UNIVERSITY OF CALIFORNIA, SAN DIEGO

Computational Tools based on Energy Landscape Theory to Predict Structurally Diverse Ensembles of Transcription Factors

A dissertation submitted in partial satisfaction of the

requirements for the degree

Doctor of Philosophy

in

Chemistry

by

Joachim Lätzer

Committee in charge:

Professor Peter G. Wolynes, Chair
Professor Gourisankar Ghosh
Professor Elizabeth A. Komives
Professor Katja Lindenberg
Professor J. Andrew McCammon
Professor José N. Onuchic

2007

The dissertation of Joachim Lätzer is approved, and it is acceptable in quality and form for publication on microfilm:

Chair

University of California, San Diego

2007

DEDICATION

I would like to dedicate this thesis to my loving family, which granted me unlimited support and encouragement. I would also like to dedicate this thesis to my great uncle Walter Lätzer, the only other Doctor of Philosphy in our family, whose life and scientific career was pre-maturely ended in World War II.

TABLE OF CONTENTS

LIST OF FIGURES

LIST OF TABLES

ACKNOWLEDGEMENTS

The word for thesis advisor or research advisor in german is "Doktorvater", which literally translates to doctor father. This word has a strong meaning in the scientific way - the Doktorvater helps the student by supplying valuable ideas and guides him to successfully complete the thesis. In this sense, Peter Wolynes is the best Doktorvater I could ask for. His continous supply of ideas and amazing scientific guidance were as much useful to me as his inexhaustible energy, patience, and support. The word Doktorvater also contains the word father. I would also like to thank Peter for his moral support, his strong believe in me, his effort to make me a better scientist, and the fact that he was always there for me, when I needed someone to talk.

I would also like to thank the current and past Wolynes group members for not only the fruitful scientific interactions, but also for great times. These include Michael Eastwood, Jin Wang, Rachel Small, Joseph Hegler, Patrick Weinkam, Michael Prentiss, Samuel Cho, Garegin Papoian, Ludovico Sutto, Diego Ferreiro, and Tracy Hogan.

Further I need to thank two incredibly intelligent and wonderful women, Dr. Betsy Komives and Dr. Katja Lindenberg. I want to thank Betsy not only for the valuable training in biochemistry and NMR, but also for countless positive interactions and great memories. I also need to thank Katja for wonderful lunches and beautiful conversations. I want to remind Katja not to forget the infinite offer of beer and lunch.

I would further like to thank my comittee member Dr. Andrew McCammon for his kindness to allow me to sit in his group meetings and Dr. Gourisankar Ghosh for his help and suggestions in the IκB/NF-κB project. I also want to remind Dr. José Onuchic that he has to come to my wedding and drink and celebrate with me (this was agreed on in germany in 2006). I also want to congratulate Dr. José Onuchic for his involvement in the Center of Theoretical Biological Physics and I want to thank for the supplied computational resources.

Finally I would like to thank my dearest friends which include Marc Bender, Frank Marquardt, Oliver Kienen, Gregor Schäfer, Benedict, Ritu Jagir, Adolfo Velazquez and Nancy Lee, the Elson-Schwab's, Andrej Grkovich, Mike and Cindy Hale, Howard and Yukako Jan, Jason Miller, the Jagir family, Talib Davis, Nate Asaro, and many more.

I would also like to thank the American Institute of Physics which kindly permitted reprint of the article from Lätzer J, Eastwood MP, Wolynes PG, JOURNAL OF CHEMICAL PHYSICS 125 (21): Art. No. 214905 DEC 7 2006, Copyright 2006, American Institute of Physics as chapter 2. Further I would like to thank Elsevier for the right to publish the article from Joachim Lätzer, Garegin A. Papoian, Michael C. Prentiss, Elizabeth A. Komives and Peter G. Wolynes, "Induced fit, folding, and recognition of the NF-κB-nuclear localization signals by IκBα and IκBβ.", Journal of Molecular Biology (2007, Vol. 367(1),262-274), which appears in this thesis as chapter 4.

VITA

December 23, 1976	Born, Düsseldorf, Germany
1999	Batchelor of Science, Physics, University of Kent, Canterbury, U. K.
2001	Master of Science, Physics, University of Kent, Canterbury, U. K.
2004	Master of Science, Chemistry, University of California, San Diego
2007	Doctor of Philosophy, Chemistry, University of California, San Diego

PUBLICATIONS

Joachim Lätzer, Michael P. Eastwood, and Peter G. Wolynes. "Simulation studies of the fidelity of biomolecular structure ensemble recreation.", Journal of Chemical Physics, (2006, Vol. 125(21), 214905).

Joachim Lätzer, Garegin A. Papoian, Michael C. Prentiss, Elizabeth A. Komives and Peter G. Wolynes. "Induced fit, folding, and recognition of the NF-κB-nuclear localization signals by IκBα and IκBβ.", Journal of Molecular Biology (2007, Vol. 367(1),262-274).

ABSTRACT OF THE DISSERTATION

Computational Tools based on Energy Landscape Theory to Predict Structurally Diverse Ensembles of Transcription Factors

by

Joachim Lätzer

Doctor of Philosophy in Chemistry

University of California San Diego, 2007

Professor Peter G. Wolynes, Chair

The NF-κB/IκB system provides a challenge to the structure-function paradigm since both binding partners are partially disordered in the monomeric form. The experimental study of this system gives rise to many interesting questions. Can one describe the kinetics of coupled folding and binding? Can one faithfully invert the available low resolution data for partially folded ensembles to provide a picture of the underlying molecular details? What happens upon phosphorylation? In this thesis I show how tools to adequately answer these questions can be obtained using energy landscape theory . These tools are validated on test systems of transcription factors where experimental data are available. I demonstrate that replica simulation algorithms based on a strict Bayesian interpretation of the data can successfully invert low resolution data into the correct partially folded ensembles. In order to study the kinetics of the NF-κB/IκB system I also show that simulations with an energy function that yields a funneled but rugged energy landscape can predict the observed binding mode of the crystal structure as well as an alternative binding mode. A method for computing the frustration of partially structured ensembles is also presented. Finally I present an energy function that can predict phosphorylation induced conformational changes for the NtrC transcritption factor.

1 Introduction

The aim of my work has been to develop useful tools based on energy landscape theory to study transcription factor systems. The NF-κB/IκB system is an exciting model system for exploring the coupled folding and binding processes of partially disordered proteins that may be much more common in biology then is commonly believed today. A plethora of interesting experimental results exist. For example, in the X-ray crystal structure of NF-κB bound to DNA the nuclear localization signal of NF-κB lacks electron density. Yet when NF-κB is bound to its inhibitor IκB, the nuclear localization signal is found to exhibit helical secondary structure. This result suggests coupled folding of the NF-κB nuclear localization signal upon binding to the specific inhibitor. Amide exchange experiments, on the other hand, show that regions of the inhibitor fold upon binding to NF-κB. The coupled folding and binding interactions in the NF-κB/IκB system are therefore complex and raise several questions for theoreticans. How can one characterize such diverse partially folded ensembles, when only low resolution structural data from the amide experiments are available?

In the first two chapters of this thesis I describe a novel method to deduce partially folded structures from low resolution data. This method is based on a strict Bayesian interpretation of the experimental input data (here taken to be transition state ϕ-values for illustration) including the known statistical and modeling uncertainties in those data. The experimental constraints act as interactions between the replicas representing the different conformers. Energy landscape ideas were used to fix the magnitude of these interactions in an

objective way based on the magnitude of the errors and the landscape yielded by the typical physical energy function used. The algorithm was first tested for validity on a completely known reference system of the λ-repressor (which is also involved in transcription) rather than applying it directly to the NF-κB/IκB system. Without such prior testing we would not be able to assure that we have a valid tool to invert low resolution data of partially folded ensembles.

While the inversion algorithm is able to deduce the diverse structural ensembles, it cannot directly be used to infer the kinetics of the coupled folding and binding. It is therefore of utmost importance to have energy functions for molecular dynamics simulations that can predict in molecular detail the folded structures of the encounter complex. The use of the Associative Memory Hamiltonian (AMH) for such studies is validated, which should then allow us to study the kinetics of the system. Interestingly, simulation studies of the coupled folding and binding of the nuclear localization signal to the inhibitor do not only reveal the X-ray crystal structure binding mode, but also suggest an alternative binding site. Does this binding site stem from frustration? The AMH is an energy function designed on the principle of minimal frustration yielding a funneled but rugged landscape. Ruggedness is a clear sign of frustration that arises from non-native contacts. I examine this question in the context of another system, IM7. The energy landscape theory acknowledges and highlights the fact that many partially folded structures have such frustrated interactions. A specifically structured long lasting non-native intermediate, such as the intermediate observed in the IM7 folding, is rare. To be able to find out whether the encountered structures are stabilized by minimally frustrated interactions, a method in collaboration with Dr. Sutto and Dr. Hegler was designed to quantify the frustration of all sites. This method was tested to explain the long-lasting non-native intermediate observed in IM7 folding.

Finally I want to note that phosphorylation is often a dominant control mechanism for transcription factors. To address this issue and enlarge the set

of tools for describing transcription factors, one needs a Hamiltonian that can reliably predict the phosphorylated conformation of a protein using information about the unphosphorylated conformation. I propose such a Hamiltonian and test it on the NtrC transcription factor and cystatin.

2 Simulation Studies of the Fidelity of Biomolecular Structure Ensemble Re-creation

2.1 Introduction

The most studied proteins in the cell fold to a reasonably well-defined, average native conformation. The fact that folding times of proteins are relatively short when compared to the time needed for the protein chain to explore all its possible conformations, leads to the conclusion that the protein must be guided towards the native state. The contacts formed in this native conformation must on average be more stabilizing than random contacts allowing the protein molecule to fold to the native conformation by trading entropy for energy. This principle of minimal frustration[1] captures the essential physics of the folding of naturally evolved proteins. The energy landscape of protein folding for proteins that fold reliably therefore resembles a rough funnel[2]. Energy landscape theory describes the folding process down the funnel as a progressive organization of ensembles of partially folded structures[3]. For 2-state folders owing to uneven compensation of entropy loss by stability gain, there is a bottleneck in the flow between the folded and unfolded minima in the free energy which represents the transition state. In the energy landscape ensemble view, the transition state is best described as an ensemble of configurations rather than a single structure[4].

Many experimental techniques have been developed to infer structural information about the structural ensembles for incompletely structured proteins along the folding funnel. With the exception of single molecule studies, those experiments that do provide structural information along the folding funnel typically provide only ensemble averaged quantities. For long lived intermediates, these measured averages directly include NMR parameters and FRET distances, and sometimes structural averages can indirectly be inferred through H/D exchange profiles which are, however, intrinsically kinetic. Using the assumption of a funneled landscape, similar information can often be obtained for the fleeting transition state. The protein-engineering method[5] developed by Fersht and co-workers provides (for smooth landscapes) structural information about the transition state ensemble analogous in many respects to NMR data obtained for long-lived intermediates. This approach assigns a ϕ_i^{exp}-value to each residue. The ϕ_i^{exp}-value is defined as the ratio of the change of the apparent free energy difference between the transition state ensemble and unfolded state ensemble upon a conservative mutation of the residue i to the change in free energy between the native and unfolded ensembles free energy with the same mutation. A ϕ_i^{exp}-value of unity for a residue would indicate that the changes in free energy made by this residue in the transition state are the same as the changes in the native state, whereas a ϕ_i^{exp}-value of zero would indicate that this residue has no native-like interactions. Assuming that the native contacts in the protein alone account for the stabilizing interactions[6], a ϕ_i^{exp}-value can then be approximated as the fraction of native contacts made[7] and this averaged structural quantity can be used as a restraint in molecular dynamics simulations[8, 9, 10, 11, 12]. Technically, this identification is only valid for a perfectly homogeneous funnel landscape. Defining a contact distance R_C for interacting amino acids, the determined ϕ-values can then be used as constraints on the ensemble of protein structures, requiring each residue to form a fraction of its native contacts to within an upper distance bound R_C. The measured constraints however do not enforce a precise distance for two residues in contact. This raises the question, whether (or when) ensembles deduced from the ϕ-value constraints are structurally

equivalent to the actual ensemble probed by the experiment. That is, can the real ensemble be faithfully recreated from experimental data alone? Also do some algorithms give greater fidelity in reconstruction than do others? In particular, the experimentally derived restraints may be applied equally to every structure encountered on a single MD trajectory (the "single replica" case); alternatively a multiple replica algorithm may be used where the restraints are applied to the ensemble of structures observed in a number of simultaneous MD simulations thereby allowing individual replicas to have fluctuations while restraining the ensemble average. Davis et al. [13] have already shown, that in the case of the β3s peptide two replicas were required to correctly predict the transition state structures from the ensemble-average set of ϕ-values. This result encourages examining more quantitatively the benefits of using multiple replica algorithms.

Here we present an extension of the multiple replica approach to the simultaneous determination of a transition state ensemble. We test this approach by attempting the re-creation of a completely known, candidate transition state ensemble of 500 structures of the λ-repressor protein. First we create several surrogates for the "experimental transition state ensembles", which we shall term reference ensembles, sampled from simulations using native structure based[14] Hamiltonians for the λ-repressor both with and without nonpairwise-additive interaction terms. From the reference ensembles we calculate the average ϕ-value for each residue. These computed ϕ-values then serve as surrogate experimental constraints for the replica simulation algorithm. Single and multiple replica molecular dynamics simulations are then performed with a Hamiltonian that biases the ensembles to match the experimental ϕ-value constraints but otherwise has no a priori biases. Since the structures of the reference ensembles are known, the success of re-creating the original ensemble using the multiple replica algorithm can be rigorously evaluated with a statistical test, the Kolmogorov-Smirnov test[15]. In the KS-test two ensemble distributions, one given by the reference ensembles, the other given by the ensembles obtained in the replica molecular dynamics simulations, are compared

and it is tested whether these two distributions are substantially the same. When the two ensembles differ, a method can be used to uncover possible matching subensembles from the ensembles obtained in the replica molecular dynamics simulations. These subensembles can be obtained by clustering the structures and selecting the most dominant cluster as a representative ensemble. These representative subensembles can then also be compared to the reference ensembles. To study whether multiple replica re-creation methods are more faithful than single copy approaches, we first analyze the principal components of the contact maps of all structures in the reference ensembles and in the ensembles obtained in the replica simulations. The principal component analysis indicates that the sampling is improved, when multiple replicas are introduced. Finally we probe the robustness of the ensemble recovery to realistic uncertainty in the input data. Experimental quantities always have errors associated with them. To mimic these errors, we assigned new ϕ-values for each residue by generating a random number drawn from a Gaussian distribution with its maximum located at the original ϕ-value of that residue and with a variance given by the variance of that ϕ-value in the reference ensemble. The new set of ϕ-values then served as input for the replica Hamiltonian ensemble reconstruction. The ensembles obtained from the replica algorithm with the new set of ϕ-values as experimental constraints were compared to the original reference ensemble. This procedure quantitatively probes how large errors in ϕ can substantially reduce the chances of faithful ensemble re-creation using replica simulation algorithms.

2.2 Methods

2.2.1 Reference Ensemble Creation

In order to rigorously test the fidelity of a reconstruction procedure a well characterized reference ensemble must first be available. An off-lattice simulation with a native structure based Hamiltonian[16] with variable strength

non-additive terms as described in detail earlier[14] was chosen to provide such reference ensembles for the reconstruction procedure. These reference ensembles are considered "gold standard" ensembles and represent the ensembles that experiments strive to determine. The energy function used to obtain these ensembles is given as the sum of a native structured based but non-additive Hamiltonian H_{na} and standard backbone energy terms

$$H = H_{backbone} + H_{na} \tag{2.1}$$

This energy function applies to a reduced set of coordinates of the heavy backbone atoms, C^{α}, C^{β} and O. In this reduced description, the positions of the nitrogen and $C^{'}$ carbons can be calculated assuming ideal protein backbone geometry. The backbone potential takes on the following form

$$H_{backbone} = \lambda_{\psi\phi} V_{\psi\phi} + \lambda_{\chi} V_{\chi} + \lambda_{ex} V_{ex} + \lambda_{harm} V_{harm} \tag{2.2}$$

The backbone terms[17] in the Hamiltonian ensure the backbone has physically allowable conformations. The planarity of the peptide bond is constrained by the SHAKE algorithm and three simple harmonic potentials, V_{harm}, which restrain the Nitrogen-C^{β}, Nitrogen-$C^{'}$ and $C^{'}$-C^{β} distances close to $2.46\mathring{A}$, $2.45\mathring{A}$ and $2.51\mathring{A}$ respectively. A chirality potential V_{χ} biases the C^{α} atoms towards the L configuration which is preferred in nature. The ϕ and ψ angles of the protein backbone are biased with the Ramachandran potential $V_{\psi\phi}$. This potential biases the torsional angles of the protein to regions allowable for a naturally occuring protein. The barriers between minima of the Ramachandran potential are intentionally set low to facilitate more rapid chain dynamics. Excluded volume effects are included between C^{α}-C^{α}, C^{α}-C^{β}, C^{β}-C^{β} and O-O pairs through the V_{ex} potential. The individual λ-parameters in the backbone Hamiltonian scale the interactions to physically reasonable values.

The H_{na} energy depends on Gaussian interaction terms for native contact pairs only. H_{na} is given as a function of pairwise energy terms raised to the power p.

$$H_{na} = -\frac{1}{2} \sum_{i} |E_i|^p \tag{2.3}$$

The parameter p in the Hamiltonian is the power of non-additivity and introduces (p+1)-body interactions as well as p,p-1,...,2-body interactions with range $r_C = 8.0\mathring{A}$. Usually increasing p results in additional cooperativity and hence in increased barrier heights for folding. The individual pairwise energy terms can be written in a normalized form containing a cut-off distance r_c.

$$\begin{aligned} E_i &= \sum_j \epsilon_{ij}(r_{ij}) \\ &= -\sum_j \left|\frac{\epsilon}{a}\right|^{\frac{1}{p}} \theta(r_c - r_{ij}^N)\gamma_{ij} \exp\left(-\frac{(r_{ij} - r_{ij}^N)^2}{2\sigma_{ij}^2}\right) \end{aligned} \tag{2.4}$$

The contribution of H_{na} to the native state energy of a protein with N residues is by definition $4N\epsilon$. This is ensured if the normalization constant a is defined as

$$a = \frac{1}{8N}\sum_i \left|\sum_j \gamma_{ij}\theta(r_c - r_{ij}^N)\right|^p \tag{2.5}$$

The weighting function γ and the well width σ depend on the sequence separation of residues i and j and are chosen such that the energy of the ground-state energy for p=1 at a cut-off distance of $r_c = 8\mathring{A}$ is evenly divided between short ($|i-j| < 5$) and long range interactions in sequence space as suggested by the analysis of Saven and Wolynes for helical proteins[18]. The parameters are

$$\begin{aligned} \sigma_{ij} &= |i-j|^{0.15}\mathring{A} \\ \gamma_{ij} &= \begin{cases} 0.125 & |i-j| < 5, \\ 0.5 & \text{otherwise} \end{cases} \end{aligned} \tag{2.6}$$

The total Hamiltonian described above can be used to infer the thermodynamic properties of a given system. To obtain useful free energy profiles, a proper reaction coordinate has to be chosen. One appropriate coordinate is Q, a measure of native-likeness

$$Q = \frac{2}{(N-1)(N-2)}\sum_{i<j-1} \exp\left(-\frac{(r_{ij} - r_{ij}^N)^2}{2\sigma_{ij}^2}\right) \tag{2.7}$$

Q is a normalized quantity that describes structural similarity of a given structure with coordinate set $\{r_{ij}\}$ to a reference structure, for folding and structure prediction usally the native structure, with coordinates $\{r_{ij}^N\}$. Free energy profiles were then obtained with the weighted histogram analysis method (WHAM) with umbrella sampling. 17 constant temperature molecular dynamics simulations were performed with a biasing potential that is a polynomial in Q of 4th order centered on different values of Q (Q_0=0.9, 0.85, 0.8,$\cdot$,0.1) to obtain good phase-space sampling along this reaction coordinate. The Q-constraint in the potential is sequentially reduced from Q=0.9, which is almost native-like, to Q=0.1. This procedure reduces the equilibration time of the system. During each of these constant temperature molecular dynamics simulations, 200 independent samples, N_s^{obs}, of Q and energy E, the backbone and H_{na} energy, were collected at regularly spaced time steps. These time steps were larger than the correlation time between sampled structures. The samples therby obtained were independent of earlier configurations sampled. The first 40 samples of each simulation run were discarded to help ensure that the system reached equilibrium, before samples were entered into the free energy calculation. A histogram $N_s(E,Q)$ for all 17 simulations was created. The density of states n(E,Q) of the system (Eastwood et al., 2001) was calculated from the histograms

$$n(E,Q) = \sum_s w_s(E,Q) \frac{N_s(E,Q)}{N_s^{obs}} Z_s(\beta_s) \exp\left(\beta_s(V_s(Q)+E)\right) \tag{2.8}$$

Here s labels the simulation and w represents a weighting function defined as

$$\begin{aligned} w_i &= \frac{A_s^{-2}}{\sum_m A_m^{-2}} \\ A_s^{-2} &= \frac{n(E,Q)}{N_s^{obs}} Z_s(\beta_s) \exp\left(\beta_s(V_s(Q)+E)\right) \end{aligned} \tag{2.9}$$

The density of states and the weighting function are functions of the partition function Z_s. The partition function, on the other hand, is also a function of the density of states

$$Z_s(\beta_s) = \sum_{E,Q} n(E,Q) \exp\left(-\beta_s(V_s(Q)+E)\right) \tag{2.10}$$

This set of equations can be used to obtain for n(E,Q) self-consistently to within a multiplicative constant and hence the free energy was obtained to within a constant as

$$F(Q,T) = -k_B T \log\left(\sum_{E,Q} n(E,Q) \exp\left(-\frac{E}{k_B T}\right)\right) \tag{2.11}$$

The free energy profile at folding temperature T_f can be inspected and ensembles for the denatured state, the transition state or any other reference state of choice, can be found by Q. Structures with the appropriate Q-value entered into the reference ensemble.

2.2.2 ϕ-Value Molecular Dynamics Replica Simulation Technique and Details

Given a set of experimental ϕ-values $\{\langle\phi_i\rangle_{exp}\}$ for the residues of the protein, we can write down a replica Hamiltonian that constrains ensemble averages to the values provided by experimental measurements for each residue. The simplest form of the replica Hamiltonian contains standard backbone terms as described above while adding the experimental biasing potential. Optionally other energy terms, H_{funnel}, that vary the protein energy landscape and encode prior theoretical expectations can also be included. In this paper H_{funnel} will be set to 0.

$$H_{rep} = H_{back} + H_{funnel} + \sum_{i=1}^{N} \lambda_i (\overline{\phi_i} - < \phi_i >_{exp})^2 \tag{2.12}$$

with N being the total number of residues. The ensemble average ϕ-value $\overline{\phi_i}$ is the arithmetic average over the realizations of the individual replicas.

$$\overline{\phi_i} = \frac{1}{N_{rep}} \sum_{\mu=1}^{N_{rep}} \phi_i^{\mu} \tag{2.13}$$

where N_{rep} is the number of simulated replicas. To perform molecular dynamics simulations, a recipe to calculate ϕ from the observed contacts must be

given. Although ϕ is a dynamical quantity measured from the ratio of thermodynamic and kinetic quantities, an often used surrogate for ϕ is the ratio of native contacts made divided by the maximum number of native contacts possible. This surrogate, of course, assumes the landscape is, in fact, reasonably funneled[19]. An explicit equation for ϕ_i^μ in terms of a contact function c_{ij} is

$$\phi_i^\mu = \frac{1}{N_{cont}^i} \sum_{<j>} c_{ij} = \frac{1}{N_{cont}^i} \sum_{<j>} \frac{1}{2}\big(1 + \tanh\left(5\left(r_C - r_{ij}\right)\right)\big) \tag{2.14}$$

The contact function considers native contacts to be formed only if they reside within some cut-off distance r_C. The cut-off distance for C^β contacts usually lies in the region of $6.5 - 8.5\mathring{A}$. In the present study cut-off distances of 6.5Å and 8.0Å have been used. We only present results for a cut-off distance of 6.5Å. The definition of the set of contacts for the completely native structure depends on the value of the cut-off distance between the C^β's. Once a value for the cut-off distance is chosen, the appropriateness of this value for defining contacts can be checked for consistency with other methods of assigning contacts such as the CSU algorithm which rely on all-atom structural information. The functional form of the contact function is a tanh function, whose continous nature prevents numerical errors in the dynamics.

The simulation scheme used is as follows: Constant temperature molecular dynamics simulations are performed at three temperatures, $T_F, 0.25T_F$ and $1.75T_F$ for 1,2,4 and 8 replicas. The folding temperature T_F corresponds to the "physiological" transition temperature for folding of the non-additive Gō-like energy function described above. The simplicity of the model allows extensive sampling to be done. It is straightforward to employ simulations of length of the order of 1ms. This time scale ensures enough sampling to compensate for topological traps in the energy landscapes. The results are checked to ensure they converged. The simulations involve different numbers of replicas, but the total number of sampled conformations is kept constant between simulation runs with different numbers of replicas. The ensembles can now be fairly compared.

We first test to make sure that the input ϕ-values are reproduced. Next a statistical test is used to decide whether ensembles generated from the replica algorithm differ from the reference emsemble or not. An appropriate statistical test for comparing ensemble distributions is the the Kolmogorov-Smirnov (KS) test. The KS-test quantifies whether two distributions differ from each other in a statistically significant way. To apply the KS-test, the ensembles are first reduced to distributions that are functions of only a single, independent variable. This single independent variable is chosen to be a structural overlap measure,q, defined analogously to Q, but where all q's of the structures in the ensemble are measured relative to each other rather than measured to one single reference structure. The KS-test requires calculation of distributions of q for all pairs of structures within the simulated ensemble ($P_B(q)$), all pairs within the reference ensemble ($P_A(q)$) and all pairs with one member chosen from each of the the two ensembles ($P_{AB}(q)$). The KS-test is then performed on the individual distributions, which tests if two distributions are statistically identical, typically we compare $P_A(q)$ with $P_{AB}(q)$. In our case where we have a large amount of data when comparing two different Hamiltonians, the result of the test indicates that the two distributions are not exactly the same. However, the KS-statistics itself provides a very useful measure for quantifying the magnitude of the difference, and simply visualizing the difference between the distributions is illuminating.

2.2.3 Principal component analysis

Contact maps for the reference ensemble and the ensembles obtained from the replica simulations were computed for all indivudual structures. Principal component analysis (PCA) of the binary contact degrees of freedom for these ensemble structures was performed [20]. The PCA we employ is not the more commonly used PCA based on Cartesian coordinates. The more commonly used PCA is based on the diagonalization of the Cartesian coordinates. This is less useful in the current problem due to the fact that the transition state ensembles generally show large anharmonic conformational differences

that go beyond simple vibrational-like fluctuations of Cartesian coordinates. This approach uses a very coarse-grained degree of freedom: the contact map, which is the simplest site specific measure of a folding progress. To facilitate the analyses, we further coarse-grained the contacts by grouping neighboring residues into groups of three residues, i.e., a coarse-grained contact matrix is calculated for each structure, with each of those independent elements either being 0 or 1. The contacts are reduced to $27 \times (27-1)/2 = 378)$ elements that are either 0 or 1. The resulting reduced covariance matrix of dimension 378×378 is diagonalized and the eigenvalues for the contact map PCA are calculated. The two most dominant principal components are plotted.

2.2.4 Structural clustering analysis

The Fitch-Margoliash algorithm[21] is a distance based bioinformatic algorithm to fit a phylogenetic tree to a distance matrix. The numerous structures obtained from the simulation runs were clustered using the FITCH program of the PHYLIP package[22]. The FITCH program can be used to create phylogenetic trees based on any given distance measure. In order to analyze the structures obtained in the simulated annealing with the bioinformatic software, a topology based distance measure d between two structures A and B was defined through the structural overlap q as $d = 1-q$. The order parameter q represents the relative similarity of two stuctures and is defined analogously to Q. Since q is a normalized measure of the fraction of overlapping contacts, d is a measure of how dissimilar two structures are in terms of their contacts. Similar structures with small d are close and dissimilar structures with large d are structurally far away. Since q, unlike ϕ, is sensitive to the correct distance between residues rather than just constraining two residues to be within a cut-off distance, the clustering should group all structures based on their local secondary structure as well as their global tertiary structure. This clustering technique helps in extracting subensembles with more narrowly defined local structures.

2.3 Results for the Transition State Ensemble of the λ-repressor

We first present the complete results of ensemble recovery with replicas for the transition state ensemble of the λ-repressor. The λ-repressor is a well studied DNA-binding regulatory protein with a four helix bundle fold. Experimental data suggest that the λ-repressor is a 2-state folder with a low barrier between folded and unfolded basins [23] . To test the inversion algorithm, reference ensembles for the λ-repressor (pdb code 1lmb [24]) were first generated with the non-additive Gō-like Hamiltonian using Q as a reaction coordinate. Although there has been controversy about the merits of Q as a reaction coordinate [25], this controversy is irrelevant for our present purpose of obtaining reference transition-state ensembles which are only to be used as test-beds for studying the reproduction of an ensemble from its averaged properties alone. To assure the reader of the validity of the ensembles obtained with this Hamiltonian, we note that it has already been tested whether these Hamiltonians with many-body interactions produce ensembles that resemble ensembles one would measure in experiments. This has been done by testing the correlation of experimental ϕ-values and experimental folding rates to the ϕ-values and folding rates obtained with the non-additive Hamiltonian. It has been shown that great agreement with experiment can be reached, when the fraction of energy arising from three-body terms in the native state is approximately 20% [124]. A power of nonadditivity in the range p = 2 to 3 for our Hamiltonian should lead to nonadditive contributions to the energy of a magnitude found in real proteins[14]. WHAM simulations with umbrella sampling were performed at the putative folding temperature $\frac{k_B T}{\epsilon} = 1.0$.

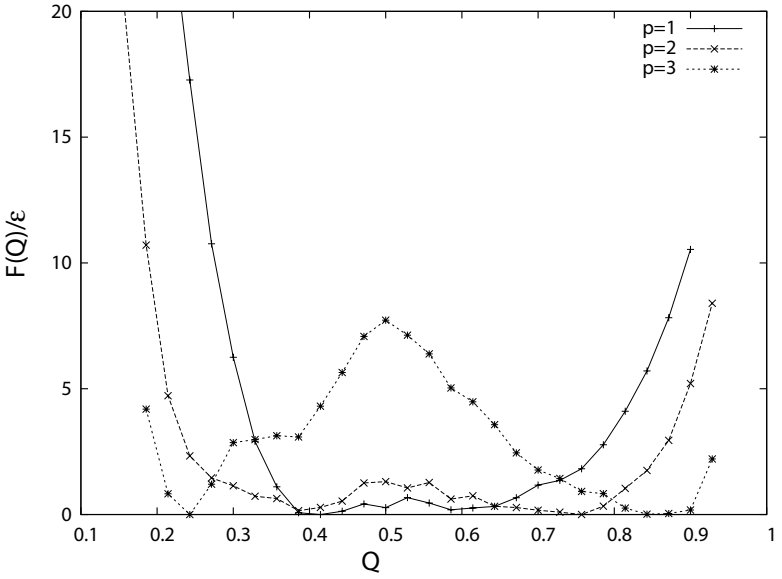
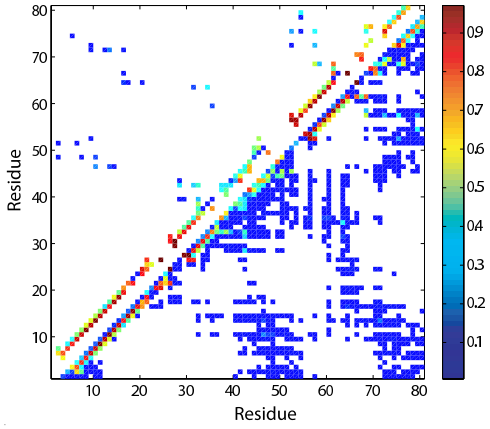


Figure 2.1: Free energy profiles at folding temperature T_F as a function of Q obtained with WHAM and umbrella sampling with the non-additive Hamiltonian. The free energy shows 2-state behaviour with increasing barrier for increasing p, the power of non-additivity. The free energy profile shows a transition state ensemble at Q≈0.5.

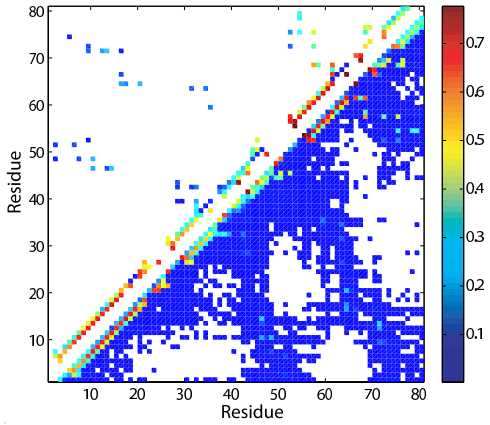
The free energy profiles were then calculated as described in the methods section. To obtain a more refined estimate of the folding temperatures the free energy profiles were extrapolated to nearby temperatures to find the temperature where the depths of the folded and unfolded basins coincide. Further WHAM simulations with umbrella sampling were then performed at this new temperature. The free energy profiles were calculated with the new data yielding a more accutate estimate of T_F. This procedure was repeated until convergence, which in practice occured after only two rounds of WHAM simulations. The free energy profiles of the λ-repressor at $T_F = 0.97, 1.02$ and 0.93 for a classical native structure based Hamiltonian with p=1, and for Gō-like Hamiltonians with many-body effects for p=2,3 respectively are shown in Figure 2.1. The free energy curves exhibit a two-state folding character with a barrier between the folded and unfolded states, that increases with increasing parameter p. For p=1,2 the barrier between unfolded and folded basins is roughly $\sim 1k_BT_F$ indicating the weak cooperative effects present. Introduction of 4-body terms corresponding to p=3 increases the barrier roughly ten-fold to about $\sim 8k_BT_F$. Various reference ensembles are then read off the free energy profile. The Q-score of the transition state ensemble was determined at the maximum value of the free energy curve between the folded and unfolded basin. The transition state ensembles had a Q-score of about $Q = 0.5$. Approximately 500 independently sampled structures were chosen for each transition state ensemble for p=1,2 and 3 to represent the reference ensemble. The contact maps for the ensembles obtained with p=1 and p=3 are shown in Figure 2.2. Contact maps of native contacts only are shown above the diagonal whereas both native and non-native contacts are shown below the diagonal. From the native contact maps it is apparent that the C-terminal and the long N-terminal helix are most ordered in the transition state ensemble. The DNA binding site (pdb residues 34-54) is most disordered in the transition state. The reference ensembles also show several non-native contacts with low contact probability in the ensemble.

The inversion algorithm derives structures only from the input ϕ-values,

which are calculated from native contacts. The ϕ-value is defined as a fraction of native contacts, where contacts are defined to fall within a certain cut-off distance. The inversion of such data might then be not accurate on the more local level due to the lack of secondary structure information in the inversion Hamiltonian and lack of knowledge of low-probability non-native contacts.



(a)



(b)

Figure 2.2: Native contact map only (above diagonal) and complete contact map (below diagonal) of the transition state reference ensembles of the λ-repressor for p=1 (a) and p=3 (b) averaged over all structures. A contact is defined when the distance of the C_β carbons are within 6.5$\mathring{A}$. A dark red contact corresponds to a contact that is on always formed in the transition state ensemble. A dark blue contact, on the other hand, is never formed in the transition state ensemble.

A set of ϕ-values denoted $\{< \phi_i >_{exp}\}$ and the corresponding (in this case, statistical) error $\delta\phi_i$ was calculated for each of the reference ensembles. The native structure of the protein, the $\{< \phi_i >_{exp}\}$ and (in some cases) the statistical errors were the only data used to infer the transition state ensembles. The simplest energy function, that can be used to recover ensembles from the given information, is a Hamiltonian which reproduces the given ensemble averaged constraints but that has no knowledge of the energy landscape of folding of the protein. Such a basic Hamiltonian is given in equation 2.12 with $H_{funnel} = 0$. The only unknown parameter in the Hamiltonian is the strength of interaction of the experimental restraint, the parameter λ. To approximately determine the parameter λ, successive ϕ-value simulations with gradually increasing values of λ were performed to set a uniform λ-value for all residues such that the experimental constraints are fullfilled. Comparison of the experimental ϕ-values to the simulation ϕ-values showed a consistent match with high correlation for different number of replicas (data not shown). This indicates that the restraints in the simulation are strong enough that ensembles with the correct ϕ-values for each residue are indeed produced. We note that the results for p=2 are very similar to those found for the p=1 case and discussion of the p=2 results will therefore be omitted. The individual ensembles obtained from the replica simulations each consisted of 1600 independently sampled structures taken from millisecond long molecular dynamics trajectories. From the structures found in the simulations with replicas and the reference ensembles probability distributions of the single independent variable q, the structural overlap reaction coordinate may be extracted. A statistical test can be performed to check whether the ensembles obtained from the ϕ-value molecular dynamics replica simulations for 1-8 replicas can be considered aprt from incomplete sampling identical to the reference ensemble which was used to generate the input ϕ values. Figure 2.3 shows the results of the KS-test for the various transition state ensembles.

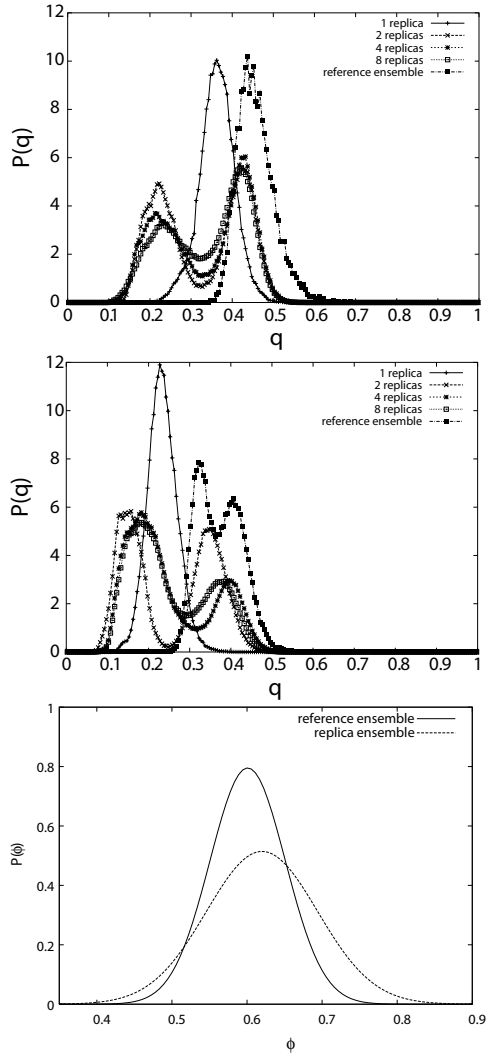


Figure 2.3: Shown are the overlap distributions at the folding temperature of the ensembles obtained from replica simulations with the reference ensemble, for the three different transition state ensembles with $p = 1$ (a) and $p = 3$ (b). The self-overlap distribution of the reference ensemble is also shown. Tthe probability distribution of the average ϕ-value of each individual realization in the recovered ensemble with eight replicas is plotted in (c)

Two structural ensembles are equal (or there is an absence of evidence that they differ) if the probability distribution of pairwise overlaps, $P(q)$, is the same irrespective of whether the pairs are drawn from the same ensemble or from distinct ensembles. For the p=1 and p=3 transition state ensembles, the probability distribution of overlaps between the one-replica and reference ensemble has some overlap with the distribution of overlaps within the reference ensemble. However, this overlap between distributions is not large, showing that the reference ensemble and the ensemble obtained from the one-replica simulation are in fact can not be considered the same, although they both have the same set of ϕ-values. It seems that the structural order parameter q since it varies more strongly with the exact distances between amino acid pairs is a more demanding similarity measure than ϕ, which depends only on whether contacts form within a specified cut-off distance. Two residues that are closer than the cut-off distance for a contact but near that limit, contribute strongly to ϕ, but lead to a low q_{ij}-value for that residue pair, if the distance of the residue pair is very near in the reference transition state ensemble. Thus we see that using ϕ-values alone for reconstruction may lead to discrepancies in short-ranged local structural elements such as the α-helical structures of the ensembles obtained with the replica simulation algorithm and the gold standard ensemble. Nevertheless both ϕ and q are adequate order parameters to quantify a conformation and its global fold.

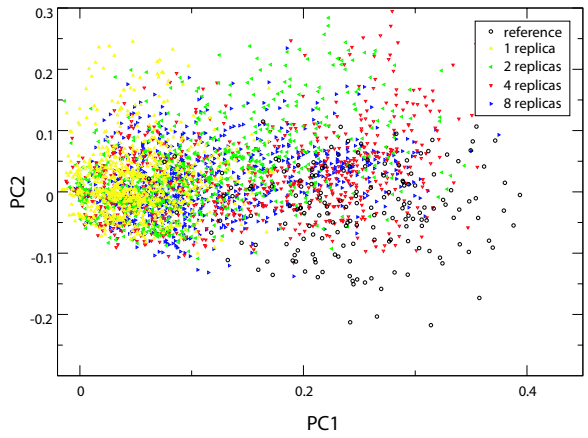
Another question we can address is whether a larger λ-parameter that would reflect the availability of more accurate data would eventually lead to precise reproduction of the reference ensemble. One might argue that increasing the strength of interaction, the λ-parameter, could force the regenerated ensemble to approach the reference ensemble. KS-tests have been performed with increasing value of λ, but they showed no noticeable improvement for the recovery of the reference ensemble with one replica. Apparently the structural imprecision of ϕ also plays a role in determining the fidelity of ensemble recovery. Without knowledge of the energy landscape of folding of the λ-repressor, the KS-test above indicates that one cannot conclude that the one-replica

Hamiltonian will reliably deduce the reference ensemble even though it reproduces the set of experimental ϕ-values. The reference ensemble is only partially reproduced using the one-replica simulation technique. Nevertheless, inversion with one replica can be judged to be partially successful.

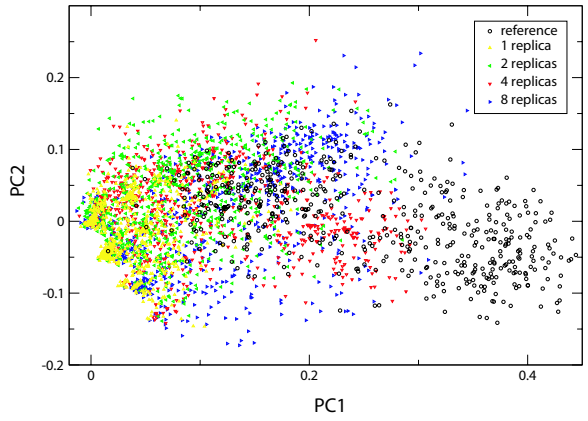
In contrast to the one replica ensemble reproduction the curves of the overlap distributions for multiple replicas are bimodal. To check whether the bimodality of the replicas is an artifact of a recovered ensemble with unfolded and folded structures, that on average match the reference ϕ-values, the probability distribution of the recovered ensemble obtained with the eight replica algorithm is plotted. The average ϕ-value of each snapshot is calculated and the results are binned in bins of size 0.005. The reference ensemble is also plotted for comparison. One of the two peaks of the overlap distribution overlaps well with the reference ensemble probability distribution suggesting that the ensembles are similar. These results suggest that the reference ensemble can be extracted from the ensemble obtained with simulations with multiple replicas. The nature of the bimodal probabilty distribution suggests that the replicas are not homogeneous but instead break the replica symmetry. It is clear from Figure 2.3 that the broken replica symmetry does not stem from a simple division of folded and unfolded structures of the recovered ensemble. The underlying replica symmetry breaking is more subtle. The distribution of the average ϕ-value of each realization (or snapshot) of the recovered ensemble is similar to the distribution of average ϕ-values of the reference ensemble.

2.4 Sampling enhancement through multiple replicas

Principal component analysis of the contact maps of each structure allows a convenient visualization of the patterns of variation in contact probabilities in the subensembles and hence allows the study of the range of conformations of all residues in those structural ensembles. Figure 2.4 displays the conformations of the structures of the reference ensemble and the regenerated ensembles for 1 to 8 replicas projected onto the first two principal components. In the p=1 case the first principal component is a good indication of the sampling of the reference ensemble. The reference ensemble shows a negative first principal component (PC1) with most conformations in the region of $PC1 = -2$ to -4. The projections of the conformations obtained with multiple replicas show much more overlap with the reference ensemble than the projections of the one replica conformations. The recreated ensemble obtained with multiple replicas is substantially shifted towards more negative PC1 when compared to the one replica ensemble. The multiple replica algorithm better samples reference-ensemble-like structures than does the single replica algorithm although the number of independent samples is kept equal between all replica simulation runs of single and multiple copies. To test whether the degree of overlap of each of the multiple replica ensembles with the reference ensemble is artificially high due to the fact that all ensembles including the single replica ensemble enter the PCA, analysis of the individual multiple replica ensembles with the reference ensemble have been performed, which shows very similar results. The advantage of the multiple replica algorithm seems even more apparent for the p=3 reference ensemble. Here the reference ensemble is bimodal as reflected in the results of the principal component analysis (Figure 2.4(b)). The reference ensemble structures projected onto the principal components show two main clusters. While the ensemble obtained with 1 replica shows only small overlap with the reference ensemble located in the PC1 region of less than 2, there is no overlap with the reference ensemble conformations



(a)



(b)

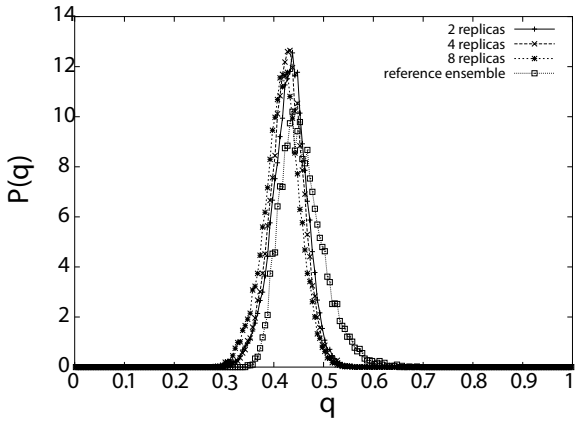
Figure 2.4: The two principal components of conformations found in the reference ensemble and ensembles obtained with the replica algorithm with 1 to 8 copies for the p=1 case (a) and the p=3 case (b) are shown.

projected along the PC1 greater than 2. The PC1 of the ensembles obtained with multiple replicas assume a wider range of PC1 values indicating the better sampling of both clusters of reference ensemble structures. The success of multiple replicas is due to the fact multiplicity of replicas allows fluctuations around the ϕ-values for individual structures, while still constraining the replicas on average to its input ϕ-values. We also projected the first two principal

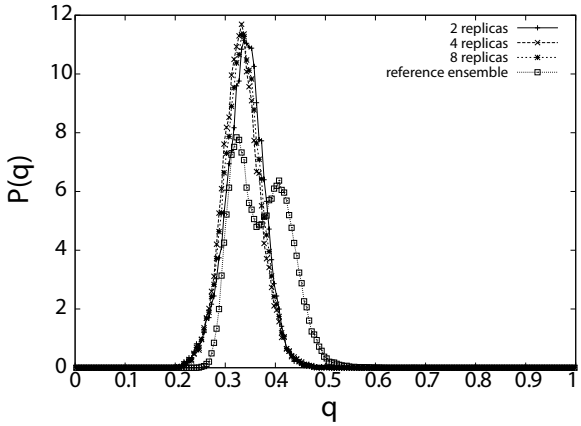
components onto the contact map of the λ-repressor (data not shown). The contact maps are convenient to visualize on a residue by residue contact basis, which residues are more reference ensemble like and which are not. Most contacts that are formed in the reference ensemble are also formed equally in the replica ensembles. Structurally, the main differences between the reference and the replica ensembles can be attributed to the different C-terminal helix contacts.

2.5 Reference Ensemble re-creation through ensemble reduction methods

A powerful adjunct for the re-creation procedure would be to have some kind of selection filter for the structures obtained in a simulation. If a postprocessing tool were to exist that allowed the selection of only those structures that truly resemble the reference ensemble, the somehow usefulness of the inversion procedure would be greatly enhanced.

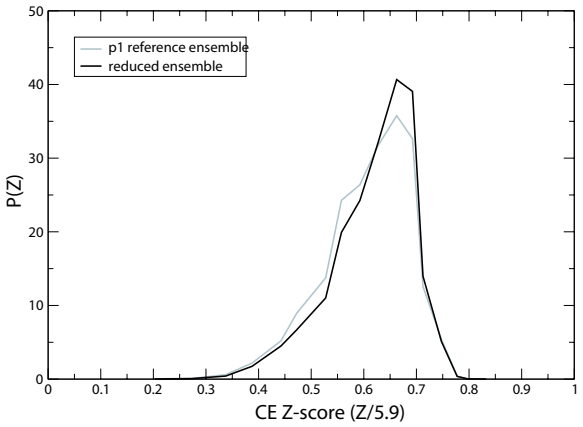


(a)

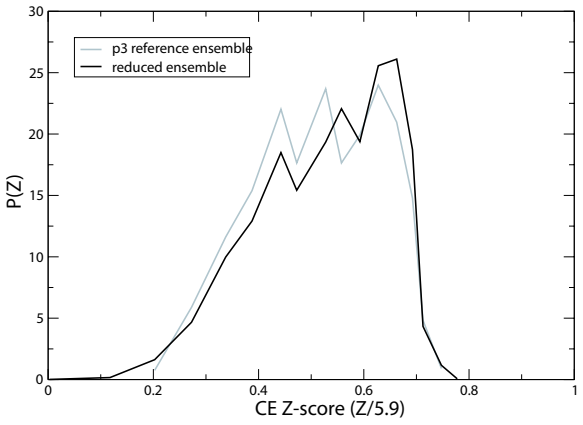


(b)

Figure 2.5: Overlap distributions of the reduced ensembles from the replica simulations for 2,4 and 8 replicas with (a) the p=1 and (b) the p=3 ensemble. The ensembles obtained from the clustering technique described in the method section improved the prediction success of the experimental ensemble for p=1 tremendously as measured by the KS-test. For the p=3 ensemble the reduced ensembles reproduce the reference ensemble as measured by the KS-test partially.



(a)



(b)

Figure 2.6: KS overlap test of the reduced ensemble obtained with eight replicas and their corresponding reference ensemble using the CE Z-score as reaction coordinate for the p=1 (a) and p=3 (b) ensemble.

There are many possible ways of partitioning the ensemble based on the structural diversity. A simple clustering algorithm that clusters structures obtained with the multiple-replica Hamiltonian allows separation of these structures into subensembles. The Fitch-Margoliash clustering algorithm uses a distance measure between all structures to generate a phylogenetic tree. The distance parameter d is given by $d = 1 - q$, where q is a normalized pairwise

measure of similarity of all structures relative to each other. For the p=1 transition state ensemble the phylogenetic tree showed clustering into two main clusters. One cluster contained structures with greater variation of the radius of gyration and less helical content. This cluster was not as homogeneous as the other cluster was. It contained lots of subclusters. The other cluster showed more compact structures with higher helical content. The structures of this cluster are denoted the "reduced ensemble". It was then confirmed, that this ensemble has on average the same set of ϕ-values as the reference ensemble. This is important in validating the choice of the most dominant subensemble as a valid representation of the reference ensemble. If the difference between the average ϕ-values of the reference ensemble and the chosen cluster are large, the cluster can not be accepted as a valid ensemble. However, there was no such difficulty for the most dominant cluster. KS-tests with the reduced ensembles were performed to test whether these ensembles overlapped with the p=1 reference ensemble. The overlap of the reduced ensemble for multiple replicas with the experimental p=1 reference ensemble suggested a successful recovery of the reference transition state ensemble (Figure 2.5(a)). The structures of the reduced ensemble (Figure 2.7(b) , 2.7(d)) exhibited the same global fold with similar disorder in the DNA binding region as the reference ensemble (Figure 2.7(a) , 2.7(c)). All structural comparisons such as RMSD, helical content, radius of gyration, secondary and tertiary structure, and Z-score from the combinatorial extensions algorithm (the CE Z-score) confirmed that these two ensembles are indeed equivalent on the basis of each of these measures.

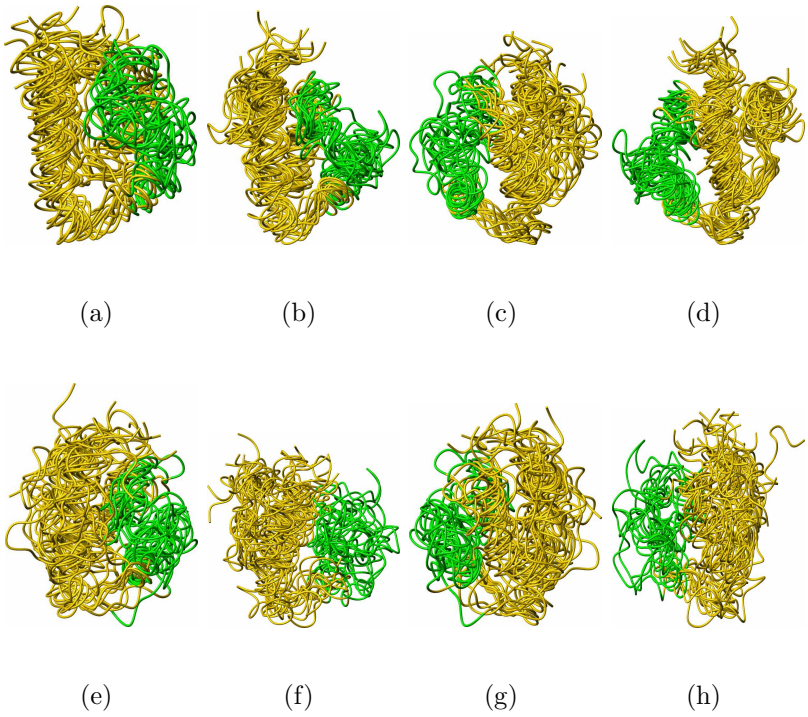


Figure 2.7: Shown are the reference transition state ensembles viewed from front (a,e) and back (c,g) and the reduced ensembles obtained with 8 replicas for p=1,3. The DNA binding region (green) is disordered in the transition state. For the p=1 case the reduced ensemble (b,d) and the reference ensemble (a,c) show the same intrinsic features such as secondary and tertiary structure and their average structure has a relative RMSD of the backbone carbons of less than 2.5$\mathring{A}$. For the p=3 case both of the reduced ensemble overlayed (f,h) show similar structural features than the reference ensemble (e,g). Pictures were made with molmol [26].

The phylogenetic tree was also obtained for structures obtained with the multiple-replica algorithm for the p=3 case. The tree showed a main cluster with a few populated sub-clusters. The structures of the sub-cluster whose average ϕ-values resemble the most the reference ensemble were taken as the reduced ensemble. The KS-test was then performed for the reduced ensemble. The result is shown in Figure 2.5(b). The probability distribution function of q overlap of the p=3 reference transition state ensemble shows a bimodal distribution. The probability distribution of the reduced ensembles overlapped well with one peak of the p=3 reference ensemble probability distribution. The structures found in the lower-q peak of the reference ensemble were compared to the structures of the reduced ensemble. The resultant structures of the replica simulations (Figure 2.7(f) , 2.7(h)) exhibit similar tertiary and secondary structures to that of reference ensemble structures (Figure 2.7(e) , 2.7(g)). Using other order parameters in the KS-test, such as the CE Z-score or RMSD, support the results of the KS-test, that the reduced ensemble and the lower-q reference ensemble are highly similar ensembles. In figure 2.6(a) and 2.6(b) we show the overlap distributions of the reduced ensemble obtained with eight replicas and their corresponding reference ensemble. We note that the CE Z-score is down-scaled by a factor of 5.9, which is the resulting score for the overlap of each of the transition state structures to themselves. This will normalize the CE Z-score axis to facilitate better comparison to the order parameter Q. The p=1 reference ensemble distribution peaks at a Z-score of about $0.67 * 5.9 = 3.95$ and most structures have a Z-score in the range of 3.65-4.19. A Z-score of 3.5 and higher is considered a criterion that the two structures share the same fold. We therefore see that in the p=1 reference transition state ensemble structures and folds are very similar to each other. The KS test using the CE Z-score as a reaction coordinate shows the high overlap of the the probability distributions of the reduced ensemble and the reference ensemble. The conclusion from the Z-score overlap test is that the two ensembles, reference and reduced ensemble, are indeed equivalent ensembles in terms of representing the same distribution of global folds. The p=3 reference ensemble showed a bimodal distribution in the order parameter q

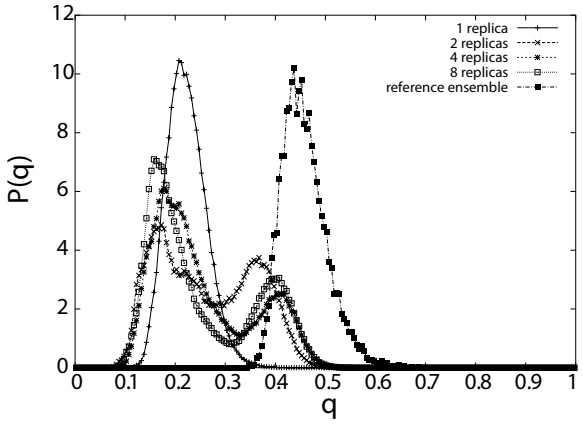
with larger variations of relative structural similarity. As one would expect, non-additivity causes the folds in this transition state ensemble to be less homogeneous than those in the p=1 reference ensemble. Indeed a wider range of CE Z-scores of $\sim 2.36 - 4.10$ is found within the reference ensemble itself. The overlap between the lower q subensemble of the p=3 reference ensemble and the reduced ensemble is excellent indicating that the folds in the reference subensemble are well represented in the reduced ensemble and that ensemble recovery judged by CE Z-score and the KS-test has been very successful for these reference structures. The CE Z-score is traditionally used for fold recognition. The probability distribution of the overlap function shows that the reduced ensemble does have the same distribution of global folds than the reference ensemble, which is not surprising since the average ϕ-values are reproduced. For the order parameter q, the dominant subensemble obtained by the clustering method only represents part (although most) of the p=3 reference transition state ensemble. Without knowledge of the energy landscape of the protein, the reduction method cannot be used to completely reproduce the reference ensembles. If further low resolution experiments are known, additional clusters can be identified that resemble the real gold standard ensemble.

2.6 Robustness of the Prediction of the Transiton State Ensemble for the λ-repressor

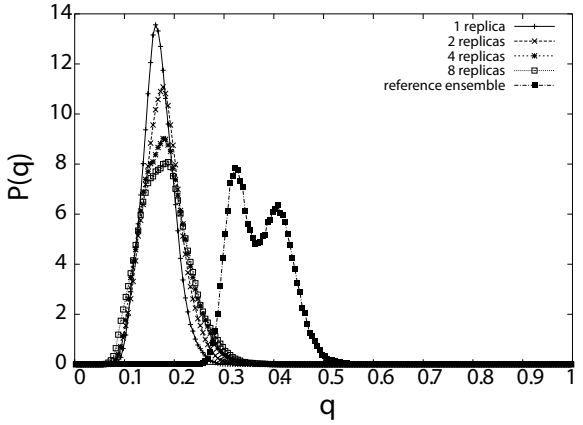
To test the robustness of the ability of replica simulations to recover reference ensembles from imperfect laboratory experiments, we introduced perturbations in the input data resembling experimental errors. To obtain these results we therefore stochastically changed the set of ϕ-values used in the reconstruction to see whether an ensemble could nevertheless be re-created that was faithful to the gold standard ensemble. To do this first a new set of ϕ-values was created by randomly picking a ϕ-value for each residue from a Gaussian distribution, having a mean value corresponding to the gold standard ϕ-value with a standard deviation given by the experimentally expected standard deviation of the measurement of that ϕ-value. The magnitude of the errors introduced was of the order of 20% for the p=1 reference ensemble ϕ-values and about 30% for the p=3 reference ensemble. These new perturbed lists of ϕ-values then served as an experimental input for the replica Hamiltonian. Replica simulations with these new ϕ-values were performed and the resulting transition state ensembles were compared to the reference ensemble with the KS overlap test. In the case of the p=1 transition state ensemble, the overlap distribution of the ensemble obtained with one replica using the new noisy ϕ-values no longer coincides with the distribution of overlaps within the gold standard reference ensemble at all (Figure 2.8(a)). The overlap distributions show that the two ensembles are rather different. Simulations using a single replica appear rather sensitive towards uncertainties in ϕ and fail to result in successful ensemble re-creation.

In contrast, the overlap distributions for ensembles obtained with multiple replicas do show considerable overlap with the distribution of the reference ensemble even when errors are introduced (Figure 2.8(a)) at least in the p=1 case. The replicas are able to compensate the uncertainties in ϕ and partially reproduce the reference ensemble. It is also found that clustering of structures with the Fitch-Margoliash clustering algorithm yields at least one cluster with

secondary and tertiary structure comparable to the reference transition state ensemble. The multiple-replica ensemble algorithm combined with selection through structural clustering therefore does successfully reproduce the p=1 reference ensemble.



(a)



(b)

Figure 2.8: Overlap distributions for the p=1 and p=3 transition state ensemble for 1,2,4 and 8 replicas. For p=3 the new set of ϕ-values generated structures, that did not overlap with the reference ensemble at all. In the p=1 case the overlap was smaller when compared to the overlap in Figure 2.3 .

However, the structures obtained in replica simulations, that should reflect the reference ensemble of the p=3 transition state ensemble, are structurally different from those in the reference ensemble. The KS-test shows that in no case can the replica simulation algorithm re-create the actual reference ensemble structures (Figure 2.8(b)). The p=3 transition state ensemble was obtained with a Hamiltonian that leads to highly cooperative behavior. The folding under this Hamiltonian should resemble the folding of a protein that folds by forming a specific, determined folding nucleus. The ϕ-values in the transition state ensemble are then expected to be less uniformly distributed with certain core residues being formed much earlier than the rest of the contacts. The errors introduced in ϕ are large and can potentially smear out the ϕ-values resulting in a mean-field like set of ϕ-values, which are more uniformly distributed like the ϕ-values of the p=1 transition state ensemble. This effect of creating a more uniform set of ϕ-values could be a possible explanation, why re-creating the transition state ensemble of the λ repressor obtained with a very non-additive Hamiltonian is more sensitive to errors in ϕ than is the re-creation for the p=1 ensemble.

2.7 Conclusion

Scientists often seek to invert hard won experimental data with the hope to obtain statistically correct structural ensembles with high fidelity. We see that the ability of successfully doing this for structural ensembles of partially folded biomolecules depends on the algorithm employed and on the quality of the measured data we seek to recreate.

First consider the p=1 reference ensemble. This ensemble has a unimodal overlap distribution and corresponds to a low transition state barrier with structures that are close in Q being close in free energy. Our simulation results with the molecular dynamics replica algorithm show that this algorithm can partially recreate the correct reference transition state ensembles from the set of ensemble-averaged ϕ-values. For the p=1 reference ensemble, structures obtained in simulations with one replica show overlap in the P(q) distribution with the reference ensemble. The KS-test shows, however, that the distributions of the reference ensemble and the ensemble obtained from the replica algorithm are not the same despite the fact that the ensembles share the same set of ϕ-values. The 1-replica algorithm partially reproduces the reference ensemble, that has a unimodal P(q) distribution with large errors in ϕ and few non-native contacts. On the other hand, the ensembles obtained with multiple replicas show a bimodal distribution in the probability distribution of overlap with the reference ensemble with only one peak strongly overlapping with the reference ensemble. The rather small overlap suggests that structural clustering could yield a small cluster of structures, that would better resemble the reference ensemble. Clustering of the structures with the Fitch-Margoliash algorithm shows two main basins of stuctures. A reduced ensemble obtained from this clustering analysis reproduces the reference ensemble as measured by the KS-test. These results suggest that when a single replica suffices to reproduce the reference ensemble, ensemble re-creation with multiple replicas does so too.

We also examined how stable the inversion is when errors mimicking those

found in experimental determinations are introduced. This study shows an additional advantage of introducing multiple replicas. Whereas the one-replica algorithm could not re-create the reference ensembles (Figure 2.8(a)) at all from error ridden input, the multiple-replica algorithm combined with structural clustering analysis is able to produce a reduced ensemble that has the same structural characteristics as the reference ensemble. Ensembles with low free energy barrier, from which a set of ϕ-values with large experimental errors is deduced, can only be inverted when multiple replicas coupled with structural clustering are introduced. For the p=3 transition state ensemble, the advantage of multiple replicas is also apparent. The ϕ-values represent ensemble averaged quantities. The reference ensemble has a bimodal probability distribution as well as two clusters of conformations when observing these conformations projected onto the principal components. Each of these subensembles have also large fluctuations in their ϕ-values. Hardly any individual structure in the reference transition state ensemble has the same set of ϕ-values as the ensemble average set of ϕ-values. In the inversion algorithm, the single replica algorithm recovers these individual structures. However a successful re-creation of the reference ensemble requires sampling of structures, that only on average reproduce the set of ϕ-values. Introduction of multiple replicas allows fluctuations of microscopic ϕ's referring to these subensembles while the ϕ-values averaged over all replicas still is constrained to its experimental value. The KS-test and PCA show that multiple replicas do sample the dominant subensemble of the p=3 reference ensemble well. Few structures sampled the minority reference subensemble (the $PC1 > 2$ region, see Figure 2.4) although the multiple replica algorithm does improve the ensemble re-creation over the single replica case. Knowledge of the energy landscape of folding is only partially encoded in the ϕ-values, thus adding additional a priori knowledge of the funneling of the energy landscape should help in inversion fidelity. Further work along these lines is planned.

Appendix

The text of this chapter, in full, is a reprint of the material as it appears in the Journal of Chemical Physics. The dissertation author was the primary researcher and author. Reprinted with permission from Lätzer J, Eastwood MP, Wolynes PG, JOURNAL OF CHEMICAL PHYSICS 125 (21): Art. No. 214905 DEC 7 2006, Copyright 2006, American Institute of Physics.

3 A Method for Inferring Partially Ordered Ensembles based on Energy Landscape Theory

3.1 Introduction

The importance of characterizing partially ordered thermodynamic states of biomolecules is becoming increasingly evident [120]. The goal of structural biologists and biophysical chemists should be to objectively infer both the mean structure and the correct magnitude of fluctuations of partially folded ensembles starting with experimental data alone and knowledge of the errors in such data. In fact, most existing direct inversion strategies are, however, biased to eliminate diversity in the ensemble of structures giving the impression of greater order than is probably correct. This is not a surprise because these strategies were originally intended for application to completely folded proteins. These are known by virtue of thermodynamic data to be separated by an energy gap form most denatured states and to be much less diverse. Such inversion strategies previously applied mostly to refining high resolution X-ray and NMR data [91, 90] may not give a faithful view of less structured ensembles. The best of the existing methods often involve simulating many copies of the protein (called "replicas" in statistical mechanical theory) and

applying the constraints provided by the intensities of the measured reflections to the average of the structure factor of the many copies [102]. Similar ensemble based constraints have also been introduced in Bayesian replica simulation methods for applying constraints derived from NMR [118, 92, 12] and from low resolution ϕ-value experiments [13, 79]. In the multiple copy approach the experimental constraints provide effective interactions between the different copies of the protein molecule. These interactions are of course not literally real. Instead they are virtual interactions representing the strength of inferences that can be made on the basis of the measured experimental observables. To the epistemological purist , the strength of these replica couplings should depend partly on the errors intrinsic to each experimental method but also on the confidence one has in an a priori model to the protein energy function - with infinite sampling and the correct energy function no experiments would have to be done at all! Statistical experimental errors are small for X-rays, moderate for most NMR experiments on highly folded proteins but are potentially large for low resolution data such as H/D exchange or ϕ-value analysis. All replica simulations so far introduced a non-physical constraint term to an energy function that otherwise is supposed to capture the essential physics and chemistry of the protein chain. Despite the fact that the strength of the inter-replica coupling should depend on the prior knowledge of the energy landscape of the system reflected in the physical energy function, in practice the interaction strengths are often cranked up to unreasonably high values. Bayesian inference methods which calculate the interaction strengths rather than adjusting them freely have been developed. These heuristic methods [112, 97, 98] determine simultaneously the weight and optimal native ensemble for a given set of NOE NMR constraints. The implementation of these more faithful methods is hindered by the computationalist - a problem we hope to overcome by developing a more analytical approach here.

In the current paper we describe a new method based on energy landscape principles to compute the inter-replica coupling without simulation. This method is based on a strict Bayesian interpretation of the experimental input

data (here chosen to be transition state ϕ-values for illustration) and can include the known statistical and modeling uncertainties in those data. Energy landscape ideas were used to fix the magnitude of these interactions in an objective way based on the magnitude of the errors and the landscape yielded by the physical energy function. We base our algorithms on free-energy functionals [114, 116, 115, 110, 111, 113], using the energy of the native contact formation and polymer physics estimates of the entropy. Such functionals have been used to successfully predict the free energy profile in the absence of experimental data. In this paper we use a similar functional to compute the interaction strength between replicas analytically, when the completely folded structure is known. The functional can assume a priori given physical energy functions either having explicit cooperativity,

After the couplings are found analytically, we then carried out replica molecular dynamics simulations with the analytically computed interaction strengths to deduce structural ensembles. To test this approach we applied the method to a known computed ensemble corresponding to a folding transition state. In this case the fidelity of the method can be objectively quantified.

3.2 Methods

3.2.1 Description of the Free Energy Functional

A replica free energy functional approach is presented in order to characterize with proper fidelity the partially structured ensembles that can be obtained for a given set of experimental constraints. The free energy functional can be written as $\mathcal{F} = \mathcal{F}_{phys} + \mathcal{F}_{exp}$. $\mathcal{F}_{phys}$ contains physics and chemistry based entropy/energy terms while the pseudo-energy term $\mathcal{F}_{exp}$ biases the ensemble to match the experimentally determined data. This biasing term can be written as a function of the normalized contact probabilities $q_{ij}^{\alpha}(\mu(\alpha))$ for residue i and

j of replica α when a total number of $\mu(\alpha)$ contacts are formed.

$$\mathcal{F}_{exp} = \sum_i \lambda_i \left(\frac{1}{N_{rep}} \sum_\alpha \frac{1}{N^i_{cont}} \sum_{<j>} q^\alpha_{ij}(\mu(\alpha)) - < \phi_i >_{exp} \right)^2 \tag{3.1}$$

The functional form of $\mathcal{F}_{exp}$ was chosen to have a quadratic form for convenience. The parameter λ_i is a measure of confidence in the experimental measurement. It weights the contribution of each constraint according to its assumed statistical error. The study uses ϕ-values for illustration purposes but other constraints that can be written in terms of contacts such as NOE or H/D exchange data can be treated similarly. We will usually assume no frustrations. The ϕ-values $< \phi_i >_{exp}$ are approximated in this functional to be equal to the average contact probability $q^\alpha_i = \frac{1}{N^i_{cont}} \sum_j q^\alpha_{ij}$ for each residue i with N^i_{cont} native contacts averaged over all N_{rep} realizations. More accurate energy weighted quantities can also be used.

We take the energetic terms as a simple two-body interaction potential for residues which are known to form contacts in the native structure.

$$\mathcal{F}_{con.pot.} = \frac{1}{N_{rep}} \sum_\alpha \sum_{i,j} \epsilon_{ij} q^\alpha_{ij}(\mu(\alpha)) \tag{3.2}$$

The ϵ_{ij} terms give the contact interaction strength between residue i and j. Suitable choices for the ϵ_{ij} terms are those from optimized potentials for protein folding such as the potential described by Goldstein [95], potentials derived from information theoretic approaches [105, 89] or simply Gō-like [117] terms if one is comfortable assuming a homogeneous funneled landscape.

The entropic costs of forming contacts must also be addressed in the functional. There are many different ways to select a sensible entropy functional, such as the entropy functional used in describing network glasses [99]. The motivation for the specific entropy functional used in this study is that this functional has already been succesfully applied to the problem of characterizing partially structured ensembles including those of the λ-repressor [115]. The entropy functional is made up of several terms including a contact entropy functional and a mixing entropy term. The contact entropy functional

used in this study has been obtrained from an interpolation of the Jacobson-Stockmayer functional [101], where the long-range entropic contributions are derived from Flory's theory of rubber elasticity [94]. The details for this functional are explained in greater detail in the work of Shoemaker, Wang and Wolynes [116].

$$S_{ij}^{\alpha} = k_B \log \left(\Delta V |i-j|^{-\frac{3}{2}} + \left(\frac{N}{\mu(\alpha)} \right)^{-\frac{3}{2}} \right) \tag{3.3}$$

One contribution to the entropy changes comes from forming a specific set of contacts. The Flory correction reflects the fact that loops are shorter once μ contacts have already been made. Yet not all contacts are made all of the time. Therefore another contribution to the total entropy in an ensemble of partially ordered structures comes from the combinatorial entropy of mixing. In a partially ordered protein ensemble, the mixing entropy term arises from the number of possible ways that these contacts can be made.

$$\begin{aligned} \mathcal{F}_{c.e.} = \frac{1}{N_{rep}} \sum_{\alpha} \sum_{ij} k_B T (q_{ij}^{\alpha}(\mu(\alpha)) \ln q_{ij}^{\alpha}(\mu(\alpha)) + \\ (1 - q_{ij}^{\alpha}(\mu(\alpha))) \ln (1 - q_{ij}^{\alpha}(\mu(\alpha)))) \end{aligned} \tag{3.4}$$

The native ensemble is obtained, when all native contacts are formed so each $q_{ij}^{\alpha} = 1$. The entropy loss to go from the completely unfolded state to the completely folded state can be estimated for a protein of N residues to be $Nlog(\nu)$, where ν is the number of conformations per residue (here we chose $\nu = 4$). This overall entropy accounting fixes the parameters in the contact

entropy functional S_{ij}^{α}. The final free energy functional can now be written as

$$\begin{aligned}\mathcal{F} = & \sum_i \lambda_i \left(\frac{1}{N_{rep}} \sum_{\alpha} \frac{1}{N_{cont}^i} \sum_{<j>} q_{ij}^{\alpha}(\mu(\alpha)) - < \phi_i >_{exp} \right)^2 \\ & + \frac{1}{N_{rep}} \sum_{\alpha} \sum_{ij} \epsilon_{ij} q_{ij}^{\alpha}(\mu(\alpha)) \\ & - T \left(\frac{1}{N_{rep}} \sum_{\alpha} \sum_{ij} S_{ij}^{\alpha} q_{ij}^{\alpha}(\mu(\alpha)) + \sum_{\mu'=1}^{\mu} \sum_{ij} \frac{\partial S_{ij}^{\alpha}(\mu')}{\partial \mu} \delta q_{ij}^{\alpha}(\mu') + k_B N log(\nu) \right) \\ & + \frac{1}{N_{rep}} \sum_{\alpha} \sum_{ij} T \left(q_{ij}^{\alpha}(\mu(\alpha)) \ln q_{ij}^{\alpha}(\mu(\alpha)) + (1 - q_{ij}^{\alpha}(\mu(\alpha))) \ln (1 - q_{ij}^{\alpha}(\mu(\alpha))) \right)\end{aligned} \tag{3.5}$$

where $\delta q_{ij}^{\alpha}(\mu'(\alpha)) = q_{ij}^{\alpha}(\mu'(\alpha)) - q_{ij}^{\alpha}(\mu'(\alpha) - 1)$. At the folding temperature T_F, this functional satisfies the condition that in the unfolded state, the entropy is $k_B N \ln \nu$, while in the folded state the entropy should vanish. It is the parameter ΔV in the contact entropy term which is tuned such that the complete entropy in the folded state is zero while in the unfolded state the total entropy is of magnitude $Nlog(\nu)$.

3.2.2 Cooperativity Effects

The free energy functional, so far, does not contain nonpairwise-additive interactions. Native structure based potentials with explicit many-body interactions have been found to capture protein folding kinetics with more accurate and realistic rates and barriers [93]. The barrier heights typically increase with increasing amount of nonadditive interactions [109, 14]. We describe three different contributions to the cooperativity. One contribution is an α-helical-local-density interaction free energy. The formation of a helical contact is facilitated, when other contacts near it have already formed. Luthey-Schulten et al. [103] have explained this based on Onsager's theory of liquid crystals - it is the so-called "induced rigidity" of Pincus and DeGennes. This interaction free energy stabilizes the native state and tends to make molten globules orientational liquid crystals. This enhancement for helical residues stems from the

fact that arrangement of a third residue to its native conformation, when two nearby residues are already formed, is entropically more favorable than the entropic cost of the sum of three pairwise terms accounting for the formation of these three contacts independently of each other.

$$F_{h-\rho} = \alpha_{h-\rho} T \frac{1}{N_{rep}} \sum_{\alpha} \sum_{i}^{hx} q^{\alpha}_{i-4,i}(\mu(\alpha)) \sum_{k} q^{\alpha}_{ik}(\mu(\alpha)) \tag{3.6}$$

Using a stabilization energy of roughly $1k_BT$ per helical residue (see helix-coil theory by Luthey-Schulten et al. [103]), one obtains an estimated value of $\alpha_{h-\rho} = 0.3$ [115] for the λ-repressor was chosen.

The second cooperative term used in this present functional arises from the fact that breaks in α-helices introduce surface energy terms. This free energy term is similar to the helical initiation free energy calculated in helix-coil theory [103].

$$F_h = \alpha_h \frac{1}{N_{rep}} \sum_{\alpha} \sum_{i}^{hx} (q^{\alpha}_{i-4,i}(\mu(\alpha)) - \frac{1}{2})(q^{\alpha}_{i,i+4}(\mu(\alpha)) - \frac{1}{2}) \tag{3.7}$$

The magnitude of interaction of this functional has been estimated by assuming that the energetic cost to form a helix relative to forming a coil is given by $F = -\ln \sigma$. Measurements of the surface tension σ [106] have been used to infer a value of approximately $\sigma = 10^{-1}$ [103].

Finally we add a free energy functional term motivated by the capillarity picture of protein folding [119]. In the capillarity picture, regions, in which side chains are either completly ordered, partially ordered or unfolded, may be separated by rather complex interfaces which may be improperly wetted. A reasonable choice for this kind of cooperative interaction is given by

$$\begin{aligned} F_c = \alpha_c \frac{1}{N_{rep}} \sum_{\alpha} \sum_{ij} \sum_{k} \sum_{l} ((q^{\alpha}_{ij}(\mu(\alpha)) - \frac{1}{2})(q^{\alpha}_{lk}(\mu(\alpha)) - \frac{1}{2}) q^{NAT}_{i,k} \\ + (q^{\alpha}_{ij}(\mu(\alpha)) - \frac{1}{2})(q^{\alpha}_{kl}(\mu(\alpha)) - \frac{1}{2}) q^{NAT}_{k,j}) \end{aligned} \tag{3.8}$$

where $q^{NAT}_{k,j}$ equals one if residues k and j form a contact and zero otherwise. It is possible to estimate the magnitude of α_c (in the range of 0.05-0.1) [116] by

matching the total surface loop entropy loss for a random conformation (where part of the protein is completely folded and the remainder of the protein is completely unfolded) calculated from polymer physics [109] to the entropy cost obtained with the functional.

3.2.3 Free Energy Profiles and the Calculation of the Parameters λ_i

To obtain the free energy profile along the folding reaction coordinate, the set of ensemble averages, the q_{ij}^{α}'s, must be calculated. The free energy profile is obtained by minimizing the free energy functional with respect to q_{ij}^{α} while imposing the boundary condition to have a given total progress along the folding reaction coordinate. This translates mathematically into solving the equation $\frac{\partial \mathcal{F}}{\partial q_{ij}^{\alpha}} = 0$ with the Lagrange multiplier $(q - q^*)\gamma = 0$, where q^* is the given value of the total degree of ordering and $q = \frac{1}{N}\sum_i q_i^{\alpha}$. This leads to the following equation to obtain the set of q_{ij}^{α}'s as the system progresses down the funnel.

$$
\begin{aligned}
q_{ij}^{\alpha} = (1 + \exp(\frac{1}{T}(\epsilon_{ij} - \gamma - TS_{\alpha}^{ij} \\
+ 2T\alpha_{h-\rho}\sum_{i}^{hx}(q_{i-4,i} + \sum_{k} q_{ik}) \\
+ \alpha_h \sum_{i}^{hx}(q_{i,i-4} + q_{i+4,i+8}) \\
+ \alpha_c \sum_{k,l}(q_{kl}\delta_{ik}^{nat} + q_{kl}\delta_{kj}^{nat}) \\
+ \frac{2T\lambda_i N_j}{N_i + N_j}(q_i - < \phi_i >_{exp}) + \frac{2T\lambda_j N_i}{N_i + N_j}(q_j - < \phi_j >_{exp})))\Big)^{-1}
\end{aligned}
\tag{3.9}
$$

Equation 3.9 can be solved self-consistently, when the λ_i's are known. For a given set of λ_i's it is straight forward to compute the free energy landscape for all values of q^*.

Often in simulations the weight parameters λ_i are estimated from an heuristic procedure that reflects the certainty in the observed data [112, 98]. Here

the free energy functional can be used to directly and quntitatively infer the λ_i's for a given set of observed data, $< \phi_i >_{exp}$, and its uncertainties $\delta\Phi_i$. To compute the λ_i's, the free energy functional $\mathcal{F}$ is minimized with respect to λ_i, i.e we seek to solve $\frac{\partial \mathcal{F}}{\partial \lambda_i} = \delta\Phi_i^2$. Since the only direct dependence of $\mathcal{F}$ is given by the experimental constraint term, we are left to solve $(Q_i - < \Phi_i >)^2 = \delta\Phi_i^2$ given the constraint that $\frac{\partial \mathcal{F}}{\partial q_{ij}^{\alpha}} = 0$. The calculated weight parameters can now be used in simulations [79] to infer structural ensembles with a precision directly determined by the experimental errors $\delta\phi_i^{exp}$.

3.3 Free energy landscape obtained with constant λ's

We first present the results for a model ensemble inversion assuming there are no cooperative terms in either the functional nor in the simulation Hamiltonians used to generate the original data. We carried out molecular dynamics simulations with a native structure based Hamiltonian [79] to obtain the free energy profile of the λ repressor. The λ repressor folded as a two-state folder with a low energy barrier between unfolded and folded ensemble. The transition state ensemble is defined by structures whose order parameter q is at the q of the barrier. The input data were derived from this completely known reference ensemble. This reference ensemble exhibits a unimodal probability distribution suggesting *one* distinct transition state ensemble. We first test the ability of our method for symmetric replicas to infer the real transition state ensemble from the given experimental data and its errors, which are taken of to be of the order of 22%. This is a reasonable magnitude for laboratory kinetic data on ϕ's.

The free energy profiles for constant λ's ($\lambda = 0..50$) in the absence of cooperativity terms were calculated at T_F by solving Equation 3.9 for the set of q_{ij}^{α}, where q^* is varied from $q^* = 0.01$ to $q^* = 0.99$. The nonlinear equations for q_{ij}^{α} were solved iteratively. The resulting free energy profiles as a function of the

total folding progress q for a Hamiltonian with all energy terms set to zero (all $\epsilon_{ij} = 0.0$), $\mathcal{F}_{bb}$, and also for a Gō-like Hamiltonian (all ϵ_{ij} are equal, non-zero and scaled as described in the method section), $\mathcal{F}_{G\bar{o}}$, are shown in Fig. 3.1a,b. At low λ the free energy profiles are dominated by $\mathcal{F}_{phys}$. Most regions are dark blue, i.e these regions are preferred regions. For the Gō-like Hamiltonian with $\lambda = 0.0$, a barrier of approximately $4k_BT$ is observed between the unfolded ensemble and the folded ensemble. This barrier is slightly higher than the barrier observed in molecular dynamics simulations. With increasing λ only regions around the average q-value of approximately $q = 0.74$ (obtained by averaging all $< \phi_i >_{exp}$-values) are preferred while especially low q regions ($q \leq 0.5$) and high q regions ($q \geq 0.85$) become unfavored.

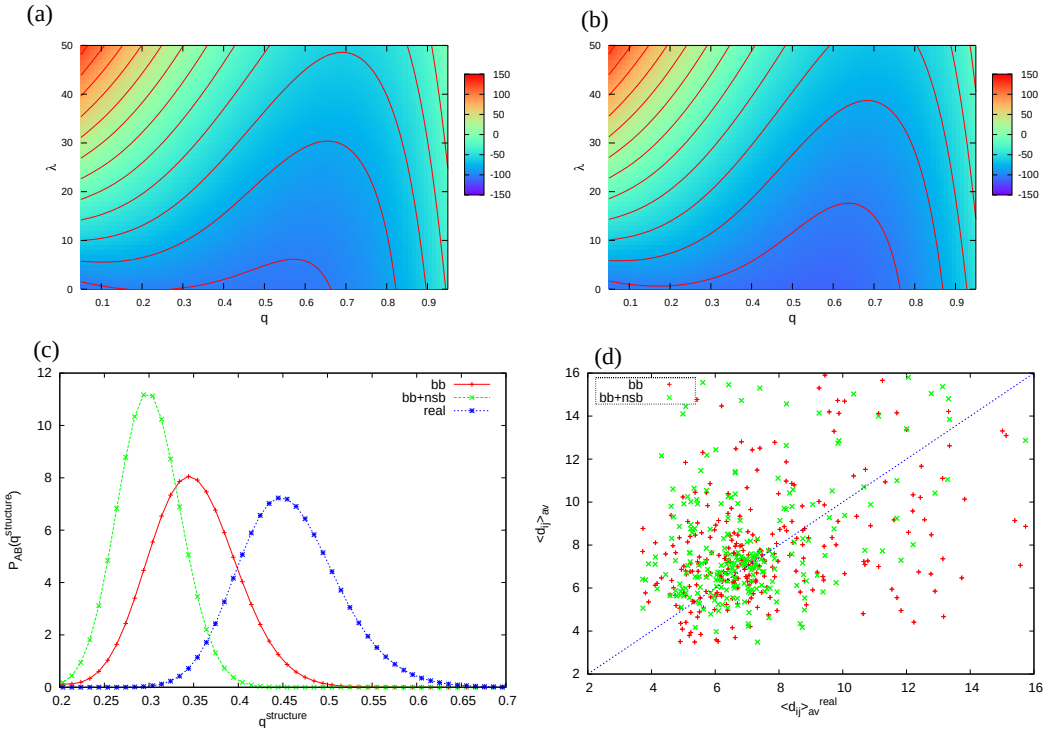


Figure 3.1: A) Free energy profiles obtained with the backbone based free energy functional for increasing, but constant weight parameter. B) Same as A) but with the native structure based functional. C) Probability distribution of the reference ensemble is shown in blue. The overlap probability distributions are shown in red for the backbone Hamiltonian and green for the funneled Gō-like Hamiltonian. D) Plot of the average distances for each residue pair ij found in the reference ensemble and the deduced ensemble. Red dots represent distances found in the backbone based Hamiltonian while green dots represent distances found in the funneled Hamiltonian.

We used molecular dynamics replica simulations [79] to obtain structural ensembles. For the deduced ensembles, the input ϕ-values were reproduced. In the ideal situation, the deduced ensembles should also be identical to the real ensemble. It is common practice to validate the deduced ensembles by calculating some different order parameters from the deduced ensembles and compare these order parameters to experimental data [118, 96]. In our case the reference ensemble is completely known and we can therefore judge directly whether the deduced ensembles faithfully represent the real ensembles or if the deduced ensemble simply represents a sub-ensemble of the real ensemble, that matches the input ϕ-values by using the Kolmogorov-Smirnoff (KS) test. The deduced ensembles were in this way able to be compared to the real "gold standard" ensemble. For the KS-test, the individual probability distributions of the reference ensemble ($P_A(q^{structure})$), the replica ensemble distributions ($P_B(q^{structure})$) and the overlap distributions ($P_{AB}(q^{structure})$) are each calculated. Here $q^{structure} = \frac{2}{(N-1)(N-2)} \sum_{i<j-1} \exp\left(-\frac{(r_{ij}^1 - r_{ij}^2)^2}{2\sigma_{ij}^2}\right)$ is a normalized order parameter that describes the structural similarity of a given structure 1 with coordinate set $\{r_{ij}^1\}$ to a second structure 2 with coordinate set $\{r_{ij}^2\}$. Unlike q, $q^{structure}$ is dependent on the exact distances for each contact pair ij. Two structural ensembles are statistically indistinguishable if the probability distribution of pairwise overlaps $P_{AB}(q^{structure})$ is equal to the reference probability distribution $P_A(q^{structure})$. The probability distribution of the "gold standard" transition state ensemble (Fig. 3.1c, dark blue curve) and the overlap distributions $P_{AB}(q^{structure})$ (Fig. 3.1c, red and green curve) with the highest overlap to the real ensemble are shown in Fig. 3.1c. Neither of the two ensembles obtained from the molecular dynamics replica simulations completely overlap with the reference ensemble. However, the ensembles obtained with $\mathcal{F}_{bb}$ show much larger overlap, i.e these ensembles are closer to the model gold-standard transition state ensemble than are the ensembles obtained with the Gō-like Hamiltonian $\mathcal{F}_{G\bar{o}}$.

To obtain a feel for the distances found in the deduced ensemble and the gold-standard ensemble, we compare the distances $< d_{ij} >_{av}^{real}$ and $< d_{ij} >_{av}$

of each residue contact pair ij observed in the two ensembles and averaged over all structures in each respective ensemble. Fig. 3.1d shows that most contacts that are formed within the cutoff distance are found along the diagonal, however a few contacts $< d_{ij} >_{av}$ can deviate quite a bit (up to 10Å) from the contact distances $< d_{ij} >_{av}^{real}$ seen in the real ensemble. The probabilty distribution P(Δ_i) with $\Delta_i = \log \frac{<d_{ij}>_{av}^{real}}{<d_{ij}>_{av}}$ exhibits an approximate Gaussian function centered at $\Delta_i = 0.0$ for both deduced ensembles (figures not shown).

3.4 Evidence of Replica Symmetry Breaking in MD simulations with constant weights

To what extent do the members of an inferred ensemble structurally cluster? In the statistical mechnical language this is the issue of whether there is replica symmetry breaking. To test whether a system of replicas exhibits replica symmetry breaking, one needs to define a replica correlation function [107] q, that measures the overlap of two states from two different replicas. In spin glass physics [104] one then computes the probability distribution $P(q)$. A trivial $P(q)$, i.e a single-peak, unimodal distribution, indicates no broken replica symmetry. The existence of replica symmetry breaking becomes apparent if there is a non-trivial distribution $P(q)$. To test for replica symmetry breaking, ensembles were deduced with a replica molecular dynamics algorithm having four replicas with the two introduced Hamiltonians, the backbone only Hamiltonian and the Gō Hamiltonian, for constant weights λ.

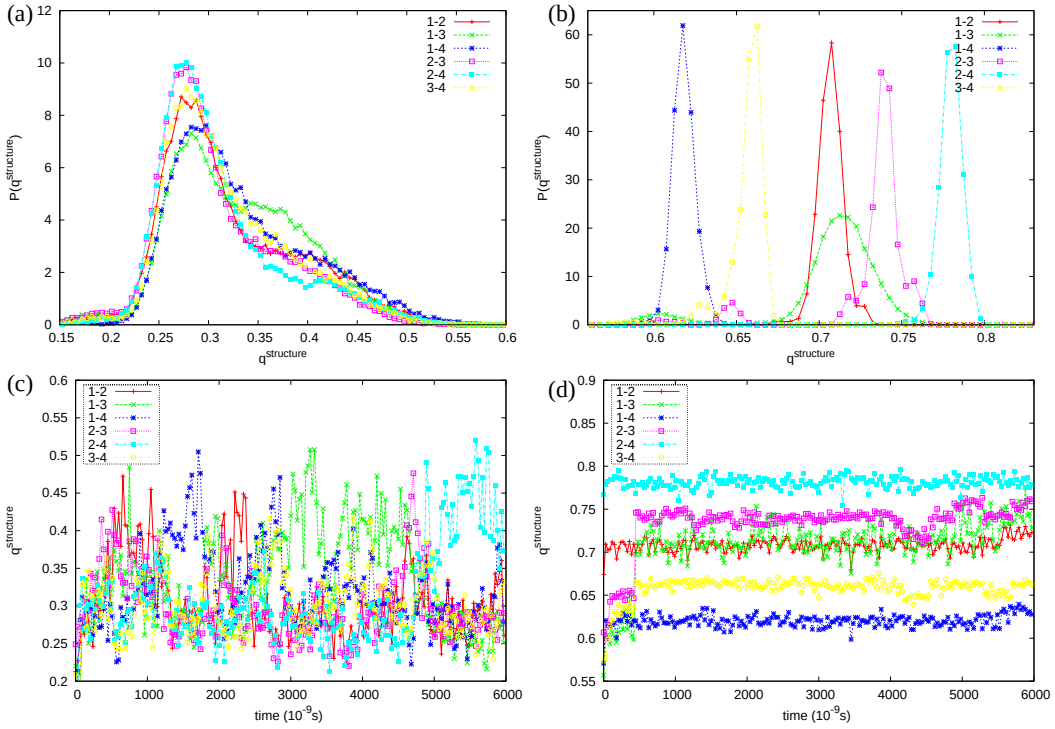


Figure 3.2: All four replicas were labelled one to four. The overlap distributions between all possible pairs of replicas for the backbone based Hamiltonian (a) and the funneled Hamiltonian (b) are shown. Also the similarity between conformations as a function of time for both Hamiltonians are plotted in (c) and (d).

The overlap distributions $P(q^{structure})$ were computed and plotted in Fig. 3.2. The replica correlation order parameter $q^{structure}$ is defined as before, but conformation 1 comes from one replica and conformation 2 comes from a different replica. The overlap distributions for the $\mathcal{F}_{G\bar{o}}$ Hamiltonian in Fig. 3.3b are fairly sharp and unimodal. There seems to be no replica symmetry breaking. Over the time course of the simulations, each replica stays correlated to another replica at about the same level of $q^{structure}$. A plot of $q^{structure}$ of all possible replica pairs for each simulation snapshot is shown in Fig. 3.2d. To assess if the four-replica $\mathcal{F}_{G\bar{o}}$ molecular dynamics simulation with no replica symmetry breaking still improved the ability to infer the real transition state ensemble, the overlap distributions $P_{AB}(q^{structure})$ were computed. A plot of the overlap distribution of the deduced ensemble with the real ensemble (figure not shown) did show the same overlap as seen in the one-replica case, i.e. there was no improvement in the ability to deduce the transition state ensemble with multiple replicas in the case of the $\mathcal{F}_{G\bar{o}}$ Hamiltonian.

The four-replica simulations with the backbone only Hamiltonian, on the other hand, displayed replica symmetry breaking. The $P(q^{structure})$ overlap distributions in Fig. 3.2a exhibit broad distributions with a peak around $q^{structure} = 0.27$ and a broad shoulder at higher $q^{structure}$. During the course of the simulation, replicas become correlated at certain times. For example, at the beginning of a simulation, the overlap between the replicas we label 1 and 3 is low , but after $3\mu s$ the replicas become overlapping for a time of $2\mu s$ before the amount of overlap decreases. The system of replicas is strongly constrained to, on average, match the experimental ϕ-values with a coupling of strength λ. At each timestep, a distinct pair of replicas shows distincly larger overlap with each other than do the remaining replica pairs, which leads to an important point: the replicas silmultaneously explore different regions of the conformational phase space, although restricted through the coupling of the replicas. Analysis of the overlap distribution of the deduced ensemble with the real ensemble shows that the four replica algorithm performs better than the one replica algorithm (see ref. [79]) when only backbone connectivity is

assumed to be a reliable a priori constraint.

3.5 Ensemble Inversion from Experimental Data and its Errors

To what extent does weigthing of the parameters λ_i improve the algorithms? The relative contributions of $\mathcal{F}_{phys}$ and $\mathcal{F}_{exp}$ to the total free energy are controlled with the weight parameters λ_i. Several choices of λ_i exist: one easy choice of λ_i is simply to set the physical energy terms equal to the constraint energy [100], i.e. $\mathcal{F}_{phys} = \mathcal{F}_{exp}$, which should yield a good approximation to obtain the "correct" structures. But this is not epistemological correct or optimal. The optimal weights can be found using different approaches based on Bayesian probability theory [97, 98, 112] and can be calculated directly with the free energy functional.

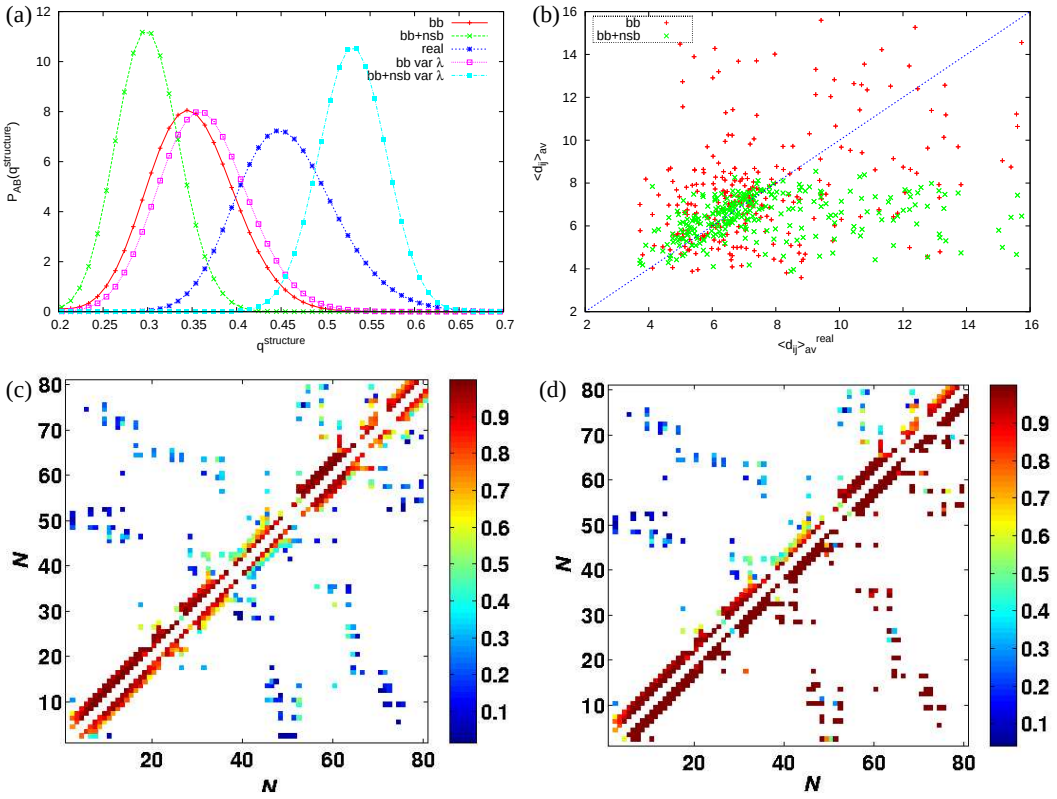


Figure 3.3: a) Various overlap probability distributions for the two Hamiltonians with analytically calculated weight paramters (pink for backbone based Hamiltonian, light blue for the funneled Hamiltonian). b) Plot of the average distances with same color coding as before but for the ensembles obtained with the analytically calculated weight parameters. (c) and (d) show the average contact map of reference ensemble (above diagonal) compared with the deduced ensembles (below diagonal) for the two Hamiltonians, backbone based Hamiltonian in (c) and funneled Hamiltonian in (d).

The free energy functional method allows us to compute the heterogeneous set of λ_i's. The error in the data dictates the strength of interaction and hence the confidence in the measurement. The free energy functional (Equation 3.5) with no cooperativity terms was minimized and solved to obtain the set of λ_i's for the given experimental errors $\delta\phi_i^{exp}$. The calculated λ_i's varied in magnitude with few λ_i-values close to zero and few values extremely high. We then performed molecular dynamics replica simulations with the two described Hamiltonians, $\mathcal{F}_{bb}$ and $\mathcal{F}_{G\bar{o}}$, with the calculated weights λ_i. As a first assessment of the fidelity of the deduced ensembles, the overlap distribution $P_{AB}(q^{structure})$ and the probability distributions $P_A(q^{structure})$ and $P_B(q^{structure})$ were computed and plotted in Fig. 3.2a. For the $\mathcal{F}_{bb}$ Hamiltonian the probability distribution of overlap for the ensemble deduced with the hetergeneous set of λ's (Fig. 3.3a, pink curve) was shifted slightly more towards the real transition state ensemble (Fig. 3.3a, dark blue curve) than the ensemble deduced with a large but constant value of λ's (Fig. 3.3a, red curve). This result suggests that the quality and fidelity of the resulting structures in the ensemble deduced with the heterogeneous set of λ_i's was improved. This is a key result because it emphasizes that it is important to know the relative strength for each constraint to optimally deduce the ensemble with highest fidelity from the input data. The plot of the average distances $< d_{ij} >_{av}$ versus $< d_{ij} >_{av}^{real}$ (Fig. 3.3b) in the deduced ensemble resembles the plot in Fig. 3.1b. The probabilty distribution P(Δ_i) is a Gaussian centered at $\Delta_i = 0.0$ with a width comparable to the width of the distribution P(Δ_i) obtained for the ensembles deduced with the constant weight parameters. An ensemble averaged contact map for beta-carbons (C^β-C^β distances that fall within 8Å) of the native contacts formed in the ensemble is shown in Fig. 3.3c. The contact map above the diagonal represents the contact map found in the gold-standard transition state ensemble while the contacts shown below the diagonal are those found in the deduced ensemble. Most of the contacts in the deduced ensemble are identical to the contacts in the real ensemble. This indicates near perfect re-creation of the real topology and ϕ-values.

The distributions $P_B(q^{structure})$ and $P_{AB}(q^{structure})$ (Fig. 3.3a, light blue curve) for the ensemble obtained with the $\mathcal{F}_{G\bar{o}}$ Hamiltonian were also computed. The overlap distribution $P_{AB}(q^{structure})$ for the ensemble obtained with the Gō Hamiltonian and the analytically computed λ's shows much larger overlap with the reference probability distribution $P_A(q^{structure})$ than the distribution $P_{AB}(q^{structure})$ (Fig. 3.3a, green curve) obtained with a high and constant λ. The free energy profile derived with the functional for constant weights already showed, that for increased λ the folded regions become stabilized (since the average q was very high, at least $\frac{3}{4}$ of the contacts should be formed). It is expected then, that the Gō Hamiltonian should mainly sample very folded conformations close to the native state. The deduced ensemble indeed exhibited very native-like secondary and tertiary structure, which explains, why the distributions $P_B(q^{structure})$ and $P_{AB}(q^{structure})$ were shifted towards higher q values. The plot of the average distances $< d_{ij} >_{av}$ (Fig. 3.3b) clearly shows that most contacts in the deduced ensemble are less than 9Å, while in the reference enesemble distances of up to 16Å are observed. The ϕ-values of the deduced ensemble were also higher than the ϕ-values of the refrence ensemble, but within error of 22%, an observation that is also manifested in the contact map (Fig. 3.2d), which shows more folded contacts especially for long range contacts (more red) in the deduced ensemble than in the reference ensemble.

3.6 Inversion of a multimodal transition state ensemble

A more challenging but interesting reference transition state ensemble for inversion is an ensemble for a protein that has multiple folding pathways. Naturally such an ensemble should exhibit a multimodal probability distribution. We test the inversion algorithms on data derived from a reference ensemble of the λ-repressor with a bimodal probability distribution [79] (see also Fig. 3.4 blue curve). This reference ensemble of the λ-repressor was obtained with a Gō like Hamiltonian with added explicit cooperativity. The bimodality stems

from *two* structurally distinct transition state ensembles that arise due to two distinct routes to the folded state. For this reference ensemble single replica methods have failed to correctly infer the transition state ensemble. We therefore present only the results for four replicas.

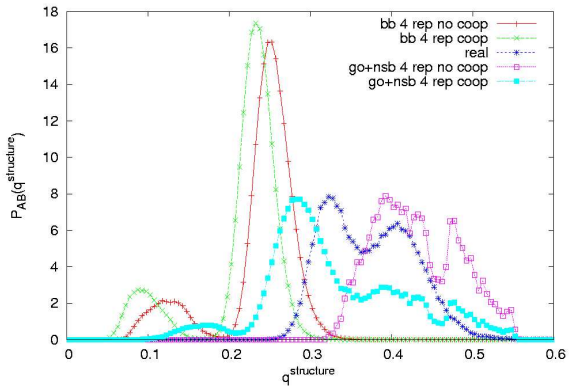


Figure 3.4: Overlap distributions with the multimodal reference ensemble (blue). The green and red curves are the overlap probability distributions of the ensembles deduced with the backbone based Hamiltonian without (red) and with (green) explicit cooperativity. The overlap probability distribution obtained with the funneled Hamiltonian are shown in pink (no explicit cooperativity) and blue (explicit cooperativity).

We calculated the various set of λ's from the free energy functional for further simulation studies. The overlap probability distributions for the deduced ensembles are shown in Fig. 3.4. All overlap probability distributions are multimodal The probability distribution of the ensembles obtained with the backbone Hamiltonian never overlaps much with the reference ensemble distribution even when a resonable level of cooperativity, which is observed in real proteins, is included in the functional and the simulations. This result is not surprising since the errors in the measurements are of the order of 40% and no guiding energy function is present in the algorithm. The algorithm with the funneled energy terms performed rather well compared to the algorithm with the $\mathcal{F}_{bb}$ Hamiltonian. This result suggests that when the physical energy function yields an energy landscape similar to the real landscape of the protein and the constraint terms are properly gauged, structure deduction with the Bayesian methods can be highly successfull despite the large errors. The presence and absence of cooperativity only changed to a small extent the ability to re-create the two transition state ensembles. In the absence of the cooperative interactions, the structures are slightly shifted towards the more folded order parameter while in the presence of cooperative interactions the structures of the transition state ensemble are less similar to each other and to the real transition state ensemble.

3.7 Conclusions

Energy landscape theory and free energy functional techniques from spin glass theory provide us useful tools to deduce structures from experimental data and its errors and to quantify our confidence. We are able to compute with a multi-replica free energy functional the inter-replica coupling strength used for structure determination. The magnitude of the experimental errors is directly related to the importance of each constraint and therefore the theory provides a measure of the fidelity of reconstruction and the strength of inference that can be made from each measurement. We infer from the simulations

to re-create the unimodal transition state ensemble that all algorithms generally work well, when the inter-replica coupling strengths are correctly gauged using the analytically functional approach.

However, for a bimodal transition state ensemble with large uncertainty in the data and two structurally distinct subensembles, ensemble re-creation can only be successfull, when the energy landscape is reasonably funneled and the strength of the constraint relative to the energetics of the physical energy function is properly gauged.

4 Induced Fit, Folding, and Recognition of the NF-κB-Nuclear Localization Signals by IκBα and IκBβ

4.1 Introduction

The import selectivity of nuclear proteins from the cytosol relies on nuclear localization signals (NLS). These generally are short sequences of 3-20 amino acid residues, normally rich in lysine and arginine, which bind to nuclear import receptors [27, 28]. Monopartite NLSs contain only a single cluster of positively charged residues, and bipartite NLSs contain two stretches of basic residues connected typically by a 10-12 residue linker. The prototypical monopartite NLS is the simian virus 40 (SV40) large T antigen (TAg) NLS of sequence PKKKRKV [29]. This NLS sequence is very specific and mutation of a single residue, K128, leads to loss of binding to the nuclear import factor resulting in cytoplasmic retention of TAg [30]. A consensus sequence that represents the diversity of NLSs is K-(K/R)-X-(K/R), where X is any amino acid [31]. This consensus sequence must be recognized by the nuclear import receptor.

For the NLS to be active, it needs to be exposed for binding to the surface

of the nuclear protein-import receptor complex. IκB, the inhibitor of NF-κB, deactivates the NF-κB nuclear localization signal by physically masking it [32, 33]. The interactions of the different NLSs of the NF-κB family members, homo- or heterodimers made from five subunits, have evolved to achieve very specific recognition and binding to members of the IκB-family. For example, IκBα, the most abundant IκB, binds and inhibits NF-κB p65 homodimers but not p50 homodimers [34]. Part of the reason for this interaction specificity lies in the recognition of the NF-κB NLS polypeptide. The NLS polypeptides of p50 and p65 are defined as the NLS-containing carboxy-terminal fragments with lowest sequence homology within the rel homology domain of the NF-κB family [35]. The specificity of IκBα for p65 comes partly from the fact that the p50 NLS does not bind to IκBα with significant affinity [36]. Comparison of the crystal structures of p50/p65 complexed to IκBα and to DNA show electron density for the p65 NLS polypeptide when bound to IκBα but not when bound to DNA [37, 38]. This result suggests that the p65 NLS polypeptide is flexible in the unbound state and becomes more ordered upon forming a complex with IκBα [36].

The advantage of the NLS being inherently flexible is that local structure can be modified in response to different molecular targets. This allows competitive binding to several different targets, affording the necessary non-linearity of a control circuit [39]. NLSs bound to importin α usually adopt extended structures [40]. The p65 NLS polypeptide, residues 289-320 of the rel homology domain of p65, binds to IκBα in a split helical conformation [38, 41]. Thus, the disordered structure of the NLS polypeptide when NF-κB is free or bound to DNA allows it to recognize either importin α or the various IκB isoforms.

In order to probe how the NF-κB NLS polypeptides achieve flexibility and specificity at the molecular level, we performed simulations to predict structures using the optimized associative memory Hamiltonian (AMH) method for the free and IκB-bound NLS polypeptides [76]. We note that the word predict can lead to misunderstanding, since the crystal structures of IκBα with

NF-κB(p50/p65) and of IκBβ with NF-κB(p65/p65) are already known and deposited with the RCSB Protein Data Bank [41, 42]. Nevertheless, we show here that our energy function, without explicit knowledge of the native structure, can capture these dominant binding modes for the full length proteins correctly, and thus predict them, and that the computational analysis allows us to elucidate how evolution has led to the necessary binding specificity. We analyze the effects by simulating shorter constructs that give insight into the individual roles of the binding partners in the binding process. The results from these simulations show that the free NLS polypeptide is thermodynamically guided to adopt a helix-turn-helix structure with the NLS itself forming the turn. Simulations of the IκB-bound NLS polypeptides show that the p50 NLS polypeptide does not interact specifically with IκBα, while the p65 NLS polypeptide is predicted to have several distinct binding modes, with the NF-κB(p50/p65)-IκBα crystal-structure-like conformation being one of them.

Simulation of the p65 NLS polypeptide interacting with IκBβ reveals two conformations, as found in the NF-κB(p65/p65)-IκBβ crystal structure. The simulations therefore make a new prediction; when the structure of NF-κB (p65/p65) bound to IκBα is determined, both NLS polypeptides will be bound on opposite faces of the IκBα.

The simulations with IκBα uncover a beautifully clear example of "induced fit", arguing for greater specificity and provide a rationale for nature's design scheme for NF-κB NLS polypeptides [43, 44]. The specific basic residues of the NF-κB NLS both interrupt the helical propensity of the signal and form crucial contacts with IκBα/β, bringing the helical portions into position to "cap" the ankyrin repeat domain.

4.2 Materials and Methods

4.2.1 Protein constructs and sequences

Simulations of the dynamics of several different constructs were performed using the associative memory hamiltonian (AMH) [76]. These constructs included the 30 residue NLS polypeptide of the p65 subunit (residues 291-320, chain C, PDB 1NFI) of the p65-p50 complex (1nfi [38], by itself, and when linked to the first three ankyrin repeats (residues 70-156, chain D, PDB 1NFI) of the ankyrin repeat domain of IκBα via a glycine linker. A truncated p65 fragment containing the NLS linked via an N-terminal glycine linker to the C terminus of the truncated IκBα fragment was also simulated, as well as a construct where the NLS polypeptide was replaced by the nucleoplasmin NLS. The initial configurations of these constructs were all built using the Biopolymer module of Insight II.

For the first construct, a simple glycine chain connects the C-terminal Lys_{320} of the NLS polypeptide to the N-terminal Ser_{70}. The advantage of using glycine residues is that they are much less sterically hindered than any other amino acid residue. The length and flexibility of the glycine chain allowed the NLS polypeptide to bind geometrically at any possible location on the IκBα protein surface. In later simulations, residues 192-320 of p65 and IκBα residues 70-156 were connected via a glycine linker connecting residue 156 of the C terminus of IκBα to residue 192 of the N terminus of p65. The purpose of this construct was to further investigate trap states in the folding and binding of the NLS polypeptide that might not be available when IκBα and p65 were arranged in crystal structure-like geometry. The same constructs with the p65 NLS polypeptide exchanged for the NLS polypeptide of p50 and for nucleoplasmin (PDB 1EE5) were also simulated.

4.2.2 Simulated annealing protocols with the associative memory Hamiltonian

The associative memory hamiltonian (AMH) is an energy function designed for making ab initio predictions of 3D protein structure from a given amino acid sequence. The AMH is an optimized energy function used for protein structure prediction even in the absence of homology information [76]. The terms of the full energy function used for the simulation contain besides the AMH sequence dependent interaction, V_{AM} for short-range and medium-range interactions, and $V_{contact}$ for long-range contact interactions, and excluded volume terms V_{ev} and basic backbone terms that include a potential $V_{\phi\psi}$, which provides a good fit of the backbone torsion angles found in a Ramachandran map, and hydrogen bonding patterns to assure correct physics and chemistry of the polypeptide chain. Here, the excluded volume potential is applied to the carbon and oxygen atoms that approach within 3.5 for $(j-i) < 5$, and 4.5 for $(j-i) \geq 5$. The chirality potential V_{chi} biases the peptide chain into the L-amino acid configuration. $V_{harmonic}$ contains three quadratic potentials along with shake constraints for the heavy backbone atoms to provide backbone rigidity. The total potential used for the AMH molecular dynamics simulations is given by:

$$V_T = V_{AM} + V_{contact} + \lambda_{\phi\psi} V_{\phi\psi} + \lambda_{ex} V_{ex} + \lambda_{harmonic} V_{harmonic} \tag{4.1}$$

The λ -terms scale the strength of interaction of the individual potential terms. The functional form of the terms in the potential has been described [76]. The AMH potential uses different interactions for pairs separated by different numbers of residues in the sequence. The short-range and medium-range potential applies to residue pairs less than 12 residue apart in the sequence. These predict formation of local secondary structure such as helices and turns. Residues that are more than 12 residues apart in the sequence interact via contact interactions that contribute to the collapse of the

protein and form tertiary structures from the shorter units. The equation for the associative memory term is:

$$V_{AM} = -\sum_{\mu}^{n}\sum_{i<j}^{N}\gamma(P_i, P_j, P_{i'}^{\mu}, P_{j'}^{\mu}\Sigma(r_{ij} - r_{i'j'}^{\mu}) \tag{4.2}$$

where μ runs over n memories. The parameters γ are learned by an optimization procedure [76]. The parameters are functions of P, where P represents the four-letter code designation assigned to each of the 20 naturally occurring amino acids [17]. The specific amino acids in each category are hydrophilic (Ala, Gly, Pro, Ser, and Thr), hydrophobic (Cys, Ile, Leu, Met, Phe, Trp, Tyr, and Val), acidic (Asn, Asp, Gln, and Glu), and basic (Arg, His, and Lys). The VAM potential encodes these sequence patterns by measuring the structural similarity to a list of memory proteins. This similarity is expressed functionally in the Σ function, which is a centered Gaussian that depends on the difference of distances in the simulated protein structure from those found in the memory protein. The contributions of the contact term to the total potential is given by a three-well potential:

$$V_{contact} = -\frac{\epsilon}{a}\sum_{i<j-12}\sum_{k=1}^{3}\gamma(P_i, P_j, k)c_k(N)U[r_{min}(k), r_{max}(k), r_{ij}] \tag{4.3}$$

These interactions are weighted by γ, depending on spatial distance and amino acid interaction type. The parameters in the potential are optimized using the quantitative form of the principle of minimal frustration, to yield the most funnel-like landscape for folding as possible, while maintaining transferability from one sequence to another. The details of the parameters of the potential have been described [76]. U is a contact function that controls the sharpness of the k = 3 well potential at the potential boundary endpoints rmin,max(k). The ck(N) terms are found from fitting the number of contacts in each of the regions as a function of sequence length of the target protein.

The physical principles of energy landscape theory apply to folding and to

binding processes. Here, the AMH energy function constructed originally for folding prediction is applied to a two-protein construct in which the binding partners are fused with a variable glycine linker. The results document that the AMH predicts the correct crystal structure of binding, and does allow some possible alternative, thermodynamically plausible, binding modes. The AMH short-range and medium-range interactions fold the local helices of the NLS polypeptide correctly, while the long-range potential contact interactions of the AMH can dock and bind the p65 NLS polypeptide to the surface of IκBα correctly.

The simulation protocols were as follows: initially, 130 annealing runs of the NLS/IκBα constructs were performed. An additional 150 annealing runs with linker lengths of five and 13 glycine residues, were performed to test for any dependence of the results on linker length. To investigate the traps, 60 additional annealing runs were carried out for all the other constructs. Each individual annealing run trajectory sampled 280 independent structures. This resulted in a total of over 100,000 structures available for analysis. In the annealing runs of the complexes, the NLS polypeptide was initially unfolded randomly and placed as far away from IκBα as possible. Initial random velocities were assigned to the protein. Temperature is quoted in units of the native state energy per residue obtained as an average over the memory terms containing only associative memory terms, which are short-range and medium-range in sequence, and the contact potential. Defining the scaled quantity ϵ, the native state energy of all the memory energy terms including contact terms, for a protein of N residues as:

$$\epsilon = \left(\frac{E_{AM+C}^{Native}}{4N} \right) \tag{4.4}$$

then a reduced temperature T^* can be defined with $k_B T = T^*$ [17]. All other energy terms, such as the backbone terms and excluded volume terms, are scaled to yield physically reasonable interaction strengths. The temperature T^*, where T^* 1 is of the order of the folding temperature, was then

reduced linearly from $T^* = 1.7 to T^* = 0.0$, resulting in trajectories of several hundred μs. This timescale was sufficient for both binding and folding of the NLS polypeptide to occur. A constraining potential assured that the three ankyrin repeats of IκBα/β did not change the topology of the backbone C^α atoms during the molecular dynamics annealing runs. The motivation for this constraint is based on the experimental fact that ankyrin repeats 2 and 3 of free IκBα show less hydrogen/deuterium exchange and thus higher protection factors when compared to the rest of the protein, suggesting an at least partially folded character of this region of the ankyrin repeat domain [46]. The side-chains of even the constrained part of the protein were free to move and interact with the flexible binding partner, which allows important interactions between the side-chains of the binding partners to occur.

4.2.3 Topological comparison

The overall topology of the protein fold and the secondary and tertiary structure of a protein can be monitored quantitatively by a variety of means. The RMSD is a standard way to probe the deviation between two structures but, since it is based on Gaussian statistics, it is best for very close structures. The fraction of overlapping structured pairs in two different structures is used as another measure of similarity, which is energetically relevant because the interactions are dominantly pairwise terms.

A useful, normalized quantity is Q, which is 1 for two identical structures and 0 for two structures that have no pair distances in common [25]. It is defined by:

$$Q = \frac{2}{(N-1)(N-2)} \sum_{i<j-1} \exp\left(-\frac{(r_{ij} - r_{ij}^N)^2}{2\sigma_{ij}^2}\right) \tag{4.5}$$

Here, the sum is over residue pairs that are separated by at least two residue in the sequence, with rij being the C^α distance of residues i and j, N is the total number of residues in the target structure and σ_{ij} is the Gaussian variance and is defined as (given in $\mathring{A}$) by $\sigma_{ij} = |i-j|^{0.15}$.

The CE Z-score obtained from a combinatorial extension (CE) algorithm is a statistical measure of how similar are the topologies and secondary structure of two proteins; a Z-score of 3.5-4 indicates possible biologically interesting similarities, naturally occurring proteins with structural Z-score greater than 4.0 are usually part of a single protein family or superfamily, and share the same overall topology and secondary structure [57].

4.2.4 Structural clustering analysis

The Fitch-Margoliash algorithm is a distance-based bioinformatics algorithm to fit a phylogenetic tree to a distance matrix [21]. The numerous structures obtained from the annealing runs were clustered using the Fitch program of the PHYLIP package. The Fitch program is an algorithm that is designed to create phylogenetic trees based on a distance measure. In order to analyze the structures obtained in the simulated annealing with the bioinformatics software, a distance measure d between two structures A and B was introduced as $d = 1 - Q$. Since Q is a normalized measure of the fraction of overlapping contacts, d is a measure of how dissimilar two structures are, which is equivalent to a distance measure, if distance is taken in the sense that similar structures with small d are close and dissimilar structures with large d are far away. In the calculation of Q, the contacts mediated by glycine residues were excluded. The clustering was carried out using different choices of Q. For example Qtotal takes into account all contacts, while Qinterface includes only the interfacial contacts between the nuclear localization signal and IκBα. Eastwood et al. point out that flexible fragments cause little change in the number of native contacts, but give rise to large fluctuations in RMSD [17].

4.2.5 Free energy calculations

Free energy profiles were obtained using the weighted histogram analysis method WHAM, which is a combination of free energy perturbation and umbrella sampling, and allows accurate and efficient calculation of the free

energy profile from a given set of simulations. The method was as follows: 17 constant-temperature molecular dynamics simulations were performed with a polynomial Q biasing potential of fourth order centered on different values of Q (Q = 0.1, 0.15, 0.2, ... ,0.9) to obtain good phase-space sampling along this reaction coordinate. During each simulation, 200 samples, Niobs, of Q and energy E, the backbone and AMH energy, were collected at regularly spaced time-points. The first 40 samples were discarded to allow for equilibration. A histogram Ni(E,Q) for all 17 simulations was created, which gave the density of states n(E,Q) of the system [17]:

$$n(E,Q) = \sum_i w_i(E,Q) \frac{N_i(E,Q)}{N_i^{obs}} \exp \beta_i E + \beta_i V Z_i(\beta_i) \tag{4.6}$$

with i being the index of simulation. The partition function is given by:

$$Z_i = \sum_{E,Q} n(E,Q) \exp -\beta_i E - \beta_i V \tag{4.7}$$

This allowed for self-consistent determination of the density of states n(E,Q) to within a multiplicative constant and, hence, the free energy was obtained to within a constant as:

$$F(Q,T) = -k_B T log \left(\sum_{E,Q} n(E,Q) \exp -\frac{E}{k_B T} \right) \tag{4.8}$$

The free energy function thus obtained is now a function of the desired order parameter Q and temperature T. Given a temperature, the free energy profile allows the identification of thermodynamically distinct states of the protein, such as the unfolded state ensemble.

4.2.6 B-factor calculations

The B-factor can be related to the mean displacement of a structure obtained in molecular dynamics simulations by the Debye-Waller equation $B = \frac{8}{3}\pi^2 \langle R^2 \rangle$ where $\langle R^2 \rangle = \langle (R_i - R_0) \rangle$ is the mean fluctuation in distance of the backbone atoms of structures sampled with the AMH relative to the refined crystal structure R_0.

4.2.7 Electron density maps

Reflection data of the PDB file 1OY3 was downloaded from the Protein Data Bank and converted into $2Fo - Fc$ electron density maps with the program CNS [58]. Electron maps were drawn with the PyMOL software package.

4.3 Results

4.3.1 Validation and benchmarking of the AMH method

This study of the IκB-NF-κB system aims to elucidate the folding and binding of a 30 residue long helical polypeptide, the NLS polypeptide, which becomes structured in the vicinity of its binding partner IκBα. In general, predicting folding and binding structures is a challenging problem when no experimental data are available. Here, we used molecular dynamics simulations with AMH as the energy function for identifying the binding interface, while folding the NLS [76]. This search problem is much simpler than total ab initio prediction, since the conformation of the binding partner, IκBα, is kept native-like but is allowed to fluctuate around the native basin with thermal energy kBT. The current study is then reduced to predicting the folding of the NLS onto a fixed protein surface. This constrained search problem is much easier than predicting folding and binding of two large proteins of unknown structure entirely from scratch. The inter and intra-residue interactions between the NLS and the binding partner IκBα are physically equivalent. Many meaningful protein dimer studies, such as G$\overline{\sigma}$-model studies, assign the same contact energies to inter and intra-residue contacts [45]. The AMH approximates a physically and chemically correct energy function and should, in principle, be appropriate to describe the complete folding and binding process of this simplified problem [76]. Water effects, which are important in protein binding, are only implicitly incorporated in the AMH. The NLS/IκBα binding interface, which has mostly hydrophobic native contacts between the NLS and

IκBα was found to be adequately described by AMH.

Several different binding and folding scenarios were simulated in the present work. In the first scenario, we present the folding of the NLS polypeptide onto the IκBα ankyrin repeat. Hydrogen/deuterium exchange experiments performed in the Komives laboratory show that the first three ankyrin repeats are protected and hence folded stably [46]. It has been established experimentally that the NLS polypeptide binds to these ankyrin repeats [47]. Here, we show that the present computation can reproduce the crystal structure of the NLS bound to IκBα when the two binding partners are connected by a glycine linker. This is not an unreasonable construct to study, as biology itself has these two parts connected by a glycine linker in the NF-κB precursor protein. We have varied the length of the glycine linker to confirm that the results are independent of linker length.

To understand the reliability of this approach, we studied whether we could successfully identify binding modes of comparable fragments in other systems. First, we constructed a shortened endonuclease dimer (PDB code 1m0i), where the two dimers were connected by a glycine linker. Eighty AMH annealing runs of the endonuclease with 75% of the protein constrained to be native-like resulted in only 17 structures misfolding, and all of the remaining structures representing the correct binding site after clustering using the Fitch-Margoliash method to generate a phylogenetic tree from the distance map, where distance is a measure of similarity (Figure 4.1(a)). The binding interface, as well as the fold of the simulated structures, resembles the crystal structure conformation found in the PDB. The AMH overfolds the endonuclear helix (green) and aligns it with a slight tilt when compared to the crystal structure (blue). The interface between the two endonucleases is predicted correctly with high overlap of native inter-residue contacts. The combinatorial extension (CE) algorithm calculates pairwise structure alignments. The CE Z-score provides a measure of statistical significance of the alignment, and structures with a Z-score of 3.5 or higher are generally judged to have a similar fold comparable to that obtained in a typical homology model [48]. The Z-score between individual

structures obtained from the molecular dynamics simulations are in the range of 3.3-3.7, again showing a great deal of similarity in overall topology and conservation of the helical secondary structural elements. The average root-mean-square deviations (RMSDs) from the crystal structure of the dimer were 2.3$\mathring{A}$ for the fragment only and 3.2$\mathring{A}$ the fragment and the interface contact residues, respectively.

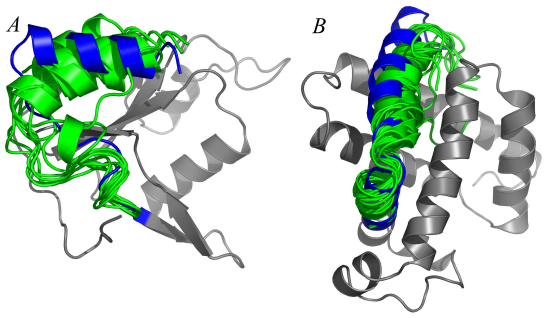


Figure 4.1: (a) Overlaid structures of endonuclease from simulation and experiment. Here, the restrained part is shown in grey, the crystal structure is shown in blue and the structures obtained with the AMH are shown in green. (b) Overlay of myoglobin structures from simulation and experiment. The coloring is the same as in (a).

Similarly, structures predicted for the folding of the N-terminal helix of 30 residues of myoglobin (PDB code 104m) when the rest of the molecule is constrained also showed profound topological agreement (Figure 4.1). In this case, first the helix was completely unfolded and allowed to dock anywhere on the protein surface of myoglobin and then 80 annealing runs were performed and the final structures were clustered. All structures obtained in these simulations docked at the correct binding site. The CE Z-score for these structures were in the range of 3.8-4.3. The RMSD from the crystal structure was $0.9\mathring{A}$ for the fragment and $1.5\mathring{A}$ for the fragment and the interface residues. These two test cases validate the present AMH approach for predicting binding conformations of modestly sized fragments like the NLS to pre-formed protein structures.

4.3.2 Predicted structure of the p50, p65, and nucleoplasmin NLS polypeptides

In order to obtain an overview of the possible structures that the free NLSs could assume when their binding partners are not present, we predicted the conformations of the free polypeptides using AMH. This study was motivated by the fact that crystallography shows the NLS polypeptide is disordered when p65 is bound to DNA with no interpretable electron density of the NLS polypeptide residues.11 Annealing runs carried out on the p65 NLS polypeptide (PDTDDRHRIEEKRKRTYETFKSIMK) alone indicated one family of structures with two helices that formed close contacts, connected by a loop (Figure 4.2). The first proline initiates the first helix of about five residues; Asp291-Arg295. The NLS ($Lys_{301}ArgLysArg_{304}$) was found in a bend and turn region. A second helix usually spanned from Glu_{307} to Met_{313}. The helices formed close contacts with a separation of about $7\mathring{A}$, with a predicted end-to-end-distance of about $5\mathring{A}$. The Q-score of 0.43, which measures the similarity of the annealed structures relative to each other, also confirmed that the structures were similar to each other (Figure 4.2(a)). The annealed

structures showed somewhat different orientations of the non-helical termini but, excluding the last two residues each of the N and C termini did result in Q-scores higher than 0.5. The Z-score between individual structures obtained from the molecular dynamics simulations are in the range of 3.3-3.7, again showing a great deal of similarity in overall topology. The secondary structure of the helical fragments was conserved. The average RMSDs for the first helix (residues 294-302) and the second helix (residues 305-314) from the NLS polypeptide in the crystal structure, were 0.87$\AA$ and 0.64$\AA$, respectively.

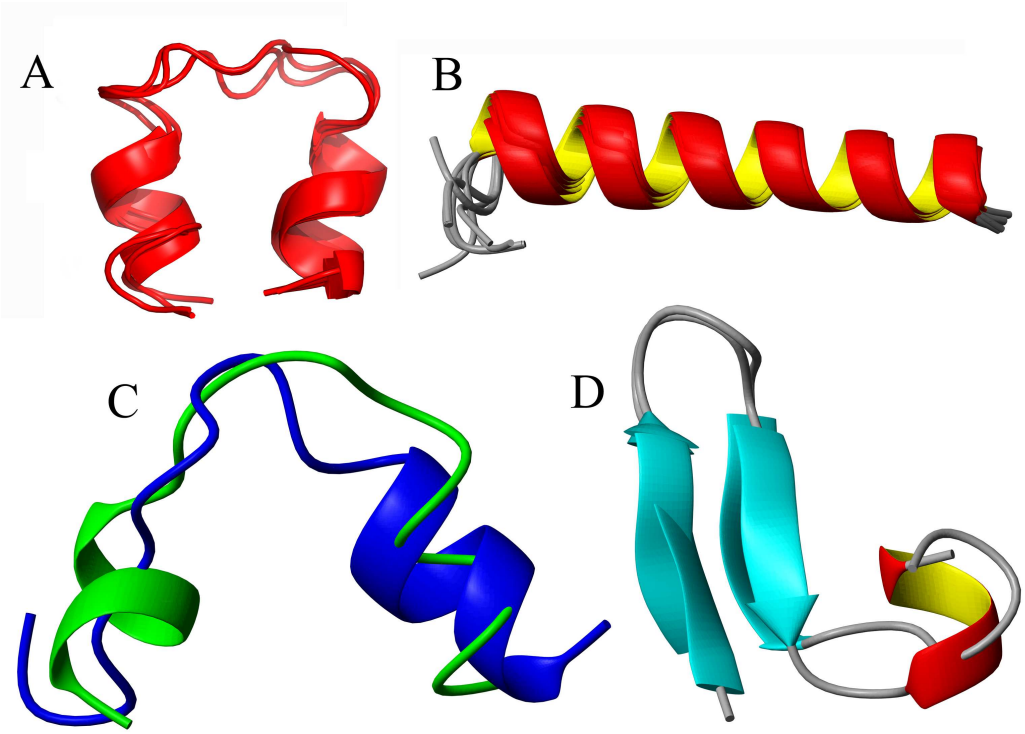


Figure 4.2: (a) Overlaid structures obtained in annealing runs to predict the structure of the p65 NLS polypeptide by itself. Two helices are connected by a breaking or kink region that contains the basic NLS residues. The structures are compact rather than elongated. (b) Structures obtained for the p50 NLS sequence folded uniquely into one straight α-helix. (c) and (d) The nucleoplasmin sequence showed frustrated folding, resulting in many structures with different folds and secondary structure content.

Structures for the p50 NLS (PLYYPEIKDKEEVQRKRQKLMPNFSDS-FGGGSGAG) and the nucleoplasmin NLS (AVKRPAATKKAGQAKKKKL) that were obtained in annealing runs following the same protocol as for the p65 NLS are also shown in Figure 4.2. All annealed structures of the p50 NLS were structurally equivalent to each other. Structural order parameters such as the CE Z-score indicate that all the predicted structures of the p50 NLS share the same fold, a long α-helix. This is interesting, because the sequence of the p50 NLS, although different from the p65 NLS, also has a basic central region, but in this case a break in the middle was not observed. The annealing runs suggested that the energy landscape of folding for the nucleoplasmin NLS sequence is more frustrated. The resulting structures vary tremendously in topology, some showing all-α secondary structures as well as $\alpha - \beta$ secondary structures. The simulations showing that the nucleoplasmin NLS adopts an extended structure agree with experiments, where the nucleoplasmin NLS binds to the import factor importin in an extended structure with hydrogen bonding interactions [40].

4.3.3 Folding of the IκBα-NLS construct

Multiple (130) annealing runs were performed using the AMH algorithm to obtain structures of the IκBα-NLS construct linked by nine glycine residues. The linker assures proximity of the NLS to IκBα in the simulation. The use of a glycine linker is biologically justified for the NF-κB/IκB system, because the precursor p105 in fact contains NF-κB and IκB linked by a glycine-rich linker region.24 Structures with various degrees of similarity to the crystal structure were obtained from these simulations. The Fitch-Margoliash method was used to cluster these structures according to their similarity as measured in distance, d, and a phylogenetic tree was created from the distance map. This phylogenetic tree showed four main clusters of structures with low relative d, and hence significant similarity. The high relative Q-scores within each cluster also indicate similar global folds that represent distinct structural families or specific binding modes. The center of each branch was chosen as the rep-

resentative structure for that individual cluster (Figure 4.3). Cluster 4, the dominant cluster with a third of the structures, is composed of those conformations that most resembled the X-ray crystal structure based on the Q-score that included contacts of the NLS polypeptide along with contributions from the interface of the NLS polypeptide with IκBα. Clusters 1, 2 and 3 all had lower Q-scores of about 0.25 relative to the crystal structure.

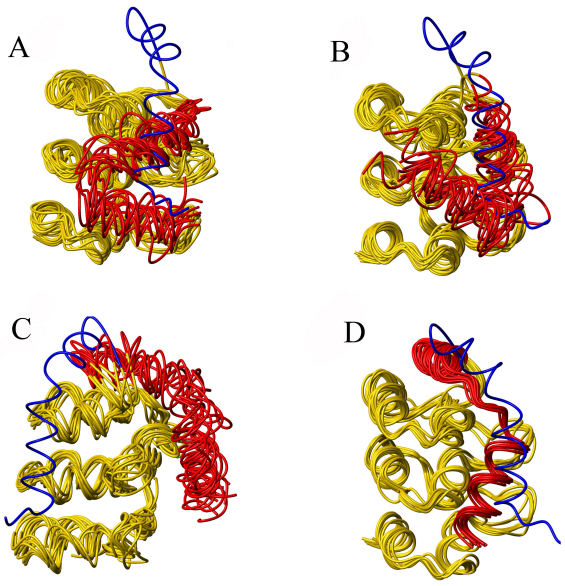


Figure 4.3: A phylogenetic tree was created for the structures obtained in the annealing runs of the IκBαp65 NLS construct using the Fitch-Margoliash algorithm. The tree showed clustering in (a)-(d) four main groups, which were analyzed as possible binding modes. The inhibitor IκB$\alpha$is colored yellow and the p65 NLS polypeptide is colored red. For comparison, the crystal structure conformation is shown in blue. (a) and (b) Clusters 1 and 2 correspond to overly compact structures that folded and then docked onto the protein surface of IκBα. (c) The cluster 3 (C) structures form a basin of "symmetric" structures, in which the NLS polypeptide docked at an alternative, biologically possible, protein surface of the inhibitor. (d) The cluster 4 structures are structurally close to those found in the X-ray crystal structure [38].

Table 4.1: Summary of RMSD for simulated constructs

Proteins	Structure	NLS	Interface	NLS+Interface
p65NLS+IκBα	Basin 1	4.9	5.9	6.4
p65NLS+IκBα	Basin 2	5.9	6.6	7.2
p65NLS+IκBα	Basin 3	3.4	6.7	6.2
p65NLS+IκBα	Basin 4	1.7	2.1	2.1
p65NLS+IκBα+p65 remainder	Basin 1	2.2	2.4	2.8
p65NLS+IκBα+p65 remainder	Basin 2	3.5	18.3	22.4
p65NLS+IκBβ	Basin 1	1.7	1.2	2.1
p65NLS+IκBβ	Basin 2	4.6	5.2	5.4

Cluster 4 had a Q-score of 0.42 of the basin center with respect to the crystal structure (Figure 4.3(d)). The cluster had an overall RMSD of $0.33(\pm 0.05)\mathring{A}$ showing very high overlap between structures. The first helix (residues 294-302) and the second helix (residues 305-314), of the clustered structures had even lower RMSDs, $0.19(\pm 0.14)\mathring{A}$ and $0.13(\pm 0.03)\mathring{A}$, respectively. Further, the CE Z-scores of the individual members of the cluster relative to the cluster center were about 5, indicating essentially identical folds and justifying the use of the cluster center as representing the structure of the cluster. Comparison of the structure from the center of basin 4 relative to the X-ray crystal structure gave an overall RMSD of the NLS residues along with the interface residues (see Materials and Methods) of 3.1Å (see also Table 1). The secondary structure elements were native-like, with an RMSD of 0.57Å for the first helix and 0.34Å for the second helix. Similar native-like features of the structures were obvious. These included the observation that the second helix caps the top of IκBα and the first helix binds appropriately to the hydrophobic fingers of IκBα. The RMSDs from the crystal structure for the structures in basin 4 are smaller than the fluctuations inferred from the measured B-factors (PDB file 1NFI) for residues 293-302. Thus, the structures predicted in this basin are consistent with observation. In the simulations, the second helix packed somewhat more closely to the hydrophobic top of IκBα than would appear to be the case in the crystal structure (Figure 4.3(d)). Clusters 1-3 had the same Q-score to the native structure, but were distinct clusters having less similarity to each other. When only atoms from the NLS were used to calculate the RMSD from the crystal structure, the results for clusters 1-3 were 4.9Å, 5.9Å and 3.4Å, respectively. The RMSDs of the two helical fragments of basin 3 structures from the corresponding crystal structure fragments were 1.1Å and 2.2Å, respectively. Thus, the basin 3 NLS had a fold similar to that of the crystal structure; however, when the interface residues were included in the calculation, the RMSD was 7.5Å, indicating clearly that basin 3 structures (Figure 4.3(c)) bind differently from the mode observed in the crystal structure. The NLS still binds to the IκBα β-hairpin fingers, and also still caps the hydrophobic top; however, the NLS polypeptide binds on the

other side of the β-hairpins of the ankyrin repeat domain than was seen in the crystal structure. Clusters 1 and 2 constituted about 55% of the observed structures. Neither of these clusters was as homogeneous as cluster 4 in terms of CE Z-score; however, the secondary structure of the helical fragments was preserved. The structures belonging to these clusters exhibited no capping of IκBα by the NLS polypeptide (Figure 4.3(a) and (b)). In the annealing simulations, structures arriving at these binding modes appeared to fold first to an overly compact structure resembling the folding found for the unbound NLS polypeptide, and this configuration later docked onto the protein surface of IκBα.

4.3.4 Important contacts between the NLS polypeptide and IκBα

Contact maps of the four clusters were constructed to identify the most important residues involved in the NLS recognition of IκBα (Table 2). As expected, the cluster 4 structures formed the largest number of native contacts between IκBα and the NLS polypeptide, while clusters 1 and 2 had only a few residues with high native contact probability, and the structures in cluster 3 showed no native contact formed at all. Clusters 1 and 2 contained mainly non-native contacts formed between helix 2 (residues 305-314), and IκBα. In these binding modes, many contacts were observed between helix 2 and IκBα, whereas helix 1 and the NLS made no contacts with IκBα at all, except for the native contact between Ile_{298} of p65 and Ile_{120} of IκBα.

Table 4.2: Important contacts for basins 1-4

	NLS	IκBα	NLS	IκBα
Basin 1	Ile298	Ile120	Phe309	Ile83
			Phe309	Ile117
			Phe309	Leu120
			Met313	His79
			Met313	Leu80
			Met313	Ile83
Basin 2	Ile298	Ile120	Phe309	Ile83
			Met313	Leu80
			Met313	Ile83
Basin 3			Tyr306	Phe87
			Phe309	Phe77
Basin 4	Asp294	Ile120	Glu299	His84
	Arg295	Ile120	Lys301	Ile112
	Arg295	Thr121	Arg302	Leu80
	Ile298	Ile83	Tyr306	Leu80
	Ile298	Leu117	Phe309	Phe77
	Ile298	Ile120	Phe309	Phe103
	Arg302	Ile83	Ile312	Phe103
	Phe309	Leu80	Met313	Leu78
			Met313	Val93

Highly probable non-native contacts involved interactions of the very hydrophobic residues Phe_{309} and Met_{313} with the also very hydrophobic residues Leu_{80} and Ile_{83} found in the first ankyrin repeat, and with Leu_{117} and Ile_{120} in the second ankyrin repeat. Cluster 3 contact maps showed that Phe_{77} in the first ankyrin repeat very often interacts strongly with residues Tyr_{306} and Phe_{309}, even though these interactions are not native contacts. Cluster 4 structures showed a high probability of forming native contacts only in helix 1. The strong native contacts of helix 1 residues Ile_{298} and Arg_{302} were complemented by non-native contacts of the specific NLS residues, Lys_{301} to Gln_{112} and Arg_{302} to Leu_{80}. Additionally, Glu_{299} formed a non-native contact with His_{84} with high probability. An important residue in helix 2 is Phe_{309}, which forms a native contact to Leu_{80} as well as non-native contacts to Phe_{77} and Ile_{94}, while Met_{313} was found to form non-native contacts to Leu_{78} and Val_{93}. On average, the NLS formed about 43% of its native interfacial contacts with IκBα. The interfacial RMSD from the crystal structure for the residue-residue pairs of the cluster 4 structures that have backbone heavy-atoms within 5Å of each other, are in the range of 3-4Å. This would be considered an acceptable result using the CAPRI criterion for identifying binding surfaces [50].

4.3.5 Folding of the IκBα-NLS-p65 construct

In the constructs studied so far, the NLS polypeptide could associate freely with IκBα. Several possible binding modes for the free NLS polypeptide were found. One of these is the native binding mode found in the crystal structure. On inspection, the alternative binding modes appeared sterically incompatible with the presence of the remainder of the p65 molecule. To confirm this explicitly, we simulated constructs that contain the remaining p65 residues and limit the geometrical space accessible to IκBα. The simulation runs with the IκBα-NLS-p65 construct, in which the relative geometry of p65 and IκBα was constrained to be like that in the crystal structure, exhibited two clusters of structures. The dominant cluster (65% of structures) resembles the binding mode found in the crystal structure (Figure 4.4) and it appeared that the

p65 helped the N-terminal helix of the NLS find its proper location. Besides yielding the X-ray crystal structure-like conformation, the simulations yielded the second cluster (35% of structures) of conformations in which the fragment of the p65 NLS binds to the main body of the p65 molecule (Figure 4.4). Thus, when the p65 dimerization domain is present, structures with an overly compact NLS polypeptide are no longer observed, although these were found in the simulations of the smaller construct. The structures of the self-interacting cluster had an RMSD from the crystal structure of 3.5Å for the NLS only, showing that the NLS adopted a similar extended structure. However, when the interface residues were included, the RMSD from the crystal structure was 22.4Å stemming from the fact that the NLS is bound to p65, not to IκBα. No basin 1, 2, or 3 structures was observed, presumably because the remainder of the p65 molecule geometrically limited the binding modes to either those resembling the crystal structure, or a self-interacting mode.

4.3.6 IκBα interactions with the nucleoplasmin NLS

To investigate whether the binding of the NLS polypeptide to IκBα is specific or whether other NLSs could function similarly, simulations of the p65-IκBα construct were performed in which the p65 NLS polypeptide was replaced by the NLS polypeptide of nucleoplasmin, a molecular chaperone whose major function is involved in the assembly of nucleosomes.

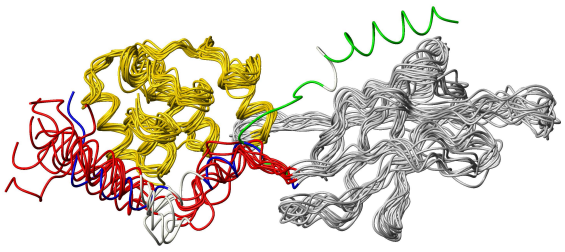


Figure 4.4: Results from simulation of IκBα-NLS-p65 show geometrical restriction of the possible binding modes. The dimerization domain of the NF-κB p65 (grey) and the IκBα(yellow), were kept fixed in the simulations. The structure of the NLS polypeptide from the crystal structure of the complex of IκBα with NF-κB is colored blue. The main cluster of structures obtained in molecular dynamics simulations with the AMH as energy function reproduced the native structure well (red). The light-grey region in the NLS polypeptide indicates the basic NLS residues, which form a break between the two helices. Some structures were formed in which the NLS polypeptide binds to the p65 dimerization domain. The center of this cluster is colored green.

The structure of the nucleoplasmin NLS polypeptide bound to importin α has been solved [40]. It is of comparable size to the p65 NLS polypeptide but the level of sequence identity is only 6.2%. The AMH simulations of the nucleoplasmin NLS tethered to IκBα and p65 yielded structures in which the NLS polypeptide did not interact with either IκBα or with p65 (data not shown). Thus, AMH simulations suggest that the basic stretch of residues, which has traditionally been the minimal description of the NLS, is insufficient for binding in the IκB/NF-αB system.

4.3.7 Specific effects of the basic NLS residues on IκBα recognition

The basic NLS residues (301-304) were seen to break the helical secondary structure of the p65 NLS in all simulations. Since these residues are required for importin α binding, we sought to define their role in binding to IκBα. Simulated annealing runs were performed on five different IκBα-p65-NLS constructs in which there were alanine substitutions in the NLS. Single mutants; K301A, R302A, K303A and R304A and the quadruple mutant with all four alanine mutations in the NLS were studied. The structures obtained in simulated annealing runs of the K301A and R302A single mutants retained almost wild type amounts (55%) of helical secondary structure. The K303A and R304A mutants showed increased helical contents of up to 77% and formation of one continuous helix with a high binding propensity towards the dimerization domain of p65. The K301A mutant formed mainly long helices in the simulated annealing runs, but the R302A mutant showed a kinked region similar to the crystal structure. None of the mutants produced a structure with native-like topology, although some structures did bind to IκBα.

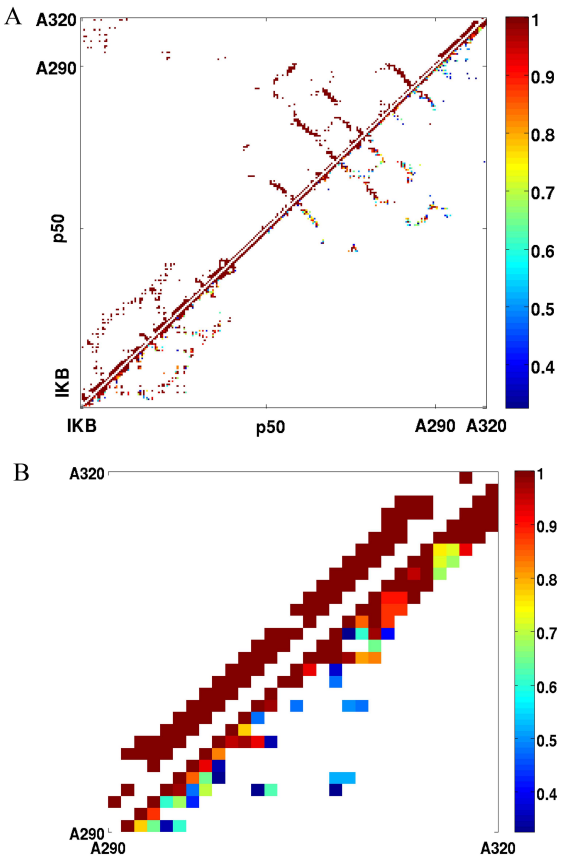


Figure 4.5: Alanine mutations in the p65 NLS polypeptide were used to probe the role of individual residues in folding and binding of the p65 NLS polypeptide to IκBα. The contact map of the structures obtained in the simulation runs (shown below diagonal) of the Lys302Ala mutant are shown and can be compared to the contact map of the crystal structure (shown above diagonal). Contact probabilities are colored red, indicating a contact that is always formed in the ensemble, while dark blue means that the contact is almost never formed. The complete contact map of the NLS, (a) the p50 remainder and the inhibitor as well as (b) the NLS only are shown.

The contact map analysis shows that mutation of R302 to Ala disrupts important native contacts of the first helix to IκBα, perhaps indicating that contacts between the first helix and IκBα are essential for correct complex formation. (Figure 4.5). To test the idea that the break in the middle of the helix is important, the R304A mutant was made as a peptide spanning residues 289-321 of NF-κB (p65) as well as in the full-length p65. Binding of the peptide was tested by isothermal titration calorimetry and binding of the full-length protein by Biacore. In both experiments, the R304A mutation decreased the binding affinity by twofold and, interestingly, the DCp was much lower for the R304A mutant than for wild type perhaps indicating a change in structure formation during binding (S. Bergqvist, unpublished data).

4.3.8 Binding of the p65 NLS polypeptide to IκBβ

Biochemically, IκBα and IκBβ share the same function and are able to replace each other in vivo [51]. The crystal structure of IκBβ bound to NF-κB(p65/p65) shows a similar binding site provided by IκBβ as compared to the binding site provided by IκBα in the complex with NF-κB(p50/p65). A second binding site is observed also in the crystal structure of IκBβ bound to NF-κB(p65/p65) with weaker electron density [52].

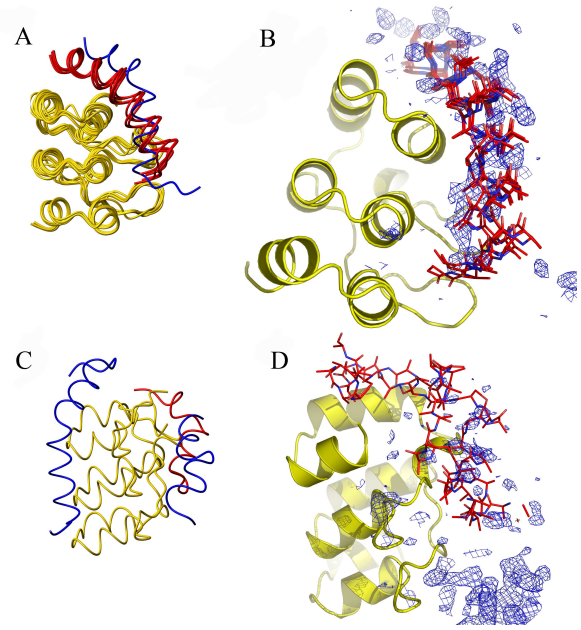


Figure 4.6: The phylogenetic tree was obtained using the Fitch-Margoliash distance-based algorithm for structures obtained in AMH simulations of the p65 NLS polypeptide bound to IκBβ. The tree shows clustering into one main group and one less prevalent alternative binding mode. (a) The dominant cluster of structures was obtained for the interaction between the p65 NLS polypeptide (red) and IκBβ(yellow). For comparison, the structure of the p65 NLS polypeptide structure from the NF-κB(p65/p65) IκBβcomplex is shown (blue) [52]. (b) Electron density map (blue) of the NLS polypeptide for $2F_0-F_c$ density using model-derived phases for 1oy3 (PDB). IκBβ is shown in cartoon style in yellow, while the p65 NLS polypeptides obtained in AMH simulations are shown as red sticks. (c) The alternative symmetric binding site of the NLS polypeptide was sampled and compared reasonably well to (d) the existing electron density derived from the crystal structure.

AMH simulations of the 30 residue p65 NLS polypeptide interacting with IκBβ yielded only one main basin of structures, in contrast to the results for IκBα. In this case, it was not necessary to include the rest of the dimerization domain of p65. The structures contained nearly all of the contacts found in the crystal structure (Figure 4.6) [52]. In 90% of the cases, the p65 NLS polypeptide assumes the crystal structure-like conformation corresponding to the strong electron density. The RMSD of the cluster center from the crystal structure considering the NLS residues alone was 1.7Å . When the interface residues were also included, the RMSD was 2.1Å. A few (8%) of the cases sampled alternative structures comparable to the cluster 3 structures seen in the simulations of the p65 NLS with IκBα and similar to the minor electron density in the crystal structure. The RMSD from the crystal structure model of this weaker binding site to the cluster center for this alternative binding mode was 4.6Å for the NLS only, and 5.4Å for the NLS and interface residues (Figure 4.6(d)). For comparison, the RMSD from the alternative binding site of the basin 3 structure was 4.6Å and 6.1Å, respectively, for NLS only and NLS with interface residues. The structure of this alternative binding site was very similar to the cluster 3 binding mode, yielding a 1.2Å RMSD from each other, and 2.8Å when comparing the deviation of the NLS and the interface residues.

Since NF-κB dimers can be homo- or heterodimers composed of both p65 and p50, we performed AMH simulations of the folding and binding of the p50 NLS polypeptide to IκBα. Simulations of the p50 NLS tethered to just IkBa gave two main clusters of structures; 32% of the structures were similar to basin 4 in the p65 NLS simulations, and the remainder were similar to basin 2 structures. When the p50 dimerization domain was included to geometrically limit the search space of the p50 NLS, some 75% of these failed to adopt stable strong interactions between the p50 NLS polypeptide and IκBα whereas the presumed "native" structure was found in the remaining 25% of the structures. Simulations of the p50 NLS polypeptide interacting with IκBβ gave no preferred structures, with less than 25% of the structures falling into any one

basin. The results suggest that the p50 NLS polypeptide can bind to IκBα with a structure similar to that observed for the p65 NLS bound to IκBα but the energy landscape is not as funneled as for the p65 NLS.

4.4 Discussion

4.4.1 The p65 NLS polypeptide has a high propensity to form helical structure

Folding simulations performed with the AMH on the p65 NLS polypeptide by itself showed that the p65 NLS polypeptide has a significant propensity to fold into a helix. The p50 NLS polypeptide also had helical propensity but did not contain the kink required for correct binding to IκBα/β. The structural role of the KRKR NLS sequence in p65 must break up this helical structure into two helices by forming a kink region. The p50 NLS contains only three basic residues flanked by Gln, and this slightly different sequence does not appear to be sufficient to break the helical structure. Binding of the polypeptide to the inhibitor requires the polypeptide to take on a fold of two helices connected by a break region. The two helices of the NLS polypeptide each bind to IκB, where the first helix forms stable contacts with the ankyrin repeats, while the second helix caps the hydrophobic top of the inhibitor. Our simulations show one role of the NLS sequence is to provide a break in the helical secondary structure for its interaction with IκB and help to explain the weaker binding affinity of the p50 NLS polypeptide [36]. The nucleoplasmin NLS polypeptide did not form one unique structure in the AMH simulations, and showed mostly extended structure consistent with the extended structure it adopts when bound to importins [40]. These results beg the question of whether the NF-κB NLS polypeptides bind to importins in an extended structure and therefore must be unraveled for binding, or whether they bind in a helical conformation.

4.4.2 Use of AMH to predict binding conformations

An important aspect of the present computer simulation study of binding and folding is that reasonable predictions of the correct binding/docking site rely on the calibration of the energy function as reliable by examining other structurally well defined binding situations. Blind trials performed in CASP have already allowed us to have an unbiased evaluation of the of the ab initio AMH energy function. These comparisons have confirmed the predictive power of the AMH to obtain the folded state of largely helical proteins up to not, vert, similar180 residues with high fidelity (M.C.P. and P.G.W., unpublished results). By constraining the problem using glycine linkers and keeping the binding partner fixed, we show here that this same energy function has predictive power for simple protein-protein interactions that involve folding as well as binding of surface helices.

Although the AMH energy function is not generally useful for inferring binding propensities, it is interesting to note that the folding of NF-κB NLS polypeptides upon binding to IκBβ could be compared directly to electron density maps for the IκBβ complex with NF-κB(p65/p65) (PDB 1oy3) in which two binding sites of the p65 NLS polypeptide are observed, one with strong electron density (Figure 4.6(a) and (b)) and a second with weak electron density (Figure 4.6(c) and (d)). The simulations are consistent with the electron density results, with some 90% of the structures binding to the site with the stronger electron density. while the rest of the predicted structures closely resemble the binding site with weaker electron density.

4.4.3 Prediction of a second p65 NLS polypeptide binding site on IκBα

Ernst et al. showed that one molecule of IκBα binds to one NF-κB dimer containing two NLS polypeptides [53]. Our simulations of the p65 NLS binding to IκBα find a major cluster that corresponds to the conformation found in the

IκBα-NF-κB(p50/p65) crystal structure and a second symmetric binding site similar to that seen in the binding to IκBβ that was observed crystallographically. The folding of NF-κB NLS polypeptides upon binding to IκBβ could be compared directly to the crystal structure of the IκBβ-NF-κB(p65/p65) homodimer [52]. The p65 NLS polypeptide binds strongly to IκBβ in the crystal structure-like conformation (Figure 4.6). Comparison of electron density maps for the IκBβ-NF-κB(p65/p65) (1oy3.pdb) with the structures obtained in our simulations demonstrate the power of the AMH to predict the location of both the strong (Figure 4.6(a) and (b)) and weak (Figure 4.6(c) and (d)) binding sites. The simulations show that the landscape of binding of the p65 polypeptide to IκBβ is strongly funneled, with some 90% of the structures binding to the site with the stronger electron density.

Thus, a new prediction made from this work is that when the X-ray crystal structure of IκBα with NF-κB(p65/p65) is completed, there will be two NLS binding sites observed on IκBα as were observed on IκBβ. This prediction awaits confirmation when the crystal structure is determined, but is consistent with the speculation that IκBα might mask both NLS polypeptides, resulting in more complete cytoplasmic localization of the p65/p65 homodimer [54].

In contrast to the p65 NLS polypeptide, the p50 NLS polypeptide interaction energy landscape was less funneled. These results are consistent with the observation that the p50 NLS polypeptide does not contribute to the binding energy of the NF-κB(p50/p65)/IκBα complex [36], and that the p50 NLS remains unbound and is responsible for the NLS-dependent import into the nucleus, resulting in continuous shuttling of the NF-κB(p50/p65)/IκBα complex [42, 54, 55]. In some simulations, the p50 NLS polypeptide interacted with the "native-like" binding site on IκBα, and this result is consistent with the report that the p50 NLS is required for NF-κB(p50/p50) binding to IκBα [56]. This prediction may help resolve a controversy in the NF-κB field because previous biochemical data have given the impression that IκBα and IκBβ are distinct largely because of how many NLS sequences they could sequester. In contrast, experiments in transgenic mice indicated that the two

proteins were functionally equivalent [51]. Our results predict that they are functionally equivalent, even in the number of NLS sequences that they sequester, and that the biochemistry was misleading due to an incomplete set of dimers tested.

4.4.4 The p65 NLS polypeptide finds its correct binding site on IκBα/β in the absence of the rest of the p65 molecule

We have not examined explicitly how IκB and DNA compete for binding to NF-κB. The simulations suggest that the NF-κB subunit NLS polypeptide, independent of the rest of the NF-κB molecule, is capable of spontaneously undergoing a disordered to ordered transition upon binding to the inhibitor. The simulations further predict that it is the p65 subunit NLS polypeptide that conveys the binding specificity observed between the IκBα and NF-κB subunits. It is possible that even when the NF-κB is bound to DNA, the unstructured p65 NLS polypeptide may interact with IκBα, ultimately facilitating the disassembly of the enhanceosome complex. The IκB/NF-κB system represents a beautiful example of how induced fit is used by Nature to achieve control of binding specificity through structural diversity.

Appendix

We thank Gourisankar Ghosh, Tom Huxford and Alexander Hoffmann for helpful discussions and critical reading of the manuscript. We thank De Bin Huang for help with creating the electron densities of the bound NLS. Support from the La Jolla Interfaces in Science (LJIS) and the R01 GM44557 NIH grant is gratefully acknowledged. J.L. thanks the National Science Foundation-sponsored Center for Theoretical Biological Physics grants PHY-0216576 and 0225630, and the CTBP for computer time on their cluster.

The text of this chapter, in full, is a reprint of the material as it appears in the Journal of Molecular Biology. The dissertation author was the primary researcher and author. The right to republish this work is retained and permitted without the need to obtain specific permission from Elsevier.

5 Consequences of frustration for the folding mechanism of the IM7 protein

5.1 Introduction

For proteins to fold to a unique state, they have evolved to form highly favorable, stabilizing native interactions. This statement suggests that proteins may have a funneled energy landscape [3]. However, real proteins should display some landscape ruggedness owing to the possible existence of favorable, non-native interactions during the folding process. If the landscape were too rugged, numerous long-lived intermediate states that energetically compete with the native state would appear. The rarity of such intermediates having significant non-native structure suggests that proteins largely conform to the principle of minimal frustration [1].

When a populated kinetic folding intermediate is observed, it is generally the case that the meta-stable thermodynamic state is not caused by energetic frustration but by a nonuniform compensation of the entropy and energy changes upon forming contacts. The existence of such productive folding intermediates can be said to be topology driven and can often be predicted from the native structure alone using a perfectly funneled landscape.

In contrast to this simple description, IM7, an 86 residues protein, is known

to fold through an on-pathway intermediate state that possesses a non native packing of three of the four helices around a specific hydrophobic core [121]. In this paper we explore the relationship between non-native interactions, the presence of the stable intermediate and the residual frustration found in the native structure. Our analysis suggests that this seeming exception to the pattern expected for a minimally frustrated protein is in fact consistent with the ideas of energy landscape theory.

We first demonstrate that the intermediate of IM7 is not a direct consequence of the topology of IM7 by simulating the folding with a sequence independent energy function which yields a perfectly funneld landscape (a so-called Gō Hamiltonian). Not only was no intermediate observed with this Hamiltonian, but none was found when contact energetic heterogeneity or non-additivity through many-body interactions were included. These more sophisticated Hamiltonians do not predict the existence of a stable intermediate. These "negative" results suggest that the experimentally observed intermediate is indeed a consequence of frustration rather than a consequence of the topology of IM7.

Can we establish how frustration changes the formation of the IM7 intermediate? To study this, we performed simulations with a transferable Hamiltonian of the type used to predict protein structure de novo. This relatively realistic Hamiltonian not only predicts the existence of an intermediate ensemble which is stabilized by non-native interactions, but also gives a predicted structure of the intermediate that compares well to experiments by Capaldi et al. [121], who have infered the existence of a three helix core of helix I, II and IV which while having numerous native-like interactions is also stabilized by some non-native interactions.

Equipped with an energy function that predicts the experimentally observed intermediate we can examine more closely the role frustration plays in the folding mechanism. Based on the same energy function we use the algorithm of Ferreiro et al. to compute local frustration for all sites in the native structure and the intermediates and predict mutants, that should reduce the

level of frustration so as to make the landscape less rugged and more funneled. We contrast two different re-design schemes, one based on the native state structure that alleviates the frustration in that structure and another re-design that attempts to specifically destabilize the intermediate states. The predicted mutants were simulated with the AMW Hamiltonian. The results show that the intermediate is indeed a sympton of residual native-state frustration.

5.2 Simulation Methods

Two different Hamiltonians are used to elucidate the folding properties of Im7. We first used an off-lative Gō-like model of the type introduced by Onuchic and coworkers [122] to investigate purely the role of topology on folding. We then performed molecular dynamics simulation with the AMW Hamiltonian [89]. This Hamiltonian allows us to obtain details on the role of non-native contacts as well as water mediated interactions at a molecular level. The AMW Hamiltonian and the simulation protocols are described.

5.2.1 Native Topolgy-based Simulations

In the first part of our study, we used a $C_\alpha$native topology-based model where a single bead centered on the C_α position represents a residue, as described previously [123], with minor changes to introduce native energetic heterogeneity. In this model, the bond and angle potentials string together the beads to their neighbors along the protein chain. The dihedral potential encodes the secondary structure. The proteins native topology defines the network of favorable long-range tertiary interactions while all other non-bonded interactions are repulsive. The energy function for a C_α native topology-based model with configuration Γ is as follows:

$$\mathcal{H}(\Gamma, \Gamma_0) = \mathcal{H}_{bb} + \mathcal{H}_{nb} \tag{5.1}$$

$$\mathcal{H}_{bb} = \sum^{bonds} K_r(r - r_0)^2 + \sum^{angles} K_\theta(\theta - \theta_0)^2 + \sum^{dihedrals} K_\phi^n[1 - cos(n(\phi - \phi_0))] \tag{5.2a}$$

$$\mathcal{H}_{nb} = \sum_{i<j-3}^{native} \epsilon_1(i,j) \left[5 \left(\frac{\sigma_{i,j}^{nat}}{r_{i,j}} \right)^{12} - 6 \left(\frac{\sigma_{i,j}^{nat}}{r_{i,j}} \right)^{10} \right] + \sum_{i<j-3}^{nn} \epsilon_2(i,j) \left(\frac{\sigma^{non}}{r_{i,j}} \right)^{12} \tag{5.2b}$$

where backbone, nonbonded and non-native are abreviated as bb, nb and nn respectively. The K_r, K_θ, and K_ϕ are the force constants of the bonds, angles and dihedral angles, respectively. The r, θ, and ϕ are the bond lengths, the angles, and the dihedral angles, with a subscript zero representing the corresponding values taken from the native configuration, Γ_0. The non-bonded contact interactions, $\mathcal{H}_{nonbonded}$, contain Lennard-Jones 10-12 terms for the non-local native interactions and a short-range steric repulsive term for the non-native pairs, corresponding to a perfectly funneled energy landscape. We chose as parameters of the energy function $K_r = 100\epsilon$, $K_\theta = 20\epsilon$, $K_\phi^1 = 1.0\epsilon$, and $K_\phi^3 = 0.5\epsilon$. The energetic weights are defined as $\epsilon_1 = \epsilon_2 = \epsilon$ in the homogeneously weighted native topology-based model. Energetic heterogeneity is introduced by having ϵ_1 equal the value of the corresponding weight according to the set of energetic weights of the Miyazawa-Jernigan potential [128], divided by the average value such that the resulting sets average is equal to ϵ. $\sigma_{i,j}^{nat}$ is the distance between the C_α atoms of the residues (i,j) in the native configuration and $\sigma^{non} = 3.0\mathring{A}$ for all non-native residue pairs. The network of native contact pairs was determined using the CSU (Contacts of Structural Units) software [129]. Multiple trajectories with numerous unfolding/folding transitions were collected and analyzed using the weighted histogram analysis method (WHAM) to calculate the free energy surface projected onto the fraction of native contacts (Q). This reaction coordinate was

previously demonstrated to accurately map to P_{fold} at the resolution of Φ-values for a funneled energy landscape [25]. We incorporated nonadditivity, which implicitly accounts for sidechain and solvent interactions ordinarily absent from the pairwise additive model, into calculations of free energy profiles by following the protocol of Plotkin and co-workers [124].

5.2.2 Molecular Dynamics Simulations with the AMW Hamiltonian

The AMW Hamiltonian is a coarse-grained, transferable potential designed to predict the global native fold of proteins. The Hamiltonian is general and contains a 20×20 contact potentials for direct and water mediated contacts that reflect modulation by the local environment. The basic mathematical form of the AMW is given by

$$\mathcal{H}_{AMW} = \mathcal{H}_{bb} + \mathcal{H}_{AM} + \mathcal{H}_{R_G} + \mathcal{H}_{contact} + \mathcal{H}_{water} + \mathcal{H}_{burial} \qquad (5.3)$$

and applies to a reduced set of coordinates of the heavy backbone atoms, C^{α}, C^{β} and oxygen. In this reduced description, the positions of the nitrogen and $C^{'}$ carbons are calculated assuming ideal protein backbone geometry. The Hamiltonian assures correct backbone chemistry and collapse of the protein. The functional forms of the individual terms of the Hamiltonian are explained in greater detail by Papoian et al. [89]. We note that for residues less than 12 apart in sequence, the Associative Memory (AM) term applies while for residues separated by more than 12 in sequence, the contact potentials, $\mathcal{H}_{contact}$, $\mathcal{H}_{water}$ and $\mathcal{H}_{burial}$, apply. The AM term [17, 126] captures local structural folding propensities. When used in structure prediction first one aligns the target sequence to memory proteins with the Local Hamiltonian [125], a sequence-structure alignment tool. The sequence is then threaded onto the memory proteins. This determines the interactions for residues close in sequence therefore introducing a local secondary structure bias. In this study we use as the memory proteins for the IM proteins the respective crystal structures. This assures that the local structure including secondary structure

in the molecular dynamics simulations will be biased towards the local structure of the native states. The contact potential terms then predict the tertiary structure of the protein by flexibly assembling supersecondary structure elements.

To obtain the free energy landscape, constant temperature molecular dynamics runs were performed. Temperature is quoted in units of the native state energy of the AM and contact terms. The native state energy for a protein of N residues is scaled in units of $\epsilon = \frac{E^{native}_{AM,contact}}{4N}$, which leads to define a reduced temperature as $k_B T = \epsilon \bar{T}$, where $\bar{T}$ is the temperature of the simulation. All other energy terms such as the backbone terms are scaled to yield physically reasonable interaction strengths. In a typical simulation run both a randomly unfolded structure as well as the x-ray structure were used as starting structures. Initial random velocities were assigned to the protein. For each temperature 2×20 trajectories were obtained. The length of each of the trajectories was $7.5{\times}10^6$ steps of approximate time length of 12ns per step [127] resulting in 90μs long trajectories. In each of the runs 3000 independent structural samples were obtained for analysis. For each sample the energies were recorded and the relevant order parameters were calculated. The free energy was then calculated at different temperatures using $F(Q_W, Rmsd) = -\bar{T} \cdot ln(P(Q_W, Rmsd))$ where T is the simulating temperature and $P(Q_W, Rmsd)$ is the probability of finding a structure with given $Rmsd$ and Q_W. These order parameters provide two different measures of the similarity to the crystal structure and both involves a sum over all (but nearest neighbour) pairs of C^β or C^α atoms:

$$Rmsd = \sqrt{\frac{1}{\mathcal{N}} \sum_{i<j-1} (r_{ij} - r^{\mathrm{N}}_{ij})^2}, \tag{5.4}$$

where r^{N}_{ij} is the C^β-C^β distance between residues i and j in the native state,

$$Q_W = \frac{1}{\mathcal{N}} \sum_{i<j-1} \exp\left[-\frac{(r_{ij} - r^{\mathrm{N}}_{ij})^2}{\sigma^2_{ij}} \right], \tag{5.5}$$

where r^{N}_{ij} is the C^α-C^α distance between residues i and j in the native state,

$\sigma_{ij} = |i-j|^{0.15}$ and the normalization $\mathcal{N} = (N-1)(N-2)/2$ is the number of non nearest neighbour pairs given the lenght N of the chain.

Thus a *Rmsd* of $0\mathring{A}$ means the examined conformation is identical to the crystal structure whereas Q_W ranges between 0 (completely unfolded) to 1 (native conformation) [17].

We define a non native contact as a C^{β}-C^{β} pair whose distance in the crystal structure is greater than $9.5\mathring{A}$ and whose backbone distance is greater than 4 residues, but whose C^{β}-C^{β} distance in the obtained ensemble is less than $9.5\mathring{A}$.

In order to calculate the folding time constant, we collected the distribution of times needed to make the transition from the unfolded state ($Rmsd < 8.0\mathring{A}$) to the native state ($Rmsd > 3.5\mathring{A}$). The exponential fit of this distribution leads to the folding time constant τ_f.

The PDB code for Im7 is 1AYI.

5.3 Localized Frustration Measurements and Design Methods

Our site-specific measure of local frustration serves to rationalize the role of specific interactions in the im7 folding mechanism. It is also used to guide our mutational designs. The definition and procedures are described.

5.3.1 Local Frustration Index

The local frustration index is a site-specific measure of the energetic fitness for a given set of residues λ_i and λ_j at residue positions i and $j > i+1$. We used a simple definition of site-specific frustration based only on the sequence-specific components of the AMW energy function ($\mathcal{H}_{contact}$, $\mathcal{H}_{water}$, $\mathcal{H}_{burial}$) (5.3). These terms depend on the identities (λ), densities (ρ) and interaction

distances (r_{ij}) of the residues involved. The local frustration is defined as:

$$F_{ij} = (\mathcal{H}_{ij}^{N} - < \mathcal{H}_{i',j'}^{U} >) / \sqrt{1/N \sum_{k=1}^{n} (\mathcal{H}_{i',j'}^{U} - < \mathcal{H}_{i',j'}^{U} >)}, \qquad (5.6)$$

where $\mathcal{H}_{ij} = \mathcal{H}_{contact}^{i,j} + \mathcal{H}_{water}^{i,j} + \mathcal{H}_{burial}^{i} + \mathcal{H}_{burial}^{j}$ is the native site energy with native parameters (λ_i, λ_j, ρ_i, ρ_j, $r_i j$). We obtain the average and standard deviation of a set of reference energies, $\mathcal{H}_{i'j'}^{U} = \mathcal{H}_{contact}^{i',j'} + \mathcal{H}_{water}^{i',j'} + \mathcal{H}_{burial}^{i'} + \mathcal{H}_{burial}^{j'}$, by randomly selecting the parameters (λ_i', λ_j', ρ_i', ρ_j', $r_i j'$) according to the native composition of the corresponding parameters.

With this definition, for a given protein sequence composition and structure, the average and standard deviation of reference site energies are the same for all interacting residue pairs i, j. For the native im7, the average and standard deviation over the reference sites converge.

When $\mathcal{H}_{ij}^{N} =< \mathcal{H}_{i',j'}^{U} >$, the native site energy is not discriminated from a typical energy at a random site, and $F_{ij} = 0$; For the present study, frustrated sites are those where $F_{ij} < 0$. Highly frustrated sites have values of $F_{ij} < -1$. Arguments from theory suggest an interaction is minimally frustrated when $F_{ij} > 1.25$. We use these cutoffs in this study. An accompanying article in the present volume presents an extensive study of the concept and alternative definitions of local frustration.

To characterize the landscape generated by AMW simulations, we randomly selected 200 structures from the native basin and 500 structure from the intermediate basin. For each structure, we calculated F_{ij} for all pairs of residues whose C^{β} atoms (C^{α} for Glycine) are within $9.5\mathring{A}$.

5.3.2 Design Procedures

In the first step of both design based on minimally-frustrating the native state and specific negative design by de-stabilizing the intermediate state, residue pairs were selected based on the local frustration index. For design, all residues involved in highly-frustrated interactions (Fig. 5.2a, red lines) in

helix region III were selected for possible mutation (12 residues, 1 outside the helix III region). All pairwise combinations of these positions define our set of double-mutant positions (28 residue pairs). All possible combinations of residue types (400) were evaluated for each pair, resulting in a total of 11200 distinct double-mutants. For our specific negative design procedure, we identified 19 minimally-frustrated interactions, involving 22 distinct residue positions, from the AMW wild-type intermediate ensemble. Unlike the design procedure based on minimizing frustration, we did not consider all combinations of these 22 positions but instead focused our efforts on the 19 pairs with minimally-frustrated interactions. For these 19 pairs, we obtain a total of 7600 distinct double-mutants.

The second step involved selecting the best double mutants. In the design procedure, we evaluated the local frustration (F_{ij}) for all interacting pairs in the crystal structure. To check that this mutant does not destabilize the protein's total energy, we compute the total energy change of the contact ($\mathcal{H}_{contact}$, $\mathcal{H}_{water}$) and burial ($\mathcal{H}_{burial}$) terms upon mutation. To minimally-frustrate the crystal structure, we filtered out those double mutants with the fewest highly-frustrated sites ($F_{ij} < -1$) and the most favorable total energy change. A similar procedure was employed for negative design. The average local frustration and total energy change (over the intermediate ensemble) was calculated. Best candidates were those which not only minimized the number of non-native minimally-frustrated ($F_{ij} > 1.25$) sites over the intermediate ensemble, but that also did not frustrate the native state interactions.

5.4 Results and discussion

5.4.1 Perfect Funnel Model

We first examine whether the existence of an intermediate in the folding landscape of IM7 could be a direct consequence of a perfectly funneled landscape. For that purpose, we simulated IM7 with a perfectly funneled

Hamiltonian, which takes into account only those interactions, that are found in the folded protein. In the simplest funneled landscape, all native contacts are weighted equally. Intermediates and barriers in the free energy profile in this case must arise from a non-uniform tradeoff between entropic terms of the chain and the stabilization energy of the formed, native contacts. Simulations with this Hamiltonian show that IM7 would clearly fold as a two-state folder for this ideal model. The folded state and the unfolded state are separated by a barrier of about $5k_BT_F$ at the folding temperature. The experimentally detected intermediate does not appear using such a perfectly funneled Hamiltonian. There is no stable thermodynamic state between the folded and unfolded state, that is stabilized by energetically homogeneous, native contacts. No local energetic traps or signs of "topological frustration" were found, if the existence of topological frustration is proved by the observation of a bimodal distribution of the transition state ϕ-values, a clear sign of multiple, distinct folding pathways. While inclusion of nonadditivity through many-body contact terms increased the barrier height, such additional explicit cooperativity did not produce an intermediate state.

To study whether there may be an effect of energetic heterogeneity, the magnitude of all the native interactions were scaled with the Miyazawa-Jernigan energies [128] rather than being kept uniform. Such an inhomogeneous but still perfectly funneled model might account for the observation of a stable intermediate. Molecular dynamics simulation with this Hamiltonian both with and without explicit non-additivity were performed but did not reveal any intermediates. The two-state behavior is preserved without any relevant traps being present. Moreover the transition state ϕ-values still exhibited unimodal probability distributions.

The intermediate observed in Im7 folding apparently cannot be captured by a perfectly funneled landscape. The absence of any populated intermediate state even with heterogenic and many-body terms in these funneled Hamiltonians implies that the real folding landscape of Im7 must be more rugged.

5.4.2 AMW - Im7 folding mechanism

To predict the folding of IM7 with an energy function that yields a more rugged but still globally funneled energy landscape, we carried out simulations with the Associative Memory Hamiltonian with water mediated interactions (AMW). This non-additive Hamiltonian not only has heterogeneous direct contact energies and water-mediated energy terms to capture solvent-mediated interactions, but makes no use of native tertiary structure information and thus also allows for non-native interactions. To quantify the amount of local ruggedness of the energy landscape yielded by the AMW, we compute the frustration of interacting residue pairs with a method developed recently by Ferreiro et al.. The energy of a site, involving interacting residues at positions i and j, depends only on their amino-acid identities (λ_i, λ_j), densities (ρ_i, ρ_j), and pairwise distance r_{ij}. In order to measure how frustrated a site is (F_{ij}), we compute for each site, the native energy and reference site energies (see appendix II for details). The ratio of the local contribution to the energy gap versus the standard deviation of the loal decoy energies (the frustration index) then gives an estimate of how favorable the native interaction is relative to randomized interactions.

We first calculate the frustration index for all interacting residue pairs in the crystal structure of IM7 (pdb id 1ayi) to characterize the local energy landscape around the native basin. Like most proteins a cluster of minimally-frustrated contacts (17% of the total in this case; Fig 5.2a, green lines) spans the protein core. Many frustrated sites (36% of total) are also present. In particular, multiple distinct highly-frustrated contact clusters (11% of total; Fig 5.2a, red lines) are revealed in three distinct regions: the loop region between helix I and II, the helix III region and the C-terminal residues in helix IV.

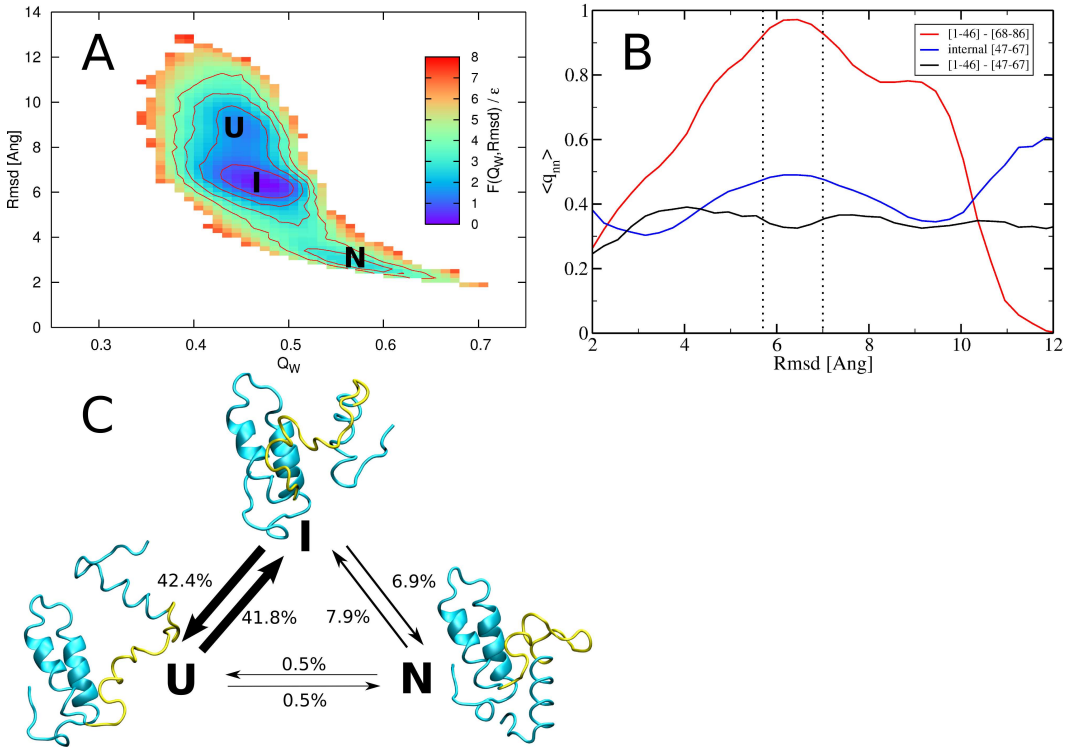


Figure 5.1: A) Free energy (in units of ϵ) of the Im7 at T=1.0 as a function of Q_W and Rmsd. The wells correspond to the unfolded (U), intermediate (I) and native (N) states. B) The average fraction of non native contacts as a function of the Rmsd. In the red curve we consider only the interactions between the first half of the protein (residues 1 to 46) with the fourth helix (residues 68 to 86). The blue curve shows the same ratio but only for the contacts internal to the helix III region (residues 47 to 67) and the black curve represents the contacts between the first half and the helix III region. Also displayed with vertical dotted lines are the boundaries used to define the intermediate state. C) The transitions observed between the three states U, I and N at T=1.0 suggest the intermediate to be on-pathway in the folding reaction. Three representative snapshots are also shown with the helix III region highlighted in yellow.

In this case the presence of frustration in the native state seems to stem from the fact that in Nature IM7 not only evolved to have a significantly funneled folding landscape, but also evolved to have a funneled binding landscape. We therefore computed the frustration for IM7 when it is bound to a binding partner and observed that most of the contacts that are found to be frustrated in the native state by itself become minimally frustrated in the bound form due to the modified burial properties of these residues.

We extracted the free energy profile from constant temperature molecular dynamics simulations with one memory term for the short-range in sequence interaction, the crystal structure of IM7. This memory term introduces a bias to form proper native secondary structure without biasing the long-range tertiary interactions. The free energy profile as a function of Q_W and $Rmsd$ of Im7 exhibits, at equilibrium, three thermodynamically stable states: a native (N), an intermediate (I) and an unfolded ensemble (U). We computed the radius of gyration R_g for the set of conformations in the respective ensembles. The three ensembles all show a similar degree of compactness relative to the X-ray crystal structure of IM7. The unfolded state is unusually compact with an average radius of gyration of $< R_g^U > /R_g^X = 1.2$. This can partly be understood by the bias of native, secondary structure in the Hamiltonian. Indeed, most of the secondary structure elements are already formed in the unfolded ensemble. Before reaching the native state, which has a crystal structure like compactness ($< R_g^N > /R_g^X = 1.0$), an intermediate ensemble is populated. At a simulation temperature of $\bar{T} = 1.0$ the intermediate is the most populated ensemble and is also on-pathway in the folding route, which is evident from the relative percentage of the transitions observed between the three states (Fig. 5.1C). The observed intermediate ensemble with $< R_g^I > /R_g^X = 1.1$ has an almost native-like compactness. In the laboratory the intermediate is also nearly as compact as the native structure as suggested by Friel et al. who measured a Tanford value of $\beta_T = 0.8$ for the intermediate of IM7 relative to the native state.

The structures of the intermediate ensemble generally fold to form a three

helix core. In this structure helix I (residues 12-24), helix II (residues 32-45) and their interface are all natively formed. Helix IV (residues 66-79), however, interacts in a non-native way with helix I. The presence of this highly structured core composed of helix I, II and IV in the intermediate was inferred experimentally by Capaldi et al. from ϕ-value analysis [121]. In our simulations, residues $77-80$ of helix IV and residues $12-24$ of helix I form numerous non-native contacts with high probability. These non-native contacts turn out to actually be minimally frustrated (see Fig. 5.3a, blue lines) - they stabilize the intermediate state through favorable contacts not found in the functioning native structure. Mutations of some of these residues (A77G, A78G and A13G, F15A, V16A, L18A, L19A, I22V) are also known from experiment to destabilize the intermediate state. From our simulations we understand that these contacts must be lost upon transition to the native state (see Fig. 5.1B, red curve) prompting helix IV to trade favorable non-native contacts for other native contacts that also are favorable.

Besides favorable interactions, the intermediate also exhibits many frustrated contacts (Fig. 5.3a, red and orange lines). While the three-helix bundle of helix I, II and IV is stabilized by favorable native and non-native interactions, the loop region (residues 47-67) surrounding helix III (residues 51-56), which we refer to as the helix III region, does not show any particular local structural preference. That is to say, native and non-native contacts internal to the region occur in almost equal proportion. Moreover, no particularly favorable interactions and many highly frustrated interactions are found in this region. The frustrated contacts and the rugged landscape of the helix III region must therefore play an important role in the formation of the IM7 folding intermediate. Upon crossing the free energy barrier separating the intermediate basin from the native one, the helix III region reduces its non-nativeness. In fact, the ratio of non-native contacts over the total number of contacts decreases by almost 40% (see Fig. 5.1B, blue curve). The helix III region also interacts non-natively with helix IV. Unlike the interactions within the helix III region, many of these are favorable.

5.4.3 Reducing Native State Frustration

Kinetic folding experiments on laboratory-designed IM7 mutants have shown that the native and intermediate populations can be shifted significantly with single and double-mutants. Based on our site-specific localization of frustration, we devised two strategies to perturb the folding landscape. In the first design scheme the goal is to find the double mutants, which defrustrate the native structure. Among the regions with significantly frustrated residues (Fig. 5.2B), we chose to focus our design strategy on the helix III region. In this frustrated region there is a large preponderance of competing non-native interactions and an absence of several native contacts in the intermediate ensemble making helix III region an optimal target for re-design. The other regions, although frustrated locally in the native structure, are substantially native-like in the intermediate (Fig. 5.2B) making them less likely successful design targets to shift the population of observed structures towards more native and less intermediate structures. We selected eight highly frustrated positions for designing a defrustrated native state and constructed a library of double mutants. For each pair of positions, the local frustration was computed for each of the 400 possible mutants (representing all combinations of amino acids). Roughly 30% (3362 of 11200) of the mutants had fewer frustrated sites than the native.

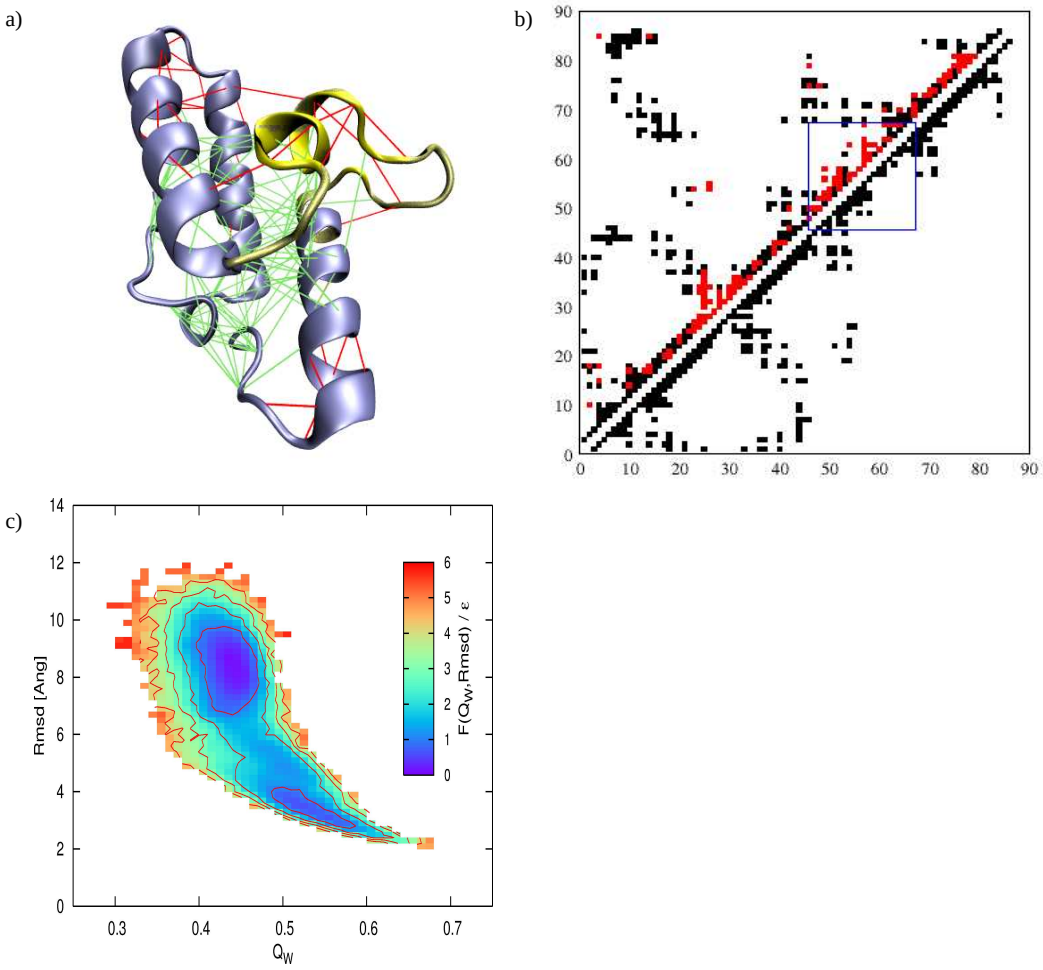


Figure 5.2: A) Local frustration is depicted on the native Im7 structure (from pdb 1ayi). A large cluster of minimally-frustrated contacts (green) defines the core of the protein yet some highly-frustrated contacts ($F_{ij} < -1$) surround the core (indicated in red). B) Highly frustrated contacts (lower-right) are primarily local compared to all contacts present in the native state (upper-left). C) Free energy (in units of ϵ) of the Im7 Designed Mutant at I=1.0 as a function of Q_W and Rmsd. The wells correspond well to the unfolded (U) and native (N) states of the wild-type protein.

Table 5.1: In total, 11300 mutants were evaluated for minimum frustration design (a) and 7600 for specific negative design (b). 26 favorable double-mutants are presented. * Selected for AMW simulation studies.

Mutated Positions	Position I	Position II	H_{seq}
52 55	K	D	172.354
	N	N	172.371
	K	E	172.523
26 55	K	K	172.712
	K	R	173.099
50 55	R	K	172.971
	R	R	173.493
55 56	K	R	173.894
*	R	N	174.001
	R	P	174.027
	R	C	174.074
	K	K	174.11
	R	H	174.17
	R	R	174.416
	S	K	174.425
	N	K	174.485
	R	K	174.632
49 55	I	K	174.761
	F	R	174.843
	C	K	175.119
	I	R	175.283
	V	K	175.456
	L	K	175.4
	C	R	175.641
	L	R	175.922
	V	R	175.978

To test whether the less frustrated mutant sequences yield a less rugged energy landscape than the wild type, we selected the top designs for further AMW simulation studies. The free energy profile obtained from constant temperature simulations with the least frustrated mutant sequence is shown in Fig. 5.2c. The free energy profile of the selected mutant sequence exhibited two-state folding behavior with no stable intermediate basin. This is a clear success for the design strategy of defrustrating the native state by mu-

tating amino acid positions which have frustrated interactions in the native state and a manifold of competing non-native interactions in the intermediate state. While the wild type sequence folded through an intermediate stabilized by non-native interactions, the designed sequence cooperatively folded from the unfolded structure to the low-energy native state without populating any intermediate traps. We computed the average folding time τ for the wild type sequence and the mutant sequences. As a consequence of being less rugged and more funneled, the less frustrated, designed sequence displayed a ten-fold speed-up in folding.The minor change of two residues in the sequence yielded an energy landscape that is similarly funneled to the energy landscapes ordinarily obtained with the "flavored" Gō-models having heterogeneous but only native contact energies.

5.4.4 Attempted Specific Negative Design of Intermediate

The goal of our second design study was to destabilize the intermediate ensemble and thereby eliminate the dominant thermodynamic trap. As we shall see, the specific negative design of the intermediate state is, however, a more difficult way to remove ruggedness from the energy landscape than is the reduction of frustration in the native target structure. As expected from energy landscape theory it is likely that even if one succeeds in destabilizing the specific ensemble of structures which constitutes the Im7 intermediate, if the native state is still frustrated one can not exclude the emergence of other thermodynamic, glassy traps. We evaluated the efficacy of a speific negative design approach by first identifying non-native contacts present in the wild-type intermediate ensemble which are minimally frustrated by our site-specific measure (blue lines in Fig. 5.3A; blue contacts in Fig. 5.3B). From this analysis, a precise candiate set of contacts emerged, namely the contacts that arise from the non-native association of helix IV with both helix I and helix III region. To disrupt these interactions, we need to find residues that frustrate the non-

native sites without significantly frustrating native contacts. We mutated to all 400 possible pairs and measured the energy change over a representative set of the intermediate ensemble (500 structures). A selection of the top mutants for negative design are presented (Table I).

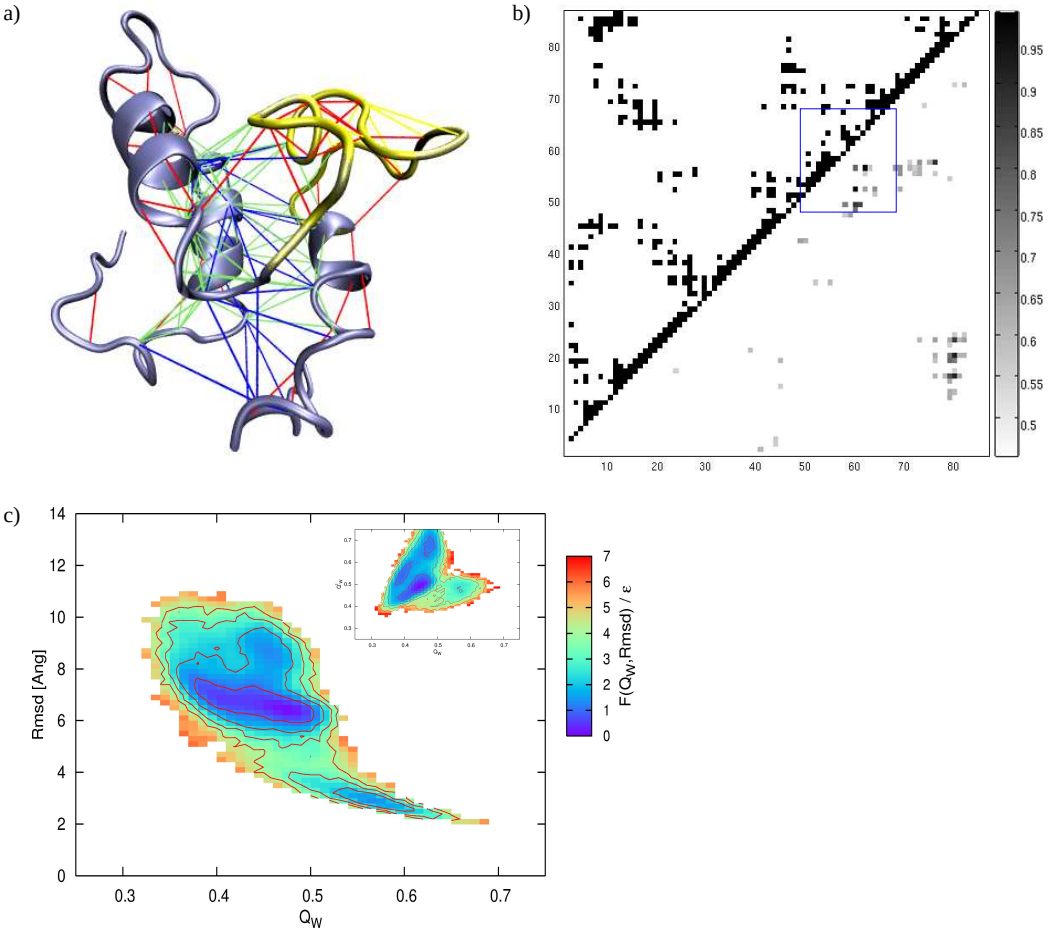


Figure 5.3: Specific Intermediate State Re-design A) Local frustration is depicted on a selected Im7 intermediate structure. Minimally-frustrated contacts present in the crystal structure (green) are distinguished from those which are non-native (blue). A distinct non-native cluster can be observed involving interactions between helix IV and the helix I-II region. Native (red) and non-native (yellow) frustrated contacts surround the core. B) Minimally-frustrated non-native contacts (lower-right) are distinguished from those present in the native state (upper-left). C) Free energy (in units of ϵ) of the Im7 re-designed mutant at I=1.0 as a function of Q_W and Rmsd. The inset shows the free energy as a function of Q_W to the crystal structure and Q_W to a representative structure from the wild-type intermediate basin.

Table 5.2: The folding time for each mutant relative to the wild type Im7 is given (see methods for further details) at $\bar{T}$=1.00.

Mutant	Mutations	τ_f'/τ_f^{wt}
I	S58C	2.5
II	D31H,D35E,E46V,S58Y,D63Y,E71R	0.7
III	K20Y,N79Y	2.4
IV	M1Y,Q17Y	1.3
V	Y55R,Y56N	0.4
VI	D49S,Y55R	0.1

Our results are consistent with ϕ-value data by Capaldi et al. [121]. In that study, mutations of residues in the ranges 3-19 and 72-78 were found to destabilize the intermediate state. Many of our top mutants involve residues in these ranges. One of the top mutants involves disrupting the interaction between the Lysine at position 20 and the Asparagine at position 79. We ran AMW simulations on this mutant. The free energy of this the double mutant (K20Y, N79Y) shown in Fig. 5.3 does display an intermediate state. This intermediate state turns out to be more widely dispersed in Q_W and *Rmsd* space than the intermediate state of the wild type still giving rise to a profound thermodynamic trap. The negative design scheme was successful in partially destabilizing the wild type intermediate, but the mutant sequence nevertheless gave rise to an alternative folding route to the native state via a different stable intermediate. Due to this alternate route, the folding time was lengthened. To quantify the destabilization of the intermediate we also plotted the free energy as a function of Q_W and Q_W^I, where Q_W^I is the structural similarity relative to a representative wild type intermediate structure (see the inset of Fig. 5.3C). There are several low free energy basins in the free energy plot. The biggest basin corresponds to an intermediate state that is not found in the wild type intermediate (low Q_W^I-value). Therefore the folding route through this alternate intermediate structure turns out to be preferred in this modified sequence.

5.5 Conclusions

From its inception the energy landscape theory acknowledged and highlighted the fact that many amino acid sequences would have frustrated interactions. Nevertheless in keeping with the principle of minimal frustration, finding a specifically structured long lasting non-native intermediate as observed in the IM7 folding has been rare. We have seen the intermediate of IM7 is not captured by perfect funnel models, even when cooperativity and contact heterogeneity are added. Yet from the landscape theoretic viewpoint this result is not surprising due to the significant clusters of highly-frustrated sites observed in the native state (Fig. 5.2). Frustrated interactions give rise to a rugged energy landscape wherein favorable non-native interactions compete. In the case of Im7, these significant traps from frustration give rise to an intermediate, which is observed both in experiment and our AMW simulations. The structural features of the intermediate from our simulations compare well with all experimental findings.

The observation of clusters of frustrated interactions in the native state points the way to a mutational strategy to reduce the ruggedness of the folding landscape. Our design strategy was a success and led to double-mutants that eliminated any significant intermediate population during AMW simulations. As expected specific negative design is not so easy or effective as minimizing frustration in the native target structure. While our mutant selected for simulations did perturb the intermediate ensemble, that ensemble was not eliminated. We note that the frustration in IM7 is relieved upon binding to its natural partner. In vivo folding and binding may occur together and should be well described by a largely unfrustrated landscape. IM7, by itself, only marginally satisfies the minimal frustration principle so the emergence of an intermediate with significant non-native structure (accompanying many native interacitons) beautifully resonates with the energy landscape theory.

6 Conformational Switching upon Phosphorylation: A predictive Framework based on Energy Landscape Principles

6.1 Introduction

Protein phosphorylation is one of the most important intracellular control mechanisms[59]. In both eukaryotic and prokaryotic cells, phosphorylation is a key step in cell cycle control, gene regulation, learning and memory[60]. Nowadays it is believed that about a third of the proteins in mammalian cells are phosphorylated at one time or another[61]. Communication in the cell by means of phosphorylation is rapid, reversible and does not require the slow production of new proteins or degradation of existing proteins. Ultimately the activities of proteins that are modified by phosphorylation must be traced to changes in the protein's conformation[62, 63, 64] that are induced by modifying the energy landscape. While native ensembles possess numerous conformational substates, the landscapes of most proteins are highly funnel-like. In many cases, phosphorylation modulates the stability of two near degenerate but structurally distinct conformational ensembles on the landscape allowing the same protein molecule to carry out different activities in the cell at different times. By modulating this near-degenerate landscape, phosphorylation can

act as a molecular switch, turning a specific conformation dependent activity on or off by tipping the balance of the population between the two ensembles.

Upon phosphorylation, a phosphate group becomes covalently attached to the side chain of a serine, threonine, tyrosine or histidine residue. Much like the more labile changes due to pH, the change of electric charge in a specific residue through phosphorylation can have several different structural consequences: it can induce local and/or global conformational change between discrete completely folded configurations, or induce order to disorder or disorder to order transitions [65]. Sometimes the effects of phosphorylation on the structure of the protein appear to be small but further recognition events essential to function, such as binding, can be profoundly affected.

To illustrate how energy landscape ideas can be used to think about phosphorylation and to devise predictive algorithms, we present a theoretical study of how phosphorylation modifies the global [66, 67, 68] rather than local [69, 70, 71] structure of two different proteins, the cysteine proteinase inhibitor cystatin and the receiver domain of the bacterial enhancer-binding protein NtrC (nitrogen regulatory protein C). These two different systems are small enough for detailed theoretical analysis but also have been structurally explored in the laboratory providing thereby the basis for a comparative study to elucidate the generality and specificity of phosphorylation effects.

Cystatins are inhibitors of cysteine proteinases, which destroy proteins by hydrolysis and hence are important in protein degradation (PDB code 1A67,1A90)[72]. Chicken cystatin has been structurally characterized in both an unphosphorylated and phosphorylated form. The phosphorylated residue, Ser80, is located in a flexible region of the protein, which is readily accessible both to protein kinases and to phosphatases. Serine phosphorylation sites in many proteins are often found to be flexible or disordered in structural studies. Phosphorylation in intrinsically disordered regions of the protein commonly results in the ordering of the structure in the vicinity of the phosphorylation site[73]. Unphosphorylated cystatin is a five-stranded β-pleated sheet which is twisted and wrapped partially around a five-turn helix. When cystatin

becomes phosphorylated, moderate structural changes occur. The overlay of the mean NMR structures of phosphorylated and unphosphorylated cystatin show an RMS deviation between the structures of 2.7Å. Cystatin thus serves as a paradigm for a system having minimal structural change induced through phosphorylation in a flexible loop region.

A more dramatic change upon phosphorylation in terms of structure occurs in another well characterized system, the receiver domain of NtrC. The receiver domain of NtrC is a conformational switch found in a bacterial "two-component" regulatory system (PDB code 1DC7,1DC8)[74]. Upon phosphorylation two β-strands as well as two α-helices are displaced away from the phosphorylation site and additionally one helix is rotated axially. The overlay of the average NMR structures of the unphosphorylated and phosphorylated conformation of NtrC shows larger RMS deviation between the structures of about 3.3Å. The amplitude of the change is thus slightly larger than for cystatin. NtrC has been regarded as a model for a conformational switch[75], in which a "large" conformational change is induced upon phosphorylation. Clearly larger proteins can exhibit still larger changes in an RMS sense, owing to a greater lever arm for hinge motion in them.

The aim of the current study is to elucidate how phosphorylation causes these observed changes in protein conformations. First we examine the free energy profiles that would be obtained by assuming an ideal landscape having as little frustration as possible. This landscape for the phosphoprotein is constructed by utilizing the information about the structures of both phosphorylated and unphosphorylated native forms. Such a model yields the free energy difference of the forms that would be expected if only the native contacts were to contribute to the energetics. Since the conformations and hence the contact maps of the unphosphorylated and the phosphorylated proteins in our study are already known from experiments, we can construct such a structure based Hamiltonian having native-only interactions for molecular dynamics simulations to obtain conformations and energies of the proteins along the reaction coordinate. This is a "vanilla" Hamiltonian because it is topology

based, not singling out any interactions as especially significant. This model treats the two different sets of input native contacts, those for the unphosphorylated conformation and those for the phosphorylated one, as independent. We can more directly extract changes in the free energy profile using the free energy perturbation method. Next, a principal component analysis of the contact maps of the simulated ensembles allows us to find the dominant components of the phosphorylation induced change and to visualize the effect that phosphorylation has on a residue-residue contact map. The contact maps of the test proteins in the unphosphorylated and phosphorylated forms show that many of the contacts formed by the phospho-residues for the test proteins are preserved, suggesting the effect of phosphorylation primarily lies in the long-range forces. This observation allows us to address a rather practical issue: Instead of needing structural information on both forms, can one predict the likely conformational changes that should occur when only one structural form is known? For example, given structural information only about the unphosphorylated protein and the sequence information of which particular residues are susceptible to phosphorylation, can one predict the dominant conformation of the phosphorylated protein? For such predictions, obviously perfect funnel, native topology based models will not suffice. Since long range interactions are expected to be dominant however, we can construct a new guided structure prediction Hamiltonian by using local structural information known from the unphosphorylated protein for residue interactions separated by a few residues (12 in this case), but use a transferable structure prediction Hamiltonian (AMH)[76, 2] having a heterogeneous through space potential for residues that are more than 12 residues apart in sequence. The transferable long range potential while transferable has been shown to yield a reasonably funneled potential which has been optimized based on a large set of generic protein structures to successfully predict the folded state of proteins of size up to 180 residues. Its predictive power has been well documented[77]. Additionally it is possible to construct a new potential in this format to evaluate the interactions of the phosphorylated residues based on the same form.

To obtain the Hamiltonian for phosphorylated proteins from that which has been optimized for normal, unphosphorylated amino acids, we earlier postulated that we can treat the interactions involving the phosphorylated residue as those of a "supercharged" glutamic acid residue[70]. The energetic interactions of the phosphorylated residue with other residues are replaced with enhanced interactions of the type ordinarily used for a glutamic acid residue with the corresponding residues[70]. Since the energy landscape of the unphosphorylated protein is known and the contact maps of the test proteins indicate there is a considerable overlap of contacts between the unphosphorylated and phosphorylated conformations, we preserve the native focussed associative memory terms biased towards the assumed known unphosphorylated structure for residues that are less than 12 residues apart in sequence space but use the transferable potential with a "supercharged glutamate" for the more distant interactions. The Hamiltonian we have constructed in this way equips us with an energy function that should reliably mimic the local structure of the unphosphorylated protein, but that nevertheless plausibly treats the effects of the long-range forces on the conformation of the protein. We show this hamiltonian correctly predicts many features of the conformational changes observed in the phosphorylated protein. To document that this can be done, we set up simulations with different strengths of the charge interactions for the phosphorylated protein, and then we project the conformations obtained in these simulations onto the first two principal components obtained earlier using the "vanilla" native structure based Hamiltonian. We also show more directly that structures rather close to the NMR structures can also be sampled.

6.2 Methods

In order to explore the issues raised above, we studied four Hamiltonians based on the native configurations of the test proteins, $\mathcal{H}_u$, $\mathcal{H}_p$, $\mathcal{H}^*{}_u$, and $\mathcal{H}^*{}_p$. We show how to construct the native-based Hamiltonians $\mathcal{H}_u$, $\mathcal{H}_p$ in

the first subsection. These two Hamiltonians are based on the information of the experimentally determined native structures of the unphosphorylated or phosphorylated form of the proteins. Note that throughout the current study, we use the subscripts p and u to indicate the phosphorylated or the unphosphorylated form respectively. We then describe how to obtain the free energy profiles from the conformations sampled with $\mathcal{H}_u$ and $\mathcal{H}_p$ and describe a principal component analysis based on the contact maps of these conformations.

Finally we describe the construction of structure prediction Hamiltonians ${\mathcal{H}^*}_u$ and ${\mathcal{H}^*}_p$, *both* of which are based on transferable interactions using the long range interaction parameters optimized for generic structure prediction but that use information about the native conformation of the unphosphorylated form to encode the short and intermediate range interactions. Note that neither ${\mathcal{H}^*}_u$ or ${\mathcal{H}^*}_p$ contains any experimental information of long-range interactions found in the unphosphorylated form; neither ${\mathcal{H}^*}_u$ or ${\mathcal{H}^*}_p$ directly make use of any (short, intermediate, or long range) experimental information on the *phosphorylated* form at all.

We also detail how we define various physical quantities for monitoring structural ensembles, such as order parameters and configurational free energy, which we adopt to analyze the results of all simulations based on these four Hamiltonians.

6.2.1 Native Structure Based Simulations

Simulations of the folding dynamics of cystatin and NtrC were performed with an off-lattice native structure based potential. The Hamiltonian used in this study contains a basic backbone Hamiltonian and a contact potential

$$\mathcal{H}_{u/p} = \mathcal{H}_{bb} + \mathcal{H}_{c,u/p} \tag{6.1}$$

and depends on the locations of the C^α, C^β and oxygen atoms. The index u/p is a simplified notation for the two cases, namely u or p. The remaining backbone atom positions can be calculated assuming ideal backbone geometry.

The backbone potential H_{bb} constrains the backbone to have chemically and physically acceptable conformations[17]. The backbone potential is given by

$$\mathcal{H}_{bb} = \lambda_{\psi\phi}\mathcal{H}_{\psi\phi} + \lambda_{\chi}\mathcal{H}_{\chi} + \lambda_{ex}\mathcal{H}_{ex} + \lambda_{harmonic}\mathcal{H}_{harmonic} \tag{6.2}$$

The Ramachandran potential $\mathcal{H}_{\psi\phi}$ provides a good fit of the backbone torsional angles based on the statistics of protein structural database. The chirality potential $\mathcal{H}_{\chi}$ biases the protein chain into the L-amino acid configuration. The algorithm SHAKE constraints for the heavy backbone atoms along with three quadratic potentials provide for backbone rigidity and planarity. To complete the picture of stereo chemically allowed protein backbones, an excluded volume potential is applied to the oxygen and carbon atoms of residue i and j. This potential applies when the heavy atoms approach within 3.5Å for residues close in sequence space such that $(j-i) < 5$, and 4.5Å for $(j-i) \geq 5$. The λ-terms scale the interactions of the individual backbone potentials.

The contact term $\mathcal{H}_c = \mathcal{H}_{c,S} + \mathcal{H}_{c,M} + \mathcal{H}_{c,L}$ is an associative memory term[78]. Through its guidance, the free energy will reach a minimum at the basin of the given native PDB structure. Since there are several structures of cystatin deposited in the PDB, all these structures were used as memory terms for the simulation. The functional form of the contact term is given by

$$\mathcal{H}_{c,u/p} = -\epsilon \sum_{i \leq j-3} \gamma[x(|i-j|)] \exp\left[-\frac{(\mathbf{r}_{ij} - \mathbf{r}_{ij}^{Nat,u/p})^2}{2\sigma_{ij}^2}\right] \tag{6.3}$$

The sum runs over all carbon atom pairs ($C^{\alpha} - C^{\alpha}$, $C^{\alpha} - C^{\beta}$, $C^{\beta} - C^{\alpha}$, $C^{\beta} - C^{\beta}$) having a sequence separation of at least three residues. The functional form of the interactions of the carbon atoms in this potential are Gaussian centered at the native distance r_{ij}^{Nat} and with a width of $\sigma_{ij} = |i-j|^{0.15}$Å. The $\mathcal{H}_c$ potential depends on the sequence separation $|i-j|$ of the residues i and j. We divide the energy into three different proximity classes $x(|i-j|)$: short range (S) for $|i-j| < 5$, medium range (M) for $5 \leq |i-j| \leq 12$ and long range (L) for $|i-j| > 12$. The $\gamma[x(|i-j|)]$-terms are weighted such that the energies in each proximity class $x(|i-j|)$ are equal to each other. Also

the energies of any contact in each proximity class are equal for all contacts formed. The total energy of the Hamiltonian is scaled to be $4N$, where N is the number of residues of the protein. The unit of energy can then be denoted as ϵ and is defined in terms of its native state energy coming from the contact term $\mathcal{H}_c$ only

$$\epsilon = \frac{\langle H_c \rangle}{4N} \tag{6.4}$$

The simulation protocol is as follows: For each protein twenty constant temperature runs were performed with the structure based Hamiltonian. The constant temperature runs sampled 800 independent structures each spaced at intervals at about 1μs corresponding to a trajectory of about 1 ms in physical time. A total of $16000 \times 2 \times 2 = 64000$ structures were obtained for various temperatures for the unphosphorylated protein as well as for the phosphorylated protein. The key thermodynamic quantity desired from the simulations is the free energy as a function of reaction coordinate Q and temperature. The normalized collective coordinate Q measures the similarity of two conformations A and B to each other.

$$Q = \frac{2}{(N-1)(N-2)} \sum_{i<j-1} \exp\left[-\frac{(r_{ij}^A - r_{ij}^B)^2}{2\sigma_{ij}^2}\right] \tag{6.5}$$

6.2.2 Free Energy Perturbation Method

We directly examine how phosphorylation changes the free energy profiles. We start by analyzing the sampling snapshots obtained in simulation with each of the two Hamiltonians. After projecting the ensembles to the desired collective coordinates $r = \{r_1, r_2, \ldots\}$, the probability distribution $\rho(r) = N(r)/N_{tot}$ is computed for a total of N_{tot} snapshots. We can then derive straight forwardly the free energy profile $F(r) = -k_B T \ln(\rho(r)/\rho_0)$, where ρ_0 is a uniform distribution, for the unphosphorylated and phosphorylated conformations. We are interested in the difference of the two free energies, so we subtract these to obtain the difference $\Delta F(r) = F_p(r) - F_u(r)$.

In the current case we use two different folding order parameters Q_u and

Q_p as the collective coordinates, i.e., $r = (Q_u, Q_p)$. We assign to each snapshot two numbers, the order parameters Q_u and Q_p, which measure how similar an individual snapshot obtained in the MD simulations is to the native structure of the unphosphorylated or the phosphorylated conformations respectively. Simulations with the native structure based Hamiltonians bias the sampled conformations strongly towards the native structure. Performing a simulation with one of the two Hamiltonians, say $\mathcal{H}_u$, results in greater sampling of structures with high Q_u but sparse sampling of structures with high Q_p.

Instead of using a brute force approach of performing a large amount of simulations to assure acceptable sampling of the 2D reaction coordinate space we use the free energy perturbation method[15] to obtain the free energy difference directly. Thus to calculate the free energy difference ΔF from the sampling of the unphosphorylated Hamiltonian $\mathcal{H}_u$, we not only project the sampled conformations to the collective Q coordinates, but also record for each conformation, what the energy $E_u = \langle \mathcal{H}_u \rangle$ of the unphosphorylated system is and also what the energy $E_p = \langle \mathcal{H}_p \rangle$ of a phosphorylated system with the *same* conformation would be. We then perform the statistics on the raw moments of the energy difference $\left\langle \Delta E^k \right\rangle (r) = \left\langle (E_p - E_u)^k \right\rangle (r)$. The free energy difference of the two systems is then simply given by the cumulant expansion equation, i.e.,

$$\Delta F(Q_u, Q_p) = -\sum_{j=1} [(-\beta)^j / j!] C_j(Q_u, Q_p) \tag{6.6}$$

Here C_j is the jth order of the expansion. We have $C_1 = \langle \Delta E \rangle$, $C_2 = \langle \Delta E^2 \rangle - \langle \Delta E \rangle^2$, etc.

6.2.3 Contact Map Principal Component Analysis

We also use a principal component analysis (PCA) based on contact maps to visualize the conformational changes induced by phosphorylation. The more commonly used principal component analysis based on the diagonal-

ization of the Cartesian coordinates is less useful for our purposes because the change in the energy is only weakly related to the changes in the linear Cartesian distances. This mismatch is due to the fact that in phosphorylation the large conformational changes are generally of a magnitude beyond the simple vibrational-like fluctuations of the Cartesian coordinates. To capture properly the conformational changes, it is necessary to employ a set of detailed, site specific, and structure based reaction coordinates that do correlate with the energy. The global order parameters Q_u or Q_p do not suffice for the detailed description. We select a set of coarse-grained yet local-information-revealing degrees of freedom encoded in the contact map. This is the simplest site specific measure properly capturing the structure of a conformation while relating directly to the energy. A contact between residue i and j is considered to be formed (given the value of 1 as opposed to 0 when no contact is formed) when the distance of the respective C^β atoms is less than 6.5Å. For each snapshot obtained in the molecular dynamics we compute the contact map. The contact principal component analysis[79] reflects the correlations between different contact forming events. The covariance matrix to be diagonalized is not based on the linear cartesian coordinates but rather on a contact map cprrelation function

$$C_{i,j,k,l} = \langle (m_{ij} - \langle m_{ij} \rangle)(m_{kl} - \langle m_{kl} \rangle) \rangle \tag{6.7}$$

This "hypermatrix" encodes how an instance in which residue i and j form a contact correlates to an instance where residues k and l form a contact. To further facilitate the analysis, we coarse-grained the contacts by grouping neighboring residues into groups of four residues, i.e a coarse-grained contract matrix is calculated for each snapshot, with each of the independent elements being either 0 or 1. The coarse grained contacts are reduced in number to $27 \times (27-1)/2 = 378$ and $31 \times (31-1)/2 = 365$ for cystatin and NtrC respectively. The resulting reduced covariance matrices of dimension 378×378 and 465×465 are diagonalized and the eigenvalues are calculated. The two most dominant principal components (PC) are plotted.

6.2.4 Linear response theory (LRT)

As an alternative to the detailed sampling of the predictive Hamiltonian in the next subsection, we can use the linear response theory to see how phosphorylation should induce conformational changes. Linear response theory suggests that the magnitude of the conformational changes is a convolution of the strength of the sequence specific perturbation times the susceptibility of the corresponding degrees of freedom to make such changes [80, 81]. Statistical thermodynamics shows the coefficient of the response of a system under small external change is also linearly related to the fluctuations of the system sampled at equilibrium. The most commonly known manifestation of this relation explains how the heat capacity, a measurement of how energy change with the temperature change of a system is related to energy fluctuations.

In our case, the linear response theory describes the changes of the contact map using a relation of the form

$$\langle \delta q_{i,j} \rangle = \sum_{k,l} C_{i,j,k,l} \langle \delta V_{k,l} \rangle \tag{6.8}$$

where $\delta V_{k,l}$ is the matrix of contact energy change upon phosphorylation. The details of δV will be spelled out in detail in the next subsection. Nevertheless it is easy to see that δV is a very local property in the contact representation. For example, say residue 7 is the only residue that undergoes phosphorylation, we will then only have nonzero contributions of δV for the elements $\delta V_{k,l}$ if $k = 7$ or $l = 7$, otherwise $\delta V_{k,l} = 0$. By bridge in with the hypermatrix $C_{i,j,k,l}$, we can see how the changes of contact energy between the pair i-j are correlated with the changes of contact probability between the pair k-l at equilibrium. Linear response analysis yields the change in probability of forming a certain pair i-j when all the input contact energies change. Since δV is very local, i.e., is an extremely sparse matrix, it follows that the structural responses are primarily a combination of the largest eigen vectors of the diagonalization of the hypermatrix C (the top PCs). The dominance of these modes reflects the fact that those eigenvectors have largest amplitude of fluctuation. The linear response theory is an efficient method to give a quick estimate of the changes

caused by a perturbation. It is more accurate for systems that undergo small changes than for systems that undergo complicated, more involved changes.

6.2.5 Modeling tertiary structure effects of phosphorylation

Can one predict the conformation of the phosphorylated protein given knowledge of the folding landscape of the unphosphorylated protein only and the changes in the modifiable residues? As a first step to answer the prediction question, we developed a set of Hamiltonians $\mathcal{H}^*$ based on the information of the unphosphorylated form alone. We use the superscript $*$ to denote the energy functions that are transferable to distinguish the two sets. We first compare the difference of the ensembles generated by $\mathcal{H}_p$ and $\mathcal{H}_u$ and the difference of the ensembles generated by ${\mathcal{H}^*}_p$ and ${\mathcal{H}^*}_u$. We thus constructed a specific Hamiltonian constructed in the following form

$$ {\mathcal{H}^*}_{u/p} = {\mathcal{H}^*}_{c,L,u/p} + \mathcal{H}_{c,S+M,u} + \mathcal{H}_{bb} \tag{6.9} $$

The only difference between ${\mathcal{H}^*}_{u/p}$ and $\mathcal{H}_u$ lies in the long range energy terms. All three Hamiltonians share the same backbone and the same short and intermediate contact terms with each other. Here $\mathcal{H}_{c,S+M,u}$ is given by Equation 6.3 and summed only over residues that are separated by twelve or less residues in sequence space. This term biases the local secondary structure of the protein by having only the native interactions of one of the forms and hence yields largely native secondary structure. The tertiary structure of the protein follows thus from the contact energy term. This contact energy term arises from an optimized energy function used previously for protein structure prediction. The details may be found in [2] and references therein. A 4-letter code is utilized and the specific amino acids in each category denoted as hydrophilic (Ala, Gly, Pro, Ser, Thr), hydrophobic (Cys, Ile, Leu, Met, Phe, Trp, Tyr, Val), acidic (Asn, Asp, Gln, Glu) and basic (Arg, His, Lys). The energy contributions of the contact potential to the total potential are given

by a three-well potential.

$$\mathcal{H}^*{}_{c,L,u/p} = -\epsilon^* \sum_{i<j-12} \sum_{k=1}^{3} \gamma^*(P_i, P_j, k) c_k(N) U[r_{min}(k), r_{max}(k), r_{ij}] \quad (6.10)$$

Here k is a function of the spatial distance r_{ij} of residues i and j and c_k is found from fitting the number of contacts of the protein in each of the regions of k as a function of sequence length of the target protein. The interactions are weighted by the interacting amino acids of class P_i and P_j and their spatial distance. The parameters γ^* have been optimized based on the principle of minimal frustration. It is critical to note that γ^* is a function of residue chemistry, thus γ_u^* and γ_p^* have different values. More specifically, γ_u^* was derived from a structural database of ordinary, unphosphorylated proteins following the training procedure for the parameters based on the quantitative form of the minimal frustration principle[17]. The training maximizes the energy gap over the variance. This quantity is a measure of how funneled the landscape is towards a properly folded structure as compared to a random ensemble of molten globule structures. The procedure for deriving the parameters has been described in greater detail by Hardin et al. The contact function U controls the shape and sharpness of the multi-well potential [17]. It is important to stress that this term is heterogeneous but generic and transferable. As for γ_p^* , we have modeled the influence of the phosphorylation of an amino acid by substituting for the phosphorylated residue a supercharged glutamic acid residue. This strategy was put forward in previous studies of phosphorylation of NFAT where the structure was entirely unknown[70]. An analogous experimental approach based on the analogy between phosphoserine and glutamate has also been demonstrated to work in several cases, notably in studies on the dematin headpiece[82] and tumor supressor protein p53 [83]. These studies show that the Ser-to-Glu mutant closely mimics the conformation of the phosphorylated protein. The details of the implementation of the hyper-charged residue and its interactions with other residues as well as robustness and caveats have already been described by Shen et al.[70].

As $\mathcal{H}^*{}_u$, $\mathcal{H}^*{}_p$ and $\mathcal{H}_u$ share the same values for all other energy terms, it

would seem to be extremely demanding to try to predict the exact changes based on this generic long range term alone. Still we will present quite a successful demonstration of the importance of the generic long range potential in predicting the phosphorylated conformation. The trends of conformational changes generated by ${\mathcal{H}^*}_p$ observed in the simulations are consistent with the trends generated by ${\mathcal{H}^*}_u$ and thus by experiments. Constant temperature MD simulations with the Hamiltonian ${\mathcal{H}^*}_{c,L}$ were performed to predict the structure of the phosphorylated protein. In these simulations the starting structure was fixed to be the average NMR structure of the unphosphorylated protein. Following this a total of $16000 \times 2 = 32000$ independent structures were sampled.

6.3 Results for the Native Structure Based Hamiltonians

6.3.1 Free energy landscape of phosphorylated proteins

To sensibly study global effects of phosphorylation using coarse-grained models, the contact maps of the unphosphorylated and phosphorylated form of the test proteins must be different, that is sufficiently large to be reflected in the contact maps of the test proteins. The contact maps of the unphosphorylated and phosphorylated conformations of cystatin and NtrC are shown in Fig. 6.1. The important conformational changes induced by phosphorylation of cystatin do indeed present themselves in the contact map. Phosphorylation however introduces rather minor perturbations to the cystatin system. The contact map of NtrC shows more substantial changes upon phosphorylation. The contacts of the phospho-residue in both the unphosphorylated and phosphorylated conformations are identical, but phosphorylation apparently introduced long-range effects that lead to the global conformational change of NtrC.

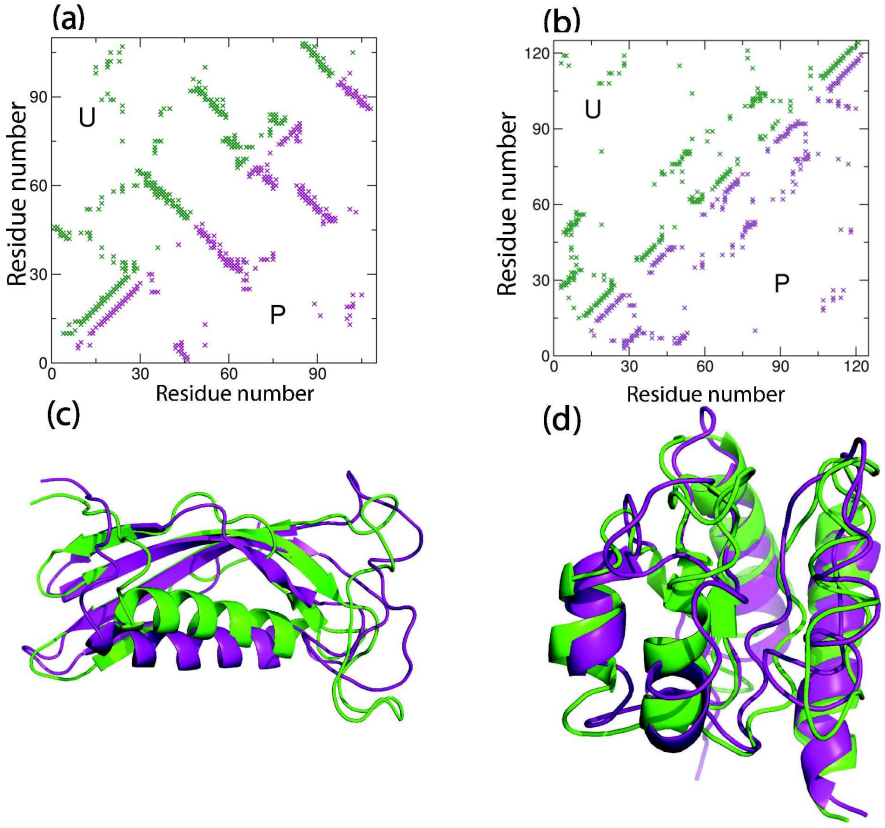


Figure 6.1: Contact maps of (a) cystatin and (b) NtrC and the corresponding structures shown in (c) and (d). The average contact map for the unphosphorylated conformations are shown in the upper-triangle, while the contacts of the phosphorylated protein forms are projected on the lower-triangle.

Molecular dynamics simulations with the native structure based Hamiltonians were performed to obtain adequate sampling of the conformations of cystatin and NtrC in their unphosphorylated and phosphorylated conformations. First, snapshots of MD simulations were sampled with the unphosphorylated native structure based Hamiltonian, $\mathcal{H}_u$. For each snapshot, the energy E_u as well as the order parameter Q_u, which measures similarity to the average structure of the unphosphorylated conformation, were calculated. The probability distribution ρ was computed. This allows calculation of the free energy, $F(r) = -k_B T \ln(\rho(r)/\rho_0)$. For the same snapshots obtained with $\mathcal{H}_u$, the energy E_p, which can be obtained from the Hamiltonian of the phosphorylated conformation, $\mathcal{H}_p$, and the order parameter Q_p were computed. The 2D free energy profiles of unphosphorylated cystatin and NtrC are plotted in Fig. 6.2, 6.3. The set of (E_u, Q_u) and (E_p, Q_p) found for snapshots at various Q_u and Q_p was used to obtain the free energy difference $\Delta F(r) = F_p(r) - F_u(r)$ via the cumulant expansion equation.

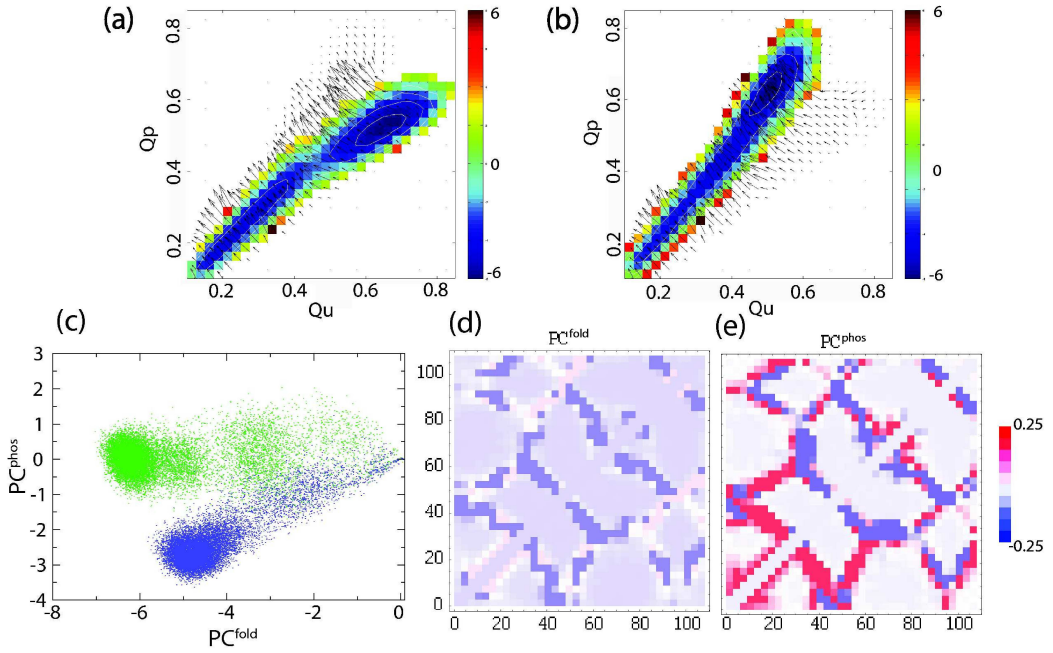


Figure 6.2: Free energy landscapes of cystatin folding for the unphosphorylated form (a) and the phosphorylated form (b). The white contour lines are drawn to facilitate observation of the native and unfolded basins in the free energy landscape. Arrows indicate the gradient of the free energy landscape pointing in the direction of phosphorylation and scaled in size to representable values. Snapshots of the conformations of unphosphorylated cystatin (green) and the phosphorylated cystatin (purple) projected along the first two dominant principal components in (c). The largest two principal components shown in the contact map form (d,e)

The gradient of $\Delta F(r)$ is also plotted in Fig. 6.2, 6.3 and is indicated by the arrows on the free energy landscape at each position along the folding order parameter. The length of the arrows indicate the relative magnitude and direction of the change of $\Delta F(r)$. Also the same procedure is applied to conformations sampled in molecular dynamics runs with the $\mathcal{H}_p$ Hamiltonian as energy function. The results are plotted in Fig. 6.2,6.3.

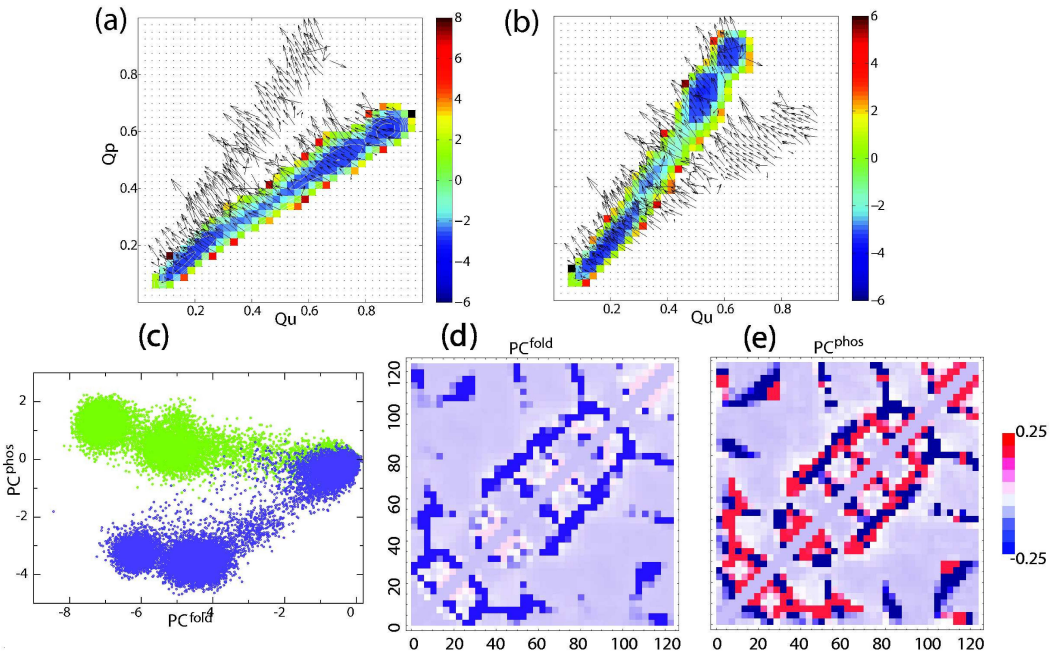


Figure 6.3: Free energy landscapes of NtrC folding for the unphosphorylated form (a) and the phosphorylated form (b). Arrows and contour lines are drawn for better visualization. Snapshots of the unphosphorylated NtrC (green) and the phosphorylated NtrC (purple) projected along the first two dominant principal components in (c). The largest two principal components are shown in the contact map form (d,e).

The free energy profile for cystatin at a simulation temperature close to the folding temperature of $T = 1.0$ shows a simple two-state folding process with an unfolded and a folded basin (Fig. 6.2) separated by a barrier of about $4k_BT$. The coordinates in Q_u, Q_p of the two free energy minima for the unphosphorylated protein are given by $(0.29, 0.25)$ for the unfolded basin and $(0.64, 0.52)$ for the folded basin. The free energy minimum for the folded state of the phosphorylated cystatin is located at $(0.49, 0.62)$. The gradient of the free energy difference $\Delta F(r)$ is also shown as a vector that gives a good indication at each value of the reaction coordinate, how phosphorylation effects the profile. In the phase space region of $Q_u \leq 0.5$ the arrows point directly into the direction of the phosphorylated protein. This is due to the fact that before reaching the transition state, the two forms of the unphosphorylated and phosphorylated protein can easily interchange. Even after crossing the transition state, the direction of the gradient of both forms is almost the same as before with the difference that most arrows do point slightly in the direction of lower Q_u, the unfolding direction. Figure 6.2 shows the free energy plot for sampling of phosphorylated conformations with $\mathcal{H}_p$. The resultant 2D free energy landscape was similar to the landscape obtained with $\mathcal{H}_u$ and using the cumulant expansion method to determine $\Delta F(r)$. Principal component analysis was performed and the conformations were projected onto the first two dominant principal components as shown in Fig. 6.2. For every projected snapshot it is known, how folded the structure is and also if the snapshot stems from a simulation of the unphosphorylated or phosphorylated protein. The principal components therefore correspond to folding and phosphorylation and we can name them the folding principal component PC^{fold} and the phosphorylation principal component PC^{phos}. PC^{fold} measures the general folding order with more negative PC^{fold} indicating a more folded set of structures. PC^{phos} measures how much a conformation is similar to the phosphorylated conformation, that is, the negative direction corresponds to the direction of conformational changes that occur upon phosphorylation. Projection of the changes of PC^{phos} onto a contact map allows inspection of phosphorylation induced contact changes. The PC^{phos} contact map shows the dominating con-

tact changes upon phosphorylation in blue, while contacts dominating in the unphosphorylated form show up in red. Direct comparison of the structural changes of the simulated ensembles (Fig. 6.2d) to the changes observed in the contact map obtained from the pdb native structures of the unphosphorylated and phosphorylated form (Fig. 6.1a) show excellent agreement, i.e. contacts that are exclusively formed in the unphosphorylated form show up as red while contacts that are solely formed upon phosphorylation show up in blue.

Three free energy minima are found in the free energy plot of unphosphorylated NtrC at temperature $T = 1.0$ (Fig. 6.3). This suggests that the unphosphorylated NtrC is not a two-state folder but has a well-ordered intermediate at coordinates in Q_u, Q_p given by $(0.7, 0.5)$. The native basin is located at $(0.87, 0.6)$ and the unfolded basin is at $(0.17, 0.16)$. The gradient of the free energy difference $\Delta F(r)$ is again plotted using arrows, that indicate the direction and magnitude of the change in $\Delta F(r)$ upon phosphorylation. The arrows show a largest gradient in the intermediate state, which would suggest that transitions from the unphosphorylated conformation to the phosphorylated conformations of NtrC are prefered in the intermediate states of folding. The free energy profile obtained from $\mathcal{H}_p$ for the phosphorylated NtrC is also shown (Fig. 6.3). The folding is also 3-state with three main free energy minima. Principal component analysis was performed on the snapshots obtained in the molecular dynamics simulations with Hamiltonians $\mathcal{H}_u$ and $\mathcal{H}_p$ (Fig. 6.3). It is apparent from the figure, that the 3-state folding behavior is well captured by the principal component analysis. The first two components are by themselves very useful in capturing the folding and the effects of phosphorylation respectively. We identify the principal component PC^{fold}, which provides a good indication of the degree of the folding order, where a more negative PC^{fold} indicates a more folded set of conformations. PC^{phos} serves to distinguish the unphosphorylated ensemble from the phosphorylated ensemble. Projection of the first two principal components of snapshots is shown in Fig. 6.3. The agreement with experiment is great, again. Fig. 6.3 proves useful in identifying the trends of contact changes upon phosphorylation. We

note that for a 3-state folder, the third principal component might also be important. Plots of combinations of any two of the first three components show 3-state behavior, however first two principal components do distinguish the global folding and phosphorylation best.

6.3.2 Changes in free energy profiles between unphosphorylated and phosphorylated protein conformations

In vivo, proteins that become phosphorylated can have two sensibly different average conformations as revealed by X-ray crystallography or NMR despite the two forms having obviously almost identical sequences (except for the phospho-residues, the two sequences are identical). Normally sequences with high sequence similarity adopt the same fold[84]. Thus it may seem obvious to assume that in fact the unphosphorylated protein itself can assume both conformations, the unphosphorylated conformation and the phosphorylated conformation. However, for phosphorylation to crisply act as a molecular switch, the two conformations should be separated by a high barrier such that the unphosphorylated protein will not likely spontaneously adopt the incorrecy structure and hence function of the phosphorylated protein. It is natural then to ask how difficult is it for the unphosphorylated protein to change from the unphosphorylated basin to the phosphorylated basin. Nature achieves this basin change by an enzymatic reaction that adds a phosphate group to the residue susceptible for phosphorylation. If the energy landscape were perfectly funneled with only a single set of native contacts (as for $\mathcal{H}_u$ and $\mathcal{H}_p$)[3], the free energy difference between the basins would be large if the two forms are very different.

In this study the sampling was performed with two different Hamiltonians. To understand the free energy profile for motion between the native (unphosphorylated and phosphorylated) basins, we use a simple approach to determine the barrier location and barrier height. We estimate an effective barrier height

by finding the minimum of the intersection of the two basins found in the free energy profiles.

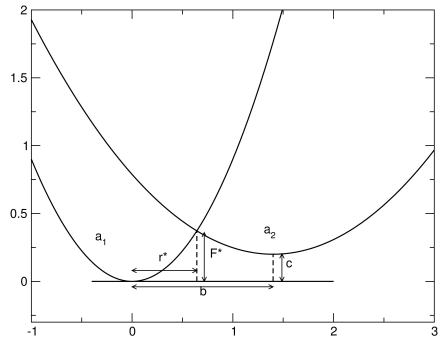


Figure 6.4: The illustration of free energy barrier estimation.

A further simplification is made assuming an isotropic, harmonic basin shape. The free energy profile around a basin with minimum position (Q_1, Q_2) is assumed to be of the form of $F(Q_u, Q_p) = \frac{a}{2}[(Q_u - Q_1)^2 + (Q_p - Q_2)^2] + F_0$. We study the profile along the reaction coordinates that link two basins (Q_1^u, Q_2^u) and (Q_1^p, Q_2^p) with a simple straight line. Without loss of generality, we assume the narrower of the two basins is at the origin, and the other basin is at distance $b = [(Q_1^u - Q_1^p)^2 + (Q_2^u - Q_2^p)^2]^{1/2}$. Their minima are at 0 and $c = \Delta F$ respectively. Along this one dimensional coordinate we have $F_1(r) = \frac{a_1}{2} r^2$ and $F_2(r) = \frac{a_2}{2}(r - b)^2 + c$ under the condition $a_1 \geq a_2$. As shown in Fig. 6.4 the intercept occurs at

$$r^{\#} = \frac{-a_2 b + [a_1 a_2 b^2 + 2c(a_1 - a_2)]^{1/2}}{a_1 - a_2}$$

The barrier height is then given by $F^{\#} = \frac{a_1}{2} r^{\#^2}$. If $a_1 = a_2 = a$, then we can compute $r^{\#} = b/2 + c/(ab)$. For the case of cystatin, we found that at $T = 1$, $Q^u = (0.64, 0.52)$ and $Q^p = (0.49, 0.62)$, we have $b = 0.57$, a rough fit gives $a = 500$ and $c = 0.01$. As a result we found that the barrier height of the free energy is $F^{\#} = 20$ for cystatin. Similarly we find at $T = 0.8 \times T^{room}$, $Q^u = (0.87, 0.6)$ and $Q^p = (0.52, 0.72)$, $c = 0.5$, $b = 1.17$, and $a = 600$, we found $F^{\#} = 90$ for NtrC. The unit of barrier height is given by $k_B \times T \sim 0.6$ kcal/mol. Note that both numbers seem rather high. As explained by Miyashita et al. [85] the local quadratic approximations are first of all quite rough and should only lead to an approximate barrier with the right order of magnitude. In reality, the barrier is much lower, because the transition state is not necessarily located on the straight line connecting the unphosphorylated basin with the phosphorylated basin. The height of the barrier should be interpreted as follows: In the context of a perfectly funneled landscape to a single minimum, the barrier located on the direct route between the unphosphorylated basin and phosphorylated basin of cystatin would be so large as to prevent an equilibrium of both conformations at the same time. We see this allows the phosphorylation event to act as a strict switch. For NtrC this barrier is several times larger and the only way for the unphosphorylated NtrC to reach the phosphorylated basin should be by means of more sophisticated

pathways including local unfolding. In our view it is clear that protein cracking motions [85, 86] are involved in the change.

Prediction of structural changes in cystatin with the Linear Response Method

Small structural changes in protein conformations upon pertubation can be predicted by a linear response method, which relates the changes in residue-residue interactions of the unphosphorylated Hamiltonian to the phosphorylated Hamiltonian. Experiments for cystatin indicate only minor, and hence small, global conformational change upon phosphorylation[72]. The main global changes of phosphorylation seen in the contact map in Fig. 6.1 include different contacts of the helical region (residues 10-28 for helix 1) with the β-like structures (residues 34-38 for strand 1, 40-46 for strand 2, 50-60 for strand 3, 80-93 for strand 4 and 100-105 for strand 5). There are also local rearrangements of contacts in the β strand 4 and the preceding loop region (residues 68-80) including the phospho-residue. These trends of structural changes were correctly captured by the PCA for the native-structure based simulations (see PC^{fold} in Fig. 6.2).

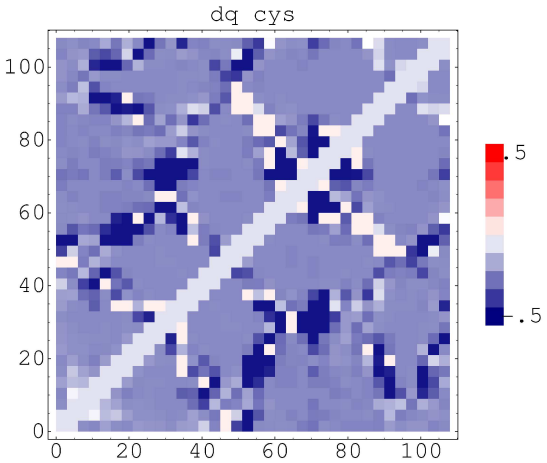


Figure 6.5: The linear response prediction of the changes of contact formation upon phosphorylation for cystatin.

We applied the linear response method to estimate the structural changes on the contact map of cystatin upon phosphorylating the protein. The result of the prediction of the change of contact formation, $\langle \delta q_{i,j} \rangle$, is shown in Fig. 6.5 as a contact map, which allows direct inspection of the residue-residue contact changes. The linear response method results are in excellent agreement with experiment. The global structural changes, i.e. the loss of contact formation between the helix and the β-like regions, were well captured. Further, the linear response method predicted the same local changes in the loop region around the phosphorylated residue as observed in experiments. Additionally, loss of loop contacts in residue region 65-75 were predicted. This region changes conformation and exhibits a 1.1Å RMS deviation of the phosphorylated native NMR structure from the unphosphorylated native NMR structure. It is clear that this linear response method developed to capture structural changes upon phosphorylation provides results consistent with experimental results.

Prediction of the phosphorylated conformation with an AMH-like contact potential

It would be desirable to have a transferable Hamiltonian, that can predict the structure of any protein before and after phosphorylation from sequence information alone. Much progress towards de novo structure prediction has already been made by our group with techniques like those employed in reference[77] and by other groups with other styles of energy function[87, 88] . However, the proteins that change under phosphorylation, as we see, probably deviate from a strictly funneled landscape. This makes the problem of complete de novo prediction more challenging than the usual. A much easier but still challenging computational problem would be to determine the structure of the phosphorylated test protein given only the structure of one form, say, the unphosphorylated conformation, or vice versa. Here we show how this can be done. To model how phosphorylation alters the tertiary structure of the protein conformation, we designed a predictive Hamiltonian $\mathcal{H}^*{}_p$, using short

range structural elements found in one form along with generic tertiary interactions. This Hamiltonian described in the method section is based on the de novo AMW prediction scheme. We call it the "phosphopredictive AMH". The Hamiltonian ${\mathcal{H}^*}_p$ uses, as the sole input, the conformation of the unphosphorylated protein for only the short and intermediate range interactions. This assures a strong bias in the short and medium class for local secondary structure to form such elements as seen in the unphosphorylated protein. To model the effect of phosphorylation we have introduced a tunable 3-well long range (in sequence space) residue-residue contact potential. This potential is modified to include interactions of the phospho-residue. The strategy to model a phospho-residue as a supercharged glutamic acid residue in the long range potential can now be tested. We call the resulting energy function the "phosphopredictive" Hamiltonian.

The first set of molecular dynamics simulations with the phosphopredictive Hamiltonian were performed with the original sequence of the unphosphorylated proteins, cystatin and NtrC. Since the input used is the contact map of the unphosphorylated protein, the predictive Hamiltonian mainly samples structures similar to those found in the folded basin of the unphosphorylated proteins when the long range term is added as a perturbation term. Additionally, the energetic contributions of short, medium and long range potentials were scaled to be equal in these simulations in keeping with estimates of the contributions of these parts of the interaction for funneled proteins. To check, whether the sampled structures were similar to the structures found in the folded basins that would be obtained with the pure native structure based Hamiltonians, these snapshots were projected onto the first two principal components obtained with the native structure based Hamiltonians. The projection of the snapshots obtained with this Hamiltonian for NtrC are shown in red in Fig. 6.6. Clearly the introduction of the long-range potential did not alter the ability to sample native unphosphorylated conformations. These projections serve as a baseline for the changes from results obtained with a pure native-structure based Hamiltonian to those from a Hamiltonian with a

heterogeneous contact potential.

Phosphorylation effects can be mimicked first by mutating the phospho-residue simply to a glutamic acid. Thus a set of molecular dynamics simulations with the predictive Hamiltonian based on a pure were performed with precisely this modification in which the phospho-residue was mutated to a glutamic acid. The snapshots for these simulations were projected onto the first two principal components and the results for NtrC were plotted in black in Fig. 6.6. Clearly, the snapshots only slightly deviate from the snapshots of the folded state of the unphosphorylated protein. To test if using a non-additive potential with water-mediated interactions will improve the quality of the prediction of the phosphorylated state, the same simulations were performed with the AMW potential [89]. Contact maps of each snapshots obtained with the AMW were computed and projected onto the pricncipal components. The AMW ensemble projection had almost identical values of PC^{phos} and PC^{fold}, and hence contact formation, as did the ensemble obtained with the simple contact based phosphopredictive Hamiltonian for the same glutamic acid mutant. The RMSD's of heavy atoms of both the predicted ensembles from the NMR structure of the phosphorylated NtrC were similar. The AMW had on average 0.1Å lower RMSDs from the NMR structure. Simulations with the AMW did show only minor improvement over the AMC in this case.

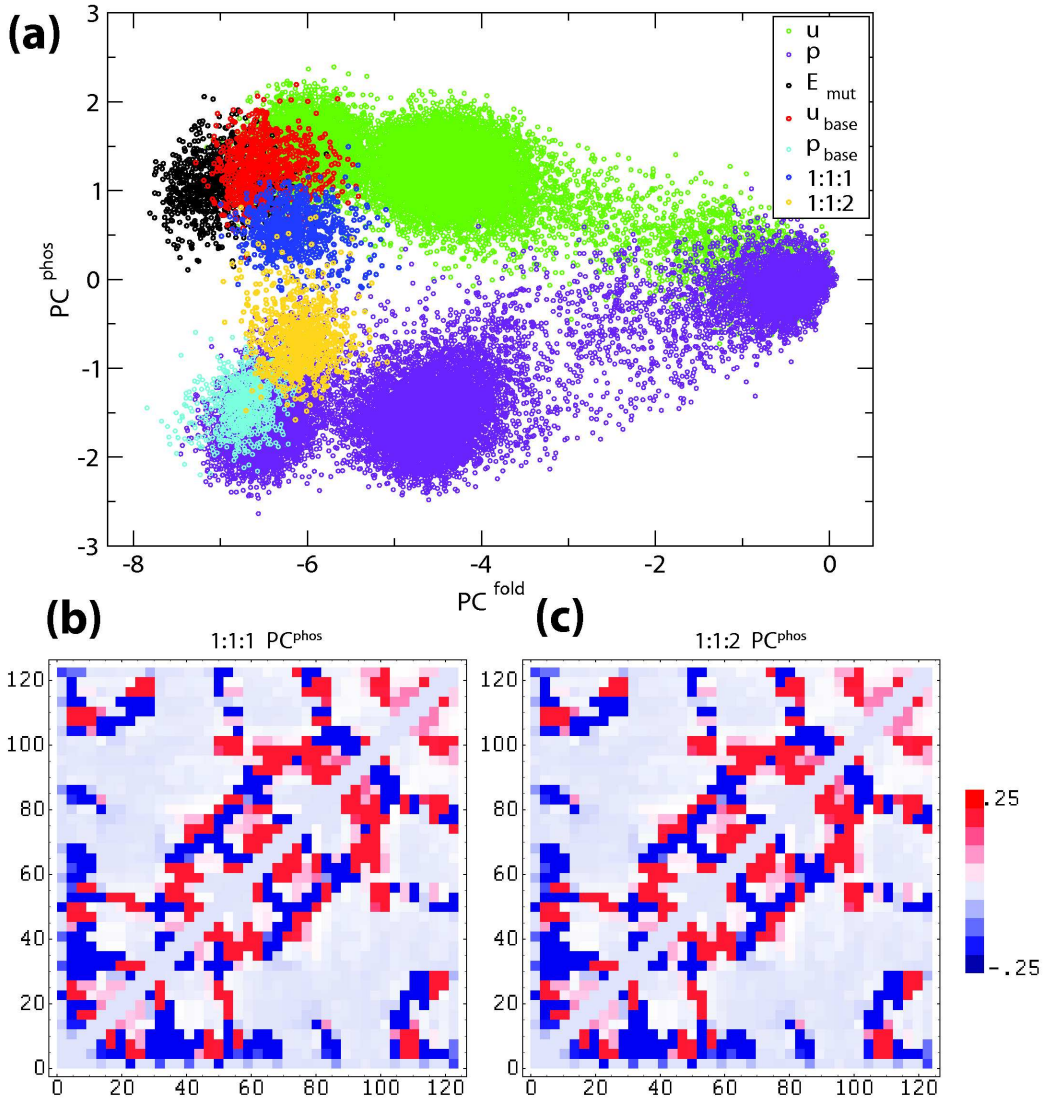


Figure 6.6: PCA of the contact maps for the conformations of NtrC obtained at T=0.75 (the lower temperature facilitates sampling of the folded structures) with the native structure based Hamiltonian and also the phosphopredictive AMH. Also shown are the contact maps for phosphorylation principal component for the ensembles obtained with the phosphopredictive Hamiltonian with short:medium:long range energy ratio of 1:1:1 and 1:1:2.

An important feature of our predictive Hamiltonian is the ability to "supercharge" the phospho-residue, that had been mutated into a glutamic acid. It is possibly to assign different weights to the strength of interaction of the supercharged residue with other residues. Simulations have been performed for two different scalings of the strength of interaction, namely 1.4 and 2.0. The difference in results obtained with Hamiltonians of these two charge scales is subtle. We will explicitly show only the results for a charge of 1.4. The contact maps of the structures sampled with the supercharged phosphopredictive AMH were computed and projected onto the folding and phosphorylation principal components (see Fig. 6.6, blue dots). The PC^{fold} values of the sampled conformations had similar PC^{fold} values to both the values of the unphosphorylated and phosphorylated ensembles. The more informative principal component, the phosphorylation principal component PC^{phos}, was shifted towards more negative values indicating enhanced formation of those contacts as seen in the phosphorylated ensemble rather than the unphosphorylated ensemble. To elucidate the predictive capability of the phosphopredictive Hamiltonian, the contact maps corresponding to PC^{phos} was plotted (Fig. 6.6). Defining four main helices in the native NMR structure of the phosphorylated form of NtrC (residues 15-27 correspond to helix 1, residues 36-42 to helix 2, residues 67-73 to helix 3 and residues 108-123 to helix 4), the contact map displays long-range contact changes for the phospho-residue (residue 54) with the turn region before helix 1, and also between the regions of helix 3 and helix 4, that are similar to the changes in contact formation seen for the vanilla Hamiltonian. The contact changes between the phospho-residue and the helix 2 region were not seen. Apart from those, the predictive Hamiltonian captured the long-range effects of the modified phospho-residue in good agreement with the experimental determinations. To measure the quality of the structures sampled with the phosphopredictive AMH, the RMSD of the heavy atoms from their native NMR structure were computed. The average RMS deviation from the NMR structure of the phosphorylated NtrC was about 2.7Å with a standard deviation of 0.1Å. Since the principal component analysis indicated a closer resemblance to the unphosphorylated ensemble rather than the phos-

phorylated ensemble, the RMSD of heavy atoms from the NMR structure of the unphosphorylated NtrC were also computed. The average RMSD was about 2.3Å with a standard deviation of 0.1Å. This result is not surprising due to the fact that the short and medium range structure is strongly biased towards the native structure of the unphosphorylated form. A more valid assesment of the quality of the predicted structure can be made by comparing the RMSD of the predicted ensemble from the respective ensembles, that would be obtained, when the NMR structures and sequences served as the sole input for the phosphopredictive Hamiltonian (see Fig. 6.6, red dots for the unphosphorylated ensemble and cyan dots for the phosphorylated ensemble). We call these ensembles the baseline ensembles. Both baseline ensembles have similar projections on the principal component space compared with their respective ensembles obtained with the vanilla Hamiltonians. The predicted ensemble (blue) has an average of about 2.5Å RMSD from both baseline ensembles, the phosphorylated and unphosphorylated one.

The simulations, so far, were performed with short, medium and long range contributions to the energy that are kept equal. This fact is motivated by the findings of Saven and Wolynes [18], who have estimated that in protein folding the contribution to the native energy arising from specific local interactions is comparable to those arising from specific tertiary interactions. It is therefore interesting to see whether different weights of the energetic contribution of the long range interactions might improve the predictions. Several sets of simulations were performed with different total strength of interactions ranging from half the original strength up to twice as large. Most simulations did not show any better structures than what could be predicted using simulations of the glutamic acid mutant only. Only the results for simulations with twice the strength of the long range interactions are therfore shown in Fig. 6.6. These results display the most improvement for the prediction results. The contact maps were computed and projected onto the folding and phosphorylation principal components (see Fig. 6.6, yellow dots). On a residue-residue contact level, this Hamiltonian best described the contact changes observed

upon phosphorylation of NtrC. The scaled long range interactions did perturb the local structure of the protein. This perturbed local structure leads to an increase of the RMS deviations from the phosphorylated NMR structure. The RMS deviations from the native NMR structure of the phosphorylated form were slightly higher than the deviations observed with the original phosphopredictive Hamiltonian with an average RMSD of 3.0Å from the NMR structure and a standard deviation of 0.1 Å. An overlay of several predicted structures is shown in Fig. 6.7 for visualization.

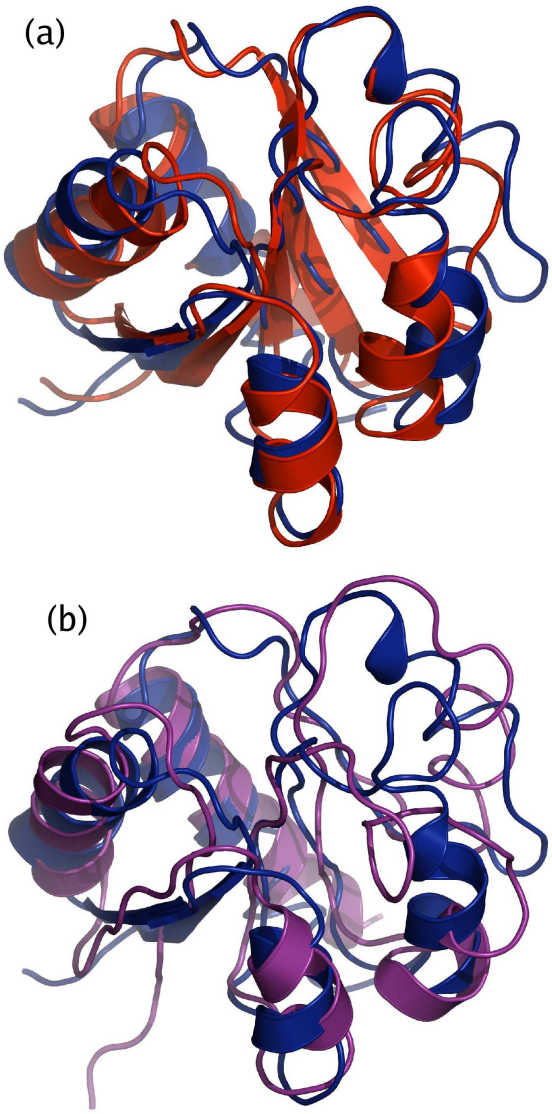


Figure 6.7: Overlay of a typical structure of NtrC (blue) obtained with the phosphopredictive Hamiltonian with the native NMR structure of the phosphorylated form of NtrC (purple) as well as the a representative structure (red), that is obtained with the same Hamiltonian for the sequence of the unphosphorylated NtrC.

Conclusions

We showed simulations with the native structure based Hamiltonians $\mathcal{H}_u$ and $\mathcal{H}_p$. While unphosphorylated and phosphorylated conformations both pre-exist on the landscape, the change of the landscape by post translational modification is needed to allow the different structure ensembles to compete. To relate the landscapes of the forms of a protein one can calculate the free energy differences using the cumulant expansion method (see Figures 6.2, 6.3). The perturbation approach shows how phosphorylation changes the free energy profile by tilting the landscape such that the phosphorylated basin was favored. The calculations show evolution has designed the unphosphorylated protein not to adopt the phosphorylated conformation (until the protein gets modified through phosphorylation) despite the fact that the RMSD between these conformations is not very large. For a simply funneled completely minimally frustrated protein landscape such as for our Hamiltonians $\mathcal{H}_u$ and $\mathcal{H}_p$, the unphosphorylated protein would rarely adopt the structure of the phosphorylated protein without post-translational modification. Partial unfolding mechanism are likely required for these dramatic conformational switching events in NtrC.

Principal component analysis allows us to visualize the conformations of the ensembles of unphosphorylated and phosphorylated test proteins by projecting all changes onto the first two dominant components. As shown in Fig. 6.2, 6.3, PC^{phos} especially indicates the major residue contacts that changed upon phosphorylation. This contact map compares quite well to the contact map obtained from the linear response theory prediction of the changes (Fig. 6.2, 6.5).

Finally we used a structure prediction Hamiltonian, $\mathcal{H}^*{}_p$, to predict the final phosphorylated conformation itself for these two systems. This algorithm successfully captures both the trends of conformational change of the unphosphorylated protein upon phosphorylation that are observed in experiments for the long range contacts of the phospho-residue and gives indeed the dominant

structures. The phospho-predictive AMH equips us with a powerful tool to predict the structure of phosphorylated proteins given information on the unphosphorylated conformation, or vice versa, and certainly pinpoints the major residue contact shifts. The Hamiltonian is general and captures the contact changes seen in small conformational changes as well as large conformational changes.

Acknowledgments

Computational support was provided in parts by the NSF based Center for Theoretical Biological Physics. This work was supported by NIH grant R01 GM044557.

Bibliography

[1] Bryngelson, J. D. and Wolynes, P. G. (1987) Spin-glasses and the statistical-mechanics of protein folding, *Proceedings of the National Academy of Sciences of the United States of America 84*, 7524–7528.

[2] Hardin, C., Eastwood, M. P., Prentiss, M., Luthey-Schulten, Z., and Wolynes, P. G. (2002) Folding funnels: The key to robust protein structure prediction, *Journal of Computational Chemistry 23*, 138–146.

[3] Onuchic, J. N., LutheySchulten, Z., and Wolynes, P. G. (1997) Theory of protein folding: The energy landscape perspective, *Annual Review of Physical Chemistry 48*, 545–600.

[4] Onuchic, J. N., Socci, N. D., LutheySchulten, Z., and Wolynes, P. G. (1996) Protein folding funnels: The nature of the transition state ensemble, *Folding & Design 1*, 441–450.

[5] Matouschek, A., Kellis, J. T., Serrano, L., and Fersht, A. R. (1989) Mapping the transition-state and pathway of protein folding by protein engineering, *Nature 340*, 122–126.

[6] Bryngelson, J. D., Onuchic, J. N., Socci, N. D., and Wolynes, P. G. (1995) Funnels, pathways, and the energy landscape of protein folding: a synthesis, *Proteins 21*, 167–195.

[7] Paci, E., Vendruscolo, M., Dobson, C. M., and Karplus, M. (2002) Determination of a transition state at atomic resolution from protein engineering data, *Journal of Molecular Biology 324*, 151–163.

[8] Lindorff-Larsen, K., Kritjansdottir, S., Teilum, K., Fieber, W., Dobson, C. M., Poulsen, F. M., and Vendruscolo, M. (2004) Determination of an ensemble of structures representing the denature state of the bovine acyl-coenzyme a binding protein, *Journal of American Chemical Society 126*, 3291–3299.

[9] Vendruscolo, M. and Dobson, C. M. (2003) Protein folding: bringing theory and experiment closer together, *Current Opinion in Structural Biology 13*, 1–6.

[10] Vendruscolo, M., Paci, E., Dobson, C. M., and Karplus, M. (2001) Three key residues form a critical netwrok in a protein folding transition state, *Nature 409*, 641–645.

[11] Lindorff-Larsen, K., Vendruscolo, M., Paci, E., and Dobson, C. M. (2004) Transition states for protein folding have native topologies despite high structural variability, *Nature structural and molecular biology 11*, 443–449.

[12] Lindorff-Larsen, K., Best, R. B., DePristo, M. A., Dobson, C. M., and Vendruscolo, M. (2005) Simultaneous determination of protein structure and dynamics, *Nature 433*, 128–132.

[13] Davis, R., Dobson, C. M., and Vendruscolo, M. (2002) Determination of the structures of distinct transition state ensembles for a beta-sheet peptide with parallel folding pathways, *Journal of Chemical Physics 117*, 9510–9517.

[14] Eastwood, M. P. and Wolynes, P. G. (2001) Role of explicitly cooperative interactions in protein folding funnels: A simulation study, *Journal of Chemical Physics 114*, 4702–4716.

[15] Eastwood, M. P., Hardin, C., Luthey-Schulten, Z., and Wolynes, P. G. (2003) Statistical mechanical refinement of protein structure prediction schemes. ii. mayer cluster expansion approach, *Journal of Chemical Physics 118*, 8500–8512.

[16] Go, N. (1983) Theoretical-studies of protein folding, *Annual Review of Biophysics and Bioengineering 12*, 183–210.

[17] Eastwood, M. P., Hardin, C., Luthey-Schulten, Z., and Wolynes, P. G. (2001) Evaluating protein structure-prediction schemes using energy landscape theory, *IBM Journal of Research and Development 45*, 475–497.

[18] Saven, J. G. and Wolynes, P. G. (1996) Local conformational signals and the statistical thermodynamics of collapsed helical proteins, *Journal of Molecular Biology 257*, 199–216.

[19] Onuchic, J. N., Socci, N. D., Luthey-Schulten, Z., and Wolynes, P. G. (1996) Protein folding funnels: the nature of the transition state ensemble, *Folding and Design 1*, 441–450.

[20] Shen, T., Zong, C., Hamelberg, D., McCammon, J. A., and Wolynes, P. G. (2005) The folding energy landscape and phosphorylation: modeling the conformational switch of the nfat regulatory domain, *FASEB Journal 19*, 1389–1395.

[21] Fitch, W. M. and Margoliash, E. (1967) Construction of phylogenetic trees, *Science 155*, 279–284.

[22] Felsenstein, J. (1993) Phylip- phylogeny inference package, *Department of Genetics, University of Washington, Seattle*.

[23] Huang, G. S. and Oas, T. G. (1995) Structural and stability of monomeric lambda repressor: Nmr evidence for two-state folding, *Biochemistry 34*, 3884–3892.

[24] Beamer, L. J. and Pabo, C. O. (1992) Refined 1.8 a crystal structure of the lambda repressor-operator complex, *Journal of Molecular Biology 227*, 177–196.

[25] Cho, S. S., Levy, Y., and Wolynes, P. G. (2006) P versus q: structural reaction coordinates capture protein folding on smooth landscapes, *Proceedings of the National Academy of Sciences 103*, 586–591.

[26] Koradi, R., Billeter, M., and Wuthrich, K. (1996) Molmol: A program for display and analysis of macromolecular structures, *Journal of Molecular Graphing 14*, 51–55.

[27] Dang, C. V. and Lee, W. M. F. (1988) Identification of the human c-myc protein nuclear translocation signal, *Molecular and Cellular Biology 8*, 4048–4054.

[28] Dingwall, C. and Laskey, R. A. (1991) Nuclear targeting sequences - a consensus, *Trends in Biochemical Sciences 16*, 478–481.

[29] Kalderon, D., Richardson, W. D., Markham, A. F., and Smith, A. E. (1984) Sequence requirements for nuclear location of simian virus-40 large-t-antigen, *Nature 311*, 33–38.

[30] Lanford, R. E. and Butel, J. S. (1984) Construction and characterization of an sv40 mutant defective in nuclear transport of t-antigen, *Cell 37*, 801–813.

[31] Leung, S. W., Harreman, M. T., Hodel, M. R., Hodel, A. E., and Corbett, A. H. (2003) Dissection of the karyopherin alpha nuclear localization signal (nls)-binding groove - functional requirements for nls binding, *Journal of Biological Chemistry 278*, 41947–41953.

[32] Baeuerle, P. A. and Baltimore, D. (1988) I-kappa-b - a specific inhibitor of the nf-kappa-b transcription factor, *Science 242*, 540–546.

[33] Baeuerle, P. A. and Baltimore, D. (1988) Activation of dna-binding activity in an apparently cytoplasmic precursor of the nf-kappa-b transcription factor, *Cell 53*, 211–217.

[34] Huxford, T., Malek, S., and Ghosh, G. (1999) Structure and mechanism in nf-kappa b/i kappa b signaling, *Cold Spring Harbor Symposia on Quantitative Biology 64*, 533–540.

[35] Malek, S., Huxford, T., and Ghosh, G. (1998) I kappa b alpha functions through direct contacts with the nuclear localization signals and the dna binding sequences of nf-kappa b, *Journal of Biological Chemistry 273*, 25427–25435.

[36] Phelps, C. B., Sengchanthalangsy, L. L., Huxford, T., and Ghosh, G. (2000) Mechanism of i kappa b alpha binding to nf-kappa b dimers, *Journal of Biological Chemistry 275*, 29840–29846.

[37] Huang, D. B., Huxford, T., Chen, Y. Q., and Ghosh, G. (1997) The role of dna in the mechanism of nf kappa b dimer formation: crystal structures of the dimerization domains of the p50 and p65 subunits, *Structure 5*, 1427–1436.

[38] Jacobs, M. D. and Harrison, S. C. (1998) Structure of an i kappa b alpha/nf-kappa b complex, *Cell 95*, 749–758.

[39] Dyson, H. J. and Wright, P. E. (2002) Coupling of folding and binding for unstructured proteins, *Current Opinion in Structural Biology 12*, 54–60.

[40] Conti, E. and Kuriyan, J. (2000) Crystallographic analysis of the specific yet versatile recognition of distinct nuclear localization signals by karyopherin alpha, *Structure with Folding and Design 8*, 329–338.

[41] Huxford, T., Huang, D. B., Malek, S., and Ghosh, G. (1998) The crystal structure of the i kappa b alpha/nf-kappa b complex reveals mechanisms of nf-kappa b inactivation, *Cell 95*, 759–770.

[42] Malek, S., Chen, Y., Huxford, T., and Ghosh, G. (2001) I kappa b beta but not i kappa b alpha, functions as a classical cytoplasmic inhibitor of nf-kappa b dimers by masking both nf-kappa b nuclear localization sequences in resting cells, *Journal of Biological Chemistry 276*, 45225–45235.

[43] Koshland, D. E. (1995) The key-lock theory and the induced fit theory, *Angewandte Chemie-International Edition 33*, 2375–2378.

[44] Papoian, G. A. and Wolynes, P. G. (2003) The physics and bioinformatics of binding and folding- an energy landscape perspective, *Biopolymers 68*, 333–349.

[45] Levy, Y., Wolynes, P. G., and Onuchic, J. N. (2004) Protein topology determines binding mechanism, *Proceedings of the National Academy of Sciences of the United States of America 101*, 511–516.

[46] Croy, C. H., Bergqvist, S., Huxford, T., Ghosh, G., and Komives, E. A. (2004) Biophysical characterization of the free i kappa b alpha ankyrin repeat domain in solution, *Protein Science 13*, 1767–1777.

[47] Bergqvist, S., Hughes, C., Huxford, T., Ghosh, G., and Komives, E. (2004) Thermodynamics and kinetics of the nf-kb/ikb and nf-kb/dna interactions, *Biophysical Journal 86*, 513a.

[48] Kim, D., Xu, D., Guo, J. T., Ellrott, K., and Xu, Y. (2003) Prospect ii: protein structure prediction program for genome-scale applications, *Protein Engineering 16*, 641–650.

[49] Liou, H. C., Nolan, G. P., Ghosh, S., Fujita, T., and Baltimore, D. (1992) The nf-kappa-b p50 precursor, p105, contains an internal i-kappa-b-like inhibitor that preferentially inhibits p50, *Embo Journal 11*, 3003–3009.

[50] Janin, J. (2002) Welcome to capri: A critical assessment of predicted interactions, *Proteins-Structure Function and Genetics 47*, 257–257.

[51] Cheng, J. D., Ryseck, R. P., Attar, R. M., Dambach, D., and Bravo, R. (1998) Functional redundancy of the nuclear factor kappa b inhibitors i kappa b alpha and i kappa b beta, *Journal of Experimental Medicine 188*, 1055–1062.

[52] Malek, S., Huang, D. B., Huxford, T., Ghosh, S., and Ghosh, G. (2003) X-ray crystal structure of an i kappa b beta center dot nf-kappa b p65 homodimer complex, *Journal of Biological Chemistry 278*, 23094–23100.

[53] Ernst, M. K., Dunn, L. L., and Rice, N. R. (1995) The pest-like sequence of i-kappa-b-alpha is responsible for inhibition of dna-binding but not for cytoplasmic retention of c-rel or rela homodimers, *Molecular and Cellular Biology 15*, 872–882.

[54] Tam, W. F., Wang, W. H., and Sen, R. J. (2001) Cell-specific association and shuttling of i kappa b alpha provides a mechanism for nuclear nf-kappa b in b lymphocytes, *Molecular and Cellular Biology 21*, 4837–4846.

[55] Johnson, C., Van Antwerp, D., and Hope, T. J. (1999) An n-terminal nuclear export signal is required for the nucleocytoplasmic shuttling of i kappa b alpha, *Embo Journal 18*, 6682–6693.

[56] Latimer, M., Ernst, M. K., Dunn, L. L., Drutskaya, M., and Rice, N. R. (1998) The n-terminal domain of i kappa b alpha masks the nuclear localization signal(s) of p50 and c-rel homodimers, *Molecular and Cellular Biology 18*, 2640–2649.

[57] Shindyalov, I. N. and Bourne, P. E. (1998) Protein structure alignment by incremental combinatorial extension (ce) of the optimal path, *Protein Engineering 11*, 739–747.

[58] Brunger, A. T., Adams, P. D., Clore, G. M., Delano, W. L., Gros, P., Grosse-Kunstleve, R. W., Jiang, J.-S., Kuszewski, J., Nilges, M., Pannu, N. S., Read, R. J., Rice, L. M., Simonson, T., and Warren, G. L. (2005) Crystallography and nmr system: a new sofftware suite for macromolecular structure determination, *Crystallography section D* pp. 905–921.

[59] Hunter, T. (2000) Signaling - 2000 and beyond, *Cell 100*, 113–127.

[60] Johnson, L. N. and Barford, D. (1993) The effects of phosphorylation on the structure and function of proteins, *Annual Review of Biophysics and Biomolecular Structure 22*, 199–232.

[61] Steen, H., Jebanathirajah, J. A., Springer, M., and Kirschner, M. W. (2005) Stable isotope-free relative and absolute quantitation of protein phosphorylation stoichiometry by ms, *Proceedings of the National Academy of Sciences 102*, 3948–3953.

[62] Radhakrishnan, I., PerezAlvarado, G. C., Parker, D., Dyson, H. J., Montminy, M. R., and Wright, P. E. (1997) Solution structure of the kix domain of cbp bound to the transactivation domain of creb: A model for activator:coactivator interactions, *Cell 91*, 741–752.

[63] Buck, M. and Rosen, M. K. (2001) Flipping a switch, *Science 291*, 2329–2330.

[64] Ramelot, T. A. and Nicholson, L. K. (2001) Phosphorylation-induced structural changes in the amyloid precursor protein cytoplasmic tail detected by NMR, *J. Mol. Biol. 307*, 871–84.

[65] Johnson, L. N. and Lewis, R. J. (2001) Structural basis for control by phosphorylation, *Chemical Reviews 101*, 2209–2242.

[66] Pufall, M., Lee, G. M., Nelson, M.L., K. H. S., Velyvis, A., Kay, L. E., McIntosh, L. P., and Graves, B. J. (2005) Variable control of ets-1 dna binding by multiple phosphates in an unstructured region, *Science 309*, 142–145.

[67] Park, K.-S., Mohapatra, D. P., Misonou, H., and Trimmer, J. S. (2006) Graded regulation of the kv2.1 potassium channel by variable phosphorylation, *Science 18*, 976–979.

[68] Okamura, H., Aramburu, J., Garcia-Rodriguez, C., Viola, J. P. B., Raghavan, A., Tahiliani, M., Zhang, X., Qin, J., Hogan, P., and Rao, A. (2000) Concerted dephosphorylation of the transcription factor nfat1 induces a conformational switch that regulates transcriptional activity, *Molecular Cell 6*, 539–50.

[69] Tholey, A., Pipkorn, R., Bossemeyer, D., Kinzel, V., and Reed, J. (2001) Influence of myristoylation, phosphorylation, and deamidation on the structural behavior of the N-terminus of the catalytic subunit of CAMP-dependent protein kinase, *Biochemistry 40*, 225–231.

[70] Shen, T., Wong, C. F., and McCammon, J. A. (2001) Atomistic Brownian dynamics simulation of peptide phosphorylation, *J. Am. Chem. Soc. 123*, 9107–9111.

[71] Groban, E. S., Narayanan, A., and Jacobson, M. P. (2006) Conformational changes in protein loops and helices induced by post-translational phosphorylation, *PLoS Comput Biol 2*, 238–250.

[72] Dieckmann, T., Mitschang, L., Hofmann, M., Kos, J., Turk, V., Auerswald, E. A., Jaenicke, R., and Oschkinat, H. (1993) The structures of native phosphorylated chicken cystatin and of a recombinant unphosphorylated variant in solution, *Journal of Molecular Biology 234*, 1048–1059.

[73] Iakoucheva, L. M., Radivojac, P., Brown, C. J., O'Connor, T. R., Sikes, J. G., Obradovic, Z., and Dunker, A. K. (2004) The importance of intrinsic disorder for protein phosphorylation, *Nucleic Acids Research 32*, 1037–1049.

[74] Kern, D., Volkman, B. F., Luginbuhl, P., Nohaile, M. J., Kustu, S., and Wemmer, D. E. (1999) Structure of a transiently phosphorylated switch in bacterial signal transduction, *Nature 402*, 894–898.

[75] Kern, D., Volkman, B. F., and Wemmer, D. E. (2001) A signaling protein 'in action' - structure and dynamics of a transiently phosphorylated switch, *Biophysical Journal 80*, 13a–13a.

[76] Hardin, C., Eastwood, M. P., Luthey-Schulten, Z., and Wolynes, P. G. (2000) Associative memory hamiltonians for structure prediction without homology: Alpha-helical proteins, *Proceedings of the National Academy of Sciences of the United States of America 97*, 14235–14240.

[77] Prentiss, M. C., Hardin, C., Eastwood, M. P., Zong, C. H., and Wolynes, P. G. (2006) Protein structure prediction: The next generation, *Journal of Chemical Theory and Computation 2*, 705–716.

[78] Hardin, C., Eastwood, M., Luthey-Schulten, Z., and Wolynes, P. G. (2001) Associative memory hamiltonians for structure prediction without homology., *Abstracts of Papers of the American Chemical Society 221*, U401–U401.

[79] Latzer, J., Eastwood, M. P., and Wolynes, P. G. (2006) Simulation studies of the fidelity of biomolecular structure ensemble recreation, *Journal of Chemical Physics 125*, 214905.

[80] Ikeguchi, M., Ueno, J., Sato, M., and Kidera, A. (2005) Protein structural change upon ligand binding: Linear response theory, *Phys. Rev. Lett. 94*, 078102.

[81] Saito, N., Hashitsume, N., Toda, M., and Kubo, R. (2003) *Statistical Physics II: Nonequilibrium Statistical Mechanics*. (Springer, New York), 2nd edition.

[82] Jiang, Z. H. G. and McKnight, C. J. (2006) A phosphorylation-induced conformation change in dematin headpiece, *Structure 14*, 379–387.

[83] Hupp, T. R. and Lane, D. P. (1995) Two distinct signaling pathways activate the latent dna binding function of p53 in a casein kinase ii-independent manner, *J. Biol. Chem. 270*, 18165.

[84] Biswas, P., Zou, J. M., and Saven, J. G. (2005) Statistical theory for protein ensembles with designed energy landscapes, *Journal of Chemical Physics 123*, 1.

[85] Miyashita, O., Onuchic, J. N., and Wolynes, P. G. (2003) Nonlinear elasticity, proteinquakes, and the energy landscapes of functional transitions in proteins, *Proceedings of the National Academy of Sciences 100*, 12570–12575.

[86] Ansari, A., Berendzen, J., Bowne, S. F., Frauenfelder, H., Iben, I. E. T., Sauke, T. B., Shyamsunder, E., and Young, R. D. (1985) Protein States and Proteinquakes, *PNAS 82*, 5000–5004.

[87] Misura, K. M. S., Chivian, D., Rohl, C. A., Kim, D. E., and Baker, D. (2006) Physically realistic homology models built with rosetta can be more accurate than their templates, *Proceedings of the National Academy of Sciences 103*, 5361–5366.

[88] Yang, J. S., Chen, W. W., Skolnick, J., and Shakhnovich, E. I. (2007) All-atom ab initio folding of a diverse set of proteins, *STRUCTURE 15*, 53–63.

[89] Papoian, G. A., Ulander, J., Eastwood, M. P., Luthey-Schulten, Z., and Wolynes, P. G. (2004) Water in protein structure prediction, *Proceedings of the National Academy of Sciences 101*, 3352–3357.

[90] A. M. Bonvin and A. T. Brünger. Conformational variability of solution nuclear magnetic resonance structures. *Journal of Molecular Biology*, 250:80–93, 1995.

[91] F. T. Burling and A. T. Brünger. Thermal motion and conformational disorder in protein crystal structures: Comparison of multi-conformer and time-averaging models. *Israel Journal of Chemistry*, 34:165–175, 1994.

[92] M. M. Dedmon, K. Lindorff-Larsen, J. Christodoulou, M. Vendruscolo, and C. M. Dobson. Mapping long-range interactions in alpha-synuclein using spin-label nmr and ensemble molecular dynamics simulations. *Journal of American Chemical Society*, 127:476477, 2005.

[93] M. R. Ejtehadi, S. P. Avall, and S. S. Plotkin. Three-body interactions improve the prediction of rate and mechanism in protein folding models. *Proceedings of the National Academy of Sciences of the United States of America*, 101:15088–15093, 2004.

[94] P. J. Flory. Theory of elastic mechanisms in fibrous proteins. *Journal of American Chemical Society*, 78:5222–5235, 1956.

[95] R. A. Goldstein, Z. A. Luthey-Schulten, and P. G. Wolynes. Optimal protein folding codes from spin glass theory. *Proceedings of the National Academy of Sciences*, 95:4299–4302, 1998.

[96] J. Gsponer, H. Hopearuoho, S. B. M. Whittaker, G. R. Spence, G. R. Moore, E. Paci, S. E. Radford, and M. Vendruscolo. Determination of an ensemble of structures representing the intermediate state of the bacterial immunity protein im7. *Proceedings of the National Academy of Sciences of the United States of America*, 103:99–104, 2006.

[97] M. Habeck, M. Nilges, and W. Rieping. Bayesian inference applied to macromolecular structure determination. *Physical Review E*, 72(3):–, 2005.

[98] M. Habeck, W. Rieping, and M. Nilges. Weighting of experimental evidence in macromolecular structure determination. *Proceedings of the National Academy of Sciences of the United States of America*, 103(6):1756–1761, 2006.

[99] R. W. Hall and P. G. Wolynes. Microscopic theory of network glasses. *Physical Review Letters*, 90(8):–, 2003.

[100] A. Jack and M. Levitt. Refinement of large structures by simultaneous minimization of energy and r-factor. *Acta Crystallographica Section A*, 34:931–935, 1978.

[101] H. Jacobson and W. H. Stockmayer. Intramolecular reaction in polycondensations. i. the theory of linear systems. *Journal of Chemical Physics*, 18:1600–1606, 1950.

[102] J. Kuriyan, K. Osapay, S. K. Burley, A. T. Brünger, W. A. Hendrickson, and M. Karplus. Exploration of disorder in protein structures by x-ray restrained molecular dynamics. *Proteins*, 10:340–58, 1991.

[103] Z. A. Luthey-Schulten, B. E. Ramirez, and P. G. Wolynes. Helix-coil, liquid crystal, and spin glass transitions of a collapsed heteropolymer. *Journal of Physcial Chemistry*, 99:2177–2185, 1995.

[104] M. Mezzard, G. Parisi, and M. A. Virasoro. *Spin Glass Theory and Beyond*. Singapore, New Jersey, Hong Kong: World Scientific, 1987.

[105] S. Miyazawa and R. L. Jernigan. Residue-residue potentials with a favorable contact pair term and an unfavorable high packing density term, for simulation and threading. *Journal of Molecular Biology*, 256:623–644, 1996.

[106] V. Munoz and L. Serrano. Elucidating the folding problem of helical peptides using empirical parameters. *Nature Structural Biology*, 1:399–409, 1994.

[107] J. N. Onuchic, J. Wang, and P. G. Wolynes. Analyzing single molecule trajectories on complex energy landscapes using replica correlation functions. *Chemical Physics*, 247(1):175–184, 1999.

[108] V. S. Pande, A. Y. Grosberg, and T. Tanaka. Heteropolymer freezing and design: Towards physical models of protein folding. *Reviews of Modern Physics*, 72:259314, 2000.

[109] S. S. Plotkin, J. Wang, and P. G. Wolynes. Statistical mechanics of a correlated energy landscape model for protein folding funnels. *Journal of Chemical Physics*, 106(7):2932–2948, 1997.

[110] J. J. Portman, S. Takada, and P. G. Wolynes. Variational theory for site resolved protein folding free energy surfaces. *Physical Review Letters*, 81:5237–5240, 1998.

[111] J. J. Portman, S. Takada, and P. G. Wolynes. Microscopic theory of protein folding rates. i. fine structure of the free energy profile and folding routes from a variational approach. *Journal of Chemical Physics*, 114:5069–5081, 2001.

[112] W. Rieping, M. Habeck, and M. Nilges. Inferential structure determination. *Science*, 309(5732):303–306, 2005.

[113] T. Y. Shen, C. P. Hofmann, M. Oliveberg, and P. G. Wolynes. Scanning malleable transition state ensembles: Comparing theory and experiment for folding protein u1a. *Biochemistry*, 44:6433–6439, 2005.

[114] B. A. Shoemaker, J. Wang, and P. G. Wolynes. Structural correlations in protein folding funnels. *Proceedings of the National Academy of Sciences of the United States of America*, 94(3):777–782, 1997.

[115] B. A. Shoemaker, J. Wang, and P. G. Wolynes. Exploring structures in protein folding funnels with free energy functionals: The transition state ensemble. *Journal of Molecular Biology*, 287(3):675–694, 1999.

[116] B. A. Shoemaker and P. G. Wolynes. Exploring structures in protein folding funnels with free energy functionals: The denatured ensemble. *Journal of Molecular Biology*, 287(3):657–674, 1999.

[117] Y. Ueda, H. Taketomi, and N. Go. Studies of protein folding, unfolding, and uctuations by computer simulation, 2. 3-dimensional lattice model for lysozyme. *Biopolymers*, 7:1531–1548, 1978.

[118] M. Vendruscolo and C. M. Dobson. Towards complete descriptions of the free-energy landscapes of proteins. *Philosophical Transactions of the Royal Society of London Series a-Mathematical Physical and Engineering Sciences*, 363(1827):433–450, 2005.

[119] P. G. Wolynes. Folding funnels and energy landscapes of larger proteins within the capillarity approximation. *Proceedings of the National Academy of Sciences of the United States of America*, 94:6170–6175, 1997.

[120] K. Sugase and H. J. Dyson and P. E. Wright. Mechanism of coupled folding and binding of an intrinsically disordered protein *Nature*, 447:1021–U11, 2007.

[121] A. P. Capaldi, C. Kleanthous, and S. E. Radford. Im7 folding mechanism: misfolding on a path to the native state. *Nature Structural Biology*, 9(3):209–215, 2002.

[122] C. Clementi, P. A. Jennings, and J. N. Onuchic. How native state topology affects the folding of dihydrofolate reductase and interleukin-beta. *Proc Natl Acad Sci*, 97:5871–5876, 2000.

[123] C. Clementi, H. Nymeyer, and J. N. Onuchic. 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. *Journal of Molecular Biology*, 298:937–953, 2000.

[124] M. R. Ejtehadi, S. P. Avall, and S. S. Plotkin. Three-body interactions improve the prediction of rate and mechanism in protein folding models. *Proceedings of the National Academy of Sciences*, 101:15088–15093, 2004.

[125] R. A. Goldstein, Z. A. Luthey-Schulten, and P. G. Wolynes. Optimal protein folding codes from spin glass theory. *Proceedings of the National Academy of Sciences*, 95:4299–4302, 1998.

[126] C. Hardin, M. P. Eastwood, M. Prentiss, Z. Luthey-Schulten, and P. G. Wolynes. Folding funnels: The key to robust protein structure prediction. *Journal of Computational Chemistry*, 23(1):138–146, 2002.

[127] Corey Hardin, Zaida Luthey-Schulten, and Peter G. Wolynes. Backbone dynamics, fast folding, and secondary structure formation in helical proteins and peptides. *PROTEINS:Structure,Function, and Genetics*, 34:281–294, 1999.

[128] S. Miyazawa and R. L. Jernigan. Residue-residue potentials with a favorable contact pair term and an unfavorable high packing density term, for simulation and threading. *Journal of Molecular Biology*, 256:623–644, 1996.

[129] V. Sobolev, A. Sorokine, J. Prilusky, E. E. Abola, and M. Edelman. Automated analysis of interatomic contacts in proteins. *Bioinformatics*, 15:327–332, 1999.