PONTIFICIA UNIVERSIDAD CATOLICA DE CHILE

SCHOOL OF ENGINEERING

DESIGN OF A METAMORPHIC PROTEIN FROM FIRST PRINCIPLES

PABLO ANTONIO GALAZ DAVISON

Thesis submitted to the Office of Graduate Studies in partial fulfillment of the requirements for the Degree of Doctor in Engineering Sciences

Advisor:
CÉSAR A. RAMÍREZ SARMIENTO

Santiago de Chile, May 2023

PONTIFICIA UNIVERSIDAD CATOLICA DE CHILE
SCHOOL OF ENGINEERING

DESIGN OF A METAMORPHIC PROTEIN FROM FIRST PRINCIPLES

PABLO ANTONIO GALAZ DAVISON

Members of the Committee:

CÉSAR RAMÍREZ	Cesar Ramirez
PEDRO SAA	Pedro Saa Higuera
ANGÉLICA FIERRO	
ARIELA VERGARA	
MATÍAS MACHADO	
GUSTAVO LAGOS	Gustavo Lagos

Thesis submitted to the Office of Graduate Studies in partial fulfillment of the requirements for the Degree of Doctor in Engineering Sciences

Santiago de Chile, May 2023

To my Friends and Parents, who supported me throughout this endeavor

ACKNOWLEDGEMENTS

I would like to thank first my advisor, César A. Ramírez Sarmiento, who aided in my professional development from my undergraduate to my PhD. My parents, Margarita Davison and Juan Galaz, and friends who have continuously supported me through difficult times, especially Beatriz Salas and Cecilia Muster, and my boyfriend Yerko Suazo.

I sincerely thank the support provided by my lab mates and lab friends, particularly Javiera Reyes, Maira Rivera, and Alejo Molina.

Also, I thank the support provided by Agencia Nacional de Investigación y Desarrollo (ANID) through the doctoral scholarship ANID-PFCHA 21181705, and to Pontificia Universidad Católica for accepting me as a doctoral candidate.

FUNDING BODIES

This thesis was supported by ANID through Fondo de Desarrollo Científico y Tecnológico (FONDECYT 11140601 and 1201684), a joint grant with Professors Elizabeth A. Komives and Majid Ghassemian at University of California San Diego through Programa de Cooperación Internacional (ANID PCI REDI170624) and ANID Millennium Science Initiative Program (ICN17_022). This thesis was also partially supported by the NVIDIA GPU Grant Program, Pontificia Universidad Católica de Chile through its Office of Research Grant (VRI Puente P1810), and by the supercomputing infrastructure of the NLHPC (ECM-02).

An internship abroad, in the laboratory of Professor Marcos Sotomayor at The Ohio State University, was funded by the Promoting Research Opportunities for Latin American Biochemists (PROLAB) program from the Pan-American Association for Biochemistry and Molecular Biology, the International Union of Biochemistry and Molecular Biology, and the American Society for Biochemistry and Molecular Biology. A second internship abroad, in the laboratory of Professor Diego U. Ferreiro at Universidad de Buenos Aires, was funded the Santander through the Graduate Mobility Program Scholarship from Red de Macro Universidades de América Latina y el Caribe, México. A third internship in the Biomolecular and Proteomics Mass Spectrometry Facility at the University of California San Diego was funded by ANID PCI REDI170624.

LIST OF PUBLICATIONS

This thesis contains research already published in literature in the **Sections 2, 3 and 4**.

Chapter 2 is adapted from the following article:

Galaz-Davison, P., Molina, J. A., Silleti, S., Komives, E. A., Knauer, S. H., Artsimovitch, I., Ramírez-Sarmiento, C. A. (2020) Differential local stability governs the metamorphic fold switch of bacterial virulence factor RfaH. *Biophysical Journal*, *118*(1), 96-104. DOI: 10.1016/j.bpj.2019.11.014.

This article was recipient of the 2020 Paper of the Year award from *Biophysical Journal*.

Chapter 3 is adapted from the following article:

Galaz-Davison, P., Román, E. A., Ramírez-Sarmiento, C. A. (2021) The N-terminal domain of RfaH plays an active role in protein fold-switching. *PLoS Computational Biology*, *17*(9), e1008882. DOI: 10.1371/journal.pcbi.1008882.

Chapter 4 is adapted from the following article:

Galaz-Davison, P. A., Ferreiro, D. U., Ramírez-Sarmiento, C. A. (2022) Coevolution-derived native and non-native contacts determine the emergence of a novel fold in a universally conserved family of transcription factors. *Protein Science*, *31*(6): e4337. DOI: 10.1002/pro.4337.

During the development of this thesis, five additional articles were also published:

Fecker, T., Galaz-Davison, P., Engelberger, F., Narui, Y., Sotomayor, M., Parra, L. P., & Ramírez-Sarmiento, C. A. (2018). Active site flexibility as a hallmark for efficient PET degradation by *I. sakaiensis* PETase. *Biophysical Journal*, *114*(6), 1302-1312. DOI: 10.1016/j.bpj.2018.02.005.

In this co-first author article, my esteemed colleague Tobias Fecker and I explored the experimental and computational dynamics and binding of the PET hydrolase from a plastic-eating bacterium. On December 2022, this article was recognized as *Biophysical Journal* most read paper.

Engelberger, F., Galaz-Davison, P., Bravo, G., Rivera, M., & Ramírez-Sarmiento, C. A. (2021). Developing and implementing cloud-based tutorials that combine bioinformatics software, interactive coding, and visualization exercises for distance learning on structural bioinformatics. *Journal of Chemical Education*, *98*(5), 1801-1807. DOI: 10.1021/acs.jchemed.1c00022.

This article is the publication of the tutorials of a computational biology course that my advisor, César Ramírez-Sarmiento, and I developed for Pontificia Universidad Católica de Chile, and my esteemed colleague Felipe Engelberger greatly improved and ported into Google Colab.

Espinosa, R., Gutiérrez, K., Ríos, J., Ormeño, F., Yantén, L., Galaz-Davison, P., Ramírez-Sarmiento C. A., Parra, V., Albornoz, A., Alfaro, I. E., Burgos, P. V., Morselli, E., Criollo, A., & Budini, M. (2022). Palmitic and stearic acids inhibit chaperone-mediated autophagy (CMA) in POMC-like neurons in vitro. *Cells*, *11*(6), 920. DOI: 10.3390/cells11060920.

In this article alongside with my advisor, César A. Ramírez-Sarmiento, we aided in the analysis and interpretation of mass spectrometry data collected by our colleague Mauricio Budini.

Rivera, M., Galaz-Davison, P., Retamal-Farfán, I., Komives, E. A., & Ramírez-Sarmiento, C. A. (2022). Dimer dissociation is a key energetic event in the fold-switch pathway of KaiB. *Biophysical Journal*, *121*(6), 943-955. DOI: 10.1016/j.bpj.2022.02.012.

In this article I collaborated with my friend and colleague Maira Rivera in estimating the fold-switching free energy of the metamorphic protein KaiB, using the methods described in Chapter 2.

Molina, J. A., Galaz-Davison, P., Komives, E. A., Artsimovitch, I., & Ramírez-Sarmiento, C. A. (2022). Allosteric couplings upon binding of RfaH to transcription elongation complexes. *Nucleic Acids Research*, *50*(11), 6384-6397. DOI: 10.1093/nar/gkac453.

In this article I collaborated with my friend and colleague J. Alejandro Molina in deconvoluting hydrogen-deuterium exchange data through computational modeling.

INDEX

TABLES INDEX

Page

FIGURE INDEX

RESUMEN

La mayoría de las secuencias de proteínas naturales se pliegan en una estructura tridimensional definida, cuya química y dinámica regula las reacciones y procesos de la materia viva. No obstante, se estima que entre 0,5 y 4% de las proteínas con estructuras resueltas experimentalmente tienen algún grado de comportamiento metamórfico. Esta es una característica de algunas proteínas que han evolucionado para tener más de una estructura, cambiando reversiblemente entre ellas en respuesta a cambios en su entorno químico.

El ejemplo más notable de una proteína metamórfica es RfaH. Esta proteína de dos dominios regula la expresión de factores de virulencia en *Enterobacteriaceae*. En su estado basal, sus dominios N-terminal (NTD) y C-terminal (CTD) interactúan estrechamente. Los operones de factores de virulencia contienen una secuencia de ADN llamada *ops*, que detiene la transcripción de la ARN polimerasa (RNAP) y sirve como una señal química para reclutar RfaH, uniéndose a RNAP al disociar sus dominios. Los dominios RfaH ahora separados se comportan de manera diferente: el NTD se une a RNAP para aumentar la procesividad de la transcripción, mientras que el CTD cambia de una horquilla α-helicoidal a un barril β, estructuralmente similar al observado en el parálogo NusG, mediante el cual RfaH recluta al ribosoma para acoplar la transcripción y traducción de genes que, de otro modo, serían pobremente expresados.

Descifrar las señales fisicoquímicas que permiten el cambio estructural de RfaH es crucial para comprender su mecanismo molecular. En esta tesis, mostramos que el CTD metamórfico alberga regiones locales de estabilidad diferencial hacia cada estructura, con la punta de la horquilla α-helicoidal estabilizando el estado basal a través de contactos interdominio. El NTD y el CTD β-plegado son menos estables que RfaH α-plegado y pueblan un intermediario β-plegado estabilizado por el NTD antes de replegarse en el estado basal. Por último, señales coevolutivas nativas y no nativas, derivadas del análisis de miles de secuencias de RfaH, son suficientes para codificar ambas estructuras, encontrando contactos interdominio e intrahelicoidales que estabilizan el CTD α-plegado y muestran evidencia directa de esta estructura secundaria en RfaH.

Estos resultados son relevantes para comprender el mecanismo de cambio de pliegue de RfaH y para postular los principios que rigen a las proteínas metamórficas.

Palabras clave: Plegamiento de proteínas; proteína metamórfica; RfaH; factor de virulencia.

ABSTRACT

Most of the natural protein sequences fold into a well-defined three-dimensional structure, whose motions and chemical cues regulate the reactions and processes of living matter. Nevertheless, it is estimated that between 0.5 and 4% of proteins with experimentally solved structures have a degree of metamorphic behavior. This is a striking feature of a few proteins that have evolved to bear more than one structure, reversibly switching between them in response to changes in their chemical environment.

The most remarkable example of a metamorphic protein is RfaH. This two-domain protein oversees the expression of virulence factors in *Enterobacteriaceae*. In its unique ground state, its N-terminal (NTD) and C-terminal (CTD) domains are closely interacting. Virulence factor operons contain a DNA sequence named *ops*, which pauses RNA polymerase (RNAP) from transcribing and that serves as a chemical cue for recruiting RfaH, binding to RNAP by dissociating its domains. The now separated RfaH domains behave differently: the NTD binds to RNAP to increase transcription processivity while the CTD fold-switches from an α-helical hairpin into a β-barrel, which is structurally similar to the one present in the paralog NusG, enabling RfaH to recruit the ribosome to couple transcription and translation of otherwise poorly expressed genes.

Deciphering the physicochemical cues that allow RfaH fold-switching is paramount to understand its molecular mechanism. In this thesis, we show that the metamorphic CTD harbors local regions of differential stability towards each fold, with the tip of the α-helical hairpin stabilizing the ground state through interdomain contacts. The NTD and the β-folded CTD are less stable than the α-folded RfaH and populate an NTD-stabilized β-folded intermediate before fold-switching back to the ground state. Lastly, both native and non-native coevolutionary signals derived from analyzing thousands of RfaH sequences are sufficient to encode both folds, with interdomain and intrahelical contacts stabilizing the α-folded CTD and showing direct evidence of this secondary structure in RfaH.

Altogether, these results are an important step in understanding the fold-switching mechanism of RfaH and in postulating the principles that govern metamorphic proteins.

Keywords: Protein folding; metamorphic protein; RfaH; virulence factor.

1. GENERAL INTRODUCTION

The structure and dynamics of proteins are encoded into their chemical composition, normally referred to as their amino acid sequence. These dynamics vary widely depending on the environment they face; salt and proton concentration, solvent polarity, water and lipid interfaces dictate whether stabilizing contacts will cooperatively form and allow protein folding and stabilization (Dill, 1990). Therefore, the native state of a protein is understood as the biologically-active conformation that a given amino acid sequence takes under physiological conditions (Anfinsen, 1973), which for most studied proteins turns out to be a well-defined and unique three-dimensional structure (Kendrew et al., 1958, 1960). Due to its complexity, it is not evident which structure will arise for a given sequence under such physiological conditions, a limitation commonly framed as the *protein folding problem* (Dill & MacCallum, 2012).

Intriguingly, some proteins seem to defy the uniqueness between sequence and structure, as their sequences encode for a varying number of stable conformations necessary for their biological role (Lella & Mahalakshmi, 2017). For example, domain-swapping proteins can fold into a given monomeric structure, yet partial or complete unfolding under increasing protein concentrations results in dimerization via the exchange of equivalent structural regions between adjacent subunits, a major topological change of domain reentrancy giving rise to an intertwined dimer (Bandukwala et al., 2011; Y. Chen et al., 2015). In an even more dramatic example, metamorphic proteins fold into a given structure in isolation, but it dramatically changes its topology after binding to a partner (Burmann et al., 2012; Knauer, Artsimovitch, & Rösch, 2012; Tseng et al., 2017). Both examples suggest that fold-switching, i.e., reversibly changing between two or more energetically favorable structures, arises in nature as an emergent property of molecular systems.

By carefully examining the known metamorphic proteins and tracing the evolution of the functional processes in which they are involved, it can be stated that these proteins have not only co-evolved alongside their binding partners, but their fold-switching behavior became an integral part of their biological function. In this regard, one of the most prominent examples corresponds to RfaH, a two-domain transcription factor that regulates the expression of pathogeny-related genes in Gram-negative bacteria (Cimini, De Rosa, Carlino, Ruggiero, & Schiraldi, 2013; A. Mitra et al., 2013; Sevostyanova, Belogurov, Mooney, Landick, & Artsimovitch, 2011) by repacking ~30% of its residues, all of them located in the C-terminal domain (CTD), into a different fold (Burmann et al., 2012).

In its ground state, both RfaH domains are interacting and hiding an RNA polymerase (RNAP) binding site located at the N-terminal domain (NTD)-side of the interface (Figure 1A) (Belogurov et al., 2007). When a specific 12-nucleotide long DNA signal found in virulence operons – known as *operon polarity suppressor* (*ops*) – pauses RNAP transcription and becomes exposed to the surface, RfaH is recruited to this complex and its NTD dissociates from its CTD, which is folded as an α-helical hairpin (αCTD), to enable binding to RNAP. While the NTD binds to RNAP to enhance transcriptional processivity, the now dissociated CTD fold-switches into a β-barrel (βCTD) structurally similar to that of its paralog NusG (Figure 1B) (Burmann et al., 2012), a non-metamorphic protein family from which RfaH diverged. This fold-switch enables RfaH CTD to physically tether trailing ribosomes by binding to the protein S10, thus coupling transcription to translation (Burmann et al., 2012).

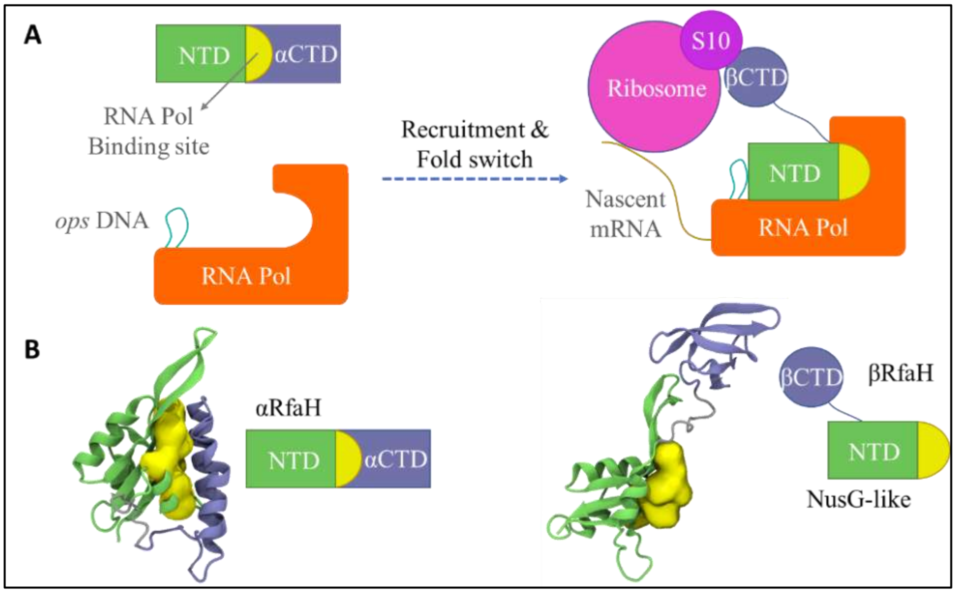


Figure 1: RfaH structural transformation.

(*A*) Overview of RfaH recruitment to the RNAP and fold switch into the NusG-like state upon recognition of the *ops* DNA mounted onto the transcription elongation complex (TEC). (*B*) Structural features of RfaH. Three-dimensional structures in cartoon representation for αRfaH and βRfaH were obtained from the solved crystal structure of isolated full-length RfaH (PDB ID 5OND) and the cryoEM structure of the RfaH-bound TEC (PDB ID 6C6S), respectively.

Multidisciplinary approaches report that the αCTD is highly unstable when not interfacing the NTD, favoring the βCTD structure instead (Burmann et al., 2012; Gc, Gerstman, & Chapagain, 2015; Ramírez-Sarmiento, Noel, Valenzuela, & Artsimovitch, 2015; Xun, Jiang, & Wu, 2016; Zuber, Schweimer, Rösch, Artsimovitch, & Knauer, 2019). Furthermore, although both RfaH and its paralog NusG are members of the only universally conserved family of transcription factors NusG/Spt5 (B. Wang & Artsimovitch, 2021), the CTD of NusG is strictly folded in the β-state, bears little and transient interdomain interactions and does not require a DNA sequence for activation, thus participating in the transcription of almost all genes in bacteria (Belogurov et al., 2007; Mooney, Schweimer, Rösch, Gottesman, & Landick, 2009; Zuber et al., 2018). Thus, the metamorphic behavior of RfaH constitutes a new layer of transcription regulation, relying on the formation of abundant interdomain contacts that are mutually exclusive with a β-fold to impede its spurious binding to RNAP (Burmann et al., 2012; Ramírez-Sarmiento et al., 2015).

Metamorphic proteins such as RfaH double down the protein folding problem: a single sequence encodes the information of two dissimilar structures. A recent bioinformatics algorithm, based on the identification of incorrect secondary structure predictions and independent folding cooperativity, rationalized the structural and sequence determinants to identify fold-switching proteins and estimated that 0.5-4% of proteins with known structure display some degree of metamorphic features (Porter & Looger, 2018). On the other hand, rational design of metamorphic proteins is in its very infancy, as attempts to develop structural multiplicity have been mostly guided by other natural sequences (Alexander, He, Chen, Orban, & Bryan, 2009; Ambroggio & Kuhlman, 2006) or disruption of native contacts, i.e. tertiary contacts between two residues in the native state of a given protein (Campos et al., 2019; Cordes, Burton, Walsh, McKnight, & Sauer, 2000; Dishman & Volkman, 2018).

For RfaH, a differential stabilization between conflicting folds is likely to explain its metamorphic behavior. A recent mathematical model, which evaluated the effect of the residue-residue interactions between α-helices forming a hairpin and β-strands forming a barrel, concluded that differential stabilization of each structure is enough to trigger temperature-induced cooperative transitions between the two (Schreck & Yuan, 2010). The mutually exclusive nature of the native contacts between both RfaH folds (Ramírez-Sarmiento et al., 2015) allows to suggest that a metamorphic transition could occur in a similar fashion if *i*) the α- and β-contacts at the CTD are differentially stabilized, and *ii*) the energy gap between the configurations can be closed by its binding to the NTD, emulating the effect of temperature in the aforementioned mathematical model.

In this context, the emergence of a folding duality within the universally conserved NusG/Spt5 protein family can be biophysically rationalized as the rearrangement of evolutionary conserved and novel interactions that must take place during RfaH fold-switch (Gc et al., 2015; Ramírez-Sarmiento et al., 2015). Thus, coevolutionary analysis measuring the conservation of interactions rather than sequences (Morcos et al., 2011) could provide valuable insights into the determinants of protein metamorphosis, and guide its computational design as an emergent property of protein-protein interaction systems (Fleishman et al., 2011; Kortemme et al., 2004).

All in all, investigating the essentials of how metamorphic proteins fold would provide a significant amount knowledge of how fold-switching occurs and even how any given protein folds at all. Based on this, the initially proposed hypothesis of this thesis project is:

Hypothesis

A fold-switching behavior can be encoded into the CTD of *E. coli* NusG protein by redesigning its folding energy landscape to display the coevolved, optimized interdomain contacts and dual secondary structure features present in its metamorphic paralog RfaH.

Main Objective

Determine the main tertiary contacts, secondary structure propensities and thermodynamics involved in RfaH metamorphosis and design from first principle a C-terminal fragment of NusG protein that bears these features.

Specific Objective

1. Evaluate the effect of interdomain contacts in stabilization of the α-fold.
2. Determine the effect of coevolutionary interdomain contacts on RfaH metamorphic equilibrium.
3. Evaluate the secondary structure propensity of the CTD of RfaH and NusG using chemically induced α-helix formation.

The proposed research was carried out and published in the form of three scientific articles. These three articles explore either one or more of these specific objectives, fulfilling the main objective and hypothesis by providing critical information regarding the tertiary contacts, secondary structure formation, and thermodynamics of the RfaH fold-switching process.

For the first specific objective, the *Biophysical Journal* article titled "Differential local stability governs the metamorphic fold-switch of bacterial virulence factor RfaH" computationally and experimentally explores the local energetics of each RfaH fold, revealing the most important NTD and CTD regions for stabilizing the α- and β-fold in terms of their sequence and structural features.

Also exploring the first specific objective, a second publication in *PLoS Computational Biology* titled "The N-terminal domain of RfaH plays an active role in protein fold-switching" relies on a physicochemical modelling of the RfaH protein through molecular dynamics that simulate its fold switch, and the refolding of each state starting from an unfolded structure.

Altogether, both articles explore the thermodynamic features of αRfaH, particularly in respect to its dependence on specific interdomain interactions with the NTD and its transformation to βRfaH.

For the second objective, a third article published in *Protein Science* explores the effect of coevolutionary-derived contacts on RfaH folding, titled "Coevolution-derived native and non-native determine the emergence of a novel fold in a universally conserved family of transcription factors". Specifically, using two methods to estimate coevolutionary contacts from a multiple sequence alignment of RfaH-like proteins predicted to be metamorphic, it was possible to retrieve a set of roughly 50% native contacts and 50% non-native contacts to either α- or β-folded RfaH. Then, we demonstrated that both sets of coevolutionary contacts are required to correctly predict both RfaH folds, as they allow the intra- and interdomain compactness necessary for the native contacts to form.

Finally, the third objective is explored in both the first and third articles mentioned above. In the first article, the secondary structure preferences are expressed in terms of the per-residue free-energies for each CTD fold, which is an indicative of two aspects of RfaH fold-switch: the stabilization due to side chain microenvironments and the main chain or backbone stabilization into a given secondary structure. Therefore, the thermodynamics calculated therein are partly indicative of the changes in secondary structure, as most hydrophobic residues are compacted either in the interdomain interface or the incipient hydrophobic core of the β-barrel CTD.

In the third article, the secondary structure preferences estimated from sequence alone are used to filter out non-metamorphic sequences from the multiple sequence alignment of RfaH-like homologs, and further utilized along with the native and non-native contacts derived from coevolutionary analysis to successfully fold RfaH into each native state via molecular dynamics. Thus, these results demonstrate that both secondary structure preferences and tertiary interactions are relevant to reach both RfaH folds.

2. DIFFERENTIAL LOCAL STABILITY GOVERNS THE METAMORPHIC FOLD-SWITCH OF BACTERIAL VIRULENCE FACTOR RFAH

Introduction

Metamorphic proteins can access multiple structurally different and yet energetically stable states in solution (Murzin, 2008), directly challenging the uniqueness of the native state considered in the thermodynamic hypothesis proposed by Anfinsen (Dishman & Volkman, 2018), typically interpreted as *one sequence - one fold*. This process takes place by major architectural rearrangements and is commonly related to changes in protein function and dynamics (Lella & Mahalakshmi, 2017).

Escherichia coli RfaH is a metamorphic protein branching from the universally conserved NusG family of transcription elongation factors (Bailey, Hughes, & Koronakis, 1997; Belogurov, Mooney, Svetlov, Landick, & Artsimovitch, 2009), which enable processive RNA synthesis by RNA polymerase (RNAP) while simultaneously coupling it to concurrent processes (Burmann et al., 2010). In NusG proteins, this coupling is achieved by two domains connected via a flexible linker. The N-terminal domain (NTD) is a structurally conserved α/β sandwich that freely binds the transcription elongation complex (TEC) by contacting the two largest RNAP subunits to form a processivity clamp around the DNA (Belogurov, Sevostyanova, Svetlov, & Artsimovitch, 2010; Hirtreiter et al., 2010; Martinez-Rucobo, Sainsbury, Cheung, & Cramer, 2011); the C-terminal domain (CTD) is commonly folded as a small five-stranded, antiparallel β-barrel able to interact with diverse cellular targets (Burmann et al., 2010; Mooney et al., 2009).

Despite sharing 41% sequence similarity, the structure of free *E. coli* RfaH displays striking differences from its paralog NusG. Instead of the canonical β-barrel, RfaH CTD is folded as an α-helical hairpin (αRfaH) which is tightly bound to the NTD, occluding the RNAP-binding site (Figure 2) (Belogurov et al., 2007; Burmann et al., 2012). This autoinhibition is relieved upon domain dissociation, which is elicited during RfaH recruitment to the TEC or when the interdomain linker is cleaved, and the released CTD spontaneously refolds into the canonical β-barrel structure (βRfaH) observed in most NusG proteins (Figure 2A) (Burmann et al., 2012; Kang et al., 2018; Tomar, Knauer, Nandymazumdar, Rösch, & Artsimovitch, 2013; Zuber et al., 2019). This unique structural transformation is required to restrict RfaH action to just a few genes: autoinhibited RfaH is specifically recruited to a paused TEC in which an *ops* sequence in the nontemplate DNA strand forms a surface-exposed hairpin (Artsimovitch & Landick, 2002; Zuber et al., 2018); subsequently, domain dissociation leads to RfaH activation by CTD fold switching to attain a NusG-like structure (Burmann et al., 2012; Kang et al., 2018) and by binding of the NTD to its high-affinity binding site on RNAP (Belogurov et al., 2010; Kang et al., 2018). Remarkably, RfaH transformation is fully reversible as the autoinhibited state is restored upon RfaH dissociation from the RNAP (Zuber et al., 2019).

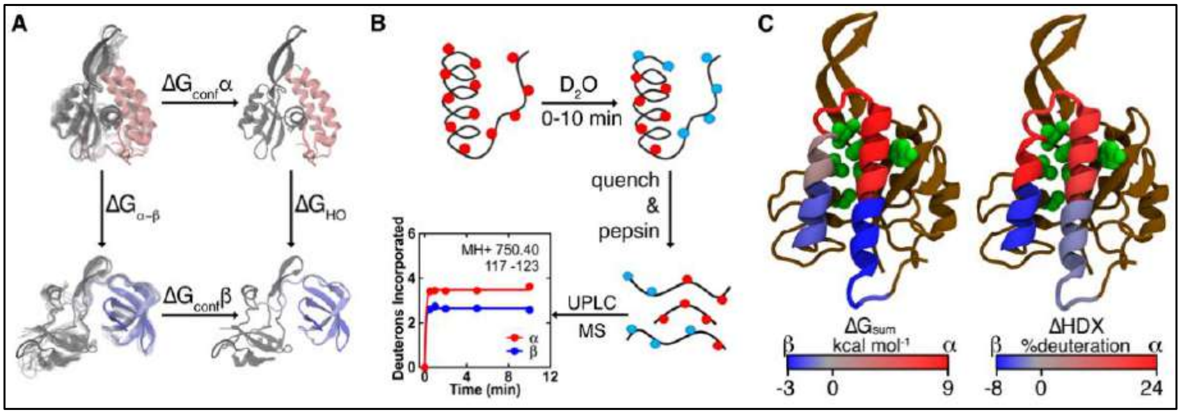


Figure 2: Computational and experimental assessment of local stability in the metamorphic protein RfaH. (*A*) Thermodynamic cycle of the confinement MD approach (Tyka et al., 2006) used to estimate the per-residue ΔG between both RfaH folds. The autoinhibited form with the CTD in the α-state (pink, PDB 5OND) and the active form with the CTD in the refolded β-state (light blue, PDB 6C6S) are confined towards a deeply minimized state through a harmonic constraint (ΔG_{conf}), allowing calculation of the difference in free energy between these structures (ΔG_{HO}). (*B*) Scheme of HDXMS experiments (Ramirez-Sarmiento & Komives, 2018). Both full-length RfaH and the isolated CTD were incubated in deuterated buffer for different reaction times, quenched and pepsin-digested for analyzing the local extent of deuteron incorporation. (*C*) Cartoon representations of the full-length αRfaH, in which the CTD covers the RNAP-binding residues from the NTD (green), summarizing our findings from simulations (left) and experiments (right) on the differential local stability towards the α- (red) and β-state (blue) of the CTD.

Since the trigger for RfaH metamorphosis is the complete *ops*-paused TEC (Zuber et al., 2018, 2019), it is challenging to study this process experimentally. Instead, most of the thermodynamic and kinetic studies have used computational approaches to directly explore this fold-switch by simulating either the isolated CTD (Balasco, Barone, & Vitagliano, 2015; Bernhardt & Hansmann, 2018; Li et al., 2014) or the entire RfaH protein (Gc et al., 2015; Ramírez-Sarmiento et al., 2015). Although the RfaH CTD is composed of only 51 residues (residues 112-162), its α-to-β transition has not been observed through conventional molecular dynamics (Balasco et al., 2015), but through the use of enhanced sampling techniques (Bernhardt & Hansmann, 2018; Gc et al., 2015; Li et al., 2014) or reduced system granularity (Ramírez-Sarmiento et al., 2015; Xiong & Liu, 2015). A way to circumvent such computing barriers is the use of confinement simulations, which rely on a discontinuous thermodynamic integration to estimate the absolute free energy of a clearly defined energy well (V. Ovchinnikov, Cecchini, & Karplus, 2013). By evaluating two alternative states within the same system, one can calculate the energy required for the structural interconversion without explicitly observing such transition (Tyka, Clarke, & Sessions, 2006).

In this work, we employed hydrogen-deuterium exchange mass spectrometry (HDXMS), [^{1}H,^{15}N] heteronuclear single quantum coherence (HSQC)-based nuclear magnetic resonance (NMR) spectroscopy, and confinement molecular dynamics (MD) to assess the differences in local stability between the autoinhibited and active folds of RfaH. By using deuterium as a probe to experimentally assess the solvent accessibility of peptides and individual backbone amides along RfaH in combination with simulations to estimate per-residue free energy changes upon refolding (Figure 2), we aimed to trace back localized regions preferentially favoring the α- or β-fold and determine how this reversible switch is encoded within RfaH sequence.

Materials and Methods

*Confinement Simulations***.** Structures of RfaH were built using the crystal structure of αRfaH (PDB 5OND) (Zuber et al., 2018) and cryo-EM composition of βRfaH (PDB 6C6S) (Kang et al., 2018). Rosetta3 suite was used to model and relax the flexible interdomain linker in both structures (Qian et al., 2007). To calculate the free energy between both structures, implicit solvent MD simulations were performed on Amber16 along with CUDA as previously reported (Cecchini, Krivov, Spichty, & Karplus, 2009; Roy, Perez, Dill, & MacCallum, 2014). Although water molecules are not being directly computed, the per-residue root mean square fluctuations (RMSF) of RfaH in both structures is comparable to explicit solvent models previously reported (Gc, Bhandari, Gerstman, & Chapagain, 2014; Joseph, Chakraborty, & Wales, 2019; Xun et al., 2016). Furthermore, this method requires an initial step of deep energy minimization of the system, suggesting that the phase change the solvent would undergo during this process would result in more artifactual dynamics than those arising from a steady solvation potential. In total, 26 independent simulations were performed per system. For each, a harmonic position-restraining potential was used to drive the atoms towards a deeply energy-minimized configuration for either the α- or β-state of RfaH (Figure 3). The stiffness of this potential was increased exponentially from mostly free ($2.5 \cdot 10^{-5}$ kcal mol^{-1} Å^{-2}) to highly restrained (419.2 kcal mol^{-1} Å^{-2}). Fluctuations and free energy were calculated for each basin as previously reported (Figure 2) (Cecchini et al., 2009).

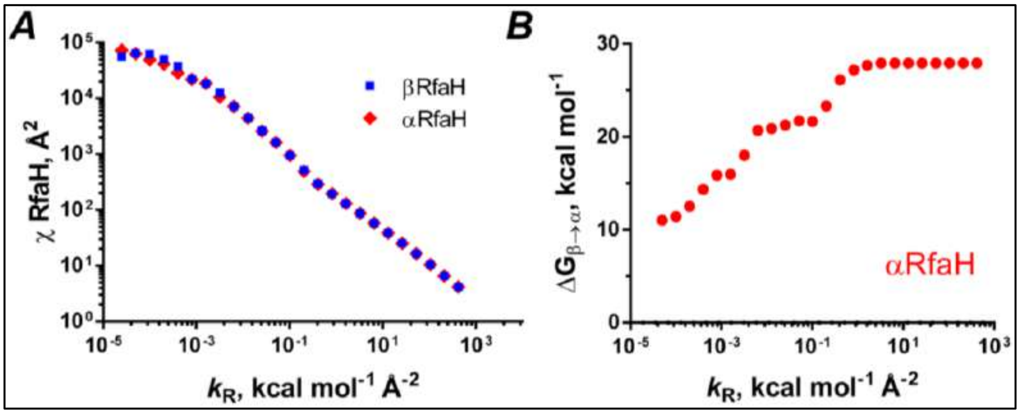


Figure 3: Dependence of fluctuations and concomitant free energy on the restraining potential. (*A*) Exponential decrease of the global squared atomic fluctuations χ of αRfaH and βRfaH with the increase of the restraining constant k_R. It can be observed that after a restraining potential of ~1 kcal mol^{-1} Å^2 both systems display the same fluctuation even for different configurations. (*B*) Free-energy difference between both RfaH states. In this, the summation of the contribution of the harmonic oscillator (11 kcal mol^{-1}) is the starting point (unconfined) and the free energy difference during confinement (area below the curve in *A*) is added at each integration step.

Briefly, the estimated free energy is the result of a thermodynamic cycle (Figure 2A) comprising two energy terms: confinement and convert. The former corresponds to the work applied by the external potential, exerted through all 26 simulations. Once confined, the magnitude of this work depends solely on the stiffness of the harmonic potential and not the configuration of the system (Figure 2). Thus, beyond this point, the confinement free-energy difference between two distinct basins converges to a single value. The convert term corresponds to the free-energy difference between the already confined states, represented as the deeply energy-minimized configuration for each basin. This is calculated by using a harmonic oscillator approximation (Cecchini et al., 2009), in which the absolute free-energy of each state is calculated from the canonical partition function for a system of vibrating particles, whose frequencies are obtained from normal-mode analysis for both deeply minimized configurations (Tyka et al., 2006). A per-residue free-energy decomposition scheme was also used as indicated previously (Roy et al., 2014).

Initial Structures. Deposited structures of αRfaH (PDB 5OND (Zuber et al., 2018)) and βRfaH (PDB 6C6S (Kang et al., 2018)) were used for confinement MD simulations. Importantly, the DNA-bound full-length RfaH structure was used instead of the extensively used free RfaH structure (PDB 2OUG (Belogurov et al., 2007)). The rationale behind it is that the latter contains a 1-residue misplacement of the last 17 NTD residues (84-100) when compared with all other RfaH NTD structures (PDB 5OND, 6C6S and 6C6T). Since the interdomain linker is expected to be flexible when the protein is isolated in solution, both structures were processed using Rosetta3 (Leaver-Fay et al., 2011). For this, fragments were generated for

RfaH sequence using Robetta (Kim, Chivian, & Baker, 2004) (available at robetta.bakerlab.org), and used along with the LoopModel protocol to generate 500 structures of α- and β-folded full-length RfaH by relaxing this linker. These structures were then minimized by Gradient Descent algorithm, and later deeply minimized (i.e. with changes in rms lower than 10^{-12} Å) using Newton-Raphson Minimization, both implemented in Amber16 through NAB (D.A. Case et al., 2016).

Free Energy Calculations through a Harmonic Oscillator Approach. Normal Mode Analysis (NMA) was performed for both structures using NAB, employing the AMBER ff14SB force field as we did with all MD/MM procedures indicated herein. The top 3N-6 positive frequencies, with N being the number of particles = 2,609, were used for computing the harmonic oscillator free energy as previously reported (Cecchini et al., 2009). Briefly, the harmonic oscillator free energy is (Equation 1):

$$(1)\ G_{HO} = -k_B T ln(z_{HO})$$

where z_{HO} is the partition function of the harmonic oscillator, k_B is the Boltzmann constant and T is the absolute temperature. The partition function corresponds to (Equation 2):

$$(2)\ z_{HO} = e^{-E/k_B T} \prod_{i=1}^{3N-6} \frac{k_B T}{v_i h}$$

where h is the Planck constant, E is the potential energy at the minimum, and v_i is the *i*-th frequency obtained from NMA, in the appropriate units. Then, free energy for each minimized structure used was calculated as (Equation 3):

$$(3)\ G_{HO} = E - k_B T \sum_{i=1}^{3N-6} ln(\frac{k_B T}{h_i v})$$

It should be noted that solving equation 3 for two harmonic oscillators having the same number of particles results in (Equation 4):

$$(4)\ \Delta G_{HO} = \Delta E - k_B T \sum_{i=1}^{3N-6} ln(\frac{v_2}{v_1})$$

This shows that only the natural logarithm of the ratio between the frequencies is relevant for the entropic contribution (rightmost summation) of the free-energy difference and implies that as long as v_1 and v_2 are expressed in the same frequency units, the energy difference can be calculated without explicitly evaluating equation 3.

Confinement Simulations and Free Energy Calculations. The aforementioned structures were used as starting configurations for implicit solvent (HCT (Hawkins, Cramer, & Truhlar, 1996)) confinement MD simulations. In these, a cartesian harmonic constraint is applied on each atom to drive it towards its deeply minimized position. These simulations are carried out for 30 ns at 298 K, using Langevin thermostat alongside SHAKE (Ryckaert, Ciccotti, & Berendsen, 1977) for hydrogens. No cutoff was used for electrostatics since no PBC was used. In these simulations, the stiffness of the harmonic potential (restraining constant, k) was increased from k_i = 2.5·10^{-5} kcal mol^{-1} Å^{-2}, doubling up 25 times until reaching k_f = 419.2 kcal mol^{-1} Å^{-2}. For calculating the energy involved in the confinement step for the entire protein as well as for each residue, the squared of the distance of each atom with respect to the minimized structure ($\chi_k = \langle N \cdot \mathrm{RMSD}^2 \rangle_k$, where N is the number of atoms) was averaged throughout each simulation for each structure. As indicated in previous works, these fluctuations decrease exponentially with the increase of the restraining constant $(\chi \approx ak^b)$ (Figure 3) (Tyka et al., 2006). Thus, the free energy was calculated simply as the area below the k, χ curve (Equation 5):

$$(5)\ \Delta G_{conf} = \int_{k_i}^{k_f} ak^b dk$$

where k is the restraining constant, a and b are unknown parameters. Since this behavior is not monotonic throughout the confinement steps, trapezoidal numerical integration for each k_i, k_{i+1} pair is used instead, which can be improved from a linear to an exponential approximation by instead using the primitive of the solution to equation 5 (V. Ovchinnikov et al., 2013; Tyka et al., 2006) (Equation 6):

$$(6)\ \int_{k_i}^{k_{i+1}} ak^b dk = \frac{ak^{b+1}}{b+1}\Big|_{k_i}^{k_{i+1}} = \frac{(ak^b)k}{b+1}\Big|_{k_i}^{k_{i+1}} = \frac{\chi_{i+1}k_{i+1} - \chi_i k_i}{b+1}$$

This shows that only b is required for the numerical integration, which can be isolated from the initial equation by evaluation between two values (Cecchini et al., 2009) (Equation 7):

$$(7)\ \chi = ak^{b};\ \left(\frac{\chi_i}{k_i^b}\right) = \left(\frac{\chi_{i+1}}{k_{i+1}^b}\right);\ b = \frac{ln(\chi_{i+1}) - ln(\chi_i)}{ln(k_{i+1}) - ln(k_i)}$$

Applying the numerical approach to equation 5 results in (Equation 8):

$$(8)\ \Delta G_{conf} = \sum_{k_i}^{k_{f-1}} \frac{(\chi_{i+1}k_{i+1} - \chi_i k_i)}{(\frac{ln(\chi_{i+1}) - ln(\chi_i)}{ln(k_{i+1}) - ln(k_i)}) + 1}$$

For a more detailed breakdown of this sum please see (Cecchini et al., 2009). This free energy can be broken down into its per-residue contribution just by considering the protein fluctuations to be the results from individual residue contributions (Equation 9):

$$(9)\ \chi = \sum_{i=1}^{L}\ r_i$$

where L is the protein length and r is the squared atom fluctuation for a residue with respect to its position in the minimized structure (Roy et al., 2014).

Free Energy Difference and Decomposition. Since we cannot decompose the contribution from the normal mode analysis, we used the same approach previously reported, consisting of calculating the change in internal free energy (ΔU) for each residue using Amber16 module *decomp* (without 1,4 long range) (Roy et al., 2014). The free energy for each structure (and residue) was calculated as (Equation 10):

$$(10) \qquad G = G_{HO} - \Delta G_{conf}$$

Therefore, the free energy difference $\Delta\Delta G$ between any pair of structures can be easily calculated as their difference (V. Ovchinnikov et al., 2013). In the case of the per-residue free energy change ($\Delta\Delta G_r$), it is calculated as (Equation 11):

$$(11) \qquad \Delta\Delta G_r = \Delta U_r - \Delta G_{conf(r)}$$

where $\Delta G_{conf\,(r)}$ is the residue free-energy difference in the confinement step, and the subscript *r* indicates single-residue potential.

*Gene Expression and Protein Purification***.** All protein sequences were encoded in plasmids harboring either a Tobacco Etch Virus protease cleaving site (TEV) or Thrombin cleaving site (HMK). Full-length *E. coli* RfaH was encoded in pIA777, a derivative of pET36b(+) containing its NTD–TEV–CTD–[His$_6$] (V. Svetlov, Belogurov, Shabrova, Vassylyev, & Artsimovitch, 2007). The isolated CTD (i.e. RfaH residues 101 to 162) was harbored in a pETGB1A vector, containing [His$_6$]-GB1-TEV-CTD (Burmann et al., 2012). *E. coli* NusG was encoded in pIA244, a derivate of pET33 (Artsimovitch & Landick, 2000), encoding [His$_6$]-HMK-NusG. For protein production, the *E. coli* BL21 (DE3) strain was used. Bacteria were grown at 37 °C until reaching an optical density at 600 nm (OD$_{600}$) = 0.6-0.7, induced with 0.2 mM IPTG (US Biological, Salem, MA, USA) at 30 °C overnight in the case of RfaH and its isolated CTD, or 30 °C for 3 hours for NusG. The cells were harvested by centrifugation at 4 °C.

Cells were disrupted by sonication at high intensity in buffer A containing 50 mM Tris-HCl pH 7.5, 150 mM NaCl and 10 mM imidazole, pH 7.5. The supernatant was obtained by centrifugation at 12,000 × *g* for 30 min, loaded onto a His-Trap HP column (GE Healthcare, Chicago, IL, USA), washed with buffer A and then eluted in gradient with the same buffer supplemented with 250 mM imidazole. For isolated CTD, this eluate was incubated in buffer A with a non-cleavable His-tagged TEV protease at 4 °C overnight in a ratio of 20:1 mg of CTD:TEV protease. This mixture was then separated using another His-Trap HP column, collecting its flow-through enriched in isolated RfaH-CTD. Purity of protein samples was verified by SDS-PAGE.

Finally, all proteins were subjected to size exclusion chromatography prior to all experiments. This was performed on a Sephadex S75 column (GE Healthcare) connected onto an ÄKTA FPLC (GE Healthcare), using 20 mM Tris-HCl pH 7.9, 40 mM KCl, 5.0 mM $MgCl_2$, 1.0 mM β-mercaptoethanol, 6.0% (v/v) glycerol as the mobile phase.

*Hydrogen-Deuterium Exchange Mass Spectrometry***.** HDXMS was performed on each protein using a Synapt G2Si system with H/DX technology (Waters Corp, Milford, MA) as in previous works (Medina et al., 2016). In these experiments, 5 µL of protein solution at an initial concentration of 11 µM were allowed to exchange at 25 °C for 0-10 min in 55 µL of deuterated buffer containing 20 mM Tris-HCl pH 7.9, 40 mM KCl, 5.0 mM $MgCl_2$, 1.0 mM β-mercaptoethanol, 6.0% (v/v) glycerol. Then, reactions were quenched for 2 min at 1 °C using an equal volume of a solution containing 2 M GndHCl, 1% formic acid, pH 2.66. The quenched samples were injected onto a custom-built pepsin-agarose column (Thermo Fischer

Scientific, Waltham, MA) and the resulting peptic peptides were separated by analytical chromatography at 1 °C. The analytes were electrosprayed into a Synapt G2-Si quadrupole time-of-flight (TOF) mass spectrometer (Waters) set to MS^E-ESI+ mode for initial peptide identification and to Mobility-TOF-ESI+ mode to collect H/DX data. Deuterium uptake was determined by calculating the shift in the centroids of the mass envelopes for each peptide compared with the undeuterated controls, using the DynamX 3.0 software (Waters). The difference in deuteron incorporation of overlapping peptides was used for calculating the incorporation of overhanging regions when the difference in mass exceeded 5 times its uncertainty. Incorporation was fitted to a single negative exponential per region to obtain the maximum deuteron incorporation per peptide, which was expressed as a percentage over the total number of amides.

Peptide sequences and deuteron incorporation. After pepsin digestion, 27 different peptic peptides were identified for the isolated CTD, 42 for the full-length RfaH protein, and 51 for the full-length NusG (Figure 4). To maximize sequence resolution, two considerations were taken: (i) incorporation was calculated for the shortest available peptic peptides; (ii) for two overlapping peptides whose sequence differs only in one overhanging bit (i.e. ACE and ACE**DF**), the deuteron uptake of the overhanging region corresponds to the difference in incorporation between the two peptides. For accuracy, the uncertainty (standard deviation, SD) of each individual peptide was considered and was propagated towards the difference peptide as the sum of their variances. If the resulting SD resulted in more than 20% of the differential uptake along the time intervals, a longer peptide was used instead. For this analysis, only the incorporating amides were considered, therefore the maximum incorporation follows the equation (Z. Zhang & Smith, 1993) (Equation 12);

$$(12) \qquad N = L_{peptide} - n_{pro} - 1$$

with $L_{peptide}$ being the length of the peptide and n_{pro} the number of proline residues contained in its sequence. The -1 arises from the fast exchange that takes place at the N-terminal of the protein or peptic peptides. However, for most overlapping peptides, the fast exchange of the N-terminal is already taken into account, thus their maximum incorporation was not corrected again for fast exchange.

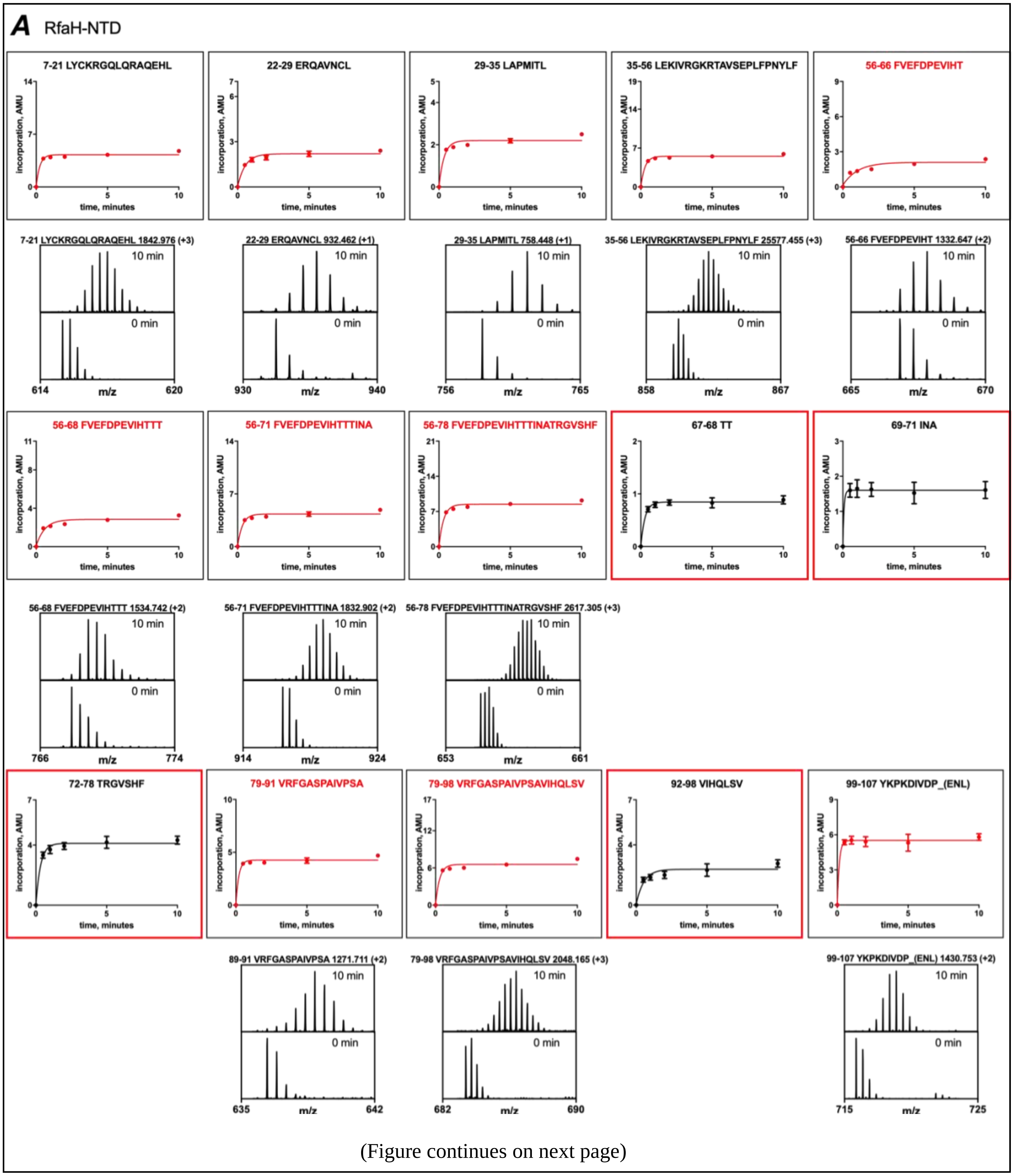
With the resulting peptic peptides and differential regions calculated (Tables Table 1, Table 2 and Table 3), their deuteron uptake was fitted to a single negative exponential as shown below (Equation 13):

$$(13) \qquad \Delta mass_t = \Delta mass_{sat} - \Delta mass_{sat} e^{-kt}$$

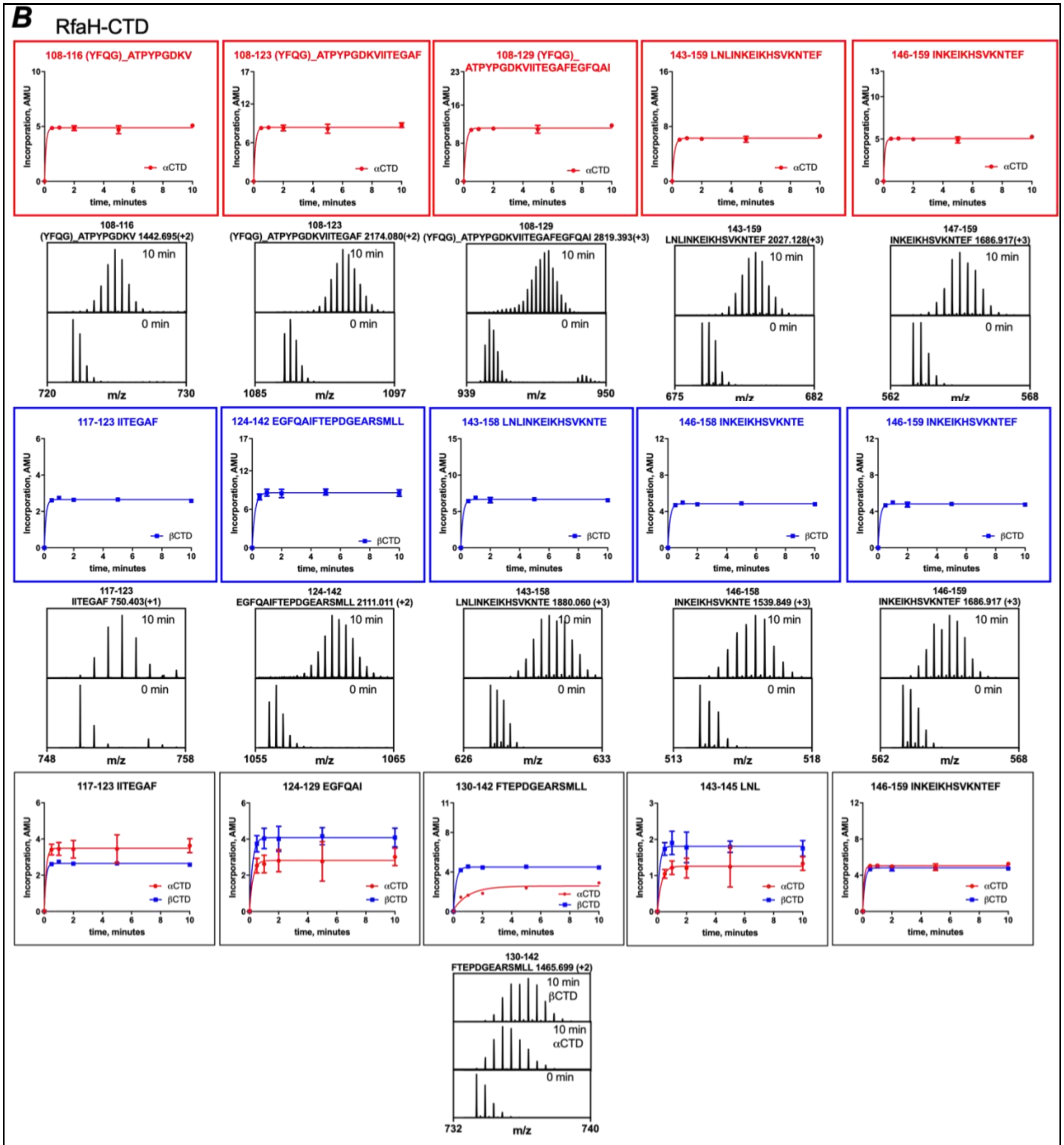
where the $\Delta mass_{sat}$ corresponds to a fitting parameter representing maximum deuteron incorporation as obtained from the experiment and k is the global rate of deuteron incorporation.

The deuteration extent (% deuteration) was calculated simply as the percentage of the maximum saturation reached by $\Delta mass_{sat}$. For a graphical representation, in the differential deuteration extents between the native forms of RfaH and between βRfaH and NusG, the free-amino ends resulting from peptic cleavage were assumed to share the same deuteration as the rest of the peptide (Figure 4).

Hydrogen-Deuterium Exchange Heteronuclear Single Quantum Coherence Spectroscopy. ^{15}N-labeled full-length RfaH and RfaH CTD were produced as described (Burmann et al., 2012). In brief, expression was carried out by growing *E. coli* in M9 minimal medium (Meyer & Schlegel, 1983; Sambrook & Russell, 2001) supplemented with $(^{15}NH_4)_2SO_4$ (Campro Scientific, Berlin, Germany) as only nitrogen source. For the HDX experiments, the proteins were in 25 mM 4-(2-hydroxyethyl)-1-piperazineethanesulfonic acids (HEPES) pH 7.5, 60 mM NaCl, 5% (v/v) glycerol and 1 mM dithiothreitol (DTT), and spectra were recorded at 288 K for full-length RfaH and 298 K for the isolated CTD on Bruker *Avance* 700 MHz and *Avance* 800 MHz spectrometers, using cryo-cooled triple-resonance bearing pulse field-gradient capabilities. After lyophilization proteins were dissolved in 500 µl D_2O (99.98%) and the decay of signal intensities was observed in a series of [^{1}H,^{15}N]-HSQC spectra over 24 hours. After the experiment, the pD was measured using a pH meter. Resonance assignment of backbone amide protons of RfaH were taken from a previous study (Burmann et al., 2012). Exchange rates were determined by fitting the signal decay to a monoexponential curve using only signals whose intensity had not completely decayed within the first 90 minutes so that at least five data points were used for fitting. The pD was corrected by adding 0.4 units to the experimentally determined value. The protection factors were calculated by dividing the experimental exchange rates (k_{ex}) by the intrinsic exchange rates calculated from the amino acid sequence (k_{rc}) and experimental conditions with tabulated parameters and were finally converted to ΔG values (Bai, Milne, Mayne, & Englander, 1993, 1994).



(Figure continues on next page)



(Figure continues on next page)

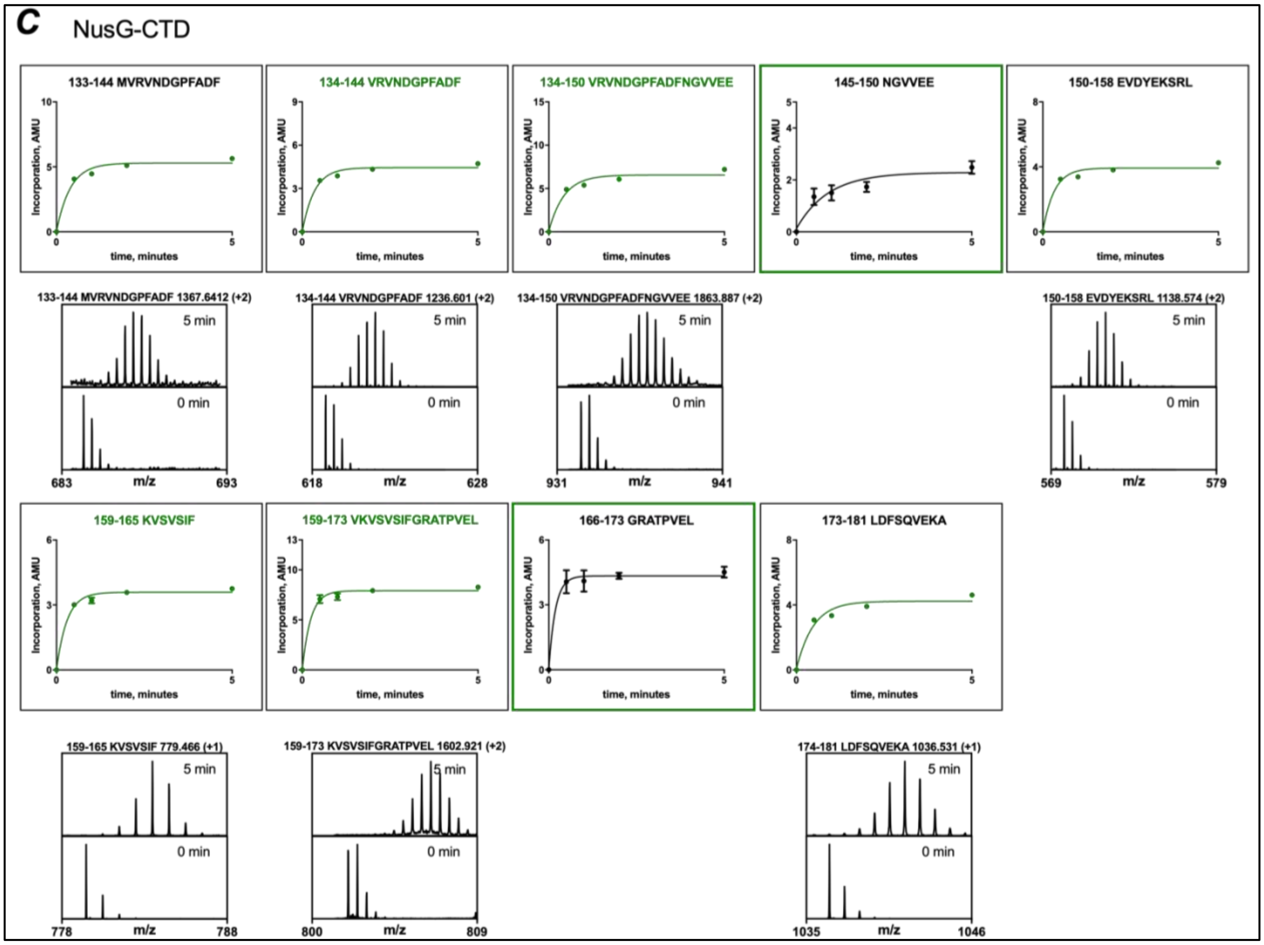


Figure 4: Deuteron incorporation and mass spectra of different regions identified for full length RfaH (red), its isolated CTD (blue) or NusG CTD (green).

Deuterium incorporation was measured between 0-10 min of incubation in deuterated buffer, except for NusG, where the maximum reaction time was 5 min. Data was fitted to a single exponential to determine the maximum extent of deuteron incorporation for each region. Only mass spectra for the minimum (0 min) and maximum (10 min) reaction times are shown. (*A*) Regions identified in the NTD of RfaH. The extent of deuteron exchange for regions indicated in red boxes and lacking mass spectra were determined based on the overlapping of two experimentally observed peptides (indicated by red titles). (*B*) Deuteron incorporation of regions of the CTD of RfaH. Peptides analyzed in the context of full-length RfaH are indicated by red boxes, whereas peptides analyzed in the context of the isolated CTD are shown in blue. These peptides were employed to calculate the extent of exchange of smaller peptide regions of RfaH in both folds by accounting for the overlapping regions between these peptides. Four peptide regions were derived using this approach (residues 117-123, 124-129, 143-145, 146-159), whereas peptide 130-142 was experimentally observed in both full-length RfaH and the isolated CTD, and its mass spectra is consistent with differences in deuterium exchange due to the topology of each native state. (*C*) Deuteron incorporation of regions of the CTD of NusG. The extent of deuteron exchange for regions indicated in green boxes and lacking mass spectra were determined based on the overlapping of two experimentally observed peptides (indicated by green titles).

Table 1: Deuteron incorporation of full-length RfaH

Position	Sequence	*k*, min^{-1}	Δmass, AMU	R^2	%Deut.
7-21	LYCKRGQLQRAQEHL	3.9 ± 1.3	4.3 ± 0.2	0.973	31
22-29	ERQAVNCL	1.8 ± 0.4	2.2 ± 0.1	0.972	31
29-35	LAPMITL	2.8 ± 0.9	2.2 ± 0.1	0.955	44
35-56	LEKIVRGKRTAVSEPLFPNYLF	3.5 ± 0.7	5.5 ± 0.1	0.988	29
56-66[a]	FVEFDPEVIHT	0.9 ± 0.4	2.1 ± 0.2	0.906	23
56-68[a,b]	FVEFDPEVIHTTT	1.7 ± 0.6	2.8 ± 0.2	0.929	25
56-71[b,c]	FVEFDPEVIHTTTINA	3.0 ± 1.0	4.3 ± 0.2	0.964	31
56-78[c]	FVEFDPEVIHTTTINATRGVSHF	3.0 ± 0.8	8.4 ± 0.4	0.979	40
67-68[a]	TT	3.5 ± 0.4	0.8 ± 0.1	0.995	40
69-71[b]	INA	12 ± 41	1.6 ± 0.1	0.996	53
72-78[c]	TRGVSHF	3.0 ± 0.5	4.1 ± 0.1	0.987	59
79-91[d]	VRFGASPAIVPSA	5.0 ± 1.9	4.2 ± 0.2	0.981	42
79-98[d]	VRFGASPAIVPSAVIHQLSV	3.5 ± 1.3	8.4 ± 0.4	0.979	49
92-98[d]	VIHQLSV	1.8 ± 0.7	2.4 ± 0.2	0.925	34
99-107	YKPKDIVDP_(ENL)	7.1 ± 3.0	5.5 ± 0.1	0.995	61
108-116[e]	(YFQG)_ATPYPGDKV	9.9±10.2	4.9 ± 0.1	0.931	49
108-123[e,f]	(YFQG)_ATPYPGDKVIITEGAF	8.6 ± 4.8	8.4 ± 0.1	0.997	49
108-129[f]	(YFQG)_ATPYPGDKVIITEGAFEGFQAI	6.6 ± 2.1	11.2 ± 0.2	0.996	49
117-123[e]	IITEGAF	7.5 ± 2.6	3.5 ± 0.1	0.997	50
124-129[f]	EGFQAI	4.3 ± 1.0	2.8 ± 0.1	0.989	47
130-142	FTEPDGEARSMLL	0.9 ± 0.4	2.6 ± 0.3	0.909	24
143-159[g]	LNLINKEIKHSVKNTEF	6.7 ± 2.0	6.3 ± 0.1	0.996	39
146-159[g]	INKEIKHSVKNTEF	11 ± 19	5.1 ± 0.1	0.997	39
143-145[g]	LNL	3.5 ± 0.5	1.3 ± 0.1	0.994	43

Expressed as average ± std. error of fit

(ENL) and (YFQG) correspond to the TEV cleaving sequence and was not considered in the sequence numbering. a-g: Deuteron incorporations in red were estimated from the difference between two overlapping peptides.

Table 2: Deuteron incorporation of the isolated CTD of RfaH

Position	Sequence	*k*, min^{-1}	Δmass, AMU	R^2	%Deut.
117-123	IITEGAF	8.6 ± 4.0	2.6 ± 0.1	0.997	43
124-142[a]	EGFQAIFTEPDGEARSMLL	5.2 ± 0.3	8.6 ± 0.1	0.999	51
130-142[a]	FTEPDGEARSMLL	5.4 ± 0.4	4.5 ± 0.1	0.999	41
124-129[a]	EGFQAI	4.9 ± 0.4	4.1 ± 0.1	0.999	68
143-158[b]	LNLINKEIKHSVKNTE	6.8 ± 1.6	6.7 ± 0.1	0.998	45
146-158[b]	INKEIKHSVKNTE	6.9 ± 1.3	4.9 ± 0.1	0.999	41
143-145[b]	LNL	6.6 ± 2.4	1.8 ± 0.1	0.995	60
146-159	INKEIKHSVKNTEF	7.0 ± 1.7	4.82 ± 0.06	0.998	37

Expressed as average ± std. error of fit

a-g: Deuteron incorporations in red were estimated from the difference between two overlapping peptides.

Table 3: Deuteron incorporation of NusG CTD

Position	Sequence	*k*, min^{-1}	Δmass, AMU	R^2	%Deut.
133-144	MVRVNDGPFADF	2.5 ± 0.6	5.3 ± 0.3	0.983	53
134-144[a]	VRVNDGPFADF	2.9 ± 0.7	4.4 ± 0.2	0.985	49
134-150[a]	VRVNDGPFADFNGVVEE	2.3 ± 0.8	6.6 ± 0.5	0.966	44
145-150[a]	NGVVEE	1.1 ± 0.5	2.3 ± 0.3	0.920	38
150-158	EVDYEKSRL	3.1 ± 1.0	3.9 ± 0. 2	0.976	49
159-165[b]	KVSVSIF	3.3 ± 0.8	3.6 ± 0.1	0.989	60
159-173[b]	KVSVSIFGRATPVEL	4.2 ± 1.1	7.9 ± 0.3	0.992	61
166-173[b]	GRATPVEL	5.3 ± 1.5	4.3 ± 0.1	0.994	61
173-181	LDFSQVEKA	2.1 ± 0.7	4.2 ± 0.3	0.964	52

Expressed as average ± std. error of fit

a-b: Deuteron incorporations in red were estimated from the difference between two overlapping peptides.

Results

Confinement Molecular Dynamics Show Localized Differential Stability. To computationally ascertain the difference in local stability between both native states of RfaH, confinement MD simulations were used to estimate their global and local free energy differences. Given that interdomain interactions are critical for the stability of the CTD in the autoinhibited state (Burmann et al., 2012), both structural states were modeled on the full-length RfaH to perform the confine-convert-release (CCR) approach (Cecchini et al., 2009; Roy et al., 2014). These models were built in both states using the crystallographic αRfaH structure (corresponding to the autoinhibited state with NTD and CTD in the α-fold (Zuber et al., 2018)) and the cryoEM βRfaH structure (corresponding to the activated, open state with NTD and CTD in the β-fold (Kang et al., 2018)), further refining the flexible loop connecting both domains using the knowledge-based Rosetta software (Qian et al., 2007). Then, these refined structures were used for confinement MD, thoroughly exploring the fluctuations from mostly free to highly restrained states, integrating and then decomposing the free energy difference between αRfaH and βRfaH required for such process (Figure 5).

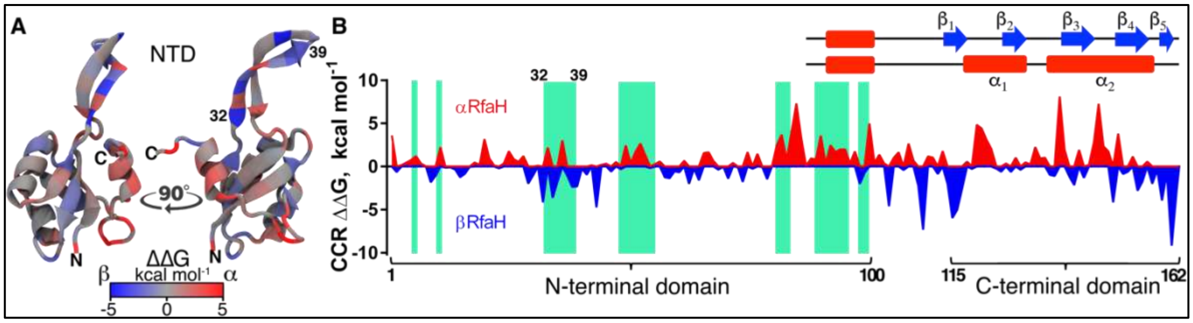


Figure 5: Confinement MD estimates anisotropic per-residue energetic contributions behind RfaH metamorphosis.

Using the CCR approximation, the contribution towards stabilizing either αRfaH (red) or βRfaH (blue) state was calculated (Cecchini et al., 2009; Roy et al., 2014) and displayed on the NTD structure (*A*) and RfaH sequence (*B*). The green stripes highlight the NTD residues forming close contacts with the CTD in the crystal structure of αRfaH (Zuber et al., 2018), with the β-hairpin residues 32-39 (also indicated in *A*) being the only NTD interface residues showing both stabilizing and destabilizing energetic contributions towards αRfaH. A visual guide for the fold-dependent secondary structure is shown for the CTD.

Free-energy decomposition at a per-residue level shows that localized groups of residues differentially stabilize either RfaH state. As shown experimentally (Burmann et al., 2012; Tomar et al., 2013), most interdomain contacts deeply stabilize the autoinhibited αRfaH, with the exception of one region comprising the first strand in the β-hairpin of the NTD (residues 32-39 in Figure 5) in which some residues are destabilizing. The behavior observed for these residues when compared to other interdomain NTD regions can be attributed to interactions between the CTD and the NTD β-hairpin observed in the structures of αRfaH (Zuber et al., 2018) and βRfaH (Kang et al., 2018) used as starting configurations for the CCR approach. As this method drives the atom positions towards highly restrained states starting from these initial structures, the CTD-NTD interactions are stably kept for βRfaH throughout the confinement MD even though for the *E. coli* paralog NusG they have been determined to be intermittent in solution (Burmann, Scheckenhofer, Schweimer, & Rösch, 2011). The persistence of these interactions could potentially lead to overestimations of the contribution of these regions towards the stability of RfaH in the β-fold in our simulations. Thus, we modelled βRfaH with a fully extended loop and lacking NTD-CTD interactions. Subsequent loop relaxation in Rosetta led to a structure similar to that determined by cryoEM for βRfaH, where CTD-NTD interactions are reestablished even when forcing Rosetta to explore extended loop configurations (Figure 6). Consequently, we consider the NTD-CTD interacting βRfaH seen in cryoEM as the functionally relevant structure for our studies. An alternative, but energetically less favored structure obtained after Rosetta relaxation shows an even larger interaction surface between NTD and CTD domains in the β-fold but contributes to changes in differential stability of the CTD in only a few residues (Figure 6).

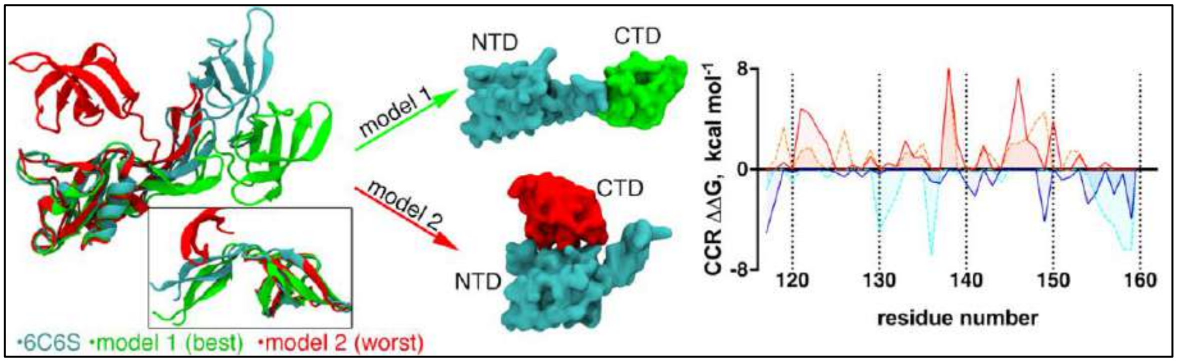


Figure 6: Effect of interdomain interactions in the active state of RfaH on the differential stability results from CCR simulations.

The NTD and CTD of RfaH in the β fold were artificially moved apart from each other and connected by an extended loop, and then Rosetta was used to relax the loop regions. In all cases, the modelled structure (green) was highly similar to the experimentally obtained structure in complex with the transcription machinery (cyan). Forcing Rosetta to explore extended loop conformations led to a less-favorable energy structure (red) with a more extensive interaction surface. Regardless, free stability differences obtained by CCR demonstrated that only a few residues (i.e. those involved in forming new interactions in the less favorable structure obtained by Rosetta) have significant changes in energetic stability towards each fold.

Although most of the CTD (residues 115-162) interacts with the NTD in αRfaH, per-residue free energy differences show localized heterogeneity in preferential stability towards each native CTD state within this region. The C-terminal region of the linker (residues 110-114) as well as initial (residues 115-119) and terminal (residues 151-162) regions of the CTD show clear preference towards forming the strands β_1 and β_4-β_5, respectively, whereas remaining residues 120-150 are more stabilized when forming the tip of the α-hairpin rather than strands β_2 and β_3 (Figure 2C and Figure 5). These results are consistent with previous simulations of the refolding pathway of RfaH in the context of the full-length protein using structure-based models (Ramírez-Sarmiento et al., 2015). This is not the first example of heterogeneous and alternating ΔG along the primary structure using the CCR method; chameleonic proteins GA30 and GB30 provide per-residue free energies that strongly correlate to the sequence content from which they were engineered, effectively tracing back conformational space information from these simulations to their primary structure (Roy et al., 2014).

Hydrogen/Deuterium Exchange Confirms Differential Stability of Metamorphic RfaH States. To determine how confinement calculations correlate with experimentally determined stabilities, HDXMS was performed with full-length RfaH, i.e. the CTD is in the α-fold, and isolated CTD, which is in the β-fold. This analysis identified a total of 43 different peptides derived from pepsin digestion of the 162 residue-long full-length RfaH (31 peptides for the NTD, 12 for the CTD), covering residues 7-159 (Figure 4, Table 1). 27 peptides were identified from pepsin digestion of the isolated CTD (residues 110-162), covering positions 117-162 (Figure 4, Table 2). Given that many of the CTD peptides have varying lengths and overlapping regions between them, backbone amide deuteron incorporation was deconvoluted into 6 unique regions that were observable for RfaH CTD in both the α-fold and the β-fold and covered almost the entirety of the CTD (residues 117-159).

As measure for flexibility, we determined the relative deuterium uptake of full-length αRfaH and the isolated βCTD (Figure 7), corresponding to the ratio between the maximum deuterium incorporation, calculated as the saturation value of an exponential fit to the deuterium uptake, and the maximum theoretical incorporation, which depends on the peptic fragment sequence and length (Z. Zhang & Smith, 1993). In the NTD, buried regions show incorporation of about 30%, while solvent-exposed regions exhibit increased deuteration of around 50%. Strikingly, in full-length RfaH almost all the peptides of the CTD display between 40-50% deuteron incorporation, with the exception of a single region covering residues 130-142, whose incorporation reaches a maximum of only 24% (Figure 5 and Figure 7B). This is reminiscent of the temperature factors observed in the crystal structure of full-length RfaH, wherein residues 115-128 and 153-156 display B-factor values over 50 while residues covering the tip of the α-hairpin display values of around 30 (Zuber et al., 2018). These results indicate that the ends of the α-hairpin are highly flexible whereas the tip is relatively more rigid.

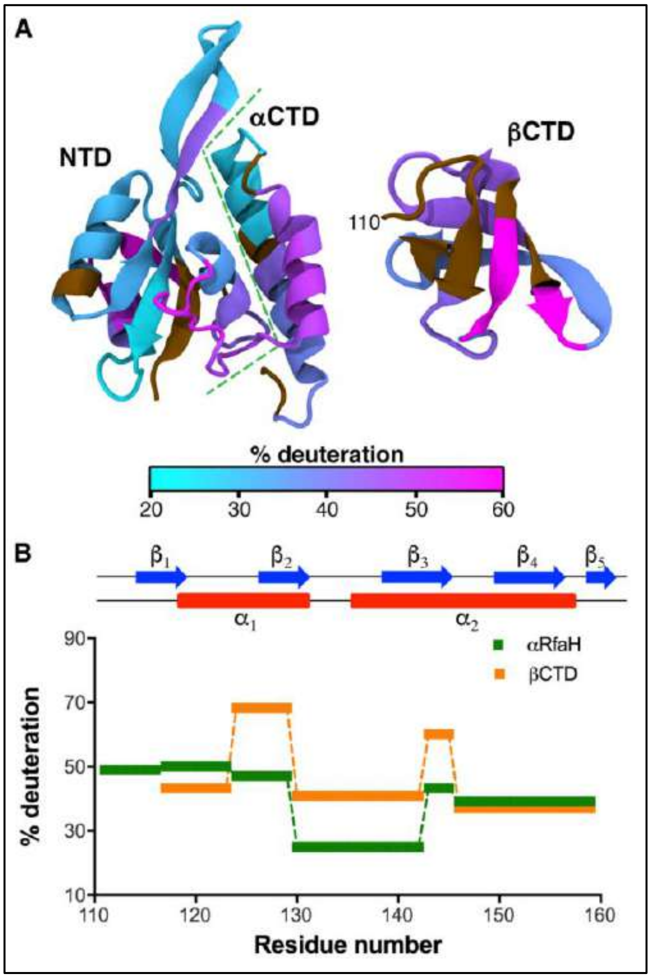


Figure 7: Mapping of the differences in local flexibility between RfaH states by HDXMS.
(*A*) Structural mapping of the relative deuterium uptake of different regions of NTD and CTD in the context of the full-length RfaH (αRfaH, PDB 5OND) or the CTD in isolation (βCTD, PDB 2LCL). Regions are colored with a gradient from cyan (solvent-protected against deuterium exchange) to magenta (solvent-accessible). Residues whose deuteron incorporation could not be determined by mass spectrometry are shown in brown. (*B*) Relative deuterium uptake for RfaH CTD in its α- and β-states (green and orange, respectively) determined by mass spectrometry.

In contrast to αRfaH, most regions of the isolated CTD display around 40% relative deuterium uptake with the exception of a short loop (residues 143-145) and strand β_2 (residues 124-129), which exhibit deuteration extents of around 60 and 68%, respectively. For comparison, the experiment was also carried out with the full-length NusG protein (Figure 8, Table 3), whose CTD is always folded as a β-barrel and that does not stably interact with its NTD (Burmann et al., 2011; Mooney et al., 2009). Relative deuterium uptake between 50-60% is observed for almost all regions within NusG CTD, slightly higher than those observed for its metamorphic paralog, except strand β_2 that exhibits 30% less exchange as compared to

RfaH (Figure 8). Despite having superimposable structures and 41% sequence similarity, RfaH CTD residues have overall larger aliphatic and more hydrophobic side chains than those of NusG, physicochemical features that are compatible with the observed lower flexibility of RfaH CTD in the β-fold. This analysis shows that, nevertheless, local flexibility of RfaH CTD in the β-fold are not drastically different from that of its paralog NusG, while highlighting the strongly reduced flexibility of the tip of the α-helical hairpin in αRfaH.

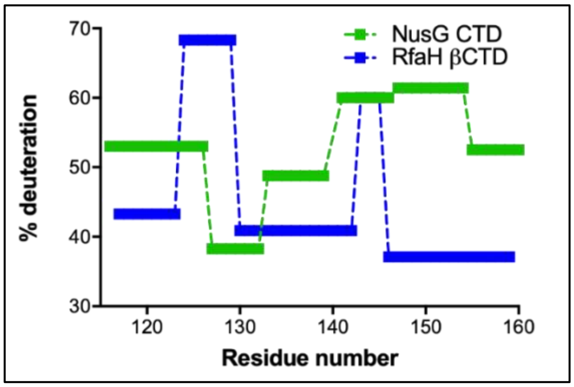


Figure 8: Maximum deuteron incorporation in NusG and in the isolated CTD of RfaH in the β-fold. In both cases, the proteins were incubated in deuterated buffer for up to 5 minutes, and its deuteron incorporation fitted to a single exponential function.

To further confirm the heterogeneity in local stability and flexibility observed for RfaH, we performed NMR-based H/D exchange experiments on full-length RfaH and the isolated CTD. The lyophilized ^{1}H,^{15}N-labeled proteins were dissolved in D_2O and HDX was monitored via the decay of signal intensities in a series of two dimensional [^{1}H,^{15}N]-HSQC spectra over 24 hours. In full-length RfaH only 17 signals in the NTD and 5 in the α-folded CTD corresponding to individual amides were detectable and analyzed. All other amide protons in full-length RfaH and all amide protons in the isolated CTD exchanged too fast (for the isolated CTD the exchange was completed within the experimental time for the first spectrum), thus suggesting that these amides are either solvent-exposed or not stably involved in hydrogen bonds to observe them in NMR-based HD exchange.

The decay rates in signal intensity of the observable amides were fitted onto a single exponential and converted into exchange protection factors (PFs). These equate to an equilibrium measurement of local stabilization in a folded conformation as compared to the unfolded state and can be further used to determine the free-energy change involved in exposing the protein amides to the solvent (Bai et al., 1993, 1994). Remarkably, all of the 22 analyzable RfaH NTD and CTD amide protons are located in regions of preferential stability towards the α-fold according to CCR (Figure 9), with all the CTD signals located on the tip of the α-helical hairpin. Also, these single amides are encompassed in peptides showing low deuteration in HDXMS experiments using full-length RfaH (Figure 9). Thus, these data confirm that the tip of the α-helical hairpin exhibits a stability comparable to that of the NTD in full-length RfaH.

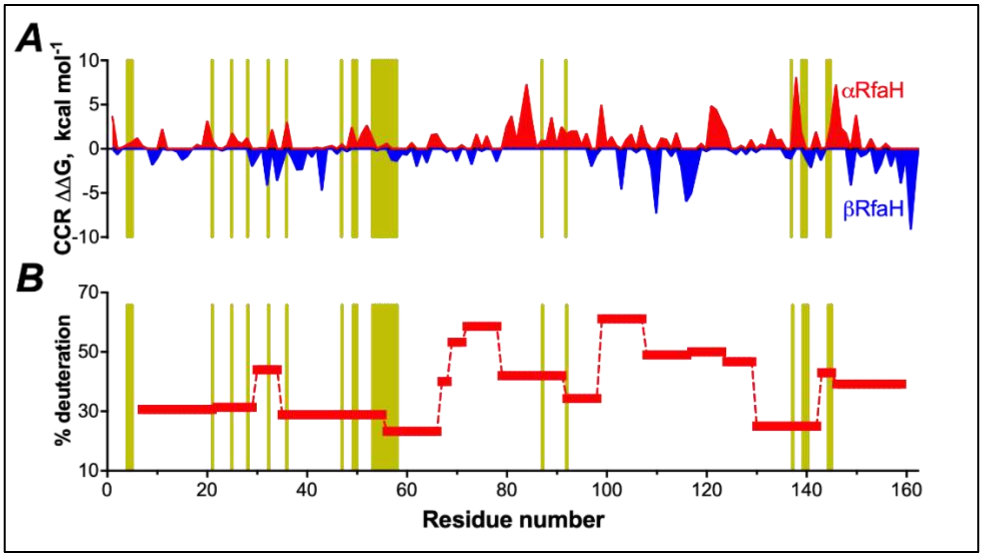


Figure 9: Highly stable single amides compared with peptic resolution and computational predictions. Amides with low deuteration due to their stable involvement in H-bonds, based on HDX measured by ^{1}H,^{15}N-HSQC, are represented as ochre lines perpendicular to the axis, compared to the per-residue preferential stability as assessed from MD simulations (*A*) or for entire peptides or regions resulting from HDXMS on full-length RfaH (*B*).

To provide greater detail into the similarities in local stability of RfaH CTD in the α-fold and the β-fold observed through HDXMS and MD, the ΔG values were summed for residues matching the 5 regions experimentally observed through HDXMS, excluding the region that contains the linker between domains. Comparison of the local differential stability patterns using both strategies revealed striking similarities in distribution as well as in magnitude (Figure 10), with the exception of CTD residues 124-129 from helix α1 that constitute part of the tip of the α-helical hairpin (Figure 2). This is partly explained by the high solvent accessibility of this region in the β-barrel fold as ascertained by HDXMS, exhibiting the highest extent of deuteration (Figure 10). In contrast, the confinement procedure of the MD simulations might stabilize the interactions within this region, thus allowing their energetics to be comparable to those estimated for the CTD in the α-fold (Figure 5).

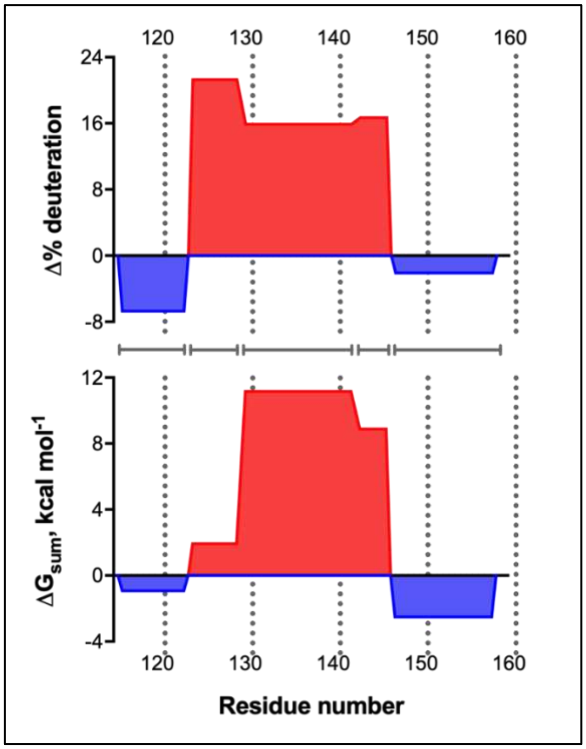


Figure 10: Comparison between computationally derived free energy (ΔG_{sum}) and experimentally determined flexibility (Δ% deuteration) between 5 indicated regions of RfaH CTD (indicated by grey lines) in its α-fold and β-fold.

Per-residue free energy differences were added accordingly to match the length of the peptides analyzed via HDXMS. Regions in red and blue have preferential stability towards the α-fold and β-fold, respectively.

Discussion

The molecular mechanism by which the transformer protein RfaH completely refolds its CTD has been experimentally elusive. Computational and experimental approaches strongly support the importance of interdomain contacts in controlling RfaH metamorphosis (Burmann et al., 2012; Ramírez-Sarmiento et al., 2015; Shi, Svetlov, Abagyan, & Artsimovitch, 2017; Tomar et al., 2013). In this work, both hydrogen-deuterium exchange and confinement MD reveal that the interaction-rich upper region of the α-helical hairpin, comprising residues 125-145, provides the highest degree of stabilization towards the α-folded CTD (Figure 2C and Figure 10). Moreover, confinement MD and [^{1}H,^{15}N]-HSQC-based HDX experiments show that interdomain contacts stabilize to a similar extent the tip of the α-helical CTD as well as the NTD (Figure 9). They also suggest that the structural metamorphosis of RfaH from the autoinhibited to the active state is controlled not only by native contacts with the other domain, but also by intrinsic CTD determinants within the aforementioned region. Thus, it comes as no surprise that the computational analysis of NusG and RfaH sequences identified seven residues that are highly conserved within the RfaH subfamily and significantly contribute to NTD-CTD binding, of which three NTD (E48, F51 and P52; *E. coli* numbering) and three CTD residues (F130, R138 and L142) are located in the vicinity of the tip of the α-helical hairpin (Burmann et al., 2012; Shi et al., 2017).

Using structure-based models, we previously simulated RNAP binding to RfaH, and concluded that contacts in the vicinity of NTD residue E48, adjacent to the tip of the hairpin, may suffice to favor CTD refolding into the β-state, and thus the relief of autoinhibition (Ramírez-Sarmiento et al., 2015). Our present results, along with the recent cryo-EM structure of RfaH bound to the complete TEC (Kang et al., 2018), are consistent with this hypothesis. Binding of the autoinhibited αRfaH, which cannot contact its high-affinity site on the β' subunit of RNAP, is thought to be mediated by its initial contacts to a hairpin that forms in the non-template DNA strand and to the β subunit gate loop (Kang et al., 2018; Sevostyanova et al., 2011; Zuber et al., 2018, 2019), resulting in an encounter complex (Zuber et al., 2019).

The preference towards the β-fold displayed by residues 110-119 and 151-162 (Figure 5) strongly suggests that refolding of the CTD towards the β-fold starts with the unfolding of the ends of the α-helical hairpin, as they fluctuate towards a locally unfolded state even before dissociation (Figure 7). Thus, the tip of the α-helical CTD seems to act as an anchor preventing its spontaneous refolding into a β-barrel. This is supported by previous observations that disruption of the E48:R138 salt bridge located in this region led RfaH to exist in equilibrium between the autoinhibited and active folds (Burmann et al., 2012). However, this view contrasts with conclusions drawn from other computational approaches (Gc et al., 2015; Li et al., 2014), which suggested that contacts involving RfaH CTD residues comprising the strand β_3 are particularly stable and nucleate the β-barrel, as solution dynamics do not display high stability in this region for RfaH or NusG CTD in the β-fold (Figure 8).

Folding of RfaH into a stable, autoinhibited structure is essential for its function. Since RfaH has a higher affinity for the TEC than NusG (Kang et al., 2018), its binding to RNAP has to be tightly controlled to prevent misregulation of NusG-dependent housekeeping genes. The emergence of the autoinhibited state, which is relieved only in the presence of a 12-bp *ops* DNA element with complex properties (Kang et al., 2018; Zuber et al., 2018), presents an elegant solution to this problem. Our results suggest that establishing interdomain contacts at the tip of the hairpin, blocking most of the RNAP-binding residues (Figure 2C), is sufficient to enable autoinhibition. Moreover, the emergence of this novel fold causes only few changes in the local stability and dynamics of the canonical β-barrel of NusG CTDs (Figure 8), supporting its ability to interact with the translational machinery (Burmann et al., 2012; Zuber et al., 2019). These interactions, established with the ribosomal protein S10, are formed through hydrophobic residues located in strands β_2 (residues 122-126) and β_4 (residues 145-148) (Burmann et al., 2012, 2010; Zuber et al., 2019) whose identities are mostly conserved between RfaH and NusG. Our results show that these residues are stably interfaced with the NTD, thus explaining why S10 is unable to elicit RfaH metamorphosis on its own (Burmann et al., 2012).

The key differences between RfaH and NusG, metamorphosis and sequence divergence of the CTD, underpin their orthogonal cellular functions. Even though NusG and RfaH bind to the same site on the TEC and display similar effects on RNA synthesis (Belogurov et al., 2009; Kang et al., 2018), they paradoxically play opposite roles in the expression of horizontally acquired genes. NusG cooperates with Rho to silence foreign DNA, an activity that explains the essentiality of *E. coli* NusG (P. Mitra, Ghosh, Hafeezunnisa, & Sen, 2017), through direct contacts between NusG-CTD and Rho (Lawson et al., 2018). In contrast, RfaH does not interact with Rho and abolishes Rho-mediated termination in its target operons, all of which have foreign origin, in part by excluding NusG (Sevostyanova et al., 2011). Remarkably, grafting a five-residue NusG loop (S^{163}IFGR) into RfaH (N^{144}LINK) creates an even more potent activator of Rho (Lawson et al., 2018). Thus, the loss of interactions with Rho, another pivotal step in the evolution of RfaH, also occurs around the tip of the helical hairpin.

Altogether, our results show that the tip of the α-helical hairpin is the main determinant for stabilizing the autoinhibited state of RfaH, and that this localized stability arises from interdomain interactions and intrinsic sequence-encoded preferences. Both our present results and other available evidence suggest that targeted substitutions in this CTD region enabled both the acquisition of the autoinhibited state, in which this region forms the tip of the stabilizing α-hairpin, and the loss of termination-promoting contacts with Rho. These changes converted a nascent paralog of NusG, an essential xenogenic silencer, into an activator of horizontally transferred virulence genes that encode capsules, toxins, and conjugation pili (Nagy et al., 2006). We hypothesize that the molecular details about RfaH mechanism can be harnessed to design ligands that interfere specifically with RfaH activity and thus virulence (D. Svetlov et al., 2018). In addition to directly inhibiting the expression of virulence genes, these ligands may also limit the spread of plasmid-encoded antibiotic-resistance determinants through conjugation and synergize with the existing drugs by compromising the cell wall integrity in Gram-negative pathogens.

3. THE N-TERMINAL DOMAIN OF RFAH PLAYS AN ACTIVE ROLE IN PROTEIN FOLD-SWITCHING

Introduction

The NusG/Spt5 family of transcription regulators is universally conserved in all three domains of life. *E. coli* NusG displays two domains in its structure, named N-terminal (NTD) and C-terminal domains (CTD) due to their location in the sequence (Artsimovitch & Knauer, 2019). The NTD is structurally conserved, folding as an α/β sandwich containing an hydrophobic depression that serves as binding site for the RNA polymerase (RNAP) (Belogurov et al., 2010), whereas the CTD folds as a small β-barrel that recruits the ribosome for coupled transcription-translation as well as other partners that regulate transcription (Figure 11) (Burmann et al., 2012, 2011; Washburn et al., 2020).

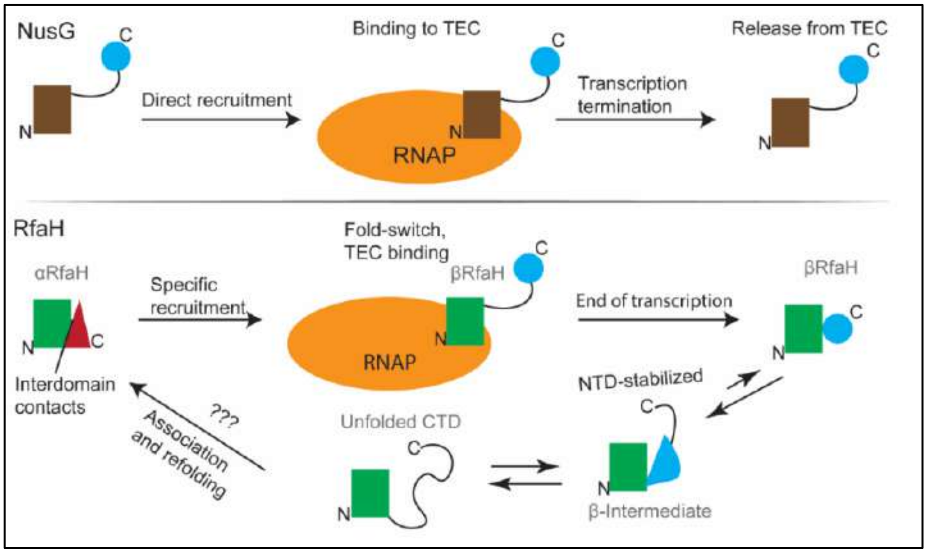


Figure 11: Schematic representation of the folding states of NusG (top) and RfaH (bottom) upon binding to and release from the transcription elongation complex (TEC).

For RfaH a fold-switch is involved in this process, in which the steps after release from the TEC corresponding to partial unfolding into a β-intermediate and transiting the unfolding state before refolding into the autoinhibited state are based on the results presented in this article.

The elongation factor RfaH of *E. coli* is a clear outlier of the NusG family of transcription factors, having an NTD with the canonical protein family structure but a CTD that is folded as an α-helical hairpin rather than the classical β-barrel (Belogurov et al., 2007). This conformation makes up the autoinhibited state of RfaH, as the α-folded CTD is blocking the RNAP binding site located at the NTD and impedes the spontaneous binding to the transcription elongation complex (TEC), i.e. the RNA polymerase in complex with DNA and RNA (Belogurov et al., 2007). This autoinhibition is relieved when the transcribing polymerase stalls at a DNA sequence named *operon polarity suppressor* (*ops*) (Bailey et al., 1997), whose exposed non-template strand forms a DNA hairpin acting as a recruiting partner for RfaH to the RNA polymerase (Artsimovitch & Landick, 2002; Belogurov et al., 2009; Zuber et al., 2018), promoting interdomain dissociation and NTD binding to the β and β' subunit of RNAP (Kang et al., 2018; Klein et al., 2011). Strikingly, the dissociated CTD refolds from the initial α-hairpin to a canonical β-barrel which serves as recruiting partner to the ribosomal protein S10, coupling transcription with translation (Figure 11) (Burmann et al., 2012; Kang et al., 2018; C. Wang et al., 2020).

Numerous studies have addressed the metamorphosis of RfaH through a computational approach, in part due to the difficulties of observing the process in solution since the trigger for RfaH interdomain dissociation is the entire TEC. There have been reports indicating the possible pathways through which the isolated CTD may refold from the α- to the β-fold (Balasco et al., 2015; Bernhardt & Hansmann, 2018; Joseph et al., 2019; Li et al., 2014), which differ from the ones proposed when the CTD is accompanied by the NTD (Gc et al., 2015; Ramírez-Sarmiento et al., 2015; Seifi, Aina, & Wallin, 2021). These results suggest that interactions formed between both domains strongly aid in stabilizing the α-fold as well as forming intermediate states that enable the transition between folds (Ramírez-Sarmiento et al., 2015). Nevertheless, these studies have focused mostly on the CTD transformation, leaving aside the details of how the NTD stabilizes the α-fold or its effects over the β-folded CTD after release from the TEC. The specifics of NTD-induced energetics on RfaH are not trivial, since the structure of RfaH-NTD (Zuber et al., 2018) displays a more hydrophobic patch than that of NusG (Kang et al., 2018; Mooney et al., 2009), which has been simultaneously associated to a tighter binding to RNAP, being RfaH NTD the only trigger required for fold-switching back from the active into the autoinhibited state (Tomar et al., 2013).

In this work, we relied on the Associative Water-Mediated Structure and Energy Model (AWSEM) to determine the effect that the NTD of RfaH has on the overall transformation energetics and the configurational space of both folds. AWSEM is a transferable force field, coarse-grained to three beads per residue (C_α, C_β and O), initially used to predict protein structure (Davtyan et al., 2012). As a force field, it has been successfully used to study the NF-κB/IκB/DNA regulatory system (Potoyan, Bueno, Zheng, Komives, & Wolynes, 2017), the nucleosome dynamics and energetics (B. Zhang, Zheng, Papoian, & Wolynes, 2016) and to determine the energy landscape of aggregation of the amyloid-β protein (Zheng, Tsai, Chen, & Wolynes, 2016), among others. Unlike common atomistic force-fields, its energy potentials and granularity have been developed for efficiently explore protein folding while robustly carrying enough information to represent up to the dihedral behavior of the main chain. This is a significant step up from our previous works on RfaH using a structure-based C_α model (Ramírez-Sarmiento et al., 2015), as not only we are now reducing the granularity but also increasing the roughness of the energy surface by including potentials for hydrogen bonding and solvent exposure propensity of each residue as well as residue-residue pairwise potentials that consider residue identity (Davtyan et al., 2012).

Using umbrella sampling, we determined the change in stability associated to interdomain separation and subsequent fold-switching, recapitulating the experimentally determined equilibrium of the system. That is, RfaH is much more stable in the α-configuration, but the β-folded CTD becomes much more stable in the absence of the NTD. Further temperature refolding simulations in the absence of information of known interdomain contacts showed that the highly hydrophobic side of the α-folded CTD consistently looks for an interaction partner and the NTD provides a suitable surface for its stabilization, recapitulating the binding orientation experimentally observed in solved structures of the autoinhibited state of RfaH. At the same time, the NTD interferes with βCTD refolding by mostly trapping it into a β-barrel intermediate, which is also observed in its metamorphic pathway. Altogether, these results suggest that the NTD favors the CTD transformation towards the α-folded CTD by simultaneously stabilizing the α-hairpin and switching the equilibrium to favor β-barrel rupture into a β-intermediate state that is part of the refolding pathway towards its autoinhibited state.

Methods

Initial structures for molecular dynamics. The structure of the full-length RfaH protein in its α-state (αRfaH hereafter) was extracted from the crystal structure deposited in the Protein Data Bank (PDB) with accession ID 5ond, and so was the α-folded CTD (αCTD hereafter). The isolated β-folded CTD (βCTD hereafter) was extracted from the first NMR solution model of the PDB accession ID 2lcl, whereas the full-length version of the active β-folded RfaH (βRfaH hereafter) was extracted from the cryoEM RfaH:TEC structure with PDB accession ID 6c6s. On the other hand, the isolated CTD of NusG was extracted from the first model of the NMR-determined structure with PDB accession ID 2jvv.

The AWSEM force field. The Associative Water-mediated Structure and Energy Model, AWSEM, (Davtyan et al., 2012) is a coarse-grained molecular dynamics (MD) protein folding model implemented in LAMMPS (Plimpton, 1995). The granularity and efficiency of this model is achieved by reducing the number of atoms per residue to three beads, the C_α C_β and O atoms, with the rest of them being calculated from ideal backbone geometry. This model contains five energy terms, which are extensively described in the work by Davtyan and cols (Davtyan et al., 2012) and are briefly summarized below (Equation 14):

$$V_{total} = V_{backbone} + V_{contact} + V_{burial} + V_{hydrogen\ bonding} + V_{DSB} + V_{memory} \quad (14)$$

Of these terms, the *backbone* energy term guides the atoms to a protein-like geometry, which is achieved using potentials that ensure atom connectivity, chirality, Ramachandran distribution, and excluded volume interaction. The *contact* term defines C_β-C_β distances and is responsible for the formation of residue-residue interactions in an amino acid-dependent manner. This potential includes pairwise direct contact potentials and many-body water-mediated contact potentials. The *burial* energy term is a many-body interaction potential that regulates solvent exposure of the protein core, depending on whether a residue has propensity to be in a low, medium, or high-density environment. The *hydrogen bonding* term replicates the contacts of carbonyl oxygen to amide nitrogen formed in α-helices, parallel β-sheets, and anti-parallel β-sheets. This potential includes additive terms for hydrogen bonding and cooperative stabilization terms for β-sheets, which we modified such that sheets of length of 3 residues can form, as the shortest strands observed in the β-barrels of NusG and RfaH are of this length. Finally, the *memory* term is a local bias applied to overlapping fragments from 3 to 9 residues that guides C_α and C_β distances to those of a reference structure, being the only native bias that is used in these simulations. This potential has the form (Equation 15):

$$(15) \qquad V_{memory} = -\lambda_{memory} \sum_m \omega_m \sum_{ij} e^{-\frac{(r_{ij}-r_{ij}^{m})^2}{2\sigma_{ij}^2}} \text{; with } \sigma_{ij} = |j-i|^{0.15} \text{ Å}$$

In this equation, the outer sum is carried out over all the aligned memory fragments, i.e., all short overlapping segments that share high sequence identity to a library of known proteins structures, with ω_m corresponding to the memory weight. The inner sum is carried out over the C_α and C_β *i,j* pairs that are separated by at least 2 residues, with r_{ij} being the distance between the atoms and r^m_{ij} the distance in the reference fragment. Finally, λ_{memory} corresponds to a scaling factor of the strength of this potential relative to the other terms. This potential can be guided to multiple structures in a simulation, or as used in this work, limited to a single or two reference structures (Davtyan et al., 2012). Also, the λ_{memory} used in this work is of 0.3 compared to the default value of 0.2, resulting in a higher cooperativity due to a decrease in the roughness of the final potential.

Calculation of Q_{diff} and umbrella sampling. Normally, MD simulations sample configurations that are very close to the initial structure, hence observing structural transitions such as RfaH fold-switching would be a rare event that would require a very long simulation time. A way to overcome this is by using enhanced sampling strategies, such as umbrella sampling. This technique enables exploring poorly sampled regions of the configurational space by applying an external bias along a reaction coordinate that describes the transition between both RfaH folds. Generally, this external bias corresponds to a harmonic potential that is applied to multiple different reaction coordinate values, such that different simulations thoroughly sample a narrow phase space while ensuring the potential energy overlap between simulations at adjacent values along the reaction coordinate. The potential energy and reaction coordinate values from multiple independent simulations are then used as input for the Weighted Histogram Analysis Method (WHAM) (Kumar, Rosenberg, Bouzida, Swendsen, & Kollman, 1992) that returns the unbiased free energy landscape of RfaH fold-switching.

For the umbrella sampling method, 51 simulations of $2.4 \cdot 10^7$ timesteps or 120 ns each were run, and energy and frames were collected every 1,000 timesteps. The initial configuration was that of the unfolded isolated CTD and a dual memory approach was used, i.e., the fragments were driven to the memory of αCTD and βCTD with equal strength. Similarly, for the full-length protein the initial state was that of the folded NTD plus unfolded CTD. The simulations sampled fractions of an order parameter called Q_{diff} which corresponds to (M. Chen, Schafer, Zheng, & Wolynes, 2018; Zheng et al., 2016) (Equation 16 & 17):

$$(16) \qquad Q_{diff} = \frac{q - q_A}{q_B - q_A}; where$$

$$(17) \qquad q = \frac{1}{(N-2)(N-3)} \sum_{j>i+2} \left[e^{-(r_{ij} - r_{ij}^{A})^2 / 2\sigma_{ij}^2} + 1 - e^{-(r_{ij} - r_{ij}^{B})^2 / 2\sigma_{ij}^2} \right]; \text{ with } \sigma_{ij} = |j - i|^{0.15} \text{ Å}$$

Where N is the sequence length, q_A and q_B are constants obtained by evaluating the q function in the two structures to which the transition is to be interpolated, and r_{ij} measures the C_α distance between residue i and j in the simulation, where superscripts A and B refer to such distance in each reference structure. This is evaluated for all contacts between $j>i+2$ residues whose C_α are at 9.5 Å or below in at least one of the reference structures. This distance is calculated from a Cα-Cα distance matrix for RfaH or the isolated CTD. In the case of the full-length protein, the NTD was excluded from the calculations for the autoinhibited and active RfaH configurations, as it does not experience a conformational change during RfaH activation. Interdomain contacts in the starting structure for βRfaH were also excluded. These exclusions were achieved by increasing the residue-residue distances within the NTD and between the NTD and CTD of βRfaH in the distance matrices to 99 Å (Figure 12).

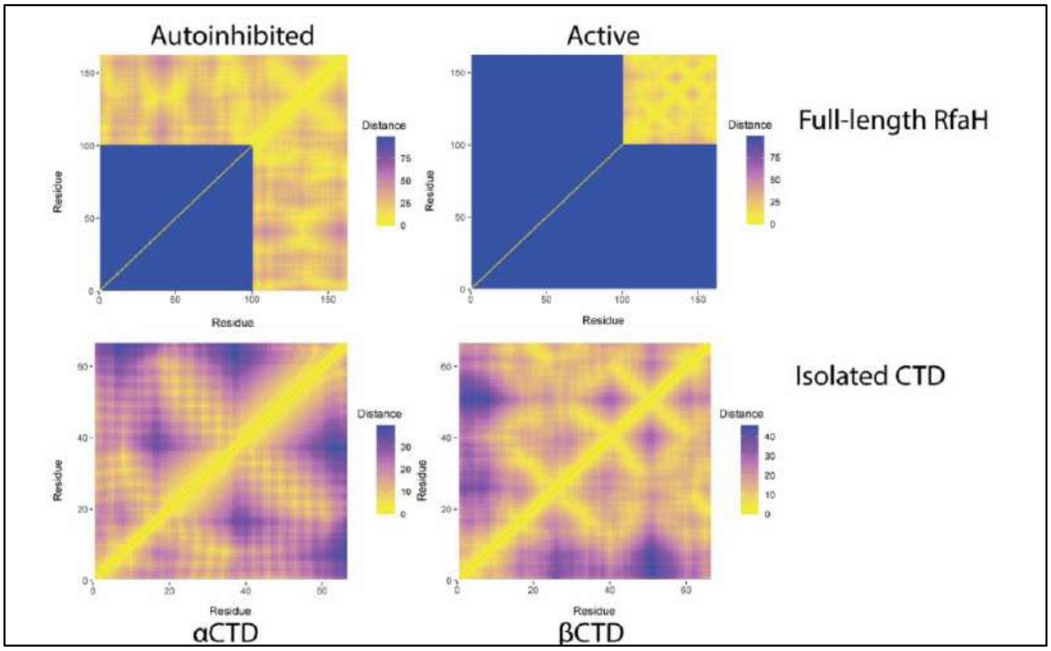


Figure 12: Interaction matrices for umbrella sampling using Q*diff*.

C_α residue-residue distance matrices for full-length RfaH and its isolated CTD. The matrices grow along the diagonal, which represents the same residue distance, in this case set to 0. Along this diagonal, contacts are formed in a 1-4 residue pattern for α-helices, antiparallel and parallel lines indicating β-strands. The blue blocks indicate regions of high distance (99 Å), which were manually set in order to exclude them from the Q_{diff} calculation.

Using the Q_{diff} value, a bias is applied by adding a new potential to the system with the form (Equation 18):

$$(18) \qquad V_{Umbrella} = \frac{1}{2}k(Q_{diff} - Q_0)^2$$

Where k is the harmonic potential constant, here 1,500 kcal·mol^{-1}, and Q_0 being the center of the distribution of a Q_{diff} value ranging from 0.00 to 1.00 by increments of 0.02. From these simulations the potential energy and Q_{diff} values were obtained for each frame, as well as the C_α RMSD of the best-fit against both reference CTD folds that were calculated using VMD (Humphrey, Dalke, & Schulten, 1996). The simulations exploring the same Q_{diff} range were run at two temperatures, 650 K and 750 K, and the AWSEM temperature units were expressed as folding temperature (T_f) by expressing these temperatures relative to the folding temperature of full-length αRfaH (~650 K). Histograms of these quantities show overlap between simulations at adjacent Q$_{diff}$ values (Figure 13).

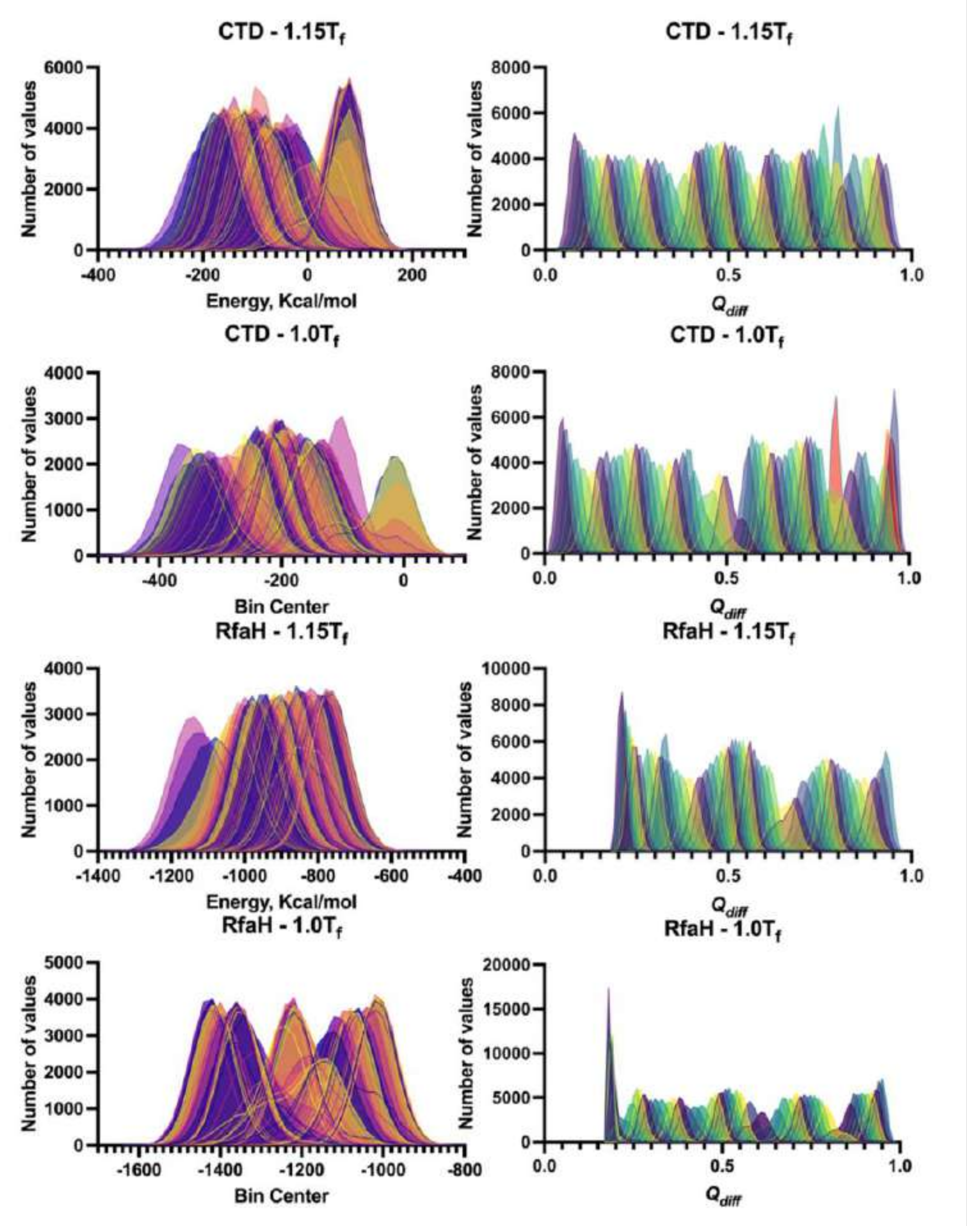


Figure 13: Histograms of the energy and Q_{diff} reaction coordinates in umbrella sampling.

In these umbrella sampling simulations, 51 simulations in Q_{diff} steps of 0.02 were run, totaling 51 simulations per system per temperature. The histograms marked in red were not used for the WHAM analysis as the simulation got trapped in a misfolded configuration. RfaH reaches the α-folded autoinhibited state when Q_{diff} = 1 and the isolated CTD reaches the β-folded state when Q_{diff} = 1.

The RMSD against αCTD and βCTD were then used as reaction coordinates for thermodynamic analysis using the WHAM algorithm (Noel, Whitford, Sanbonmatsu, & Onuchic, 2010) implemented in Java (Noel et al., 2016). For this analysis, the first 4,000 frames or 20 ns were excluded as this was the equilibration time from the unfolded state to the desired biased configuration.

Refolding simulations. For these simulations, random initial unfolded configurations for each system were generated by running 100,000 timesteps of 5 fs of a simulation without any potential but the backbone energy term, saving a simulation restart configuration every 10,000 timesteps. The restart configuration with the lowest Q_W value, which in all cases was below 0.1, was used as a starting configuration for the refolding simulations. All 100 simulations were randomly assigned initial velocities and run for $3 \cdot 10^7$ timesteps of 5 fs, totaling 150 ns each, during which the temperature linearly decreased from $1.5T_f$ to $0.6T_f$, where the temperature is expressed relative to the T_f of αRfaH, the predominant state in solution for full-length RfaH (Figure 14).

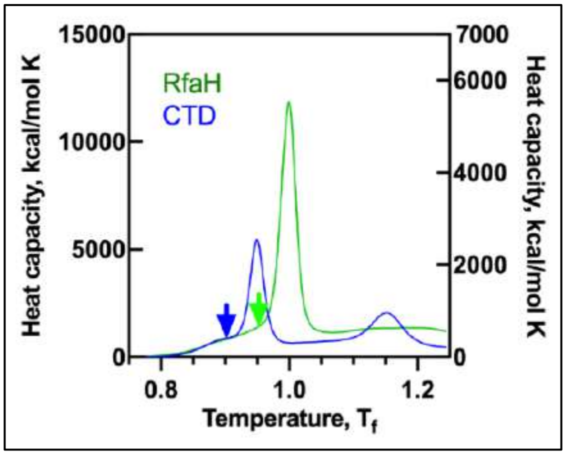


Figure 14: Heat capacity of RfaH.

Heat capacity calculated from umbrella simulations on the full-length RfaH and the isolated CTD. The blue arrow indicates the temperature selected for presenting the free energy landscape of the isolated CTD in Figure 15A, and the blue arrow indicates the temperature selected for presenting the free energy landscape of the full-length RfaH in Figure 15B. The values on the left y-axis correspond to RfaH, whereas the values on the right y-axis correspond to the isolated CTD.

All constructs were completely unfolded at the initial temperature and either completely refolded, trapped into an intermediate state or misfolded at the final temperature. The final structures of these simulations were clustered by calculating pairwise best-fit RMSD (Kelley, Gardner, & Sutcliffe, 1996) using Chimera (Pettersen et al., 2004). For the representative member of each cluster, as well as for non-clustered models, the secondary structure assignment was calculated using STRIDE (Frishman & Argos, 1995). These secondary structure assignments are summarized in Table 4 alongside the corresponding Q_W, which is a measure of structural similarity to a given structure and obtained using the formula (Sirovetz, Schafer, & Wolynes, 2017) (Equation 19):

$$(19)\qquad Q_w = \frac{2}{(N-2)(N-3)} \sum_{j>i+2} e^{-(r_{ij}-r_{ij}^{N})^2/2\sigma_{ij}^2} ; with\ \sigma_{ij} = |j-i|^{0.15}\text{Å}$$

Where, similarly to Q_{diff}, r_{ij} measures the C_α distance between residues *i* and *j* for the current and reference (superscript N) structure, given that the distance in the latter is lower than 9.5 Å, and *N* stands for the number of residues in the protein.

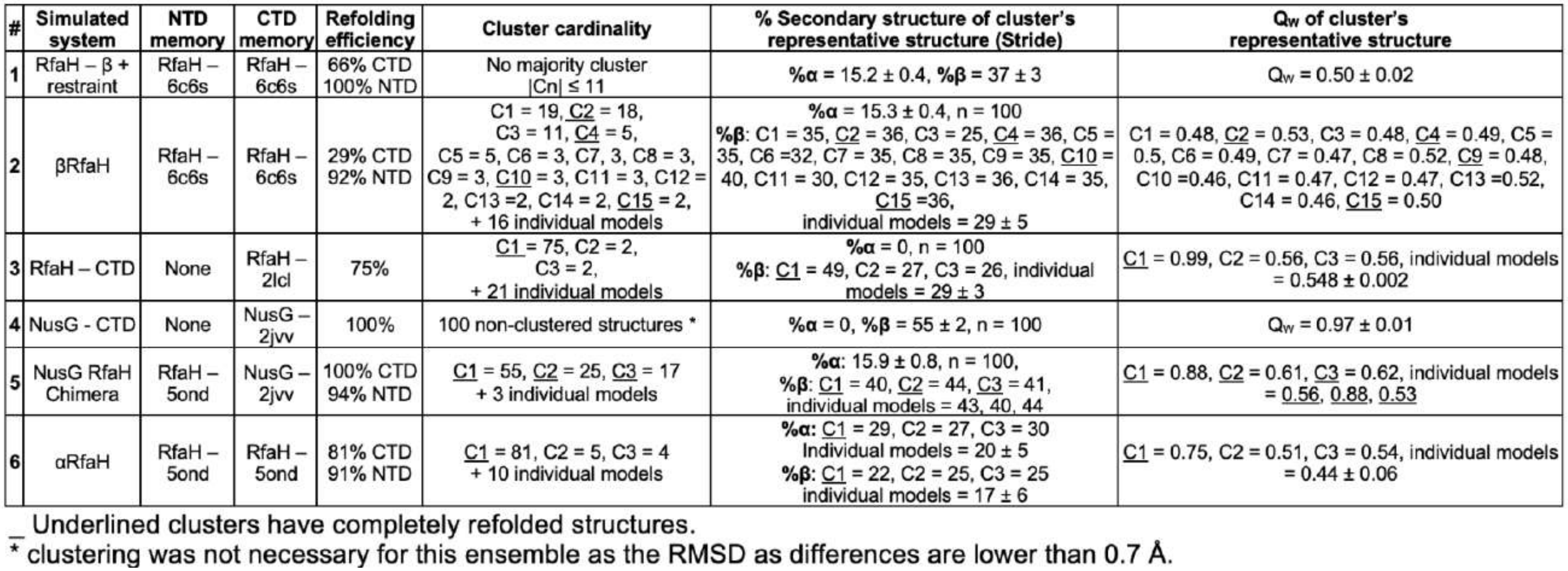
Results

MD simulations of RfaH and its isolated CTD recapitulate their experimental states. The simplest question that can be asked to an energy model about RfaH is whether it can replicate the experimentally observed CTD populations of α and β folds. More precisely, the strong predominance of αRfaH for the full-length protein in solution (Belogurov et al., 2007; Zuber et al., 2018), and of βCTD when this domain is isolated as the result of the NTD-CTD linker being cleaved or by purifying only this domain in solution (Burmann et al., 2012).

To explore this scenario, we set up umbrella sampling simulations that guide the transformation of RfaH for two systems: one in which we modeled the transition in the context of the full-length protein, that is αRfaH and βRfaH, and another in which only its CTD is modeled transitioning between αCTD and βCTD. Specifically, 51 umbrella simulations were generated for each system at two temperatures, 1.0 and 1.15T_f, where each simulation is energetically biased to explore a fraction of the configurations determined by a reaction coordinate named Q_{diff}, resulting in a gradual exploration of the configurational space between the α- and β-states of either full-length RfaH or its isolated CTD.

This exploration of the transformation was then analyzed using WHAM (Noel et al., 2010), and the heat capacity was visually inspected. To evaluate the change in stability between RfaH folds, free energy surfaces were calculated at a temperature just below the first peak in heat capacity for each system to ascertain the preferred folded state (Figure 15A and B).

Table 4: Refolding summary for protein systems.



#	Simulated system	NTD memory	CTD memory	Refolding efficiency	Cluster cardinality	% Secondary structure of cluster's representative structure (Stride)	Q_W of cluster's representative structure
1	RfaH – β + restraint	RfaH – 6c6s	RfaH – 6c6s	66% CTD 100% NTD	No majority cluster |Cn| ≤ 11	**%α** = 15.2 ± 0.4, **%β** = 37 ± 3	Q_W = 0.50 ± 0.02
2	βRfaH	RfaH – 6c6s	RfaH – 6c6s	29% CTD 92% NTD	C1 = 19, C2 = 18, C3 = 11, C4 = 5, C5 = 5, C6 = 3, C7, 3, C8 = 3, C9 = 3, C10 = 3, C11 = 3, C12 = 2, C13 =2, C14 = 2, C15 = 2, + 16 individual models	**%α** = 15.3 ± 0.4, n = 100 **%β**: C1 = 35, C2 = 36, C3 = 25, C4 = 36, C5 = 35, C6 =32, C7 = 35, C8 = 35, C9 = 35, C10 = 40, C11 = 30, C12 = 35, C13 = 36, C14 = 35, C15 =36, individual models = 29 ± 5	C1 = 0.48, C2 = 0.53, C3 = 0.48, C4 = 0.49, C5 = 0.5, C6 = 0.49, C7 = 0.47, C8 = 0.52, C9 = 0.48, C10 =0.46, C11 = 0.47, C12 = 0.47, C13 =0.52, C14 = 0.46, C15 = 0.50
3	RfaH – CTD	None	RfaH – 2lcl	75%	C1 = 75, C2 = 2, C3 = 2, + 21 individual models	**%α** = 0, n = 100 **%β**: C1 = 49, C2 = 27, C3 = 26, individual models = 29 ± 3	C1 = 0.99, C2 = 0.56, C3 = 0.56, individual models = 0.548 ± 0.002
4	NusG - CTD	None	NusG – 2jvv	100%	100 non-clustered structures *	**%α** = 0, **%β** = 55 ± 2, n = 100	Q_W = 0.97 ± 0.01
5	NusG RfaH Chimera	RfaH – 5ond	NusG – 2jvv	100% CTD 94% NTD	C1 = 55, C2 = 25, C3 = 17 + 3 individual models	**%α**: 15.9 ± 0.8, n = 100, **%β**: C1 = 40, C2 = 44, C3 = 41, individual models = 43, 40, 44	C1 = 0.88, C2 = 0.61, C3 = 0.62, individual models = 0.56, 0.88, 0.53
6	αRfaH	RfaH – 5ond	RfaH – 5ond	81% CTD 91% NTD	C1 = 81, C2 = 5, C3 = 4 + 10 individual models	**%α:** C1 = 29, C2 = 27, C3 = 30 Individual models = 20 ± 5 **%β:** C1 = 22, C2 = 25, C3 = 25 individual models = 17 ± 6	C1 = 0.75, C2 = 0.51, C3 = 0.54, individual models = 0.44 ± 0.06

_ Underlined clusters have completely refolded structures.

* clustering was not necessary for this ensemble as the RMSD as differences are lower than 0.7 Å.

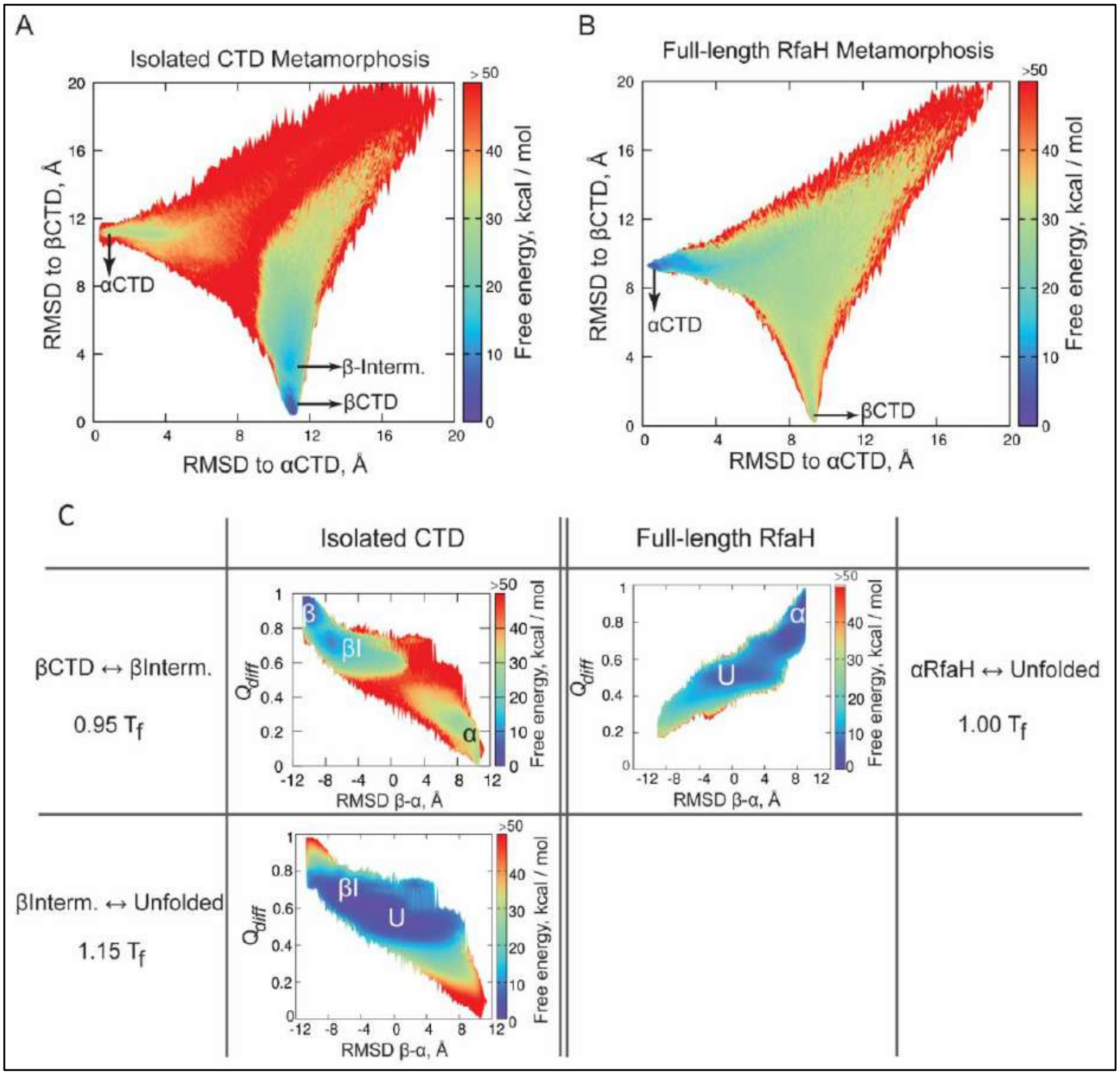


Figure 15: Energetics of RfaH transformation.

(*A*, *B*) Free energy surface for the transformation of RfaH CTD in the full-length protein (*A*) or the isolated domain (*B*). The RMSD against the experimental αCTD and βCTD were used as reaction coordinates. (*C*) Free energy surface of the transitions of RfaH CTD in the context of the full-length protein with folded NTD or the isolated CTD, projected onto the transformation reaction coordinates Q_{diff} and RMSD β-α. Here, βI corresponds to a folding intermediate, and U corresponds to the unfolded state.

The isolated CTD free energy surface displays two minima of similar free energy at low RMSD of βCTD and a higher free energy minimum at low RMSD of αCTD (Figure 15A). This suggests that the isolated CTD (residues 100 to 162) exist predominantly as a β-barrel, and it needs to cross an energy barrier of over 50 kcal·mol^{-1} to reach the α-folded state. On the other hand, the energy landscape of the CTD in the context of the full-length protein displays a major free energy minimum that expands between 1 and 4 Å in RMSD to αCTD (Figure 15B), indicating that RfaH exists predominantly in the autoinhibited state. These results are consistent with the experimental evidence for full-length RfaH and the isolated CTD in solution (Burmann et al., 2012).

The fold-switching path explored in our simulations is best observed when projecting the free energy surface onto coordinates that directly measure the structural transition of RfaH CTD, such as Q_{diff} and the difference in RMSD between βCTD and αCTD. These transitions, shown in Figure 15C, were obtained at temperatures where the peak in heat capacity is observed for each system.

In the case of the isolated CTD, the first peak in heat capacity is observed at $0.95T_f$ and corresponds to the transition between the folded βCTD and a folding intermediate. Meanwhile, a second peak in heat capacity is observed at 1.15Tf and corresponds to the transition between the β-intermediate and the unfolded state. In the first of these landscapes, the αCTD minimum is shown as a high and broad free energy minimum similarly to its basin observed in Figure 15B, a characteristic that likely arises from the structuredness of the helices, which have been ascertained in both simulations (Ramírez-Sarmiento et al., 2015; Seifi & Wallin, 2021) and experiments (Galaz-Davison et al., 2020).

By projecting the free energy into a single coordinate, namely Q_{diff}, the energy barriers involved in the fold-switching process can be observed more clearly. The transition between the β-barrel and β-intermediate has an estimated free energy barrier of 6.4 kcal/mol, whereas the transition between the β-intermediate and the unfolded minimum has a free energy barrier of mere 1.5 kcal/mol. At 1.15Tf only the β-intermediate and unfolded states are observed, while at $0.95T_f$ the transition to the αCTD is better observed, separated by a free energy barrier of 30 kcal/mol with a transition state sitting at unfolded configurations (Figure 16).

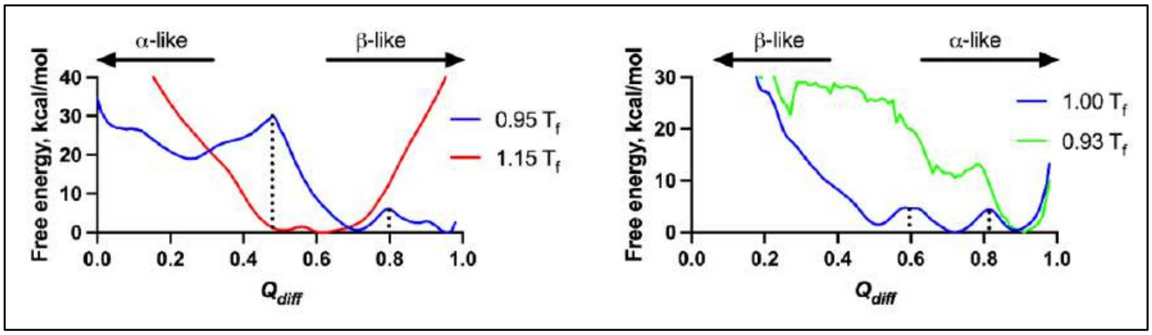


Figure 16: Free-energy landscapes of RfaH over Q_{diff}.

The free energy landscapes of isolated CTD (left) or full-length protein (right) were projected onto the Q_{diff} reaction coordinate alone, which describes the transition between α-folded and β-folded CTD.

In the free energy surfaces for the full-length RfaH protein only one transition is observed at its T_f. By analyzing its free energy barriers, it can be noted that a transition occurs between Q_{diff} 0.7 and 0.9, with a barrier of 4.4 kcal/mol. Closer inspection of the structural characteristics of this second minimum show that it has a RMSD of around 2.5 Å to αCTD, indicating that the cooperative decrease in Q_{diff} is explained by the dissociation and partial rupture of the αCTD. The second energy barrier observed separates the folded state from the unfolded configurations and has a similar energy of 4.6 kcal/mol. The free energy basin for βRfaH is not observed at this temperature.

Altogether, these results recapitulate the experimentally predominant folded state for each simulation system in solution, which is separated by a significant energy gap from their alternative native states. Our results also show that both folded states of RfaH are connected by the unfolded state as well as by a hypothetical three-strand intermediate observed in the simulations for the isolated CTD, thus proposing the following fold-switching mechanism (Equation 20):

$$\text{(20)} \qquad \alpha\text{RfaH} \rightleftarrows \text{Unfolded} \rightleftarrows \beta\text{-intermediate} \rightleftarrows \beta\text{RfaH}$$

The NTD of RfaH strongly stabilizes the α-fold and hinders proper βRfaH refolding. One disadvantage of the umbrella sampling simulations is that it directly employs the number of native contacts of the system in αRfaH and βRfaH as collective variables to drive the structural interconversion of RfaH. Then, it becomes difficult to calculate the likelihood of other configurations that, albeit having a significant number of native contacts, may also display an important number of non-native contacts that could be relevant for its stabilization. Consequently, one is unable to directly evaluate, for example, how the appropriate binding configuration between the NTD and CTD is guided by sequence features in RfaH.

AWSEM allows to restrict the use of structural biases only towards local-in-sequence interactions by using the fragment memory potential that limits the configurational exploration of short segments of the protein to those of a reference structure (Davtyan et al., 2012). By not providing information about contacts between the NTD and CTD, these simulations freely explore the interdomain interaction landscape. A similar simulation strategy has been previously employed to correctly predict binding interfaces of both homodimers and heterodimers (Zheng, Schafer, Davtyan, Papoian, & Wolynes, 2012).

Using a temperature gradient through long MD simulations ($3 \cdot 10^7$ timesteps of 5 fs, compared to previously reported folding annealing simulations of $4 \cdot 10^6$ timesteps (Davtyan et al., 2012) and $6 \cdot 10^6$ timesteps (Tsai et al., 2016)), 100 models with fragment memory to a single reference structure were allowed to refold starting from random unfolded conformations ($Q_W < 0.1$). In these single-memory models, only the NTD and CTD of RfaH, but not the linker connecting both domains, were given memory, and these memories are withdrawn from a single reference structure, either αRfaH or βRfaH. This approach leaves the linker that connects both domains with a major conformational freedom and results in the C- and N-terminal domains being structurally uncoupled, as the 10-residue long connector that exist between them is not part of the structural bias and therefore disrupts memory continuity. Therefore, any interdomain interaction formed in these simulations is the result of stabilizing residue-residue contacts encoded by the transferable part of the AWSEM force field, and not due to fragment memory or any other external potentials to favor its exploration. Using this approach, we simulated the refolding of αRfaH and calculated the amount of native tertiary contacts reached at the end of the simulation (Figure 17, Table 4).

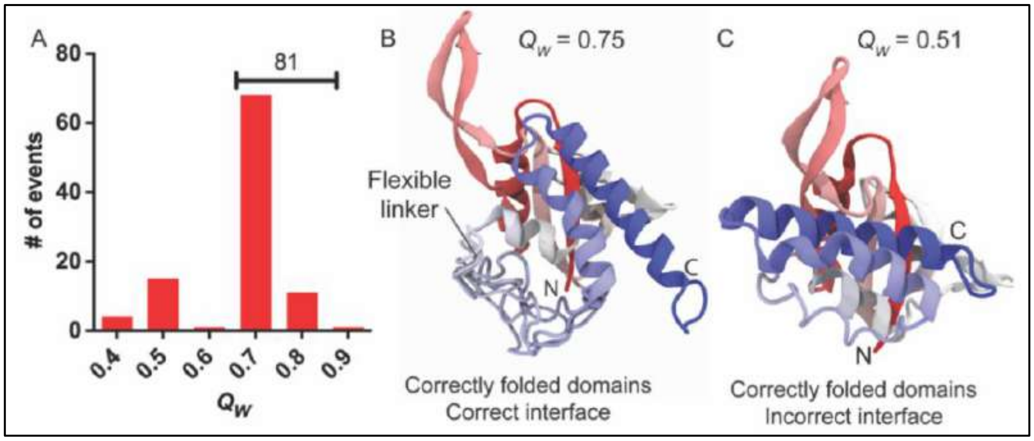


Figure 17: Refolding efficiency of αRfaH.

(*A*) Distribution of tertiary contacts (Q_W) in the final structure of the 100 refolding simulations generated for αRfaH using a single memory. (*B*-*C*) Representative final structures after αRfaH refolding with high (*B*) and low (*C*) Q_W respectively. The images are colored in gradient from red (N-terminus) to blue (C-terminus).

Refolding simulations employing αRfaH as the single memory reference structure show that 81% of the trajectories reach the native state (Q_W = 0.75, Figure 17A). These predicted structures are characterized by the proper orientation and binding of the αCTD against the NTD (Figure 17B), recapitulating the experimentally solved structure of RfaH in its autoinhibited state (Belogurov et al., 2007), and is compatible with the observation that the full-length protein successfully refolds to this state on its own (Tomar et al., 2013). This specificity is achieved despite the lack of structural biases on the interdomain interface and linker regions, and thus a result of sequence determinants in both the NTD and CTD of RfaH encoding this behavior. In fact, the linker is not stabilized in a particular conformation and does not form stable contacts with any domain. In all other trajectories the interdomain interface is formed incorrectly, although both the NTD and αCTD reach their native conformations mostly due to the fragment memory bias. Observation of the refolding traces show that the αCTD is only stabilized upon or after NTD folding, suggesting that the NTD is responsible for the stabilization and orientation of the αCTD.

To further assess the effect of the NTD hydrophobic patch in CTD folding, the same refolding experiment was performed for βRfaH extracted from the cryo-EM structure. To enlighten the effect that the NTD could have on βCTD refolding, the resulting structures are compared with equivalent refolding of the solved structure of the isolated βCTD. The results of βRfaH and βCTD refolding experiments are summarized in Figure 18 and Table 4.

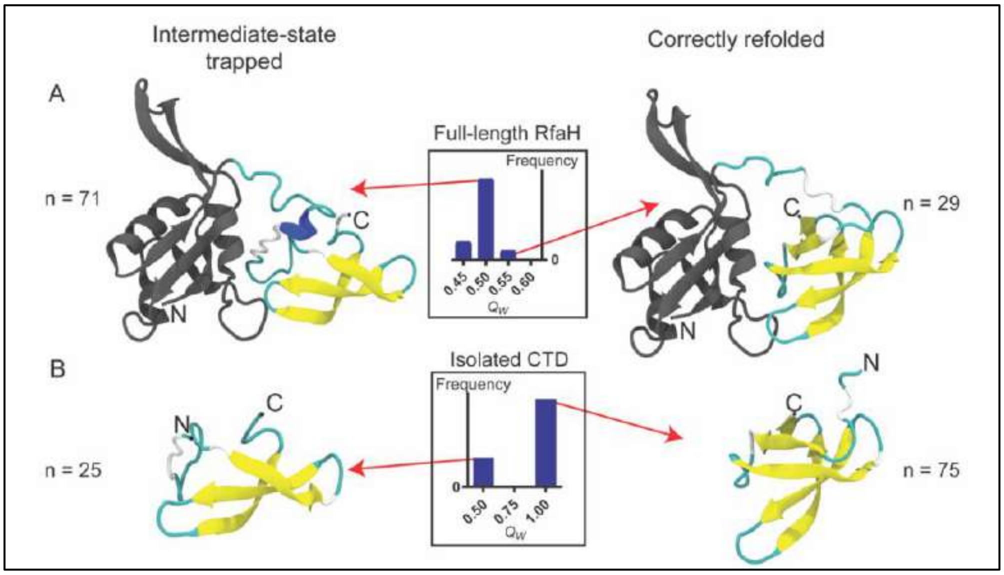


Figure 18: Refolding of βCTD in the context of the full-length protein and in isolation.
Representative final structures after βCTD refolding in the context of the full-length protein (*A*) and in isolation (*B*). The histograms represent the Q_W distribution of the final structures. The intermediate state is formed by the three largest β-strands that form the CTD barrel, namely strands β2, β3 and β4.

For the isolated domain, the βCTD refolds with a similar efficiency than αRfaH (75%), with the remainder of the simulations reaching an intermediate state characterized by a lower Q_W, in which only the three larger β-strands of the barrel are folded (Figure 18B). In stark contrast, the presence of RfaH NTD reduces the CTD refolding efficiency to only 29%, whereas all other refolding trajectories become trapped in the same β-intermediate observed for the isolated βCTD (Figure 18A). These results suggest that the stabilization of this intermediate is a result of specific NTD-CTD interactions established during the folding process of βRfaH.

To determine that the βCTD intermediate is stabilized by specific interactions between both RfaH domains, a harmonic potential was used to maintain the NTD and CTD domains away from each other during refolding simulations of βRfaH. Upon keeping both domains apart throughout the simulation, the βCTD mostly refolds as if it was isolated, with 66% of cases achieving complete refolding (Table 4). Also, two additional systems were used for refolding simulations: i) the isolated CTD of NusG, a protein that shares almost identical secondary and tertiary structure but lacks any observable metamorphic feature, and ii) a chimeric protein connecting the NTD of RfaH with the CTD of NusG, in which it is expected that no specific interdomain interactions are formed given the divergent evolution of RfaH and NusG (B. Wang, Gumerov, Andrianova, Zhulin, & Artsimovitch, 2020). Remarkably, when the isolated CTD of NusG and its fusion to RfaH NTD were used as input for refolding simulations, the totality of the simulations reached the β-folded state of NusG CTD, regardless of the presence of the NTD (Figure 19, Table 4).

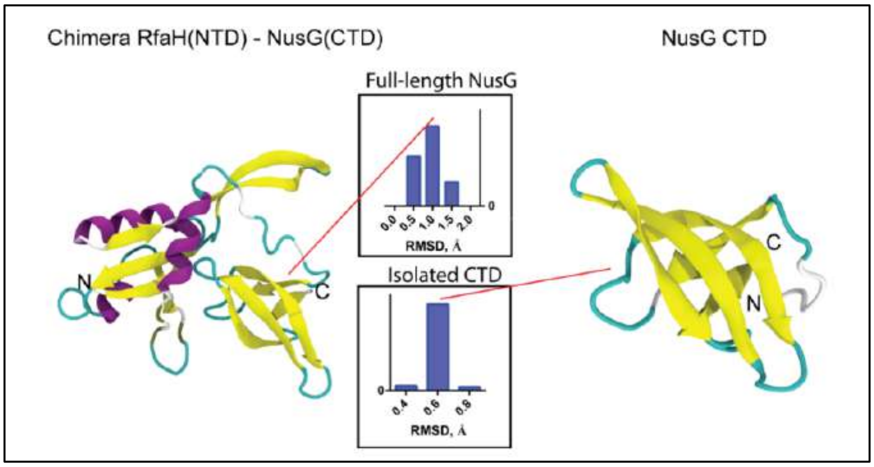


Figure 19: Refolding of NusG βCTD alone and its fusion to RfaH NTD.

Representative final structures after NusG βCTD refolding in a RfaH NTD – NusG CTD chimera and in isolation. The histograms represent the RMSD distribution of the final structures. All simulations reached the β-folded state of NusG CTD.

Although NusG CTD also traverses through a three-strand intermediate state during refolding, it does not become trapped in this configuration as it does the βCTD of RfaH. Altogether, these data strongly suggest that an interruption in the β-barrel folding process is caused by specific interactions established between RfaH domains (Figure 20).

To understand what interactions are arising between the βCTD of RfaH and its NTD, the majority cluster of the intermediate-folded βRfaH was analyzed in more detail to gain insights into what interactions stabilize this state (Figure 21A). A C_α contact map with a threshold of 9.5 Å was calculated for the interaction between the β-intermediate and NTD, as well as αCTD and NTD. In this map, three distinct interaction regions between the βCTD intermediate and the NTD were identified (Figure 21B). Among these, one set comprises native contacts found in the α-fold, corresponding to residues that form the helix α_2 of αCTD, or the loop between strands β_3-β_4 in the βCTD. Apart from this, the region comprising strand β_1 (residues 114-123) contains most contacts with the NTD, all of which are absent in the autoinhibited state of RfaH.

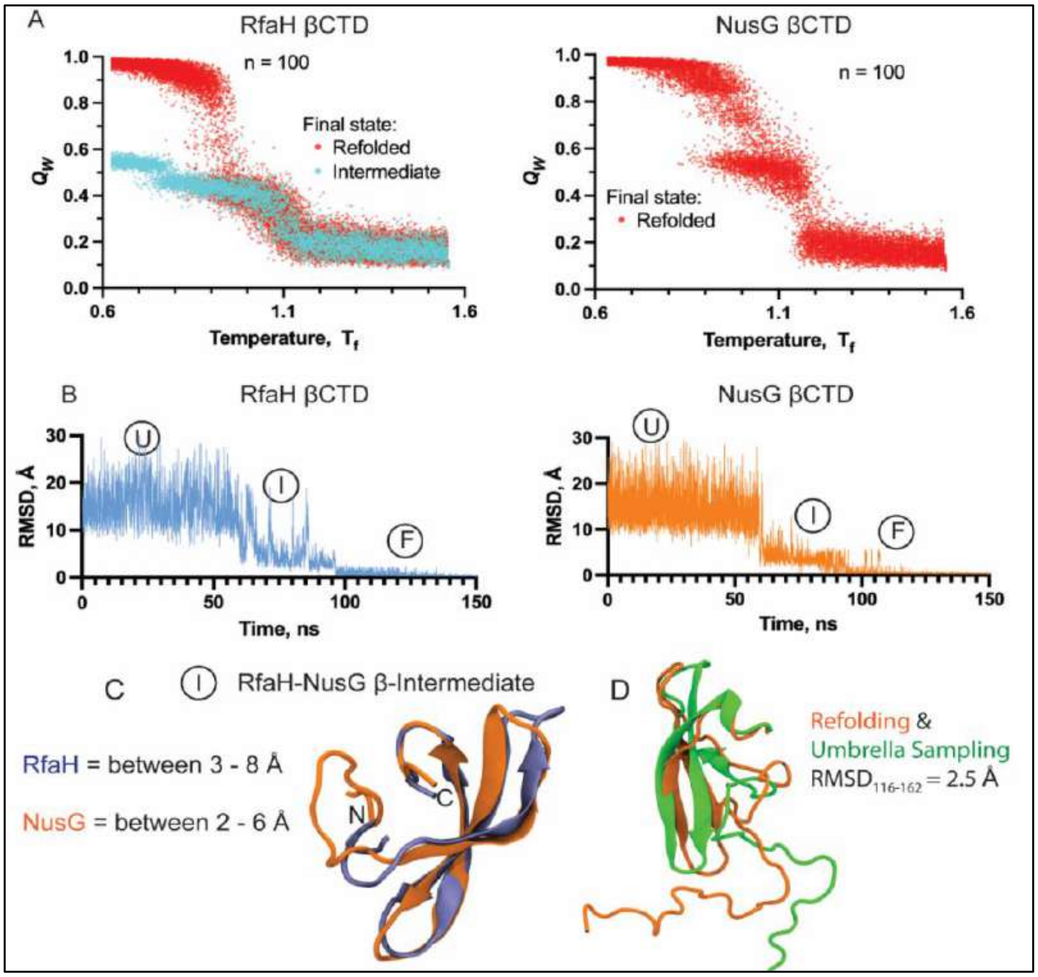


Figure 20: The intermediate of the RfaH CTD is the same for NusG CTD.

(*A*) Annealing plots of RfaH CTD and NusG CTD. Each point was taken every 2,000 steps of $3 \cdot 10^7$ step trajectories that ramped down from 1.6 T_f to 0.6 T_f. For both RfaH and NusG, an intermediate is observed at $0.4 \leq Q_W \leq 0.6$. (*B*) Comparison of refolding traces and intermediate structures of RfaH CTD and NusG CTD. The folding states of both traces was visually inspected. For each trace, the unfolded state is denoted as U, while the intermediate state is denoted as I and the folded state is denoted as F. (*C*) Structural alignment via STAMP of the intermediate states observed for RfaH and NusG and the RMSD to the folded state for RfaH and NusG. (*D*) Structural alignment via STAMP of the β-intermediate state observed for RfaH CTD in umbrella sampling and refolding simulations.

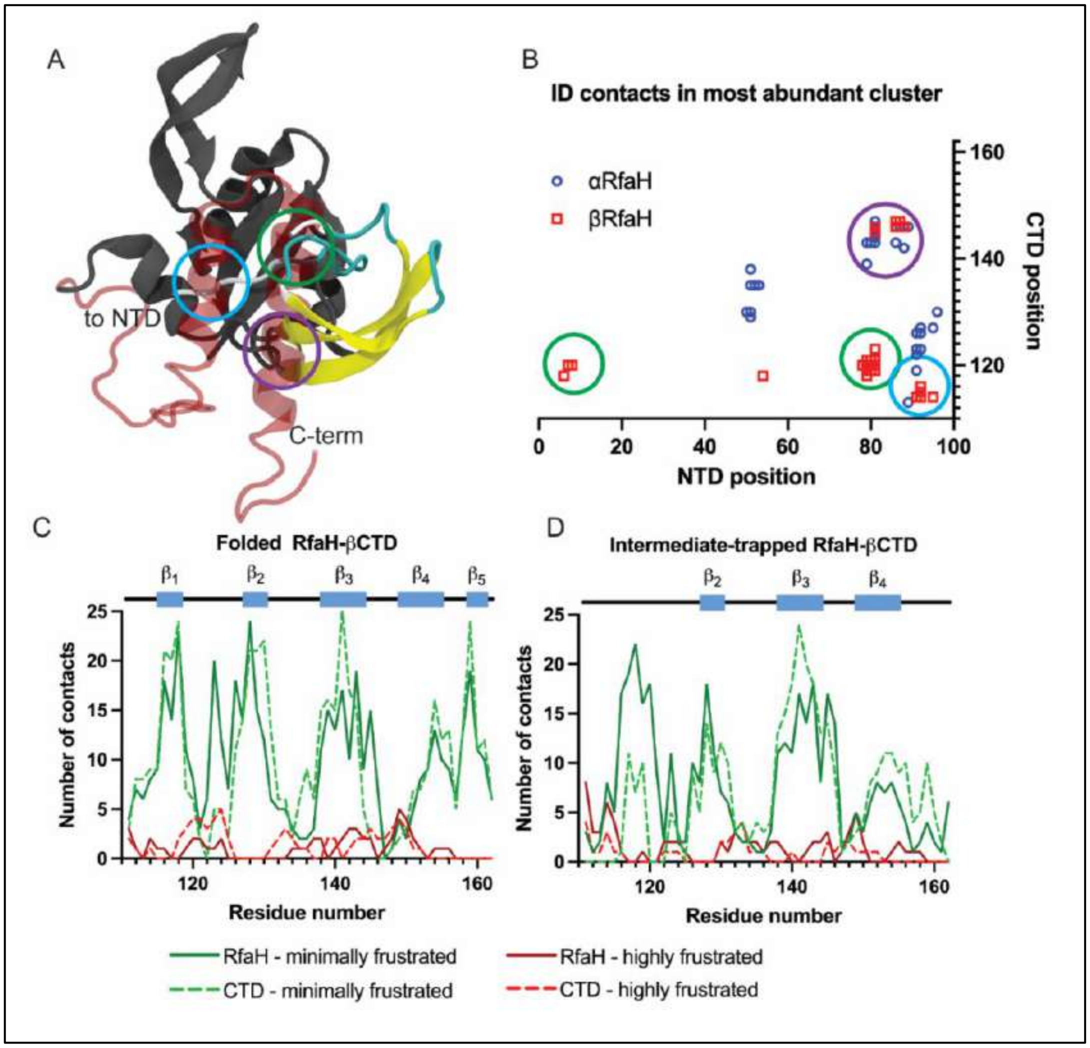


Figure 21: Contact and frustration analysis of the RfaH β-intermediate and its interaction with the NTD. (*A*) Superimposed structures of the αCTD (diffuse, red) and β-intermediate (yellow) on the aligned NTD (gray). The three major points of contacts are circled in different colors. (*B*) Contact map of the interdomain interface observed in αRfaH (blue) and in the β-intermediate (red). The number of highly (red) and minimally frustrated (green) contacts is shown for the CTD in isolation (dashed line) and in the context of full-length RfaH (solid line) for the completely folded CTD (*C*) or the β-intermediate (*D*).

To get further information of the nature of the interactions established between the CTD and NTD, we calculated the per-residue tertiary contacts that are minimally or highly frustrated using the protein frustratometer (Parra et al., 2016). For this end, the representative structure of the most populated cluster of the intermediate-trapped or completely refolded βCTD, both in isolation and in the context of full-length RfaH, were analyzed using the web version of the protein frustratometer (http://frustratometer.qb.fcen.uba.ar) (Figure 21C and D).

When the CTD is successfully refolded, most of the minimally and highly frustrated contacts in the CTD residues are the same throughout this domain, except for residues 123, 145 and 130. Residues 123 and 145 show an increase of more than 10 minimally frustrated contacts when refolded in the full-length RfaH, whereas residue 130 has more minimally frustrated contacts in the isolated CTD. These sets of residues have been identified to be relevant for the stability of the βCTD in previous simulations using dual-basin structure-based models (Ramírez-Sarmiento et al., 2015) and also for the stability of the autoinhibited state of RfaH in recent NMR experiments of the transformation of RfaH (Zuber et al., 2019). In contrast, the β-barrel intermediate of the CTD forms more minimally frustrated contacts when in the presence of the NTD than in isolation (Figure 21), particularly doubling the number of these type of contacts in the region corresponding to strand β_1 and the loop preceding strand β_2 (residues 114-123). Despite not forming the strand β_1, such region becomes highly stabilized by bridging interactions between the NTD and the β-barrel intermediate and serves as the interface between the two domains.

The non-native, minimally frustrated interactions that stabilize the β-intermediate in the full-length protein are formed against a hydrophobic patch in the NTD, comprising residues 78-82 and 91-93. It is worth noting that these NTD residues are solvent-protected when RfaH is bound to the TEC (Kang et al., 2018). This patch is flanked by a charged and a polar residue, namely H77 and Q95, that are at close distance from two acidic residues of the CTD, E120 and D114. Most of the other CTD residues in between these positions are non-polar and form interactions either with the incipient hydrophobic core of the three-strand intermediate or the hydrophobic patch of the NTD. Of these residues, the only non-polar residue that does not form part of the hydrophobic core in the folded βCTD is I117.

We also observe a decrease in minimally frustrated contacts in strand β_3, β_4 and the C-terminus of the β-intermediate upon binding to the NTD. Upon careful inspection of the contacts taking place in these regions, we noted that the region corresponding to strand β_1 forms a core of contacts with the C-terminus and the three β-strands in the isolated β-intermediate. This core decreases its amount of intradomain contacts when strand β_1 encounters the NTD hydrophobic patch rich in minimally frustrated contacts.

Discussion

E. coli RfaH is known as one of the most dramatic examples of protein fold-switching. In solution, RfaH folds into an autoinhibited state in which the αCTD tightly binds to the NTD. This contrasts to the dynamics of its active state, which is only feasible in its full length in the presence of *ops*-stalled TEC (Zuber et al., 2019), in which case both domains dissociate and fluctuate independently. In contrast, the non-metamorphic *E. coli* NusG only transiently forms interdomain interactions, existing always in solution as a protein with two independently moving domains (Burmann et al., 2012, 2011). Our simulations using the AWSEM MD and force field package correctly model RfaH in all its conformations and recapitulate its thermodynamic behavior in solution, evidenced as the switching of the energetic minimum between αCTD and βCTD when breaking interdomain interactions. This switch has also been observed in previous computational works on full-length RfaH using various simulation strategies (Ramírez-Sarmiento et al., 2015; Seifi & Wallin, 2021).

More importantly, our refolding simulations show that the number of trajectories that successfully reach the β-folded CTD in the context of full-length RfaH is a minority when compared to the cases in which the CTD becomes trapped in a three-strand β-barrel intermediate, and almost three times less successful than refolding of αRfaH. We also demonstrate that a significant number of minimally frustrated NTD-CTD interactions, some of which are also observed in the autoinhibited state of RfaH, interfere with proper β-fold formation by stabilizing its intermediate state. These results suggest that the thermodynamic stability of the autoinhibited state of RfaH is not only due to the compatibility between the αCTD and NTD but also due to a selective stabilization of the β-intermediate by the NTD, which increases the probability of the β-barrel being trapped in a three β-strands intermediate. Moreover, while refolding of the CTD of the non-metamorphic RfaH paralog NusG successfully reaches the β-folded state, the transient observation of a structurally similar intermediate state also suggests that it is the nature of the NTD and CTD sequence of RfaH that drives the interdomain interaction and ultimate trapping into this state.

Of importance in the refolding process is the configuration that the interdomain linker may take. As it has been previously reported (Xun et al., 2016), including our own research (Galaz-Davison et al., 2020), the linker does play a role in interdomain stability by favoring and stabilizing the αCTD in the hairpin conformation. During our experiments the linker was not given a memory potential, not being stabilized in a particular conformation other than that which arises from the force field for its sequence. We observed the linker to be flexible, not acquiring any degree of secondary structure during our refolding or umbrella sampling simulations. Based on our results and the literature, we hypothesize that αRfaH stabilization by the linker is due to it acting as an entropic spring, i.e., when both domains are close together the linker accesses to a higher number of configurations than when the domains are separated. A similar process may be responsible for allowing the interactions between the β-barrel intermediate and the NTD.

Multiple reports have studied the metamorphic process of RfaH CTD in the context of the isolated domain (Balasco et al., 2015; Bernhardt & Hansmann, 2018; Joseph et al., 2019; Li et al., 2014) and the full-length protein (Gc et al., 2015; Ramírez-Sarmiento et al., 2015; Seifi et al., 2021), but only a few have described the β-intermediate observed here during βCTD refolding. One of such works corresponds to the computational study of the α-to-β transition of the isolated CTD of RfaH through targeted MD and Markov state models using an adaptive seeding method, in which several en-route ensembles collectively suggests that strands β2, β3 and β4 are relatively stable and form earlier during refolding towards the β-state (Li et al., 2014). Additionally, our previous work with full-length RfaH using dual-basin structure-based models also identified a βCTD-like intermediate that is either free or interacting with the NTD, but with a different topology (Ramírez-Sarmiento et al., 2015). Lastly, recent unbiased explicit solvent simulations of the spontaneous α-to-β fold-switch of RfaH CTD using a replica exchange with hybrid tempering method exhibits three-stranded and four-stranded intermediates before reaching the β-folded CTD (Appadurai, Nagesh, & Srivastava, 2021). Nevertheless, none of these works described the active role of the NTD in stabilizing such intermediate state nor characterized its role as part of the β-barrel folding process.

We believe that this three-strand intermediate and its NTD-dependent stabilization has been overlooked due to either the granularity of the model used, the absence of sequence-dependent potentials or the velocity with which the system is being driven out of the equilibrium. In fact, the sequence-dependent potential embedded on AWSEM shows its capabilities when simulating the correct refolding of αRfaH to a high fraction of native contacts Q_W even in the absence of knowledge-based contact information of the interdomain interface and the linker connecting both domains, meaning that these simulations are robust enough to discriminate the interactions arising from RfaH sequence in terms of NTD-CTD association. The observation that NusG CTD, unlike RfaH βCTD, is not affected by RfaH NTD in these simulations is confirmation of the latter.

These arguments, alongside the observation of this intermediate in both NusG and RfaH βCTD folding pathways, also suggest that this intermediate is likely a topological solution to the small β-barrel folding process, which could also be necessary for the transition between the α- and β-folds of RfaH. While our previous work using Hydrogen-Deuterium exchange mass spectrometry show no apparent differences between NusG CTD and RfaH CTD and no indications of intermediate states under native conditions (Galaz-Davison et al., 2020), it is possible that the intermediate state observed here requires the addition of chaotropic agents to favor its abundance. It can be presumed that the destabilization of the native state using such approaches not only would favor the intermediate population but also the unfolded state.

All in all, our simulations indicate that the NTD actively participates in thermodynamically favoring the autoinhibited α-state by properly orienting the αCTD and correctly specifying the interactions occurring upon interdomain interface formation and by switching the equilibrium from the β-folded CTD into a folding intermediate. Such intermediate could be potentially observed by studying the equilibrium unfolding of the isolated CTD, as it was observed here during the refolding process of the isolated CTD of RfaH and NusG as well as part of the metamorphic pathway in full-length RfaH. We also hypothesize that stabilization into the β-intermediate by the NTD is the initial step for RfaH to fold-switch back into the autoinhibited state, as the intermediate states observed through umbrella sampling and temperature annealing are structurally the same, i.e., both have three β-strands and share an RMSD value of 2.5 Å. This idea is compatible with the observation of RfaH stably binding the ribosomal protein S10 through its βCTD when bound to the TEC (Zuber et al., 2019), as in such state the NTD hydrophobic patch is blocked by

RNAP. Therefore, the effect of the NTD over the βCTD can only be observed when releasing the active state of RfaH from the TEC, hence the role of the NTD to fold-switch back into the autoinhibited state.

4. COEVOLUTION-DERIVED NATIVE AND NON-NATIVE CONTACTS DETERMINE THE EMERGENCE OF A NOVEL FOLD IN A UNIVERSALLY CONSERVED FAMILY OF TRANSCRIPTION FACTORS

Introduction

Protein evolution is at the cornerstone of organism adaptation and gain of function. It diversifies proteins into entire families, whose members can branch out into proteins carrying distinct functions than its predecessors. This is the case of RfaH, a transcription and virulence factor in enterobacteria that evolved from a highly conserved family of transcription regulators called NusG (Artsimovitch & Knauer, 2019). This protein family is universally conserved in all domains of life, regulating transcription by directly binding to RNA polymerase (RNAP) (Washburn et al., 2020), and an ancestor of this protein family is thought to have been present in the last universal common ancestor (LUCA) (B. Wang & Artsimovitch, 2021).

In *Escherichia coli*, NusG is an essential protein that regulates virtually all transcription processes (B. Wang & Artsimovitch, 2021). Meanwhile, its RfaH paralog is quite unique as it regulates transcription in a sequence-dependent manner (Bailey et al., 1997). RfaH, unlike NusG, is not directly recruited to RNAP, but to the entire *ops*-paused transcription elongation complex (TEC) (Kang et al., 2018), with the *ops* (*operon polarity suppressor*) corresponding to a DNA sequence commonly found in pathogenicity islands and xenogenes incorporated by Enterobacteria (Artsimovitch & Knauer, 2019).

This striking feature of RfaH is achieved by its three-dimensional structure, which differs from that of the canonical NusG fold. As in NusG, it consists of an N-terminal domain (NTD) comprising a hydrophobic depression that binds to the polymerase, but that in RfaH is blocked by its own C-terminal domain (CTD) folded as an α-helical hairpin, constituting an autoinhibited state (αRfaH). Notably, when RfaH encounters the *ops*-paused TEC, it binds to it and relieves itself from autoinhibition, upon which the released CTD refolds into a β-barrel (βRfaH) in a process that is known as fold-switching (Burmann et al., 2012) (Figure 22).

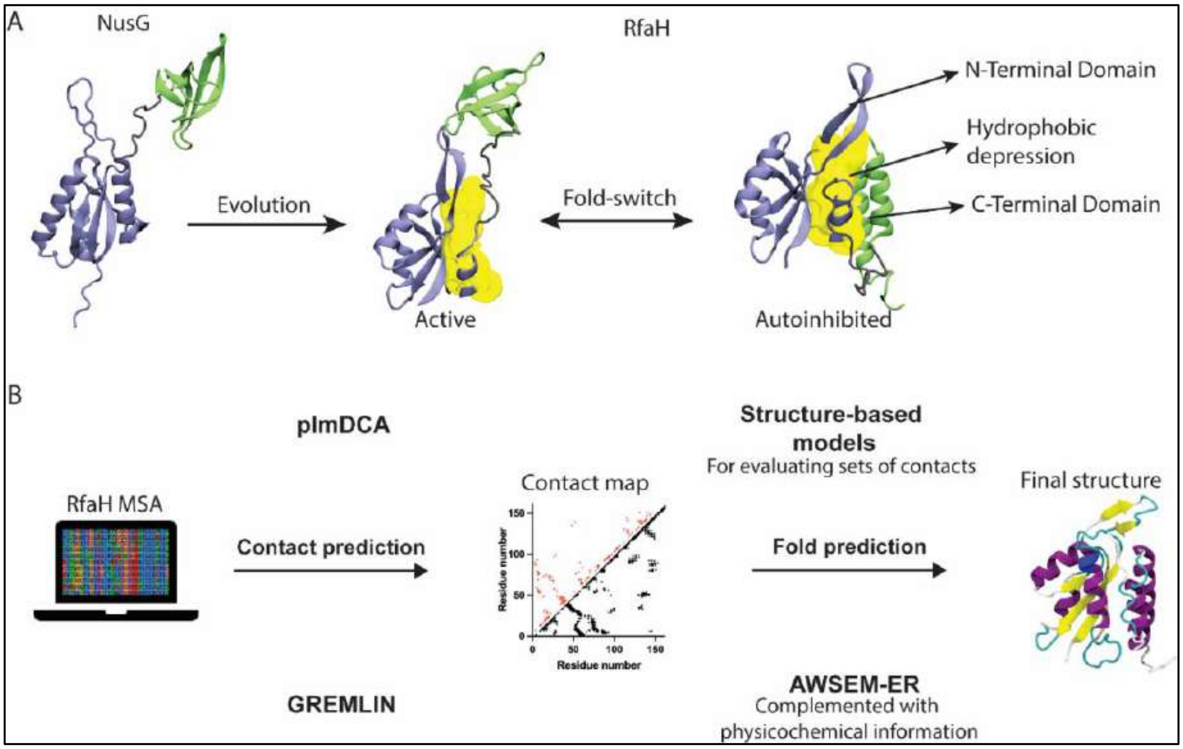


Figure 22: Summary of the research and methods used in this work.

NusG, a non-metamorphic protein, evolved into its paralog RfaH, whose fold-switching is characterized to take place between an autoinhibited fold (αRfaH) that has interdomain contacts at the hydrophobic patch (yellow) and an active fold that does not (βRfaH) (*A*). We hypothesized that the emergence of these interdomain contacts can be inferred via coevolutionary analysis. (*B*) By constructing a metagenomic-enriched multiple sequence alignment (MSA) of RfaH and filtering out non-metamorphic sequences based on secondary structure predictions, we inferred a contact map of coevolving residue pairs that we used to predict the structures of the autoinhibited and active states of RfaH through molecular dynamics using two different pipelines, namely DCA/SBM and GREMLIN/AWSEM-ER, that capture the distinctive features of RfaH folding.

It is estimated that between 0.5 and 4% of proteins whose structures are deposited in the PDB are likely to exhibit fold-switching, or metamorphic, behavior (Porter & Looger, 2018). Among them, RfaH is one of the most studied cases due to its dramatic all α-helix (αCTD) to all β-strand (βCTD) conversion of a whole domain. Computational approaches have sought to determine the fold-switching mechanism of this protein (Balasco et al., 2015; Bernhardt & Hansmann, 2018; Joseph et al., 2019; Li et al., 2014; Ramírez-Sarmiento et al., 2015; Seifi et al., 2021; Xun et al., 2016), and experimental structural work has been performed to characterize its binding to the TEC (Belogurov et al., 2007; Zuber et al., 2019). Recent reports suggest that during evolution, protein metamorphosis emerges as the connecting path between two distinct folds (Tian & Best, 2020). Nevertheless, the fold-switching process of RfaH is key for its function, as it allows RfaH to become active upon highly specific recognition of a DNA sequence while avoiding its spurious binding to RNAP (Zuber et al., 2019).

Experimental and computational approaches have shown that interdomain contacts formed between the CTD and the RNAP-binding interface of the NTD are essential for the formation of the autoinhibited state of RfaH. Particularly, the electrostatic interaction E48-R138 is responsible for stabilizing the α-folded state, as its disruption leads to roughly equally populated βRfaH and αRfaH in solution (Burmann et al., 2012). Furthermore, removal of interdomain contacts in coarse-grained simulations of the full-length protein is enough to give rise to βRfaH (Galaz-Davison et al., 2020; Ramírez-Sarmiento et al., 2015). Consequently, the autoinhibiting interdomain interactions, absent in the RfaH paralog NusG, are essential to stabilize the novel αRfaH fold, giving rise to its structural duality.

We sought to determine if intradomain and interdomain interactions stabilizing both RfaH folds can be inferred from the coevolutionary analysis of their amino acid sequences, and further evaluate their sufficiency to encode both RfaH folds via molecular dynamics (MD) that explicitly incorporate this information. Coevolutionary inference methods, such as direct coupling analysis (DCA) (Morcos et al., 2011) and generative regularized models of proteins (GREMLIN) (S. Ovchinnikov, Kamisetty, & Baker, 2014), have been developed for the statistical analysis of large multiple protein sequence alignments in two essential terms: sequence conservation and correlated mutations. Given that spatially proximate residues in the native state of a given protein family tend to coevolve (Ekeberg, Lövkvist, Lan, Weigt, & Aurell, 2013), these methods have been widely successful in inferring the structural proximity of coevolving residue pairs that are fundamental for folding, function, and dynamics (dos Santos, Jiang, Martínez, & Morcos, 2019; Morcos, Jana, Hwa, & Onuchic, 2013).

In this work, RfaH sequences deposited in the Interpro database (Blum et al., 2021) were used to predict coevolutionary contacts with pyDCA (Zerihun, Pucci, Peter, & Schug, 2020) and GREMLIN (S. Ovchinnikov et al., 2014) (Figure 22). The number of RfaH sequences was further increased by constructing a hidden Markov model (HMM) profile to use as input for a subsequent search of RfaH sequences in the metagenomic database metaclust (Steinegger & Söding, 2018). Finally, in line with recent works (Porter et al., 2022), all sequences were filtered using secondary structure prediction in JPred (Drozdetskiy, Cole, Procter, & Barton, 2015) to select only metamorphic candidates. This metamorphic enrichment protocol yielded an alignment of 3,570 non-redundant sequences that display 4 coevolving pairs of residues involved in interdomain interactions in the experimentally solved structure of αRfaH.

The inferred contacts for RfaH were used as restraints for protein structure prediction in simulations based on coevolutionary structure-based models (Jana, Morcos, & Onuchic, 2014) and coarse-grained force fields (Sirovetz et al., 2017), whose final configurations largely reproduced the experimentally solved structures of both RfaH folds and the NTD-CTD binding of αRfaH. Furthermore, choosing subsets of coevolutionary interactions to guide these simulations led to the observation that contacts between residue pairs not observed in the crystallographic structure of RfaH, i.e., non-native contacts, are important to reach a compact native state having the correct topology and that CTD compactness is essential for forming the autoinhibiting interface of RfaH.

In summary, our results effectively demonstrate that coevolutionary signals encode the metamorphic behavior of RfaH, replicating the distinct features of the active and autoinhibited folds that are essential for the biological function of this transcription factor.

Methods

Sequence search: All initial RfaH sequences were retrieved from the Interpro database of protein families (Blum et al., 2021). The choice of this database for sequence retrieval is due to a recent study that employed this database to characterize the sequence conservation in both RfaH and NusG and further experimentally tested substitutions of these residues *in vitro* (Shi et al., 2017). The retrieved sequences were also used to construct an HMM (Eddy, 2011) profile that was employed to retrieve more RfaH sequences from the metaclust database (Steinegger & Söding, 2018), using e-values for recovery of 10^{-30}, 10^{-25} , and 10^{-20}. Lastly, the sequences from Interpro and metaclust were combined and filtered based on the duality of their secondary structure propensity, similarly to previous works (Porter et al., 2022). To do this, a region of the CTD sequence of each protein (starting from the residue pattern FQAIF, corresponding to residue number 126 in *E. coli* RfaH) of each protein was used as input for secondary structure prediction on the JPred4 server using default settings (Drozdetskiy et al., 2015). Each unaligned sequence that had at least four consecutive helical residues in this trimmed version of the CTD was included in the *Metamorphics* alignment.

Coevolutionary analysis: The three datasets obtained before, *Interpro* with 1,005 sequences, *Interpro + MG* with 5,379 sequences and *Metamorphics* with 3,570 sequences, were used as input for the pyDCA algorithm implemented in Google Colaboratory or submitted at the GREMLIN webserver (gremlin.bakerlab.org). The retrieved residue-pair list was analyzed using a homemade script to calculate the Cα distance at the target PDB file of either αRfaH or βRfaH from PDB ID 5OND or 6C6S, respectively.

Structure-based MD (DCA-SBM): A SBM was generated based on the secondary structure of either RfaH fold and the coevolutionary information obtained by plmDCA, following a protocol already reported (dos Santos et al., 2019). The models for αRfaH with all DCA contacts, only native DCA contacts, all DCA contacts except those of the βCTD, βRfaH with all DCA contacts, βCTD with predicted DCA contacts of

L = 55 and αRfaH with native contacts plus randomly generated non-native contacts equal to the number of non-native coevolving pairs were produced. All these models were run for 2 × 10^7 steps with a timestep 0.0005 τ in reduced units, over which a temperature gradient lowered the temperature from 200 to 0 reduced temperature units. All the simulations were performed with a modified version of GROMACS (Noel et al., 2016).

GREMLIN-AWSEM-ER simulated annealing: Following the AWSEM-ER protocol for the GREMLIN-derived RfaH contacts (Sirovetz et al., 2017), a simulated annealing was produced by decreasing the temperature from 450 to 350 temperature units over 4 × 10^6 steps using a timestep of 5 fs. The default AWSEM forcefield was used, except for the fragment memory potential which was turned off and the evolutionary restraints derived from GREMLIN that were added.

Results

Sequence retrieval and coevolutionary analysis. Retrieving enough sequences to predict robust evolutionary signals is not trivial. Research using the GREMLIN algorithm suggests that a number of non-redundant sequences at least 20 times the length of the protein (L) is needed to achieve a true positive (TP) rate (i.e. coevolving pairs forming a native contact in an experimentally solved structure) of ~0.7 for the top L/2 contact predictions (Kamisetty, Ovchinnikov, & Baker, 2013).

This is an issue for the RfaH subfamily, considering that its most studied representative from *E. coli* is 162 residues long. The predicted members deposited in the Interpro database (Blum et al., 2021) make up nearly 3,000 sequences that, when clustered at 90% identity to reduce redundancy, decreases to ~1,000 sequences. This is less than one third of what is needed according to the criteria above.

Therefore, HMMER (Eddy, 2011) was used to build an HMM profile with the non-redundant Interpro sequences, allowing to search for additional RfaH protein sequences in metaclust, a large metagenomic database of 1.59 billion sequences (Steinegger & Söding, 2018). Using several e-value cutoffs, 3 sets were retrieved, containing 3,865 (e-value = 10^{-30}), 5,378 (e-value = 10^{-25}) and 8,516 (e-value = 10^{-20}) sequences clustered at 90% identity.

Using pyDCA (Zerihun et al., 2020) as a python script in Jupyter Notebooks for execution in Google Colaboratory (Engelberger, Galaz-Davison, Bravo, Rivera, & Ramírez-Sarmiento, 2021), coevolutionary interactions were calculated using pseudo-likelihood maximization direct coupling analysis (plmDCA) on MSAs generated using *hmmalign* from HMMER (Eddy, 2011). Two MSAs were analyzed: one with only non-redundant RfaH sequences from Interpro database (*Interpro*, 1,005 sequences), and another one complemented with the sequences found in metaclust (*Interpro + MG*, 4,853 sequences). Given that GREMLIN has been shown to be more accurate that other coevolution-based residue-residue contact prediction methods (Kamisetty et al., 2013), we also performed our coevolutionary analyses with this algorithm.

Using the Interpro database alone, plmDCA correctly predicts 17 CTD contacts below 8 Å that are formed either in αCTD or βCTD, 1 interdomain (ID) contact below 10 Å and 34 NTD contacts below 8 Å, reaching a fraction of TP (with L = 162) of 0.32, whereas GREMLIN correctly predicts 4 additional contacts for RfaH CTD and 10 additional contacts for the NTD, reaching a TP of 0.40 (Figure 23, Table 5 and Table 6). The choice of 10 Å for ID contacts was due to the observation that the average NTD-CTD distance in randomized residue-residue pairs is 32 Å, in contrast with CTD and NTD contacts that take place at 10-12 Å (Figure 24 and Figure 25).

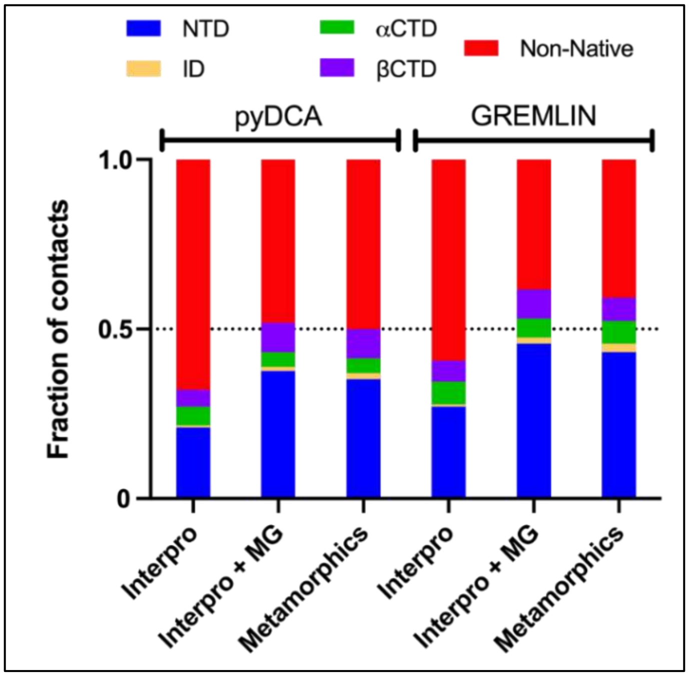


Figure 23: Summary of the coevolutionary analysis results for the RfaH subfamily.

The stacked bar graphs show the fraction of coevolving residue pairs (L = 162) that are found forming a native contact (Cα distance < 8 Å) in the NTD, αCTD, or βCTD; an interdomain (ID) contact (Cα distance < 10 Å) or a non-native contact. The contact distances are calculated based on the crystal structure of full-length αRfaH (PDB 5OND) and the cryo-EM structure of βRfaH (PDB 6C6S).

Table 5: plMDCA results for RfaH subfamily

Dataset	Number of sequences	NTD predicted < 8 Å	CTD predicted < 8 Å	ID predicted < 10 Å	TP
Interpro	1,001	34	17	1	0.32
Interpro + MG	5,379	61	21	2	0.52
Metamorphics	3,570	57	21	3	0.50

Table 6: GREMLIN results for RfaH subfamily

Dataset	Number of sequences	NTD predicted < 8 Å	CTD predicted < 8 Å	ID predicted < 10 Å	TP
Interpro	1,001	44	21	1	0.40
Interpro + MG	5,379	74	23	3	0.62
Metamorphics	3,570	70	22	4	0.59

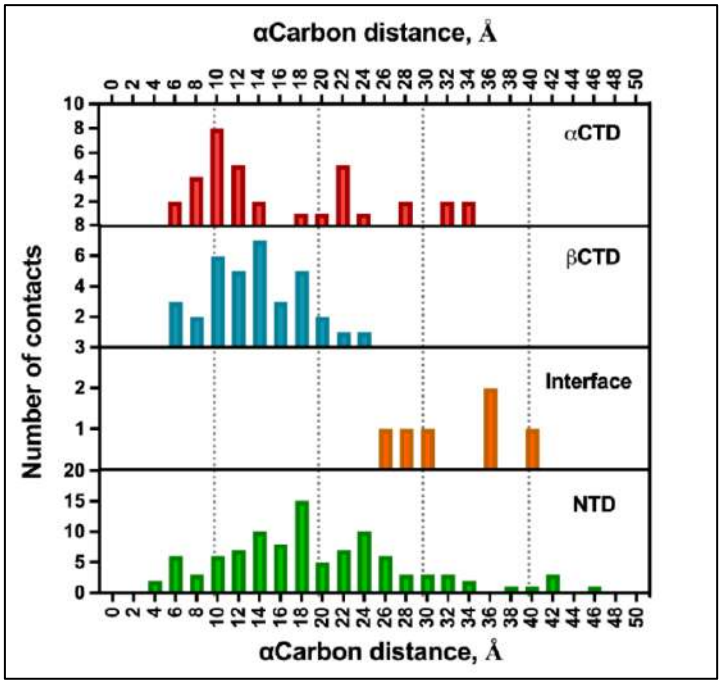


Figure 24: Distance distribution of random contacts inside RfaH.

Random i,j pairs were generated with j > i+2, and their distance was measured in the αRfaH structure (PDB 5OND) or the βCTD (PDB 2LCL) for L = 162 contacts. It is observed that most αCTD contacts take place at around 10 Å, while NTD contacts take place at around 12 Å. This is a feature of the compactness of the domain itself, i.e., the closer the residues are the more likely that two random residues would be at a short distance. In comparison, the distribution for random interdomain contacts taking place at the NTD-CTD interface is centered at a significantly higher distance of ~32 Å. This analysis suggests that native contacts at a distance shorter than 10 Å that are correctly predicted by coevolutionary analysis are significant despite their slightly longer distance than other predictions.

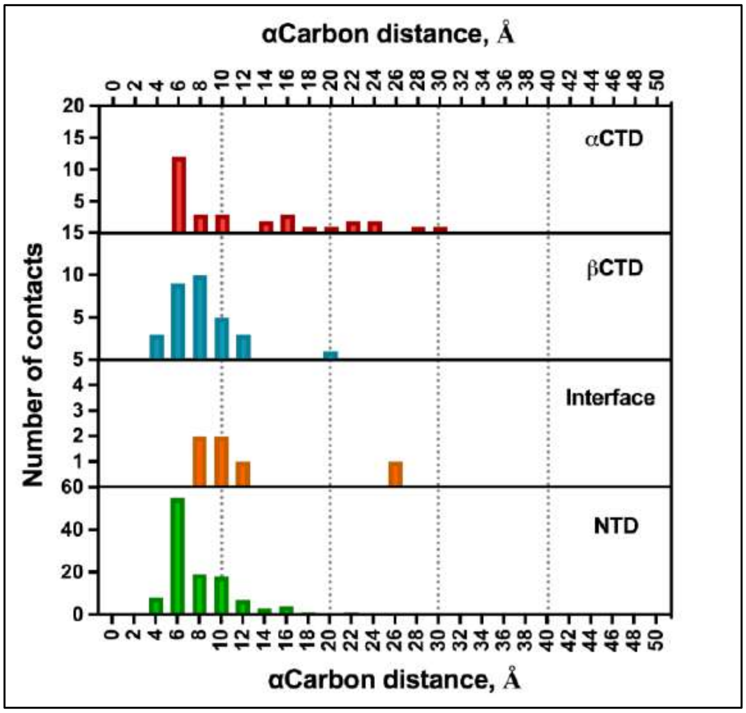


Figure 25: Distance distribution for coevolutionary-derived contact pairs.

The distance between coevolutionary-predicted contact pairs was measured in the crystal structure of αRfaH (PDB 5OND) or βCTD (PDB 2LCL). Most CTD contacts correspond to closely interacting βCTD pairs, whereas αCTD pairs are mostly helical contacts with no interhelical contacts present.

The addition of metagenomic RfaH sequences from metaclust led to an increase in TP up to 0.52 for plmDCA and 0.62 for GREMLIN (Figure 23), including 1 additional ID contact for plmDCA (Table 5) and 2 additional ID contacts for GREMLIN (Table 6). Special attention was paid to the increase in ID contacts due to their relevance in stabilizing the autoinhibited state of RfaH (Galaz-Davison et al., 2020; Ramírez-Sarmiento et al., 2015; Tomar et al., 2013).

In this regard, it is worth noting that although increasing the number of metagenomic sequences beyond those retrieved by HMM search with an e-value of 10^{-30} increases the TP rate for both RfaH folds, i.e. from 0.63 to 0.71 for αRfaH and from 0.70 to 0.82 for βRfaH, there is no increase in the number of predicted ID contacts (Figure 26).

Therefore, we opted to use this MSA to avoid leakage of NusG sequences into our alignment at lower e-values, which could potentially disrupt the coevolutionary signals between the NTD and CTD of RfaH (dos Santos et al., 2019), and to reduce the computing time of coevolutionary contacts through the plmDCA algorithm.

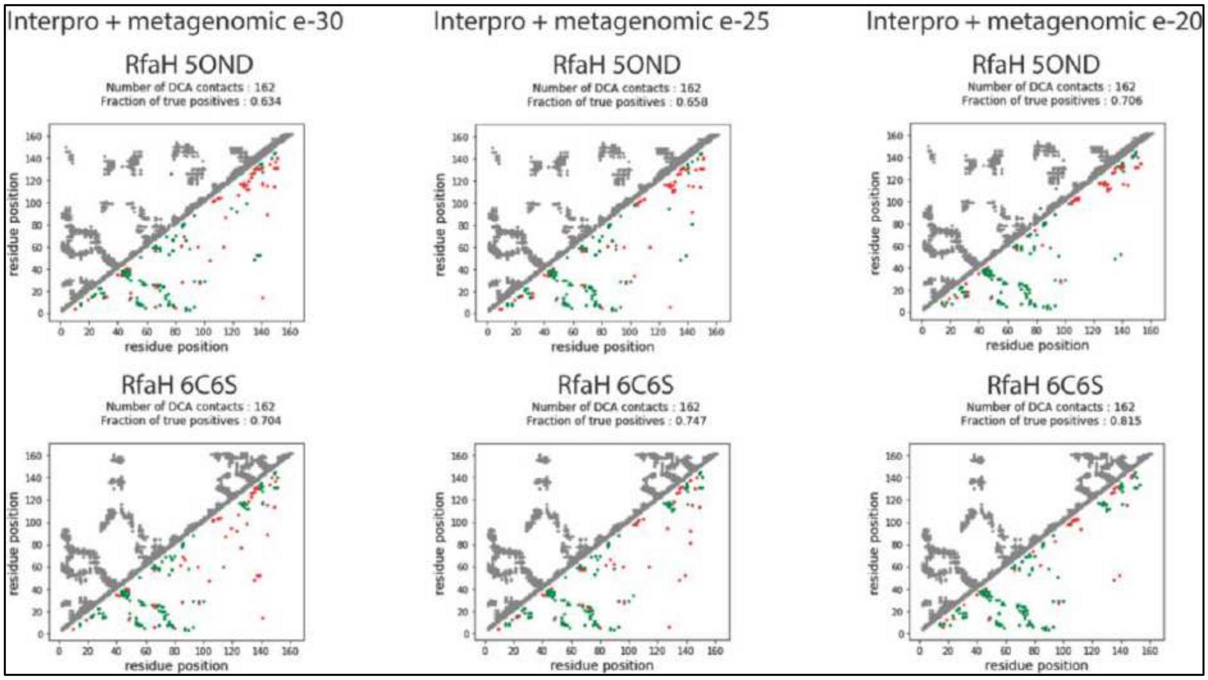


Figure 26: Coevolution-based contact maps of Interpro + MG datasets of RfaH generated with different e-value cutoffs for the retrieval of metagenomic sequences.

The coevolutionary analyses were performed using plmDCA in pyDCA. Upon increasing the number of metagenomic RfaH sequences used as input (i.e., from e-value 10^{-30} to e-value 10^{-20}), an increment in correctly predicted native contacts (green) is observed for the NTD (residues 1-100) but not for the CTD or the interdomain region. Indeed, the amount of false positive interdomain contacts (red) decreases as the number of metagenomic sequences increases. For comparison, the contact maps based on the experimental structures of αRfaH (PDB 5OND) and βRfaH (PDB 6C6S) are shown in grey.

In fact, addition of the metagenomic RfaH sequences led to a 70% increase in the number of correctly predicted NTD contacts whereas the number of CTD contacts remained roughly the same, which could be an indication that non-metamorphic protein sequences have leaked into the alignment, as the NTD fold is highly conserved through the NusG family.

A recent work employed a secondary structure prediction approach to filter out potential non-metamorphic RfaH homologs (Porter et al., 2022). Based on this work, we used JPred (Drozdetskiy et al., 2015) to identify which protein sequences from the *Interpro + MG* MSA exhibit both β-strand and α-helical propensity in a short section of the CTD, comprising residues 126-162 in the representative sequence of *E. coli* RfaH, that reports its metamorphic duality. This filtering process led to a third MSA containing 3,570 sequences clustered at 90% identity (*Metamorphics*), a reduction of ~1,000 sequences from the starting MSA.

Despite this important reduction in the number of sequences, the resulting MSA almost completely replicates the coevolutionary information and TP obtained using either plmDCA or GREMLIN on the *Interpro + MG* MSA while correctly predicting an additional ID contact (Figure 23, Table 5 and Table 6). Altogether, our best TP for L contacts (L = 162) with the highest number of ID contacts was achieved with the *Metamorphics* MSA and GREMLIN, obtaining 70 contacts for the NTD, 22 for the CTD in either fold, and 4 ID contacts (Figure 23 and Table 6). In comparison, a recent work on coevolutionary analysis on RfaH using EVcouplings (Hopf et al., 2019) on sequences collected using iterative BLAST (Altschul et al., 1997) and filtered by secondary structure propensity with JPred (Drozdetskiy et al., 2015) led to the prediction of CTD and NTD contacts but did not report any correctly predicted ID contacts (Porter et al., 2022).

To rationalize the successful prediction of ID contacts in our coevolutionary analysis, we took into consideration the differences between pyDCA and GREMLIN and the increase and identity of the ID contacts predicted upon addition of metagenomic RfaH sequences and secondary structure filtering (Table 7).

Table 7: TP contacts found for each dataset

Dataset	**pyDCA j > i+2**	**pyDCA j > i+3**	**GREMLIN**
Interpro	V92 – I146	V92 – I146	P52 – S139
Interpro + MG	P52 – S139, E48 – G135	P52 – S139, E48 – G135	P52 – S139, P52 – A137, E48 – G135
Metamorphics	P52 – S139, E48 – G135	P52 – S139, P52 – A137, E48 – G135	P52 – S139, P52 – A137, N53 – S139, E48 – G135

One relevant difference between both coevolution-based methods is that pyDCA recommends using a sequence separation for residue pairs of *j > i+3*, while the sequence separation used in GREMLIN is *j > i+2*. Setting the residue separation for plmDCA at *j > i+2* show that most predicted contacts are false positives or short ranged (Figure 27). Despite this observation, 1 correctly predicted ID contact (residue pair 92-146) was retrieved from *Interpro* MSA and 2 from *Interpro* + MG and *Metamorphics* MSAs, corresponding to residue pairs 48-135 and 52-139 that are also obtained using GREMLIN. Thus, filtering out short-range contacts at low sequence separation is required for predicting long-range ID contacts in pyDCA.

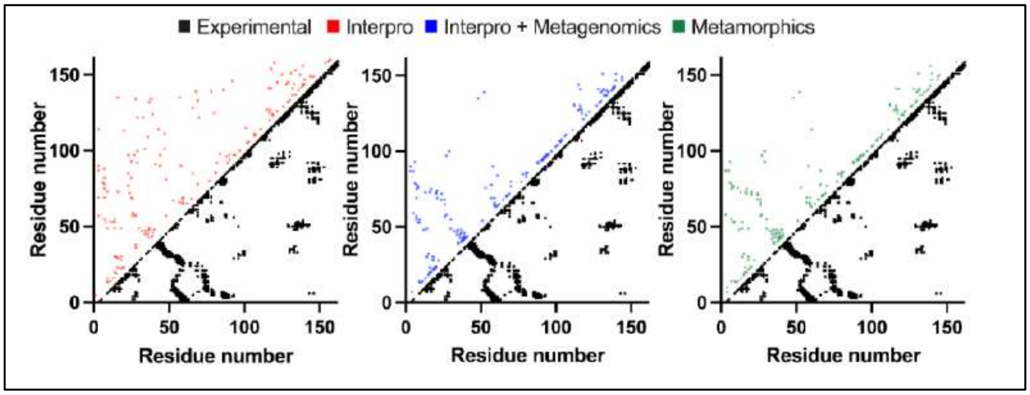


Figure 27: Experimental and predicted contact maps using pyDCA and a sequence separation of j > i+2. It can be clearly seen that the coevolution-based contact map generated with the Interpro dataset (red) has well-distributed but mostly erroneous contacts when compared to the experimental contact map for αRfaH (PDB 5OND). Conversely, coevolutionary analysis of the Interpro + Metagenomics and Metamorphics datasets predicts contacts that closely resemble the experimental features, but most contacts occur at very short sequence separations, and thus are not fully informative of the tertiary structure of the protein.

Analysis of the *Interpro* MSA using plmDCA led to the correct prediction of ID residue pair 92-146. The addition of metagenomic sequences led to the disappearance of the previous residue pair and the correct prediction of ID residue pairs 48-135 and 52-139. Lastly, upon filtering the sequences via Jpred, one additional ID contact is correctly predicted for residue pairs 52-137 (Table 7). This analysis suggests that some of the ID contacts that are likely important for αRfaH may have low DCA scores because they are buried in the dominant coevolutionary signals of the canonical βCTD found in both metamorphic and non-metamorphic NusG family members.

Besides the increase in ID contacts upon enriching the number of RfaH sequences and their subsequent filtering based on secondary structure predictions, we were also interested on the intradomain contacts predicted by these methods, as they may be key in stabilizing each fold. For αRfaH, it is consistently observed that both plmDCA and GREMLIN only yield helical contacts, i.e., with sequence separations of 3 or 4 residues, and no interhelical contacts. Meanwhile, coevolutionary contacts in the β-folded CTD are formed between β2-β3, β3-β4 and β4-β5. It is also worth noting that most of the helical contacts inferred for αCTD are exclusive to this fold, i.e., the interacting residue pairs are significantly more separated in distance in the βCTD (Figure 28). Therefore, these findings suggest that the helical propensity of RfaH CTD is encoded within its sequence coevolution, unlike the hairpin formation, which likely results from compaction of this domain against the hydrophobic ID surface in the NTD.

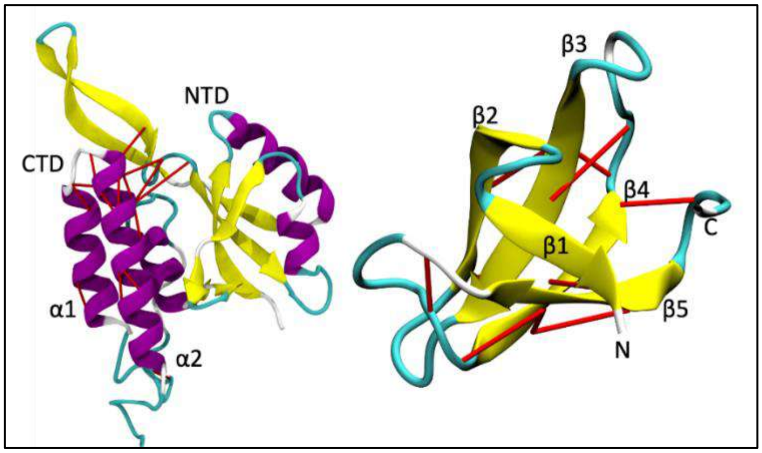


Figure 28: Coevolutionary-predicted intradomain and interdomain contacts in RfaH CTD using the *Metamorphics* MSA and the GREMLIN algorithm.

The structure of full-length αRfaH (PDB 5OND) is shown alongside the isolated RfaH βCTD (PDB 2LCL) in cartoon representation, with strands colored in yellow and helices colored in purple. The coevolutionary contacts predicted for the CTD in each fold are shown as red lines, illustrating the lack of intrahelical contacts for the αCTD in full-length RfaH.

For both plmDCA and GREMLIN, there are about 40-50% coevolutionary signals that do not correspond to any known contact in αRfaH or βRfaH when using the ~3,500 sequences of the *metamorphics* MSA. These apparent non-native interactions are all significant and contribute to a large fraction of the total predicted interactions. To assess if the same rate of false positive contacts is observed in the non-metamorphic NusG protein family members, the NusG sequences deposited in Pfam (Mistry et al., 2021) were clustered at 90% identity (10,593 sequences), aligned and used as input for plmDCA (Figure 29). The results show that out of the top L = 181 residues, 157 contacts are correctly predicted, representing a TP of 0.86, which is higher than the TP rate observed for αRfaH. However, this TP rate is similar to the one obtained for βRfaH (0.82) using the Interpro and metagenomic sequences retrieved at e-value 10^{-20} (8,516 sequences, Figure 26). These results suggest that it would be required to further increase the number of true metamorphic RfaH sequences to overwhelm the coevolutionary signals coming from βRfaH or non-metamorphic homologs.

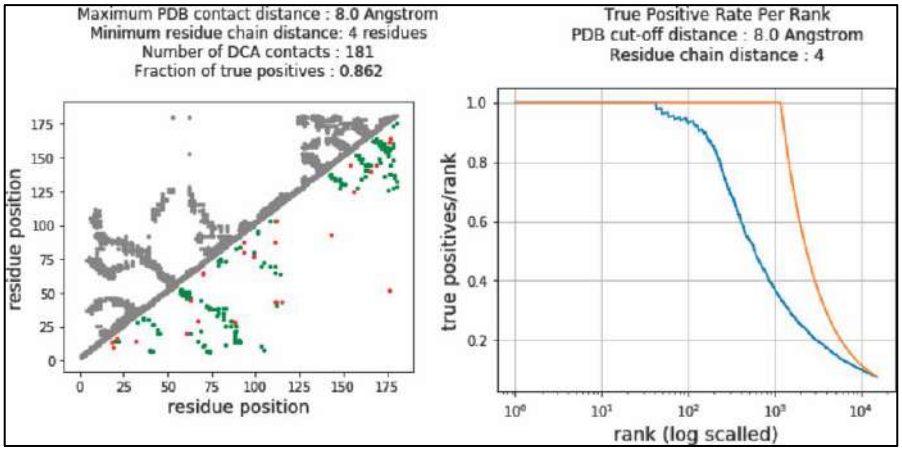


Figure 29: Coevolution-based contact map and true positive rate per rank for the predicted NusG 181 contacts.

The full-length NusG structure was derived from the ColabFold implementation of AlphaFold2 (colabfold.com), which closely resembles the experimental structures of NusG NTD (PDB 2K06) and CTD (PDB 2JVV). On the left, the contact map for the NusG structure is shown in grey, and the coevolution-based predicted true positive and false positive contacts are shown in green and red, respectively. On the right, the plot shows the change in true positive rate per rank as a function of the rank position of each predicted contact in blue, and the theoretically maximum possible true positive rate in orange.

Structure prediction through coevolution-based MD. To determine if the coevolutionary signals identified for RfaH are enough to correctly predict the αRfaH fold, which is the novel topology in the NusG family, two MD pipelines were used: Cα coarse-grained simulations using structure-based models (SBM) guided by DCA-predicted RfaH contacts (dos Santos et al., 2019), and Cβ coarse-grained semi-empirical simulations using AWSEM-ER guided by GREMLIN-predicted RfaH contacts (Sirovetz et al., 2017). Briefly, each coevolutionary contact is used as a bias to guide the MD simulation to form such contact in a simulated annealing, i.e., a descending temperature gradient that allows the formation of the selected contacts during the protein folding. These coevolution-guided MD simulations also rely on secondary structure biases for higher accuracy. Thus, we employed the secondary structure observed in PDB 5OND and 6C6S to model αRfaH and βRfaH, respectively.

A total of 10 simulations for αRfaH and βRfaH were produced for each pipeline. Regardless of the MD pipeline and the low number of ID contacts obtained with either coevolutionary analysis, most of the final configurations of αRfaH after simulated annealing exhibit the formation of an incipient NTD-CTD interaction in the same location of the hydrophobic depression of the NTD (Figure 30 and Figure 31). For the best predicted structure, the RMSD of the NTD and CTD against αRfaH reached 4.2 and 4.1 Å, respectively.

Compaction of the CTD allows for the formation of an incipient NTD-CTD interaction in 80% of the simulations with the GREMLIN-AWSEM-ER pipeline and 70% for the DCA-SBM pipeline. However, the hairpin αCTD structure is only achieved in a few cases, supporting the idea that NTD-CTD binding can occur even in the absence of an αCTD with all native intrahelical contacts formed, as observed in previous simulations using dual-basin SBM (Ramírez-Sarmiento et al., 2015) and experiments using hydrogen-deuterium exchange mass spectrometry (Galaz-Davison et al., 2020), which would likely give rise to RfaH autoinhibition.

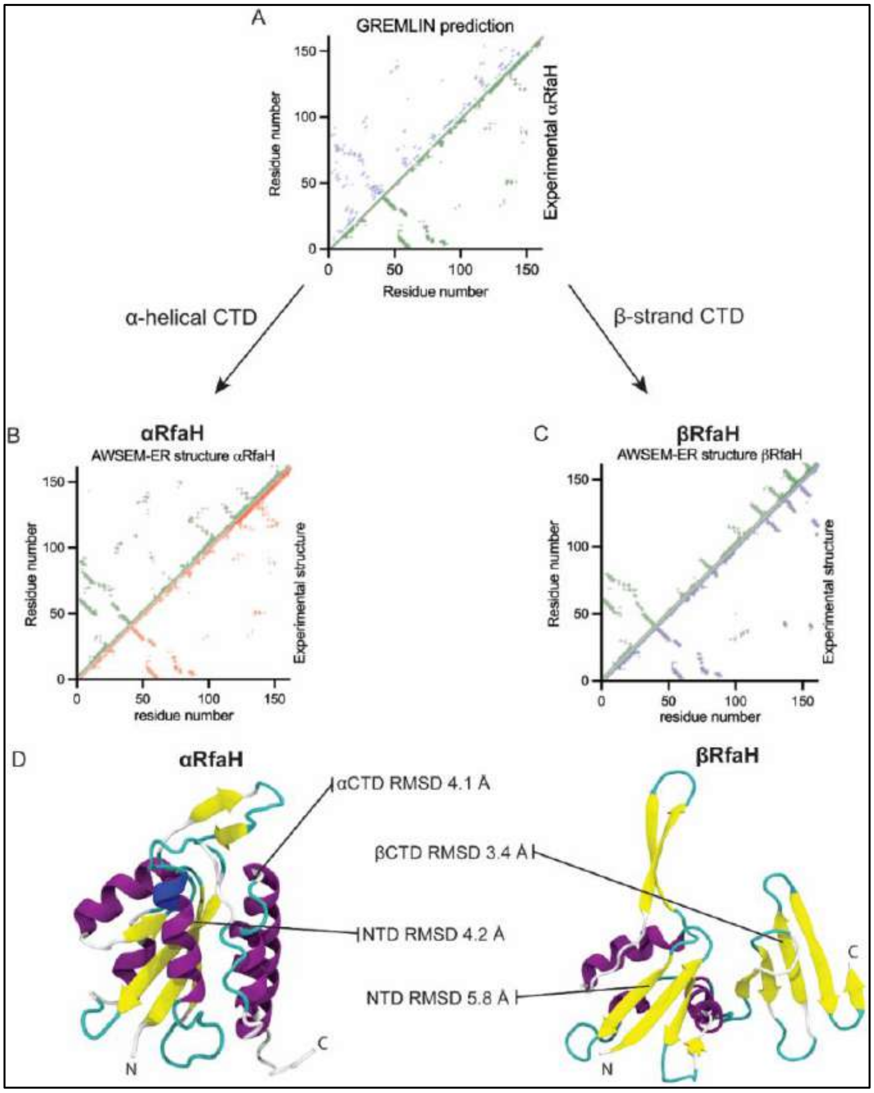


Figure 30: Best predicted structures for αRfaH and βRfaH based on the GREMLIN-AWSEM-ER pipeline. (*A*) Comparison of the coevolution-based and the experimental contact maps of αRfaH (PDB 5OND). (*B*, *C*) Comparison of the contact maps generated using the AWSEM-ER-predicted structures and experimental structures for αRfaH (PDB 5OND, *B*) and βRfaH (PDB 6C6S, *C*). (*D*) Cartoon representation of the best predicted structures for both RfaH folds, and their respective RMSD to the experimental structures.

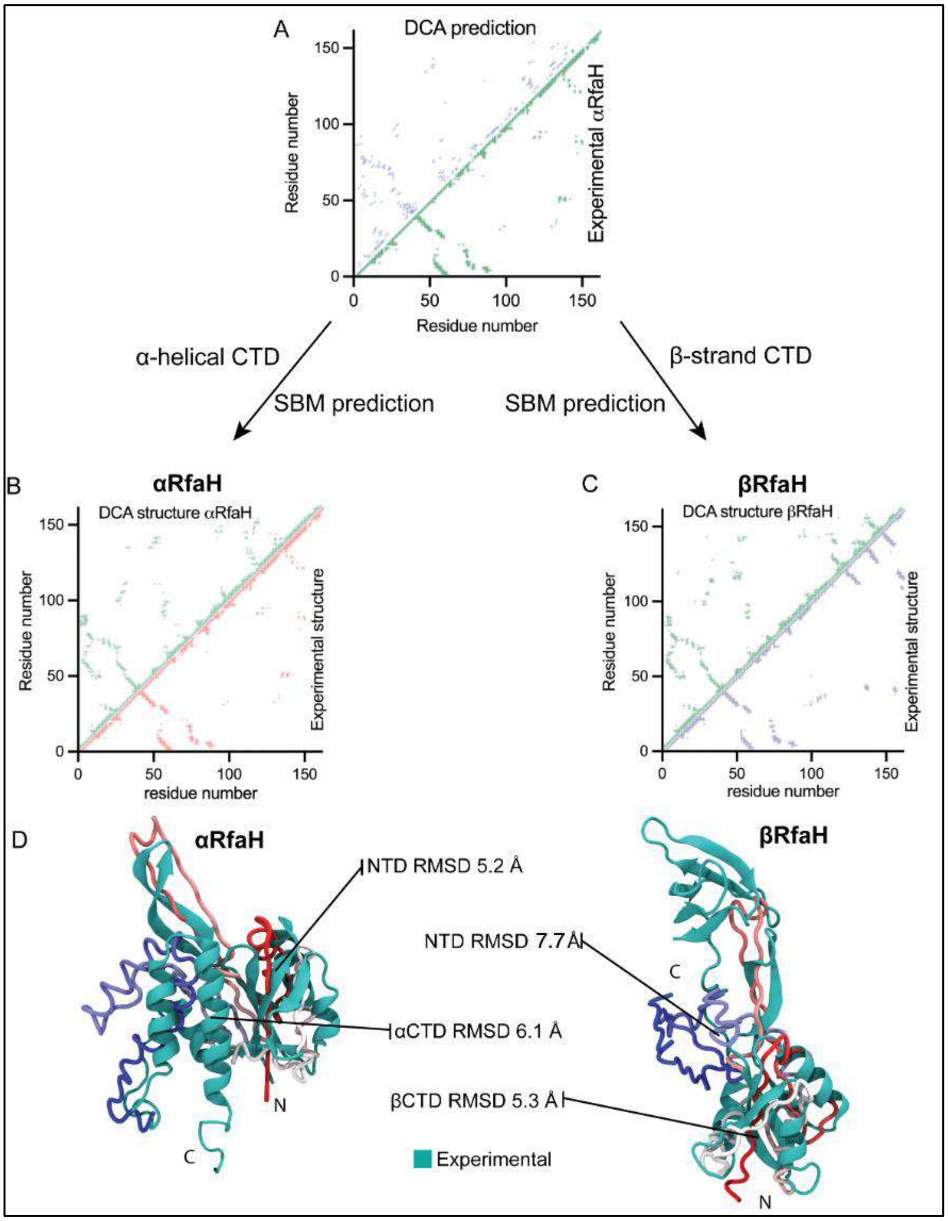


Figure 31: Best predicted structures for αRfaH and βRfaH from DCA-SBM.

(*A*) Comparison of coevolution-based and experimental contact maps of αRfaH (PDB 5OND). (*B, C*) Comparison of contact maps generated using the DCA-SBM predicted structures and experimental structures for αRfaH (PDB 5OND, *B*) and βRfaH (PDB 6C6S, *C*). (*D*) Cartoon representation of the best predicted structures for both RfaH folds and their RMSD to the experimental structures.

In the case of βRfaH, it was observed that the structure of the βCTD is too distorted to be properly folded with either MD pipeline (Table 8 and Table 9), reaching instead the folding state of a three-strand intermediate chaperoned by the NTD, in which only strands β2, β3 and β4 are formed and that has been described through multiple computational approaches in previous works (Appadurai et al., 2021; Galaz-

Davison, Román, & Ramírez-Sarmiento, 2021; Li et al., 2014). Also, the RMSD of the βCTD is higher for the DCA-SBM pipeline. As a control, we compared our DCA-SBM simulations for full-length βRfaH with simulations of the isolated βCTD using coevolutionary contacts predicted only for the 55 C-terminal residues of RfaH (Table 10). While in all simulations of the isolated CTD its RMSD against the β-folded state stayed above 8 Å, its RMSD against the reported β-intermediate reaches down to 4.6 Å. This evidence suggests that βCTD folding is impeded by the idealized fully extended β-strands that are being used as secondary structure bias in these SBM-models, but that its main features are correctly modelled.

Table 8: Summary of simulated annealing results for βRfaH using DCA-SBM

Number	RMSD NTD (Å)	Correct NTD?	RMSD CTD (Å)	NTD bound?	RMSD Intermediate (Å)
1	5.3	yes	9.3	yes	7.2
2	5.4	yes	10.2	yes	7.9
3	5.5	yes	9.1	yes	7.3
4	5.2	yes	7.2	yes	5.3
5	5.3	yes	7.7	yes	6.2
6	10.4	no	6.5	no	5.3
7	10.6	no	6.4	no	5.1
8	5.3	yes	6.7	no	5.3
9	10.9	no	8.3	no	8.0
10	4.9	yes	8.9	yes	7.3
AVG	**6.9 ± 2.6**	**70%**	**8.0 ± 1.3**	**60%**	**6.5 ± 1.1**

Table 9: Summary of simulated annealing results for βRfaH using GREMLIN-AWSEM-ER.

Number	RMSD NTD (Å)	Correct NTD?	RMSD CTD (Å)	NTD bound?	RMSD Intermediate (Å)
1	7.9	no	4.2	no	2.9
2	5.8	yes	3.4	no	2.0
3	8.1	no	5.6	no	3.1
4	9.2	no	4.3	no	3.3
5	11.3	no	5.8	no	3.6
6	6.3	yes	4.6	no	2.6
7	7.3	yes	5.4	no	3.3
8	8.5	no	4.0	yes	1.8
9	12.6	no	3.4	no	1.6
10	6.6	yes	5.6	yes	3.6
AVG	**8.4 ± 2.2**	**40%**	**4.6 ± 0.9**	**20%**	**2.8 ± 0.7**

Table 10: Summary of simulated annealing results for βCTD using DCA-SBM.

Number	RMSD CTD (Å)	RMSD Intermediate (Å)
1	8.7	4.6
2	10.7	6.8
3	10.8	7.1
4	11.1	5.4
5	11.1	7.2
6	11.6	7.5
7	12.2	7.8
8	9.8	5.0
9	9.9	5.3
10	9.6	5.4
AVG	**10.6 ± 1.0**	**6.2 ± 1.2**

For the NTD, we observed that the β2-β3 hairpin, which is largely responsible for the RNAP binding, is highly flexible in the final configurations obtained using the GREMLIN-AWSEM-ER pipeline, particularly in the βRfaH configuration, giving rise to higher RMSD values than in αRfaH. It should be noted, however, that the RMSD between the NTD for both experimental structures of RfaH (PDB 5OND and 6C6S) is 3.0 Å, largely due to structural differences in this hairpin. Regardless, the physicochemical potential of the AWSEM forcefield allows proper compaction of the NTD, thus exhibiting lower RMSD values for the best predicted structure than the DCA-SBM pipeline.

To determine the effect of coevolution-based native and non-native contacts on the accuracy of RfaH structure prediction, we further employed the DCA-SBM pipeline in which only bonded interactions and non-bonded coevolution-based contacts are involved, in contrast to GREMLIN-AWSEM-ER that also incorporates physicochemical and knowledge-based potentials. We first determined the fraction of native and non-native contacts as a function of time for the DCA-SBM simulation that reflects the lower RMSD to the αRfaH fold (Figure 32). We observed that the fraction of native contacts reaches 0.7 in the final native ensemble, as expected for a simulated annealing of a globular protein, whereas the fraction of non-native contacts only reaches, indicating that the non-native interactions inferred by coevolution are not compatible with the native contacts in this structural form.

Next, we chose different subsets of coevolution-based predicted contacts for subsequent DCA-SBM simulations. First, we performed simulations in which we deleted DCA-predicted contacts that exhibit shorter distances in βCTD than in αCTD (Table 11). Second, we performed additional simulations only considering TP native contacts present in the experimental αRfaH structure, effectively disregarding non-native interactions (Table 12). In the first case, we observed that the RMSD to the NTD was similar as the one achieved with the whole set of DCA-predicted contacts, whereas an overall increase in RMSD for the NTD and a less compact global architecture was observed for the second scenario. Regardless, in both cases the CTD was no longer binding the hydrophobic depression at the NTD and it was much less compact when compared with the initial simulations.

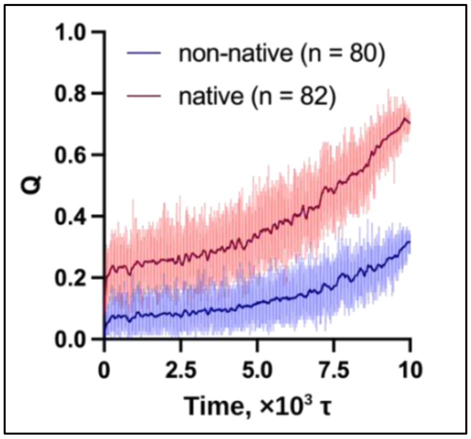


Figure 32: Fraction of native and non-native contacts (Q) as a function of simulation time.

The Cα distances were calculated using GROMACS, and a cutoff of 8 Å for intradomain contacts and 10 Å for interdomain contacts was used for calculating the contacts of native and non-native αRfaH.

Table 11: Summary of simulated annealing results for αRfaH using all DCA contacts except those of the βCTD on DCA-SBM.

Number	RMSD NTD (Å)	Correct NTD?	RMSD CTD (Å)	NTD bound?
1	5.0	yes	12.5	no
2	5.0	yes	11.2	no
3	5.0	yes	15.8	no
4	5.2	yes	15.4	no
5	5.2	yes	15.0	no
6	5.7	yes	14.3	no
7	10.8	no	9.0	no
8	10.9	no	10.2	no
9	10.9	no	10.5	no
10	11.0	no	15.4	no
AVG	**7.5 ± 3.0**	**60%**	**12.9 ± 2.6**	**0%**

Table 12: Summary of simulated annealing results for αRfaH using only true-positive αRfaH contacts on DCA-SBM.

Number	RMSD NTD (Å)	Correct NTD?	RMSD CTD (Å)	NTD bound?
1	6.1	yes	11.6	no
2	7.9	yes	13.8	no
3	8.2	no	11.1	no
4	8.6	yes	14.3	no
5	9.6	no	11.1	no
6	10.7	no	13.6	no
7	11.1	no	13.8	no
8	11.7	no	13.1	no
9	11.9	no	10.5	no
10	12.2	no	12.9	no
AVG	**9.8 ± 2.0**	**30%**	**12.6 ± 1.4**	**0%**

Finally, given that most coevolutionary-predicted native and non-native contacts exhibit shorter interaction distances than randomly generated contacts, we tested if replacing these non-native contacts by an equal number of randomly selected non-native interactions would replicate the compaction caused by the coevolved pairs obtained through DCA. As summarized in Table 13, we observe a similar behavior for the NTD as in the simulations in which only TP contacts for RfaH were considered, except that the topology of this domain was, in most cases, incorrect. For the CTD, we observed compaction of this domain, which is likely caused by the native interactions of the βCTD, but ID interactions occur away from the hydrophobic patch of the NTD probably due to the random nature of the non-native contacts.

Table 13: Summary of simulated annealing results for αRfaH using only true-positive αRfaH contacts and adding randomly generated pairs up to 162 contacts on DCA-SBM.

Number	RMSD NTD (Å)	Correct NTD?	RMSD CTD (Å)	NTD bound?
1	6.1	yes	11.6	no
2	7.9	yes	13.8	no
3	8.2	no	11.1	no
4	8.6	yes	14.3	no
5	9.6	no	11.1	no
6	10.7	no	13.6	no
7	11.1	no	13.8	no
8	11.7	no	13.1	no
9	11.9	no	10.5	no
10	12.2	no	12.9	no
AVG	**9.8 ± 2.0**	**30%**	**12.6 ± 1.4**	**0%**

Altogether, these results suggest that non-native coevolutionary contacts may be important to reach a compact architecture, and that compactness at the CTD may be essential for having clustered interactions that simultaneously bind the NTD at the hydrophobic depression, enabling RfaH the to reach its autoinhibited state.

Discussion

RfaH is one of the most prominent examples of metamorphic proteins, exhibiting fold-switch of an entire domain from an α-helical hairpin to a small β-barrel. Since this protein subfamily is thought to have originated from non-metamorphic NusG transcription factors, its new metamorphic fold should be stabilized by ID interactions emerging during its evolution. To assess such scenario, we sought to find RfaH sequences to infer coevolution-based residue-residue interactions essential for the novel RfaH autoinhibited fold.

It is worth noting that current state-of-the-art structural predictors, such as AlphaFold2 (Jumper et al., 2021), can predict the metamorphic behavior of RfaH. For example, using the sequence of full-length *E. coli* RfaH as input into ColabFold (colabfold.com) (Mirdita et al., 2022), a Google Colaboratory implementation of AlphaFold2, yields the αRfaH fold as a result, whereas using only the CTD of the same protein as input yields the canonical NusG-like β-barrel. However, this approach is not as straightforward as the coevolutionary analysis of thousands of protein sequences in defining the key interactions stabilizing each fold and how these interactions emerged during the evolution of the NusG protein family.

Using the metagenomic database metaclust to increase the available number of RfaH sequences over those deposited in the Interpro database and then filtering these sequences by using a secondary structure predictor to ascertain the duality of their structure propensity, it was possible to increase the number of ID contacts by enriching our coevolutionary analysis with true metamorphic sequences.

This coevolutionary information, in combination with different secondary structure biases, was sufficient to predict both the predominant and autoinhibited structure of RfaH in solution and to retrieve key contacts involved in the stabilization of the βCTD in the active state and the formation of a recently described β-intermediate. In fact, the β-intermediate that precedes βCTD folding (Galaz-Davison et al., 2021) is formed even in the presence of ID contacts. These findings suggest that the duality in secondary structure propensity of RfaH is essential for the stabilization of both folds.

We have also shown that it is not necessary to infer all interhelical CTD contacts to predict a compact αCTD that inhibits the NTD. In fact, recent experimental work has demonstrated that the ends of the αCTD hairpin are largely unstructured in solution (Burmann et al., 2012; Galaz-Davison et al., 2020). Furthermore, it has been shown that an exactly solvable model of helical-coil-sheet transitions displays cooperativity in its temperature-induced folding from helical to extended configurations, prior to reaching the coil state (Schreck & Yuan, 2010). Altogether, these precedents suggest that nucleation of the tip of the αCTD hairpin at the NTD hydrophobic patch by coevolutionary ID contacts could trigger the formation of the autoinhibited fold of RfaH. Although the stability of the autoinhibited over the active state of RfaH in solution cannot be derived from these coevolution-guided simulations, its thermodynamic favorability has been thoroughly analyzed in simulations and experiments that explore its protein folding landscape in detail (Galaz-Davison et al., 2021; Ramírez-Sarmiento et al., 2015; Seifi & Wallin, 2021).

Our results also showed that non-native interactions of either RfaH state were relevant to produce compactness during protein folding, particularly non-native contacts from the β-barrel CTD were essential to ensure compactness of the αCTD in the autoinhibited state. Even though they account for nearly half of all coevolution-based interactions, only 30% of these non-native contacts are present in the final annealed configurations of αRfaH, while 70% of the correctly predicted native contacts are formed.

Although being marginally formed, some of the non-native contacts that guide αRfaH folding are in close spatial proximity in the native state, and hence help compacting the protein structure. Thus, the local frustration, i.e. the roughness of the potential energy landscape for protein folding arising from conflicting interactions (Mouro, de Godoi Contessoto, Chahine, Junio de Oliveira, & Pereira Leite, 2016), brought by non-native contacts is expected to play a fundamental role in enhancing the folding process of RfaH, as has been seen in globular proteins (Clementi & Plotkin, 2004; Contessoto et al., 2013). These pairs of significantly coevolving residues not involved in direct physical contacts in RfaH may correspond to interactions necessary to some functional aspects, presumably even fold-switching.

5. OVERALL CONCLUSIONS

As it has been shown through these articles, the metamorphic protein RfaH displays striking structural features that enable its fold-switching. Specifically, it has been demonstrated that the features responsible for its transformation can be traced back to its primary structure, since the tip of the α-hairpin is the main responsible for its autoinhibition, as described in Chapter 2, meanwhile the rest of the helices display mostly unstructured behavior, hence favoring the folding into a β-barrel.

The different computational and experimental strategies utilized here have a twofold purpose. First, they ascertain the effect of the interdomain contacts in stabilizing αRfaH (or destabilizing βRfaH, as described in Chapter 3), in many times by comparing the changes in stability between the full-length protein and the isolated CTD. Second, they highlight the relevance of secondary structure propensities both for metamorphic predictions and for structural stability into each fold. This is facilitated by the fact that the fold-switch of RfaH CTD goes from being all α-helical into a fully β-stranded barrel during activation.

Despite being key in the stabilization of the autoinhibited fold of RfaH, the interdomain contacts also are highlighted in the refolding process of RfaH from the active, fully β-folded CTD to the ground αRfaH state. Specifically, these interdomain interactions stabilize an intermediate state that is observed during RfaH CTD refolding from a fully unfolded state of the isolated domain or the full-length protein. Hence, this three-stranded β-folded state is likely a thermodynamic or kinetic intermediate of the fold-switching process. We further described that its stabilization is due to both native and non-native interactions with the NTD.

Finally, in Chapter 4 we demonstrated that the pairing of coevolutionary native and non-native contacts derived from the analysis of thousands of RfaH-like homologs with appropriate secondary structure propensities are sufficient for the stabilization of RfaH into its autoinhibited state. Surprisingly, most predicted contacts that match the physical contacts observed in the experimentally solved structures of autoinhibited RfaH, so-called "true positive contacts", are mostly responsible for the helicity of the CTD and only described a few interdomain interactions at the NTD hydrophobic patch, which likely would impede spontaneous binding to the RNAP.

The non-native contacts detected by coevolutionary analysis, i.e. false positive contacts with high statistical support but contact distances in the experimental RfaH structures that are too large to be considered as direct native contacts, had a distribution that was much closer in distance than that of randomly generated contacts, suggesting that the DCA-derived non-native contacts have physical implications during folding of the protein. Furthermore, replacing the non-native contacts by random interaction resulted in misfolding of RfaH during its refolding via simulated annealing.

In addition, calculating the fraction of contacts formed relative to the distance in the native structure, show that what we determined to be "native contacts" were formed in an abundance similar to any other protein being simulated (~70%, or Q ~ 0.7), meanwhile the non-native contacts were formed to a much lower extent (~30%, or Q = 0.3). These data suggest that subsets of non-native contacts are effectively being formed during the folding process, but they themselves are insufficient to fold into any given structure and are likely just aiding as redundant contacts in the formation of native contacts.

The original hypothesis and goal of this thesis is to use all the information collected regarding what makes RfaH a metamorphic protein and use this information for the design of a novel RfaH based on the sequence of its non-metamorphic paralog NusG. Although we were able to determine most of the factors in the evolution, thermodynamics and structural preferences that make RfaH metamorphic, due to time limitations it is no longer possible to go forward and use this information as input for the NusG redesign. Although partially evaluated, the hypothesis remains as future work for other researchers.

The redesign or simply design from scratch of protein sequences has become easier in the time of artificial intelligence tools such as Alphafold2 (Jumper et al., 2021), that can efficiently predict structures from an aminoacidic sequence, and even perform the opposite work of "hallucination", in which it evaluates the possible sequences that might fit a given set of structural constraints. Even though these tools are available now for testing the proposed hypothesis of this thesis, the experimental validation of the sequences is not an easy task. Although a plan can be devised for approximating whether the resulting protein is folded as the α- or β-state of RfaH using such simple tools as circular dichroism (Porter et al., 2022), the time it would require to perform such experiments exceeds the allotted time for a doctoral thesis, thus it will remain as future work.

The compilation of these three works is key in understanding the evolution and structuring of metamorphic proteins, as apart from being one of the most studied proteins, RfaH is an straightforward example of the emergence of metamorphosis in a protein family. First, it evolved in a highly evolutionarily pressured environment, i.e., a binding partner for the transcriptional machinery, where defective and hyper-active mutations can significantly alter its host fitness. Second, the evolutionary counterpart of a non-metamorphic protein is retained in the bacteria, therefore the structural and functional effect of mutagenic assays on RfaH metamorphosis can always be correlated or normalized by performing the corresponding mutations on NusG. Third, the metamorphosis of RfaH is as intricate as it can be, meaning that the whole 52-residue long domain refolds into a completely different structure. Other examples, such as KaiB or lymphotactin, do not undergo a structural rearrangement as big as the one of RfaH, nor the secondary structure is as different. Hence, studying RfaH metamorphosis is central to understanding how metamorphic proteins work and evolve.

Future perspectives of this work, apart from the *de novo* design of a metamorphic protein, includes finding a straightforward computational prediction of metamorphic proteins. Recent work has been able to determine the “structural diversity landscape” of metamorphic proteins using the Alphafold2 artificial intelligence, but the authors mention that this tool is very limited for the recognition of new metamorphic proteins (Wayment-Steele, Ovchinnikov, Colwell, & Kern, 2022), hence more sensitive “metamorphic trained” methods are required to perform robust predictions on the sequence-to-structure relationship of these proteins. These tools could be used for essential tasks in understanding the emergence of a metamorphic protein family through ancestral reconstruction, which in principle would reveal the complete story of the emergence of RfaH since its duplication from the NusG gene.

All in all, in these three articles the main drivers of RfaH structural duality were examined, its energetic propensity, mechanism and evolutionary determinants were all studied with the end of explaining RfaH metamorphosis. Despite being unable to design a new protein based on these determinants, I amassed enough information to guide the rational design of a new RfaH protein, something that will likely occur in the future using recent developments in artificial intelligence for protein hallucination and can be further demonstrated by studying the structure of ancestral RfaH sequences that branched out of the NusG family.

6. REFERENCES

Alexander, P. A., He, Y., Chen, Y., Orban, J., & Bryan, P. N. (2009). A minimal sequence code for switching protein structure and function. *Proceedings of the National Academy of Sciences of the United States of America*, *106*(50), 21149–21154. https://doi.org/10.1073/pnas.0906408106

Altschul, S. F., Madden, T. L., Schäffer, A. A., Zhang, J., Zhang, Z., Miller, W., & Lipman, D. J. (1997). Gapped BLAST and PSI-BLAST: A new generation of protein database search programs. *Nucleic Acids Research*. https://doi.org/10.1093/nar/25.17.3389

Ambroggio, X. I., & Kuhlman, B. (2006). Computational design of a single amino acid sequence that can switch between two distinct protein folds. *Journal of the American Chemical Society*, *128*(4), 1154–1161. https://doi.org/10.1021/ja054718w

Anfinsen, C. B. (1973). Principles that govern the folding of protein chains. *Science*, *181*(4096), 223–230. https://doi.org/10.1126/science.181.4096.223

Appadurai, R., Nagesh, J., & Srivastava, A. (2021). High resolution ensemble description of metamorphic and intrinsically disordered proteins using an efficient hybrid parallel tempering scheme. *Nature Communications*, *12*(1), 958. https://doi.org/10.1038/s41467-021-21105-7

Artsimovitch, I., & Knauer, S. H. (2019). Ancient transcription factors in the news. *MBio*, *10*(1), 1–16. https://doi.org/10.1128/mBio.01547-18

Artsimovitch, I., & Landick, R. (2000). Pausing by bacterial RNA polymerase is mediated by mechanistically distinct classes of signals. *Proceedings of the National Academy of Sciences of the United States of America*, *97*(13), 7090–7095. https://doi.org/10.1073/pnas.97.13.7090

Artsimovitch, I., & Landick, R. (2002). The transcriptional regulator RfaH stimulates RNA chain synthesis after recruitment to elongation complexes by the exposed nontemplate DNA strand. *Cell*, *109*(2), 193–203. https://doi.org/10.1016/S0092-8674(02)00724-9

Bai, Y., Milne, J. S., Mayne, L., & Englander, S. W. (1993). Primary structure effects on peptide group hydrogen exchange. *Proteins: Structure, Function, and Bioinformatics*, *17*(1), 75–86. https://doi.org/10.1002/prot.340170110

Bai, Y., Milne, J. S., Mayne, L., & Englander, S. W. (1994). Protein stability parameters measured by hydrogen exchange. *Proteins: Structure, Function, and Bioinformatics*, *20*(1), 4–14. https://doi.org/10.1002/prot.340200103

Bailey, M. J. A., Hughes, C., & Koronakis, V. (1997). RfaH and the ops element, components of a novel system controlling bacterial transcription elongation. *Molecular Microbiology*, *26*(5), 845–851. https://doi.org/10.1046/j.1365-2958.1997.6432014.x

Balasco, N., Barone, D., & Vitagliano, L. (2015). Structural conversion of the transformer protein RfaH: New insights derived from protein structure prediction and molecular dynamics simulations. *Journal of Biomolecular Structure and Dynamics*, *33*(10), 2173–2179. https://doi.org/10.1080/07391102.2014.994188

Bandukwala, H. S., Wu, Y., Feuerer, M., Chen, Y., Barboza, B., Ghosh, S., … Chen, L. (2011). Structure of a Domain-Swapped FOXP3 Dimer on DNA and Its Function in Regulatory T Cells. *Immunity*, *34*(4), 479–491.

https://doi.org/10.1016/j.immuni.2011.02.017

Belogurov, G. A., Mooney, R. A., Svetlov, V., Landick, R., & Artsimovitch, I. (2009). Functional specialization of transcription elongation factors. *EMBO Journal*, *28*(2), 112–122. https://doi.org/10.1038/emboj.2008.268

Belogurov, G. A., Sevostyanova, A., Svetlov, V., & Artsimovitch, I. (2010). Functional regions of the N-terminal domain of the antiterminator RfaH. *Molecular Microbiology*, *76*(2), 286–301. https://doi.org/10.1111/j.1365-2958.2010.07056.x

Belogurov, G. A., Vassylyeva, M. N., Svetlov, V., Klyuyev, S., Grishin, N. V., Vassylyev, D. G. G., & Artsimovitch, I. (2007). Structural Basis for Converting a General Transcription Factor into an Operon-Specific Virulence Regulator. *Molecular Cell*, *26*(1), 117–129. https://doi.org/10.1016/j.molcel.2007.02.021

Bernhardt, N. A., & Hansmann, U. H. E. (2018). Multifunnel Landscape of the Fold-Switching Protein RfaH-CTD. *Journal of Physical Chemistry B*, *122*(5), 1600–1607. https://doi.org/10.1021/acs.jpcb.7b11352

Blum, M., Chang, H. Y., Chuguransky, S., Grego, T., Kandasaamy, S., Mitchell, A., … Finn, R. D. (2021). The InterPro protein families and domains database: 20 years on. *Nucleic Acids Research*, *49*(D1), D344–D354. https://doi.org/10.1093/nar/gkaa977

Burmann, B. M., Knauer, S. H., Sevostyanova, A., Schweimer, K., Mooney, R. A., Landick, R., … Rösch, P. (2012). An α helix to β barrel domain switch transforms the transcription factor RfaH into a translation factor. *Cell*, *150*(2), 291–303. https://doi.org/10.1016/j.cell.2012.05.042

Burmann, B. M., Scheckenhofer, U., Schweimer, K., & Rösch, P. (2011). Domain interactions of the transcription-translation coupling factor Escherichia coli NusG are intermolecular and transient. *Biochemical Journal*, *435*(3), 783–789. https://doi.org/10.1042/BJ20101679

Burmann, B. M., Schweimer, K., Luo, X., Wahl, M. C., Stitt, B. L., Gottesman, M. E., & Rösch, P. (2010). A NusE:NusG complex links transcription and translation. *Science*, *328*(5977), 501–504. https://doi.org/10.1126/science.1184953

Campos, L. A., Sharma, R., Alvira, S., Ruiz, F. M., Ibarra-Molero, B., Sadqi, M., … Muñoz, V. (2019). Engineering protein assemblies with allosteric control via monomer fold-switching. *Nature Communications*, *10*(1). https://doi.org/10.1038/s41467-019-13686-1

Cecchini, M., Krivov, S. V., Spichty, M., & Karplus, M. (2009). Calculation of free-energy differences by confinement simulations. Application to peptide conformers. *Journal of Physical Chemistry B*, *113*(29), 9728–9740. https://doi.org/10.1021/jp9020646

Chen, M., Schafer, N. P., Zheng, W., & Wolynes, P. G. (2018). The Associative Memory, Water Mediated, Structure and Energy Model (AWSEM)-Amylometer: Predicting Amyloid Propensity and Fibril Topology Using an Optimized Folding Landscape Model. *ACS Chemical Neuroscience*, *9*(5), 1027–1039. https://doi.org/10.1021/acschemneuro.7b00436

Chen, Y., Chen, C., Zhang, Z., Liu, C. C., Johnson, M. E., Espinoza, C. A., … Chen, L. (2015). DNA binding by FOXP3 domain-swapped dimer suggests mechanisms of long-range chromosomal interactions. *Nucleic Acids Research*, *43*(2), 1268–1282. https://doi.org/10.1093/nar/gku1373

Cimini, D., De Rosa, M., Carlino, E., Ruggiero, A., & Schiraldi, C. (2013). Homologous overexpression of rfaH in

E. coli K4 improves the production of chondroitin-like capsular polysaccharide. *Microbial Cell Factories*, *12*(1), 46. https://doi.org/10.1186/1475-2859-12-46

Clementi, C., & Plotkin, S. S. (2004). The effects of nonnative interactions on protein folding rates: Theory and simulation. *Protein Science*, *13*(7), 1750–1766. https://doi.org/10.1110/ps.03580104

Contessoto, V. G., Lima, D. T., Oliveira, R. J., Bruni, A. T., Chahine, J., & Leite, V. B. P. (2013). Analyzing the effect of homogeneous frustration in protein folding. *Proteins: Structure, Function and Bioinformatics*, *81*(10), 1727–1737. https://doi.org/10.1002/prot.24309

Cordes, M. H. J., Burton, R. E., Walsh, N. P., McKnight, C. J., & Sauer, R. T. (2000). An evolutionary bridge to a new protein fold. *Nature Structural Biology*, *7*(12), 1129–1132. https://doi.org/10.1038/81985

D.A. Case, R.M. Betz, Cerutti, D. S., T.E. Cheatham, I., T.A. Darden, Duke, R. E., … Kollman, P. A. (2016). Amber 2016. *University of California, San Francisco*. University of California, San Francisco.

Davtyan, A., Schafer, N. P., Zheng, W., Clementi, C., Wolynes, P. G., & Papoian, G. A. (2012). AWSEM-MD: Protein structure prediction using coarse-grained physical potentials and bioinformatically based local structure biasing. *Journal of Physical Chemistry B*, *116*(29), 8494–8503. https://doi.org/10.1021/jp212541y

Dill, K. A. (1990). Dominant Forces in Protein Folding. *Biochemistry*, *29*(31), 7133–7155. https://doi.org/10.1021/bi00483a001

Dill, K. A., & MacCallum, J. L. (2012). The protein-folding problem, 50 years on. *Science*, *338*(6110), 1042–1046. https://doi.org/10.1126/science.1219021

Dishman, A. F., & Volkman, B. F. (2018). Unfolding the Mysteries of Protein Metamorphosis. *ACS Chemical Biology*, *13*(6), 1438–1446. https://doi.org/10.1021/acschembio.8b00276

dos Santos, R. N., Jiang, X., Martínez, L., & Morcos, F. (2019). Coevolutionary Signals and Structure-Based Models for the Prediction of Protein Native Conformations. *Methods in Molecular Biology*, *1851*, 83–103. https://doi.org/10.1007/978-1-4939-8736-8_5

Drozdetskiy, A., Cole, C., Procter, J., & Barton, G. J. (2015). JPred4: A protein secondary structure prediction server. *Nucleic Acids Research*, *43*(W1), W389–W394. https://doi.org/10.1093/nar/gkv332

Eddy, S. R. (2011). Accelerated profile HMM searches. *PLoS Computational Biology*, *7*(10). https://doi.org/10.1371/journal.pcbi.1002195

Ekeberg, M., Lövkvist, C., Lan, Y., Weigt, M., & Aurell, E. (2013). Improved contact prediction in proteins: Using pseudolikelihoods to infer Potts models. *Physical Review E - Statistical, Nonlinear, and Soft Matter Physics*, *87*(1), 012707. https://doi.org/10.1103/PhysRevE.87.012707

Engelberger, F., Galaz-Davison, P., Bravo, G., Rivera, M., & Ramírez-Sarmiento, C. A. (2021). Developing and Implementing Cloud-Based Tutorials That Combine Bioinformatics Software, Interactive Coding, and Visualization Exercises for Distance Learning on Structural Bioinformatics. *Journal of Chemical Education*, *98*(5), 1801–1807. https://doi.org/10.1021/acs.jchemed.1c00022

Fleishman, S. J., Corn, J. E., Strauch, E. M., Whitehead, T. A., Karanicolas, J., & Baker, D. (2011). Hotspot-centric de novo design of protein binders. *Journal of Molecular Biology*, *413*(5), 1047–1062. https://doi.org/10.1016/j.jmb.2011.09.001

Frishman, D., & Argos, P. (1995). Knowledge-based protein secondary structure assignment. *Proteins: Structure, Function, and Bioinformatics*, *23*(4), 566–579. https://doi.org/10.1002/prot.340230412

Galaz-Davison, P., Molina, J. A., Silletti, S., Komives, E. A., Knauer, S. H., Artsimovitch, I., & Ramírez-Sarmiento, C. A. (2020). Differential Local Stability Governs the Metamorphic Fold Switch of Bacterial Virulence Factor RfaH. *Biophysical Journal*, *118*(1), 96–104. https://doi.org/10.1016/j.bpj.2019.11.014

Galaz-Davison, P., Román, E. A., & Ramírez-Sarmiento, C. A. (2021). The N-terminal domain of RfaH plays an active role in protein fold-switching. *PLoS Computational Biology*, *17*(9), e1008882. https://doi.org/10.1371/journal.pcbi.1008882

Gc, J. B., Bhandari, Y. R., Gerstman, B. S., & Chapagain, P. P. (2014). Molecular dynamics investigations of the α-helix to β-Barrel conformational transformation in the RfaH transcription factor. *Journal of Physical Chemistry B*, *118*(19), 5101–5108. https://doi.org/10.1021/jp502193v

Gc, J. B., Gerstman, B. S., & Chapagain, P. P. (2015). The Role of the Interdomain Interactions on RfaH Dynamics and Conformational Transformation. *Journal of Physical Chemistry B*, *119*(40), 12750–12759. https://doi.org/10.1021/acs.jpcb.5b05681

Hawkins, G. D., Cramer, C. J., & Truhlar, D. G. (1996). Parametrized models of aqueous free energies of solvation based on pairwise descreening of solute atomic charges from a dielectric medium. *Journal of Physical Chemistry*, *100*(51), 19824–19839. https://doi.org/10.1021/jp961710n

Hirtreiter, A., Damsma, G. E., Cheung, A. C. M., Klose, D., Grohmann, D., Vojnic, E., … Werner, F. (2010). Spt4/5 stimulates transcription elongation through the RNA polymerase clamp coiled-coil motif. *Nucleic Acids Research*, *38*(12), 4040–4051. https://doi.org/10.1093/nar/gkq135

Hopf, T. A., Green, A. G., Schubert, B., Mersmann, S., Schärfe, C. P. I., Ingraham, J. B., … Marks, D. S. (2019). The EVcouplings Python framework for coevolutionary sequence analysis. *Bioinformatics*, *35*(9), 1582–1584. https://doi.org/10.1093/bioinformatics/bty862

Humphrey, W., Dalke, A., & Schulten, K. (1996). VMD: Visual molecular dynamics. *Journal of Molecular Graphics*, *14*(1), 33–38. https://doi.org/10.1016/0263-7855(96)00018-5

Jana, B., Morcos, F., & Onuchic, J. N. (2014). From structure to function: The convergence of structure based models and co-evolutionary information. *Physical Chemistry Chemical Physics*, *16*(14), 6496–6507. https://doi.org/10.1039/c3cp55275f

Joseph, J. A., Chakraborty, D., & Wales, D. J. (2019). Energy Landscape for Fold-Switching in Regulatory Protein RfaH. *Journal of Chemical Theory and Computation*, *15*(1), 731–742. https://doi.org/10.1021/acs.jctc.8b00912

Jumper, J., Evans, R., Pritzel, A., Green, T., Figurnov, M., Ronneberger, O., … Hassabis, D. (2021). Highly accurate protein structure prediction with AlphaFold. *Nature*, *596*(7873), 583–589. https://doi.org/10.1038/s41586-021-03819-2

Kamisetty, H., Ovchinnikov, S., & Baker, D. (2013). Assessing the utility of coevolution-based residue-residue contact predictions in a sequence- and structure-rich era. *Proceedings of the National Academy of Sciences of the United States of America*, *110*(39), 15674–15679. https://doi.org/10.1073/pnas.1314045110

Kang, J. Y., Mooney, R. A., Nedialkov, Y., Saba, J., Mishanina, T. V., Artsimovitch, I., … Darst, S. A. (2018). Structural Basis for Transcript Elongation Control by NusG Family Universal Regulators. *Cell*, *173*(7), 1650–1662. https://doi.org/10.1016/j.cell.2018.05.017

Kelley, L. A., Gardner, S. P., & Sutcliffe, M. J. (1996). An automated approach for clustering an ensemble of NMR-derived protein structures into conformationally related subfamilies. *Protein Engineering*, *9*(11), 1063–1065. https://doi.org/10.1093/protein/9.11.1063

Kendrew, J. C., Bodo, G., Dintzis, H. M., Parrish, R. G., Wyckoff, H., & Phillips, D. C. (1958). A three-dimensional model of the myoglobin molecule obtained by x-ray analysis. *Nature*, *181*(4610), 662–666. https://doi.org/10.1038/181662a0

Kendrew, J. C., Dickerson, R. E., Strandberg, B. E., Hart, R. G., Davies, D. R., Phillips, D. C., & Shore, V. C. (1960). Structure of myoglobin: A three-dimensional fourier synthesis at 2 . resolution. *Nature*, *185*(4711), 422–427. https://doi.org/10.1038/185422a0

Kim, D. E., Chivian, D., & Baker, D. (2004). Protein structure prediction and analysis using the Robetta server. *Nucleic Acids Research*, *32*(WEB SERVER ISS.). https://doi.org/10.1093/nar/gkh468

Klein, B. J., Bose, D., Baker, K. J., Yusoff, Z. M., Zhang, X., & Murakami, K. S. (2011). RNA polymerase and transcription elongation factor Spt4/5 complex structure. *Proceedings of the National Academy of Sciences of the United States of America*, *108*(2), 546–550. https://doi.org/10.1073/pnas.1013828108

Knauer, S. H., Artsimovitch, I., & Rösch, P. (2012). Transformer proteins. *Cell Cycle*, *11*(23), 4289–4290. https://doi.org/10.4161/cc.22468

Kortemme, T., Joachimiak, L. A., Bullock, A. N., Schuler, A. D., Stoddard, B. L., & Baker, D. (2004). Computational redesign of protein-protein interaction specificity. *Nature Structural and Molecular Biology*, *11*(4), 371–379. https://doi.org/10.1038/nsmb749

Kumar, S., Rosenberg, J. M., Bouzida, D., Swendsen, R. H., & Kollman, P. A. (1992). THE weighted histogram analysis method for free-energy calculations on biomolecules. I. The method. *Journal of Computational Chemistry*, *13*(8), 1011–1021. https://doi.org/10.1002/jcc.540130812

Lawson, M. R., Ma, W., Bellecourt, M. J., Artsimovitch, I., Martin, A., Landick, R., … Berger, J. M. (2018). Mechanism for the Regulated Control of Bacterial Transcription Termination by a Universal Adaptor Protein. *Molecular Cell*, *71*(6), 911-922.e4. https://doi.org/10.1016/j.molcel.2018.07.014

Leaver-Fay, A., Tyka, M., Lewis, S. M., Lange, O. F., Thompson, J., Jacak, R., … Bradley, P. (2011). Rosetta3: An object-oriented software suite for the simulation and design of macromolecules. *Methods in Enzymology*, *487*(C), 545–574. https://doi.org/10.1016/B978-0-12-381270-4.00019-6

Lella, M., & Mahalakshmi, R. (2017). Metamorphic Proteins: Emergence of Dual Protein Folds from One Primary Sequence. *Biochemistry*, *56*(24), 2971–2984. https://doi.org/10.1021/acs.biochem.7b00375

Li, S., Xiong, B., Xu, Y., Lu, T., Luo, X., Luo, C., … Jiang, H. (2014). Mechanism of the all-α to all-β conformational transition of RfaH-CTD: Molecular dynamics simulation and markov state model. *Journal of Chemical Theory and Computation*, *10*(6), 2255–2264. https://doi.org/10.1021/ct5002279

Martinez-Rucobo, F. W., Sainsbury, S., Cheung, A. C. M., & Cramer, P. (2011). Architecture of the RNA polymerase-Spt4/5 complex and basis of universal transcription processivity. *EMBO Journal*, *30*(7), 1302–1310. https://doi.org/10.1038/emboj.2011.64

Medina, E., Córdova, C., Villalobos, P., Reyes, J., Komives, E. A., Ramírez-Sarmiento, C. A., & Babul, J. (2016). Three-Dimensional Domain Swapping Changes the Folding Mechanism of the Forkhead Domain of FoxP1. *Biophysical Journal*, *110*(11), 2349–2360. https://doi.org/10.1016/j.bpj.2016.04.043

Meyer, O., & Schlegel, H. G. (1983). Biology of aerobic carbon monoxide-oxidizing bacteria. *Annual Review of Microbiology*, *37*(1), 277–310. https://doi.org/10.1146/annurev.mi.37.100183.001425

Mirdita, M., Schütze, K., Moriwaki, Y., Heo, L., Ovchinnikov, S., & Steinegger, M. (2022). ColabFold: making protein folding accessible to all. *Nature Methods*, *19*(6), 679–682. https://doi.org/10.1038/s41592-022-01488-1

Mistry, J., Chuguransky, S., Williams, L., Qureshi, M., Salazar, G. A., Sonnhammer, E. L. L., … Bateman, A. (2021). Pfam: The protein families database in 2021. *Nucleic Acids Research*, *49*(D1), D412–D419. https://doi.org/10.1093/nar/gkaa913

Mitra, A., Loh, A., Gonzales, A., Łaniewski, P., Willingham, C., Curtiss, R., & Roland, K. L. (2013). Safety and protective efficacy of live attenuated Salmonella Gallinarum mutants in Rhode Island Red chickens. *Vaccine*, *31*(7), 1094–1099. https://doi.org/10.1016/j.vaccine.2012.12.021

Mitra, P., Ghosh, G., Hafeezunnisa, M., & Sen, R. (2017). Rho Protein: Roles and Mechanisms. *Annual Review of Microbiology*, *71*(1), 687–709. https://doi.org/10.1146/annurev-micro-030117-020432

Mooney, R. A., Schweimer, K., Rösch, P., Gottesman, M., & Landick, R. (2009). Two Structurally Independent Domains of E. coli NusG Create Regulatory Plasticity via Distinct Interactions with RNA Polymerase and Regulators. *Journal of Molecular Biology*, *391*(2), 341–358. https://doi.org/10.1016/j.jmb.2009.05.078

Morcos, F., Jana, B., Hwa, T., & Onuchic, J. N. (2013). Coevolutionary signals across protein lineages help capture multiple protein conformations. *Proceedings of the National Academy of Sciences of the United States of America*, *110*(51), 20533–20538. https://doi.org/10.1073/pnas.1315625110

Morcos, F., Pagnani, A., Lunt, B., Bertolino, A., Marks, D. S., Sander, C., … Weigt, M. (2011). Direct-coupling analysis of residue coevolution captures native contacts across many protein families. *Proceedings of the National Academy of Sciences of the United States of America*, *108*(49), E1293–E1301. https://doi.org/10.1073/pnas.1111471108

Mouro, P. R., de Godoi Contessoto, V., Chahine, J., Junio de Oliveira, R., & Pereira Leite, V. B. (2016). Quantifying Nonnative Interactions in the Protein-Folding Free-Energy Landscape. *Biophysical Journal*, *111*(2), 287–293. https://doi.org/10.1016/j.bpj.2016.05.041

Murzin, A. G. (2008). Biochemistry: Metamorphic proteins. *Science*, *320*(5884), 1725–1726. https://doi.org/10.1126/science.1158868

Nagy, G., Danino, V., Dobrindt, U., Pallen, M., Chaudhuri, R., Emödy, L., … Hacker, J. (2006). Down-regulation of key virulence factors makes the Salmonella enterica serovar typhimurium rfaH mutant a promising live-attenuated vaccine candidate. *Infection and Immunity*, *74*(10), 5914–5925. https://doi.org/10.1128/IAI.00619-06

Noel, J. K., Levi, M., Raghunathan, M., Lammert, H., Hayes, R. L., Onuchic, J. N., & Whitford, P. C. (2016). SMOG 2: A Versatile Software Package for Generating Structure-Based Models. *PLoS Computational Biology*, *12*(3), e1004794. https://doi.org/10.1371/journal.pcbi.1004794

Noel, J. K., Whitford, P. C., Sanbonmatsu, K. Y., & Onuchic, J. N. (2010). SMOG@ctbp: Simplified deployment of structure-based models in GROMACS. *Nucleic Acids Research*, *38*(SUPPL. 2). https://doi.org/10.1093/nar/gkq498

Ovchinnikov, S., Kamisetty, H., & Baker, D. (2014). Robust and accurate prediction of residue-residue interactions across protein interfaces using evolutionary information. *ELife*, *2014*(3). https://doi.org/10.7554/eLife.02030

Ovchinnikov, V., Cecchini, M., & Karplus, M. (2013). A simplified confinement method for calculating absolute free energies and free energy and entropy differences. *Journal of Physical Chemistry B*, *117*(3), 750–762. https://doi.org/10.1021/jp3080578

Parra, R. G., Schafer, N. P., Radusky, L. G., Tsai, M. Y., Guzovsky, A. B., Wolynes, P. G., & Ferreiro, D. U. (2016). Protein Frustratometer 2: a tool to localize energetic frustration in protein molecules, now with electrostatics. *Nucleic Acids Research*, *44*(W1), W356–W360. https://doi.org/10.1093/NAR/GKW304

Pettersen, E. F., Goddard, T. D., Huang, C. C., Couch, G. S., Greenblatt, D. M., Meng, E. C., & Ferrin, T. E. (2004). UCSF Chimera - A visualization system for exploratory research and analysis. *Journal of Computational Chemistry*, *25*(13), 1605–1612. https://doi.org/10.1002/jcc.20084

Plimpton, S. (1995). Fast parallel algorithms for short-range molecular dynamics. *Journal of Computational Physics*, *117*(1), 1–19. https://doi.org/10.1006/jcph.1995.1039

Porter, L. L., Kim, A. K., Rimal, S., Looger, L. L., Majumdar, A., Mensh, B. D., … Strub, M. P. (2022). Many dissimilar NusG protein domains switch between α-helix and β-sheet folds. *Nature Communications*, *13*(1), 2021.06.10.447921. https://doi.org/10.1038/s41467-022-31532-9

Porter, L. L., & Looger, L. L. (2018). Extant fold-switching proteins are widespread. *Proceedings of the National Academy of Sciences of the United States of America*, *115*(23), 5968–5973. https://doi.org/10.1073/pnas.1800168115

Potoyan, D. A., Bueno, C., Zheng, W., Komives, E. A., & Wolynes, P. G. (2017). Resolving the NFκB Heterodimer Binding Paradox: Strain and Frustration Guide the Binding of Dimeric Transcription Factors. *Journal of the American Chemical Society*, *139*(51), 18558–18566. https://doi.org/10.1021/jacs.7b08741

Qian, B., Raman, S., Das, R., Bradley, P., McCoy, A. J., Read, R. J., & Baker, D. (2007). High-resolution structure prediction and the crystallographic phase problem. *Nature*, *450*(7167), 259–264. https://doi.org/10.1038/nature06249

Ramirez-Sarmiento, C. A., & Komives, E. A. (2018). Hydrogen-deuterium exchange mass spectrometry reveals folding and allostery in protein-protein interactions. *Methods*. https://doi.org/10.1016/j.ymeth.2018.04.001

Ramírez-Sarmiento, C. A., Noel, J. K., Valenzuela, S. L., & Artsimovitch, I. (2015). Interdomain Contacts Control Native State Switching of RfaH on a Dual-Funneled Landscape. *PLoS Computational Biology*, *11*(7), e1004379. https://doi.org/10.1371/journal.pcbi.1004379

Roy, A., Perez, A., Dill, K. A., & MacCallum, J. L. (2014). Computing the relative stabilities and the per-residue components in protein conformational changes. *Structure*, *22*(1), 168–175. https://doi.org/10.1016/j.str.2013.10.015

Ryckaert, J. P., Ciccotti, G., & Berendsen, H. J. C. (1977). Numerical integration of the cartesian equations of motion of a system with constraints: molecular dynamics of n-alkanes. *Journal of Computational Physics*, *23*(3), 327–341. https://doi.org/10.1016/0021-9991(77)90098-5

Sambrook, J., & Russell, D. W. (2001). *Molecular Cloning: A Laboratory Manual*. *Cold Spring, Harbor Laboratory, New York, 3th edn* (3rd ed.). New York: Cold Spring Harbor Laboratory Press.

Schreck, J. S., & Yuan, J. M. (2010). Exactly solvable model for helix-coil-sheet transitions in protein systems. *Physical Review E - Statistical, Nonlinear, and Soft Matter Physics*, *81*(6). https://doi.org/10.1103/PhysRevE.81.061919

Seifi, B., Aina, A., & Wallin, S. (2021). Structural fluctuations and mechanical stabilities of the metamorphic protein RfaH. *Proteins: Structure, Function and Bioinformatics*, *89*(3), 289–300. https://doi.org/10.1002/prot.26014

Seifi, B., & Wallin, S. (2021). The C-terminal domain of transcription factor RfaH: Folding, fold switching and energy landscape. *Biopolymers*, *112*(10). https://doi.org/10.1002/bip.23420

Sevostyanova, A., Belogurov, G. A., Mooney, R. A., Landick, R., & Artsimovitch, I. (2011). The β Subunit Gate Loop Is Required for RNA Polymerase Modification by RfaH and NusG. *Molecular Cell*, *43*(2), 253–262. https://doi.org/10.1016/j.molcel.2011.05.026

Shi, D., Svetlov, D., Abagyan, R., & Artsimovitch, I. (2017). Flipping states: A few key residues decide the winning conformation of the only universally conserved transcription factor. *Nucleic Acids Research*, *45*(15), 8835–8843. https://doi.org/10.1093/nar/gkx523

Sirovetz, B. J., Schafer, N. P., & Wolynes, P. G. (2017). Protein structure prediction: making AWSEM AWSEM-ER by adding evolutionary restraints. *Proteins: Structure, Function and Bioinformatics*, *85*(11), 2127–2142. https://doi.org/10.1002/prot.25367

Steinegger, M., & Söding, J. (2018). Clustering huge protein sequence sets in linear time. *Nature Communications*, *9*(1). https://doi.org/10.1038/s41467-018-04964-5

Svetlov, D., Shi, D., Twentyman, J., Nedialkov, Y., Rosen, D. A., Abagyan, R., & Artsimovitch, I. (2018). In silico discovery of small molecules that inhibit RfaH recruitment to RNA polymerase. *Molecular Microbiology*, *110*(1), 128–142. https://doi.org/10.1111/mmi.14093

Svetlov, V., Belogurov, G. A., Shabrova, E., Vassylyev, D. G., & Artsimovitch, I. (2007). Allosteric control of the RNA polymerase by the elongation factor RfaH. *Nucleic Acids Research*, *35*(17), 5694–5705. https://doi.org/10.1093/nar/gkm600

Tian, P., & Best, R. B. (2020). Exploring the sequence fitness landscape of a bridge between protein folds. *PLoS Computational Biology*, *16*(10), e1008285. https://doi.org/10.1371/journal.pcbi.1008285

Tomar, S. K., Knauer, S. H., Nandymazumdar, M., Rösch, P., & Artsimovitch, I. (2013). Interdomain contacts control folding of transcription factor RfaH. *Nucleic Acids Research*, *41*(22), 10077–10085. https://doi.org/10.1093/nar/gkt779

Tsai, M. Y., Zheng, W., Balamurugan, D., Schafer, N. P., Kim, B. L., Cheung, M. S., & Wolynes, P. G. (2016). Electrostatics, structure prediction, and the energy landscapes for protein folding and binding. *Protein Science*, *25*(1), 255–269. https://doi.org/10.1002/pro.2751

Tseng, R., Goularte, N. F., Chavan, A., Luu, J., Cohen, S. E., Chang, Y. G., … Partch, C. L. (2017). Structural basis of the day-night transition in a bacterial circadian clock. *Science*, *355*(6330), 1174–1180. https://doi.org/10.1126/science.aag2516

Tyka, M. D., Clarke, A. R., & Sessions, R. B. (2006). An efficient, path-independent method for free-energy calculations. *Journal of Physical Chemistry B*, *110*(34), 17212–17220. https://doi.org/10.1021/jp060734j

Wang, B., & Artsimovitch, I. (2021). NusG, an Ancient Yet Rapidly Evolving Transcription Factor. *Frontiers in Microbiology*, *11*, 3382. https://doi.org/10.3389/fmicb.2020.619618

Wang, B., Gumerov, V. M., Andrianova, E. P., Zhulin, I. B., & Artsimovitch, I. (2020). Origins and molecular evolution of the nusg paralog rfah. *MBio*, *11*(5), 1–19. https://doi.org/10.1128/mBio.02717-20

Wang, C., Molodtsov, V., Firlar, E., Kaelber, J. T., Blaha, G., Su, M., & Ebright, R. H. (2020). Structural basis of transcription-translation coupling. *Science*, *369*(6509), 1359–1365. https://doi.org/10.1126/SCIENCE.ABB5317

Washburn, R. S., Zuber, P. K., Sun, M., Hashem, Y., Shen, B., Li, W., … Frank, J. (2020). Escherichia coli NusG Links the Lead Ribosome with the Transcription Elongation Complex. *IScience*, *23*(8), 101352. https://doi.org/10.1016/j.isci.2020.101352

Wayment-Steele, H. K., Ovchinnikov, S., Colwell, L., & Kern, D. (2022). Prediction of multiple conformational states by combining sequence clustering with AlphaFold2. *BioRxiv*, 2022.10.17.512570. https://doi.org/doi.org/10.1101/2022.10.17.512570

Xiong, L., & Liu, Z. (2015). Molecular dynamics study on folding and allostery in RfaH. *Proteins: Structure, Function and Bioinformatics*, *83*(9), 1582–1592. https://doi.org/10.1002/prot.24839

Xun, S., Jiang, F., & Wu, Y. D. (2016). Intrinsically disordered regions stabilize the helical form of the C-terminal domain of RfaH: A molecular dynamics study. *Bioorganic and Medicinal Chemistry*, *24*(20), 4970–4977. https://doi.org/10.1016/j.bmc.2016.08.012

Zerihun, M. B., Pucci, F., Peter, E. K., & Schug, A. (2020). Pydca v1.0: A comprehensive software for direct coupling analysis of RNA and protein sequences. *Bioinformatics*, *36*(7), 2264–2265. https://doi.org/10.1093/bioinformatics/btz892

Zhang, B., Zheng, W., Papoian, G. A., & Wolynes, P. G. (2016). Exploring the Free Energy Landscape of Nucleosomes. *Journal of the American Chemical Society*, *138*(26), 8126–8133. https://doi.org/10.1021/jacs.6b02893

Zhang, Z., & Smith, D. L. (1993). Determination of amide hydrogen exchange by mass spectrometry: A new tool for protein structure elucidation. *Protein Science*, *2*(4), 522–531. https://doi.org/10.1002/pro.5560020404

Zheng, W., Schafer, N. P., Davtyan, A., Papoian, G. A., & Wolynes, P. G. (2012). Predictive energy landscapes for protein-protein association. *Proceedings of the National Academy of Sciences of the United States of America*, *109*(47), 19244–19249. https://doi.org/10.1073/pnas.1216215109

Zheng, W., Tsai, M. Y., Chen, M., & Wolynes, P. G. (2016). Exploring the aggregation free energy landscape of the amyloid-β protein (1-40). *Proceedings of the National Academy of Sciences of the United States of America*, *113*(42), 11835–11840. https://doi.org/10.1073/pnas.1612362113

Zuber, P. K., Artsimovitch, I., NandyMazumdar, M., Liu, Z., Nedialkov, Y., Schweimer, K., … Knauer, S. H. (2018). The universally-conserved transcription factor RfaH is recruited to a hairpin structure of the non-template DNA strand. *ELife*, *7*. https://doi.org/10.7554/elife.36349

Zuber, P. K., Schweimer, K., Rösch, P., Artsimovitch, I., & Knauer, S. H. (2019). Reversible fold-switching controls the functional cycle of the antitermination factor RfaH. *Nature Communications*, *10*(1), 702. https://doi.org/10.1038/s41467-019-08567-6