Structured and Unstructured binding of an Intrinsically Disordered Protein as revealed by atomistic simulations

Raúl Esteban Ithuralde,[†] Adrián Enrique Roitberg,[‡] and Adrián Gustavo Turjanski[*,†]

[†]*Departamento de Química Biológica/Departamento de Química Inorgánica, Analítica y Química Física, Facultad de Ciencias Exactas y Naturales, Universidad de Buenos Aires & INQUIMAE-UBA/CONICET, Intendente Güiraldes 2160, Pabellón II, Ciudad Universitaria, ciudad de Buenos Aires, Argentina, C1428EGA*

[‡]*Department of Chemistry. University of Florida. PO Box 117200. Gainesville, FL 32611-7200*

E-mail: adrian@qi.fcen.uba.ar

Phone: +54 11 45763380 ext 123. Fax: +54 11 45763380 ext 123

Abstract

Intrinsically Disordered Proteins(IDPs) are a set of proteins that lack a definite secondary structure in solution. IDPs can acquire tertiary structure when bound to their partners, and therefore the recognition process must also involve protein folding. The nature of the transition state(TS), structured or unstructured, determines the binding mechanism. The characterization of the TS has become a major challenge for experimental techniques and molecular simulations approaches since diffusion, recognition and binding is coupled to folding. In this work we present atomistic Molecular

Dynamics(MD) simulations that sample the free energy surface of the coupled folding and binding of the transcription factor c-myb to the co-transcription factor CREB Binding Protein(CBP). This process has been recently studied and became a model to study IDPs. Despite the plethora of available information we still do not know how c-myb binds to CBP. We performed a set of atomistic biased MD simulations running a total of 15.6μs. Our results show that c-myb folds very fast upon binding to CBP with no unique pathway for binding. The process can proceed through both structured or unstructured TS with similar probabilities. This finding reconciles previous, seemingly different, experimental results. We also performed G$\bar{o}$-Type coarse grained MD of several structured and unstructured models that indicate that coupled folding and binding follows a native contact mechanism. To the best of our knowledge this is the first atomistic MD simulation that samples the free energy surface of the coupled folding and binding processes of IDPs.

Introduction

In contrast to folded proteins, intrinsically disordered proteins (IDPs) and IDP regions do not form a stable structure in solution but they nevertheless exhibit biological activity.[1–5] We now know that many of these proteins are involved in protein-protein and protein-nucleic acid interactions and can structure when bound to their partners. The structure of several of these complexes has been solved in the last few years.[6–9] Due to their intrinsic flexibility, it is a difficult task to study the mechanism that IDPs follow in order to recognize their targets and acquire a folded structure.[10] Two limiting mechanisms have been proposed.[11–13] One is conformational selection, involving an structured transition state(TS) and binding can only occur if contacts are made between a previously pre-ordered molecule and its binding partner. The other mechanism is known as induced folding where the TS is unstructured, so the protein recognizes its partner in a disordered state, and only then folds over the surface of the other protein.[12,14] The ability to study the processes of coupled folding-and-binding either by experiments or simulations requires following the binding in time and analyzing the conformational changes that occur along the path. Recently, several experiments have been published that offer insights into the structure of the Transition State and Transition Paths.[8,15,16] Those studies were done by using stopped flow techniques with fluorescent probes, by performing mutational studies, single molecule studies and also with NMR dispersion. However, due to the complexity of these pathways, there is no technique that can follow the binding pathways at the atomic level. Molecular Dynamics(MD) simulations allow a detailed description of protein folding and binding dynamics but the large number of degrees of freedom involved in the process means that usually coarse grained G$\bar{o}$-type potentials are used. Recently, Shaw's group conducted long atomistic MD that sampled the folding of several protein domains[17] by using the Anton machine. However, the application of atomistic MD to study the folding and binding process of IDPs is even more complicated than single domains as we need to sample protein-protein interactions. So the large number of degrees of freedom involved preclude (at this pointin time) the use of long equilibrium simulations.

Association rates of IDPs vary greatly, but it is not yet clear that they bind faster than structured proteins. IDPs have larger capture radii, which is the basis for the fly-casting mechanism,[18,19] but it has also been proposed that larger Rg diminishes the diffusion coefficient.[7] The conformational flexibility of IDPs may allow them to bind to many distinct partners. Moreover, IDPs in general and transcriptions factors in particular, are subject to post-translational modifications, adding complexity to regulatory networks and to the regulation of the coupled folding and binding process.[20]

CREB Binding Protein (CBP) binds to specific Transcription Factors (TF) and the polymerase II complex, enhancing transcription of target genes.[21] One of these TFs is c-myb, involved in the regulation of hemapoietic cells life cycle.[22] The Transactivation Domain (TAD) of c-myb and the KIX domain of CBP are mostly responsible for binding interactions.[23]

The TAD region of c-myb is intrinsically disordered, retaining only 30% of helicity in solution, and folds upon binding to the KIX domain of CBP.[23] The NMR structure of the complex shows c-myb forming an α-helix with a kink around Leu302, which allows this residue to be deeply buried in the hydrophobic groove between α1 helix and α3 helix of KIX.[24]

Kinetic data available for c-myb - KIX binding is consistent with a two-state, one barrier process, with no accumulation of intermediates.[23,25] The association process has an apparent activation energy of approximately 11 kcal/mol and the dissociation process an enthalpic barrier of almost 20 kcal/mol.[23]

Association and equilibrium experiments carried out at different Trifluroethanol (TFE) concentrations suggest that the proportion of secondary structure acquired by the c-myb peptide in solution has no effect over the association rates (kon) but decreases koff, as the activation energy for the dissociation process increases about 1.2 kcal/mol by adding 10% TFE, (and thus, Kd decreases).[25] Giri and coworkers performed a Φ value analysis of the binding of c-myb and KIX by means of fluorescence change upon binding (using a Y652W KIX mutant) and proposed a transition state slightly more disordered than the native bound state[26] pointing to a conformational selection mechanism. However, new studies by Clarke suggest that conformational selection does not play a major role in c-myb KIX coupled folding and binding process and thus a folding after binding mechanism is proposed instead.[27]

A recent article by Peter Wright's group states that the αA helix of free c-myb has a higher helical tendency than αB and that while kon kinetic constants are almost the same for both helices, the koff constants of αA helix are significantly higher than those of αB helix.[28] This data highlights the relationship between the preorganized helical content and the kon (koff) changes. Reanalysing Giri et al data for the αA helix[26] and due its higher propensity for helicity, they hypothesize that there could be a very fast pre-folding step in the mechanism that cannot be detected by NMR, or at least that a majority of the flux of the reaction undergoes the proposed mechanism.[28] They also find that the main binding site for c-myb is the KIX site. They identified binding to KIX to another site, a site known to bind the transcription factor MLL, with a binding constant 180 times lower than the final binding site only when extremely high c-myb concentrations are used. So, at physiological concentrations binding to the MLL site is not relevant. Previous discrepancies create an opportunity to study the folding and binding process of c-myb to KIX by means of molecular simulations.

In this work we perform all atom Molecular Dynamics simulations of c-myb-KIX. We carry out umbrella sampling simulations of the binding of the two proteins, to provide an atomistic understanding of the free energy landscape of the coupled folding and binding process and of the transition state. We also perform long equilibrium Gō-type Coarse Grain simulations in order to capture the role of the native contacts and their effect in the kinetics of the binding process.

Results and discussion

Folding and binding of c-myb upon binding to KIX at atomic resolution

The sequences of c-myb-TAD and the KIX domain of CBP are shown in Figure 1A-B highlighting the secondary structure elements present in the bound state. The NMR structure of the complex of c-myb-KIX shows that KIX is formed by three helices $\alpha 1$, $\alpha 2$ and $\alpha 3$ and that c-myb forms two almost continuous helices, αA and αB, due to a kink located at Leu 302 in the middle of the peptide (Figure 1C). c-myb Leu 298, Leu 301 and Leu 302 anchor in a KIX hydrophobic groove formed by $\alpha 1$ and $\alpha 3$ helices. Previous experimental work[26] points out to this Leu302 and Leu298 as major contributors to the binding free energy. Native interactions of c-myb residues with KIX are shown in Figure 1. Polar, charged and hydrophobic contacts are depicted in Figure 1D, 1E and 1F respectively.

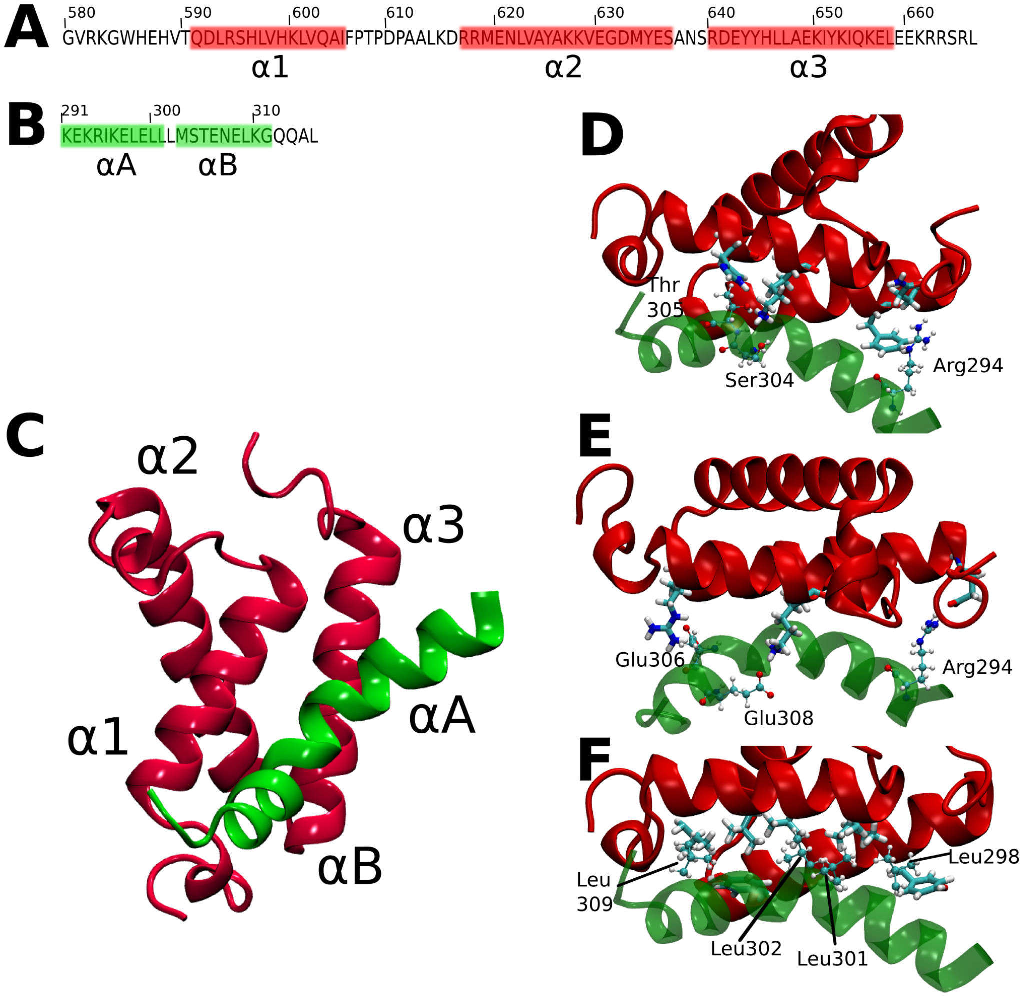


Figure 1: Sequence and interactions of the c-myb/KIX system. (a) Amino Acid sequence of KIX. The three helices are highlighted in red. (b) Amino Acid sequence of c-myb. The helices are highlighted in green. (c) The structure of c-myb/KIX. Cartoon 3D structure of the c-myb-KIX complex obtained from the experimental NMR structure pdbid: 1SB0. c-myb AB interacts with the face of KIX formed by helices 1 and 3. (d) Polar native contacts of the complex. Cartoon 3D structure of the complex c-myb/KIX showing polar intermolecular contacts represented in balls and sticks. c-myb Ser304 and Thr305 form a hydrogen bond with Lys600 of KIX 1. c-myb Arg 294 forms a hydrogen bond with Tyr 652 of KIX. (e) Charged native contacts of the complex. Cartoon 3D strcuture of the complex c-myb/KIX showing charged intermolecular contacts represented in balls and sticks. c-myb Glu308 forms a salt bridge with KIX Lys600. c-myb Glu306 interacts with KIX Arg640. c-myb Arg294 forms a salt bridge with Glu659 of KIX. (f) Hydrophobic Native contacts of the complex. cartoon 3D strcuture of the complex c-myb/KIX showing hydrophobic intermolecular contacts represented in balls and sticks. c-myb Leu298, Leu301, Leu302 and Leu309 are located in an hydrophobic pocket formed by helices 1 and 3 of KIX including residues Leu593, Leu597, Leu601, Ala604, Leu647 and Ile651.

It has been shown that long atomistic MD simulations are able to sample the folding of several globular domain proteins.[17,29,30] However, sampling coupled binding and folding is a more difficult task due to protein diffusion. Simulations that study IDPs recognition mechanisms have usually been done using Gō-Type coarse grained models.[6,14,31,32] To tackle this problem we decided to carry out long all atom umbrella sampling simulations starting at the bound complex and slowly increased the distance between the proteins reaching a separation where there are no contacts among them. We used as sampling coordinate the distance between KIX CB Ile 651 and c-myb CB Leu 301 (Figure 2A). No other bias was applied during the simulations. We ran 39 windows of 400ns each increasing the CB-CB reference distance by 0.5 Å comprising a total of 15.6 μs all atom simulation.

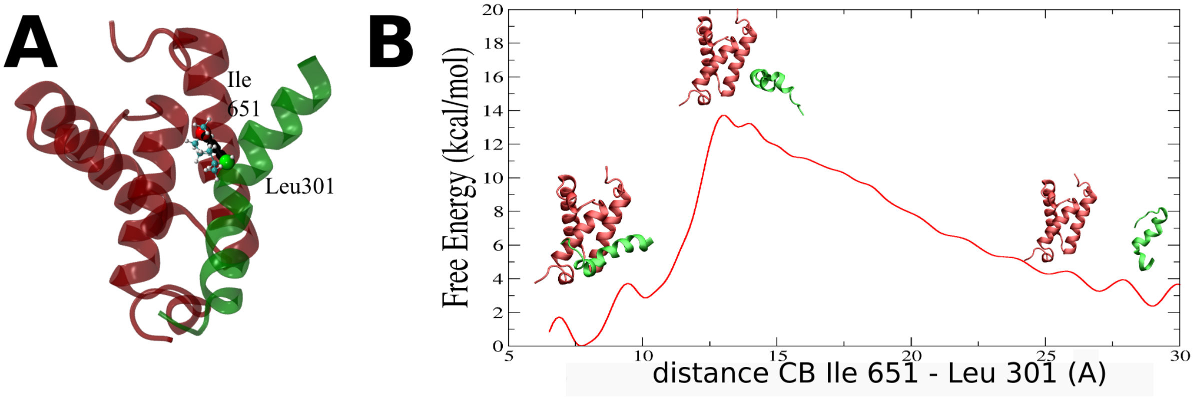


Figure 2: Umbrella sampling simulations of the folding and binding process of c-myb to KIX. (a) New cartoon drawing of the structure of the c-myb/KIX complex showing in red and green vdw spheres the atoms used for the umbrella sampling and in balls and sticks the residues involved. (b) Potential of Mean Force of Binding of KIX and c-myb in the distance coordinate (distance of KIX CB Ile 651 and c-myb CB Leu 301), obtained using 8000 frames from each of the 39 all-atom biased simulations with the vFEP algorithm.[49] The activation energy for the coupled folding/binding process is 12.9 kcal/mol and occurs at 13 Å. The result is in very good agreement with experimental data. We observe only one maximum in the Free Energy Surface with no intermediates in the free energy curve. The graphic shows a steady increase in the free energy from 25 Å to 13Å. A steep descent is observed in the free energy right after the top of the barrier due to a rapid increase in the native contacts between KIX and c-myb.

The potential of mean force of the folding and binding process is presented in Figure 2B. The binding activation free energy is 12.9 kcal/mol and no intermediates are observed.

These results are consistent with the experimental data.[23] The top of the free energy curve is located at 13 Å in the distance coordinate. The conformational ensemble at this point represent an apparent transition state(ATS) as we cannot guarantee that the simple distance bias coordinate is a "good" coordinate for the whole process.The PMF obtained through umbrella sampling simulations should almost always be corrected for a Jacobian term due to changing from cartesian to spherical coordinates, when sampling a distance. This term is usually written as +2 kbT log(distance), which is derived by assuming full spherical sampling in the angle terms. The effect of this correction on the barrier for kon is 0.43 kcal/mol and for the barrier for koff is 0.25 kcal/mol. In protein-protein interactions the assumption of full spherical symmetry is unwarranted since there are steric clashes between the systems at short distances. Given the fact that the corrections are very small and the underlying assumption might not fully apply, we note that the correction does not change the physical insights from the simulations.[54]

We also performed Coarse Grain equilibrium molecular dynamics simulations to gain insight into the dynamics and kinetics of the folding/binding and unfolding/unbinding pathways. Each simulation is 45 μs long and samples more than 20 binding and 20 unbinding events (Figure S1). We ran 8 simulations, which allowed us to compute equilibrium and kinetic properties.

Figure 3 shows different representations of the free energy landscape for this concerted reaction. Figs 3 A, C and E represent the potential of mean force obtained by means of atomistic simulations and Figs 3 B, D and F by coarse grain analysis. Qinter (fraction of native intermolecular contacts that are formed) versus Qintra c-myb (fraction of native c-myb intramolecular contacts that are formed) are depicted in Figure 3A and 3B for the all atom and coarse grain simulations respectively. In Figures 3C and 3D Qintra c-myb is plotted versus the sampling coordinate of the umbrella sampling distance(KIX CB Ile 651-c-myb CB Ile 301) and in panels 3E and 3F Qinter is plotted versus the umbrella reaction coordinate.

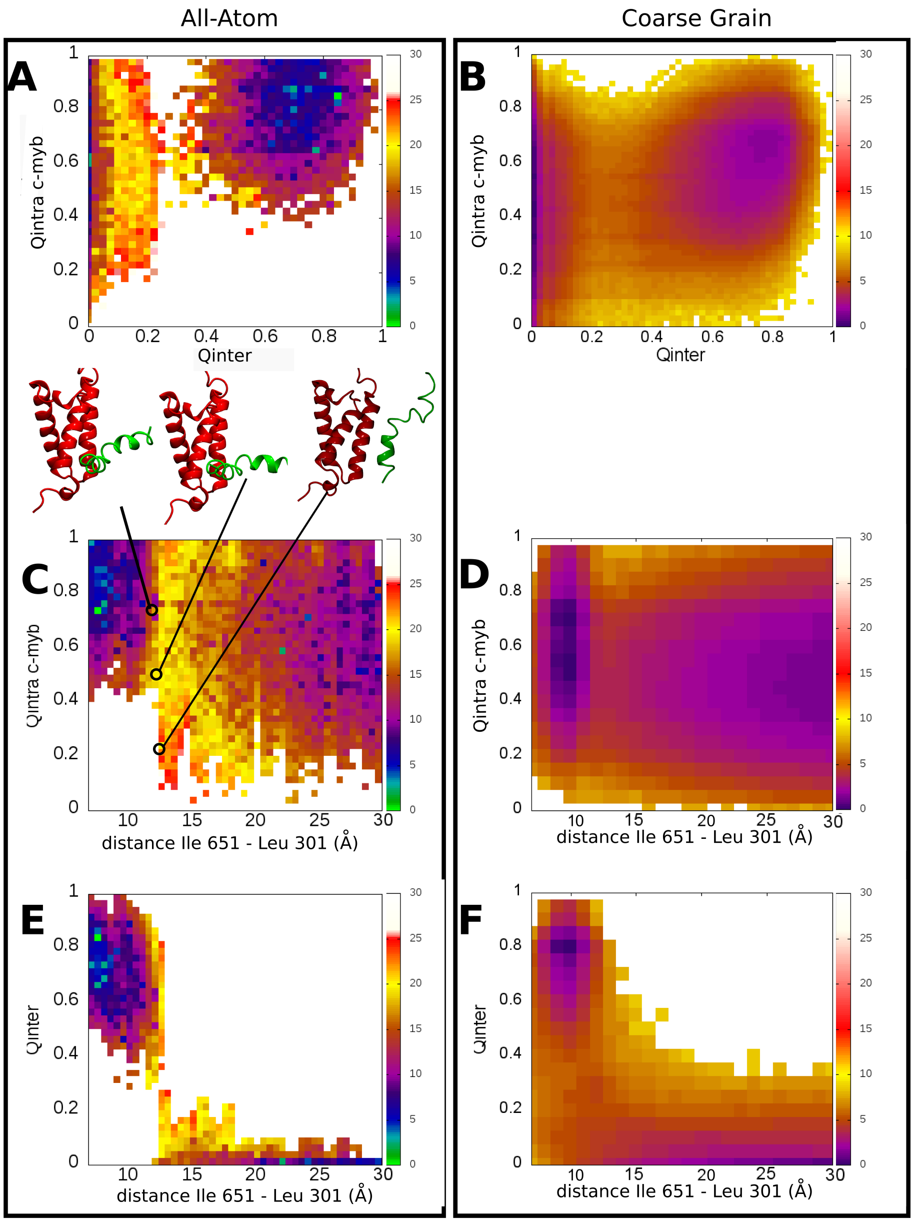


Figure 3: 2-Dimensional Potential of Mean Force for Folding and Binding of KIX-c-myb, obtained from biased all-atom simulations and equilibrium Coarse Grain simulations. (A) The fraction of intramolecular native contacts of c-myb (Qintra) is plotted versus the fraction of native contacts between c-myb and KIX (Qinter) is plotted for the all atom simulations and (B) Coarse Grain simulations. Both simulations are in agreement showing an increase in Qintra c-myb upon binding. Bound complexes have a broad population of Qintra c-myb as c-myb seems to have flexible regions even upon binding. When projected into these variables the TS is not well localized in the all atom MD. A broad transition state region is seen along the Qintra c-myb coordinate, around 0.23 in the Qinter coordinate in the Gō-type graph. In panels (C) and (D) Qintra c-myb is plotted versus the distance used in the umbrella sampling simulations (Ile651-Leu301). In these plots we clearly observe a broad TS in the Qintra c-myb coordinate which is around 13 in the distance coordinate. Three selected structures from the TS region are plotted showing how Qintra c-myb increases despite having the same free energy. In panels (E) and (F) Qinter is plotted versus the distance Ile651-Leu301. As in the umbrella sampling curve the native contacts between the proteins increase fast upon binding indistictintangly of the structure of c-myb. All intermolecular contacts collapse when a contact is made between the two proteins.

The results of the all atom and coarse grain simulations are in agreement with respect to the folding and binding mechanism. There is a clear increase in Qintra c-myb upon binding, but we observe significant population of native-like structures even in the unbound complex, Qinter <0.1 shows a broad distribution of Qintra c-myb from 0.2 to 0.8 in panels A-B, the same is observed at long distances(d >20 Å) in panels C-D. In the atomistic simulations we calculated an average helix population of 39% for unbound c-myb, in good agreement with previous experiments by Clarke's group where they estimated that the c-myb peptide in solution has an helical content of the order of 28% to 38% by measuring the CD spectra.[23] Once c-myb is bound, Qinter c-myb >0.6 or distances <11 Å the distribution of Qintra reduces but we still have a broad population of Qintra c-myb, from 0.5 to 0.8, this is due to flexible regions that explore different conformations even when c-myb is bound, mostly from helical to disordered.

When we plot Qinter versus the bias distance for the all atom simulation we observed that in the transition state ensemble, located around 13 Å, Qintra C-myb has a broad distribution with values going from 0.2 to 0.8. This result clearly shows that c-myb can bind with the same barrier with Qintra c-myb as low as 0.2 or as high as 0.8. We show three selected structures from the TS region, Qintra c-myb 0.2, 0.5 and 0.75, to characterize how c-myb can bind to KIX. Similar results are observed in panels B and D, corresponding to coarse grain simulations. Care must be taken when comparing Figure 2 with Figure 3, as the barriers measured at the highest point in the 2D surface can be different from the 1D ones. This is because in Figure 2 the 1D profile is obtained by integrating Figure 3 along the non-distance coordinate. Since the value of this integral is different for each value of the distance, the simple reading of the 2D plot cannot really give the same value for the barrier when projected onto 1D. Moreover, despite the fact that Figure 3 gives very valuable information, the statistics for the event count when going from 1D to 2D have significant noise for some values of the coordinates and one can not simply look at the free energy value in each point.

One key aspect of the folding and binding process is that when residues in the unbound structures make contact with the surface of KIX fast folding occurs as can be observed following the change in Qintra c-myb along the Ile651-Leu301 distance in figure 3C. We also observe a fast decrease in the free energy in Figure 2B. Our results support that the mechanism of binding is not unique, and many different reaction paths (combinations of unstructured and structured binding) can be followed with the same energetic cost.

We have shown that we can sample the folding and binding process with the all atom biased simulations and we have identified the key features of the folding mechanism. We will now describe the specific interactions that occur during the binding process.

We can see in Figure 4 (which shows which are the most probable contacts formed at different distances between the proteins) that the central residues of c-myb seem relevant for binding and folding but are not the only residues that have interactions in the process. In panel A we show the results obtained at a Ile651-Leu301 8 Å distance, the bound complex. The contacts that have red color imply that a contact is formed between those residues in most of the structures obtained along the simulation and white means that the contact is not observed. The region of c-myb that spans from aminoacid 294 to 302, central αA and the hinge region, has strong interactions with the KIX region spanning aminoacids 649 to 660 encompassing helices $\alpha2$ and $\alpha3$ of KIX. In particular, as depicted in figure 1, Arg294, Leu298, Leu301 and Leu302 of c-myb contact with Leu647, Ile651 and Glu659 of KIX. Interestingly, during the simulations these residues of c-myb, that are located in the face of the helix that looks towards KIX, interact with nearby residues besides the expected native contacts. This means that due to its intrinsic flexibility c-myb also establishes other contacts beside the contacts previously observed in the NMR structure. The other region that has interactions involves helix 1 of KIX and residues 302 to 308 of c-myb. Residues 302 and 306 interact with both regions of KIX. Specific interactions depicted in figure 1 involve Ser304, Thr305, Glu306, Glu308 and Leu309 of c-myb and residues Leu593, Leu597, Lys600, Leu601, and Ala604 of KIX. Again, we observe in these regions that nearby residues, also have interactions during the simulations.

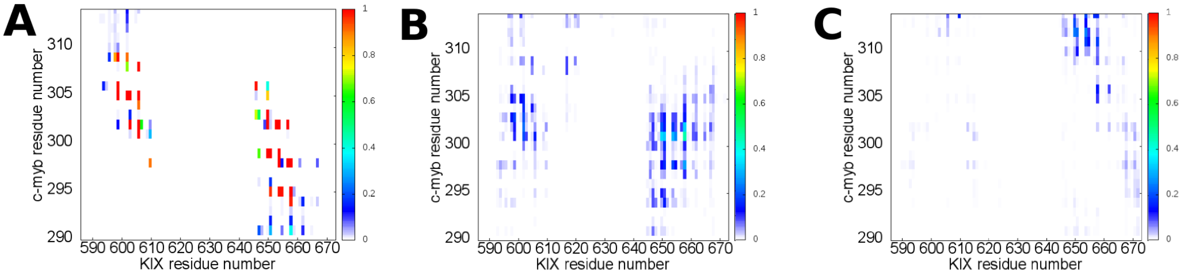


Figure 4: Fraction of time each possible intermolecular contact is formed for different restraint distances: (A) 8 Å, complex formed. The figure displays the native interactions in the bound and folded complex; (B) 13 Å, the transition state distance. The panel shows relevant contacts formed by the centre of c-myb and a pattern similar to panel A and (C) 21 Å, dissociated complex. The figure depicts the existence of contacts even at far distances. At distances longer than 22 Å almost no contact is observed.

In panel B (Figure 4) we show the contacts at the transition state distance of 13 Å. In this panel all contacts are blue colored, meaning that no contact is observed more that 40% of the time during the simulation, but most of the contacts resemble the ones observed in the bound state. Strong native contacts of c-myb αB, with populations of around 20%, are located in the region spanning residues 301 to 305 of c-myb. Again, as in the bound complex, other interactions are observed with residues that are located in the same face of KIX but now due to the fact that the helix is not well formed the effect is stronger. Interestingly, these other populated interactions are observed ”around” the native ones, meaning that they do not conform an intermediate but are transient nearby interactions. This contributes to the roughness of the potential energy surface but still dominated by the native contacts. This tendency is more pronounced in the region that includes αA and the hinge of c-myb, with stronger interactions in the native core, with populations up to 40%, but with several interactions spanning residues 647 to 667 of KIX. In brief we observe that the transition state involves mainly native interactions (within a broad spectrum) with a higher contribution of αA and the hinge of c-myb.

At distances as far as 22 Å depicted in figure 4C, we observe contacts with populations as high as 20% that are not native contacts. At higher distances no relevant interactions are observed. At the 22 Å distance the change in free energy is relatively small as compare to higher distances where no contacts are observed. Interestingly, the initials contacts observed are non-native, indicating that they have a role in the recognition process of KIX and c-myb. As we will discuss in the next section, we believe that non-native interactions are relevant in the context of the fly casting mechanism because they can dramatically affect binding kinetics as has been previously proposed.[53]

As we have shown in Figure 4 c-myb is able to form other contacts besides the native ones. During our simulation c-myb initiated from the bound conformation and the CB of Leu 302 was set appart as far as 30 Å from the CB of Ile 651 of KIX. To show the accesible interactions for c-myb during the umbrella sampling simulations we calculated the probability density corresponding to the CB of residue Leu 301 in c-myb at the three distances representing the TS unsemble at 13 Å, the distance depicted in Figure 4C with non-nantive interactions at 22 Å and a distance with no contacts between both proteins at 30 Å (Figure S2). Our results clearly show that even though we have used a distance restraint for our umbrella sampling calculations, c-myb can sample significant angular space and make many different contacts with KIX.

Do Native contacts determine the folding and binding mechanism of c-myb?

Our coarse grain equilibrium molecular dynamics simulations, which are in agreement with previously published results by Brooks,[32] sampled more than 200 binding and unbinding events allowing a good estimate of the free energy surface. In figure 3B we plotted Qintra c-myb vs Qinter for the G$\bar{o}$-type potential. A comparison with the atomistic simulation clearly shows similar results, describing a broad transition state with Qintra c-myb going from 0.2 to 0.7.

To gain further insight of folding/binding mechanism we computed the conditional probability p(TP|Qinter) of being on a transition path (TP) given a particular fraction of intermolecular contacts, Qinter, to identify the TS ensemble of the coupled folding-and-binding process. TP are defined as trajectory segments that connect unbound conformations (Qinter <0.05) with bound conformations (Qinter >0.7), and vice versa, as observed in Figure 3B, without re-crossings (Figure S3). The value of Qinter with the highest p(TP|Qinter) is most indicative of being on a transition path, and is used to identify the transition states. The largest value of p(TP|Qinter) is obtained for Qinter from 0.2 to 0.3. This result nicely correlates with the Free energy showed in figure 3B, were Qintra c-myb is plotted vs Qinter and figure 5A, where the free energy curve is shown, in all figures the TS is located at Qinter around 0.2-0.3. More interestingly, a similar result is observed in figure 3E for the all atom simulation around the 13 Å value in the distance coordinate, where the all atom TS is located in Figure 2, that shows Qinter values around 0.2-0.3 again indicating that both type of simulations gives similar results.

The obtained maximum p(TP|Qinter) is around 0.2, a low value as compared to theoretical maximum of 0.5 for a perfect reaction coordinate of a diffusive process, showing that even though the Gō-type potential reproduces the general pathway the broad transition state, global intermolecular fraction of native contacts is not a good reaction coordinate. We tried many other one dimensional variables (i.e. one intermolecular native contact) and collective variables (i.e. a subset of intermolecular native contacts) with no better results. A combination of 1D collective variables (Qinter,QinterA) was used to get a good picture of the transition state (Figure S4), and the maximum obtained for p(TP|Qinter,QinterA) is around 0.4. This maximum for p(TP| Qinter,QinterA) is located at around 0.25 in Qinter and between 0.3 and 0.7 in QinterA. As there is a bottleneck in the free energy surface at around 0.25 in Qinter and 13 Åin the distance coordinate, this provides evidence that the TS should be located at 13 Åin the distance coordinate, which is also the top in the free energy apparent barrier. Non-native interactions could also contribute to the pathway as shown using the all-atom force field in Figure 4. As we stated before, the non-native interactions are residues that are like a second shelf of the native interactions and seem to be formed due to high flexibility of the apparent TS. We believe they contribute in an unspecific way to the folding and binding pathway just to accelerate the process as previously proposed in the context of the fly-casting mechanism. Even though Wolynes predicted a rate enhancement of up to 2-fold, we previously obtained a similar value previously for the pkid/KIX case, he formulated the theory by only considering native contacts. Indeed, when non-native interactions were included a more dramatic effect in the binding kinetics was shown.[33,53]

As we obtained similar pathways and free energy surfaces with both all-atom and Gō-type potential we conclude that each folding and binding pathway is governed by native contacts but due to the heterogeneity of the transition state ensemble we do not have an specific native contact that is always observed in most of the transition paths. Moreover, in each transition path we observed that also non-native contacts are formed but again there is no key non native interaction observed neither in the Gō-type or all atom simulations.

Folding before binding, binding before folding or both?

Two general mechanisms have been proposed for protein coupled folding and binding: conformational selection and induced folding. The conformational selection model argues that the unfolded/unbound state is in a dynamical equilibrium among many conformations, but one or more of those conformations, that resemble the native bound structure, preferentially bind to its partner. These conformational states might be weakly populated in the unbound state. This mechanism implies that folding occurs before binding, and that these are sequential steps.

The induced folding model proposes that when weak interactions between both partners are formed, a shift is produced in the conformational ensemble towards the bound native state. Thus, some degree of binding precedes folding, and there this a one step process as folding is coupled to binding.

Based on the previous calculated 2-D potential of mean force (Figure 3) we observed that the unbound peptide explores a broad number of conformations, some very unstructured (with Qintra c-myb being almost zero) and others quite ordered (with Qintra c-myb above 0.6). Interestingly, transition paths obtained in the Coarse Grain simulations can be initiated from any of the states in the unbound basin, high and low Qintra, and cross the transition state region throughout a wide range of Qintra c-myb. We calculated the distribution of lengths of transition paths for the ones crossing the TS through the low Qintra c-myb region (Qintra c-myb <0.6) and through the high Qintra c-myb region (Qintra c-myb >0.6). Probabilities of Transition Paths crossing the transition state region trough the high Qintra c-myb region and the ones starting at the low Qintra c-myb region are similar and, moreover, the mean time of these transition paths lengths are also similar (Figure S5). The number of transitions that start at Qintra >0.6 is 420, and the ones starting at Qintra <0.6 is 384, this means 52% and 48% respectively. The distributions of duration of transitions paths is broad but very similar as depicted in Figure S5, with a mean time of approximately 0,11ns for both transitions starting at Qintra <0.6 and starting at Qintra >0.6. This broad TS however is shifted towards higher helicity values and Qintra c-myb, indicating that there is some more order in the TS ensemble than in the unbound ensemble, and in this respect is more bound-like, so a more structured TS is involved in the folding and binding mechanism. We observe that the change in Qintra c-myb is accompanied by the formation of intermolecular contacts. However, instead of observing one specific contact we identify several contacts that induce folding of c-myb each other independently of the others.

To gain a further insight into the recognition mechanism we performed Gō-type Coarse Grain simulations where the intramolecular contacts of c-myb were strengthened, so the protein is more structured in the unbound state. This was done to correlate our simulations with the experimental data obtained in the presence of TFE, which is known to increase secondary structure in the unfolded state. As can be seen when comparing the plots of Qintra c-myb vs Qinter for the temperature of Folding set at 365 K (Figure S6A) and 410 K (Figure S6C) the unbound state in the latter case shifts towards more structured states. As can be seen in Figure S6 the main mechanism of folding and binding is similar in both cases, but we observe as expected a more structured transition state like region in the 410 K case.

These simulations can shed light into the effect that a more structured c-myb can have on the kon and koff of the process and how our results correlate with previous experiments. In Figure 5 we show the free energy curve projected on both Qinter c-myb-KIX and the distance used in the umbrella sampling simulations. As expected, there are several differences between the G$\bar{o}$-type and the atomistic free energy surfaces, shown in Figure 5B and Figure 2B respectively. The first one is calculated with a coarse grained unbiased simulation that only has favourable interactions between the native contacts as compared to the later that is calculated with a biased simulation with an atomistic potential that evaluates interactions between all the atoms in the system. These differences account, for example,for the fact that the G$\bar{o}$-type curve is smoother than the atomistic curve, differences in the free energy between states and in the activation energy. But both agree qualitatively that free energy increases slowly when decreasing distance at the long distance region and decreases very fast upon crossing the top of the apparent free energy barrier. When we set the folding temperature to 410K (Figure 5A red curve) the free energy curve shows an activation barrier both for binding/unbinding around 5.0 kcal/mol. On the other hand, in the case when the folding temperature is set to 365K (Figure 5A black curve) the top of the free energy curve for binding is again around 5.0 kcal/mol but the activation barrier for unfolding is 3.7 kcal/mol. Similar results are observed when we plotted with respect to the umbrella sampling distance. We also estimated the mean residence times for the bound state and the unbound state for both models (Table 1). When structuring c-myb, we observe that kon is similar but a high effect is observed on koff that strongly increases the folding temperature. These results implicate that inducing structure in c-myb stabilize the bound complex structure, but make no changes in the kon. As in the apparent transition state ensemble coexist both structured and unstructured c-myb when binding to CBP, if we induce structure in c-myb the protein follows the structured pathways with a similar activation barrier as before. These results are in agreement with previous experiments[25,27] and consistent with a broad Transition State. Previous experiments done by adding 10% TFE showed no effect over the kon but decreased the koff, as the activation energy for the dissociation process increases about 1.2 kcal/mol.[25] Thus, structuring c-myb does not accelerate the coupled folding and binding process.

Table 1: Kinetic data obtained from coarse grain simulations

	c-myb Tf 360 K	c-myb Tf 410 K
mean first passage time (ns)	466	286
mean residence time in bound state (ns)	204	1357
mean Transition Path length (ns)	0.30	0.17
Transmission coefficient bound to unbound	0.45	0.35
Transmission coefficient unbound to bound	0.27	0.37

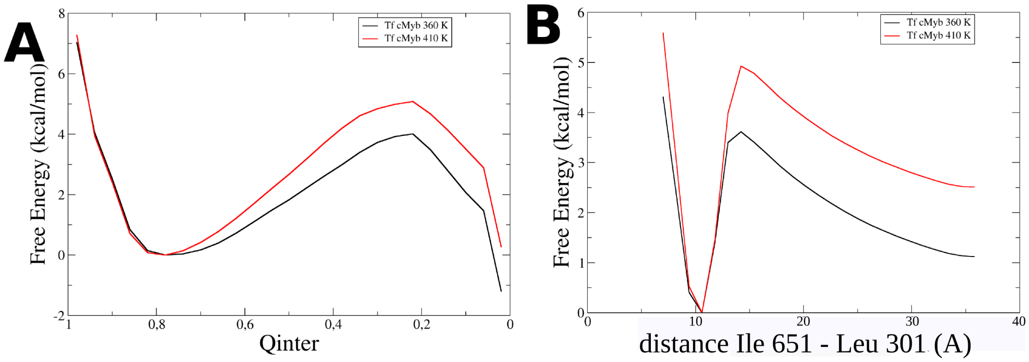


Figure 5: Potentials of Mean Force of Binding of KIX and c-myb, obtained from Coarse Grain equilibrium simulations. Activation Energy for the Binding process seems not to be affected. (A) Potential of Mean Force versus Qinter. (B) Potential of Mean Force versus distance of KIX Ile651 and c-myb Leu301.

Giri and coworkers performed a Φ value analysis for c-myb binding to CBP measuring the change in fluorescence of a tryptophan, they mutated Y652 to W.[26] We also performed a Φ value analysis for each amino acid residue of c-myb that has native contacts with KIX in the all atom simulations. We estimated the Φ values by using a simple definition, Φ = Qx(TS)/Qx(bound) where Qx is the number of native contacts in the respective state between KIX and residue x of c-myb. Our results indicate that relevant contacts formed in the apparent TS are present in the native bound state. However they do not indicate that the apparent transition state is structured as previously proposed (Table S1). Is important to clarify that our Φ values are estimated based on the simple idea that more prevalent contacts are more relevant, but do not actually mean that these residues will indeed contribute more strongly to the free energy of the transition or the bound state. Therefore we only expect qualitative agreement with experiments. All our calculated Φ values are low as we have a broad ensemble of conformations in the apparent transition state region. Two of the highest estimated Φ values in our simulation correspond to residues Leu298 and Leu302 (0.30 and 0.23, respectively), and they were labelled as essential for binding in previous experiments (mutants Leu298Ala and Leu302Ala bind so weakly to KIX that binding cannot be measured). Ser304, Glu308 and Arg294 were reported to have experimental Φ values higher than 0.5, and in our simulations they all have calculated Φ values higher than or around 0.1. Glu299 and Met303 have calculated Φ values lower than 0.05 in our simulations and low experimental Φ values (lower than 0.4). Glu306 cannot be compared because in previous experiments had very low $\Delta\Delta$G of binding. Despite the general agreement, a TS ensemble where native contacts dominate binding but due to the variety of conformations in each one a different native contact is used could produce the Φ value graphs obtained by Gianni and coworkers,[26] and would also explain the results obtained from mutants by the Clarke group[23,27] and the TFE experiments done by Brunori's group.[25]

We also made several graphs of 2D potentials of mean force, of QinterA (fraction of native intermolecular contacts made by αA helix) and Qinter for the all-atom and the Coarse Grain simulations, and the same for QinterB (fraction of native intermolecular contacts made by αB helix) and Qinter coordinates (Figure S7). We show that αA helix binds preferentially structured, but there are paths for mostly unstructured αA helix (QinterA $<$0.5) of similar activation energy that those structured ones. αB helix appears much more unstructured, both in the free c-myb state and the TS (Figure S7). We can see again excellent agreement between biased all-atom and equilibrium coarse grain simulations. This is in agreement with recent NMR results.[28] However, we still observe that c-myb as a whole may go via structured or unstructured TPs, and that the flux of the two possible pathways is similar. Even more, there is a minority, but significant flux through unstructred αA helix conformations.

Discussion

The debate on mechanisms on protein dynamics that could give insights into protein function is old in biochemistry. Induced fit model was first proposed by Koshland in 1959[35,36] to explain how protein dynamics could account for allosteric effects in enzyme catalysis. The assumption that only few (or one) of the protein conformational states are responsible for the observed activity exists at least since the Monod-Wyman-Changeaux (MWC) model was published in 1965.[37] There is still great debate on the subject[18,36,38] and it has extended to many other fields and, in our case, to the studies of Intrinsically Disordered Proteins. Currently, it is proposed that both mechanisms may be operative and the specific characteristics of the unstructured protein determine which mechanism dominates.[18,39–42]

In our case, neither conformational selection nor induced fit mechanisms dominate the coupled folding and binding process. C-myb may bind unstructured to KIX and fold over the surface or may acquire a bound like structure before binding and then recognize KIX. In our simulations both pathways turn out to be equally probable and have similar kon.

The transition state ensemble comprises a wide and extended portion of the energy surface when projected onto a variety of different collective variables. Only in the distance KIX Ile 651 - c-myb Leu 301 and Qinter we can observe a bottleneck for transition paths, but still those states sample a large number of c-myb conformations. A common feature is that contacts at the center of c-myb are the most populated ones in the transition states.

Our results are consistent with experimental data, stating that the highest individual Φ values are located at the center of c-myb,[26] that the activation energy for coupled folding and binding is above 10 kcal/mol and that kon is not significantly affected by shifting free c-myb conformational ensemble towards more ordered states, either by addition of Trifluoroethanol[25] or by mutational studies.[27] They are also in agreement with R_2 dispersion relaxation NMR data stating that αA helix is much structured in the free and transition states and that αB conformational ensemble is more disordered in both states.[28]

The excellent agreement between G$\bar{o}$-type Coarse Grain simulations and all-atom simulations carried out with a transferable force field indicate that the mechanism of coupled binding/folding is governed by native contacts.

Conclusions

We studied in detail by the folding and binding mechanism of the transcription factor c-myb to the KIX domain of the co-transcription factor Creb Binding Protein. Our results provide insights into the mechanisms of coupled folding and binding for Intrinsically Disordered Proteins, one of the few works to do so at atomistic level.

We compared atomistic simulations using a transferable potential with a G$\bar{o}$-type coarse grain model and conclude that native contacts determine the coupled binding/folding process. The apparent transition state ensemble identified in the atomistic simulation and the TS region of the coarse grain simulations are very broad with both unstructured and structured conformations. Previous experiments were controversial regarding the recognition mechanism. Our results are able to explain previous experiments and contribute to give a clear picture of how c-myb binds to CBP.

Methods

All atom simulations

The initial structure files were obtained from the structure of the complex of c-myb (291-315) and KIX (580-666) from the Protein Data Bank (PDB ID: 1SB0). We selected the first structure of the 10 models for the simulations.[24] The AMBER99SB force field was used[43] and tleap program was used to create the topology and coordinate files. Implicit solvent conditions were used. The system was equilibrated in the NVT ensemble by running a 25 ps long MD simulation using the Berendsen thermostat[44] and then the temperature was slowly raised to 300K while running another 25 ps long simulation. During these processes the CA atoms were restrained using a harmonic potential with a 20 kcal/mol constant for the thermalization. The temperature was kept constant by the Berendsen thermostat algorithm set at 300 K with a 0.1 ps coupling constant.[44] The SHAKE algorithm was used for constraining the bonds that contained an H atom.[45]

400 ns long molecular dynamics simulations were run with a 4 fs timestep, using a hydrogen mass repartitioning scheme[46] with an harmonic potential in the distance between KIX Ile651 CB and c-myb Leu301 CB, each with a different reference distance and the AMBER package.[47] This coordinate was chosen as it was close to the center of the KIX hydrophobic groove and to c-myb Leu302 which anchors in it, but did not form a contact between both peptides. The minimum reference distance was 7 Å and then it was increased by 0.5 Å until 21 Å, when it was increased by 1 Å until the last simulation which was carried out with a 30 Å reference distance on the harmonic potential. We used a force constant of 32 kcal/mol. Histograms in the reaction coordinate are provided as to assess the overlapping sampling of the windows (Figures S8-S10). The vFEP program was used for performing WHAM.[48,49]

Coarse Grain Simulations

The initial structure files were obtained from the structure of the complex of c-myb (291-315) and KIX (580-666) from the Protein Data Bank (PDB ID: 1SB0). We selected the first structure of the 10 models for the simulations.[24] We created the topology and initial coordinate files using the Karanicolas-Brooks standard protocol for a G$\bar{o}$-Type coarse grain model.[31,34]

Two sets of simulations were carried out. One with a 365 K folding temperature and another one where c-myb intramolecular native contacts were strengthened to a folding temperature of 410 K. Eight 45 μs long Langevin molecular dynamics simulations set at 300

K were performed for each set under gromacs4.0.5, using a 15 fs timestep (a total of 720 μs),[50,51] with a friction coefficient of 0.2 ps-1 and all Cα-Cα bonds were constrained using LINCS algorithm.[52]

Acknowledgement

This research is part of the Blue Waters Sustained-Petascale Computing Project, which is supported by grants from the National Science Foundation (Award No. OCI 07-25070) and the state of Illinois. Blue Waters is a joint effort of the University of Illinois at UrbanaChampaign and its National Center for Supercomputing Applications. Research funding comes from National Science Foundation (OCI-1440031). The authors thank the High-Performance Computing Center at the University of Florida for providing further computational resources. Adrian G. Turjanski is member of the CONICET. PICT-PRH 2010 grant from the ANCyPT.

Supporting Information Available

A graph of Qinter vs time of a typical Coarse Grain simulation; probability density graph of the position of c-myb with respect to KIX, tp(TP|Qinter) vs Qinter; p(TP|QinterA,Qinter); frequence hystogramas of structured and unstructured TPs; 2D Potential of Mean Force plots for the structured c-myb coarse grain force field; 2D Potential of Mean Force plots for the folding and binding of c-myb to KIX; hystograms of restraint distance.

This material is available free of charge via the Internet at `http://pubs.acs.org/`.

References

(1) Schweers O, Schnbrunn-Hanebeck E, Marx A and Mandelkow E (1994). *J Biol Chem 269(39)*: 24290-24297.

(2) Weinreb PH, Zhen W, Poon AW, Conway KA and Lansbury PT Jr (1996). *Biochem 35(43)*: 13709-13715.

(3) Dunker AK, Lawson JD, Brown CJ, Williams RM, Romero P, Oh JS, Oldfield CJ, Campen AM, Ratliff CM, Hipps KW, Ausio J, Nissen MS, Reeves R, Kang C, Kissinger CR, Bailey RW, Griswold MD, Chiu W, Garner EC and Obradovic Z (2001). *J Mol Graph Model 19(1)*: 26-59.

(4) Uversky V N (2002). *Eur J Biochem 269(1)*: 2-12.

(5) Tompa P (2005).*FEBS Lett 579(15)*:3346-3354

(6) Ganguly D and Chen J (2011). *Proteins 79(4)*: 1251-1266.

(7) Huang Y and Liu Z (2009).*J Mol Biol 393(5)*: 1143-1159.

(8) Sugase K,Dyson HJ and Wright PE (2007). *Nature 447(7147)*: 1021-1025.

(9) Wells M, Tidow H, Rutherford TJ, Markwick P, Jensen MR, Mylonas E, Svergun DI, Blackledge M and Fersht AR. (2008). *Proc Natl Acad Sci USA 105(15)*: 5762-5767.

(10) Uversky VN (2011). *Int J Biochem Cell Biol 43(8)*: 1090-1103.

(11) Iesmantavicius V, Dogan J, Jemth P, Teilum K and Kjaergaard M (2014). *Angew Chem Int Ed Engl 53(6)*: 1548-1551.

(12) Dogan J, Gianni S and Jemth, P (2014). *Phys chem chem phys 16(14)*: 6323-6331.

(13) Dogan J, Mu X and Jemth P (2013). *Science Reports 3*: 2076.

(14) Turjanski AG, Gutkind JS, Best RB and Hummer G (2008). *PLoS Comp Biol 4(4)*: e1000060.

(15) Zhang Z, Mazouchi A, Chong A, Forman-Kay J and Gradinaru C. (2014). *Biophys J 106(2)*: 50a.

(16) Yuwen T and Skrynnikov NR (2014). *J Magnetic Resonance 241*: 159-166.

(17) Piana S, Lindorff-Larsen K and Shaw DE (2012). *Proc Natl Acad Sci USA 109(44)*: 17845-17850.

(18) Boehr DD, Nussinov, R and Wright PE (2009). *Nat chem biol 5(11)*: 789-796.

(19) Shoemaker BA, Portman JJ and Wolynes PG(2000). *Proc Natl Acad Sci USA 97(16)*: 8868-8873.

(20) Dyson HJ and Wright PE (2015). *Nat Rev Mol Cell Biol 16(1)*: 18-29.

(21) Shao Y, Zhang G, Lu J and Huang B(2004). *Chinese Science Bulletin 49(24)*:2555-2562.

(22) Greig KT, Carotta S and Nutt SL (2008). *Seminars in immunology 20(4)*: 247-256.

(23) Shammas SL, Travis AJ and Clarke J (2013). *J of Phys Chem B 117(42)*: 13346-13356.

(24) Zor T, De Guzman RN, Dyson HJ and Wright PE. (2004). *J Mol Biol 337(3)*: 521-534.

(25) Gianni S, Morrone A, Giri R and Brunori M. (2012). *Biochem Biophys Res Comm 428(2)*: 205-209.

(26) Giri R, Morrone A, Toto A, Brunori M and Gianni S (2013). *Proc Natl Acad Sci USA 110(37)*: 14942-14947.

(27) Shammas SL, Travis, AJ and Clarke J (2014). *Proc Natl Acad Sci USA 111(33)*: 12055-12060.

(28) Araia M, Sugase K, Dyson HJ and Wright PE (2015). *Proc Natl Acad Sci USA 112(31)*: 9614-9619.

(29) Freddolino PL, Liu F, Gruebele M and Schulten K (2008).*Biophys J 94(10)*: L75-L77.

(30) Nguyen H, Maier J, Huang H, Perrone V and Simmerling C (2014). *J Am Chem Soc 136(40)*: 13959-13962.

(31) Karanicolas J and Brooks CL (2002). *Prot sci 11(10)*: 2351-2361.

(32) Law SM, Gagnon JK, Mapp AK and Brooks CL. (2014). *Proc Natl Acad Sci USA 111(33)*: 12067-12072.

(33) De Sancho D and Best RB. (2012). *Mol Biosyst 8*: 256-267.

(34) Karanicolas J and Brooks CL (2003). *J Mol Biol 334(2)*: 309-325.

(35) Koshland, DE (1958). *Proc Natl Acad Sci USA 44(2)*: 98-104.

(36) Koshland, DE (1995). *Angew Chem Int Ed Engl 33(2324)*: 2375-2378.

(37) Monod J, Wyman J and Changeux JP (1965). *J Mol Biol 12*: 88-118.

(38) Changeux JP and Edelstein S (2011). *F1000 biol rep 3*: 19.

(39) Wang Y, Chu X, Longhi S, Roche P, Han W, Wang E and Wang J (2013). *Proc Natl Acad Sci USA 110(40)*: E3743-E3752.

(40) Hammes GG, Chang YC and Oas TG (2009). *Proc Natl Acad Sci USA 106(33)*: 13737-13741.

(41) Wlodarski T and Zagrovic B (2009). *Proc Natl Acad Sci USA 106(46)*:19346-19351.

(42) Ganguly D, Zhang J and Chen (2013). *Plos Comp Biol 9(11)*: e1003363.

(43) Hornak V, Abel R, Okur A, Strockbine B, Roitberg A and Simmerling C (2006). *Proteins 65(3)*: 712-725.

(44) Berendsen HJC, Postma JPM, van Gunsteren WF, DiNola A and Haak JR. (1984). *J Chem Phys 81(8)*: 3684-3690.

(45) Forester TR and Smith W (1998).*J Comput Chem 19(1)*: 102-111.

(46) Hopkins CW, Le Grand S, Walker RC and Roitberg AE. (2015). *J Chem Theory Comput 11(4)*: 1864-1874.

(47) Case DA, Berryman JT, Betz, Cerutti DS, Cheatham, III TE, Darden TA, Duke RE, Giese TJ, Gohlke H, Goetz HW, Homeyer N, Izadi S, Janowski P, Kaus J, Kovalenko A, Lee TS, LeGrand S, Li P, Luchko T, Luo R, Madej B, Merz KM, Monard G, Needham P, Nguyen H, Nguyen HT, Omelyan I, Onufriev A, Roe DR, Roitberg A, Salomon-Ferrer R, Simmerling CL, Smith W, Swails J, Walker RC, Wang J, Wolf RM, Wu X, York DM and Kollman PA (2015). University of California.

(48) Kumar S, Rosenberg JM, Bouzida D, Swendsen RH and Kollman PA. (1992). *J Comput Chem 13(8)*: 1011-1021.

(49) Lee TS, Radak BK, Huang M, Wong KY and York DM. (2014).*J Chem Theory Comput 10(1)*: 24-34.

(50) Berendsen HJC, van der Spoel D and van Drunen R (1995). *Computer Physics Communications 91(1-3)*: 43-56.

(51) Hess B, Kutzner C, Van Der Spoel D and Lindahl E. (2008). *J Chem Theory and Comput 4(3)*: 435-447.

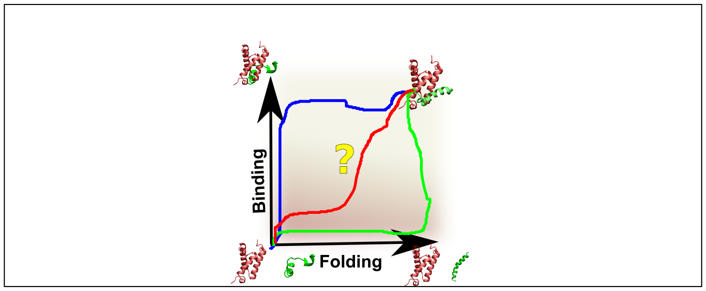
(52) Hess B, Bekker H, Berendsen HJC and Fraaije JGEM. (1997).*J Comput Chem 18(12)*: 1463-1472.

(53) Zhou H and Szabo A (2004). *Phys Rev Lett Volume 93(17)*: 178101-178104.

(54) Trzesniak D, Kunz AE, and van Gunsteren WF (2007). *ChemPhysChem Volume 8(1)*: 162-169.

(55) Bomblies R, Luitz, MP and Zacharias, M. (2016). *The Journal of Physical Chemistry B*

Graphical TOC Entry


Binding
?
Folding