the head. Thus, by putting essential perturbations coming from MT binding and ATP binding, we can understand the mechanochemistry, directionality, and most importantly processivity of this complex molecular motor.

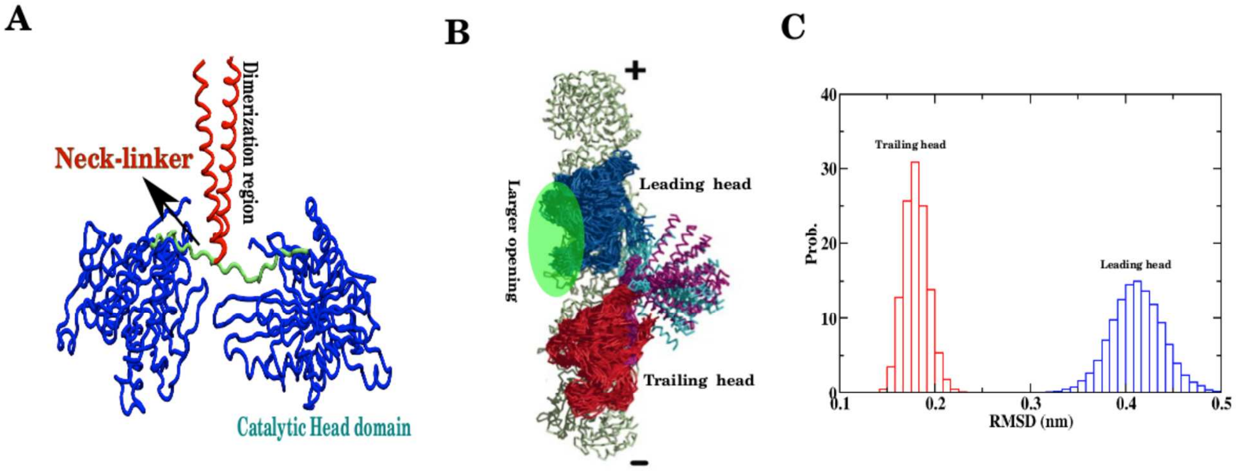

Figure 1: Asymmetric fluctuations in two kinesin-1 motor heads. (A) Different structural elements of the kinesin-1 motor homo-dimer. (B) Ensemble of superimposed structures of the two-head bound kinesin-1 motor. Note the higher opening in the NT binding site for the leading head. (C) Distribution of root mean square deviation (RMSD) of the two heads from the simulation. Note the higher RMSD of the leading head compared to trailing head.

Blending of multiple-basin SBM and perturbations to understand function

In the previous sections, we have shown how multiple basin SBM and a perturbation approach can be used to understand functions of different proteins/enzymes. Integrating these two approaches has also proved to be useful in understanding functionalities of certain proteins. This is the case when exploring mechanochemistry and directionality of a molecular motor called Ncd that has a –ve end MT directionally (kinesin-1 has +ve end directionality) and is a non-processive motor (19). Ncd motors are homo-dimers with two catalytic motor head domains and an extended dimerization region connected to the motor head (Figure 2B). Interestingly, the motor heads of both kinesin-1 and Ncd are very similar structurally and share conserved MT and ATP binding sites. Therefore, naturally a question arises; why these two motors have different directionality? To understand this functional difference, we first notice that there are structural dissimilarities between kinesin-1 and Ncd. The former has a neck-linker and the latter has a more extended dimerization domain (Figure 2B). Ncd has two conformations in solution without a MT: a symmetric dimer where both monomers have exactly the same structure and an asymmetric dimer where one of the monomers is rotated by ~180° around the neck-head junction (see Figure 2A). We constructed a dual-basin SBM for each of the heads in the dimer (19). A simulation using this Hamiltonian shows that the transition from symmetric to asymmetric dimer is correlated with the enhanced structural fluctuation near the NT binding domain (Figure 2C). Such a correlation implies that whenever a dimer goes to an asymmetric state, the rotated head can more efficiently dissociate ADP that is needed to eventually bind to the MT. In order to understand the population distribution of these two types of dimers in the MT bound state prior to ATP binding, we perturbed our SBM with MT binding site interactions. Since the structure of Ncd bound state is not available in the literature, we exploited the fact that the structure of motor head of kinesin-1 and Ncd are structurally similar and extracted the interaction pairs from a superposition of known structures. This perturbed Hamiltonian provides evidence that a symmetric state ensemble is stable on the MT and the tip of the coiled-coil dimerization region orients towards +ve end in this situation. ATP binding to the MT bound head is incorporated by changing a few contacts near the neck-head junction region (similar to kinesin-1). This perturbation makes a change in the population distribution with more stabilization towards its asymmetric state (this step is called powerstroke, see Figure 2D). In the asymmetric state, the coiled-coil dimerization region orients towards –ve end. In this entire cycle, the neck-helix region goes though an order-disorder transition depending on the NT state of the head. Finally, after ATP hydrolysis and phosphate release, the head with ADP detaches and the cycle restarts. It is also derived from the structural comparison that the absence of an unstructured neck-linker region in Ncd makes impossible to bind both heads simultaneously to the MT. Therefore Ncd motor is a non-processive motor.

In summary, using multiple basins and appropriate perturbations, we can elucidate complex motor actions. These results, for Ncd and kinesin-1, reveal that the kinesin family motors work using similar physical principles. Changing small structural elements (neck-linker for kinesin-1 and neck-helix for Ncd) peripheral to catalytic domain is sufficient to achieve a host of different functionalities.

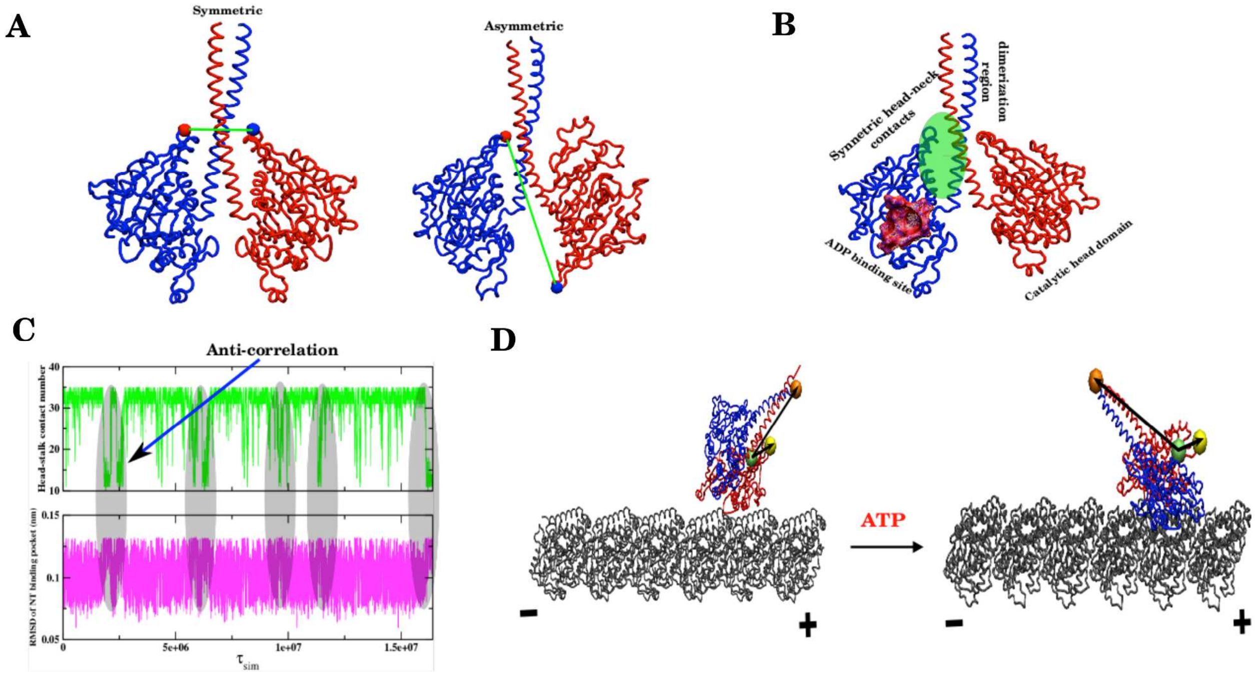

Figure 2: Enhanced fluctuations near the NT binding region upon symmetric to asymmetric transition in Ncd. (A) Symmetric and asymmetric dimers are in equilibrium without MT in solution. (B) Different structural elements of the Ncd motor dimer. The symmetric head-neck contact region and NT binding region of Ncd dimer. (C) The correlation between the head-neck contact and fluctuations near the NT binding region. The decrease in contact indicating the transition from symmetric to asymmetric state induces larger fluctuations. (D) The powerstroke of Ncd upon ATP binding.

Direct couplings in protein sequence data uncover residue contacts in protein structures

The SBM is a useful tool to explore the mechanisms of folding and function of proteins/enzymes. However, the method is restricted to the knowledge of experimental structures in order to obtain the key parameters for its energy Hamiltonian. This applies to both single and dual basin models. One of the most important parameters for the SBM is the set of residue pairs accounting for physical contacts between amino acid distant in the protein chain. Recently, this restriction has been alleviated with the introduction of better approximations to global statistical models of protein sequences (21, 22, 26, 45, 46).

One of such global formulations, known as DCA (21), models the joint probability distribution of amino acid occupancies as an exponential distribution that depends on pairwise residue energy couplings and single site terms or local fields. These terms are computed using statistics from a multiple sequence alignment (MSA) of protein families with an abundant number of sequences (47). The computed pairwise "direct" couplings help disentangle stronger couplings due to compensatory mutations required to maintain physical interactions (see Figure 3A) from other indirect correlations related to phylogeny or chain effects. As a result, amino acid pairs with high direct probabilities, assessed using the metric Direct Information (DI), tend to be good predictors of residue-residue physical interactions in the 3D fold of a protein that is member of a family with a large number of sequences. Figure 3B shows the performance in terms of the mean true positive (TP) rate of the *mean field* formulation of DCA in predicting residue-residue contacts. This analysis is done for several families and predictions come from more than 800 PDB structures. For more details on DCA please refer to Morcos *et al*. (21).

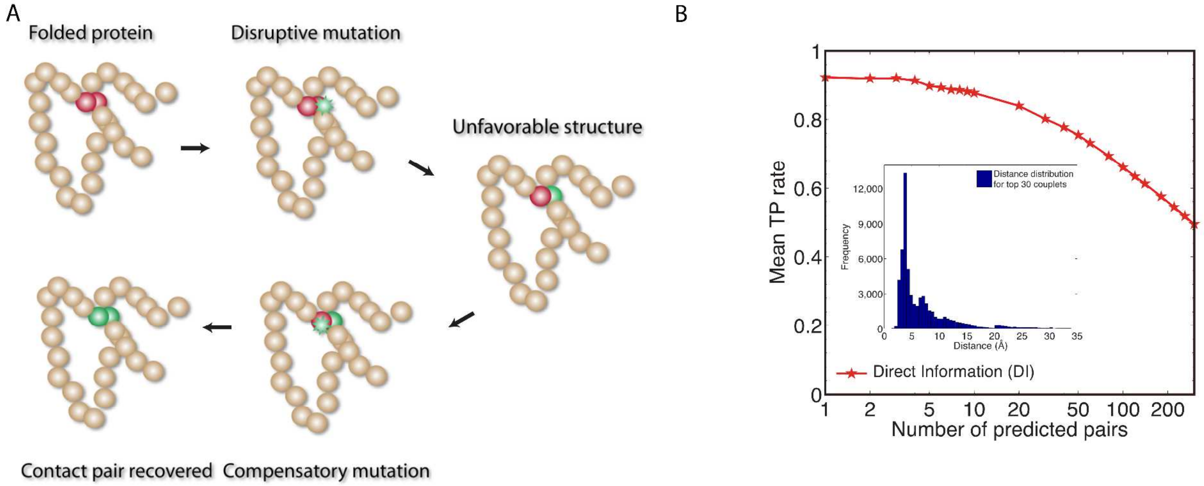

Figure 3: Direct couplings are good predictors of residue-residue couplings. (A) Compensatory mutations that recover stability and function in proteins of the same family are recorded in a collection of protein sequences in the form of MSA. (B) Using direct coupling analysis in MSA, the top couplings are a proxy of residue-residue contacts. The mean true positive rate as a function of the ranking of pairs shows that many contacts can be extracted from co-evolutionary information. The inset shows that most of the top 30 predicted pairs are separated by less than 8Å in the crystal structures.

DCA is not the only formulation capable of predicting contacts with high accuracy; there is an increasing number of methodologies that aim to improve contact prediction using statistical inference on MSA (See a review by De Juan & Valencia (48)). Conceptually all of them have the same goal of extracting residue interactions from abundant protein sequences using correlations imposed by co-evolution.

Co-evolving residue pairs give insights on protein structure prediction

The high fraction of correctly predicted contacts using DCA as well as their distribution across the complete contact map of several proteins provides enough constrains to fold proteins using a modification of a SBM. These co-evolutionary constraints complemented with knowledge-based potentials used to estimate local secondary structure and typical residue-residue distances are able to predict structures that are comparable with experimentally determined structures. Figure 4 shows two examples of proteins predicted using DCAfold (27). One is the TerR transcriptional regulator and the second is a Peptidoglycan-Associated lipoprotein PAL. These examples show the potential for high-resolution structure prediction when there is an accurate knowledge of local information. Both estimates have an RMSD less than or equal to 1.5Å. The tertiary fold determination is driven by the number of estimated DCA contacts that are shown as links in the figure.

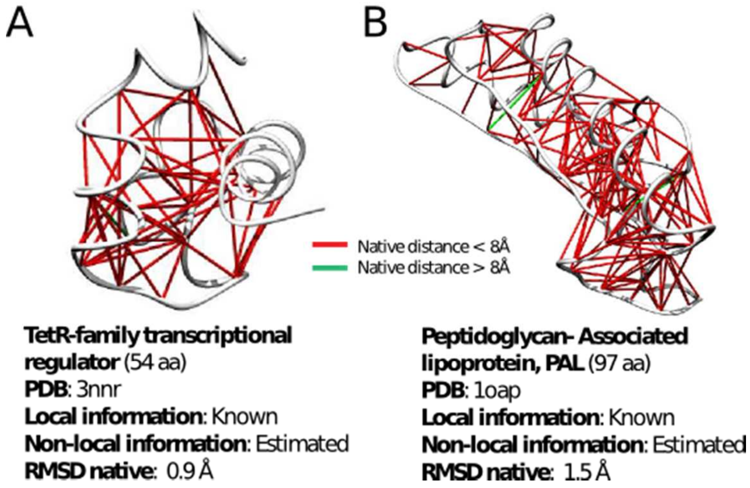
A
B
Native distance < 8Å
Native distance > 8Å
TetR-family transcriptional regulator (54 aa)
PDB: 3nnr
Local information: Known
Non-local information: Estimated
RMSD native: 0.9 Å
Peptidoglycan- Associated lipoprotein, PAL (97 aa)
PDB: 1oap
Local information: Known
Non-local information: Estimated
RMSD native: 1.5 Å

Figure 4: Two examples of protein structures predicted using DCAfold. (A) The TetR transcriptional regulator can be predicted to a resolution of 0.9Å if a very accurate estimate of the local structures is known. (B) The structure of the Peptidoglycan associated protein (97aa) is predicted with an RMSD of 1.5Å. These predictions are driven by long distance contact predictions (links) obtained using DCA. The red links illustrate predictions that are less than 8Å apart in the native structure. Green links are false positives above the 8Å threshold.

The key element in these methodologies is the ability to predict non-local contacts that are required in the tertiary structure of a given protein. Although it is still unclear if these co-evolutionary constrains can be used to investigate folding mechanisms, as it has been done previously with native contacts, their relevance in protein structure prediction is evident and has become a topic of vigorous research in the field of structural biology (49, 50). The combination of physical contacts estimated from sequence and the use of physical models like SBM has the potential to help us understand not only structural features of proteins but to provide insights of their function.

The functional conformational landscape of proteins can be explored with co-evolutionary signatures

As mentioned in the previous sections, when proteins perform their function, they generally explore a spectrum of conformations that deviates from the single native structure. When these alternative conformations form new residue-residue contacts then co-evolving interactions should also induce couplings in the collection of sequences belonging to a protein family. When analyzing direct couplings using DCA in a protein family with members known to have diverse functional conformations, we were able to show that highly ranked residue pairs not only are predictors of contacts in a given configuration but also are important contacts found in different functional states (21, 28). For instance, the L-leucine binding protein undergoes a conformational change upon ligand binding, forming new contacts between two of its subdomains. Contacts that overlap with both states are found as highly ranked couplings. However, a series of contacts unique to the closed state are also found by DCA. If we enrich a SBM of the open state with co-evolving constrains from DCA then a transition towards the closed state is induced. This structural transition is only possible due to these additional DCA contacts. Figure 5 shows how supplementing a SBM with DCA contacts for the L-leucine binding protein promotes transitions towards its ligand bound state. The centroid of the newly formed cluster has an RMSD of 2Å with respect to the experimentally determined structure (PDB 1usg/1usi).

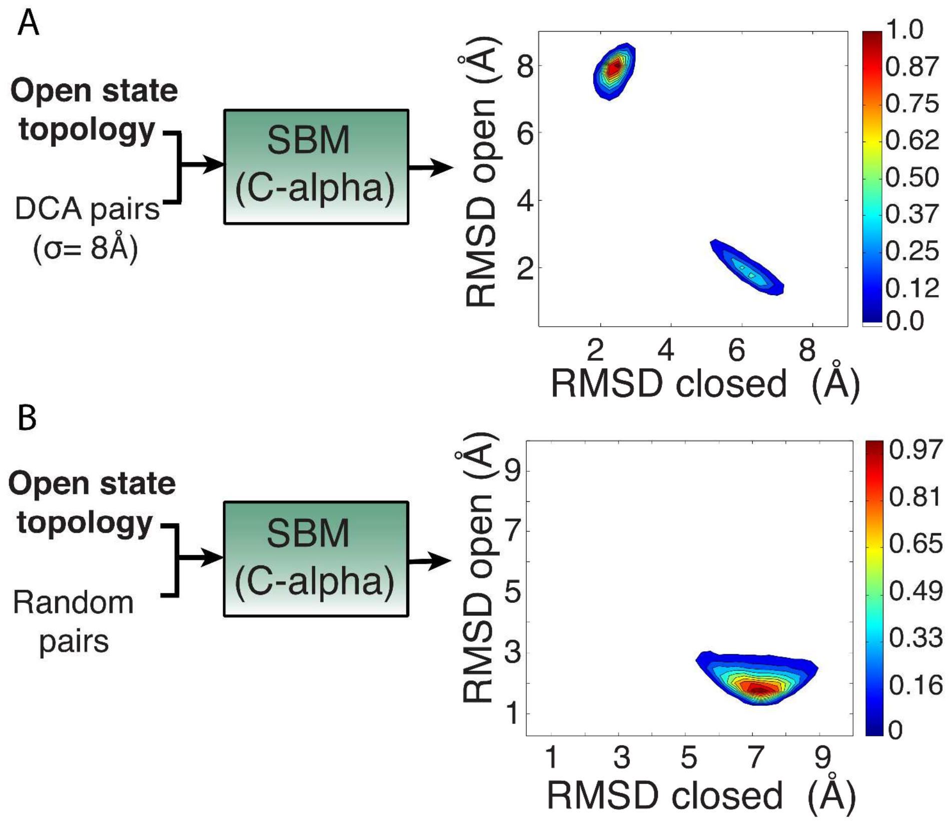

Figure 5: SBM+DCA explores the conformational landscape of proteins. (A) If a SBM with parameters of an open state topology is enriched with co-evolutionary couplings from DCA, then the molecular dynamics simulation visits both open and closed states. The centroid of the predicted closed state distribution is very close (~2Å) to the experimental structure of the closed state. (B) A control simulation with the same number of constrains but randomly distributed produces only a fluctuation of the open state conformation.

This combination of DCA+SBM uncovers not only holo states for ligand binding proteins, but it is also able to uncover intermediates or hidden states that are also important for function. Notable examples are the GluR2 glutamate binding receptor for which a partial agonist was found and the D-ribose binding protein that has a putative twisted state hypothesized to promote faster ligand release (28).

The Case of FtsH AAA protease

We have compared models that have structures for all functional conformations with a model that used only one structure and was complemented with coevolutionary information. Before this study, previous systems have been restricted mostly to monomeric single domain proteins. We now investigate larger systems where multidomain proteins interact in a macro molecular complex. A particularly interesting example is the FtsH AAA peptidase. This motor acts as a complex machine that degrades proteins into small peptides (51, 52). This macro complex forms a ring with six monomeric units. Each of the monomers consists of two domains: an AAA domain, which hydrolyses ATP and a peptidase domain that hydrolyses peptide bonds to degrade proteins into small peptides. The outer ring consists of six AAA domains that grab the peptide chain and then translocates it down towards the inner peptidase chain ring with the help of ATP hydrolysis. As it is evident from this mechanism, the ring goes through a large-scale conformational change during the translocation process. We show how we can use available crystal structures of the hexameric ring and direct coupling analysis to understand the mechanochemical cycle of this large macro molecular machine. This demonstrates the power of the integration of genomic information with structural models of proteins.

Experimental structures and DCA inter-domain residue pair predictions

There are two experimental structures available for FtsH AAA motors. One is a conformation with no NT in the AAA domain (PDB ID 3kds) (53). In this conformation, both the outer ring radius and the distance between the outer and inner ring are large. We refer to this state as the open state. This state has a C_6 symmetry and each monomer has the same conformation. Another available crystal structure (PDB ID 2cea) has six ADP in its six AAA domains (54). In this conformation both the outer ring radius and the distance between the outer and inner rings are smaller. We refer to this state as the closed state. This state has a C_2 symmetry indicating the presence of three types of monomeric conformations. In Figure 6A, the residue contact maps for all three monomeric conformations in the closed state are compared with the open state monomeric conformation. In all these four monomeric conformations (1 open and 3 closed), the contact maps are similar for individual domains and changes occur mainly in the inter-domain regions. Therefore, we have shown in Figure 6B, the super-imposed inter-domain contact map for three closed conformations. In the right panel of the figure 6B, we have also shown the DCA predicted highly ranked inter-domain contact map. The region between amino acids 380 and 120 is found to be highly coupled by DCA. This set of inter-domain contacts is also present in the native inter-domain contact maps. We have observed that these interactions are responsible for bringing down the AAA domain towards the peptidase domain. While not all inter-domain interactions are predicted by DCA, they are sufficient to induce the closing transition. It should be noted here that the complete set of residue interactions of the ring also includes important variations in the interfacial contacts between monomers.

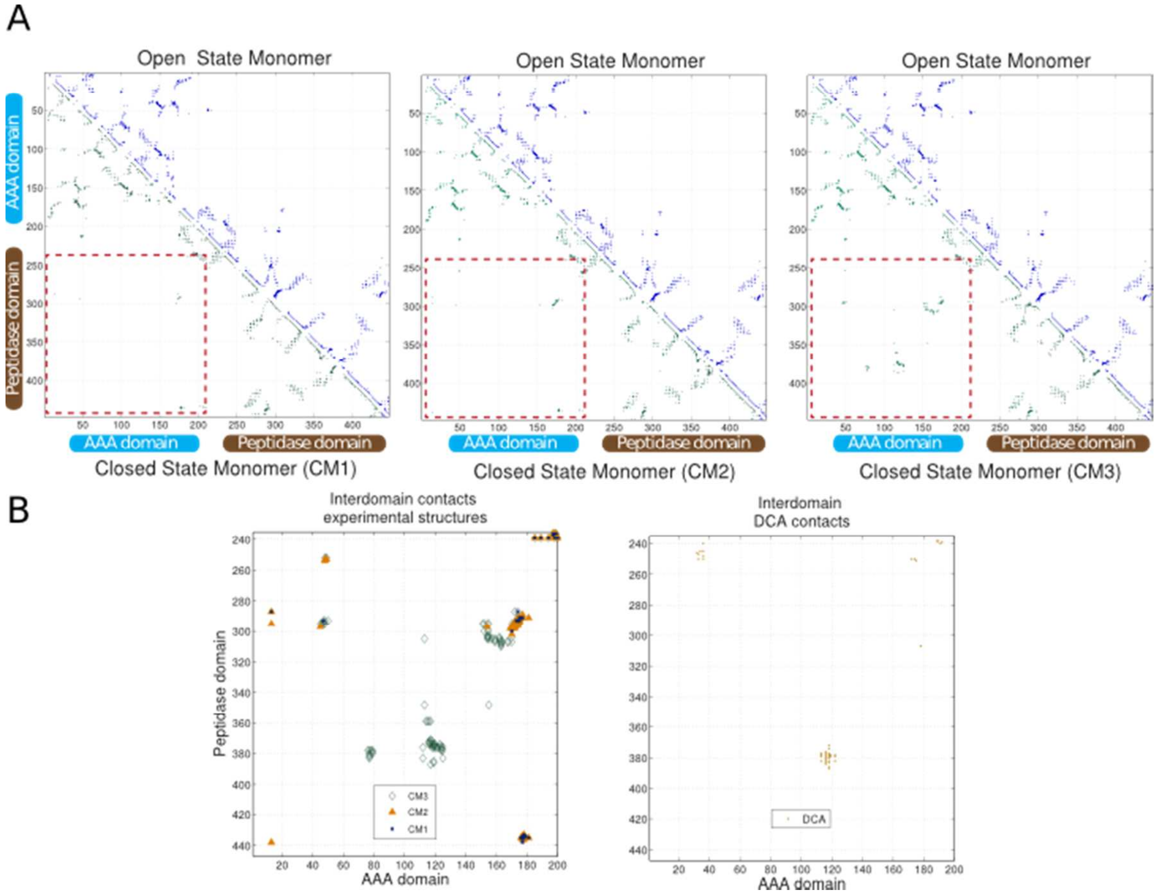

Figure 6: The residue contact map of the AAA-peptidase interdomain region is predicted by DCA. (A) Residue contact maps of three kinds of monomeric conformation in the C_2 symmetric closed state of AAA motors are compared with open state monomer conformations. The interdomain regions are marked with red dashed boxes. (B) Superimposed interdomain contact map from three conformations are compared with the DCA predicted

interdomain region. Note that DCA has predicted the region around 120 and 380 with high couplings. These contacts appear when the AAA domain and the peptidase domain form a closed structure.

Understanding the FtsH functional mechanism using a dual basin SBM

We constructed a SBM of the hexameric ring using the open and closed state crystal structures. This is a dual basin SBM. We have derived the local part of the Hamiltonian (bond, angle and dihedral) from the open state. It is important to note that if the local information is chosen from the closed state, the qualitative mechanistic aspects of the catalytic cycle remains unchanged. The contacts have been divided into three parts: contacts exclusive to the open state, contacts exclusive to the closed state and shared contacts. We used these shared contacts to stabilize a particular state over the other one. Within this strategy, we have observed both open to close and close to open transitions. In the following section, we discuss the nature of these transitions.

Emergence of a paddling mechanism

To understand the mechanism of closing, we have simulated the Hamiltonian with shared contacts from the closed state starting from the open state. To measure the progression, we have defined two order parameters: the fraction of the outer pore radius opening and the fraction of the opening distance between the outer and inner rings (See Figure 7A for all the parameters). We find that in the initial phase, the pore radius of the outer ring decreases while keeping the distance between the two rings still high (see closing intermediates in Figure 7B). In this stage the pore radius still fluctuates. In the next stage, the distance between the two rings decreases to its closed state along with the smaller pore radius leading to a stable closed state configuration. To model the opening mechanism, we derived a Hamiltonian with shared contacts from the open state and simulate the Hamiltonian starting from the closed state. Initially, the pore radius increases to a greater extent than the ring separation distance (see opening intermediates in Figure 7B). Then both pore radius and ring separation distance achieve the final open state values. Such a process has been proposed before and is called the “paddling mechanism” (55). This mechanism seems essential for the function of this biomachine. In the closing process, first the decrease in pore radius could promote the peptide chain to bind the ring. In the next step, this bound peptide segment could be brought down towards the peptidase ring for the hydrolysis step. During opening, an increase in pore size could induce peptide release before going to a higher ring separation (Figure 7). In the final step, returning to a completely open state is needed for the continuation of the catalytic cycle. The representative intermediate from the open-to-closed transition (centroid of points labeled “closing” in Figure 7B) has an RMSD of 5.3Å with respect to the open state and an RMSD of 15.2 Å with respect to the closed state. Conversely, the representative intermediate (centroid of points labeled “opening” in Figure 7B) found in the closed-to-open transition has an RMSD of 5Å compared to the closed state and an RMSD of 14.5Å compared to the open state.

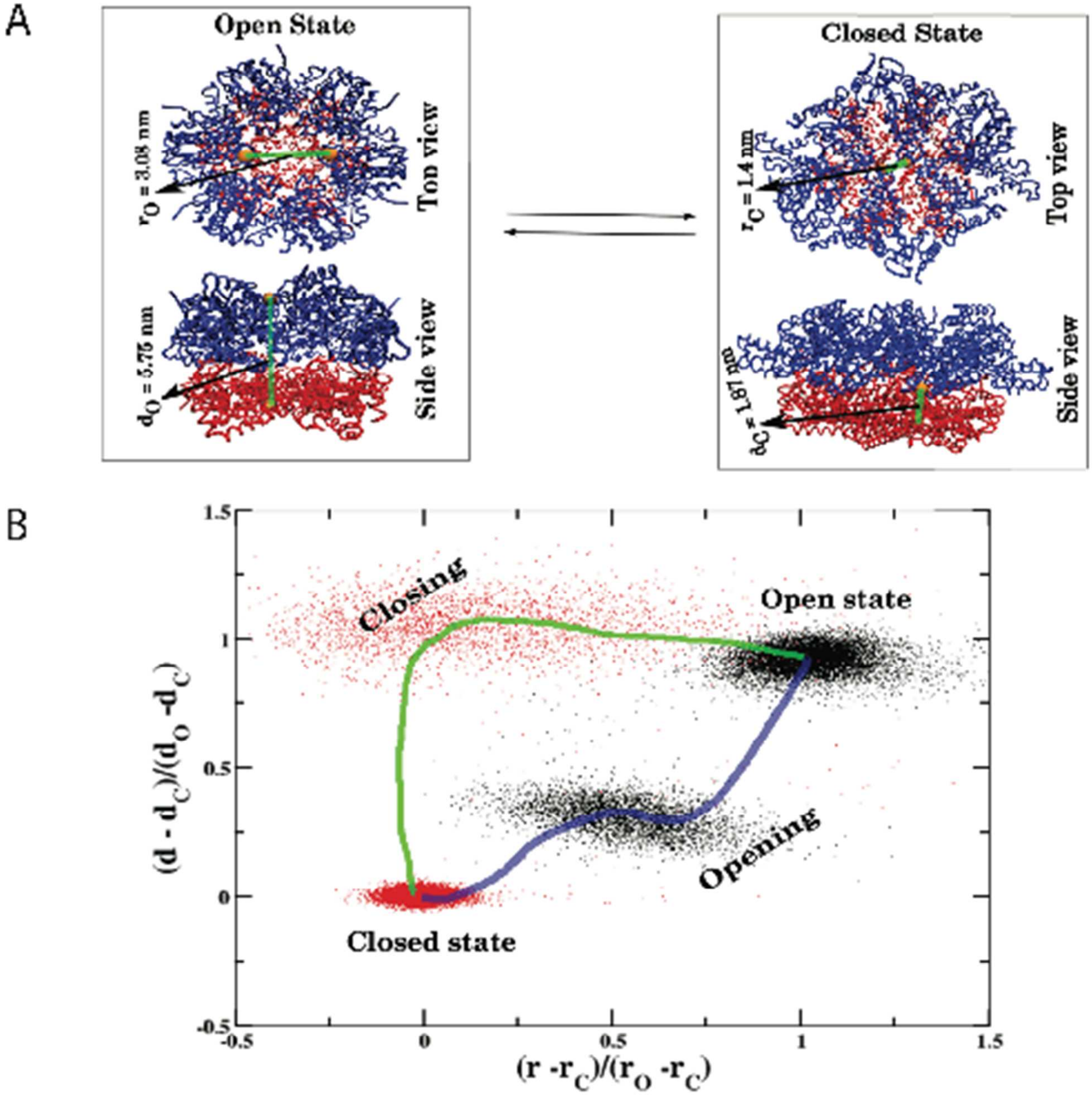

Figure 7: Putative paddling mechanism of AAA motor function. (A) Top and side views of AAA motor in the open and closed states. The pore radius (r) and ring separation distance (d) in both the conformations are also shown. These parameters are used to build our reaction coordinates. (B) Closing and opening mechanisms are shown in terms of the two reaction coordinates. Data points from 14 different trajectories have been aggregated into this plot. Note that in both cases, pore radius changes first followed by the ring separation distance providing further evidence to the paddling mechanism hypothesis.

Discussion

In this work we have described the evolution of the use of SBM starting from protein folding and moving towards functional mechanisms of proteins. A similar progression is observed with the use co-evolutionary couplings inferred from protein families. Initially those couplings were found to be good predictors of residue-residue interactions, then those predictions were utilized to accurately estimate protein structure (27). More recently, directly coupled residues have been merged with SBM to better explore the functional landscape of proteins (28). A new successful example was presented for the case of the FtsH integral membrane protease. A predicted inter-domain contact map (Figure 6) of the AAA domain (PF00004) and the peptidase M41 domain (PF01434) combining only the open structure and DCA contacts is sufficient to bring together these two domains and close the hexameric pore. Alternatively, a dual basin SBM provides evidence of a paddling mechanism hypothesized to be responsible of peptide trapping, translocation and release. A molecular dynamics trajectory shows how the ring first closes its pore and subsequently reduces the separation between the upper AAA hexameric ring and the lower peptidase ring. The reverse pathway promotes first a distance separation between the rings and then an opening of the pore. This series of steps agree well with the paddling hypothesis of the FtsH peptidase (55).

Having both structures, however, is not necessary and in many cases not available. It is possible to investigate the functional landscape of this macro complex, by combining SBM with co-evolutionary constrains from DCA. We can build a SBM with only the open state crystal structure and incorporate coupled DCA pairs by considering the fact that these interactions can also extend across different monomers. A rationale for such assumption is the fact that the peptidase domain of one monomer gets in close proximity with the AAA domain of the next monomer in the ring and DCA predictions are not restricted to monomeric interactions. We hypothesize that DCA predictions can lead us to the functional closed state conformation of such a big complex. As the combined SBM+DCA method seems to sample structural conformations constrained by evolution, we believe that this methodology predict novel structural states that are relevant for function. This is the topic of current investigations in our group. The problem of limited structural information can be alleviated by the imminent increase of sequence data and the use of SBM to study the function of very large macro-complexes. This opens the door to study systems that until recently were barely accessible to molecular dynamics research but with the potential of exploring much longer time scales needed in many cases to observe functional transitions.

Methods

For the study of the AAA motor, we have used a SBM in the SMOG (33) protocol implemented in GROMACS in combination with DCA. SBMs are derived combining two structural forms of FtsH AAA, open (PDB ID 3kds) and closed states (PDB ID 2cea). The direct coupling analysis was performed on a multiple sequence alignment to identify important residue contacts for structural transition.

Structure based models of the FtsH protease

We built our SBM from native structures (open and closed states) by placing a single bead of unit mass for each amino acid at the location of the Cα atom. The energy function used for this model is given below,

$$H_{SBM}(\{\vec{r}_i\}) = H_b^O + H_{nb}^O(uniq) + H_{nb}^C(uniq) + H_{nb}\left(shared\right) \tag{1}$$

Here, the superscripts O and C refer to the open and closed states, respectively. H_b^O represents the local bonded component of the Hamiltonian,

$$H_b^O = \sum_{i=1}^{N-1} \frac{K_r}{2}(r_{i,i+1} - r_{i,i+1}^{0(O)})^2 + \sum_{i=1}^{N-2} \frac{K_\theta}{2}(\theta_i - \theta_i^{0(O)})^2 + \sum_{i=1}^{N-3} \sum_{n=1,3} K_\phi^{(n)}(1-\cos[n(\phi_i - \phi_i^{0(O)}])) \tag{2}$$

The first term in H_b^O ensures that the bond distance $r_{i,i+1}$ between the neighboring residues *i* and *i+1* to be contained harmonically with respect to its native bond distance $r_{i,i+1}^{0(O)}$ by a spring constant $K_r = 20(kJ / mol.A^2)$. The second term constrains the angle θ_i among the residues *i*, *i+1*, and *i+2* with respect to its native value $\theta_i^{0(O)}$ by a harmonic spring constant $K_\theta = 20(kJ / mol.rad^2)$. The third term represents the dihedral angle potential with $K_\phi^{(1)} = 2K_\phi^{(3)}$ that describes the rotation of the backbone involving successive residues from *i* to *i+3*. The native values $r_{i,i+1}^{0(O)}$, $\theta_i^{0(O)}$ and $\phi_i^{0(O)}$ are taken from the open conformation crystal structure. The non-local terms of the Hamiltonian ($H_{nb}^O(uniq)$, $H_{nb}^C(uniq)$ and $H_{nb}(shared)$) have the following general form:

$$H_{nb} = \sum_{i=1}^{N-4} \sum_{j=i+4}^{N} \left[\varepsilon \left(6\left(\frac{r_{ij}^0}{r_{ij}}\right)^{12} - 5\left(\frac{r_{ij}^0}{r_{ij}}\right)^{10} \right) \Delta_{ij} + \varepsilon_r \left(\frac{\sigma}{r_{ij}}\right)^{12} (1-\Delta_{ij}) \right] \tag{3}$$

The 10-12 Lennard-Jones (LJ) potential is used in H_{nb} to describe the interactions that stabilize the non-bonded native contacts. Native contact pairs (*i* and *j*) are obtained using the shadow contact map that is implemented in SMOG. If *i* and *j* residues are in contact in the native state, $\Delta_{ij} = 1$; otherwise $\Delta_{ij} = 0$. Native contact pair distances $r_{i,j}^{0}$ are obtained from the native state. Non-native pairs with $\Delta_{ij}^{O} = 0$ are under repulsive potential with a distance parameter $\sigma = 4$Å. $H_{nb}^{O}(uniq)$ and $H_{nb}^{C}(uniq)$ represent only the unique contacts present in the open and closed states, respectively. $H_{nb}(shared)$ is the Hamiltonian that describes the shared contact pairs present in both the crystal structures. To simulate open to close transition, we derived $H_{nb}(shared)$ from the closed state, which ensures the overall stability of the closed state. To simulate the closed to open transition, we derived $H_{nb}(shared)$ from the open state. We have also modulated the strength of the contacts from both states for these two kinds of simulations. For open to close transition, ε (closed) = 0.8 KJ/mole and ε (open)=0.7 KJ/mole and for close to open transition, ε (closed) = 0.7 KJ/mole and ε (open)=0.8 KJ/mole. We have used a constant value for ε_r =0.75 KJ/mole.

Simulations

Initial structures in each case were relaxed under the SBM Hamiltonian and subsequently Langevin dynamics simulations at low-friction limit were performed at T = 90K to simulate the kinetic pathway of the transition. The equation of motion for the Langevin dynamics used for integration is

$$m\ddot{\vec{r}}_i = -\zeta\dot{\vec{r}}_i - \partial_{\vec{r}} H\left(\left\{\vec{r}_i\right\}\right) + \vec{\Gamma}_i\left(t\right) \qquad (4)$$

where ζ is the friction coefficient, $-\partial_{\vec{r}} H(\{\vec{r}_i\})$ is the conformational force. $\vec{\Gamma}_i(t)$ is the random force satisfying $\left\langle \vec{\Gamma}_i(t)\cdot\vec{\Gamma}_j(t')\right\rangle = (6\zeta k_B T / h)\delta_{ij}(t-t')$ where integration time h is discretized. In this dynamics we chose ζ = 0.05 τ_L^{-1} and h = 0.0025 τ_L with $\tau_L = (m\sigma^2/\varepsilon_r)^{1/2}$.

Direct Coupling Analysis

To identify direct correlations related to physical contacts, we used a statistical framework termed Direct Coupling Analysis (DCA) and specifically, an efficient formulation called mean field DCA (mfDCA) (21) that allows us to work with large number of sequences and protein lengths. In this framework, frequency counts for individual sites and couplets in multiple sequence alignments are defined as single and pairwise probabilities. An application of the maximum entropy principle gives rise to the following model of the joint probability function of amino acid composition of a complete protein family:

$$P(A_1,\square\ ,A_L) = \frac{1}{Z}\exp\left\{\sum\nolimits_{i<j} e_{ij}(A_i,A_j) + \sum\nolimits_i h_i(A_i)\right\} \qquad (5)$$

where $e_{ij}(A,B)$ are the pairwise couplings and $h_i(A)$ are the single site biases. In order to obtain accurate estimates of direct coupling among residue pairs, it is necessary to compute the pairwise couplings. After a small-coupling expansion explained in detail in (21), the pairwise couplings have the following form:

$$e_{ij}(A,B) = -(C^{-1})_{ij}(A,B) \qquad (6)$$

where C is the connected correlation matrix and depends only on single and pairwise frequency counts: $C_{ij}(A,B) = f_{ij}(A,B) - f_i(A)f_j(B)$. Finally, a metric to rank such direct couplings, termed Direct Information (DI) can be computed as follows:

$$DI_{ij} = \sum\nolimits_{AB} P_{ij}^{(dir)}(A,B) \ln \frac{P_{ij}^{(dir)}(A,B)}{f_i(A) f_j(B)} \qquad (7)$$

The direct probabilities of pairs only depend on the couplings and the local fields that are computed with the frequency counts. We can use these probabilities to rank residue pairs according to their potential of being in contact in its 3-D fold.

Multiple Sequence Alignments

In order to compute couplings to study the FtsH system we used MSA of domain families for which FtsH is a member. We obtained sequence data from domains (PF00004) and the peptidase M41 domain (PF01434) stored in the Pfam database (47). The number of sequences is variable for both domains, however we focused on sequences having the same domain architecture as FtsH. The goal of this is to be able to infer interdomain coevolving residues rather than intradomain contacts for which structural information is already available. The final dataset consisted of 4389 protein sequences with both AAA and peptidase M41 domains in the same protein.

Acknowledgments. This work was supported by the Center for Theoretical Biological Physics sponsored by the Welch Foundation Grant C-1792 and by the National Science Foundation (NSF) (Grants PHY-1308264 and MCB-1214457). J.N.O. is a CPRIT Scholar in Cancer Research sponsored by the Cancer Prevention and Research Institute of Texas.

References

1. Anfinsen CB (1973) Principles That Govern Folding of Protein Chains. *Science* 181(4096):223-230.
2. Levinthal C (1968) Are There Pathways for Protein Folding. *J Chim Phys Pcb* 65(1):44-&.
3. Zwanzig R, Szabo A, & Bagchi B (1992) Levinthals Paradox. *Proceedings of the National Academy of Sciences of the United States of America* 89(1):20-22.
4. Bryngelson JD & Wolynes PG (1987) Spin-Glasses and the Statistical-Mechanics of Protein Folding. *Proceedings of the National Academy of Sciences of the United States of America* 84(21):7524-7528.
5. Leopold PE, Montal M, & Onuchic JN (1992) Protein folding funnels: a kinetic approach to the sequence-structure relationship. *Proceedings of the National Academy of Sciences of the United States of America* 89:8721-8725.
6. Onuchic JN & Wolynes PG (2004) Theory of protein folding. *Current Opinion in Structural Biology* 14(1):70-75.
7. Dill KA & MacCallum JL (2012) The Protein-Folding Problem, 50 Years On. *Science* 338(6110):1042-1046.
8. Dill KA & Chan HS (1997) From Levinthal to pathways to funnels. *Nat Struct Biol* 4(1):10-19.
9. Baker D (2000) A surprising simplicity to protein folding. *Nature* 405(6782):39-42.
10. Fleishman SJ & Baker D (2012) Role of the biomolecular energy gap in protein design, structure, and evolution. *Cell* 149(2):262-273.
11. Veitshans T, Klimov D, & Thirumalai D (1997) Protein folding kinetics: timescales, pathways and energy landscapes in terms of sequence-dependent properties. *Fold Des* 2(1):1-22.
12. Straub JE & Thirumalai D (1993) Exploring the Energy Landscape in Proteins. *Proceedings of the National Academy of Sciences of the United States of America* 90(3):809-813.
13. Bryngelson JD, Onuchic JN, Socci ND, & Wolynes PG (1995) Funnels, Pathways, and the Energy Landscape of Protein-Folding - a Synthesis. *Proteins-Structure Function and Genetics* 21(3):167-195.
14. Socci ND, Onuchic JN, & Wolynes PG (1996) Diffusive dynamics of the reaction coordinate for protein folding funnels. *J Chem Phys* 104(15):5860-5868.
15. Clementi C, Nymeyer H, & Onuchic JN (2000) Topological and energetic factors: what determines the structural details of the transition state ensemble and "en-route" intermediates for protein folding? An investigation for small globular proteins. *J Mol Bio* 298:937-953.
16. Whitford PC*, et al.* (2009) An all-atom structure-based potential for proteins: Bridging minimal models with all-atom empirical forcefields. *Proteins-Structure Function and Bioinformatics* 75(2):430-441.

17. Okazaki K, Koga N, Takada S, Onuchic JN, & Wolynes PG (2006) Multiple-basin energy landscapes for large-amplitude conformational motions of proteins: Structure-based molecular dynamics simulations. *Proc Natl Acad Sci U S A* 103(32):11844-11849.
18. Whitford PC, Gosavi S, & Onuchic JN (2008) Conformational transitions in adenylate kinase - Allosteric communication reduces misligation. *Journal of Biological Chemistry* 283(4):2042-2048.
19. Jana B, Hyeon C, & Onuchic JN (2012) The origin of minus-end directionality and mechanochemistry of Ncd motors. *PLoS Comput Biol* 8(11):e1002783.
20. Hyeon C & Onuchic JN (2007) Internal strain regulates the nucleotide binding site of the kinesin leading head. *Proceedings of the National Academy of Sciences of the United States of America* 104(7):2175-2180.
21. Morcos F*, et al.* (2011) Direct-coupling analysis of residue coevolution captures native contacts across many protein families. *Proc Natl Acad Sci USA* 108(49):E1293-E1301.
22. Taylor WR & Sadowski MI (2011) Structural constraints on the covariance matrix derived from multiple aligned protein sequences. *PLoS ONE* 6(12):e28265.
23. Göbel U, Sander C, Schneider R, & Valencia A (1994) Correlated mutations and residue contacts in proteins. *Proteins Struct Funct Genet* 18:309-317.
24. Altschuh D, Lesk A, Bloomer A, & Klug A (1987) Correlation of co-ordinated amino acid substitutions with function in viruses related to tobacco mosaic virus. *J Mol Bio* 193:693-707.
25. Lockless SW & Ranganathan R (1999) Evolutionarily conserved pathways of energetic connectivity in protein families. *Science* 286:295-299.
26. Jones DT, Buchan DW, Cozzetto D, & Pontil M (2012) PSICOV: precise structural contact prediction using sparse inverse covariance estimation on large multiple sequence alignments. *Bioinformatics* 28(2):184-190.
27. Sulkowska JI, Morcos F, Weigt M, Hwa T, & Onuchic JN (2012) Genomics-aided structure prediction. *Proc Natl Acad Sci U S A* 109(26):10340-10345.
28. Morcos F, Jana B, Hwa T, & Onuchic JN (2013) Coevolutionary signals across protein lineages help capture multiple protein conformations. *Proc Natl Acad Sci U S A* Online Early Edition.
29. Onuchic JN & Wolynes PG (2004) Theory of protein folding. *Curr Opin Struc Biol* 14:70-75.
30. Frauenfelder H, Sligar SG, & Wolynes PG (1991) The Energy Landscapes and Motions of Proteins. *Science* 254(5038):1598-1603.
31. Chavez LL, Onuchic JN, & Clementi C (2004) Quantifying the roughness on the free energy landscape: Entropic bottlenecks and protein folding rates. *Journal of the American Chemical Society* 126(27):8426-8432.
32. Davtyan A*, et al.* (2012) AWSEM-MD: protein structure prediction using coarse-grained physical potentials and bioinformatically based local structure biasing. *The journal of physical chemistry. B* 116(29):8494-8503.
33. Noel JK, Whitford PC, Sanbonmatsu KY, & Onuchic JN (2010) SMOG@ctbp: simplified deployment of structure-based models in GROMACS. *Nucleic Acids Res* 38:W657-661.
34. Fersht AR (1995) Characterizing Transition-States in Protein-Folding - an Essential Step in the Puzzle. *Current Opinion in Structural Biology* 5(1):79-84.
35. Noel JK, Sulkowska JI, & Onuchic JN (2010) Slipknotting upon native-like loop formation in a trefoil knot protein. *Proceedings of the National Academy of Sciences of the United States of America* 107(35):15403-15408.
36. Levy Y, Cho SS, Onuchic JN, & Wolynes PG (2005) A survey of flexible protein binding mechanisms and their transition states using native topology based energy landscapes. *Journal of Molecular Biology* 346(4):1121-1145.
37. Schug A & Onuchic J (2007) Symmetric mutations and their asymmetric effect on a dual-funneled energy landscape: Modelling the Rop-dimer. *Biophys J*:216A-216A.

38. Levy Y, Wolynes PG, & Onuchic JN (2004) Protein topology determines binding mechanism. *Proceedings of the National Academy of Sciences of the United States of America* 101(2):511-516.

39. Dill KA & Fersht AR (1996) Folding and binding - Editorial overview. *Current Opinion in Structural Biology* 6(1):1-2.

40. Levy Y, Cho SS, Shen T, Onuchic JN, & Wolynes PG (2005) Symmetry and frustration in protein energy landscapes: A near degeneracy resolves the Rop dimer-folding mystery. *Proceedings of the National Academy of Sciences of the United States of America* 102(7):2373-2378.

41. Noel JK*, et al.* (2012) Mirror Images as Naturally Competing Conformations in Protein Folding. *Journal of Physical Chemistry B* 116(23):6880-6888.

42. Whitford PC, Miyashita O, Levy Y, & Onuchic JN (2007) Conformational transitions of adenylate kinase: Switching by cracking. *Journal of Molecular Biology* 366(5):1661-1671.

43. Hyeon C, Jennings PA, Adams JA, & Onuchic JN (2009) Ligand-induced global transitions in the catalytic domain of protein kinase A. *Proceedings of the National Academy of Sciences of the United States of America* 106(9):3023-3028.

44. Hyeon C & Onuchic JN (2007) Mechanical control of the directional stepping dynamics of the kinesin motor. *Proceedings of the National Academy of Sciences of the United States of America* 104(44):17382-17387.

45. Kamisetty H, Ovchinnikov S, & Baker D (2013) Assessing the utility of coevolution-based residue-residue contact predictions in a sequence- and structure-rich era. *Proceedings of the National Academy of Sciences of the United States of America* 110(39):15674-15679.

46. Ekeberg M, Lovkvist C, Lan YH, Weigt M, & Aurell E (2013) Improved contact prediction in proteins: Using pseudolikelihoods to infer Potts models. *Phys Rev E* 87(1).

47. Finn RD*, et al.* (2010) The Pfam protein families database. *Nucleic Acids Research* 38:D211-D222.

48. de Juan D, Pazos F, & Valencia A (2013) Emerging methods in protein co-evolution. *Nature reviews. Genetics* 14(4):249-261.

49. Marks DS*, et al.* (2011) Protein 3D Structure Computed from Evolutionary Sequence Variation. *PLoS ONE* 6:e28766.

50. Marks DS, Hopf TA, & Sander C (2012) Protein structure prediction from sequence variation. *Nat Biotechnol* 30(11):1072-1080.

51. Langklotz S, Baumann U, & Narberhaus F (2012) Structure and function of the bacterial AAA protease FtsH. *Bba-Mol Cell Res* 1823(1):40-48.

52. Yamada-Inagawa T, Okuno T, Karata K, Yamanaka K, & Ogura T (2003) Conserved pore residues in the AAA protease FtsH are important for proteolysis and its coupling to ATP hydrolysis. *Journal of Biological Chemistry* 278(50):50182-50187.

53. Bieniossek C, Niederhauser B, & Baumann UM (2009) The crystal structure of apo-FtsH reveals domain movements necessary for substrate unfolding and translocation. *Proceedings of the National Academy of Sciences of the United States of America* 106(51):21579-21584.

54. Bieniossek C*, et al.* (2006) The molecular architecture of the metalloprotease FtsH. *Proc Natl Acad Sci U S A* 103(9):3066-3071.

55. Koga N, Kameda T, Okazaki K, & Takada S (2009) Paddling mechanism for the substrate translocation by AAA plus motor revealed by multiscale molecular simulations. *Proceedings of the National Academy of Sciences of the United States of America* 106(43):18237-18242.