The C-terminal domain of transcription factor RfaH:

Folding, fold switching and energy landscape

Bahman Seifi[†] and Stefan Wallin^{†,*}

*[†]Department of Physics and Physical Oceanography,
Memorial University of Newfoundland,
St Johns, NL A1B 3X7, Canada.*

*Correspondance. E-mail: swallin@mun.ca

Abbreviations: N-terminal domain, NTD; C-terminal domain, CTD; Monte Carlo, MC; structure-based model, SBM.

Abstract

We study the folding and fold switching of the C-terminal domain (CTD) of the transcription factor RfaH using a hybrid sequence-structure based model. We show that this model captures the essential thermodynamic behavior of this metamorphic domain, i.e., a switch in the global free energy minimum from an α -helical hairpin to a 5-stranded β -barrel upon separating the CTD from the rest of the protein. Using this model and Monte Carlo sampling techniques, we analyze the energy landscape of the CTD in terms of progress variables for folding towards the two folds. We find that, below the folding temperature, the energy landscape is characterized by a single, dominant funnel to the native β -barrel structure. The absence of a deep funnel to the α -helical hairpin state reflects a negligible population of this fold for the isolated CTD. We observe, however, a significantly higher α -helix structure content in the unfolded state compared to results from a similar but fold switch-incompetent version of our model. Moreover, in folding simulations started from an extended chain conformation we find transient α -helix structure that disappears as the chain progresses to the thermally stable β -barrel state.

1 Introduction

Proteins are increasingly being discovered with a remarkable ability to switch from one native fold into another one with widely different structure [1–3]. While it is not uncommon for proteins to undergo large-scale motions after their initial folding, such as domain-swapping [4] or other hinge-like motions [5], fold switching is a distinct phenomenon. By definition, it involves a reorganization of the protein at the most basic structural level at play in folding, i.e., secondary structure (α -helices and β -sheets). Despite the remarkable complexity of these molecular transformations, fold switching is reversible and thereby controlled by the system’s free energy. In this sense, metamorphic proteins can be said to adhere to Anfinsen’s thermodynamic principle (or hypothesis) of protein folding [6]. Clearly, however, fold switching fundamentally challenges the idea of a “unique” native conformation, which was a central aspect of the classic view of folding since emerging from the pioneering refolding experiments on ribonuclease A [7]. It is important to note that fold switching typically occurs only when triggered by specific changes to the local environment or milieu of the protein, such as salt concentration [8], redox condition [9] or oligomerization state [10]. In the absence of such a trigger, metamorphic proteins typically fold to an apparently unique structure, which masks their fold switching capabilities. As a result, metamorphic proteins often go unrecognized [11].

Metamorphic proteins thus encode two different folds within a single amino acid sequence even though, as mentioned, they typically adopt a single fold for a given (constant) local milieu. It is natural, then, to ask: what impact does this dual encoding have on their folding? This question is in fact related to a classic line of inquiry in protein folding, namely whether the mechanism of folding is conserved among homologous proteins or, more generally, among sequences adopting the same fold [12, 13]. It was observed, remarkably, that sequences with low sequence similarity (but still adopting the same native fold) often fold in a very similar manner [14–16], however, this conservation break down at very low sequence similarity [17]

Figure 1: **The two different folds of the C-terminal domain (CTD) of RfaH.** Native contacts in the two folds, determined using the shadow map method [54], are shown as thin lines in the C_α -traces (left and right) and as solid squares in the contact maps (center). The two contact sets, denoted C^β (β -barrel) and C^α (α -helical hairpin), have 13 common contacts (green).

and does not extend to all fold classes [17–19]. The sequences of metamorphic proteins have diverged from their homologs to the point of nearly adopting a new unique fold. Characterizing the folding of metamorphic proteins may shed light on the limits of how the amino acid sequence can shape the energy landscape of proteins. It may also be practically important in the field of protein metamorphism because any identifying feature, including folding behavior, can be used to discover as of yet unknown fold switching events [10, 11, 20].

Here we study the C-terminal domain (CTD) of the transcriptional antiterminator protein RfaH, a prototypical metamorphic protein [21]. This fold switching exhibited by this protein has been studied extensively both experimentally [22–25] and computationally [26–33]. The RfaH CTD, on its own, i.e., excised from the rest of protein, folds spontaneously into a β -barrel-like fold with 5 strands, as shown in Fig. 1. This structure is virtually identical to the CTD structure of other members of the NusG/Ste5 family of transcription factors to which RfaH belongs [34]. However, as part of the full-length RfaH, the CTD adopts instead an α -helical hairpin fold [22]. This entirely different folded state is stabilized by favorable interactions with the RfaH N-terminal domain (NTD) [23]. The switch in structure from

all- α to all- β is triggered when RfaH binds to RNA polymerase in a paused state, which underpins RfaH's regulatory function in transcription and translation [24, 35].

Simulating the structural transitions of metamorphic proteins on the computer has the potential for detailed insight but is technically challenging. Some studies have focused on alleviating the formidable problem of achieving sufficient conformational sampling for these large-scale transitions by various enhanced sampling techniques [26, 28, 32, 36, 37]. This approach assumes, however, an underlying model that accurately captures the delicate free energy balance between folds. This is not guaranteed even for modern explicit-water molecular dynamics force fields [38], although it has been achieved within the framework of coarse-grained 3-letter protein model [39–41]. Other studies have relied on so-called structure-based models (SBM). The basic idea of SBMs is to make the contacts present in a given target (typically native) structure artificially favorable, while non-native interactions are left neutral or even made repulsive [42]. SBMs can be constructed with several basin of attractions, each one representing a different target structure. Single-basin SBMs have a long tradition in protein folding studies [43] and has had notable successes in reproducing detailed experimental data on specific proteins, such as folding rates [44, 45] and transition-state structures [46], despite the *ad hoc* nature of the approach. Dual- or multi-basin SBMs are the logical extension to metamorphic proteins, and such models have been applied to both the CTD of RfaH [27, 30] and the mutation-induced fold switch between the GA and GB proteins [47, 48].

To study the energy landscape of the RfaH CTD, we take a hybrid approach that combines a physics-based model for protein folding with a dual-basin SBM. We aim to develop this hybrid sequence-structure based model based on its basic thermodynamic behavior, which we determine using extensive Monte Carlo simulations. Hence, we require that this model captures the switch in global free energy minimum between the two folds of RfaH CTD, depending on whether it is part of the full-length RfaH or considered as an isolated fragment. We also consider a single-basin version of our model that does not exhibit proper fold

switching. By comparing closely with this model, we probe the impact of encoding for two different folds on the energy landscape of RfaH CTD.

2 Materials and Methods

2.1 Physics-based computational protein model

All simulations are carried out using the software package PROFASI [49]. The physics-based model implemented in this package is described in Ref. [50]. Briefly, the model combines an all-atom protein representation with an effective potential energy function with 4 terms: $E^{(0)} = E_{\text{loc}} + E_{\text{ev}} + E_{\text{hb}} + E_{\text{sc}}$. Solvent molecules are implicitly taken into account by the energy function. The term E_{ev} implements excluded-volume effects between all atom pairs using a $1/r^{12}$ -function. The term E_{loc} includes interactions between atoms close along the chain, e.g., between partial charges on neighboring peptide planes, which help provide a good local chain description. The remaining two terms, E_{hb} and E_{sc} , represent hydrogen bonding and sidechain-sidechain interactions, respectively. Hydrogen bonding is implemented through orientationally dependent attractions between donor and acceptor groups. The term E_{sc} includes both pairwise interactions between partial charges on sidechains and effective hydrophobic attractions.

2.2 Monte Carlo simulations

Simulated tempering Monte Carlo simulation were used to determined equilibrium behavior (thermodynamics). Both global (pivot) and small-step (Biased Gaussian Steps [51] and sidechain rotations) Monte Carlo moves are included to speed up efficient sampling of conformational space. Our small-step “kinetic” Monte Carlo runs differ from our simulated

tempering runs in two ways: (1) global MC moves are turned off and (2) the temperature is held fix. Statistical errors are estimated from independent simulations. Thermodynamic simulations of the isolated RfaH CTD are obtained using 20 independent runs of each 3×10^7 MC cycles, where each cycle is 239 elementary MC moves (the number of turnable torsion angle of the simulated chain). For the full-length RfaH, 10 independent runs of at least 1×10^7 MC cycles, where a cycle is 740 elementary steps.

2.3 Representative experimental structures

As representative structures of the two RfaH folds, we use the X-ray structure of the full-length RfaH with PDB id 2oug [22] and the NMR derived structure of the isolated CTD with PDB id 2lcl [21]. The missing residues in 2oug, including the linker region (101-114) and flexible C-terminal tail (157-162), were added back using a homology modeling tool [52], as described previously [33]. The 2lcl structure of the isolated CTD was truncated to retain the ordered part (residues 113-162).

2.4 Native contact maps

The representative structures were submitted to the SMOG webserver [53] to obtain sets of residue-residue contacts according to the shadow map algorithm [54]. For 2lcl, we obtained a set of 129 native contacts, C^β . For 2oug, retaining only contacts with both i and j within the segment 115-156, we obtained a set of 69 contacts, C^α . The C^α and C^β contact maps are shown in Fig. 1 and used for the simulations of the isolated CTD. Simulations of the full-length RfaH are carried out with the contact map $C^{\alpha_{\text{RfaH}}}$, with 157 contacts. $C^{\alpha_{\text{RfaH}}}$ includes all contacts in C^α and, in addition, all the NTD-CTD inter-domain contacts, i.e., contacts between one residue in the segment 1-100 and another in the segment 115-156.

2.5 Observables

The progress variables, Q_α and Q_β , are the fraction of the native contacts formed in C^α and C^β , respectively (see Fig. 1). A contact between residues i and j is considered formed if $r_{ij} < 1.2r_{ij}^0$, where r_{ij} is the distance between residues i and j (as defined in section 3.1), and r_{ij}^0 is the value r_{ij} in the representative structure. The root-mean-square deviations, RMSD_α and RMSD_β , are taken with respect to two representative structures of the all- α and all- β CTD folds, respectively, and determined over their C_α atoms.

2.6 Correction term for the dual-basin SBM

In our dual-basin SBM, the idea is to include attractions between pairs of residues that contact each other in the all- α fold (C^α) and attractions between pairs of residues that contact each other in the all- β fold (C^β). However, C^α and C^β have 13 common contacts, as shown in Fig. 1, and it would be unreasonable to include both attractions for these contacts. Therefore, we use the convention that, for each of these 13 contacts, only the most favorable of the two contact energies contributes to the total energy. This is achieved by a correction term included in the potential energy function of the dual-basin SBM (see section 3.3), which is given by

$$E_{\text{corr}}(\lambda^\alpha, C^\alpha; \lambda^\beta, C^\beta) = -\epsilon \sum_{ij} \min[\lambda^\alpha g(r_{ij}, r_{ij}^{\alpha,0}), \lambda^\beta g(r_{ij}, r_{ij}^{\beta,0})], \quad (1)$$

where the sum goes over the 13 contacts ij that are present in both C^α and C^β . In this equation, the distance r_{ij} , the function g and the strengths λ^α and λ^β are defined as in sections 3.2 and 3.3. The “native” distances $r_{ij}^{\alpha,0}$ and $r_{ij}^{\beta,0}$ are the values of r_{ij} in the all- α and all- β representative structures, respectively.

3 Results and Discussion

3.1 Hybrid sequence-structure-based simulation approach

An ideal computational model for simulating metamorphic proteins would be capable of capturing the large-scale structural transitions of these proteins and, at the same time, be computationally tractable. However, as mentioned in Introduction, such an approach is out of reach at present. We pursue here a hybrid approach in which an implicit solvent physics-based model is augmented with an SBM. Similar approaches have been tested before [55–57], although not for the system under study here. Specifically, we construct the potential energy function of the hybrid model as a linear combination of the physics-based and the structure-based potentials. Our aim is to pick the relative strength of the SBM as small as possible, while requiring that the hybrid model as a whole exhibits a thermodynamic behavior in agreement with available experimental data.

As our starting point, we use the protein model developed in Ref. [50]. This model is based on an effective all-atom (solvent free) potential energy function, which includes terms for the major driving forces of protein structure formation, including hydrophobic and electrostatic attractions and hydrogen bonding. Parametrization of this model was done by requiring that a set of 17 different amino acid sequences exhibit global free energy minima corresponding to their respective experimentally determined native structures. Interestingly, this “top-down” approach to parameterization also lead to thermodynamic behaviors, such as melting temperatures, that for many of the sequences were in quantitative agreement with experimental data [50].

With this model in mind as a baseline, we formulate a structure-based, or Gō-like [43], potential, which provides an energetic bias towards a single target structure defined by a set of residue-residue contacts, C (the contact map of the structure). We pick this energy term

to have the form

$$E_{\text{SB}}(C) = -\epsilon \sum_{ij \in C} g(r_{ij}, r_{ij}^0), \quad (2)$$

where the sum goes over all contacts ij in C , ϵ is the energy unit of our baseline model [50], r_{ij} is the C_β - C_β distance between residues i and j , and r_{ij}^0 is the value of r_{ij} in the native fold. For any contact ij for which one (or both) positions is a glycine residue, thus lacking a C_β atom, r_{ij} is instead determined using the C_α -atom position. The quantity $g(r_{ij}, r_{ij}^0)$, where

$$g(r, r^0) = e^{-(r-r^0)^2/2\xi^2}, \quad (3)$$

measures the extent to which the native contact ij is correctly formed. The width of the Gaussian function g is controlled by the parameter ξ , which we set to $\xi = 1 \text{ \AA}$. Previous work [55–57] that have similarly combined physics-based and SBMs, have exclusively formulated their SBMs using C_α - C_α distances. Here we primarily use C_β -atoms for quantifying native contact formation because we expect that C_β - C_β distances, through the function $g(r_{ij}, r_{ij}^0)$, to provide a higher specificity towards the native structure.

In the following, we apply Eq. 2 to the two possible folds of the RfaH CTD. This gives two structure-based potentials, $E_{\text{SB}}(C^\alpha)$ and $E_{\text{SB}}(C^\beta)$, where C^α and C^β are the contact maps of the two experimentally determined structures displayed in Fig. 1. In section 3.2, we combine our physics-based model with each term separately to create two hybrid models in which the SBM has a single basin of attraction. In section 3.3, we combine these potentials into a hybrid model with a dual-basin SBM.

3.2 Single-basin SBM

Combining the SBM defined by Eq. 4 with our physics-based model results in a potential energy function of the form,

$$E^{(1)} = E^{(0)} + \lambda E_{\text{SB}}(C), \quad (4)$$

where $E^{(0)}$ is the energy function defined in Ref [50] and λ is the strength of the structure-based term, which should be seen as a free parameter in this approach.

We apply Eq. 4 to the CTD of RfaH, with either $C = C^\alpha$ or C^β (see Fig. 1). Quite generally, we expect that increasing the relative strength of the SBM should increase the “nativeness” of the generated conformational ensemble. In other words, increasing λ should increase Q_α in the case $C = C^\alpha$ and increase Q_β in the case $C = C^\beta$, where Q_α and Q_β are the fraction of native contacts formed with regards to the α -helical hairpin and β -barrel folds, respectively (see Methods). We indeed observe such a trend, as shown in Fig. 2. However, beyond this general trend, we find stark qualitative differences. The term $E_{\text{SB}}(C^\beta)$ has a relatively large effect on the population of the all- β native state. For large enough λ , the melting curves are characterized by a sharp decrease in Q_β with increasing temperature. Moreover, the folding transition occurs at increasingly higher temperatures, as indicated by a shift in the peaks of the heat capacity, C_v . Interestingly, the height of the C_v peaks (C_v^{max}) also increases with λ , suggesting also an increasing cooperativity of the transitions. By contrast, the single-basin SBM applied to the α -helical hairpin fold ($C = C^\alpha$) has a very different effect. The fraction of native contacts, Q_α , does not reach 0.5, even at much lower temperatures. Strikingly, C_v^{max} decreases with increasing λ in this case.

It must be noted that $E^{(1)}$ in Eq. 4 does not preserve the energy scale of the underlying model [50], because $E^{(1)}$ depends on λ . As a result, temperatures and energy quantities can therefore no longer be directly interpreted as real physical units. For example, obtained folding temperatures, T_f , which we define using the C_v peaks, can in general not be compared

Figure 2: **Equilibrium behavior of hybrid models with single-basin SBM.** Shown are results obtained with the energy function in Eq. 4 with (A, C) a β -barrel SBM, $C = C^\beta$, and (B, D) a α -helical hairpin SBM, $C = C^\alpha$, and the strengths $\lambda = 0.20$ – 0.40 . As functions of the temperature: (A) Q_β , (B) Q_α and (C, D) heat capacity C_v/k_B , where $C_v = (\langle E^2 \rangle - \langle E \rangle^2)/k_B T^2$, E is the energy from Eq. 4, and k_B is Boltzmann’s constant. Lines between points in (A) and (B) are drawn to guide the eye. Solid curves in (C) and (D) are obtained using multiple-histogram reweighting [58].

with experimental data. However, comparing the T_f s for different proteins should still inform on their relative stabilities, as long as results are obtained for the same choice of λ .

3.3 Dual-basin SBM

We now combine the two structure-based terms, $E_{SB}(C^\alpha)$ and $E_{SB}(C^\beta)$, into a dual-basin SBM. Following our general approach, the energy function of the resulting hybrid model

becomes

$$E^{(2)} = E^{(0)} + \lambda^\alpha E_{\text{SB}}(C^\alpha) + \lambda^\beta E_{\text{SB}}(C^\beta) - E_{\text{corr}}(\lambda^\alpha, C^\alpha; \lambda^\beta, C^\beta), \quad (5)$$

where strengths of the two structure-based terms, λ^α and λ^β , are selected to be $\lambda^\alpha = \lambda^\beta = 0.30$ based on the results of the previous section. The last term, E_{corr} , is a correction term that is necessary in order to avoid a double counting of common contacts in C_α and C_β (see Fig. 1), which receive contributions from both $E_{\text{SB}}(C^\alpha)$ and $E_{\text{SB}}(C^\beta)$. The role of E_{corr} is to eliminate the weaker of these two interactions (see Methods).

We apply Eq. 5 to the CTD as an isolated fragment and to the full-length RfaH. For the isolated CTD, we find that Q_β increases with decreasing temperature, as shown in Fig. 3. In other words, despite the dual-basin nature of the SBM in our hybrid model, the isolated CTD folds into a stable β -barrel at low T . As a comparison, we show also in Fig. 3A the results for the single-basin SBM, which, interestingly, gives a higher stability compared to the dual-basin SBM. The simulations of full-length RfaH are carried out in the following way. First, the contact map C^α in Eq. 5 is replaced with $C^{\alpha_{\text{RfaH}}}$, which in addition to all contacts in C^α also include NTD-CTD interdomain contacts, as shown in Fig. 3B. Second, for computational reasons, we keep the NTD backbone fixed in its native conformation while the linker and CTD regions are free to move. All sidechains are also left free. From these simulations we find, in contrast to the isolated CTD, that Q_α increases with decreasing temperature. Hence, within the context of the full-length RfaH, the CTD switches into an α -helix rich state.

Figure 3: Equilibrium behavior of hybrid model with dual-basin SBM.

(A) Temperature dependence of the fraction of native contacts (Q_β or Q_α) obtained from simulations of the isolated CTD or full-length RfaH. For comparison, results for the single-basin β -barrel SBM is re-drawn from Fig. 2A. as described below. (B) Structure in cartoon representation (right) and native contact map (left) of full-length RfaH (PDB ID 2oug). Intra-CTD, inter-domain and other native contacts are shown in blue, yellow and gray, respectively. Blue and yellow contacts make up the contact set, $C^{\alpha\text{RfaH}}$. (C) The average α -helix content as a function of sequence position for full-length RfaH, taken at 420 K (NTD region not shown). $Q_\alpha^{(49)}$ is determined as Q_α but taken over a reduced set of 49 contacts that excludes contacts in the segment Val116–Gly121 (orange shade). All results are obtained with $\lambda^\alpha = \lambda^\beta = 0.30$.

The fraction of native contacts, Q_α , does not reach values higher than around 0.5 at the lowest simulated T , however, (see Fig. 3A). To examine the reason for this, we plot the propensity for forming α -helix structure as a function of residue position, as shown in Fig. 3C. The α_4 helix (residues 116—130) is in our model significantly less structured than the α_5 helix (residues 134—156). According to the NMR experiments of Burmann et al. [21], the segment Val116-Gly121 displays chemical shifts that are more in line with a random coil than an α -helix, suggesting that the 6 N-terminal residues of α_4 might be mainly disordered in the solution state. Therefore, we construct an alternative nativeness measure, $Q_\alpha^{(49)}$, which is the same as Q_α but all contacts involving 116-121 are ignored. Indeed, $Q_\alpha^{(49)}$ is consistently higher than Q_α across all temperatures indicating that the α -helical hairpin fold is formed at low T with the exception of the N-terminal part of α_4 , which, in line with experiments, remain largely unstructured.

3.4 Folding energy landscape with a single dominant funnel

Having showed that our hybrid model with a dual-basin SBM captures the basic thermodynamic behavior of the RfaH CTD, i.e., its all- α -to-all- β switch in global free energy minimum upon excision from the rest of the protein, we turn to the folding behavior of this domain. To delineate the impact of the fold switching capability of the RfaH CTD, we compare with results from the hybrid model with a single-basin SBM (applied to the β -barrel contact map C^β). Both these hybrid models fold the CTD into a stable β -barrel, but only the dual-basin SBM gives proper fold switching. We therefore reason that differences in the respective energy landscapes can be attributed to the unique fold switching behavior of the CTD.

We first determine the free energy surface $F(Q_\alpha, Q_\beta)$ close to the folding temperature, T_f , as shown in Fig. 2A. At T_f , one might expect our dual-basin SBM to produce three distinct free energy minima corresponding to the unfolded state, which must be substantially populated

Figure 4: Free energy surfaces at the folding midpoint: single-basin SBM vs dual-basin SBM. Shown are free energy surfaces $F(X_1, X_2) = -\ln P(X_1, X_2)$, with (A, B) $X_1 = Q_\alpha$ and $X_2 = Q_\beta$ or (B, D) X_1 is the total energy and $X_2 = \text{RMSD}_\beta$. The probability distributions $P(X_1, X_2)$ are obtained close to the respective T_f s, which are (nominally) 384 K for the dual-basin SBM and 377 K for the (β -barrel) single-basin SBM.

at T_f , and two ordered states, the all- α fold and all- β folds. However, there are only two major minima in the Q_α - Q_β plane: (1) a low- Q_α , low- Q_β minimum, i.e., the unfolded state; and (2) a high- Q_β , low- Q_α minimum, i.e., the β -barrel state. The absence of a minimum for the α -helical hairpin fold is also clear from the free energy surface $F(E, \text{RMSD}_\beta)$ in Fig. 4B, which show no distinct low-energy states that compete with the native β -barrel fold. At $T < T_f$, i.e., under folding conditions, the population of the unfolded state is substantially decreased as ordered states with lower energies are increasingly favored. As a result, the energy landscape becomes dominated by a single funnel toward the β -barrel fold, as shown by the one-dimensional free energy profiles, $F(Q_\alpha)$ and $F(Q_\beta)$, in Fig. 5.

Figure 5: **Free energy profiles under folding conditions.** Shown is the free energy as function of (A) Q_α and (B) Q_β for hybrid models with either dual-basin (solid curves) or single-basin (dashed curves) SBM, taken at $T \approx 0.96T_f$.

Recent computational studies of the RfaH CTD have found energy landscapes with clear two-funnel character [32,37], in apparent contradiction with our results. These studies relied on molecular dynamics simulations with either implicit or explicit water in combination with advanced sampling and analysis techniques. For example, Bernhardt et al. [37] achieved enhanced sampling using a Hamilton replica-exchange method that couples their physical model to a Gō potential, which seeds the sampling of conformational space. It is possible that these types of techniques are able to detect fine features of the energy landscape that are not apparent in our calculations. However, the lack of a major competing minimum corresponding to the α -helical hairpin fold in the energy landscape of the isolated CTD is consistent with the $[^1\text{H}, ^{15}\text{N}]$ -HSQC NMR data of Burmann et al. [21]. Upon protease-induced cleavage of the NTD-CTD linker, thus releasing the CTD into the solution, they observed a complete disappearance of the spectra from the α -helical hairpin and the emergence of the spectra from the β -barrel fold [21]. Although the α -helical hairpin is clearly encoded in the amino acid sequence of the CTD, these experiment results limit the possible “depths” of this funnel and hence its significance in comparison with the large number of other local minima present in any protein energy landscape.

3.5 The impact of encoding two folds on the unfolded state

While the differences between the free energy surfaces $F(Q_\alpha, Q_\beta)$ at $T \approx T_f$ (see Fig. 3A and B), obtained for the single- and dual-basin SBMs are not large, it is interesting to examine them in detail. We find that, for the single-basin SBM, the unfolded state is a rather narrow minimum, especially in the Q_α direction. For the dual-basin SBM, the unfolded state is broader in the Q_α -direction, indicating larger fluctuations. Many of the native contacts in C^α are local, such as α -helical $i, i+4$ -contacts (see Fig. 1). The main impact of the encoding the α -helical hairpin within our dual-basin SBM, which nonetheless spontaneously fold into the β -barrel fold as we have shown, is to produce residual α -helix structure in the unfolded state. Indeed, we quantify the α -helix content in the unfolded state, as defined by $Q_\beta < 0.5$, for both models and find that it is significantly higher for the dual-basin SBM (20%) than for the single-basin SBM (7.8%).

A similar picture is obtained from the free energy profiles $F(Q_\alpha)$ and $F(Q_\beta)$ (see Fig. 5) at lower temperatures, $T \approx 0.96T_f$. We note, in particular, that the shape of $F(Q_\beta)$ is remarkably similar between the two models at large Q_β . This similarity likely reflects that contacts unique to the α -helical fold are either highly unstable or otherwise impossible to form due to topological constraints, once folding has progressed towards the β -barrel beyond a certain point (roughly $Q_\beta \approx 0.5$).

3.6 Is there an activation barrier to the all- α -to-all- β fold switch?

Although we find in our hybrid approach, with a dual-basin SBM, a dominant funnel to the β -barrel fold, conformations sufficiently close to the α -helical hairpin will, by construction, be energetically biased towards this attractor. Therefore, a conformation prepared in this all- α state might still need to overcome an activation barrier before it can proceed downhill the energy landscape towards the β -barrel native state. Such a situation might occur *in vivo*

Figure 6: **Fold switching and folding simulations.** Shown is the evolution of the α -helix content, $\langle \alpha \rangle$, and the fraction of native contacts, $\langle Q_\beta \rangle$, in small-step Monte Carlo simulations started in either the α -helical hairpin fold (purple circles) or in an open chain conformation (green triangles), where the average $\langle \rangle$ is taken over 100 independent trajectories. Simulations are carried at $T = 365$ K. Error bars show standard errors.

when RfaH binds to RNA polymerase, which triggers the CTD to detach from the NTD [24].

In order to explore the potential barrier of fold switching in the all- α -to-all- β direction, we carry out small-step “kinetic” MC simulations. We consider two different starting points: (1) the α -helical hairpin and (2) an extended, open conformation. The temperature is held fixed at $T = 365$ Kelvin, for which the β -barrel is the thermodynamically dominant state (see Fig. 3). Hence, both starting points should eventually transform into the β -barrel fold, although the timescale of the transformation could be different. These two sets of simulations started from point (1) and (2) thus probe respectively the fold switching and folding of the RfaH CTD.

In Fig. 6, we show the relaxation of the α -helix content, $\langle \alpha \rangle$, and the fraction of native contacts, $\langle Q_\beta \rangle$, where $\langle \rangle$ indicates an average taken over 100 independent runs, towards their respective equilibrium values at this T , i.e., $\langle \alpha \rangle_{\text{equil}} = 0.08$ and $\langle Q_\beta \rangle_{\text{equil}} = 0.70$. In terms of $\langle Q_\beta \rangle$, the folding simulations, which start at $\langle Q_\beta \rangle = 0$, rapidly “overtake” the fold switching simulations. This behavior indicates some degree of kinetic trapping early in the

Figure 7: **Example of fold switching trajectory.** Shown is the evolution of the root-mean-square deviation determined with respect to the α -helix fold (RMSD_α) or the β -barrel fold (RMSD_β) for one of the 100 runs in Fig. 6.

fold switching process. However, the associated barrier cannot be large because the two sets of simulations become statistical indistinguishable at around 10^6 MC cycles, when $\langle Q_\beta \rangle$ is still far from its equilibrium end point. A convergence of the two sets of simulations occurs also in terms of $\langle \alpha \rangle$ at roughly the same point. Interestingly, and perhaps surprisingly, the folding simulations exhibit an initial increase in α -helix structure content to a maximum of $\langle \alpha \rangle \approx 0.25$, and thereafter a much slower decrease following closely the trend of the fold switching simulations. This gradual decrease comes from the conversion of more and more of the trajectories into the β -barrel fold. Taken together, these simulations suggest a fold switching process from the α -helical hairpin state (N_α) to the β -barrel state (N_β) proceeding as $N_\alpha \rightarrow U \rightarrow N_\beta$, where U is the unfolded state of the CTD.

This basic scheme is confirmed when examining in more structural detail individual fold switching events. Figure 7 shows RMSD_α and RMSD_β as functions of MC time for a typical fold switching trajectory, where RMSD_α and RMSD_β are the root-mean-square deviations taken with respect to the representative all- α and all- β CTD structures, respectively. After only a few MC cycles, RMSD_α increases rapidly, and the chain settles into an intermediate

state with large conformational fluctuations. Residual α -helical structure are found upon inspection of conformations in this state, as shown in Fig. 7. Eventually, the chain switches abruptly to an all- β state, characterized by low RMSD_β values and much smaller chain fluctuations. Overall, we find that of all the trajectories that switch folds within the simulation time (47 out of 100), we find that none proceed directly from the all- α to the all- β state but instead proceed via an intermediate state. Because of the presence of residual α -helix structure, we identify this intermediate state with the unfolded state.

4 Conclusions

Metamorphic proteins are often discussed in terms of an energy landscape with two separate and “coexisting” funnels in order to rationalize their ability to adopt two different folds. However, most naturally occurring metamorphic proteins adopt a unique fold under a given constant local environment and switch to a different fold only upon a change in the environment. To examine the energy landscape of the CTD of RfaH, we developed and tested a hybrid all-atom model that combines a physics-based model with a dual-basin structure-based potential (dual-basin SBM). We showed that this model captures the required change in global free energy minimum upon the excision of the CTD from RfaH. Applying this model to the isolated CTD, we found an energy landscape that is characterized by a single dominant funnel towards the β -barrel fold, with no sign of a second funnel towards the α -helical hairpin fold. Our model thus suggests that a multifunneled energy landscape cannot be assumed for metamorphic proteins. Further, by comparing with another hybrid model (with a single-basin SBM) that is similar but incapable of fold switching, we found that all major differences could be associated with the unfolded state. Specifically, this comparison indicated a relatively high content of α -helix structure in the unfolded state. Biophysical characterizations, e.g., using circular dichroism, of the CTDs of RfaH and other members

of the general NusG/Spt5 family of transcription factors under weakly unfolding conditions, would provide an interesting experimental test of the computational results of this work.

Acknowledgements

This work was supported by grants from Memorial University and Natural Sciences and Engineering Research Council of Canada (NSERC), and by computational resources provided by Compute Canada.

Conflict of Interest

None to declare.

References

- [1] P. N. Bryan and J. Orban. Proteins that switch folds. *Curr Opin Struct Biol*, 20:482–488, 2010.
- [2] A. F. Dishman and B. F. Volkman. Unfolding the mysteries of protein metamorphosis. *ACS Chem Biol*, 13:1438–1446, 2018.
- [3] M. Lella and R. Mahalakshmi. Metamorphic proteins: emergence of dual protein folds from one primary sequence. *Biochemistry*, 56:2971–2984, 2017.
- [4] Y. Liu and D. Eisenberg. 3D domain swapping: as domains continue to swap. *Protein Sci*, 11:1285–1299, 2002.

- [5] S. Hayward. Structural principles governing domain motions in proteins. *Proteins*, 36:425–435, 1999.
- [6] J. A. Vila. Metamorphic proteins in light of Anfinsen’s dogma. *J Phys Chem Lett*, 11:4998–4999, 2020.
- [7] C. B. Anfinsen. Principles that govern the folding of protein chains. *Science*, 181:223–230, 1973.
- [8] R. L. Tuinstra, F. C. Peterson, S. Kutlesa, E. S. Elgin, M. A. Kron, and B. F. Volkman. Interconversion between two unrelated protein folds in the lymphotactin native state. *Proc Natl Acad Sci USA*, 105:5057–5062, 2008.
- [9] D. R. Littler, S. J. Harrop, W. D. Fairlie, L. J. Brown, G. J. Pankhurst, S. Pankhurst, M. Z. DeMaere, T. J. Campbell, A. R. Bauskin, R. Tonini, M. Mazzanti, S. N. Breit, and P. M. Curmi. The intracellular chloride ion channel protein CLIC1 undergoes a redox-controlled structural transition. *J Biol Chem*, 279:9298–9305, 2004.
- [10] Y. G. Chang, S. E. Cohen, C. Phong, W. K. Myers, Y. I. Kim, R. Tseng, J. Lin, L. Zhang, J. S. Boyd, Y. Lee, S. Kang, D. Lee, S. Li, R. D. Britt, M. J. Rust, S. S. Golden, and A. LiWang. A protein fold switch joins the circadian oscillator to clock output in cyanobacteria. *Science*, 349:324–328, 2015.
- [11] L. L. Porter and L. L. Looger. Extant fold-switching proteins are widespread. *Proc Natl Acad Sci USA*, 115:5968–5973, 2018.
- [12] A. Zarrine-Afsar, S. M. Larson, and A. R. Davidson. The family feud: do proteins with similar structures fold via the same pathway? *Curr Opin Struct Biol*, 15(1):42–49, Feb 2005.
- [13] A. A. Nickson and J. Clarke. What lessons can be learned from studying the folding of homologous proteins? *Methods*, 52(1):38–50, Sep 2010.

- [14] D. S. Riddle, V. P. Grantcharova, J. V. Santiago, E. Alm, I. Ruczinski, and D. Baker. Experiment and theory highlight role of native state topology in SH3 folding. *Nat Struct Biol*, 6(11):1016–1024, Nov 1999.
- [15] F. Chiti, N. Taddei, P. M. White, M. Bucciantini, F. Magherini, M. Stefani, and C. M. Dobson. Mutational analysis of acylphosphatase suggests the importance of topology and contact order in protein folding. *Nat Struct Mol Biol*, 6(11):1005–1009, Nov 1999.
- [16] J. C. Martinez and L. Serrano. The folding transition state between SH3 domains is conformationally restricted and evolutionarily conserved. *Nat Struct Biol*, 6(11):1010–1016, Nov 1999.
- [17] D. Trotter and S. Wallin. Effects of topology and sequence in protein folding linked via conformational fluctuations. *Biophys J*, 118:1370–1380, 2020.
- [18] S. S. Cho, Y. Levy, and P. G. Wolynes. Quantitative criteria for native energetic heterogeneity influences in the prediction of protein folding kinetics. *Proc Natl Acad Sci USA*, 106(2):434–439, Jan 2009.
- [19] A. Kluber, T. A. Burt, and C. Clementi. Size and topology modulate the effects of frustration in protein folding. *Proc Natl Acad Sci USA*, 115(37):9234–9239, 09 2018.
- [20] S. Mishra, L. L. Looger, and L. L. Porter. Inaccurate secondary structure predictions often indicate protein fold switching. *Protein Sci*, 28:1487–1493, 2019.
- [21] B. M. Burmann, S. H. Knauer, A. Sevostyanova, K. Schweimer, R. A. Mooney, R. Landick, I. Artsimovitch, and P. Rösch. An α helix to β barrel domain switch transforms the transcription factor RfaH into a translation factor. *Cell*, 150:291–303, 2012.

- [22] G. A. Belogurov, M. N. Vassilyeva, V. Svetlov, S. Klyuyev, N. V. Grishin, D. G. Vassilyev, and I. Artsimovitch. Structural basis for converting a general transcription factor into an operon-specific virulence regulator. *Mol Cell*, 26:117–129, 2007.
- [23] D. Shi, D. Svetlov, R. Abagyan, and I. Artsimovitch. Flipping states: a few key residues decide the winning conformation of the only universally conserved transcription factor. *Nucleic Acids Res*, 45:8835–8843, 2017.
- [24] P. K. Zuber, K. Schweimer, P. Rösch, I. Artsimovitch, and S. H. Knauer. Reversible fold-switching controls the functional cycle of the antitermination factor RfaH. *Nat Commun*, 10:702, 2019.
- [25] P. Galaz-Davison, J. A. Molina, S. Silletti, E. A. Komives, S. H. Knauer, I. Artsimovitch, and C. A. Ramírez-Sarmiento. Differential local stability governs the metamorphic fold switch of bacterial virulence factor RfaH. *Biophys J*, 118:96–104, 2020.
- [26] S. Li, B. Xiong, Y. Xu, T. Lu, X. Luo, C. Luo, J. Shen, K. Chen, M. Zheng, and H. Jiang. Mechanism of the all- α to all- β conformational transition of RfaH-CTD: Molecular dynamics simulation and Markov State model. *J Chem Theory Comput*, 10:2255–2264, 2014.
- [27] C. A. Ramirez-Sarmiento, J. K. Noel, S. L. Valenzuela, and I. Artsimovitch. Interdomain contacts control native state switching of RfaH on a dual-funneled landscape. *PLOS Comput Biol*, 11:e1004379, 2015.
- [28] J. B. GC, Y. R. Bhandari, B. S. Gerstman, and P. P. Chapagain. Molecular dynamics investigations of the α -helix to β -barrel conformational transformation in the RfaH transcription factor. *J Phys Chem B*, 118:5101–5108, 2014.
- [29] J. B. GC, B. S. Gerstman, and P. P. Chapagain. The role of the interdomain interactions on RfaH dynamics and conformational transformation. *J Phys Chem B*, 119:12750–12759, 2015.

- [30] L. Xiong and Z. Liu. Molecular dynamics study on folding and allostery in RfaH. *Proteins*, 83:1582–1592, 2015.
- [31] S. Xun, F. Jiang, and Y. D. Wu. Intrinsically disordered regions stabilize the helical form of the C-terminal domain of RfaH: A molecular dynamics study. *Bioorg Med Chem*, 24:4970–4977, 2016.
- [32] J. A. Joseph, D. Chakraborty, and D. J. Wales. Energy landscape for fold-switching in regulatory protein RfaH. *J Chem Theory Comput*, 15:731–742, 2019.
- [33] B. Seifi and S. Wallin. Structural fluctuations and mechanical stabilities of the metamorphic protein RfaH. *Proteins (in press)*.
- [34] J. Y. Kang, R. A. Mooney, Y. Nediaklov, J. Saba, T. V. Mishanina, I. Artsimovitch, R. Landick, and S. A. Darst. Structural basis for transcript elongation control by NusG family universal regulators. *Cell*, 173:1650–1662, 2018.
- [35] P. K. Zuber, I. Artsimovitch, M. NandyMazumdar, Z. Liu, Y. Nediaklov, K. Schweimer, P. Rösch, and S. H. Knauer. The universally-conserved transcription factor RfaH is recruited to a hairpin structure of the non-template DNA strand. *Elife*, 7, 2018.
- [36] A. Roy, A. Perez, K. A. Dill, and J. L. Maccallum. Computing the relative stabilities and the per-residue components in protein conformational changes. *Structure*, 22:168–175, 2014.
- [37] N. A. Bernhardt and U. H. E. Hansmann. Multifunnel landscape of the fold-switching protein RfaH-CTD. *J Phys Chem B*, 122:1600–1607, 2018.
- [38] J. R. Allison, M. Bergeler, N. Hansen, and W. F. van Gunsteren. Current computer modeling cannot explain why two highly similar sequences fold into different structures. *Biochemistry*, 50:10965–10973, 2011.

- [39] C. Holzgräfe and S. Wallin. Smooth functional transition along a mutational pathway with an abrupt protein fold switch. *Biophys J*, 107:1217–1225, 2014.
- [40] C. Holzgräfe and S. Wallin. Local versus global fold switching in protein evolution: insight from a three-letter continuous model. *Phys Biol*, 12:026002, 2015.
- [41] A. Aina and S. Wallin. Multisequence algorithm for coarse-grained biomolecular simulations: Exploring the sequence-structure relationship of proteins. *J Chem Phys*, 147:095102, 2017.
- [42] R. D. Hills and C. L. Brooks. Insights from coarse-grained Gō models for protein folding and dynamics. *Int J Mol Sci*, 10:889–905, 2009.
- [43] N. Gō and H. Taketomi. Respective roles of short- and long-ranged interactions in protein folding. *Proc Natl Acad Sci USA*, 75:559–563, 1978.
- [44] S. Wallin and H. S. Chan. Conformational entropic barriers in topology-dependent protein folding: perspectives from a simple native-centric polymer model. *J Phys Condens Matter*, 18:S307, 2006.
- [45] L. L. Chavez, J. N. Onuchic, and C. Clementi. Quantifying the roughness on the free energy landscape: entropic bottlenecks and protein folding rates. *J Am Chem Soc*, 126(27):8426–8432, Jul 2004.
- [46] J. S. Yang, S. Wallin, and E. I. Shakhnovich. Universality and diversity of folding mechanics for three-helix bundle proteins. *Proc Natl Acad Sci USA*, 105:895–900, 2008.
- [47] L. Sutto and C. Camilloni. From A to B: a ride in the free energy surfaces of protein G domains suggests how new folds arise. *J Chem Phys*, 136:185101, 2012.
- [48] M. Kouza and U. H. Hansmann. Folding simulations of the A and B domains of protein G. *J Phys Chem B*, 116:6645–6653, 2012.

- [49] A. Irback and S. Mohanty. PROFASI: A Monte Carlo simulation package for protein folding and aggregation. *J Comput Chem*, 27(13):1548–1555, Oct 2006.
- [50] A. Irback, S. Mitternacht, and S. Mohanty. An effective all-atom potential for proteins. *PMC Biophysics*, 2:2, 2009.
- [51] G. Favrin, A. Irback, and F. Sjunnesson. Monte Carlo update for chain molecules: biased Gaussian steps in torsional space. *J Chem Phys*, 114:8154–8158, 2001.
- [52] A. Sali, E. Shakhnovich, and M. Karplus. How does a protein fold? *Nature*, 369:248–251, 1994.
- [53] J. K. Noel, M. Levi, M. Raghunathan, H. Lammert, R. L. Hayes, J. N. Onuchic, and P. C. Whitford. SMOG 2: A versatile software package for generating structure-based models. *PLOS Comput Biol*, 12:e1004794, 2016.
- [54] J. K Noel, P. C. Whitford, and J. N. Onuchic. The shadow map: a general contact definition for capturing the dynamics of biomolecular folding and function. *J Phys Chem B*, 116:8692–8702, 2012.
- [55] J. H. Meinke and U. H. E. Hansmann. Protein simulations combining an all-atom force field with a Go term. *J Phys Condens Matter*, 19:285215, 2007.
- [56] T. Sikosek, H. Krobath, and H. S. Chan. Theoretical insights into the biophysics of protein bi-stability and evolutionary switches. *PLOS Comput Biol*, 12:e1004960, 2016.
- [57] Y. Wang, P. Tian, W. Boomsma, and K. Lindorff-Larsen. Monte Carlo sampling of protein folding by combining an all-atom physics-based model with a native state bias. *J Phys Chem B*, 122:11174–11185, 2018.
- [58] A. M. Ferrenberg and R. H. Swendsen. Optimized Monte Carlo data analysis. *Phys Rev Lett*, 63(12):1195–1198, 1989.