UC Berkeley
UC Berkeley Electronic Theses and Dissertations

Title
Understanding and engineering protein function from an energy landscape-based perspective

Permalink
https://escholarship.org/uc/item/3wm1c532

Author
Hart, Kathryn Ming

Publication Date
2013

Peer reviewed|Thesis/dissertation

eScholarship.org Powered by the California Digital Library
University of California

Understanding and engineering protein function from an energy landscape-based perspective

by

Kathryn Ming Hart

A dissertation submitted in partial satisfaction of the

requirements for the degree of

Doctor of Philosophy

in

Chemistry

in the

Graduate Division

of the

University of California, Berkeley

Committee in charge:

Professor Susan Marqusee, Co-chair
Professor Judith P. Klinman, Co-chair
Professor Bryan Krantz
Professor J. Christopher Anderson

Spring 2013

ABSTRACT

Understanding and engineering protein function from an energy landscape-based perspective

by

Kathryn Ming Hart

Doctor of Philosophy in Chemistry

University of California, Berkeley

Professor Susan Marqusee, Co-chair
Professor Judith Klinman, Co-chair

The three projects discussed in this thesis are unified by the common goal of understanding and manipulating protein function from an energy landscape-based perspective.

First, I explore how the energy landscapes of ribonucleases HI have evolved over time by resurrecting and characterizing extinct ancestors to the modern-day homologs from *E. coli* and *T. thermophilus*. Our results suggest that thermostability is a finely tuned property, which has adapted along each evolutionary lineage of RNase H to accommodate diverse environments. The thermodynamic mechanisms by which these changes occur, however, are found to be highly variable.

Then, I describe the construction of an unfolded maltose-binding protein and its subsequent analysis using neutron scattering to probe pico-nanosecond dynamics on the protein's surface. We find that our model for the unfolded state is more dynamic than its folded state and, perhaps more surprisingly, also more dynamic than an intrinsically disordered protein, tau. This interesting result highlights the difference between proteins that have evolved to be disordered and the unfolded state of proteins that have a well-defined native state.

Finally, I design and characterize an enzymatic switch that responds allosterically to a novel effector. The design is based on the principle of mutually exclusive folding and involves fusing a ligand-binding protein with an enzyme to create a construct in which only one domain is folded at a time. Several of the constructed chimeras are inhibited by ligand, and the strengths and weaknesses of our design are discussed.

For my father, who insisted I take risks; and for my mother, who taught me how to manage them

TABLE OF CONTENTS

Abstract **1**

List of Figures and Tables iv
Acknowledgements vi

Chapter 1. Introduction **1**

1.1 Protein Energy Landscapes 2
1.2 Mesophile-Thermophile Comparisons 3
1.3 Ancestral Sequence Resurrection 6
1.4 Unfolded States *versus* Intrinsically Disordered Proteins 7
1.5 Designing Protein Switches 9
1.6 Overview of Thesis 11
1.7 References 12

Chapter 2. Exploring evolution of protein energy landscapes using ancestral protein resurrection **15**

2.1 Abstract 16
2.2 Introduction 16
2.3 Results
2.3.1 Ancestral sequence resurrection 20
2.3.2 Structural characterization by CD and X-ray diffraction 23
2.3.3 Catalytic activity 25
2.3.4 Thermodynamic characterization 28
2.3.5 Unfolding and folding kinetic chacterization 46
2.4 Materials and Methods
2.4.1 Ancestral sequence resurrection 48
2.4.2 Expression and purification 48
2.4.3 Circular dichroism spectroscopy 48
2.4.4 Crystallization and structure determination of Node C 49
2.4.5 Activity Assays 49
2.4.6 Denaturant-induced and thermal denaturation 50
2.4.7 Denaturation and stability curve analysis 50
2.4.8 Unfolding and refolding kinetics 51
2.5 Discussion 52
2.6 References 54

Chapter 3. Probing dynamics of the unfolded state under native conditions using neutron scattering **58**

3.1 Abstract 59
3.2 Introduction 59
3.3 Results

3.3.1 Design of an unfolded MBP 60
3.3.2 Circular dichroism spectroscopy 61
3.3.3 Size exclusion chromatography 62
3.3.4 SAXS experiments 63
3.3.5 Neutron scattering of soluble proteins 64
3.3.6 Neutron scattering of aggregated proteins 65
3.4 Materials and Methods
3.4.1 Expression and purification of wt MBP 65
3.4.2 Design, expression and purification of MBP0 66
3.4.3 Circular dichroism spectroscopy 66
3.4.4 Size exclusion chromatography 66
3.4.5 Small angle X-ray scattering 67
3.4.6 Elastic incoherent neutron scattering 68
3.5 Discussion
3.5.1 Enhanced sidechain flexibility of unfolded MBP 69
3.5.2 Similar sidechain flexibility of aggregates 70
3.5.3 Conclusions and next steps 70
3.6 References 71

Chapter 4. Using the design principle of mutually exclusive folding to introduce novel allosteric control of enzymatic activity 72

4.1 Abstract 73
4.2 Introduction 73
4.3 Results
4.3.1 Design of mutually exclusive folding allosteric switches 74
4.3.2 Protein models for the regulatory and active domains 75
4.3.3 Characterization of MBP-SNase chimeras 77
4.3.4 Characterization of MBP-RNase H* chimeras 91
4.4 Materials and Methods
4.4.1 Construction, expression and purification of switches 100
4.4.2 Circular dichroism spectroscopy 101
4.4.3 Activity assays 101
4.5 Discussion 103
4.6 References 105

Appendix. Multiple sequence alignment of RNases H

A1.1 FASTA alignment of RNases H used to reconstruct ancestors 107

LIST OF FIGURES AND TABLES

Chapter 1

Figure 1.1	Funnel-shaped protein energy landscape	3
Figure 1.2	Thermodynamic strategies for increasing thermostability	5
Figure 1.3	Mutually exclusive folding switch	10

Chapter 2

Figure 2.1	Thermodynamic strategies for increasing thermostability	17
Figure 2.2	WebLogo representation of the RNase H multiple sequence alignment	20
Figure 2.3	Phylogenetic tree built from RNase H alignment	21
Figure 2.4	Node alignment and pairwise identity matrix	22
Table 2.1	Statistical support for resurrected ancestors	22
Figure 2.5	Posterior probabilities for Node 1	23
Figure 2.6	Circular dichroism spectra	24
Figure 2.7	Node C crystals and diffraction pattern	24
Table 2.2	Node C data collection and refinement statistics	25
Figure 2.8	Node C structure and superposition with ecRNH and ttRNH	25
Figure 2.9	Hyperchromic activity assay	26
Figure 2.10	Fluorescence activity assay	27
Table 2.3	Michaelis-Menten parameters	28
Figure 2.11	Thermal denaturation melts	29
Figure 2.12	Pre- and post-melt CD spectra	29
Table 2.4	Melting temperatures of nodes	30
Figure 2.13	Trends in melting temperature	30
Table 2.5	Melting temperatures of extant RNases H	31
Figure 2.14	Correlations between T_m and T_{env}	32
Figure 2.15	GdmCl-induced denaturation melts at 25 °C	33
Table 2.6	ΔG and m values at 25 °C	33
Figure 2.16	Trends in ΔG and m values at 25 °C	34
Figure 2.17	Correlations between ΔG and T_{env}	35
Figure 2.18	Representative denaturant melts for Node 1 at multiple temperatures	36
Figure 2.19	Stability curves	36-37
Table 2.7	Thermodynamic parameters from stability curve fits	38
Figure 2.20	Schematic of RNase H topology with labels	40
Figure 2.21	Calculated formal charges at pH 5.5	41
Figure 2.22	ΔC_p *versus* branch length and superimposed stability curve fits	42
Figure 2.23	Asymmetry in ecRNH stability data	43
Figure 2.24	Stability data for single and double cysteine variants of ecRNH	44
Table 2.8	Effects of protein concentration and TCEP on ecRNH C_m values	45
Figure 2.25	Representative folding and unfolding data for Node 2	46
Figure 2.26	Chevron of Node 2	47
Table 2.9	Kinetic fit parameters for Node 2	47
Figure 2.27	Final emission amplitudes of beacon substrate	49

Chapter 3

Figure 3.1	Location of destabilizing substitutions in MBP	61

Figure 3.2 Maltose-induced conformational change in MBP0 61
Figure 3.3 Urea denaturation profiles of MBP0 and wtMBP 62
Figure 3.4 Size exclusion chromatograms for soluble and aggregated MBP0 63
Figure 3.5 SAXS intensity profiles and Kratky plots 64
Figure 3.6 Mean square displacements for soluble MBP0 64
Table 3.1 Apparent force constants extracted from MSD for T > 260 K 65
Figure 3.7 Mean square displacements for aggregated MBP0 65
Table 3.2 Molecular masses and hydrodynamic radii for reference proteins 67

Chapter 4
Figure 4.1 Mutually exclusive folding switch 75
Figure 4.2 Boltzmann diagrams in the absence and presence of ligand 75
Table 4.1 Measures and calculated stabilities for model switch domains 77
Figure 4.3 Urea-induced denaturation profiles of active domain models 78
Table 4.2 Measured stabilities of regulatory domain models 78
Figure 4.4 Urea-induced denaturation profiles of regulatory domain models 79
Figure 4.5 Michaelis-Menten plot of wild-type SNase 79
Figure 4.6 Blue plate activity assay for wt SNase 80
Figure 4.7 MBP-SNase chimera, insertion at residue 286 81
Figure 4.8 Circular dichroism spectra of MBP-SNase short chimera 82
Figure 4.9 Circular dichroism spectra of hyperstable MBP-SNase chimera 83
Figure 4.10 Circular dichroism spectra of hypostable MBP-SNase chimera 83
Figure 4.11 Circular dichroism spectra of MBP-SNase short chimeras 84
Figure 4.12 Spectroscopic activity of MBP-SNase short chimeras 85
Figure 4.13 Blue plate activity of MBP-SNase short chimeras 86
Figure 4.14 Spectroscopic activity of truncated SNase 86
Figure 4.15 Circular dichroism spectra of MBP-SNase long chimera 87
Figure 4.16 Circular dichroism spectra of hyperstable MBP-SNase long chimera 88
Figure 4.17 Blue plate activity of hyperstable MBP-SNase long chimera 88
Figure 4.18 Blue plate activity of MBP-SNase long chimera 89
Figure 4.19 Blue plate activity of hypostable MBP-SNase long chimeras 89
Figure 4.20 MBP-SNase chimera, insertion at residue 169 90
Figure 4.21 MBP-RNase H* chimera, insertion at residue 286 91
Figure 4.22 Circular dichroism spectra of MBP-RNase H chimeras 92
Figure 4.23 Circular dichroism spectra of hypostable MBP-RNase H chimeras 92
Figure 4.24 Activity of MBP-RNase H chimera 93
Figure 4.25 Activity of hypostable MBP-RNase H chimera 94
Figure 4.26 MBP-RNase H* chimera, insertion at residue 169 94
Figure 4.27 Circular dichroism spectra of Δloop MBP-RNase H chimera 95
Figure 4.28 Circular dichroism spectra of Δloop MBP-RNase H I25A chimera 96
Figure 4.29 Activity of Δloop MBP-RNase H chimera 96
Figure 4.30 Michaelis-Menten analysis of Δloop MBP-RNase H chimera 97
Figure 4.31 Activity of hypostable Δloop MBP-RNase H chimera 98
Figure 4.32 MBP-RNase H* "flip-flopped" chimera 98
Figure 4.33 Activity of flip-flopped MBP-RNase H chimera 99
Figure 4.34 Circular dichroism spectra of unfolded MBP and wt MBP 103

ACKNOWLEDGEMENTS

First and foremost, I'd like to acknowledge all of the people who contributed directly to the work presented here. Specific collaborators are named in each chapter and include FX Gallat, Mike Harms, Bryan Schmidt and Tracy Young. Additionally, several first-year graduate students and one undergraduate worked on various aspects of these projects and include Lia Ball, Ava Brozovich, Caleb Cassidy-Amstutz, Stephanie Davis, Carolyn Elya, Naeem Hussain and Jeannette Tenthorey. I would especially like to thank Carolyn Elya for her hard work and persistence, which led to the crystallization of an ancestral RNase H.

I'd also like to acknowledge some of the people who have contributed indirectly to this work. The Marqusee lab has played the most significant role in my growth as a scientist. While people have come and gone in my time here, the environment is consistently defined by smart, community-minded individuals, who do great work and actively discuss and debate each other's research. In particular, I'd like to thank Katie Tripp for her mentorship and leadership. She was always the first person I consulted for scientific expertise, and she played an important role in fostering the lab's atmosphere of cooperation. Katie has shaped the way I think about science and problem-solving in general, and I know I will continue to count on her guidance and friendship in the future.

Choosing one's adviser is the most important decision a graduate student makes. Susan is the reason I finished graduate school and the reason I will continue to do research as postdoc. She was encouraging in the face of failure, insisted upon being convinced by the data and always pushed me to communicate the research clearly and concisely. She has gone above and beyond scientific mentorship in encouraging my interest in teaching, and I am grateful for her investment in me as a person as well as a scientist.

Graduate school is an emotional challenge as much as an intellectual one, so I would be remiss to overlook those who have supported and sustained me outside of lab. My best friend and partner, Jake Chu, brings joy and love to my life every day, which is only enhanced by his remarkable ability to distinguish between the times I need a hug and the times he should stay far, far away. My "West Coast family" of friends provides inspiration as well as diversion. I'm humbled by their thoughtfulness and honesty, and I feel privileged that my most important role models are also my dearest friends. Finally, I'd like to thank my family, especially my parents, Kathryn and Les Hart. Having contributed both genetics and environment, they are surely responsible for all my accomplishments and all my failures.

CHAPTER 1

Introduction

1.1 Protein Energy Landscapes

Proteins mediate nearly all processes necessary to sustain life. Their specific functions range from maintaining the structural integrity of cells to catalyzing the chemical reactions of central metabolism. It is a testament to the power of evolution that such diversity of function can be achieved simply by varying the linear arrangement of 20 naturally occurring amino acids.

A typical protein is comprised of 200 to 300 amino acids linked end-to-end in a single chain [2]. While this represents an astronomical number of theoretical unique sequences, in reality, evolved proteins occupy a much smaller region of sequence space. One distinguishing property of evolved, versus random, sequences is the ability to encode a thermodynamically well-defined structure. Anfinsen first described this phenomenon in 1973 by observing that a fully denatured and reduced enzyme would regain activity when returned to physiological conditions [3]. The process of a protein folding into its active structure, he noted, was driven entirely by its free energy. The so-called "thermodynamic hypothesis" informs all protein structure-function analyses and has led to an understanding of protein function based primarily on the lowest energy conformation, or native state.

In the years since Anfinsen's foundational contribution to understanding structure-function relationships, notable exceptions have emerged. Prions and other amyloidogenic proteins adopt stable conformations that are not only non-functional but typically deleterious to the organism [4]. Another exception is a class of proteins that lack a well-defined native state called intrinsically disordered proteins. Proteome-wide predictors of intrinsic disorder have revealed that more than a third of eukaryotic proteins contain unfolded regions, which means these sequences encode an ensemble of isoenergetic conformations [5]. Anfinsen's hypothesis is further complicated by the fact that proteins are dynamic systems, whose large- and small-scale motions are required for proper functioning. For these reasons, it is necessary to modify the one-sequence-one-structure paradigm into a new model that encompasses more complex behaviors. By considering that sequence encodes not just protein structure but its entire energy landscape, it is possible to reconcile these observations.

An energy landscape is a theoretical description of all the conformations accessible to a polypeptide sequence and their relative energies and rates of interconversion (**Figure 1**). For proteins with a well-defined native state, the landscape is funnel-shaped with the native structure occupying the lowest energy well and the unfolded ensemble defining the upper rim [6]. States higher in energy than the native state, but more stable than the unfolded ensemble appear as bumps and divots in the landscape. These high-energy states direct the stability and folding of a protein and can be involved directly in function. For instance, it has been demonstrated in a number of systems that undergo conformational changes upon binding that the binding-competent state exists in pre-equilibrium with the unbound state [7, 8]. In other words, this conformation appears as a high-energy, lowly populated state on the landscape, but in the presence of its substrate or ligand, the landscape is altered such that it becomes the lowest energy, most populated state. It is increasingly accepted that such conformational selection is a general phenomenon, and the more classical induced-fit mechanism represents a limiting case within its framework [9].

Non-native regions of the landscape are thought to play a role in misfolding and aggregation. Many of the high-energy states are partially unfolded, allowing them to self-associate into canonical amyloid structures or less organized protein aggregates [10]. Landscapes are also useful for rationalizing the potential functional relevance of intrinsic disorder. Proteins lacking a well-defined native state, for instance, have a relatively flat energy landscape and exist as an ensemble of rapidly equilibrating isoenergetic states. It has been observed that intrinsically disordered proteins often have multiple binding partners, so by maintaining a diversity of conformations, they are poised to interact with a diversity of ligands.

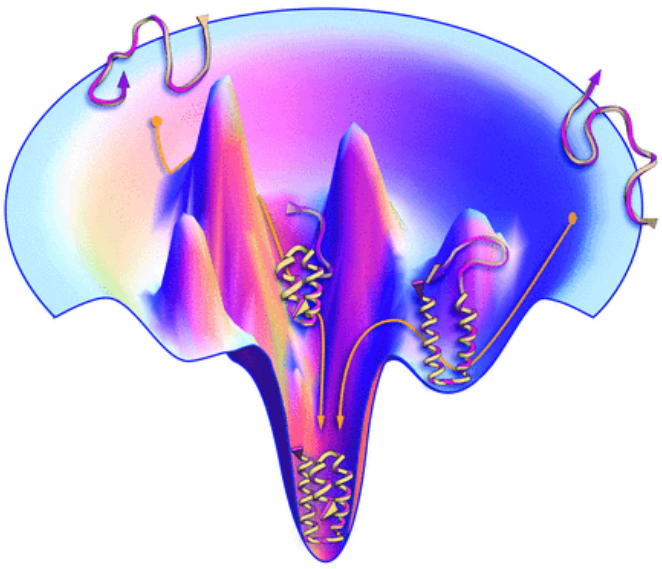


Figure 1. Funnel-shaped protein energy landscape. The folded, or native, state is the most energetically favorable state on the landscape. Image taken from [1].

Understanding sequence-landscape relationships is crucial, because while small changes in sequence rarely disrupt a protein's structure, the consequences for a protein's energy landscape can be significant. Even a conservative amino acid substitution can decrease the global stability of a protein on the order of 1-2 kcal/mol. It is true that in a two-state system, such destabilization has little effect on the overall population of the native state. For instance, assuming a typical protein is 10 kcal/mol stable, a destabilization of 2 kcal/mol corresponds with a 0.0001% change in the native state population, relative to the unfolded state. But this seemingly small change is misleading, because altering sequence also has the potential to affect populations in regions of the landscape other than native and unfolded. The two-state approximation, while certainly valid under specific conditions and for particular sorts of analyses, fails to capture the inherent roughness of protein energy landscapes. Site-specific variants, many of which retain the native structure, can have an impact on function by increasing or decreasing accessibility to higher energy states that could, for instance, be aggregation-prone or relevant for catalysis.

In this chapter, I will introduce three distinct, but related, topics in an order parallel to my experimental chapters. While my graduate work has involved a diversity of systems and techniques, all of the projects are unified by a common goal of understanding the biological relevance of protein energy landscapes.

1.2 Mesophile-Thermophile Comparisons

The biological significance of modulating the energy landscape through sequence variation can be illustrated by comparing homologous proteins that share a common native structure but function under vastly different environmental conditions. Proteins have evolved to function in a wide variety of environments. Remarkably, organisms that thrive at temperatures approaching water's boiling point utilize the same fundamental cellular machinery as those living in the human gut or on polar icecaps. Restricted to life's existing chemistries and building blocks, proteins must modulate their energetics in order to preserve function. The resulting biophysical adaptations can be elucidated by comparing proteins that share an evolutionary heritage but differ in functional tolerance to temperature.

One challenge for thermophilic proteins is to retain their native structure at high temperatures. For instance, the melting temperature of the enzyme *E. coli* ribonuclease H (ecRNH) is 68 °C, but this also represents the optimal growth temperature for *T. thermophilus*, which contains a close structural homolog [11]. Replacing the enzyme in *T. thermophilus* (ttRNH) with its mesophilic homolog would result in half the enzymes being unfolded. Not only would this prove inefficient, as half of the synthesized proteins would not function, it could have direct detrimental effects if the unfolded molecules aggregate in the cell. Thermophilic proteins must, at the very least, have sufficient global stabilities at high temperatures to preserve their native structures.

Comparisons between homologous proteins from mesophiles and thermophiles have revealed several strategies for increasing thermostability. Most commonly, thermophilic proteins are more stable at all temperatures relative to their mesophilic homologs [12]. This manifests as an upward shifting of their stability curves, which are plots of how global stability changes as function of temperature (**Figure 2**). Curvature in the plot results from the change in heat capacity associated with protein folding reactions [13]. Stability curves can be described using the Gibbs-Helmholtz equation [13]:

$$\Delta G\ (T) = \Delta H_m\ \left(1 - \frac{T}{T_m}\right) - \Delta C_p\ [\ (T_m - T) + \ T\ ln\left(\frac{T}{T_m}\right)\ ] \qquad (1)$$

where the global stability at any temperature ($\Delta G\ (T)$) is a function of the melting temperature of the protein (T_m), the change in enthalpy at the T_m (ΔH_m) and the change in heat capacity upon unfolding (ΔC_p). An upshifted curve describes higher stabilities across all temperatures, relative to a reference protein. Because the curve crosses the *x*-axis, where $\Delta G = 0$, at both higher and lower temperatures than the reference, its thermal melting temperature is higher and its cold denaturation temperature is lower. In this thermodynamic strategy, stability is increased at high temperatures without changing the heat capacity upon unfolding or the temperature of maximum stability, T_s. One way to adjust the thermodynamics of a system in this way is by increasing enthalpic interactions in the folded state without causing a compensatory entropic change. In protein design studies, thermostability is engineered almost exclusively by focusing on favorable interactions in the folded state either through the introduction of disulfide bonds, salt bridges or optimized hydrophobic packing. In nature, however, this strategy is typically used in combination with one or two of the following approaches.

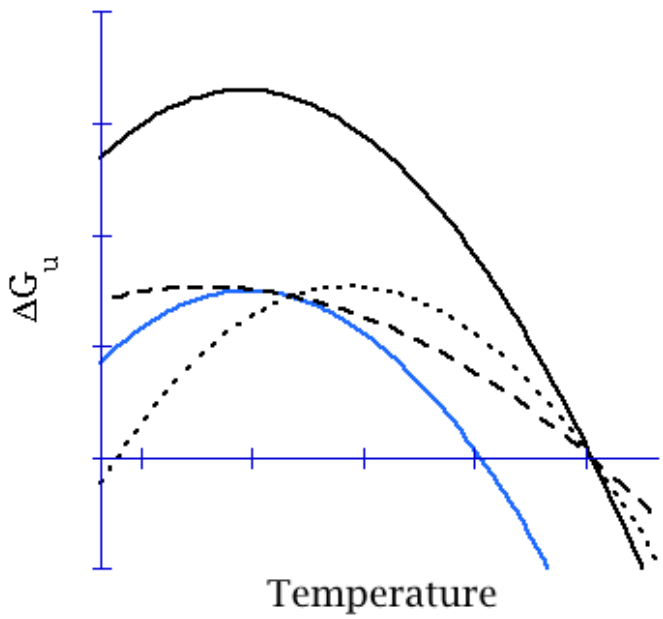


Figure 2. Thermodynamic strategies for increasing T_m. Relative to the reference state (blue line), the stability curve can be upshifted (solid black line), broadened (dashed line) or right-shifted (dotted line).

Simply right-shifting a stability curve to higher temperatures increases stability at temperatures above the intersection of the two curves. This results in a higher thermal melting temperature, higher temperature for cold denaturation and higher temperature of maximum stability. Because the T_s is also the temperature at which the unfolding reaction has a $\Delta S = 0$, right-shifting the curve requires either decreasing entropy of the unfolded state, increasing entropy of the folded state or some combination of the two. Of all the strategies, right-shifting the curve is the least commonly observed in natural systems [12].

Another way to increase stability at high temperatures is by broadening the stability curve, which is reflected in a decreased ΔC_p. Using this strategy, stability is increased at all temperatures except the T_s, with the greatest stabilization occurring near the melting temperatures. Thus, an increased T_m is observed, while the T_s remains unchanged. The positive heat capacity change upon unfolding results from the unfolded state having a higher absolute heat capacity than the folded state. ΔC_p correlates with the change in solvent exposed hydrophobic surface area [14]. In the unfolded state, more hydrophobic surface area is solvated than in the folded state, and more solvated hydrophobic surface creates a higher absolute heat capacity. One structural mechanism for reducing the ΔC_p without changing the folded state is by having residual structure in the unfolded state, which effectively reduces its solvated hydrophobic surface area. One example of this strategy is observed in the comparison between ecRNH and ttRNH [15]. While the crystal structures of the thermophilic and mesophilic enzymes overlay with an RMSD less than 2 Å, indicating that the two share a similar folded state, ttRNH has a significantly smaller ΔC_p [11]. By swapping structural domains between the two homologs, it was found that the low ΔC_p tracks with the core domain, evidently because it contains residues involved in the residual structure [16]. A later study demonstrated that a site-specific variation within the core domain was sufficient to increase ttRNH's ΔC_p to match that of ecRNH [15]. Differential scanning calorimetry studies are consistent with hydrophobic clusters in the unfolded state of ttRNH, providing further evidence for residual structure [17]. Based on the approximately 20 existing

case studies, thermophilic proteins tend to have both up-shifted and broadened curves relative to their mesophilic counterparts [12].

1.3 Ancestral Protein Resurrection

Investigating how proteins modulate their energetics in order to function under various conditions is, at its core, a study in molecular evolution. All living things are descended from a shared common ancestor. It follows that all the biomolecules necessary for life evolved from a small pool of monomers and polymers intermingling in the primordial soup. Thus, existing diversity can be rationalized from a historical perspective, because evolution is a directional, irreversible process.

Sequence conservation within a family of homologous proteins can be used to identify functional residues, because such residues are largely preserved over evolutionary time. Since the early days of genomic sequencing, this fact has been exploited to identify catalytically active residues in enzymes and protein binding interfaces [18, 19]. More ambitious efforts also have been made to use conservation to discern allosteric pathways and folding nuclei [20, 21], although both of these applications have been refuted in the literature [22, 23]. Conservation has proven to be a powerful tool in deciphering structure-function relationships, and its utility can be extended even further in the pursuit of reconstructing extinct proteins. For instance, folding studies of homologous proteins suggest that energy landscapes have largely been conserved over evolutionary time [24, 25]. The question remains whether shared evolutionary history or simply common topology is responsible for this conservation. Studying existing sequences can only address this question indirectly, but by resurrecting extinct proteins, we can begin to probe how biophysical characteristics have changed over evolutionary time.

Using computational methods, such as maximum parsimony and maximum likelihood, ancestral states can be inferred from aligning an existing protein family. By incorporating a time axis into the analysis, this methodology captures more information from the alignment than just conserved sites alone and enables more nuanced relationships within families to be explored.

Ancestral protein resurrection (APR) can been used to ask how evolution occurs on the molecular scale. The Thornton lab resurrects ancient vertebrate steroid receptors to address a classic chicken-or-egg problem: which came first, the hormone or the receptor? In this particular case, the ancient receptor demonstrates promiscuous affinity for aldosterone, which would not yet have evolved as a signaling molecule when this ancestor existed. The authors suggest that binding is a result of aldosterone's structural similarities with more ancient ligand, and thus conclude that the receptor evolved first [26]. A different study, however, illustrates that evolution can follow different trajectories. Thornton *et al.* resurrect an ancient hormone receptor and compare its binding affinity with existing receptors in vertebrates and invertebrates. The authors conclude that the receptor evolved affinity for other small molecules, which already existed as intermediates in the estrogen biosynthetic pathway, and ultimately became the receptors present in modern organisms [27].

Biochemical and biophysical analysis of ancestral proteins has been used in other studies to gain insight to changes in the global environment. Thomson *et al.* examine the predecessor of two

alcohol dehydrogenase paralogs in yeast. The modern-day enzymes have diverged in function such that one specializes in converting acetaldehyde to ethanol, leading to its accumulation in rich sugar sources. Once the sugar is consumed, the second enzyme converts ethanol back into acetaldehyde, which is ultimately channeled into the energy-harnessing pathways of central metabolism. Because the ancestral dehydrogenase shares kinetic behavior with the former, the authors conclude that the enzyme first evolved as a defense mechanism for yeast feeding on fleshy fruit, whose appearance may have coincided with the gene duplication event [28]. An earlier study from the same lab examines the temperature-dependent binding of EF-Tu ancestors to its modern ligand, GTP. All the resurrected proteins show maximal binding at higher temperatures than the extant proteins. The authors suggest this is consistent with an origin-of-life hypothesis that posits ancient life existed at elevated temperatures on the early Earth [29, 30]. Studies of Precambrian thioredoxins reveal that resurrected enzymes share mechanistic traits with their descendants but differ in other properties like melting temperature [31]. A more recent study of resurrected enzymes involved in leucine biosynthesis also uses the proteins' melting temperatures to infer ancient environmental conditions [32]. It finds that some of the ancestors have thermophile-like T_ms, while others do not, leading to the conclusion that thermophilicity has evolved multiple times during the course of this enzyme's evolution. While a limited number of studies have used the APR methodology to measure properties of ancient proteins, further work is needed to elucidate how protein energetics, and thus functions, are tuned over evolutionary time.

The mechanistic foundations for thermodynamic differences between homologous proteins, whether extant or extinct, can be attributed to the folded state, the unfolded states and all other accessible conformations. Therefore, it is imperative to interrogate native and non-native regions of the energy landscape in order to gain a full understanding of how sequence encodes function.

1.4 Unfolded State versus Intrinsically Disordered Proteins

Characterizing lowly populated, high energy states on a protein's energy landscape is experimentally challenging but essential for understanding how sequence encodes function. What dictates the energetics of various states depends not only on interactions within the protein sequence but also on interactions between the protein and water. One of the major driving forces for protein folding, for instance, is water's differential solvation of the folded versus unfolded states.

The water molecules residing on a protein's surface are particularly relevant for protein function. While it has been observed that some enzymes demonstrate activity in organic solvents and even *in vacuo*, they are not completely dehydrated [33]. For example, the specific activity of pig liver esterase in vapor phase is zero in the absence of water and increases as a function of hydration [34]. Enzymes require a minimal amount of hydration water, estimated from 0.2-0.4 g H_2O per gram of protein, presumably to facilitate the dynamics necessary for catalysis [33]. Specific waters, which are often bound specifically in enzyme active sites, can also be directly involved in chemical reactions by serving as general acids and bases or through hydrolysis. Hydration water is also important for non-enzymatic proteins, as some protein-protein and protein-ligand interactions are facilitated by bridging waters [33]. Depending upon the context, waters at

interfaces can either promote plasticity by adapting a surface to accommodate various binding partners or specificity if the waters are sufficiently restricted.

Further evidence for the importance of water is that computer simulations of protein dynamics are significantly improved when explicit waters are included. In a comprehensive analysis of existing forcefields, the Pande lab demonstrated that explicit water models outperform implicit solvent in the molecular simulation's ability to recapitulate NMR measurements [35]. Consistent with this result is the observation that the dynamics of a protein and its hydration waters are closely coupled [36]. Neutron scattering experiments measure an averaged parameter for the system that reflects the pico- nanosecond dynamics of hydrogen atoms in either a protein's sidechains or its hydration waters. It has been observed using this technique that the sidechain hydrogens of hydrated proteins undergo a characteristic dynamical transition around 200 K, which is thought to coincide with the unfreezing of water from the protein's surface [37]. Dehydrated proteins lack this dynamical transition [38]. Because the technique is essentially blind to deuterons, data collected on perdeuterated proteins will report specifically on the hydration waters' hydrogen atoms. Comparing perdeuterated maltose-binding protein (MBP) in H_2O with hydrogenated MBP in D_20 reveals that the protein and its hydration water undergo a dynamical transition at the same temperature [39]. The dynamics of both the protein and its waters, as measured by mean square displacement (MSD), share the same magnitude and temperature dependence at low temperatures. Both species deviate from linear behavior around 220 K, resulting in the dynamical transition. Above this critical temperature, water becomes more dynamic than protein, as one might expect [39].

One curious result from neutron scattering studies is that the magnitude and temperature dependence of MSDs vary little among different proteins [40]. The data for MBP, RNase A and myoglobin, for instance, overlay completely, despite significant differences in sequence, structure and size. This can be rationalized by the fact that these proteins are all soluble and globular with similar expanses of hydrophobic and hydrophilic surface. Membrane proteins, on the other hand, can show distinctive behavior, perhaps due to the character of their solvent exposed surfaces. In a study of purple membrane, which contains both proteins and lipids, it was observed that the system and its hydration waters underwent dynamical transitions at different temperatures [41]. In fact, there was no measured temperature range over which the dynamics matched. Extrapolating from this study, it seems reasonable that membrane-bound proteins represent a unique class of proteins whose dynamics are decoupled from their hydration waters' dynamics.

Another class of proteins that show distinctive dynamics is intrinsically disordered proteins (IDPs). IDPs differ from both globular and membrane-bound proteins in their sequence and the character of the solvent exposed surface area. They contain a larger proportion of polar and charged residues, relative to hydrophobic, and they exist as an expanded conformational ensemble [42]. Neutron scattering studies reveals that the dynamics of tau protein, an IDP that binds to microtubule, are even more closely coupled to hydration-water dynamics than well-folded proteins [40]. Not only do the dynamical transitions occur at the same temperature, the MSDs track quite closely at temperatures even above the transition. This is in contrast to MBP and its waters, which diverge at temperatures above the dynamical transition. The coincident dynamics are achieved both through higher MSDs for tau, relative to MBP, and lower MSDs for

tau's waters, relative to MBP's waters. Increased protein dynamics may be a reflection of tau's relatively flat energy landscape. Tau binds microtubules in a number of partially folded conformations, and perhaps the interconversion between these states is facilitated by the observed enhanced dynamics [40].

Protein-water dynamic coupling exists on a continuum, with membrane-bound proteins exhibiting the least coupling, well-folded proteins exhibiting intermediate coupling and IDPs exhibiting the greatest coupling of all. Additionally, MSDs measured at room temperature indicate that IDPs are more dynamic overall than well-folded proteins, reflecting fundamental differences in their energy landscapes.

1.5 Designing Protein Switches

The true test for understanding how a protein's sequence encodes its energetics is designing a novel energy landscape. Being able to design protein function also has direct practical applications in therapeutics and the emerging field of synthetic biology. Current strategies in synthetic biology involve modifying microorganisms by mixing and matching "parts," which can be genetic elements or proteins, from different organisms. In their quest to engineer increasingly novel metabolic pathways, however, synthetic biologists push the boundaries of existing natural diversity. If, for instance, a desired chemical transformation cannot be catalyzed by any known enzyme, the synthetic route must be redesigned or the target molecule might need to be reconsidered all together. The ability to rationally engineer protein function would enable the creation of increasingly complex organisms capable of producing chemicals or exhibiting behaviors that do not exist naturally. Protein design can be used to introduce complexity in the form of regulation. Integrating complex input signals to produce a desired output represents a major challenge to the field, which is why designing new regulatory mechanisms for protein function is critical. One way to introduce regulation is by engineering allostery.

Enzymes that are regulated allosterically respond to effector molecules with either increased or decreased activity. The classic definition of allostery derives from studies of hemoglobin and describes the non-additive behavior of ligand binding [43]. Although the concept of allostery was formalized originally to describe binding in multimeric systems, it has since been extended to include monomeric systems in which binding, or simply an energetic perturbation, at one site elicits change at a distal site. The goal of engineering allosteric control is to affect an enzyme's activity with a ligand to which it does not normally respond.

One consequence of the natural ruggedness of energy landscapes is that proteins are poised for allosteric regulation [44]. In the Monod-Wyman-Changeux (MWC) model of allostery, ligand binding modulates populations within the pre-existing conformational ensemble described by the landscape. There is evidence that the alternative, induced-fit model does not represent a competing model but rather a specialized case of MWC [9]. Thus, allostery results when perturbations to the landscape cause one conformation to be populated preferentially in the presence of ligand. Such ligand-induced switching can be introduced by creating a new binding site [45] or through domain insertion. Guntas and Ostermeier, for instance, created a library of chimeric fusion proteins and selected for those exhibiting functions of both parents [46]. The chimeras were generated by randomly inserting the sequence for TEM-1 β-lactamase, an enzyme

that hydrolyzes β-lactam antibiotics, into the sequence of maltose-binding protein (MBP) and selecting for survival on plates containing both maltose and ampicillin. Remarkably, two of the bifunctional chimeras exhibit allosteric behavior; however, this corresponds to a 0.02% hit rate, so clearly more rational approaches are needed.

Simply introducing allostery is one thing, but forward engineering switching behavior without the use of combinatorial screens or selections requires understanding mechanism. There is much debate in the field as to whether allosteric mechanisms require physical connectivity between sites or if thermodynamic linkage is sufficient [47] [48]. One design platform that incorporates both types of mechanism is called mutually exclusive folding.

Switching behavior can be engineered into a chimeric protein system using mutually exclusive folding [49]. In this design, two proteins are fused together such that only one domain can fold into its native conformation at any given time (**Figure 3**). Exclusivity is achieved by creating a discrepancy in end-to-end distance at the attachment point, resulting in one domain's folding geometrically precluding the folding of the other domain. Ultimately, thermodynamics govern which protein is folded [50]. If one of the domains is a ligand-binding protein and the other is an enzyme, then enzymatic activity should be inhibited by ligand. This gross allostery results from preferential stabilization of the ligand-binding domain, which folds in the presence of ligand causing the enzymatic domain to unfold. Ha *et al.* demonstrated proof of this principle by inserting the ligand-binding peptide, GCN4, into the enzyme barnase [51]. GCN4's ligand was found to inhibit barnase activity in a concentration-dependent manner both *in vitro* and *in vivo*. More work is required to determine if mutually exclusive folding represents a generalizable strategy for engineering allosteric control of enzymes.

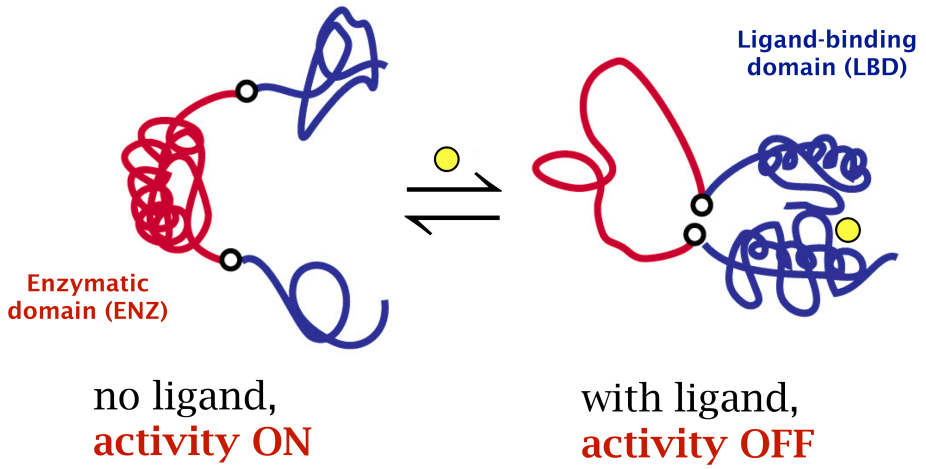


Figure 3. Mutually-exclusive folding switch. Only one domain can be folded at a time due to steric constraints, and ligand controls which state is preferentially populated. In the absence of ligand, the enzymatic domain is more stable, but in the presence of ligand, the ligand-binding domain is more stable.

1.6 Overview of Thesis

In Chapter 2, I explore how the energy landscapes of ribonucleases H have evolved over time by carrying out extensive thermodynamic and kinetic analysis of extinct ancestors to the modern homologs from *E. coli* and *T. thermophilus*. Our results suggest that thermostability is a finely tuned property, which has adapted along each evolutionary lineage of RNase H to accommodate diverse environments. The thermodynamic mechanisms by which these changes occur, however, are found to be highly variable. This work is being prepared for submission. I will be first author, but the work was carried out in collaboration with Michael Harms (U. Oregon, Eugene) and Joseph Thornton (U. Chicago), who contributed to tree construction and protein resurrection, and Bryan Schmidt (UCSF), who solved the structure for one of the extinct enzyme.

In Chapter 3, I describe the construction and validation of an unfolded maltose-binding protein and its subsequent analysis using neutron scattering, which was performed by our collaborators Martin Weik and Francois-Xavier Gallat at the Institut de Biologie Structurale in Grenoble, France. We find that our model for the unfolded state is more dynamic than its folded state and, perhaps more surprisingly, also more dynamic than an intrinsically disordered protein, tau. This interesting result highlights the difference between proteins that have evolved to be disordered and the unfolded state of proteins that have a well-defined native state. This work is being prepared for submission; I will be second author and Francois-Xavier Gallat will be first.

In Chapter 4, I design and characterize several chimeric fusions enzymes designed to respond allosterically to maltose via a mechanism of mutually exclusive folding. The design is based on the principle of mutually exclusive folding and involves fusing a ligand-binding protein with an enzyme to create a construct in which only one domain is folded at a time. Several of the constructed chimeras are inhibited by ligand, and the strengths and weaknesses of our design are discussed.

1.7 References

1. Dill, K.A. and J.L. MacCallum, *The Protein-Folding Problem, 50 Years On.* Science, 2012. **338**(6110): p. 1042-1046.
2. Brocchieri, L. and S. Karlin, *Protein length in eukaryotic and prokaryotic proteomes.* Nucleic Acids Res, 2005. **33**(10): p. 3390-400.
3. Anfinsen, C.B., *Principles that govern the folding of protein chains.* Science, 1973. **181**(4096): p. 223-30.
4. Baldwin, A.J., et al., *Metastability of native proteins and the phenomenon of amyloid formation.* J Am Chem Soc, 2011. **133**(36): p. 14160-3.
5. Ward, J.J., et al., *Prediction and functional analysis of native disorder in proteins from the three kingdoms of life.* J Mol Biol, 2004. **337**(3): p. 635-45.
6. Dill, K.A. and H.S. Chan, *From Levinthal to pathways to funnels.* Nat Struct Biol, 1997. **4**(1): p. 10-9.
7. Masterson, L.R., et al., *Dynamics connect substrate recognition to catalysis in protein kinase A.* Nat Chem Biol, 2010. **6**(11): p. 821-8.
8. Volkman, B.F., et al., *Two-state allosteric behavior in a single-domain signaling protein.* Science, 2001. **291**(5512): p. 2429-33.
9. Hammes, G.G., Y.C. Chang, and T.G. Oas, *Conformational selection or induced fit: a flux description of reaction mechanism.* Proc Natl Acad Sci U S A, 2009. **106**(33): p. 13737-41.
10. Buell, A.K., et al., *Population of nonnative states of lysozyme variants drives amyloid fibril formation.* J Am Chem Soc, 2011. **133**(20): p. 7737-43.
11. Hollien, J. and S. Marqusee, *A thermodynamic comparison of mesophilic and thermophilic ribonucleases H.* Biochemistry, 1999. **38**(12): p. 3831-6.
12. Razvi, A. and J.M. Scholtz, *Lessons in stability from thermophilic proteins.* Protein Sci, 2006. **15**(7): p. 1569-78.
13. Becktel, W.J. and J.A. Schellman, *Protein stability curves.* Biopolymers, 1987. **26**(11): p. 1859-77.
14. Myers, J.K., C.N. Pace, and J.M. Scholtz, *Denaturant m values and heat capacity changes: relation to changes in accessible surface areas of protein unfolding.* Protein Sci, 1995. **4**(10): p. 2138-48.
15. Robic, S., et al., *Role of residual structure in the unfolded state of a thermophilic protein.* Proc Natl Acad Sci U S A, 2003. **100**(20): p. 11345-9.
16. Robic, S., J.M. Berger, and S. Marqusee, *Contributions of folding cores to the thermostabilities of two ribonucleases H.* Protein Sci, 2002. **11**(2): p. 381-9.
17. Guzman-Casado, M., et al., *Energetic evidence for formation of a pH-dependent hydrophobic cluster in the denatured state of Thermus thermophilus ribonuclease H.* J Mol Biol, 2003. **329**(4): p. 731-43.
18. Sankararaman, S., B. Kolaczkowski, and K. Sjolander, *INTREPID: a web server for prediction of functionally important residues by evolutionary analysis.* Nucleic Acids Res, 2009. **37**(Web Server issue): p. W390-5.
19. Chung, J.L., W. Wang, and P.E. Bourne, *Exploiting sequence and structure homologs to identify protein-protein binding sites.* Proteins, 2006. **62**(3): p. 630-40.
20. Lockless, S.W. and R. Ranganathan, *Evolutionarily conserved pathways of energetic connectivity in protein families.* Science, 1999. **286**(5438): p. 295-9.

21. Mirny, L. and E. Shakhnovich, *Evolutionary conservation of the folding nucleus.* J Mol Biol, 2001. **308**(2): p. 123-9.
22. Chi, C.N., et al., *Reassessing a sparse energetic network within a single protein domain.* Proc Natl Acad Sci U S A, 2008. **105**(12): p. 4679-84.
23. Plaxco, K.W., et al., *Evolutionary conservation in protein folding kinetics.* J Mol Biol, 2000. **298**(2): p. 303-12.
24. Nickson, A.A. and J. Clarke, *What lessons can be learned from studying the folding of homologous proteins?* Methods, 2010. **52**(1): p. 38-50.
25. Zarrine-Afsar, A., S.M. Larson, and A.R. Davidson, *The family feud: do proteins with similar structures fold via the same pathway?* Curr Opin Struct Biol, 2005. **15**(1): p. 42-9.
26. Bridgham, J.T., S.M. Carroll, and J.W. Thornton, *Evolution of hormone-receptor complexity by molecular exploitation.* Science, 2006. **312**(5770): p. 97-101.
27. Thornton, J.W., E. Need, and D. Crews, *Resurrecting the ancestral steroid receptor: ancient origin of estrogen signaling.* Science, 2003. **301**(5640): p. 1714-7.
28. Thomson, J.M., et al., *Resurrecting ancestral alcohol dehydrogenases from yeast.* Nat Genet, 2005. **37**(6): p. 630-5.
29. Gaucher, E.A., et al., *Inferring the palaeoenvironment of ancient bacteria on the basis of resurrected proteins.* Nature, 2003. **425**(6955): p. 285-8.
30. Gaucher, E.A., S. Govindarajan, and O.K. Ganesh, *Palaeotemperature trend for Precambrian life inferred from resurrected proteins.* Nature, 2008. **451**(7179): p. 704-7.
31. Perez-Jimenez, R., et al., *Single-molecule paleoenzymology probes the chemistry of resurrected enzymes.* Nat Struct Mol Biol, 2011. **18**(5): p. 592-6.
32. Hobbs, J.K., et al., *On the origin and evolution of thermophily: reconstruction of functional precambrian enzymes from ancestors of Bacillus.* Mol Biol Evol, 2012. **29**(2): p. 825-35.
33. Ball, P., *Water as an active constituent in cell biology.* Chem Rev, 2008. **108**(1): p. 74-108.
34. Lind, P.A., et al., *Esterase catalysis of substrate vapour: enzyme activity occurs at very low hydration.* Biochim Biophys Acta, 2004. **1702**(1): p. 103-10.
35. Beauchamp, K.A., et al., *Are Protein Force Fields Getting Better? A Systematic Benchmark on 524 Diverse NMR Measurements.* J Chem Theory Comput, 2012. **8**(4): p. 1409-1414.
36. Gabel, F., et al., *Protein dynamics studied by neutron scattering.* Q Rev Biophys, 2002. **35**(4): p. 327-67.
37. Schiró, G., F. Natali, and A. Cupane, *Physical Origin of Anharmonic Dynamics in Proteins: New Insights From Resolution-Dependent Neutron Scattering on Homomeric Polypeptides.* Physical Review Letters, 2012. **109**(12): p. 128102.
38. Wood, K., et al., *A benchmark for protein dynamics: Ribonuclease A measured by neutron scattering in a large wavevector-energy transfer range.* Chem Phys, 2008. **345**(2–3): p. 305-314.
39. Wood, K., et al., *Coincidence of dynamical transitions in a soluble protein and its hydration water: direct measurements by neutron scattering and MD simulations.* J Am Chem Soc, 2008. **130**(14): p. 4586-7.
40. Gallat, F.X., et al., *Dynamical coupling of intrinsically disordered proteins and their hydration water: comparison with folded soluble and membrane proteins.* Biophys J, 2012. **103**(1): p. 129-36.

41. Wood, K., et al., *Coupling of protein and hydration-water dynamics in biological membranes.* Proc Natl Acad Sci U S A, 2007. **104**(46): p. 18049-54.
42. Tompa, P., *Unstructural biology coming of age.* Curr Opin Struct Biol, 2011. **21**(3): p. 419-25.
43. Monod, J., J. Wyman, and J.P. Changeux, *On the Nature of Allosteric Transitions: A Plausible Model.* J Mol Biol, 1965. **12**: p. 88-118.
44. Gunasekaran, K., B. Ma, and R. Nussinov, *Is allostery an intrinsic property of all dynamic proteins?* Proteins, 2004. **57**(3): p. 433-43.
45. Wright, C.M., R.A. Heins, and M. Ostermeier, *As easy as flipping a switch?* Curr Opin Chem Biol, 2007. **11**(3): p. 342-6.
46. Guntas, G. and M. Ostermeier, *Creation of an allosteric enzyme by domain insertion.* J Mol Biol, 2004. **336**(1): p. 263-73.
47. Suel, G.M., et al., *Evolutionarily conserved networks of residues mediate allosteric communication in proteins.* Nat Struct Biol, 2003. **10**(1): p. 59-69.
48. Tsai, C.J., A. del Sol, and R. Nussinov, *Allostery: absence of a change in shape does not imply that allostery is not at play.* J Mol Biol, 2008. **378**(1): p. 1-11.
49. Radley, T.L., et al., *Allosteric switching by mutually exclusive folding of protein domains.* J Mol Biol, 2003. **332**(3): p. 529-36.
50. Cutler, T.A. and S.N. Loh, *Thermodynamic analysis of an antagonistic folding-unfolding equilibrium between two protein domains.* J Mol Biol, 2007. **371**(2): p. 308-16.
51. Ha, J.H., et al., *Modular enzyme design: regulation by mutually exclusive protein folding.* J Mol Biol, 2006. **357**(4): p. 1058-62.

CHAPTER 2

Exploring evolution of protein energy landscapes using ancestral protein resurrection

Conducted in collaboration with Michael Harms and Joseph Thornton (U. Oregon, Eugene); Bryan Schmidt and James Berger (U. California, Berkeley)

2.1 Abstract

In this study, we use ancestral protein resurrection to ask whether the observed thermodynamic differences between *Escherichia coli* (ecRNH) and its homolog from *Thermus thermophilus* (ttRNH) can be understood from an evolutionary perspective. Several evolutionary intermediates, including the most recent common ancestor of ecRNH and ttRNH, are reconstructed and characterized. We observe pronounced trends in melting temperature, reversibility and global stability along each lineage; however other features, most notably ΔC_p, are shown to fluctuate stochastically in a "neutral corridor" defined by the extant proteins. The distinctive trends in melting temperature between the mesophilic and thermophilic lineages show that the mesophilic T_m was maintained over evolutionary time, and thermostability of ttRNH developed via a gradual, potentially adaptive process.

2.2 Introduction

Studies of homologous proteins can be used both to explore how an organism's living environment relates to its proteins' energetics and to dissect how these biophysical properties are encoded by sequence. Comparing thermophiles and mesophiles has revealed that proteins from thermophiles tend to have higher melting temperatures than their mesophilic counterparts [4], which is consistent with the idea that proteins must be folded to function. Because homologs generally share similar folds, such differences in behavior are not always apparent from their three-dimensional structures. Instead, energetic properties emerge from subtle sequence variation that does not affect the overall fold. It has been observed, for instance, that increased melting temperatures can be accomplished by strengthening enthalpic interactions in the folded state with the addition of disulfide bonds or salt bridges [5]. A resulting model from these studies is that thermostability represents a global property, which is distributed throughout the structure and can be encoded by sequence in a number of different ways. While several thermodynamic strategies can be used to ensure folding at elevated temperatures, the questions remains whether a protein's evolutionary history informs which specific strategies are employed by a given protein family.

The finer details of how temperature affects protein function are largely system-specific and may be expected to differ between classes of proteins, such as enzymes and structural proteins; however, some general principles emerge from the fact that proteins must be folded in order to function. For instance, thermophilic proteins appears to be more stable than their mesophilic homologs at all temperatures [6], and a protein's melting temperature has been shown to correlate with the optimal growth temperature of its organism [7]. Interestingly, it has also been observed that at their respective source-organisms' growth temperatures, homologous proteins share similar global stabilities. One example is ribonuclease H, a RNA-DNA duplex hydrolase involved in DNA replication. Ribonuclease HI from the mesophile *E. coli* (ecRNH) has a stability of 6.5 kcal/mol at 37 °C, while its homolog in *T. thermophilus* (ttRNH) has a stability of 5.5 kcal/mol at its living temperature of 68 °C [8]. This small difference in stability is remarkable considering the two proteins differ by 20 °C in their melting temperatures, and it suggests that the enzyme's energetics have been tuned to accommodate environmental conditions.

Energetic properties emerge from interactions throughout a protein's structure and are not simply encoded by one or even a few specific residues. Thus it is useful to describe how a protein tunes its energetics in terms of thermodynamic strategies rather than focusing strictly on specific interations. Thermodynamic strategies for increasing thermostability in particular can be illustrated with changes in a protein's stability curve, which is described by the Gibbs-Helmholtz equation [9]:

$$\Delta G\ (T) = \Delta H_m\ \left(1 - \frac{T}{T_m}\right) - \Delta C_p\ [\ (T_m - T) + \ T\ ln\left(\frac{T}{T_m}\right)\ ] \qquad (1)$$

where the global stability at any temperature ($\Delta G\ (T)$) is a function of the change in enthalpy at the T_m (ΔH_m), the change in heat capacity upon unfolding (ΔC_p), and the melting temperature of the protein (T_m) [9]. Stability curves demonstrate how protein stability depends upon temperature, and changes to a curve's shape or position are directly related to changes in a protein's structure and sequence. The curve intersects the abscissa at two points, which reflect a protein's two melting temperatures, one at a low temperature and another at the more familiar thermal-denaturation temperature. Cold denaturation most often occurs at non-physiological conditions, so the more relevant parameter for understanding a thermophilic protein's resistance to denaturation is its high melting temperature, or T_m. A protein's T_m can increase in one of three ways (**Figure 1**). Razvi and Scholtz formalize these strategies as "method I," where the entire curve is upshifted; "method II," where the curve is broadened; and "method III," where the curve is right-shifted [4]. All transformations result in an increased T_m and correspond with distinct changes in the other thermodynamic parameters, and thus structural interactions.

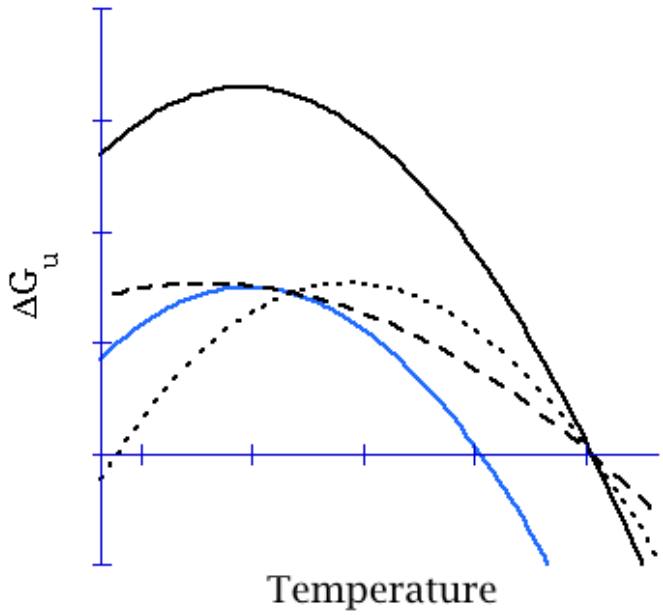


Figure 1. Thermodynamic strategies for increasing T_m. Relative to the reference state (blue line), the stability curve can be upshifted (method I, solid black line), broadened (method II, dashed line) or right-shifted (method III, dotted line).

In order to intuit how the entropic and enthalpic components of free energy change for each method, it is useful to first consider changes in the temperature of maximum stability, T_s. At this temperature, the change in entropy between the folded and unfolded states is zero, and folding is enthalpically driven. It has been observed that most proteins, regardless of thermostability or species of origin, are maximally stable near 20 °C [10]. This is attributed to the fact that the major entropic contribution of folding comes from transferring hydrophobic residues from water

to the protein core. Because this so-called hydrophobic effect influences all globular proteins, the value for T_s does not vary much between proteins. Accordingly, right-shifting the curve (method III) is the strategy least commonly observed in thermophilic proteins, because it requires decreasing the ΔS for folding without compensatory changes in ΔH. On the other hand, increasing the change in enthalpy at the T_s (ΔH_s), which results in an upshifted curve (method I), can be achieved simply by strengthening interactions in the folded state. This method is the most commonly observed strategy in thermophilic proteins, perhaps due to the large number of ways of increasing non-covalent interactions through sequence variation. Another common thermophilic strategy is broadening the curve (method II), which requires lowering the ΔC_p. Curvature in the plot, which arises from the fact that the heat capacity of the unfolded state is larger than that of the folded state, has been found to correlate with changes in the solvent accessible surface area upon folding. Lowering the ΔC_p requires structural changes that effect the change in solvent accessible surface area, such as retaining some residual structure in the unfolded state. Based on the approximately 20 existing case studies of natural homologs, thermophilic proteins tend to have both up-shifted (method I) and broadened (method II) curves relative to their mesophilic counterparts [4].

One of the first examples of method II thermostabilization was identified in the enzyme RNase H [8]. While the crystal structures of the thermophilic and mesophilic enzymes overlay with an RMSD less than 2 Å, indicating that the two share a similar folded state, ttRNH has a significantly smaller ΔC_p. [8]. By swapping structural domains between the two homologs, it was found that the low ΔC_p tracks with the core domain, evidently because it contains residues involved in the residual structure [11]. A later study demonstrated that a single point mutation within the core domain was sufficient to increase ttRNH's ΔC_p to the value measured for ecRNH [12]. Differential scanning calorimetry was used to confirm the presence of hydrophobic clusters in the unfolded state of ttRNH [13].

While several thermodynamic strategies theoretically can be used to ensure foldedness at elevated temperatures, the question remains open as to why particular proteins prefer one method, or combination of methods, over the others. It is tempting to draw conclusions from homology studies based on fold topology or functional considerations. Perhaps, for instance, there are more ways to increase enthalpic interactions in the folded states of spherical rather than cylindrical proteins, creating a preference for method I. But this analysis ignores how a protein family's evolutionary history informs which specific strategies are employed. Physical determinants certainly play a role in shaping the evolution of macromolecules, but shared ancestry is often the most compelling explanation for existing similarities. Each of the strategies described fundamentally depends upon sequence variation, which arises via evolutionary mechanisms. A few biophysicists argue that evolution has sampled every possible sequence for a protein of average length [14]. Having fully explored sequence, and thus structural, space, it follows that every protein is optimized for its particular environment and function, in which case evolutionary lineage is irrelevant. The more common view in evolutionary biology, however, is that evolution is largely a contingent process, where solutions depend upon their ancestors' sequences [15]. Just as evolutionary biologists do not assume life exists in its current forms because it *must*, protein biochemists should be similarly cautious when interpreting *why* particular proteins use particular biophysical strategies.

Studying only existing homologs is limiting when trying to understand how proteins evolved. This is especially true for deciphering how global biophysical traits are encoded, because homologs can vary at many functionally irrelevant positions, making it difficult to identify the source for these subtly encoded traits. But there is another way to harness information from phylogenetic relationships in a more directed fashion. Using computational methods, such as maximum parsimony and maximum likelihood, extinct states can be inferred from an alignment of an existing gene family in a process known as ancestral protein resurrection (APR) [16]. These ancestors can then be synthesized and studied experimentally. By incorporating a time axis into the analysis, this methodology captures more information from the alignment than conserved sites alone and enables more nuanced relationships within families to be explored. The resulting ancestral proteins represent relevant reference states against which their divergent descendants can be compared. In a study of EF-tus, all the resurrected proteins demonstrate maximal binding to GTP, the modern-day ligand, at higher temperatures than extant proteins. The authors suggest that the bacteria containing these ancient EF-Tus were thermophilic, possibly reflecting ancient extreme environments [17, 18]. Consistent with the idea of a hotter early Earth, the melting temperatures of Precambrian thioredoxins are as much as 32 °C higher than current thioredoxins. The resurrected enzymes share mechanistic traits with their descendants, but their activities are more tolerant of acidic conditions, which might have been necessary in the low pH environment of ancient oceans [19]. In contrast, the melting temperatures of enzymes involved in leucine biosynthesis do not decrease monotonically as a function of evolutionary time [20]. Instead, some of the ancestors have thermophile-like T_ms, while others do not, leading the authors to conclude that thermophilicity has evolved multiple times during the course of this enzyme's evolution. These studies demonstrate the utility of the APR methodology in understanding how protein energetics are shaped by evolutionary history, but do not address whether thermodynamic mechanisms are conserved within a family.

Using computational methods, such as maximum parsimony and maximum likelihood, ancestral states can be inferred from an alignment of an existing gene family [16]. These ancestors can then be synthesized and studied experimentally. By incorporating a time axis into the analysis, this methodology, known as ancestral protein resurrection (APR), captures more information from the alignment than conserved sites alone and enables more nuanced relationships within families to be explored. The resulting ancestral proteins represent relevant reference states against which their divergent descendants can be compared.

This study aims to understand from a historical perspective the thermodynamic differences between a mesophilic and thermophilic ribonuclease HI (RNase H). Much work already has been done to elucidate the similarities and differences between *Escherichia coli* (ecRNH) and its homolog from *Thermus thermophilus* (ttRNH), and here we resurrect their most recent common ancestor as well as evolutionary intermediates along both lineages to identify how traits, such as melting temperature, stability and heat capacity, changed over evolutionary time.

2.3 Results

2.3.1 Ancestral protein resurrection

RNase H is particularly well suited for APR due its conservation throughout all domains of life and the resulting plethora of available sequences. A phylogeny was constructed using 409 representative bacterial and archael RNase H sequences, which were curated from the NCBI protein database. Redundancy was minimized using a 95% similarity cutoff, and RNases H comprising only one domain within a larger protein were not included. Multiple sequence alignment was performed initially with MUSCLE and further refined manually. While RNases H vary in length, many positions are well conserved, including the three active site residues D10, E48 and D70 denoted with a star in **Figure 2**.

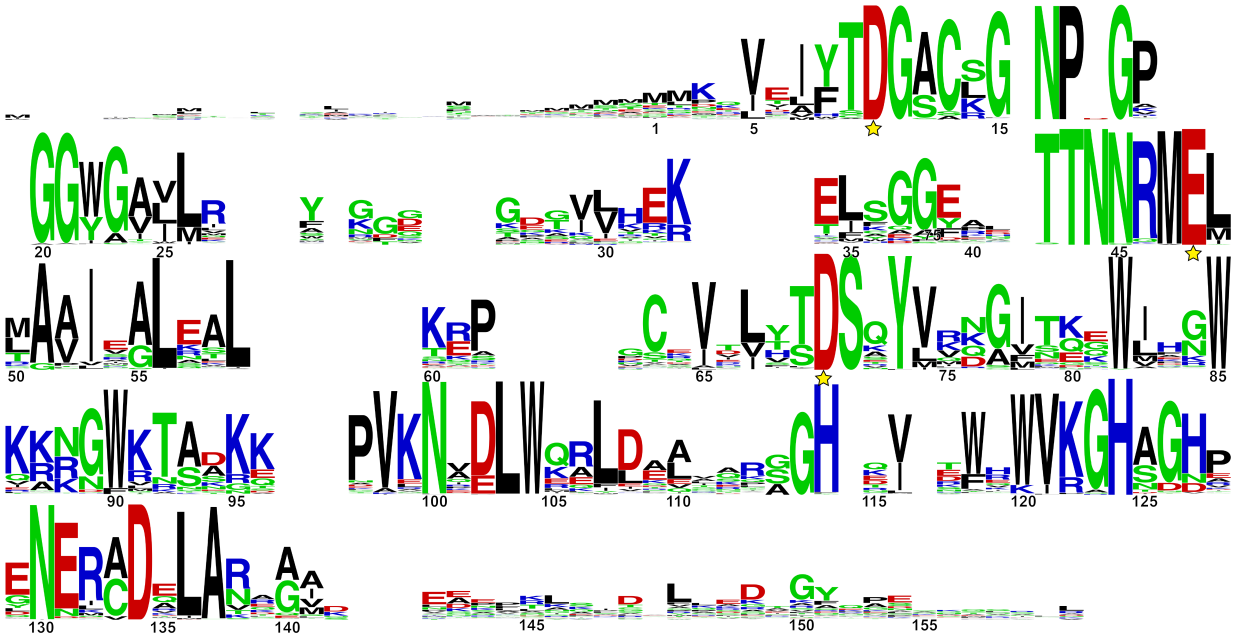


Figure 2. WebLogo representation of the RNase H multiple sequence alignment [3]. Conservation is reflected by the overall height of the stack at each position. Height of individual letters within the stack indicates the relative frequency of a residue at the position. Numbering is based on ecRNH. Active site residues are starred.

The maximum likelihood phylogenetic tree was constructed using the JTT+Γ_8 substitution model and SPR moves as implemented in PhyML 3.0 [21, 22]. Branch supports were estimated using the approximate likelihood ratio test [23]. Maximum likelihood ancestral RNases H were reconstructed with the maximum likelihood topology, branch lengths and phylogenetic model using PAML 3.14 [24, 25]. The locations within the tree for the 8 resurrected sequences are indicated with stars in **Figure 3**. The archaeal sequences cluster together to the exclusion of all other clades, allowing the tree to be rooted. Branch lengths indicate the sequence distance between nodes and are given in units of average substitutions per position. For example, the branch length between Node 1 and Node A is 0.1205, which means the maximum number of positions that could differ between the sequences is 12%; however, a pairwise alignment reveals

that the sequences only differ by 8%. This discrepancy is due to the fact that branch lengths are estimates based on a model of sequence evolution that corrects for unobserved changes. Some positions turnover multiple times, resulting in longer branch lengths. Thus, branch length is not strictly equal to observed sequence differences; rather, it should be thought of as the amount of time a particular sequence has had to evolve. Strictly speaking, we have not attempted to convert branch lengths into time, because it would require making questionable assumptions about bacterial evolution. Because the tree is rooted with an uncontroversial outgroup, the order of the nodes does reflect their order of appearance.

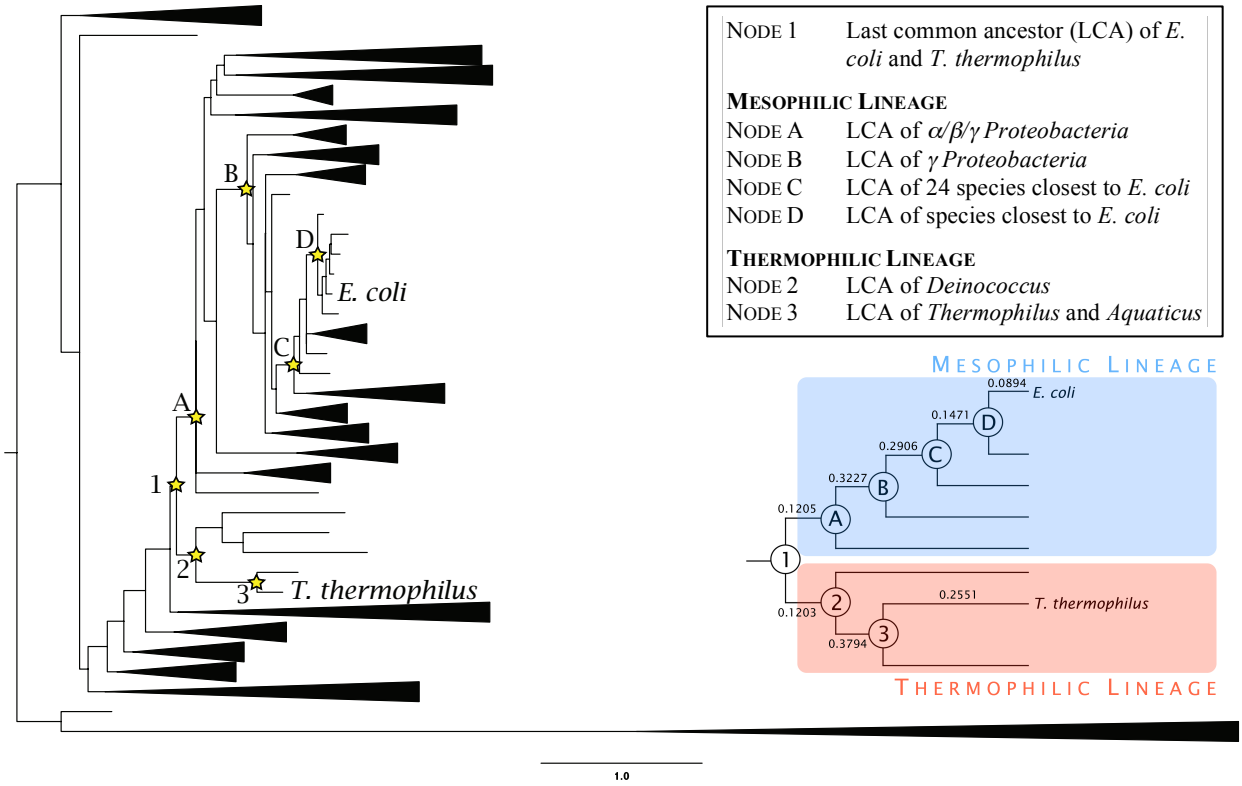


Figure 3. Phylogenetic tree built from RNase H alignment. Branch length reflects sequence distance, as indicated by the scale bar, in average number of substitutions per position. Resurrected nodes are starred. (Inset, top) Description of each resurrected node (Inset, bottom) Cladogram version of tree labeled with branch lengths.

Node 1 represents the most recent common ancestor to ecRNH and ttRNH. Other nodes were chosen for their statistical support as well as their analogous spacing along the two lineages. Nodes 2 and A, for example, both share 92% sequence identity with Node 1 but are only 82% identical to one another. Similarly, Nodes 3 and B are equally distanced from Node 1 with identities of 77% and 70%, respectively (**Figure 4A**). Aligning the ancestral sequences makes it evident that the nodes are quite similar, and conserved positions do not appear to correspond with secondary structure elements observed in ecRNH (**Figure 4B**).

All resurrected nodes are reasonably well supported (**Table 1**). Support was estimated using the approximate likelihood ratio (LR) test, in which the likelihood of a tree is divided by the

A

	ttRNH	3	2	1	A	B	C	D	ecRNH
ttRNH		87	73	69	64	57	54	55	52
3			82	77	70	61	55	56	53
2				92	85	66	61	61	60
1					92	70	63	63	63
A						74	66	67	65
B							76	75	73
C								89	86
D									93
ecRNH									

B

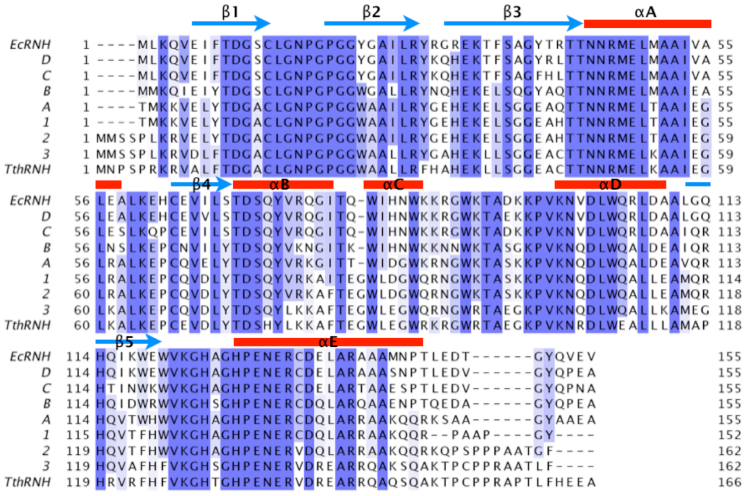


Figure 4. (A) Alignment of ancestors with ecRNH and ttRNH. Secondary structure elements are based on ecRNH. (B) Sequence identity matrix for ancestors, ecRNH and ttRNH. Ancestors that are analogously spaced along the thermophilic and mesophilic lineages appear in the same color.

likelihood of the next most likely tree that does not contain the node of interest. This statistic has a straightforward interpretation. For example, the LR of Node B means that it is 1.3 x 10^6 times more likely that this node existed than it did not given the experimental data and amino acid substitution model. The most weakly supported node, Node 1, is the last common ancestor of ecRNH and ttRNH and has an LR of 30.5. The low value arises from the small number of thermophilic sequences in the data set, which causes ambiguity in the placement of the clade containing *T. thermophilus*. This is relatively poor support. It is not expected to alter the sequence of the reconstructed ancestor, however, because the node immediately preceding it is both well supported (LR = 1043) and separated from node 1 by a short internal branch of length 0.0367 substitutions per site. It has been demonstrated that in these situations, reconstructed sequences are robust to ambiguities in the tree topology [20]. For all of the nodes, in fact, most positions in the sequence are unambiguous. The highest posterior probability for each position is averaged across the entire sequence to give the mean posterior probability. A value of 0.9 means that for each position, the probability of the chosen residue is 90%. Many positions have a posterior probability of 1, but the average is lowered by greater uncertainty in the termini, which are unstructured in crystal structures [21, 22].

Table 1. Statistical support for resurrected ancestors

Ancestor	Likelihood Ratio	Mean Posterior Probability
1	30.5	0.925
2	148	0.858
3	1.74 x 10^{16}	0.912
A	6.62 x 10^4	0.947
B	1.30 x 10^6	0.957
C	8.45 x 10^6	0.954
D	4.90 x 10^4	0.977

The distribution of posterior probabilities across the sequence of Node 1 is shown in **Figure 5**. In all cases, the residue with the highest posterior probability was chosen. In many instances where the posterior probabilities for the first and second most probable residue were similar, the amino acids were also chemically similar. For example, at position 57, the posterior probabilities were 0.55 and 0.45 for arginine and lysine, respectively, so an arginine was used. Genes encoding the ancestoral proteins were codon optimized for expression in *E. coli* and synthesized by GENEART (Regensburg, Germany).

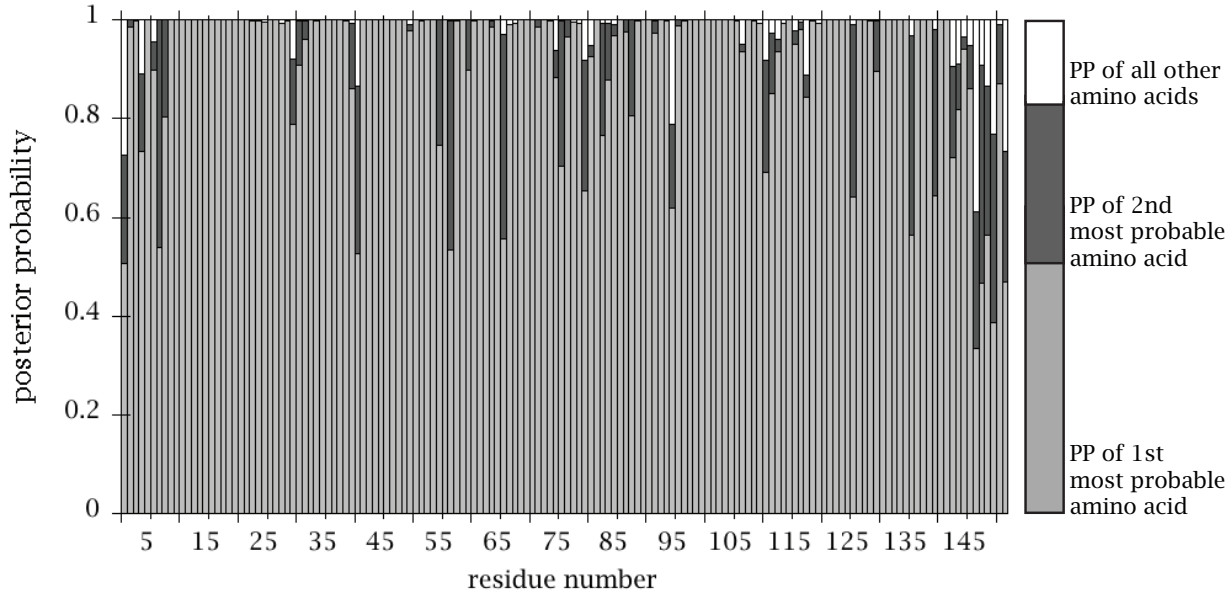


Figure 5. Distribution of posterior probabilities across the sequence for Node 1. Many positions are unambiguous with a PP = 1, and the greatest uncertainty is at the termini.

2.3.2 Structural characterization by CD and X-ray diffraction

Our first goal was to characterize the structural and functional properties of the ancestors. All ancestors were subcloned into pET27 vectors and expressed in BL21(DE3) pLys S cells. Proteins expressed solubly and were purified over a heparin column followed by an S column to >95% purity. Far-UV circular dichroism (CD) spectroscopy was used to determine whether the ancestral proteins fold. While the exact CD profiles vary slightly between the ancestors, their spectra indicate that they are folded and contain significant secondary structure at 25 °C (**Figure 6**). Far-UV CD primarily probes secondary structure content but is also sensitive to small structural changes, such as twists in β-sheets and specific environment of tryptophan residues. This makes it difficult to assess the overall structural similarity between proteins that differ by more than a few residues; however, all the ancestors display spectra that are consistent with variability in the extant RNases H.

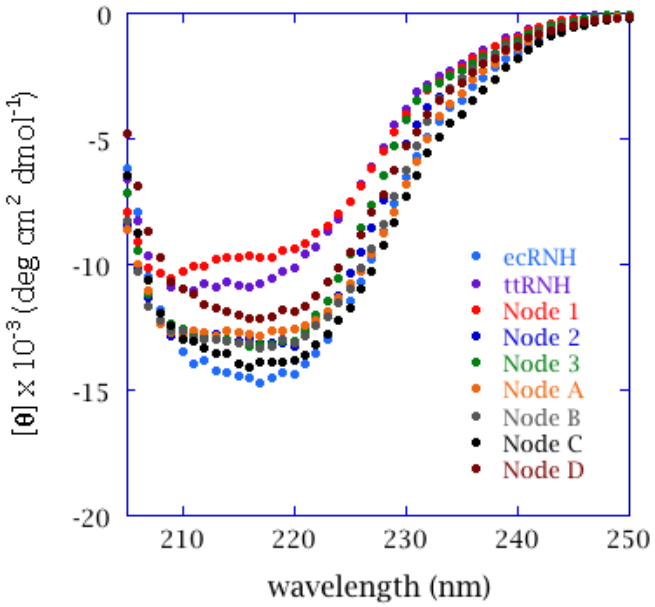


Figure 6. CD spectra of ancestors, ecRNH and ttRNH at 25 °C in 20 mM NaOAc (pH 5.5), 50 mM KCl and 1 mM TCEP.

The most robust analysis of structural similarity requires high-resolution techniques such as NMR or X-ray crystallography. The latter was used here to determine whether a representative ancestor, Node C, adopts the same RNase H fold observed in ecRNH and tRNH. Attempts to grow crystals with Node 1 were met with limited success, as the crystals assumed a stacked-plate morphology that proved unsuitable for diffraction studies. Node C, however, formed long, rod-shaped crystals that diffracted to 1.3 Å (**Figure 7**).

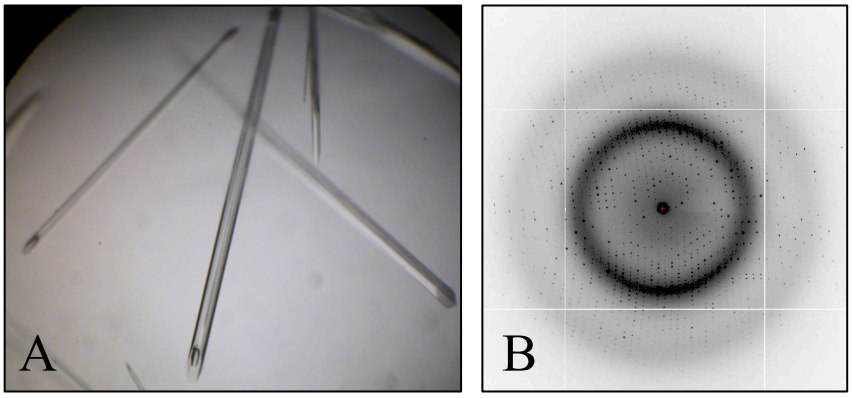


Figure 7. (A) Rod-shaped crystals of Node C. (B) Diffraction pattern from Node C crystal.

Initial phases were calculated by molecular replacement (MR) using PHASER [26]. The search model was the extant RNase H from *E. coli* (PDB ID code 2RN2). Sequence variations between ecRNH and Node C were readily apparent in the initial positive and negative difference density, thus enabling rapid modeling of the differences. The final model included residues 1-153 and 2 sulfate ions and was refined to an R_{work}/R_{free} of 13.36/16.60. 98.7% of main-chain torsion angles fell within favored regions of Ramachandran space, as calculated by MolProbity, with only 1.3% in allowed regions and 0% in disallowed (**Table 2**). Node C's backbone, as defined by the α-carbons, overlay well with structures from both ecRNH, RMSD of 0.8 Å, and ttRNH, RMSD of 1.3 Å (**Figure 8**). The crystal structure of Node C confirms that this ancestor adopts the canonical RNase H fold.

Table 2. Data collection and refinement statistics

Data Collection Statistics	
space group	$P6_1$
cell dimensions	
a, *b*, *c* (Å)	82.38, 82.38, 44.65
a, b, g (°)	90, 90, 120
resolution (Å)	50 – 1.36 (1.41 – 1.36)
no. of reflections	37219
R_{sym}	7.9 (85.8)
$\langle I / \sigma I \rangle$	26.1 (2.7)
completeness (%)	99.1 (99.7)
redundancy	7.6 (7.2)
Refinement Statistics	
resolution (Å)	50 – 1.36
R_{work} / R_{free}	13.5 / 16.7
B-factors	
protein	18.7
water	28.6
sulfate	25.5
no. of protein atoms	2642
no. of water atoms	237
no. of sulfates	2
rmsd	
Bond lengths (Å)	0.012
Bond angles (°)	1.3
Ramachandran (favored/disallowed)	98.0/0.0

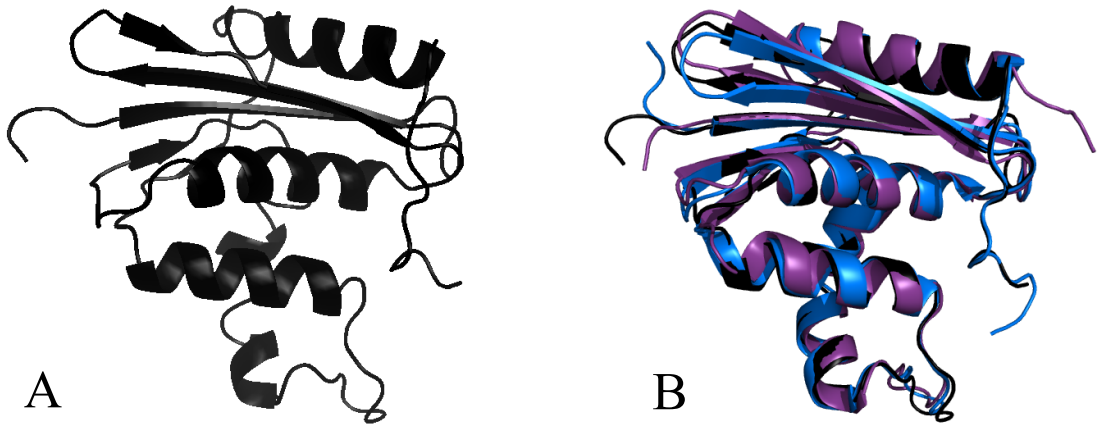


Figure 8. (A) Ribbon representation for Node C. (B) Superposition of Node C with ecRNH (2RN2) and ttRNH (1RIL).

2.3.3 Catalytic activity

Ancestors were also assayed for their ability to degrade RNA-DNA hybrids. RNase H hydrolyzes 3'-phosphodiester bonds of RNA in RNA-DNA hybrids in a Mg^{2+}-dependent manner [27]. The enzyme is a processive nuclease that first cleaves the RNA strand endonucleolytically

and then continues to hydrolyze RNA exonucleolytically. It also has minimal sequence specificity [28]. Activity was measured using two different assays, one based on the intrinsic hyperchromicity of the reaction products and another using a fluorescent beacon substrate.

All ancestors demonstrated RNase H activity using the hyperchromic assay, which takes advantage of the fact that liberated bases absorb more strongly at 260 nm than either single- or double-stranded nucleic acids. Substrate is prepared by annealing dT_{20} oligomers to poly-rA strands, and the given concentrations assume complete annealing. Reactions were performed with 5 nM enzyme in 10 mM Tris (pH 8), 50 mM NaCl, 10 mM $MgCl_2$, 1 mM TCEP and 16.7 µg/mL substrate. All ancestors are active at 25 °C, which suggests that they share similar native structures with ecRNH, ttRNH and Node C (**Figure 9**).

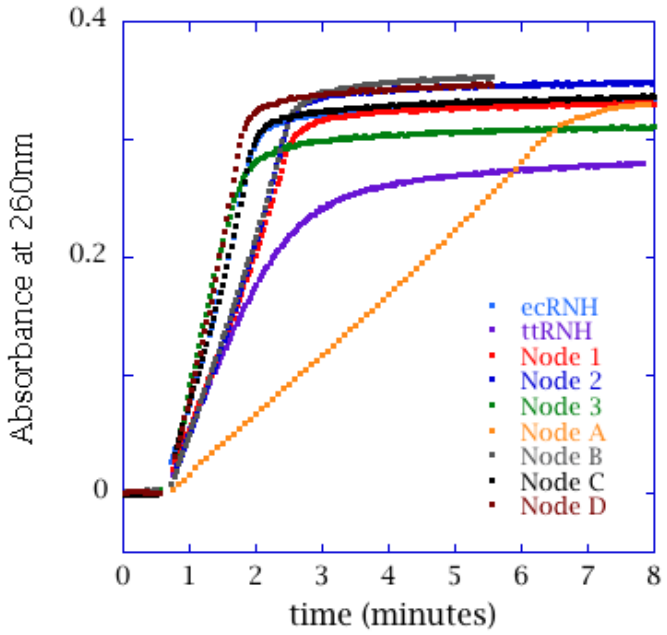


Figure 9. Hyperchromic activity assay for all ancestors, ecRNH and ttRNH at 25°C in 10 mM Tris (pH 8), 50 mM NaCl, 10 mM $MgCl_2$, 1 mM TCEP and 16.7 µg/mL poly-rA:dT_{20} substrate.

In order to compare catalytic efficiencies, it is necessary to perform full Michaelis-Menten analysis. Heterogeneity of the dT_{20}:poly-rA substrate and the mixed endo- and exonucleolytic activity of RNase H complicate this analysis. Thus, a more straightforward, fluorescence-based assay was developed. This assay, based on a published method, is sensitive only to the first endonucleolytic cleavage per molecule of substrate, because the short RNA-DNA chimera dissociates post-hydrolysis [29]. The substrate is prepared by annealing a 5'-fluorescein-labeled DNA 10-mer to a complementary 3'-DABCYL-labeled RNA 10-mer. As the RNA is cleaved, the strands separate and fluorescein is liberated from the quencher, DABCYL, resulting in increased fluorescence. Michaelis-Menten analysis of ecRNH, ttRNH and nodes 1 and 2 were performed using this assay (**Figure 10**). At 25 °C, ttRNH is the least efficient enzyme; ecRNH has twice the catalytic efficiency of ttRNH, and nodes 1 and 2 have three times its efficiency (**Table 3**). Interestingly, Node 2 has a significantly lower K_m than the other enzymes, perhaps reflecting a higher affinity for the substrate. Both ancestors contain all known substrate-binding residues, which are identified based on contacts with a 12-mer RNA-DNA hybrid visible in the co-crystal of an RNase H from *B. halodurans* [30]. It is unclear how Node 2's lower K_m is specifically encoded by its sequence.

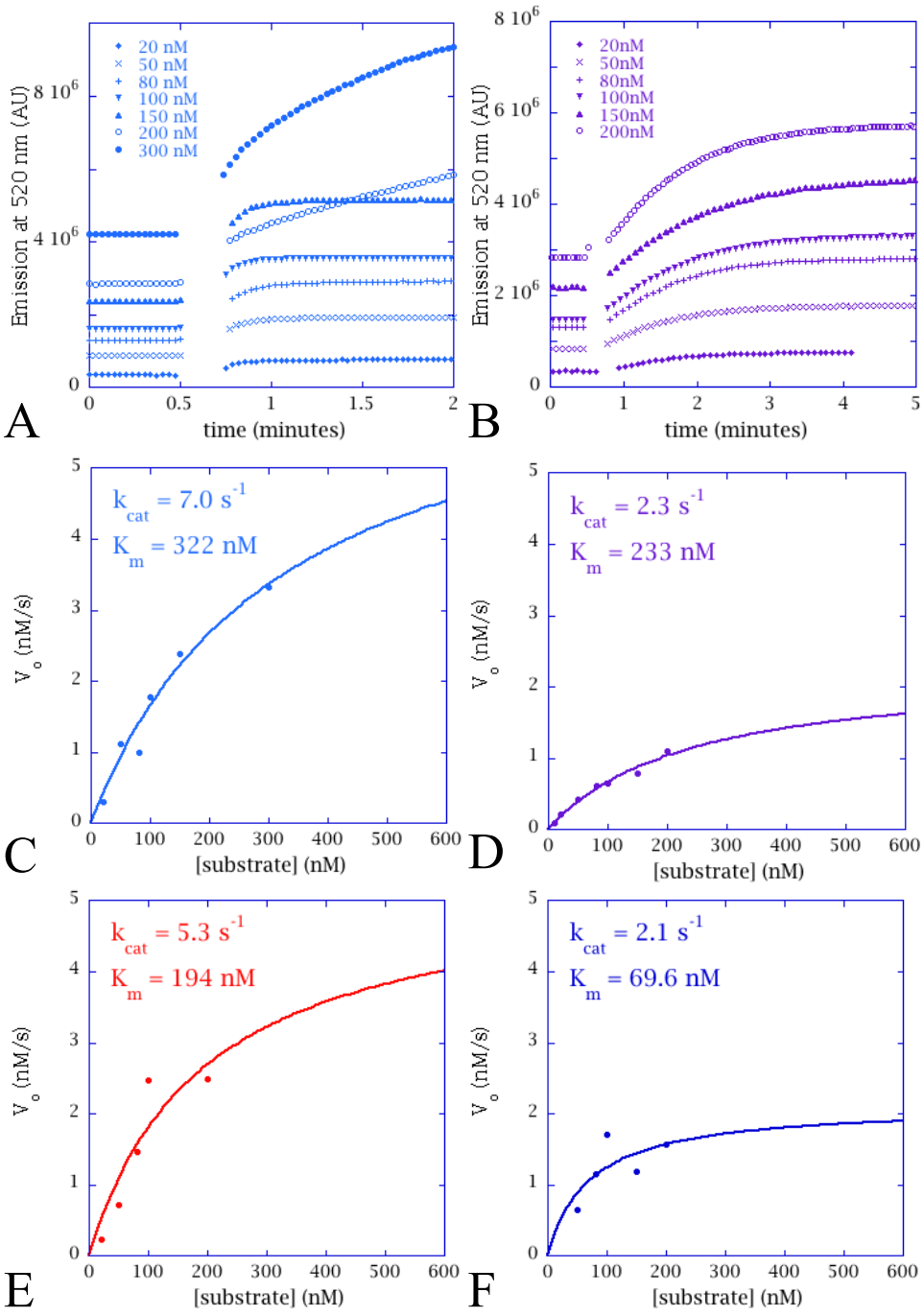


Figure 10. Fluorescence activity assay. Representative reactions with variable concentrations of beacon substrate for (A) ecRNH and (B) ttRNH at 25°C. Michaelis-Menten plots for (C) ecRNH (D) ttRNH (E) Node 1 and (F) Node 2.

Table 3. Catalytic efficiencies with beacon substrate at 25 °C

	k_{cat} (s^{-1})	K_m (nM)	k_{cat} / K_m (s^{-1} nM^{-1})
ecRNH	7.0	322	0.022
ttRNH	2.3	233	0.010
Node 1	5.3	194	0.027
Node 2	2.1	69.6	0.030

Temperature-dependence of activity could not be assessed using either assay, because both substrates denature considerably at temperatures much above 25 °C. The T_m of the beacon substrate under activity conditions is predicted to be approximately 35 °C [31], and the dT_{20}:poly-rA substrate is expected to be even less thermostable. Ongoing work will determine whether another substrate can be used for assaying activity at higher temperatures.

2.3.4 Thermodynamic characterization

2.3.4.1 Thermal denaturation

Our next goal was to understand the evolutionary path that gave rise to the distinctive energetics observed in modern-day mesophilic ecRNH and thermophilic ttRNH. To this end, we rigorously characterized the energetics of the ancestors by measuring stabilities. Thermostabilities were measured by tracking changes in CD signal at 222 nm as a function of temperature (**Figure 11**). The midpoint of the transition, or T_m, represents the temperature at which half the molecules are folded and half are unfolded. This interpretation assumes that only two states are populated at each temperature, and that each measurement is taken at equilibrium. ttRNH unfolds reversibly with temperature, but ecRNH visibly crashes out of solution during the course of the experiment. Three of the ancestors also precipitate, indicating that at some temperature the measurements are no longer taken at equilibrium. For these ancestors and ecRNH, the midpoints of the transition represent apparent T_ms and are not true thermodynamic parameters. Reversibility is defined here as 80% recovery of signal at 222 nm when cooled to 25 °C post-thermal denaturation. By this definition, ttRNH and nodes 3, 2, 1 and A all unfold reversibly (**Figure 12A**), while eRNH and nodes D, C and B do not (**Figure 12B**). While caution must be used when interpreting apparent T_ms, we consider them here only in comparison with true T_ms measured for the reversible species. (**Table 4**)

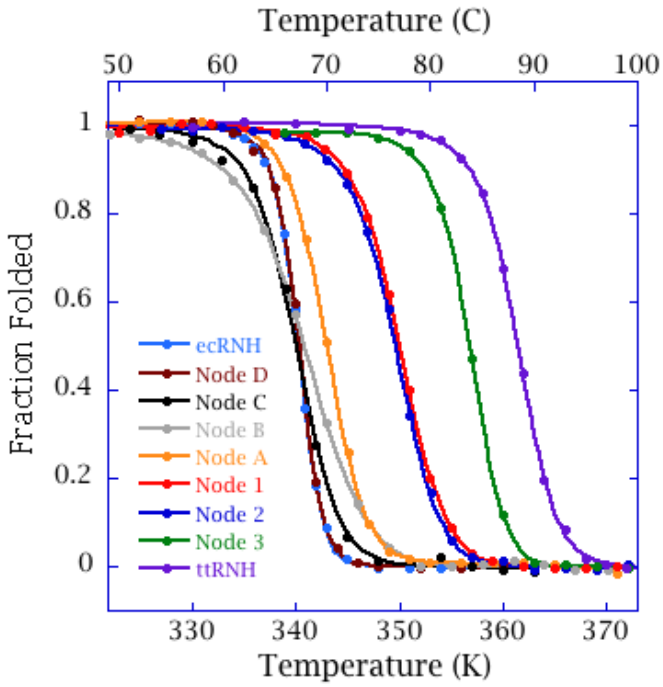


Figure 11. Thermal denaturation of ecRNH, ttRNH and all nodes as probed by CD signal at 222 nm.

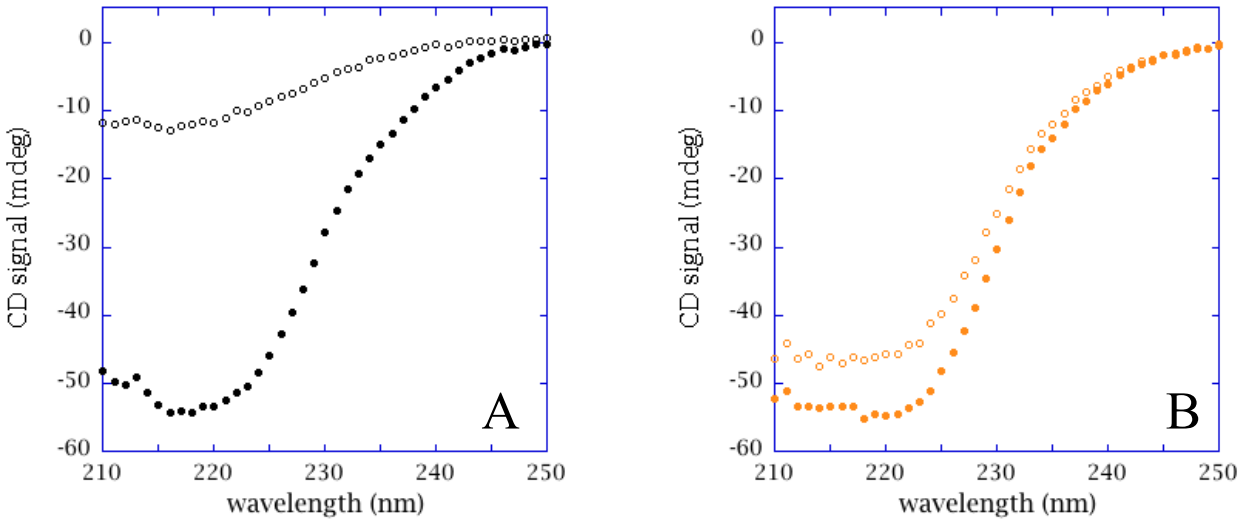


Figure 12. CD spectra at 25 °C before (solid circles) and after (open circles) thermal denaturation. (A) Node C is irreversible. (B) Node A is reversible, because greater than 80% of the signal at 222 nm is retained post-thermal denaturation.

Table 4. Measured T_ms and apparent T_ms

	T_m (°C)	Reversibility
ttRNH	88	reversible
Node 3	83	reversible
Node 2	77	reversible
Node 1	77	reversible
Node A	70	reversible
Node B	68	irreversible
Node C	67	irreversible
Node D	68	irreversible
ecRNH	68	irreversible

The shared ancestor, Node 1, refolds upon cooling post-thermal denaturation. Reversible thermal unfolding is lost along the mesophilic branch with Node B but maintained throughout the thermophilic lineage. Trends in melting temperatures are apparent when T_ms are plotted a function of branch length (**Figure 13**). Starting with Node 1 and moving along the mesophilic lineage, the T_m initially drops with Node A, but then it holds at a relatively constant value that is also shared with the modern-day protein, ecRNH. Along the thermophilic branch, the T_m initially holds steady with Node 2, but then proceeds to increase with each successive ancestor, culminating with ttRNH, which has the highest thermostability.

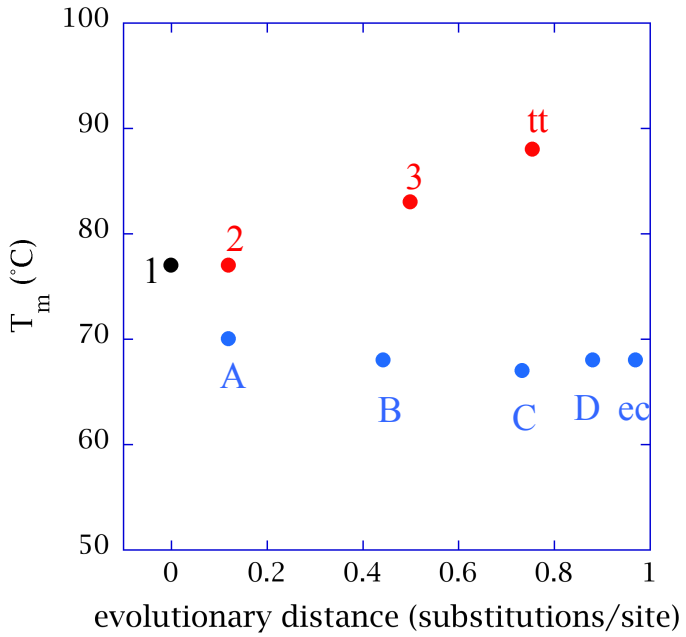


Figure 13. Melting temperature as a function of evolutionary distance from the last common ancestor, Node 1. Starting with Node 1 (black), melting temperature increases along the thermophilic lineage (red) and stays constant along the mesophilic lineage (blue). Distance is calculated by summing the branch lengths between the ancestor of interest and Node 1 on the maximum likelihood tree.

Thermostability is of particular interest for understanding the evolution of biophysical traits, because protein melting temperatures have been shown to correlate with organismal growth temperature [7]. Gromiha *et al*. observe the following relationship, with a correlation coefficient of 0.91, in a set of 56 globular proteins: $T_m = 24.4 + 0.93\ T_{env}$, where T_{env} is the organism's optimal growth temperature [7]. This correlation suggests that T_m adjusts to accommodate environmental temperatures, perhaps in response to selective pressure on foldedness. It is somewhat puzzling, however, that such a relationship is observed given that melting temperatures for proteins from a single organism can vary widely. In fact, it has been suggested that only a small subset of an organism's proteome is responsible for temperature sensitivity [32, 33].

To determine if a trend is observed in extant RNases H, T_ms were measured for proteins from 6 additional species and compared with ecRNH, ttRNH and two previously studied RNases H [34, 35] (**Table 5**). Growth temperatures were culled from the literature [36, 37]. Neither optimal growth nor environmental temperatures could be found for two mesophiles, *Enterobacter sp. 638* and *Candidatus Hamiltonella defensa*, so laboratory culturing temperatures were used. The former, a plant endophyte, was successfully cultured at 30 °C, and the latter, an aphid symbiont was cultured at 25.6 °C [38, 39].

Table 5. T_ms and T_{env}s for extant RNases H

	T_{env} (°C)	T_m (°C)
Thermus thermophilus	68	88
Chlorobium tepidum	48*	67*
Escherichia coli	37	68
Klebsiella pneumoniae	37	68
Citrobacter sp. 30_2	37	62
Cronobacter turicensis	37	64
Salmonella enterica	37	59
Enterobacter sp. 638	30	63
Shewanella oneida	30†	53†
C. Hamiltonella defensa	27	53

* Taken from reference [35].
† Taken from reference [34].

Despite potential discrepancies in exact values of T_{env}, a clear trend is observed between the T_ms of RNases H and their organismal growth temperatures (**Figure 14A**). The trend is in close agreement with the one observed by Gromiha *et al*. Assuming this correlation holds for the ancestral RNases H, then growth temperatures for the extinct species can be predicted based on measured T_ms (**Figure 14B**). Trends observed in T_ms predict trends in organismal growth temperatures, which are consistent with a slow, adaptive process toward thermophilicity.

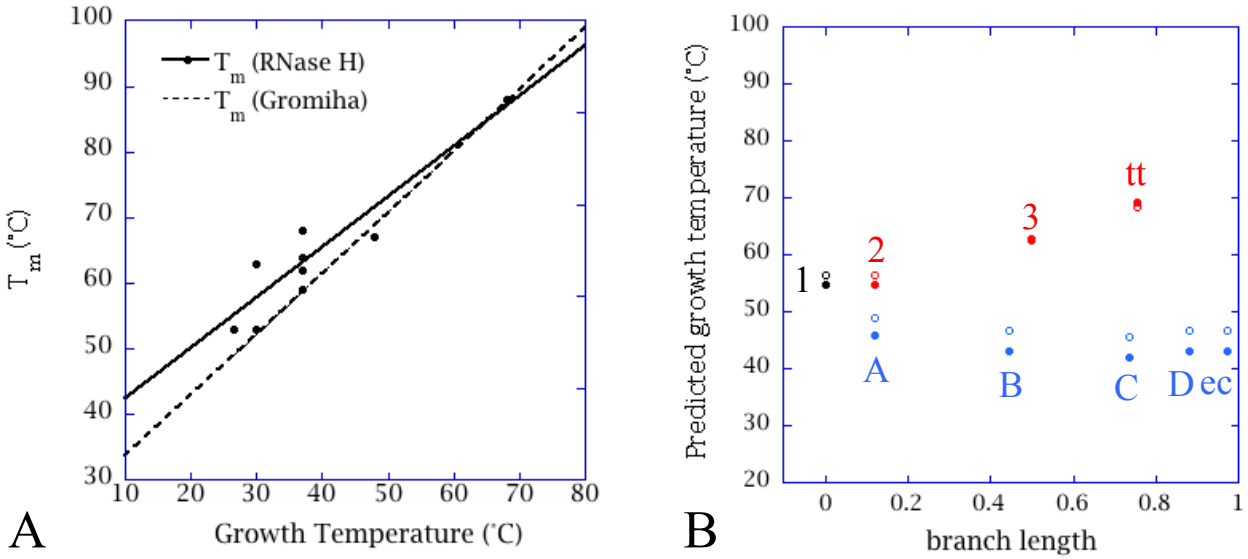


Figure 14. Correlation between growth temperature and T_m (A) Linear regression from Gromiha *et al.* (dotted line) and linear regression for extant RNases H (solid line, circles) are similar. (B) Predicted growth temperatures for node organisms based on RNase H linear regression (open circles) or linear regression from Gromiha *et al*. (closed circles).

2.3.4.2 Chemical denaturation

Global stabilities for ecRNH, ttRNH and all the nodes were measured using GdmCl-induced denaturation and monitored by the change in CD signal at 222 nm (**Figure 15**). Individual melts were fit to a two-state model with the assumption that ΔG varies linearly with GdmCl concentration [40]:

$$\Delta G_u = \Delta G_{H2O} - m\,[\,\text{denaturant}\,] \qquad (2)$$

where ΔG_{H2O} is the extrapolated stability in water and the *m*-value is the slope of ΔG_u 's dependence on denaturant. The *m-value* has been shown empirically to correlate with changes in solvent exposed surface area upon unfolding [41]. ΔG and *m*-values at 25 °C are collected in **Table 6**.

Notably, Node B has the lowest ΔG and *m*-value. CD and activity data suggest Node B adopts a similar folded structure to ecRNH, ttRNH and Node C, which means that its *m*-value should fall within the range established by these proteins. One plausible explanation could be residual structure in its unfolded state; however, it would have to be more extensive than that observed in ttRNH. Alternatively, partially folded intermediates could be populated just enough to effect stability and *m*-values without causing obvious deviation from two-state behavior in the melts. In fact, it has been demonstrated that an anomalously low *m*-value observed in a destabilized

variant of ecRNH is due to an equilibrium intermediate [42]. In this particular variant, alanine substitutions at R46, D102 and D148 lead to the elimination of a partially buried salt-bridge network. While Node B retains these specific residues, it does lack another residue involved in a nearby salt bridge at position 57. Position 57 is expected to be involved in salt bridges in all of the other proteins studied here. It is either lysine or arginine in ttRNH, nodes 3, 2, 1 and A. In the ttRNH crystal structure, it interacts with E54. ecRNH and nodes C and D have a glutamic acid at position 57, and it interacts with R106. Node B has an asparagine at position 57, which is likely to destabilize the conserved salt-bridge network. Therefore, we interpret the low *m-value* observed in Node B as reflecting the presence of equilibrium intermediates. Because it violates the two-state assumption, its stability cannot be measured by these methods. Node B cannot be directly compared with the other proteins, and it is not included in further analyses.

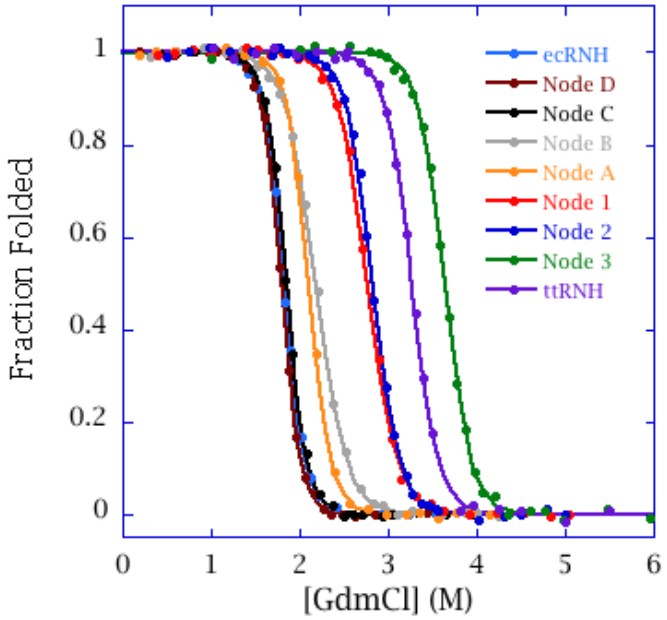


Figure 15. GdmCl-induced denaturation melts of all ancestors, ecRNH and ttRNH at 25 °C as monitored by CD at 222 nm.

Table 6. *ΔG*s and *m*-values at 25 °C

	ΔG * **(kcal mol^{-1})**	***m*-value** * **(kcal mol^{-1} M^{-1})**
ttRNH	12.8	3.93
Node 3	13.3 ± 1.2	3.84 ± 0.12
Node 2	10.6 ± 0.3	3.80 ± 0.10
Node 1	10.9 ± 1.1	4.23 ± 0.09
Node A	9.9 ± 0.2	4.69 ± 0.08
Node B	*7.0 ± 0.4*†	*3.17 ± 0.24*†
Node C	9.2 ± 0.3	4.99 ± 0.15
Node D	9.4 ± 0.1	5.25 ± 0.07
ecRNH	9.0 ± 0.3	4.99 ± 0.17

* Errors reported for replicated experiments.

† Values reflect a two-state assumption, which may not be valid for Node B. See section 2.3.4.2.

Plotting ΔG and *m*-values as a function of branch length reveals trends in both parameters at 25 °C (**Figure 16**). Stabilities within the mesophilic lineage are similar to one another and lower than stabilities in the thermophilic lineage. The largest increase in stability occurs between nodes 2 and 3 and is maintained in ttRNH. Similar trends are observed at all tested temperatures (data not shown), and the relationships between these trends and trends in T_m are explored in depth in section 2.3.4.3. Interestingly, trends are also observed in the *m*-values at all temperatures. Because *m*-values correlate with changes in solvent exposed surface area upon unfolding [41], it is possible the lower *m*-values along the thermophilic lineage are reporting on residual structure in their unfolded states. Indeed, *m*-value effects in the protein SNase can be attributed to changes in the unfolded state [43]. If this were true for RNase H, we would expect to see a similar trend in ΔC_p, which also relates to changes in solvent exposed surface area; however, no such trends are observed (see section 2.3.4.3). Discrepancies between ΔC_p an *m*-values are not unprecedented [11, 12, 44]. A single amino acid substitution, I53D, in the cysteine-free version ttRNH causes a dramatic increase in ΔC_p but no change in *m*-value. The interpretation for these effects is that replacing a buried hydrophobic residue with a polar sidechain disrupts residual structure in the unfolded state, leading to an increase in ΔC_p. Robic *et al.* suggest the lack of a parallel change in *m*-values is not problematic, because *m*-values are measured in high denaturant where the unfolded state can differ from that of the unfolded state under native conditios. DSC, which was used to identify residual structure in ttRNH [13], was not performed on I53D ttRNH; however, it seems the most likely model for the increase in ΔC_p. Thus, while it is tempting to interpret divergent *m*-values between the two lineages as divergence in residual structure in the unfolded state, it is unlikely this is the case. Future DSC studies, particularly of nodes A, 1, 2, and 3, which unfold reversibly, are needed to determine if the *m*-values are related to residual structure in their unfold states.

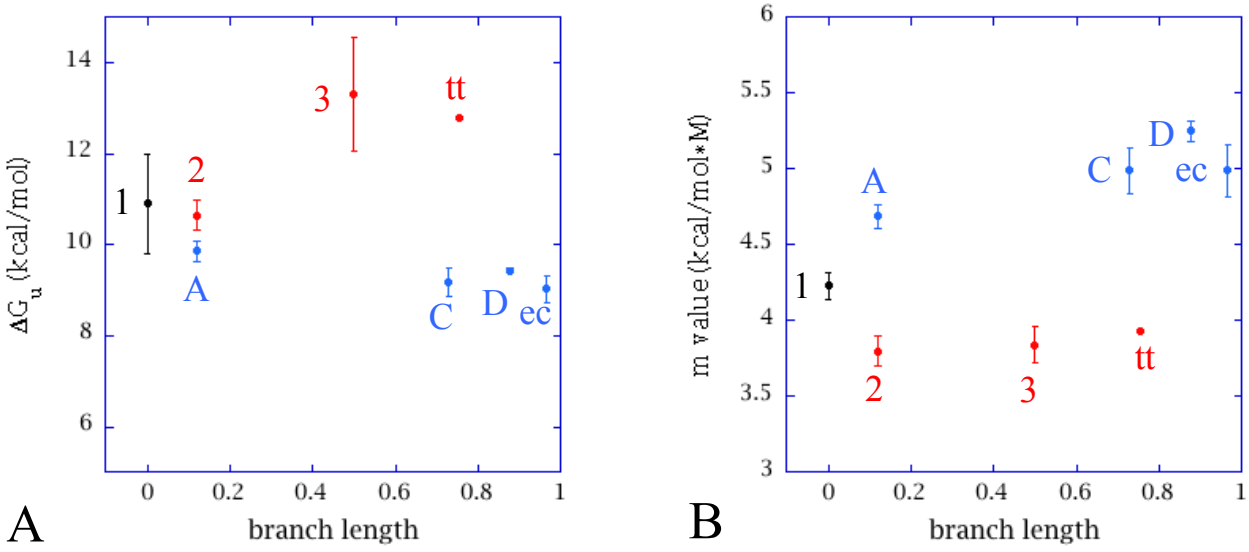


Figure 16. Average (A) ΔGs and (B) *m*-values at 25 °C as a function of branch length. Errors are standard deviations from replicated experiments. Stabilities are higher in the thermophilic lineage (red) than in the mesophilic lineage (blue), but *m*-values are lower.

Extant RNases H share a similar global stability at their respective environmental growth temperatures, suggesting that this value may have functional relevance (**Figure 17A**). Stabilities for two additional mesophilic RNases H, from *Citrobacter sp.30_2* (cbRNH) and *Klebsiella pneumoniae* (kpRNH), were measured by GdmCl-induced denaturation and fit to a two-state model (data not shown). At their growth temperature of 37 °C, cbRNH has a ΔG of 6.2 kcal/mol and kpRNH has a ΔG of 7.8 kcal/mol. None of the other extant RNases H for which T_ms were measured in Figure 13A was suitable for analysis. Other stabilities included in Figure 16 were curated from the literature [34, 35]. The average ΔG for RNases H at their respective organismal growth temperatures is 6.28 ± 1.48 kcal/mol. If the trend holds for the ancestral RNases H, then their growth temperatures can be predicted based on the temperatures at which they achieve this global stability (**Figure 17B**). The predicted growth temperatures are consistent with those predicted based on T_ms in Figure 13B.

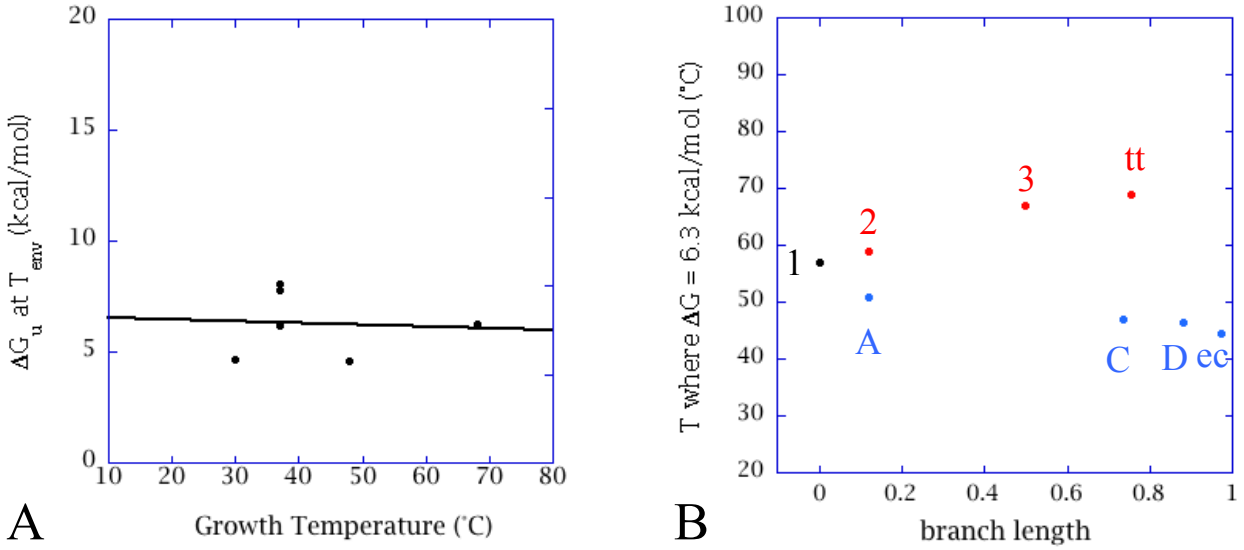


Figure 17. Conserved ΔG at organismal growth temperature. (A) Extant RNases H have an average ΔG of 6.28 ± 1.48 kcal/mol at their respective growth temperatures (circles, solid line). (B) Predicted growth temperatures for node organisms is based on temperature at which the node has a ΔG = 6.3 kcal/mol.

2.3.4.3 Stability curve determination

To determine the mechanism of thermostabilization observed in the ancestors, stability curves were generated for all of them. Stabilities were determined as described above at multiple temperatures. Representative data are shown in **Figure 18**. Protein samples were equilibrated overnight or longer for temperatures below 25 °C. For higher temperatures, proteins were equilibrated overnight only if solubility was not compromised. Otherwise, shorter equilibration times were used, typically 2-3 hours for melts performed manually and 10 minutes for melts performed using an automated titrator. Data from titration experiments were only used if the measured C_m was within 0.1 M of the value collected manually at the same temperature.

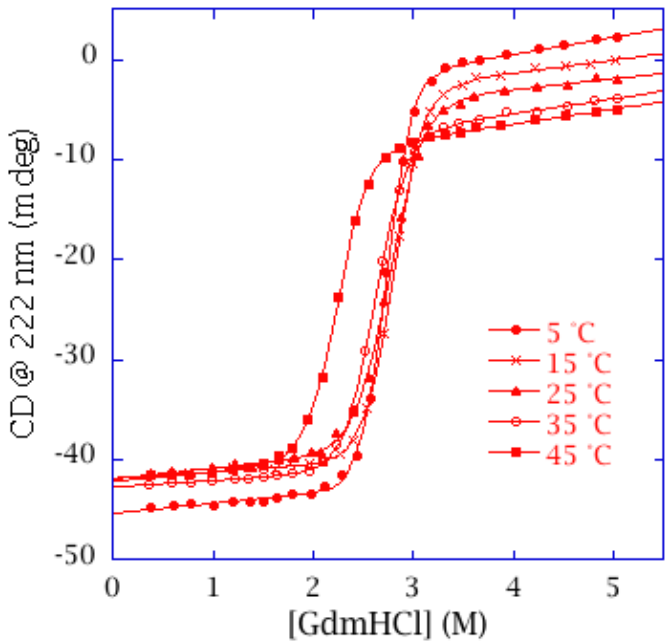
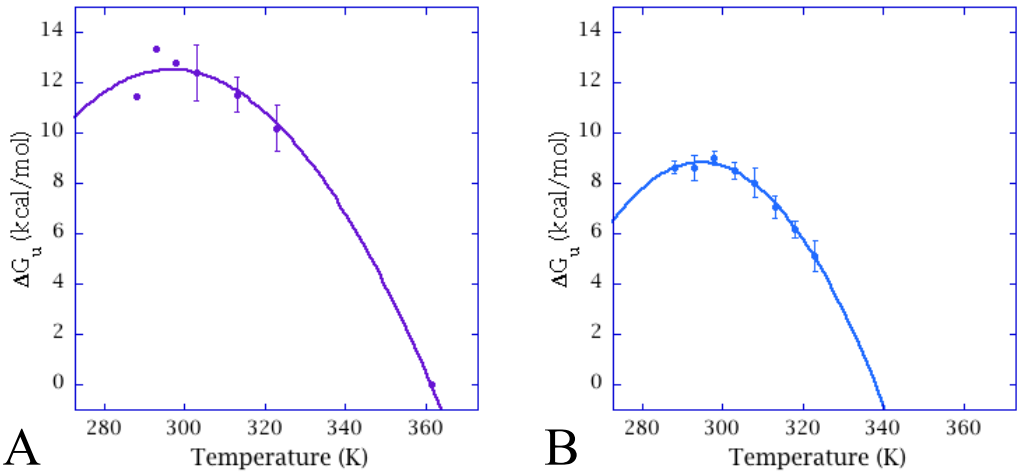


Figure 18. Representative GdmCl-induced denaturant melts for Node 1 at multiple temperatures

For each protein, the ΔG is plotted versus temperature to generate a stability curve (**Figure 19**). Most denaturant melts were performed in duplicate or triplicate, in which case an average ΔG is used for the fits, and standard deviations are reported. Some melts were performed only once, so no error is reported. Melts for ecRNH, particularly at low temperatures, were repeated more than three times and are discussed at length below in section 2.3.4.4. The T_m determined from thermal denaturation is included as a single point at $\Delta G = 0$ for proteins demonstrating reversible thermal unfolding. The criterion for reversibility is defined as 80% recovery of signal at 222 nm when cooled to 25 °C post-thermal denaturation. Nodes A, 1, 2, 3 and ttRNH all demonstrate reversible thermal unfolding. Stability curves were fit to the Gibbs-Helmholtz equation (**Equation 1**) in order to extract the thermodynamic parameters ΔH_m, ΔC_p and T_m. In general, the T_ms from the fits are in close agreement with measured values, even for proteins that do not unfold reversibly with temperature. The one notable exception is ecRNH, which has a measured T_m of 68 °C, but the fit T_m is 65 °C.



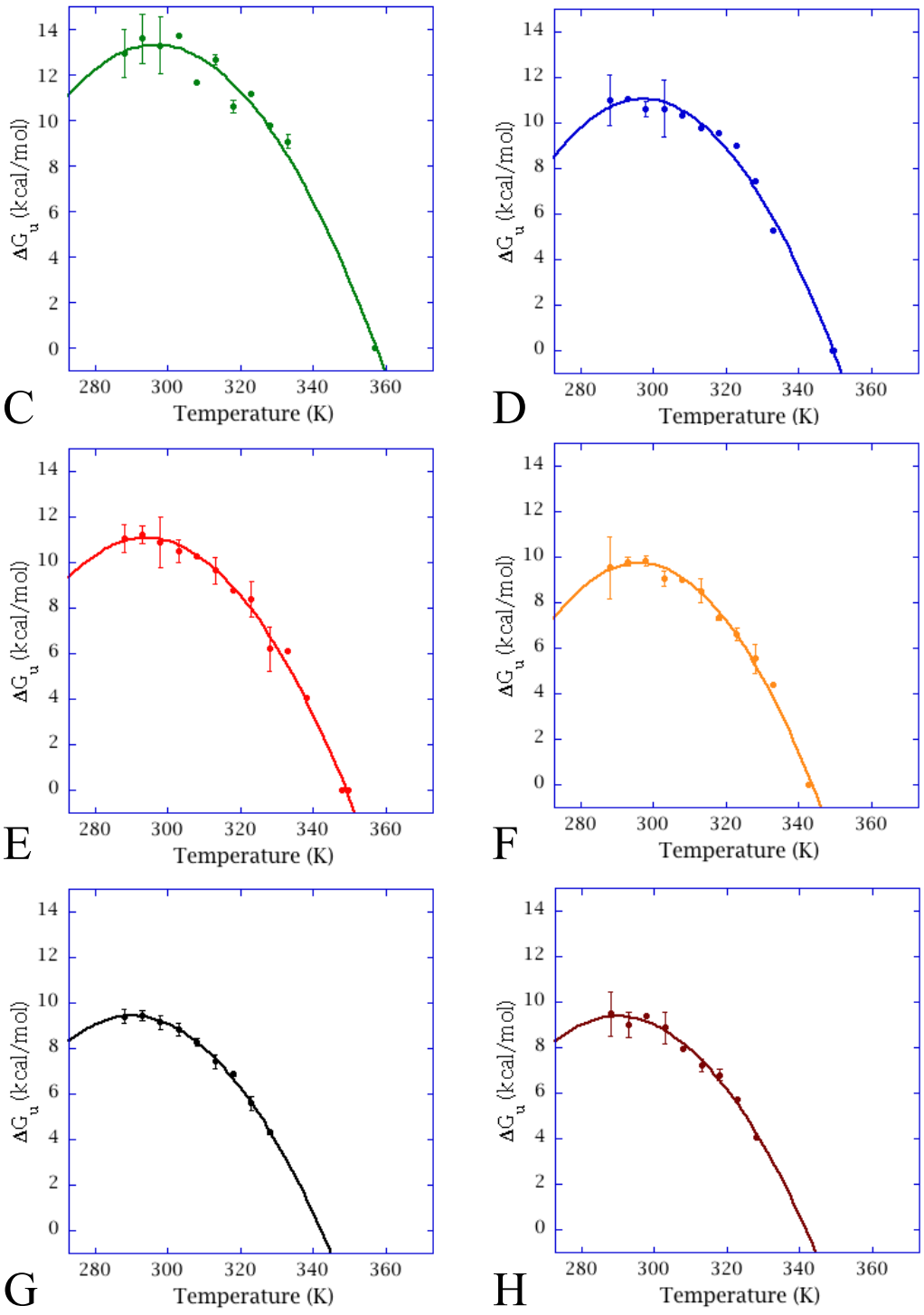


Figure 19. Stability curves for (A) ttRNH (B) ecRNH (C) Node 3 (D) Node 2 (E) Node 1 (F) Node A (G) Node C (H) Node. Average ΔG is used for the fits, and errors are standard deviations from replicate experiments. Melts performed do not have error bars.

The data were also fit using an alternative version of the Gibbs-Helmholtz equation (**Equation 3**) in order to extract additional parameters relevant for classifying the various thermodynamic mechanisms. This version of Gibbs-Helmholtz is derived using reference temperatures T_s and T_h, or the temperatures at which ΔS and ΔH are zero, respectively.

$$\Delta G\ (T) = \Delta C_p\ (T - T_h) - T\ \Delta C_p\ ln\left(\frac{T}{T_s}\right) \quad (3)$$

The T_s represents the temperature of maximum stability, and because $\Delta S = 0$, folding is entirely driven by enthalpy at this temperature. Enthalpies at T_s, (ΔH_s) are calculated using the following relationship:

$$\Delta H_s = \Delta C_p\ (T_s - T_h) \quad (4)$$

For nearly all thermodynamic parameters, including T_m, values for the shared ancestor, Node 1, fall in between those of ecRNH and ttRNH (**Table 7**). The notable exception is T_s, which is lowest for Node 1 at 21 °C, slightly higher for ecRNH at 22 °C and highest for ttRNH at 24 °C. How the stability curves change along the mesophilic and thermophilic lineages can be understood in terms of the thermodynamic strategies outlined above in section 2.2.

Table 7. Thermodynamic parameters from stability curve fits

	ΔC_p (kcal mol^{-1} K^{-1})	$T_{m,\,fit}$ / $T_{m,\,meas}$ (K)	ΔH_m (kcal mol^{-1})	T_s (K)	ΔH_s (kcal mol^{-1})
ttRNH	1.91 ± 0.30*	361‡ / 361†	136	297 ± 5*	12.5
Node 3	2.29 ± 0.30*	357‡ / 356†	152	297 ± 3*	13.3
Node 2	2.53 ± 0.20*	350‡ / 350†	144	297 ± 2*	11.1
Node 1	2.28 ± 0.19*	349‡ / 350†	137	294 ± 2*	11.1
Node A	2.67 ± 0.24*	344‡ / 343†	137	296 ± 2*	9.7
Node C	2.14 ± 0.15*	342‡ / 340†	121	290 ± 1*	9.4
Node D	2.14 ± 0.40*	342‡ / 341†	121	290 ± 3*	9.4
ecRNH	2.89 ± 0.31*	338‡ / 341†	135	295 ± 1*	8.8

* Errors from stability curve fit.
† Extracted from thermal melt fit.
‡ Extracted from stability curve fit.

Beginning with Node 1 and moving along the thermophilic lineage, T_m increases (**Figure 11**). Analyzing the stability curves for nodes 1, 2, 3 and ttRNH reveals that the observed trend is not driven by a single thermodynamic strategy. T_m increases by 1 °C from Node 1 to Node 2, and this change is primarily due to an right-shifting of the stability curve (method III) as indicated by an increase in T_s from 21 °C to 24 °C. Such a dramatic shift in T_s could have resulted in an even larger increase in T_m if changes had not also occurred in the ΔC_p, which increase from 2.28 kcal mol^{-1} K^{-1} in Node 1 to 2.53 kcal mol^{-1} K^{-1} in Node 2. In fact, the measured T_ms do not change at all, indicating the shift in T_s is largely compensated by broadening of the curve. The ΔH_s is similar between the two ancestors, reflecting the fact that an upshift of the curve is not

responsible for the modest increase in T_m. An increase in T_s indicates a smaller ΔS, which can be accomplished either by reducing entropy of the unfolded state or increasing entropy of the folded state. Thus, it is interesting to note that Node 2 contains ten proline residues, while Node 1 contains only 7. In certain structural contexts, such as loops, prolines can restrict conformational entropy more in the unfolded state than in the folded state leading to a reduction in overall ΔS. The additional prolines in Node 2, however, appear at the N- and C-termini, regions that are not well defined in the crystal structures, so it is not clear if or how they are contributing to reduced ΔS.

Continuing along the thermophilic lineage, T_m increases by 7 °C from Node 2 to Node 3. This is the largest increase in T_m observed between adjacent nodes and is primarily the result of stability curve broadening and upshifting (methods II and I). The ΔC_p of Node 3 is 2.29 kcal mol^{-1} K^{-1}, which is 0.24 kcal mol^{-1} K^{-1} lower than the ΔC_p of Node 3 and results in a more broad curve. Interestingly, the ΔC_p of Node 3 is quite similar to that of Node 1. In this case, however, the lower ΔC_p leads to an increased T_m, because Node 3 retains Node 2's high T_s. It is difficult to attribute the high T_s values observed in nodes 2 and 3, and also in ttRNH, to specific sequence differences. All of these proteins contain more prolines, 10, 9 and 12 respectively, than Node 1; however similar T_s values along the mesophilic branch are observed in Node A and ecRNH, which each contain only five.

The second major contributing factor to Node 3's increased T_m is its upshifted stability curve, which is reflected in the increase of ΔH_s from 11.1 to 13.3 kcal/mol. The energetic consequences of electrostatic interactions on a protein's surface are highly context-dependent, but there are a couple notable sequence changes from Node 2 to Node 3 that may contribute to stabilization. First, a salt bridge between K50 in helix A and E39 in strand 3 is present in ttRNH and may form in Node 3, since it also contains these residues. Helices and strands and named according to convention and are labeled along with other notable features in **Figure 20**. Node 2 could not have this interaction, because it has a threonine at position 50. Second, the substrate recognition loop contains a high density of basic residues, presumably to aid in binding RNA-DNA hybrids, but there are certainly energetic costs for concentrating positive charges in close proximity. In this stretch of residues, extending from position 80 to 100, ttRNH has 6 basic residues but also 2 acidic residues, which may help relieve any local strain. Node 3 is even more charge-balanced with 4 basic and 2 acidic residues. All of the other ancestors have up to 8 basic residues in this region but only one acidic residue, which is likely to be destabilizing.

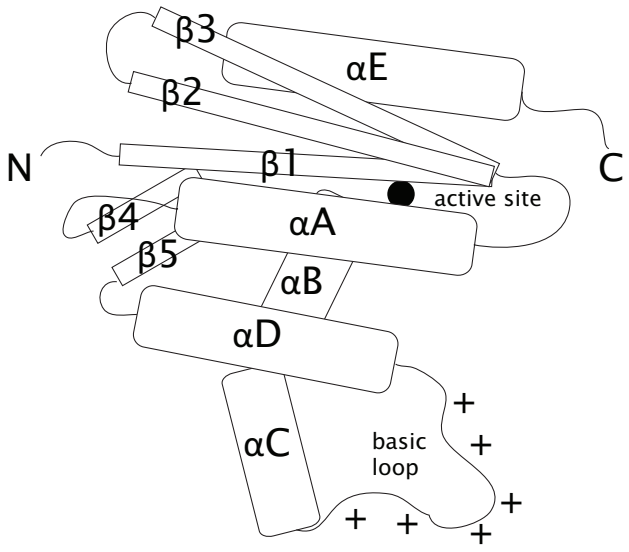


Figure 20. Schematic of RNase H topology labeled with elements of secondary structure, the substrate-recognition basic loop and active site Mg^{2+}.

The most recent development of thermostability in ttRNH occurred exclusively via stability curve broadening (method II). The ΔC_p of Node 3 is 2.29 kcal mol^{-1} K^{-1}, but ttRNH has a ΔC_p value of 1.91 kcal mol^{-1} K^{-1}. There is no change in T_s, and the ΔH_s actually goes down slightly from 13.3 to 12.5 kcal/mol. Previous studies of a cysteine-free variant of ttRNH identified residual structure in the unfolded state [13], which causes a reduction in ΔC_p and is expected to be slightly destabilizing. Thus, these data are consistent with appearance of residual structure in ttRNH but not its immediate ancestor; however, additional studies are needed to confirm that low ΔC_p in ttRNH is a result of residual structure in the unfolded state and to determine if any such structure is retained in other ancestors. While it is not known which residues are responsible for residual structure, sequence differences might help identify other distinguishing features such as electrostatic interactions in the native state. For instance, ttRNH contains more charged residues, particularly arginines and lysines. In fact, assuming the pK_as of the sidechains match amino acids in solution, the formal charge for ttRNH at pH 5.5 is 15.5, which is much higher than that of the other ancestors (**Figure 21**). These additional charged residues create three new salt bridge interactions that are not accessible to Node 3. Salt bridges are defined here as forming between any two oppositely charged residues within 4 Å of each other. The first salt bridge is between H29, located in the loop between strands 2 and 3, and E61, located in the loop between helix A and strand 4. It might also be form in Node D, which contains K29, and ecRNH, which contains R29, but not the remaining ancestors, which contain glycines at position 29. The second novel salt bridge is between H72 and D10. D10 is one of three active site residues that chelate the Mg^{2+} ion necessary for catalysis. Both the structures and stabilities were measured in the absence of Mg^{2+}, so the three acidic sidechains are clustered without a counterion. Removing D10 by replacing it with an alanine results in global stabilization by 3 kcal/mol for the cysteine-free ecRNH [45], which demonstrates that this cluster is destabilizing. H72 may interact with D10 to relieve destabilization in a similar manner. A third novel salt bridge in ttRNH forms between E64 in strand 4 and R115 in strand 5. Lastly, R101 and E105 are appropriately spaced along helix D, but are not close enough in the crystal structure to form a salt bridge. It is possible, however, that this interaction does occur in solution. Recent studies of ribosomal protein L30e show that 2 salt bridges on its surface not only contribute to increased stability, but also reduce ΔC_p by 0.2 kcal mol^{-1} K^{-1} [46]. Thus, it is possible that the lower ΔC_p observed in ttRNH relative

to Node 3 is due to more charge interactions in the folded state and does not reflect additional residual structure in the unfolded state.

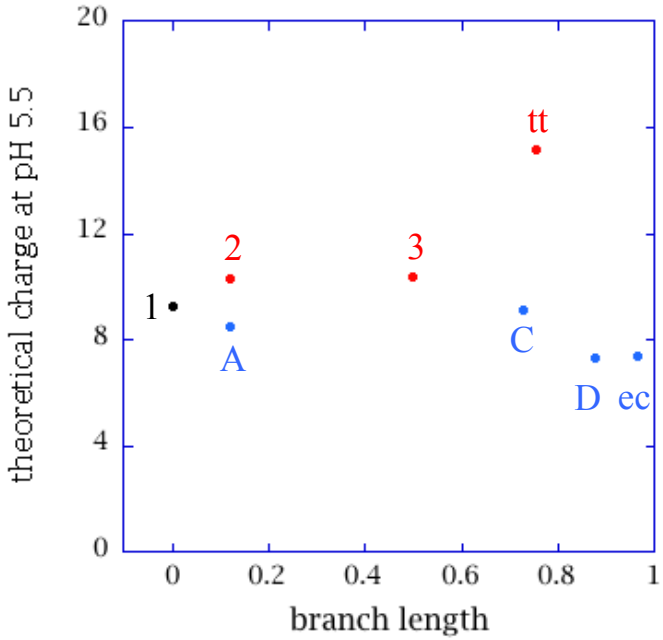


Figure 21. Calculated formal charge on ancestors, ecRNH and ttRNH at pH 5.5 using pK_as from individual amino acid sidechains.

Beginning with Node 1 and proceeding along the mesophilic lineage, T_m drops from 76 °C to 71 °C and then levels off (**Figure 13**). The 5 °C decrease in T_m from Node 1 to Node A occurs through a narrowing (method II) and downshifting (method I) of its stability curve. There is also a small increase in T_s, which might be due to the introduction of a glycine between helices C and D. This inserted glycine is conserved in thermophilic RNases H and has been shown to be increase dynamics in the native state [47]. Increased entropy in the folded state is expected to reduce ΔS and increase in T_s, but the inserted glycine alone cannot be the only determinant as it is also present in nodes C and D, which have significantly lower T_ss.

After the initial drop in thermostability from Node 1 to Node A, which has a T_m of 71 °C, melting temperatures remain fairly constant along mesophilic lineage. Detailed analysis of their stability curves, however, demonstrates that even when maintaining a given T_m, RNases H employ different thermodynamic strategies. For instance, the stability curves of both nodes C and D are left-shifted (method III) relative to Node A, reflecting a decrease in T_s of 6 degrees. This effect is countered by a large decrease in the ΔC_p of 0.53 kcal mol^{-1} K^{-1} (method II), which results in an overall modest 1-3 degree decrease in T_m. ecRNH shows an increase in T_s, restoring it to a value more similar to those the other ancestors (method III), but a dramatic increase in ΔC_p by 0.75 kcal mol^{-1} K^{-1} (method II). Again, the effects cancel one another, leading to an overall 0-4 degree drop in T_m. Remarkably, these thermodynamic changes are encoded by very small changes in sequence. Part of the reason is that thermodynamic parameters are inextricably linked to each other, resulting, for instance, in compensations between entropy and enthalpy. Thus, sequence changes have inconsistent effects in different backgrounds. Node D and ecRNH differ by only 11 residues, but each substitution may need to be interrogated in isolation in order to decipher how each parameter is encoded.

In a previous study comparing ecRNH and ttRNH, methods I and II were identified as the relevant mechanisms for evolving thermophilicity in RNase H [8]. Much attention has been paid to ΔC_p, in particular, due to its demonstrated relationship with residual structure in the unfolded state [13]. While we do observe trends in T_m along the evolutionary lineages, no trend is observed in ΔC_p (**Figure 21A**). Instead, comparing stability curves between ancestral proteins and their modern-day descendants reveals that alternate thermodynamic strategies were used to tune thermostability in ancient RNases H (**Figure 21B**).

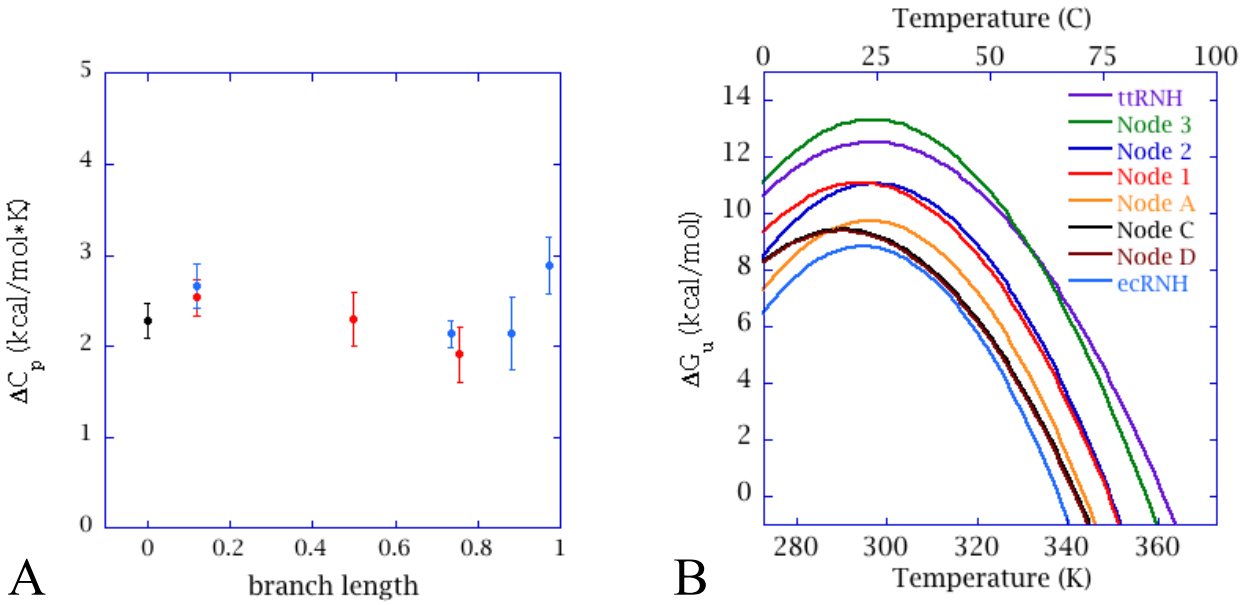


Figure 22. (A) ΔC_p as function of branch length. There is no clear trend in this parameter (B) Superimposed stability curve fits.

2.3.4.4 Deviation from two-state behavior at low temperatures

Only data collected at temperatures above 15 °C are used in the stability curve fits due to unexpected behavior observed at low temperatures. Below 15 °C, measured stabilities plateau such that the resulting stability curve appears asymmetric. The effect is the most pronounced for ecRNH (**Figure 22**). Because curvature is described by a single parameter, ΔC_p, asymmetry in the curve is an indication that one of the assumptions made in deriving the stability curve is incorrect.

The first possibility is that ΔC_p is not constant. In fact, ΔC_p is unlikely to be constant over all temperatures, because the heat capacities of both the folded and unfolded states must, by definition, approach zero at absolute zero. Thus, it is unlikely a constant difference in heat capacity is maintained over all temperatures [9]. It has been found, however, that assuming ΔC_p is constant over the measured temperature range is reasonable, especially given the experimental error in the parameter [9, 48, 49]. Values of ΔC_p can be determined to within 5-10% error, which is comparable to the maximum amount of temperature-dependent variation expected [9]. Therefore, changes in ΔC_p are expected to be too small to measure accurately. Nicholson and Scholtz demonstrated this experimentally with histidine-containing phosphocarrier protein by performing temperature melts in various concentrations of denaturant [50]. They found a linear dependence of enthalpy of unfolding on melting temperature, indicating that ΔC_p is constant over

a 60 °C range. Since then, this method for measuring ΔC_p has been employed for a diversity of proteins [8, 51-53].

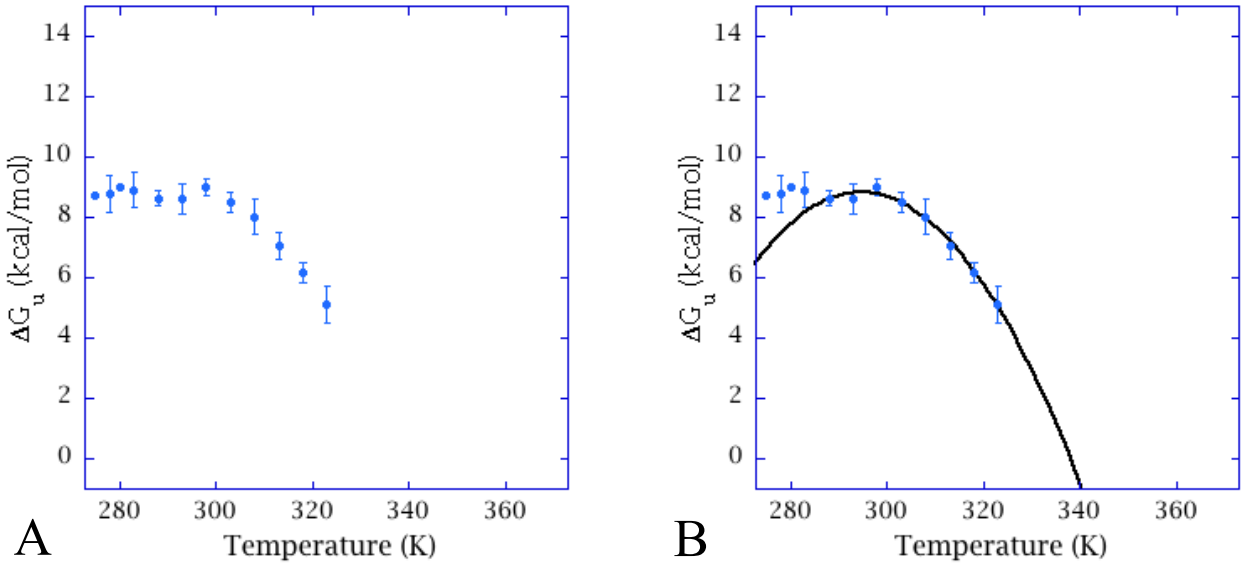


Figure 22. Asymmetry in ecRNH stability data. (A) ΔG values unexpectedly plateau below 15 °C. Data at 5 °C and 10 °C reflect the averages of 12 and 7 independent experiments, respectively. (B) Average ΔG values superimposed with fit from Figure 22B to highlight deviations at low temperatures.

The second potential erroneous assumption is that the measured ΔC_p reflects only the folding reaction with no contribution from denaturant molecules binding to the protein. While we use a simple linear extrapolation to determine ΔG from denaturant melts, binding models where denaturant molecules interact with specific sites on the protein represent an alternative method. In practice, however, binding models are less accurate due to ambiguities in binding constants [41]. Binding of denaturant molecules to the protein would mean ΔC_p reflects not only the folding reaction but the binding reaction, which is likely to have a temperature-dependence [54].

Another fundamental assumption underlying the analysis is that the protein folds in a two-state manner. All of the melts are fit assuming the signal at each denaturant concentration represents a weighted sum of two states. The presence of additional states would invalidate the extracted ΔG values and, by extension, the stability curves. Deviations at low temperature might reflect the temperature-dependence of populating these additional states.

Curiously, the asymmetric behavior is not observed in a cysteine-free version of *E. coli* RNase H (**Figure 24**), denoted with an asterisk as ecRNH*, which is consistent with previous studies [8]. ecRNH* differs from ecRNH at 3 positions: C13A, C63A and C133A. To identify the underlying cause for the differential behavior, all possible single and double cysteine variants were constructed and studied. Two of the double cysteine variants, C13A/C63A and C133A/C63A, have symmetric curves like ecRNH* (**Figure 24B**). The effect is also significantly diminished in the single cysteine variant C63A (**Figure 24A**), leading to the conclusion that C63 is the residue responsible for asymmetry in the stability curve of ecRNH.

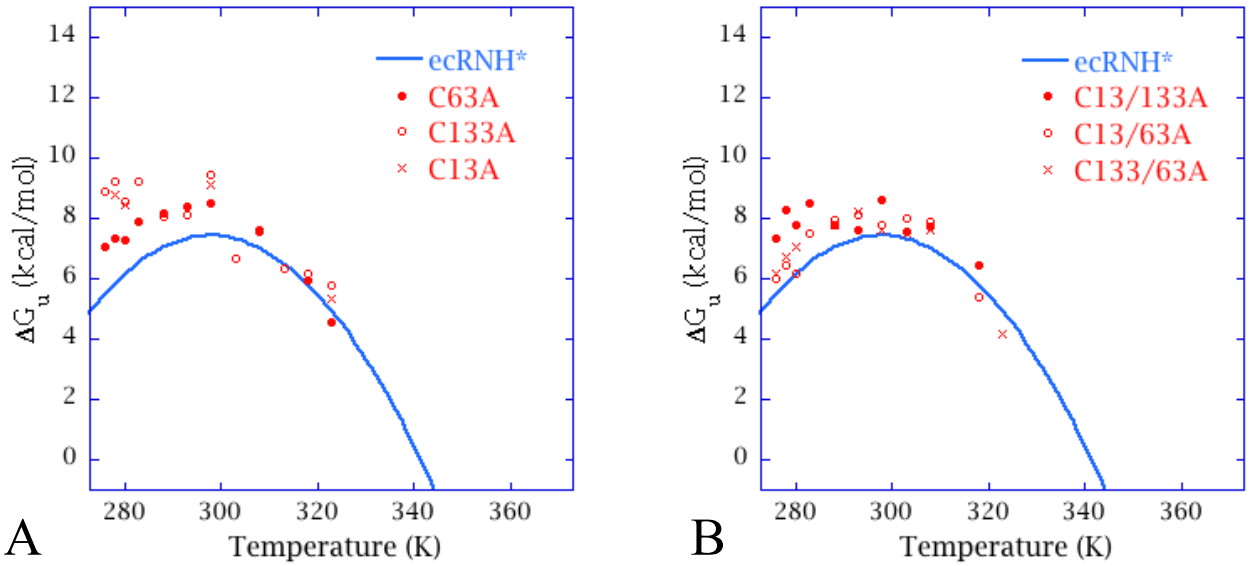


Figure 24. Stability curve fit from cysteine-free ecRNH* superimposed with data from (A) single cysteine variants and (B) double cysteine variants. Variants with cysteine at position 63 show the unexpected asymmetry in their stabilities.

Because changing a single amino acid drastically changes the shape of the curve, it is unlikely that the assumptions about ΔC_p are the problem. If, for instance, ΔC_p were not constant over the measured temperature range or if ΔC_p reflected denaturant binding, then the effects would be observed in both ecRNH and ecRNH*. Also, if denaturant binding were a more appropriate model, then urea-induced denaturation might be expected to give different results. While a full stability curve was not generated in urea for ecRNH, urea melts at 5 °C and 25 °C measure similar stabilities (data not shown), which is analogous to the behavior observed by Gdml-induced denaturation. What is more likely is that the single cysteine causes deviations in two-state behavior. The most obvious mechanism would be dimerization through the formation of a disulfide bond; however, all experiments were performed using 1 mM TCEP, and neither native gel analysis nor size exclusion chromatography gives any indication of dimer formation at low temperatures (data not shown). It is possible that dimer formation does not depend upon cysteine oxidation, in which case it might be difficult to detect by these techniques. Interaction between molecules, however, is likely to effect stability in a concentration-dependent manner. Several denaturant melt experiments were performed to test if stability is sensitive to protein concentration, the presence of TCEP and/or alternative structural probes. The mid-point for denaturation, or C_m, represents the most well determined parameter from the fits, so this the best value to compare between experiments (**Table 8**). In order to restore symmetry to ecRNH's stability curve, the C_m at 5 °C would need to decrease to approximately 1.4 - 1.5 M, assuming *m*-values remain constant.

Table 8. Effects of protein concentration and TCEP on the C_m of ecRNH

Probe	[ecRNH] (μM)	Temperature (°C)	[TCEP] (mM)	C_m (M)
Fluorescence	2.8	5	1	1.77
CD	2.8	5	1	1.79 ± 0.01*
Fluorescence	2.8	25	1	1.79
CD	2.8	25	1	1.79 ± 0.03
Fluorescence	0.05	5	1	1.81 ± 0.03
Fluorescence	0.05	25	1	1.79 ± 0.01
CD	28.4	5	1	1.82
CD	28.4	25	1	1.84
CD	2.8	5	0	1.74
CD	2.8	25	0	1.62

* Errors reported for replicated experiments.

Stabilities were measured using protein concentrations that varied from 0.05 μM to 28.4 μM, a greater than 500-fold range. Fluorescence was necessary for measuring the low concentration samples, because ecRNH has a weak CD signal but strong intrinsic fluorescence. CD reports on conformation of the entire polypeptide chain, whereas fluorescence reports on the local environment of tryptophan residues. Coincident probes are often used to support a two-state assumption. Importantly, the two probes were found to be coincident for melts performed using the same samples at 2.8 μM, so comparisons between CD and fluorescence experiments are justified. ecRNH's C_m is found to be independent of protein concentration over the tested range at both 5 °C and 25 °C. The results also show that TCEP effects C_m, but the observed decrease in C_m at low temperatures is insufficient to eliminate asymmetry in the stability curve. Melts of ecRNH*, which contains no cysteines, were also performed in the presence of TCEP, so the low temperature behavior would have to depend not just on TCEP but its effects on C63 specifically.

C63 is the residue responsible for asymmetry in the stability curve of ecRNH, and it is possible that TCEP contributes to its effect. All of the ancestors and ttRNH also contain C63, and all the stability curves were measured in the presence of TCEP. For ecRNH, ignoring stabilities measured at temperatures below 15 °C restores symmetry to the curve and results in thermodynamic parameters that agree with those measured for ecRNH* in the presence of TCEP and the published values of ecRNH* acquired without reducing agent [8]. For consistency, only data collected at temperatures above 15 °C are used in the stability curve fits for all of the proteins.

2.3.5 Unfolding and refolding kinetic characterization

Folding studies are used to further probe conservation of the RNase H energy landscape. ecRNH and ttRNH fold via a conserved three-state mechanism; however a single amino acid substitution is sufficient to convert ecRNH into a two-state folder [55]. The biological relevance of folding intermediates is the subject of much debate in the literature. One hypothesis is that intermediates contribute to efficient folding [56], and therefore they might be subject to positive selective pressures. By interrogating the folding pathways of the ancestors, we hope to learn whether the RNase H intermediate has been conserved throughout the evolution of this enzyme.

Refolding experiments of Node 2 were conducted using a stopped-flow CD spectrophotometer, and unfolding experiments were performed manually. CD signal at 222 nm was fit by a single exponential to give observed rate constants (**Figure 25**). The dependence of these rate constants on urea concentration creates an inverted V-shape characteristic of chevron plots (**Figure 26**). Data were fit by a three-state mechanism as described in section 2.4.8. It represents the same mechanism used to fit data for ecRNH* and ttRNH*, which assumes an initial pre-equilibrium between the unfolded state and an on-pathway intermediate that forms in the deadtime of the instrument (~15 ms) [1, 2].

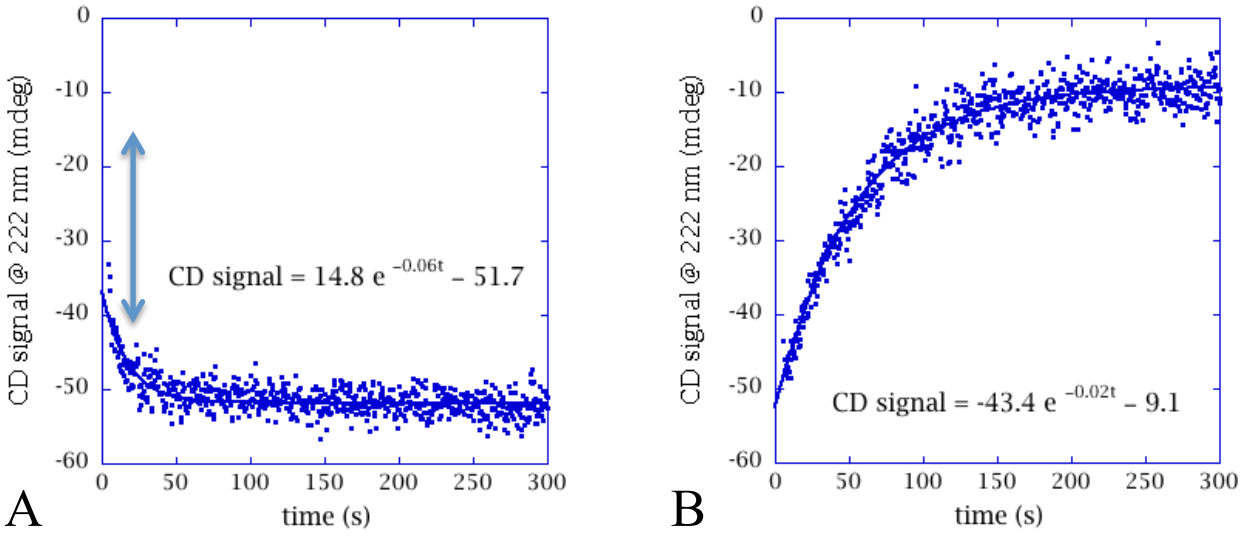


Figure 25. Representative kinetic data for Node 2 (A) refolding in 5M urea and (B) unfolding in 7M urea. Missing amplitude in the refolding experiment is indicated by the arrow.

Parameters extracted from Node 2's fit are compared with those published for ecRNH* and ttRNH* in **Table 9**. The parameters ΔG_{ui} and m_{ui} describe the equilibrium between the unfolded state and the folding intermediate. All three proteins share a similar m_{ui} value, suggesting the intermediate is similar in size; however the stability of Node 2's intermediate more closely resembles that of ttRNH*, and it is nearly twice as stable as ecRNH*'s. The parameters k_{in} and k_{ni} are the folding and unfolded microscopic rate constants, respectively, between the folded state and the intermediate. In these two parameters, Node 2 differs from both ecRNH* and ttRNH*. Node 2's k_{in} is 10-fold larger, which means its intermediate folds faster to the native state. Node

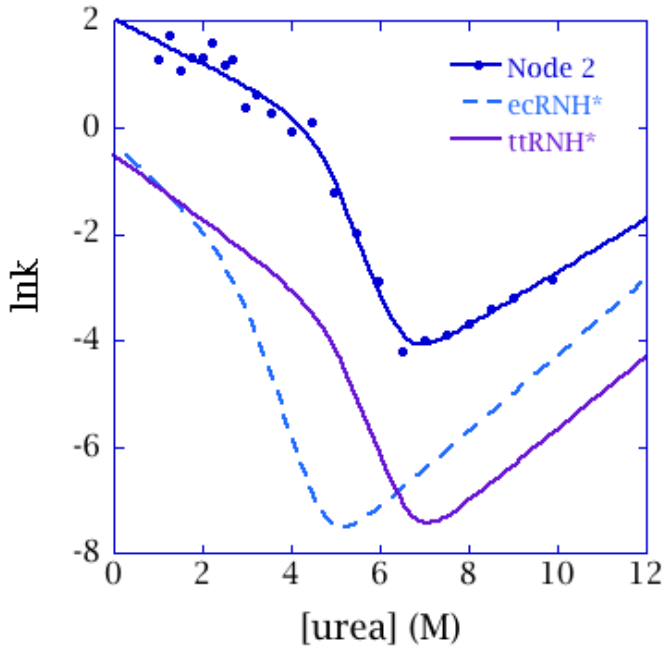


Figure 26. Chevrons of Node 2 (dark blue circles), ecRNH* (dashed, light blue line) and ttRNH* (solid, purple line). Node 2 data is fit to a three-state model (solid, dark blue line), as are the other two fits. The fit for ecRNH* is taken from [1], and the fit for ttRNH* is for its slow phase and is taken from [2].

2's k_{ni} is 100-fold larger than that of ttRNH* and 40-fold larger than that of ecRNH*, indicating it unfolds faster, as well.

Preliminary folding studies reveal that Node 2 folds by the same three-state mechanism observed in ecRNH* and ttRNH*. Thus, it appears that the intermediate is conserved in this ancestor. Further work is need to determine if it is similar structurally to the intermediate observed in extant RNases H and whether it exists in the other ancestors. Surprisingly, the ancestor also folds and unfolds several orders of magnitude faster than either extant protein without major changes to overall global stability. Future studies with the other ancestors are needed to determine if kinetic instability is a general feature of ancestral RNases H.

Table 9. Kinetic fit parameters for Node 2 compared with ecRNH* and ttRNH*

	Node 2	**ttRNH***	**ecRNH***
ΔG_{ui} (kcal mol^{-1})	6.33	6.2 ± 0.9†	3.55‡
m_{ui} (kcal mol^{-1} M^{-1})	1.34 ± 0.27	1.2 ± 0.2†	1.24‡
k_{in} (s^{-1})	7.66 ± 1.97	0.6 ± 0.1†	0.74 ± 0.02‡
m_{in} (kcal mol^{-1} M^{-1})	0.25 ± 0.07	0.36 ± 0.04†	0.454‡
k_{ni} (s^{-1})	4 x 10^{-4} ± 5 x 10^{-4}	4 x 10^{-6} ± 6 x 10^{-6}	1.1 x 10^{-5}‡
m_{ni} (kcal mol^{-1} M^{-1})	-0.30 ± 0.08	-0.4 ± 0.1†	-0.422‡

Errors are standard deviations from fits.
† Parameters from slow phase fit. Taken from reference [2].
‡ Taken from reference [1].

2.4 Materials and Methods

2.4.1 Ancestral protein resurrection

Bacterial and archaeal RNaseH sequences were identified by BLAST against the NCBI non-redundant protein database using RNases H from *E. coli* and *T. thermophilus* sequences as seed sequences [57, 58]. Redundant sequences were removed using cdhit 4.6 [59]. In total, 439 sequences were kept for further analysis. Sequences were aligned using MUSCLE 3.8.31 [60], followed by manual refinement using Mesquite 2.75 (Maddison and Maddison). Alignment quality was verified by checking for alignment of 3 universally conserved acidic residues that make up the RNase H active site. The final alignment is available in the appendix A1.1. The maximum likelihood phylogenetic tree was constructed using the JTT+Γ_8 substitution model and SPR moves as implemented in PhyML 3.0 [21, 22]. Branch supports were estimated using the approximate likelihood ratio test [23]. Maximum likelihood ancestral RNases H were reconstructed with the maximum likelihood topology, branch lengths, and phylogenetic model using PAML 3.14 [24, 25].

2.4.2 Expression and purification

Genes encoding the ancestral ancestors were codon optimized for expression in *E. coli* and synthesized by GENEART (Regensburg, Germany). The genes were subcloned using NdeI and HindIII restriction sites into the multiple cloning site of a pET27 vector (Life Technologies). Other site-specific variants were constructed via site-directed mutagenesis and verified by sequencing.

Plasmids were transformed into BL21(DE3)pLysS cells for expression under T7 promoter control. Cells were induced with 1 mM IPTG at OD = 0.6 and grown at 37 °C for 3 hours before harvesting. Cells were lysed in buffer via sonication. All ancestors expressed predominantly in the soluble fraction, though some partitioned into inclusion bodies as well. Only proteins expressing solubly were purified for further analysis. Lysate was purified first over a HiTrap Heparin column (GE Healthcase at pH 8. Peak fractions were pooled and diluted two-fold with doubly deionized water. Then the pH of the solution was adjusted to 5.5 using dilute NaOAc, and the sample was purified over a HiTrap S column (GE Healthcare). Protein was then concentrated and dialyzed against either ammonium bicarbonate for subsequent freeze-drying and storage or appropriate buffer conditions for immediate use. Each protein's purity and molecular weight were confirmed by SDS-PAGE and electrospray mass spectrometry. Protein concentrations were determined in Edelhoch buffer using extinction coefficients calculated based on the number of tryptophan and tyrosine residues [61].

2.4.3 Circular dichroism spectroscopy

CD spectra were collected on an AVIV 410 spectrophotometer using protein samples at 0.5 mg/ml in a 0.1 cm quartz cuvette at 25 °C. Data points were collected from 250-200 nm at 1-nm intervals, and each data point represents signal averaged over 5 seconds. Data for which the dynode voltage exceeded 500 V were discarded. Buffer conditions were 20 mM NaOAc (pH 5.5), 50 mM KCl and 1 mM TCEP.

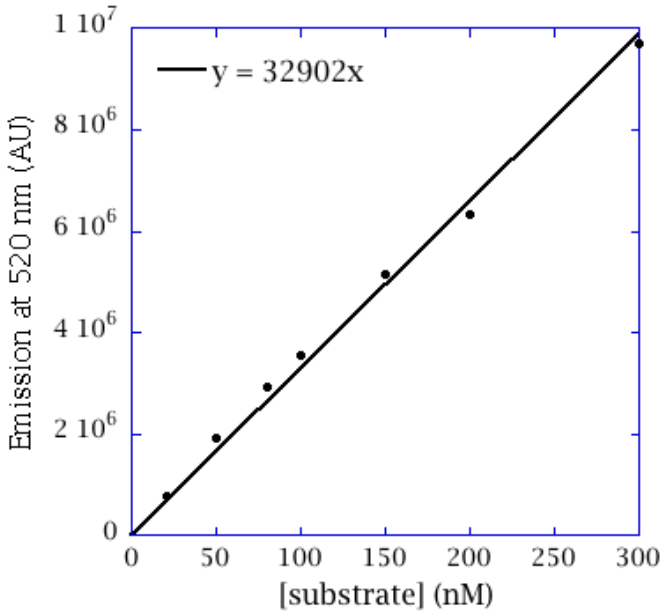


Figure 27. Final emission amplitudes at 520 nm for beacon substrate post-hydrolysis. The extracted scaling factor was used to determine K_ms and catalytic efficiencies.

2.4.4 Crystallization and structure determination of Node C

Crystals were grown at 18°C in hanging drop format by mixing 1 µl protein solution with 1 µl well solution containing 20% PEG 3350, 20-50 mM Li_2SO_4, 1 mM TCEP and 100 mM Bis-tris (pH 6.5). For harvesting, crystals were transferred for one minute to well solution containing 10% glycerol for cryoprotection, and then looped and flash frozen in liquid nitrogen. Data were collected at Beamline 8.3.1 (wavelength 1.1159 Å) under a cryo-stream at the Advanced Light Source (ALS) at Lawrence Berkeley National Laboratory, and integrated using HKL2000[62]. Initial phases were calculated by molecular replacement (MR) using PHASER [26]. The search model was the extant RNase H from *E. coli* (PDB ID code 2RN2). Building of the model was carried out in COOT [63], followed by a refinement strategy using PHENIX [64] that consisted of an initial round of rigid-body refinement, followed by individual-atom positional and anisotropic ADP refinement including hydrogens. Structure validation was assisted by MolProbity [65], and figures were rendered using PyMOL [66].

2.4.5 Activity assays

2.4.5.1 Fluorescent beacon assay

RNase H activity was assayed in 50 mM NaCl, 50 mM Tris (pH 8.0), 10 mM $MgCl_2$ at 25 °C. Substrate was prepared by mixing equimolar amounts of 5'-fluorescein-labeled DNA oligomer (CCGTCATCTC) and 3'-DABCYL-labeled RNA oligomer (GAGAUGACGG) (IDT), heating to 95 °C for 5 min, then slowly cooling to room temperature for one hour before storing at 4 °C. Substrate concentrations assume complete hybridization and are given in moles of the RNA-DNA hybrid. The reaction was initiated with addition of enzyme, excited at 490 nm and monitored at 520 nm using a Fluoromax 3 fluorimeter (Horiba). Increasing fluorescence at 520 nm indicates the release of fluorescein as RNA is hydrolyzed. Final emission amplitudes post-hydrolysis from known amounts of substrate were used to convert signal to K_ms and catalytic

efficiencies. Complete hydrolysis of 1 nM substrate results in a signal of 32,902 counts at 520 nm (**Figure 8**).
Initial velocities were measured in the 40-60 second range, which represents the first 10-30 seconds post-initiation, and fit to the Michaelis-Menten equation using KaleidaGraph (version 4.1.2):

$$v = \frac{k_{cat}[\,\mathrm{E}\,][\,\mathrm{S}\,]}{K_m + [\,\mathrm{S}\,]} \tag{5}$$

2.4.5.2 Hyperchromic assay
RNase H activity was assayed in 50 mM NaCl, 50 mM Tris (pH 8.0), 10 mM $MgCl_2$ at 25 °C. Substrate concentration is given in internucletotide bonds, due to the heterogeous nature of the substrate, using ε_{260} = 8250 $M^{-1}cm^{-1}$ and 330 g/mol for the average nucleotide molecular weight. Substrate was prepared by mixing equal parts dT_{20} oligomers (IDT) and poly-rA (Sigma), heating to 95 °C for 5 min, then slowly cooling to room temperature for one hour before storing at 4 °C. The reaction was initiated with the addition of enzyme and monitored at 260 nm using a Cary UV spectrophotometer. Increasing absorbance at 260 nm indicates the release of nucleotides as they are hydrolyzed.

2.4.6 Denaturant-induced and thermal denaturation

Thermal and denaturation melts were performed in 20 mM NaOAc (pH 5.5), 50 mM KCl and 1mM TCEP. Melts monitored by CD were followed at 222 nm used 50 μg/mL protein in a 1-cm pathlength quartz cuvette, unless otherwise noted, and 60 seconds of signal were averaged for each data point. Denaturant melt samples were prepared individually and allowed to equilibrate at the appropriate temperature overnight. Samples were allowed to stir in the instrument for 1-2 minutes before data were collected. Alternatively, denaturant titrations were used at higher temperatures with 5-15 minutes of equilibration.

To measure CD signal at 222 nm as a function of temperature, samples were allowed to equilibrate for 5 minutes at each temperature and data were collected every 3 °C. Spectra were taken at 25 °C before and after the thermal melt to test for reversibility. Temperature melts were fit to a two-state model using the Gibbs-Helmholtz relationship (**Equation 1**).

Melts monitored by tryptophan fluorescence were excited at 280 nm, and emission spectra were recorded from 300 - 400 nm. Fluorescence at 340 nm as well as the center of mass were analyzed. Denaturant concentrations were verified using a refractometer. Data were fit using a two-state approximation and assume a linear dependence of ΔG on denaturant concentration.

2.4.7 Denaturation and stability curve data analysis

To generate stability curves, average global stabilities derived from GdmCl-induced denaturation melts were plotted as a function of temperature. T_ms extracted from thermal denaturation experiments were used as single points at ΔG = 0 for ttRNH, nodes A, 1, 2, and 3, which all unfold reversibly. Data were fit to two versions of the Gibbs-Helmholtz equation. Equation 1, which uses T_m as the reference temperature, was used to extract ΔC_p, ΔH_m and T_ms. Equation 3,

which uses T_s as the reference temperature, was used to extract ΔC_p, T_s, T_h and to calculated ΔH_s (**Equation 4**).

2.4.8 Unfolding and refolding kinetics

Unfolding and refolding kinetics were measured at 25 °C. Unfolding was initiated manually in a 1-cm pathlength quartz cuvette by a 30-fold dilution of folded stock [2 mg/mL Node 2, 4 M urea, 20 mM NaOAc (pH 5.5), 20 mM KCl, 1 mM TCEP] into unfolding buffer containing the appropriate concentration of urea. Final denaturant concentrations were verified using a refractometer.

Refolding experiments were performed in an AVIV 202SF stopped-flow CD with a 1-mm pathlength. Refolding was initiated by an 11-fold dilution of unfolded stock [6 mg/mL Node 2, 7.5 M urea, 20 mM NaOAc (pH 5.5), 20 mM KCl, 1 mM TCEP] into folding buffer containing the appropriate concentration of urea.

Data were fit to a single exponential using KaleidaGraph (version 4.1.2):

$$\text{signal} = A\, \mathrm{e}^{-k_{obs}t} + C \tag{6}$$

where C is the final amplitude, A is the amplitude of the observable phase, t is time and k_{obs} is the observed rate constant. Refolding experiments had missing amplitude in the dead time of the instrument. It was assumed that these burst phase amplitudes describe a two-state system (U ↔ I), so the observed rates were fit to a three-state on-pathway model (U ↔ I ↔ N) using the following equations:

$$\ln k_{obs} = \ln\, [[\frac{K_{ui}}{1+K_{ui}}]\, k_{in}\, + k_{ni}\,] \tag{7}$$

$$k_{in} = k_{in,H2O}\, e^{-m_{in}[urea]/RT} \tag{8}$$

$$k_{ni} = k_{ni,H2O}\, e^{-m_{ni}[urea]/RT} \tag{9}$$

where K_{ui} and m_{ui} describe the equilibrium constant and *m-value* between U and I, respectively; k_{in} and k_{ni} are the folding and unfolding microscopic rate constants between N and I, respectively; m_{in} and m_{ni} are the denaturant dependencies of these rate constants; R is the universal gas constant and T is temperature.

2.5 Discussion

In this study, we asked whether the observed thermodynamic differences between *Escherichia coli* (ecRNH) and its homolog from *Thermus thermophilus* (ttRNH) could be understood from a historical perspective by resurrecting their most recent common ancestor as well as evolutionary intermediates. We observe pronounced trends in melting temperature, reversibility and stability along each lineage, but the related feature of ΔC_p fluctuates in a stochastic manner within upper and lower bounds established by the extant proteins.

Proteins are physical entities subject to the physiochemical realities of their constituent atoms; however, they are also historical artifacts descended from, and thus dependent on, their evolutionary predecessors. While it seems unlikely, in our view, that sequence space has been exhaustively sampled in every biological context, certain traits may have been. For instance, a given protein's stability can be encoded by many different sequences, but introducing a single amino acid substitution is also likely to change its stability. This apparent contradiction means that properties like stability have the potential to converge on optimal solutions. Indeed it has been suggested that over evolutionary time, amino acids within a sequence have been sampled in accordance to a Boltzmann distribution of their individual energetic contributions [67, 68]. This pseudo-equilibrium, or Boltzmann hypothesis, stems from the observation that the strength of a particular interaction within a polypeptide chain can be predicted based on its frequency within a collection of known protein structures. Likewise, experimental evidence from studies of thioredoxin demonstrates that the destabilizing effects of conservative amino acid substitutions correlate with their frequency within the protein family [69, 70]. Clearly, however, not all residues can conform to the Boltzmann hypothesis. Active site residues, for example, might be able to accommodate one or two different amino acids, but most substitutions will destroy function. Thus, the pseudo-equilibrium applies only to neutral or nearly neutral substitution, and it implies that the range of acceptable evolutionary solutions, defining the so-called "neutral corridor" [71], is established, in part, by the protein being stable and folded.

Our results indicate that the melting temperatures and stabilities of ecRNH and ttRNH are similar to traits observed in their more recent ancestors; however, the T_ms of the ancestors are not modulated using the same thermodynamic strategies observed in their descendants. We had hypothesized that the ancestors of ttRNH might share its low ΔC_p, which results from structure in the unfolded state [12]. What we find, however, is that the recent ancestors share high T_ms but have variable ΔC_ps. The lack of correlation between T_m and ΔC_p, while somewhat unexpected for this lineage, is consistent with another, more distantly related RNase H homolog from *Chlorobium tepidum*. *C. tepidum* is a moderate thermophile with an optimal growth temperature of 48 °C. Its RNase H has a low, mesophilic-like T_m of 66.5 °C but also a low ΔC_p of 1.7 kcal mol^{-1} K^{-1} similar to that observed in ttRNH . Thus, in the context of the RNase H fold, T_m and ΔC_p are not correlated. Instead, conservation of thermostability along ecRNH's and ttRNH's respective lineages is achieved via a combination of thermodynamic mechanisms. Because it is impossible to know the environment in which the ancestral states functioned, shared traits are consistent with models of either contingent or deterministic evolution. Future studies are needed to assess the ancestors' impacts on organismal fitness and to compare the resurrected ancestors with RNases H evolved *in vitro* in predetermined environments.

Previous work led us to hypothesize that ΔC_p, which reports on residual structure in the unfolded state of RNase H, might represent such a feature; however, the apparently random fluctuations observed in ΔC_p along both lineages has led us to conclude that it is not subject to strong selective pressures. We believe, however, that the distinctive trends in melting temperature between the mesophilic and thermophilic lineages, which show that the mesophilic T_m was maintained over evolutionary time and thermostability of ttRNH developed via a gradual process, represent evidence that protein energetics are subject to selective pressures. It remains unclear, however, whether T_m itself is the parameter under selection or whether it reflects selection on another related feature of the protein's energy landscape, such as unfolding rate or the accessibility of high-energy states. Our results suggest that T_m, and possibly global stability, represent constrained phenotypes, but the mechanism for tuning thermostabilty is highly plastic and non-deterministic.

2.6 References

1. Raschke, T.M., J. Kho, and S. Marqusee, *Confirmation of the hierarchical folding of RNase H: a protein engineering study.* Nat Struct Biol, 1999. **6**(9): p. 825-31.
2. Hollien, J. and S. Marqusee, *Comparison of the folding processes of T. thermophilus and E. coli ribonucleases H.* J Mol Biol, 2002. **316**(2): p. 327-40.
3. Crooks, G.E., et al., *WebLogo: a sequence logo generator.* Genome Res, 2004. **14**(6): p. 1188-90.
4. Razvi, A. and J.M. Scholtz, *Lessons in stability from thermophilic proteins.* Protein Sci, 2006. **15**(7): p. 1569-78.
5. Karshikoff, A. and R. Ladenstein, *Ion pairs and the thermotolerance of proteins from hyperthermophiles: a "traffic rule" for hot roads.* Trends Biochem Sci, 2001. **26**(9): p. 550-6.
6. Kumar, S., C.J. Tsai, and R. Nussinov, *Thermodynamic differences among homologous thermophilic and mesophilic proteins.* Biochemistry, 2001. **40**(47): p. 14152-65.
7. Gromiha, M.M., M. Oobatake, and A. Sarai, *Important amino acid properties for enhanced thermostability from mesophilic to thermophilic proteins.* Biophys Chem, 1999. **82**(1): p. 51-67.
8. Hollien, J. and S. Marqusee, *A thermodynamic comparison of mesophilic and thermophilic ribonucleases H.* Biochemistry, 1999. **38**(12): p. 3831-6.
9. Becktel, W.J. and J.A. Schellman, *Protein stability curves.* Biopolymers, 1987. **26**(11): p. 1859-77.
10. Baldwin, R.L., *Temperature dependence of the hydrophobic interaction in protein folding.* Proc Natl Acad Sci U S A, 1986. **83**(21): p. 8069-72.
11. Robic, S., J.M. Berger, and S. Marqusee, *Contributions of folding cores to the thermostabilities of two ribonucleases H.* Protein Sci, 2002. **11**(2): p. 381-9.
12. Robic, S., et al., *Role of residual structure in the unfolded state of a thermophilic protein.* Proc Natl Acad Sci U S A, 2003. **100**(20): p. 11345-9.
13. Guzman-Casado, M., et al., *Energetic evidence for formation of a pH-dependent hydrophobic cluster in the denatured state of Thermus thermophilus ribonuclease H.* J Mol Biol, 2003. **329**(4): p. 731-43.
14. Dryden, D.T., A.R. Thomson, and J.H. White, *How much of protein sequence space has been explored by life on Earth?* J R Soc Interface, 2008. **5**(25): p. 953-6.
15. Luisi, P.L., *Chemical aspects of synthetic biology.* Chem Biodivers, 2007. **4**(4): p. 603-21.
16. Harms, M.J. and J.W. Thornton, *Analyzing protein structure and function using ancestral gene reconstruction.* Curr Opin Struct Biol, 2010. **20**(3): p. 360-6.
17. Gaucher, E.A., et al., *Inferring the palaeoenvironment of ancient bacteria on the basis of resurrected proteins.* Nature, 2003. **425**(6955): p. 285-8.
18. Gaucher, E.A., S. Govindarajan, and O.K. Ganesh, *Palaeotemperature trend for Precambrian life inferred from resurrected proteins.* Nature, 2008. **451**(7179): p. 704-7.
19. Perez-Jimenez, R., et al., *Single-molecule paleoenzymology probes the chemistry of resurrected enzymes.* Nat Struct Mol Biol, 2011. **18**(5): p. 592-6.

20. Hobbs, J.K., et al., *On the origin and evolution of thermophily: reconstruction of functional precambrian enzymes from ancestors of Bacillus.* Mol Biol Evol, 2012. **29**(2): p. 825-35.
21. Guindon, S., et al., *New algorithms and methods to estimate maximum-likelihood phylogenies: assessing the performance of PhyML 3.0.* Syst Biol, 2010. **59**(3): p. 307-21.
22. Jones, D.T., W.R. Taylor, and J.M. Thornton, *The rapid generation of mutation data matrices from protein sequences.* Comput Appl Biosci, 1992. **8**(3): p. 275-82.
23. Anisimova, M. and O. Gascuel, *Approximate likelihood-ratio test for branches: A fast, accurate, and powerful alternative.* Syst Biol, 2006. **55**(4): p. 539-52.
24. Yang, Z., S. Kumar, and M. Nei, *A new method of inference of ancestral nucleotide and amino acid sequences.* Genetics, 1995. **141**(4): p. 1641-50.
25. Yang, Z., *PAML 4: phylogenetic analysis by maximum likelihood.* Mol Biol Evol, 2007. **24**(8): p. 1586-91.
26. McCoy, A.J., et al., *Phaser crystallographic software.* J Appl Crystallogr, 2007. **40**(Pt 4): p. 658-674.
27. Keck, J.L., E.R. Goedken, and S. Marqusee, *Activation/attenuation model for RNase H. A one-metal mechanism with second-metal inhibition.* J Biol Chem, 1998. **273**(51): p. 34128-33.
28. Crooke, S.T., et al., *Kinetic characteristics of Escherichia coli RNase H1: cleavage of various antisense oligonucleotide-RNA duplexes.* Biochem J, 1995. **312 (Pt 2)**: p. 599-608.
29. Rizzo, J., et al., *Chimeric RNA-DNA molecular beacon assay for ribonuclease H activity.* Mol Cell Probes, 2002. **16**(4): p. 277-83.
30. Nowotny, M., et al., *Crystal structures of RNase H bound to an RNA/DNA hybrid: substrate specificity and metal-dependent catalysis.* Cell, 2005. **121**(7): p. 1005-16.
31. Dumousseau, M., et al., *MELTING, a flexible platform to predict the melting temperatures of nucleic acids.* BMC Bioinformatics, 2012. **13**: p. 101.
32. Ghosh, K. and K. Dill, *Cellular proteomes have broad distributions of protein stability.* Biophys J, 2010. **99**(12): p. 3996-4002.
33. Lepock, J.R., *Measurement of protein stability and protein denaturation in cells using differential scanning calorimetry.* Methods, 2005. **35**(2): p. 117-25.
34. Tadokoro, T., et al., *Structural, thermodynamic, and mutational analyses of a psychrotrophic RNase HI.* Biochemistry, 2007. **46**(25): p. 7460-8.
35. Ratcliff, K., J. Corn, and S. Marqusee, *Structure, stability, and folding of ribonuclease H1 from the moderately thermophilic Chlorobium tepidum: comparison with thermophilic and mesophilic homologues.* Biochemistry, 2009. **48**(25): p. 5890-8.
36. Parte, A., *Bergey's manual of systematic bacteriology*2012, New York: Springer.
37. Iversen, C., et al., *Cronobacter gen. nov., a new genus to accommodate the biogroups of Enterobacter sakazakii, and proposal of Cronobacter sakazakii gen. nov., comb. nov., Cronobacter malonaticus sp. nov., Cronobacter turicensis sp. nov., Cronobacter muytjensii sp. nov., Cronobacter dublinensis sp. nov., Cronobacter genomospecies 1, and of three subspecies, Cronobacter dublinensis subsp. dublinensis subsp. nov., Cronobacter dublinensis subsp. lausannensis subsp. nov. and Cronobacter dublinensis subsp. lactaridi subsp. nov.* Int J Syst Evol Microbiol, 2008. **58**(Pt 6): p. 1442-7.

38. Darby, A.C., et al., *Aphid-symbiotic bacteria cultured in insect cell lines.* Appl Environ Microbiol, 2005. **71**(8): p. 4833-9.
39. Taghavi, S., et al., *Genome sequence of the plant growth promoting endophytic bacterium Enterobacter sp. 638.* PLoS Genet, 2010. **6**(5): p. e1000943.
40. Bolen, D.W. and M.M. Santoro, *Unfolding free energy changes determined by the linear extrapolation method. 2. Incorporation of delta G degrees N-U values in a thermodynamic cycle.* Biochemistry, 1988. **27**(21): p. 8069-74.
41. Myers, J.K., C.N. Pace, and J.M. Scholtz, *Denaturant m values and heat capacity changes: relation to changes in accessible surface areas of protein unfolding.* Protein Sci, 1995. **4**(10): p. 2138-48.
42. Spudich, G. and S. Marqusee, *A change in the apparent m value reveals a populated intermediate under equilibrium conditions in Escherichia coli ribonuclease HI.* Biochemistry, 2000. **39**(38): p. 11677-83.
43. Wrabl, J. and D. Shortle, *A model of the changes in denatured state structure underlying m value effects in staphylococcal nuclease.* Nat Struct Biol, 1999. **6**(9): p. 876-83.
44. Baskakov, I.V. and D.W. Bolen, *The paradox between m values and deltaCp's for denaturation of ribonuclease T1 with disulfide bonds intact and broken.* Protein Sci, 1999. **8**(6): p. 1314-9.
45. Goedken, E.R. and S. Marqusee, *Native-state energetics of a thermostabilized variant of ribonuclease HI.* J Mol Biol, 2001. **314**(4): p. 863-71.
46. Chan, C.H., T.H. Yu, and K.B. Wong, *Stabilizing salt-bridge enhances protein thermostability by reducing the heat capacity change of unfolding.* PLoS One, 2011. **6**(6): p. e21624.
47. Butterwick, J.A. and A.G. Palmer, 3rd, *An inserted Gly residue fine tunes dynamics between mesophilic and thermophilic ribonucleases H.* Protein Sci, 2006. **15**(12): p. 2697-707.
48. Pace, C.N. and D.V. Laurents, *A new method for determining the heat capacity change for protein folding.* Biochemistry, 1989. **28**(6): p. 2520-5.
49. Privalov, P.L., *Stability of proteins: small globular proteins.* Adv Protein Chem, 1979. **33**: p. 167-241.
50. Nicholson, E.M. and J.M. Scholtz, *Conformational stability of the Escherichia coli HPr protein: test of the linear extrapolation method and a thermodynamic characterization of cold denaturation.* Biochemistry, 1996. **35**(35): p. 11369-78.
51. Zweifel, M.E. and D. Barrick, *Studies of the ankyrin repeats of the Drosophila melanogaster Notch receptor. 2. Solution stability and cooperativity of unfolding.* Biochemistry, 2001. **40**(48): p. 14357-67.
52. Deutschman, W.A. and F.W. Dahlquist, *Thermodynamic basis for the increased thermostability of CheY from the hyperthermophile Thermotoga maritima.* Biochemistry, 2001. **40**(43): p. 13107-13.
53. Agashe, V.R. and J.B. Udgaonkar, *Thermodynamics of denaturation of barstar: evidence for cold denaturation and evaluation of the interaction with guanidine hydrochloride.* Biochemistry, 1995. **34**(10): p. 3286-99.
54. Zweifel, M.E. and D. Barrick, *Relationships between the temperature dependence of solvent denaturation and the denaturant dependence of protein stability curves.* Biophys Chem, 2002. **101-102**: p. 221-37.

55. Spudich, G.M., E.J. Miller, and S. Marqusee, *Destabilization of the Escherichia coli RNase H kinetic intermediate: switching between a two-state and three-state folding mechanism.* J Mol Biol, 2004. **335**(2): p. 609-18.
56. Brockwell, D.J. and S.E. Radford, *Intermediates: ubiquitous species on folding energy landscapes?* Curr Opin Struct Biol, 2007. **17**(1): p. 30-7.
57. Altschul, S.F., et al., *Gapped BLAST and PSI-BLAST: a new generation of protein database search programs.* Nucleic Acids Res, 1997. **25**(17): p. 3389-402.
58. Pruitt, K.D., T. Tatusova, and D.R. Maglott, *NCBI Reference Sequence (RefSeq): a curated non-redundant sequence database of genomes, transcripts and proteins.* Nucleic Acids Res, 2005. **33**(Database issue): p. D501-4.
59. Huang, Y., et al., *CD-HIT Suite: a web server for clustering and comparing biological sequences.* Bioinformatics, 2010. **26**(5): p. 680-2.
60. Edgar, R.C., *MUSCLE: a multiple sequence alignment method with reduced time and space complexity.* BMC Bioinformatics, 2004. **5**: p. 113.
61. Edelhoch, H., *Spectroscopic determination of tryptophan and tyrosine in proteins.* Biochemistry, 1967. **6**(7): p. 1948-54.
62. Otwinowski, Z. and W. Minor, *Processing of X-ray diffraction data collected in oscillation mode.* Methods Enzymol., 1997. **276**: p. 307-326.
63. Emsley, P., et al., *Features and development of Coot.* Acta Crystallogr D Biol Crystallogr, 2010. **66**(Pt 4): p. 486-501.
64. Adams, P.D., et al., *PHENIX: a comprehensive Python-based system for macromolecular structure solution.* Acta Crystallogr D Biol Crystallogr, 2010. **66**(Pt 2): p. 213-21.
65. Chen, V.B., et al., *MolProbity: all-atom structure validation for macromolecular crystallography.* Acta Crystallogr D Biol Crystallogr, 2010. **66**(Pt 1): p. 12-21.
66. Spudich, G., S. Lorenz, and S. Marqusee, *Propagation of a single destabilizing mutation throughout the Escherichia coli ribonuclease HI native state.* Protein Sci, 2002. **11**(3): p. 522-8.
67. Shortle, D., *Propensities, probabilities, and the Boltzmann hypothesis.* Protein Sci, 2003. **12**(6): p. 1298-302.
68. Finkelstein, A.V., A.M. Gutin, and A. Badretdinov, *Boltzmann-like statistics of protein architectures. Origins and consequences.* Subcell Biochem, 1995. **24**: p. 1-26.
69. Godoy-Ruiz, R., et al., *Relation between protein stability, evolution and structure, as probed by carboxylic acid mutations.* J Mol Biol, 2004. **336**(2): p. 313-8.
70. Godoy-Ruiz, R., et al., *A stability pattern of protein hydrophobic mutations that reflects evolutionary structural optimization.* Biophys J, 2005. **89**(5): p. 3320-31.
71. Shih, P., et al., *Reconstruction and testing of ancestral proteins.* Methods Enzymol, 1993. **224**: p. 576-90.

CHAPTER 3

Probing dynamics of the unfolded state under native conditions using neutron scattering

Conducted in collaboration with Francois-Xavier Gallat and Martin Weik, Institut de Biologie Structurale (Grenoble, France)

3.1 Abstract

The dynamics of an intrinsically disordered protein are compared with that of an unfolded protein variant that typically has a well-defined native structure. While the IDP tau, a microtubule-bound protein, and the unfolded variant of maltose-binding protein share similar dimensions, as probed by SAXS in low denaturant conditions, they show distinctive pico-nanosecond sidechain motions in the powder state.

3.2 Introduction

Proteins are dynamic systems whose motions are critical for function. Large-scale conformational changes, occurring on the millisecond-second timescale, are required for folding to the native state and are often involved in complex responses such as allostery and molecular recognition. While these large, slow motions are the subject of much experimental study, less is known about the equally important fast dynamics that facilitate protein-water interactions.

Protein and hydration-water dynamics mutually affect each other [1]. It has been observed that the extent of coupling differs between protein classes, which may reflect distinct functional requirements. For instance, the intrinsically disordered protein tau shows more coupling to its hydration waters than folded, globular proteins; whereas the membrane protein bacteriorhodopsin shows almost no coupling to solvent [2, 3]. Although there exists a gradient of coupling across protein classes, all protein dynamics are influenced by solvent interactions.

One technique used to probe these protein-water dynamics, which occur on the pico-nanosecond timescale and at angstrom (Å) length scales, is elastic incoherent neutron scattering (EINS). The signal from EINS is dominated by neutrons' interactions with hydrogens mainly situated on protein sidechains. The resulting mean square displacements (MSDs) represent an averaged parameter of the system and thus reflect some feature of the protein's global flexibility [4]. MSDs are typically measured as a function of temperature. At low temperatures (< 200 K), MSDs from hydrated and dehydrated proteins are indistinguishable. Around 200-220 K, however, hydrated proteins undergo the so-called "dynamical transition," as evidenced by a marked increase in the slope of MSD versus temperature. This dynamical transition has been noted to coincide with the minimum temperature necessary for protein function, with at least one notable exception [5]. Ribonuclease A, for instance, is able to bind its substrate at 228 K, which is just above its dynamical transition, but not at 212 K [6]. Dehydrated proteins lack a dynamical transition all together, which results in significantly dampened dynamics at physiological temperatures [7]. The rigidity observed in dehydrated proteins is consistent with the fact that most proteins cannot function in the absence of hydration water, because protein flexibility is necessary to facilitate catalysis and binding [8]. The physical basis for the dynamical transition observed in the hydrated protein remains unclear. The temperature of its onset is highly dependent on the instrument resolution, giving the impression of a "window effect" and not a real transition; however, analysis by orthogonal techniques, such as terahertz and Mössbauer spectroscopies, also reveal a transition at 200 K. Mössbauer spectroscopy measures dynamics on on the 100-ns timescale, which is between 3 to 6 orders of magnitude slower than those measured by EINS. Coincidence of temperatures measured by disparate techniques in spite of

their different timescales suggests that the transition reflects real dynamical changes in the system.

Curiously, the MSDs of similarly sized proteins are nearly identical [2]. This may be due to the fact that most proteins studied by EINS fall into the folded, globular class of proteins. Recently, MSDs were measured for an intrinsically disordered protein, human tau protein isoform 40 [2]. Tau's dynamics track closely with those of a representative folded protein, maltose-binding protein (MBP), until about 270 K, which is well after the dynamical transition. Above this temperature, tau's dynamics increase more sharply than MBP's such that at room temperature, the MSDs of tau are 50% greater than those of MBP. These enhanced sidechain dynamics could arise either from differences in amino acid composition, increased conformational freedom of the entire polypeptide chain or both. Amino acid sidechains are known to have unique, intrinsic pico-nanosecond dynamics [9], and intrinsically disordered proteins have distinctive amino acid compositions [10]; however, there is evidence that motions probed above the dynamical transition largely reflect conformational flexibility of the backbone [9]. Furthermore, extreme backbone flexibility is consistent with other structural studies of tau. Detailed NMR analysis, for instance, reveals that tau protein exists as a heterogeneous conformational ensemble [11], and tau has been shown to bind microtubules in several partially folded conformations [12]. Thus, it is reasonable to interpret high MSDs as reflecting motions associated with interconversion between isoenergetic states.

In this study, we employ EINS to investigate the dynamics of the unfolded state of MBP populated under the same "native-like" conditions. By comparing dynamics between an intrinsically disordered protein and both the folded and unfolded state of a foldable protein under the same conditions, we can distinguish the contribution of backbone flexibility from those intrinsic to the amino acid sequence. We find, as expected, that the unfolded state of MBP is more dynamic than folded MBP. To our surprise, however, unfolded MBP is also more dynamic than the intrinsically disordered protein tau. Potential functional implications for this discrepancy are addressed in the discussion section.

3.3 Results

3.3.1 Design of an unfolded MBP

An unfolded MBP was designed by introducing five destabilizing amino acid substitutions to the wild-type MBP sequence. The resulting construct, referred to as MBP0 throughout this work, is a quintuple site-specific variant. The specific substitutions, I59A/L115A/L147A/I161A/I226A, were chosen based on their published individual changes to protein stability *(*$\Delta\Delta G$ values) [13]. Assuming the sites are independent, the global stability of MBP0 is expected to be close to 0 kcal/mol, corresponding to a 1:1 ratio of folded:unfolded molecules, assuming a two-state system. Three of the positions, L115A/L147A/I226A, are clustered within van der Waals contact range in the crystal structure (**Figure 1**) [14]. Thus it is likely that the substitutions cause an even greater destabilization than implied by simply adding their individual $\Delta\Delta G$ values, which would cause a corresponding increase in the population of unfolded molecules under native conditions.

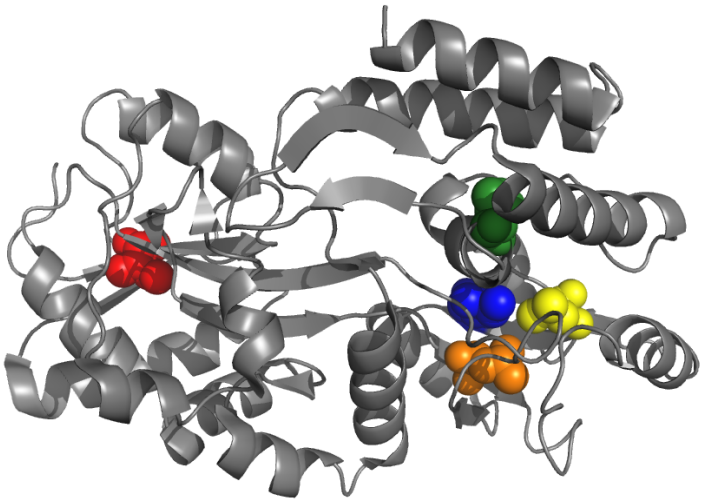


Figure 1. Location of destabilizing substitutions in MBP (1OMP). Residues I59 (red), L115 (orange), L147 (yellow), I161 (green) and I226 (blue) were all changed to alanine to create the variant MBP0. Image was made using MacPyMOL version 1.3.

3.3.2 Circular dichroism spectroscopy

The secondary structure content of MBP0 was probed using far-UV circular dichroism (CD) spectroscopy. The spectrum for MBP0 in the absence of maltose (**Figure 2A**) shows very little signal at 222 nm, consistent with an unfolded protein that lacks significant secondary structure. Addition of 100 mM maltose reveals a time-dependent change in the CD spectrum consistent with the expected increase in stability in the presence of ligand (**Figure 2A**). The CD signal at 222 nm decreases with a k_{obs} = 0.06 hours^{-1}, which corresponds to a mean lifetime of 15.9 hours (**Figure 2B**). This signal change indicates that the protein folds in the presence of maltose, albeit at a very slow rate.

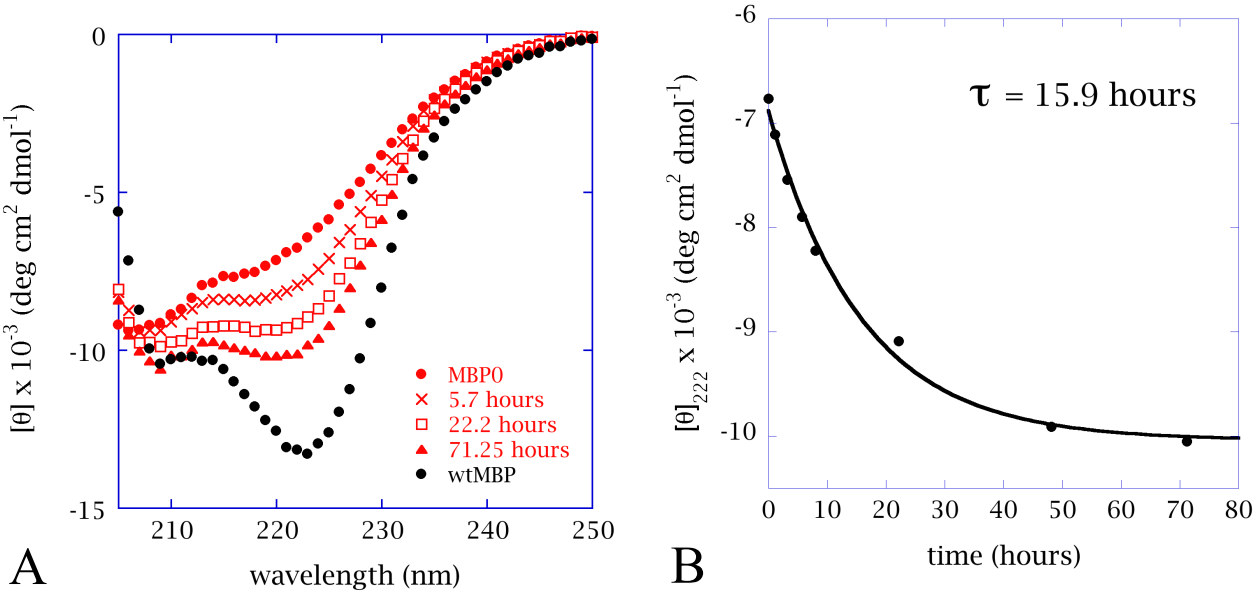


Figure 2. (A) CD spectra of MBP0 (red circles), wt MBP (black circles) and MBP0 after the addition of 100 mM maltose (5.7 hours, red x's; 22.2 hours, red open squares; 71.25 hours, red triangles). (B) Kinetics of MBP0 folding in the presence of 100 mM maltose fit to a single exponential (black curve).

The chemical denaturation profile of MBP0 confirms that the stability of MBP0 is severely compromised compared to the wild-type protein (**Figure 3)**. The data cannot be fit with a two-state transition, due to the lack of any folded baseline; however, a comparison with the folded signal for the wild-type protein suggest that MBP0 is predominantly unfolded even the absence of denaturant (**Figure 3**). The CD signal, however, does not match the expected signal based on the extrapolated value from wt MBP's unfolded baseline suggesting that MBP0 does show some residual secondary structure in the absence of denaturant. The lack of a cooperative unfolding transition, however, indicates that this structure is likely non-native and not well-defined.

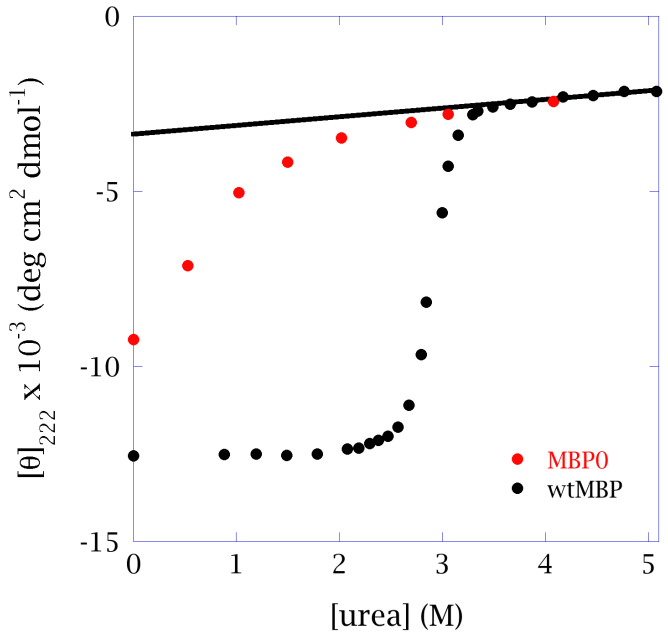


Figure 3. Urea denaturation profiles of MBP0 (red circles) and wt MBP (black circles) overlaid with the extrapolated unfolded baseline for wt MBP (black line).

3.3.3 Size exclusion chromatography

MBP0 is prone to aggregation, as evidenced by visible precipitation at high concentrations (~0.5 mg/ml). Both the aggregated and soluble forms of MBP0 (see section 3.4.4) were evaluated by size exclusion chromatography (**Figure 4**). Protein from the aggregated sample eluted at a volume of 46 mL just after blue dextran (44 mL), which is a large molecule used to demarcate the dead volume of a column. The absence of peaks at longer elution volumes indicates the absence of any small (monomer or small oligomers) of the protein within the aggregated sample. Conversely, the injection of the soluble MBP0 sample eluted at a volume of 70 mL, indicating a monodisperse solution. This elution volume corresponds to an apparent molecular mass of 110 kDa and an apparent hydrodynamic radius of 4.4 nm. This is larger than the calculated molecular mass of MBP0 (40.6 kDa). For comparison, ovalbumin at 43 kDa elutes at 80 mL with a corresponding hydrodynamic radius of 3 nm. The smaller elution volume and corresponding larger hydrodynamic radius of MBP0 is consistent with expectations for an unfolded protein.

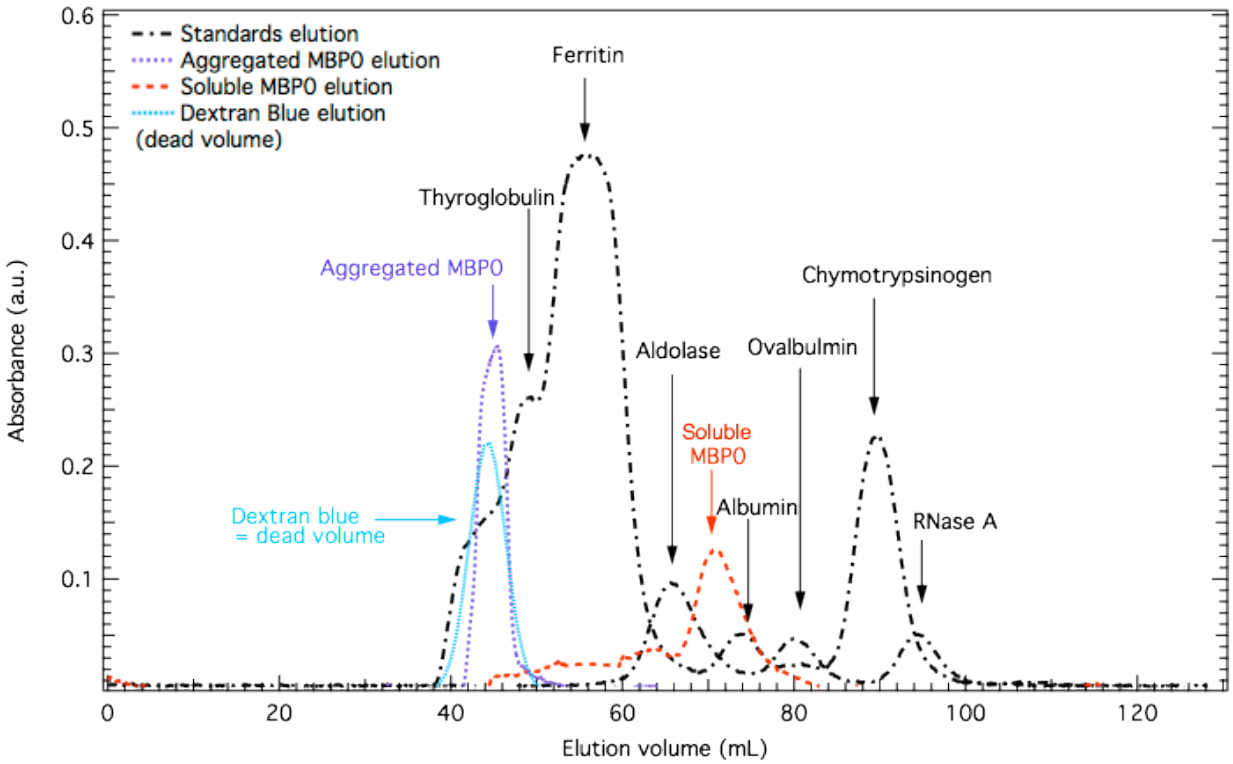


Figure 4. Elution profiles from size exclusion chromatography of standards (dashed black lines), soluble MBP0 (red dashed curve) and aggregated MBP0 (purple dotted curve) on a Superdex S200 (Amersham), equilibrated in H_2O. The dead volume was defined as the elution volume of blue dextran at 42 mL. *Data from FX Gallat.*

3.3.4 SAXS experiments

The radii of gyration (R_g) for wt MBP, tau, and MBP0 were determined by small angle X-ray scattering (SAXS). Measurements were taken in 0.5 M GdmCl to ensure monodispersity. Values of 69 Å, 62 Å, and 23 Å were extracted for tau, MBP0 and wt MBP, respectively (**Figure 5A**, inset). The Kratky plot (**Figure 5B**) of wt MBP is bell-shaped, consistent with the behavior of a globular, folded protein. The plots of MBP0 and tau, however, lack a bell shape and plateau at high Q values, which is characteristic of unfolded proteins.

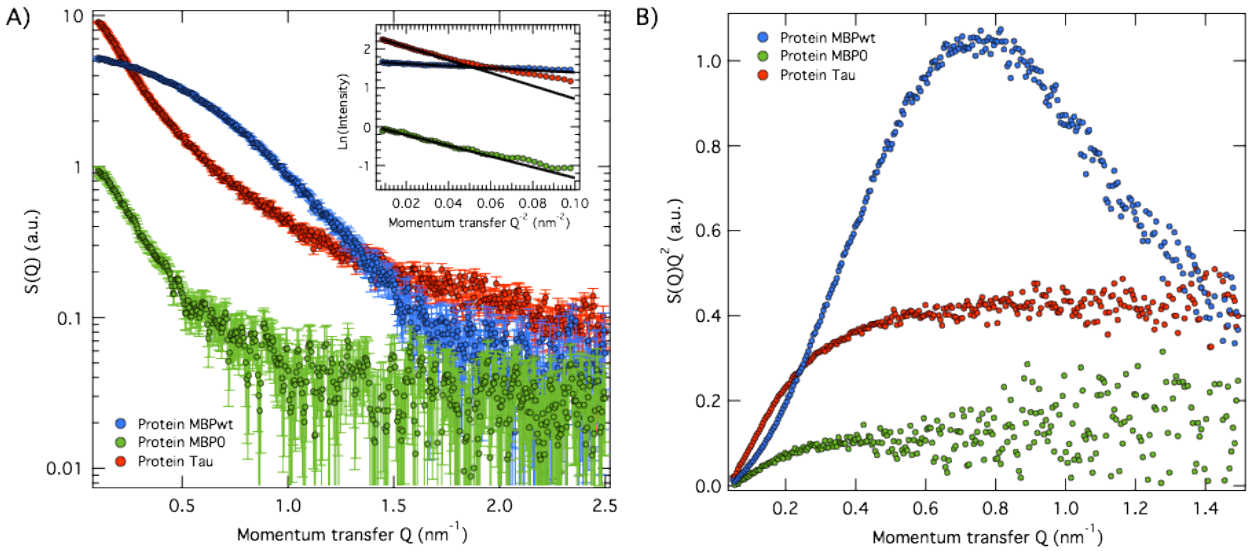


Figure 5. (A) SAXS intensity profiles of wt MBP (blue circles), MBP0 (green circles) and tau (red circles) proteins. Inset: Guinier plot of the intensities. Radii of gyration of 69 Å (tau), 62 Å (MBP0) and 23 Å (wt MBP) were extracted from linear fits (dark lines) in the range 0.01 - 0.02 nm^{-2} for tau and MBP0, and 0.01 - 0.36 nm^{-2} for wt MBP. **(B)** Kratky plots of wt MBP (blue circles), MBP0 (green circles) and tau (red circles). *Data from FX Gallat.*

3.3.5 Neutron scattering of soluble wt MBP, tau protein and MBP0

The MSDs of MBP0 are higher than those of wt MBP and tau protein at temperatures above 260 K (**Figure 6**). The similarities in the measured radii of gyration for tau and MBP0 indicate that they should also experience similar confinement effects. Thus, the observed difference in dynamics between tau protein and MBP0 reflect the difference in each protein's disorder and dynamics. Extracted pseudo-force constants, *k'*, describe the system's resilience to changing temperature (**Table 1**). MBP0 has a four-fold smaller *k'* than tau protein, corresponding to a lower resilience.

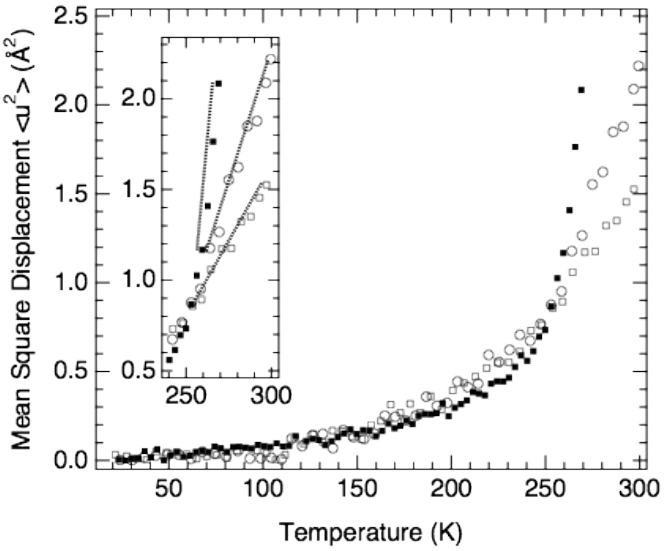


Figure 6. MSDs of tau (empty circles), wt MBP (empty squares) and MBP0 (filled squares) extracted from EINS experiments using the backscattering spectrometer IN16 at 0.9 μeV resolution. MSDs of wt MBP and tau were taken from [1] and [2], respectively. MSDs of MBP0 were extracted in the range 0.2 - 1.40 Å^{-2}. (Inset) MSDs at high temperatures with apparent force constant fits (solid lines).

Table 1. Apparent force constant extracted from the mean squares displacements for T > 260 K

	tau protein	wt MBP	$MBP0_{soluble}$
Apparent force constant, *k'*	0.096 N/m	0.185 N/m	0.024 N/m

Data from FX Gallat.

3.3.6 Neutron scattering of aggregated wt MBP and MBP0

To ensure the enhanced dynamics observed for MBP0 were not related to aggregation during sample drying, an intentionally aggregated sample was also studied by EINS (**Figure 7**). Sample prepared from the aggregated form of MBP0 showed distinct dynamics, validating the previous comparisons between soluble MBP0 and tau protein. At high temperatures, aggregated MBP0 is less dynamic than soluble MBP0, as might be expected. An aggregated wt MBP sample was also prepared (see section 3.4.4). Curiously, aggregated wt MBP shows increased dynamics relative to its soluble form. Furthermore, both aggregated wt MBP and aggregated MBP0 have equivalent MSDs over the whole temperature range. This surprising result remains to be explored and is beyond the main scope of this study.

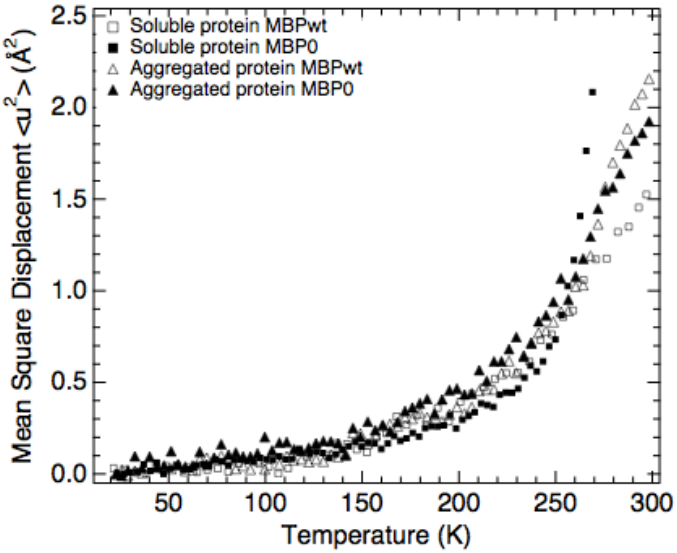


Figure 7. MSDs of soluble wt MBP (open squares), aggregated wt MBP (open triangles), soluble MBP0 (solid squares), and aggregated MBP0 (solid triangles) extracted from EINS experiments. Data for soluble wt MBP are taken from [1].

3.4 Materials and Methods

3.4.1 Expression and purification of wt MBP

A plasmid containing the wt MBP gene (without its leader sequence) was kindly provided by C. Park (Purdue University). The protein was expressed in *E. coli* BL21(DE3)pLysS under the control of the T7 promoter and purified using a Q column at pH 7 (GE Healthcare). The protein's purity and molecular weight were confirmed by SDS-PAGE and electrospray mass spectrometry.

3.4.2 Design, expression and purification of MBP0

MBP0 was constructed by introducing five amino acid substitutions, I59A/L115A/L147A/I161A/I226A, to a wt MBP-containing plasmid via multi-site QuikChange (Agilent) mutagenesis. Successful variants were verified by DNA sequencing and then transformed into BL21(DE3)pLysS cells for expression. Cells were induced with 1mM IPTG at OD = 0.6 and grown at 37º C for 3 hours before harvesting. Cells were lysed in buffer via sonication; then inclusion bodies were isolated and washed with non-ionic detergent. Inclusion bodies were solubilized in 6 M urea, and the protein purified using a Q column in 6 M urea at pH 7. Protein was diluted and dialyzed against either 100 mM ammonium bicarbonate for subsequent freeze-drying and storage) or 100 mM NaCl, 10 mM Tris-HCl buffer (pH 8) for immediate use. The protein's purity and molecular weight were confirmed by SDS-PAGE and electrospray mass spectrometry.

3.4.3 Circular dichroism spectroscopy

CD measurements were collected on an AVIV 410 spectrophotometer. Spectra of MBP0 before and after the addition of 100 mM maltose were taken in a 0.1-cm pathlength quartz cuvette at 25 °C containing 0.48 mg/mL of protein in 100 mM NaCl, 10 mM Tris-HCl buffer (pH 8). Data were collected from 250-200 nm at 1-nm intervals, and each data point represents signal averaged over 5 seconds. Only data at wavelengths above the point where the dynode voltage was below 500 V were used.

Maltose-induced folding kinetics were followed by measuring the CD signal at 222 nm averaged over 60 seconds. The data were fit to a single exponential (k_{obs}) using KaleidaGraph (Version 4.1 by Synergy Software):

$$\text{CD signal} = \text{A} * \exp(-k_{obs} * t) + C \quad (1)$$

where C is the final signal, A is the amplitude of the observable phase, and t is time.

Urea denaturation of MBP0 was performed in a 0.1-cm pathlength quartz cuvette at 25 °C by monitoring the CD signal at 222 nm and averaging the signal over 60 seconds for each data point. Individual samples containing 0.48 mg/mL of protein in 100 mM NaCl, 10 mM Tris-HCl buffer (pH 8) and varying concentrations of urea were equilibrated at 25 °C overnight. The urea denaturation melt of wt MBP was performed in a 1-cm pathlength cuvette containing 48 μg/mL protein in the same buffer conditions. Urea concentrations were verified using a refractometer [15].

3.4.4 Size exclusion chromatography

3.4.4.1 Soluble sample preparation

Lyophilized MBP0 powder was resuspended in D_2O at 0.1 mg/ml concentration. No buffer was used to resuspend in order to avoid salt in the final sample. This step the labile D/H to exchange before lyophilization. The solution was then filtered through a 0.22 μm pore size membrane to remove possible aggregates. The protein solution was then flash frozen in liquid nitrogen and lyophilized.

3.4.4.2 Aggregated sample preparation

Lyophilized MBP0 powder was resuspended in D_2O at 5 mg/ml concentration to ensure complete aggregation. Again, no buffer was used. The solution was then concentrated by solvent evaporation until complete protein drying, which took overnight.

Aggregation states of the protein samples (soluble and aggregated solutions) were verified by size exclusion chromatography, using a Superdex S200 (Amersham) equilibrated in H_2O. Model proteins standards (**Table 2**) were injected prior to soluble and aggregated MBP0 solutions in order to calibrate the column and extract information on the hydrodynamic radius of MBP0. The dead volume of the column was determined by an injection of blue dextran, a glucose-derived polymer with averaged molecular weight of 2 MDa, which elutes in the dead volume of the column.

Table 2. Molecular masses and hydrodynamics radii of reference proteins

Protein	Molecular mass (kDa)	Hydrodynamic radius (nm)
Thyroglobulin	669	7.85
Ferritin	440	6.80
Aldolase	158	4.65
Albumin	67	3.37
Ovalbumin	43	3.00
Chymotrypsinogen	25	2.10
RNase A	14	1.64

Data from FX Gallat.

3.4.5 Small angle X-ray scattering

Small angle X-ray scattering (SAXS) measurements on tau, wt MBP and soluble MBP0 were recorded using the ID14-3 BioSAXS beamline at the European Synchrotron Radiation Facility (ESRF Grenoble, France) in 10 mM Tris pH 8.0, 100 mM NaCl, 500 mM GmdCl (guanidinium chloride). The absence of radiation damage was verified by 10 successive exposures of 10 seconds each. Radii of gyration, R_g, were extracted from a Guinier plot, which plots the logarithm of the intensities S(Q) against Q^2, in the range 0.01 – 0.02 nm^{-2} for MBP0 and tau, and 0.01- 0.36 nm^{-2} for wt MBP, in such a way that $Q_{max}.R_g < 1$ for MBP0 and tau, and $Q_{max}.R_g < 1.3$ for wt MBP. Kratky plots were obtained by plotting the quantities $S(Q)Q^2$ against Q. The protein concentrations used were 2.0 mg/mL for tau, 0.89 mg/mL for wt MBP, and 0.2 mg/mL for MBP0.

3.4.6 Elastic incoherent neutron scattering

To prepare soluble protein samples, protein powder was resuspended in D_2O at 0.1 mg/ml. No buffer was used to resuspend the protein in order to avoid salt in the final sample. This step allows exchange of labile hydrogens, most of which come from amide protons on the protein backbone. The solution was then filtered using a 0.22 μm pore size membrane to remove possible aggregates. The protein solution was then freeze-dried by flash freezing in liquid nitrogen and then lyophilizing.

To prepare aggregated protein samples, protein powder was resuspended in D_2O at 5 mg/ml to ensure complete aggregation. The solution was then concentrated by solvent evaporation until complete protein drying was achieved.

The D/H-exchanged protein powders were dried over P_2O_5 for two days on a 4 x 3 cm^2 flat aluminum sample holder. The resulting hydration level was defined as corresponding to 0 g water/g protein. Powders were then rehydrated over pure D_2O to a hydration level of 0.44 g D_2O/g protein. This hydration level has been shown to correspond, at least for globular proteins, to full coverage by a monolayer of water. Previous experiments on the intrinsically disordered protein tau indicate that this hydration level is also appropriate for studying disordered proteins [2]. Dynamics of these samples were measured on the IN16 backscattering spectrometer (Institut Laue-Langevin) with a resolution of 0.9 μeV, which is associated with motions on the nanosecond timescale. Elastic intensities were recorded while the temperature was continuously increased from 20 to 300 K. Atomic mean square displacements (MSD, $\langle u^2 \rangle$) were extracted from the Q-dependence of the elastic intensity, which can be described in the Gaussian approximation by:

$$I\,(Q, \boldsymbol{\omega} = 0) = \mathrm{I}_0 * \exp\left(-\frac{1}{6} * < u^2 > * Q^2\right) \qquad (2)$$

where $I\,(Q, \boldsymbol{\omega} = 0)$ is the elastically scattered intensity and I_0 is the value of the scattering at Q = 0. This expression remains valid as long as $(Q^2 * < u^2 >) \leq 2$. MSDs were extracted in the range $0.2 < Q^2 < 1.40$ $Å^{-2}$ for all samples. Apparent force constants, $< k' >$, were extracted in the high temperature region, according the following relation:

$$< k' > = 2\mathrm{k}_B\mathrm{T} / \left(\frac{d<u^2>}{dT}\right) \qquad (3)$$

3.5 Discussion

3.5.1 Enhanced sidechain flexibility of unfolded MBP

A variant of MBP with five site-specific mutations (MBP0) populates the unfolded state under native conditions. MBP0, shows increased sidechain flexibility relative to an intrinsically disordered protein, tau. At temperatures above 260 K, the neutron diffraction mean square displacements for MBP0 increase with temperature more dramatically than those of tau. This is reflected in MBP0's effective force constant, *k'*, which is fourfold smaller than tau's and indicates its reduced structural resilience over this temperature range.
Previous work comparing tau protein with several folded proteins demonstrated that intrinsic disorder is associated with greater flexibility [2]. It was unknown, however, if the difference in dynamics was due to the unique amino acid composition of tau relative to the folded, globular model proteins or increased conformation freedom of the entire polypeptide chain. Structural analysis by NMR indicates that tau exists as a conformational ensemble [11], but intrinsically disordered proteins are also known to contain an unusually high ratio of polar to hydrophobic residues [10]. MSDs reflect both the dynamics intrinsic to each sidechain and overall dynamics of the backbone, which propagate to increase sidechain flexibility [9]. Our data from MBP0, however, suggest it is the latter effect that dominates. MBP0 differs from wt MBP, our folded protein model, by five conservative amino acid substitutions. By replacing three isoleucines and two leucines with alanines, we effectively populate the unfolded state without radically changing the molecule's intrinsic sidechain dynamics. In fact, MBP0 has fewer hydrogens than wt MBP, which might be expected to effect average MSDs in the opposite direction than is observed.

Unfolded states of proteins also exist as conformational ensembles. From this perspective, it is perhaps not surprising that MBP0 is more dynamic than wt MBP; however previous neutron scattering experiments with alkali-denatured lysozyme [16] and a partially folded mutant staphylococcal nuclease (SNase) [17] showed no difference in MSDs between the unfolded and folded proteins. These studies conclude that dynamics probed by EINS are insensitive to the conformational state of the protein. Our seemingly contradictory conclusions might be reconciled by considering difference in compactness. MBP0 is very expanded with an R_g more than twice that of wt MBP (62 Å *versus* 23 Å). The truncated mutant of SNase, however, is more compact and has a similar R_g to the full-length, folded protein (21 Å *versus* 16 Å) [18]. Thus, it is possible we are studying a fundamentally different unfolded state than previous studies and/or that enhanced dynamics of the unfolded state are only obvious when it is also sufficiently extended.

What is less clear is why the unfolded state would be more dynamic than an intrinsically disordered protein (IDP). MBP0 and tau protein are similarly expanded with R_g values in 0.5M GdmCl of 62 Å and 69 Å, respectively. Therefore, the observed difference in MSDs is either due to differences in amino acid content, conformational flexibility of the backbone or some combination. One potential explanation could be that sidechain dynamics are finely tuned for biological activity, and while IDPs are functional, the unfolded states of natively folded proteins are not. For instance, it is thought that tau's conformational ensemble is relevant for function, because the protein binds microtubules in multiple partially folded forms. Thus, tau's landscape is broader compared to a natively folded protein like MBP. The energetic landscape for unfolded state of MBP, on the other hand, appears to be even flatter. Such extreme heterogeneity could

prove deleterious, for instance by facilitating the non-specific interactions leading to aggregation. Tau protein is disordered but can be concentrated nearly a hundred fold more than MBP0 without visibly precipitating (data not shown). Thus the unfolded state is important for modulating the stability of other functional states on the landscape, but it itself is likely not involved in functional interactions that might require finely tuned dynamics.

3.5.2 Similar sidechain flexibility of aggregates

The MSDs for aggregated MBP0 and aggregated wt MBP are equivalent over the entire temperature range measured, indicating that the systems share similar sidechain dynamics. This result suggests that, from this perspective, the aggregated state is the same for both proteins, which we might expect given the subtle difference in sequence. Soluble MBP0 is more dynamic than the aggregates and soluble wt MBP is less, so the convergent behavior is a result of opposite dynamical changes to the proteins upon aggregation. The change in MBP0's behavior can be rationalized by reduced freedom of motion in the aggregated state. It is more difficult to rationalize the increased dynamics observed in wt MBP aggregates. This phenomenon merits further investigation, however, especially given that the process of soluble, folded proteins forming aggregates is implicated in the progression of many human diseases.

3.5.3 Conclusions and next steps

Based on MSDs from EINS experiments, unfolded MBP is more dynamic than tau protein, which is more dynamic than folded MBP. Taken together, these results suggest that MSDs provide a reasonable measurement for conformational flexibility, even though the scattered neutrons are directly probing sidechain-hydrogen dynamics.

Studies of additional unfolded proteins and IDPs are needed to adequately interpret the observation that unfolded MBP is more dynamic than tau protein. It would interesting, for instance, to consider whether IDPs from other functional classes share this behavior or whether it is optimized to accommodate tau's multiple binding modes. Also, further studies are needed to illuminate the role of water in facilitating MBP0's enhanced dynamics and to address whether unfolded states are as intimately coupled to hydration-water dynamics as has been observed for tau protein.

3.6 References

1. Wood, K., et al., *Coincidence of dynamical transitions in a soluble protein and its hydration water: direct measurements by neutron scattering and MD simulations.* J Am Chem Soc, 2008. **130**(14): p. 4586-7.
2. Gallat, F.X., et al., *Dynamical coupling of intrinsically disordered proteins and their hydration water: comparison with folded soluble and membrane proteins.* Biophys J, 2012. **103**(1): p. 129-36.
3. Wood, K., et al., *Coupling of protein and hydration-water dynamics in biological membranes.* Proc Natl Acad Sci U S A, 2007. **104**(46): p. 18049-54.
4. Gabel, F., et al., *Protein dynamics studied by neutron scattering.* Q Rev Biophys, 2002. **35**(4): p. 327-67.
5. Daniel, R.M., et al., *Enzyme activity below the dynamical transition at 220 K.* Biophys J, 1998. **75**(5): p. 2504-7.
6. Rasmussen, B.F., et al., *Crystalline ribonuclease A loses function below the dynamical transition at 220 K.* Nature, 1992. **357**(6377): p. 423-4.
7. Wood, K., et al., *A benchmark for protein dynamics: Ribonuclease A measured by neutron scattering in a large wavevector-energy transfer range.* Chem Phys, 2008. **345**(2–3): p. 305-314.
8. Ball, P., *Water as an active constituent in cell biology.* Chem Rev, 2008. **108**(1): p. 74-108.
9. Schiro, G., et al., *Direct evidence of the amino acid side chain and backbone contributions to protein anharmonicity.* J Am Chem Soc, 2010. **132**(4): p. 1371-6.
10. Tompa, P., *Intrinsically disordered proteins: a 10-year recap.* Trends Biochem Sci, 2012.
11. Mukrasch, M.D., et al., *Structural polymorphism of 441-residue tau at single residue resolution.* PLoS Biol, 2009. **7**(2): p. e34.
12. Barre, P. and D. Eliezer, *Folding of the repeat domain of tau upon binding to lipid surfaces.* J Mol Biol, 2006. **362**(2): p. 312-26.
13. Chang, Y. and C. Park, *Mapping transient partial unfolding by protein engineering and native-state proteolysis.* J Mol Biol, 2009. **393**(2): p. 543-56.
14. Sharff, A.J., et al., *Crystallographic evidence of a large ligand-induced hinge-twist motion between the two domains of the maltodextrin binding protein involved in active transport and chemotaxis.* Biochemistry, 1992. **31**(44): p. 10657-63.
15. Warren, J.R. and J.A. Gordon, *On the refractive indices of aqueous solutions of urea.* The Journal of Physical Chemistry, 1966. **70**(1): p. 297-300.
16. Mamontov, E., H. O'Neill, and Q. Zhang, *Mean-squared atomic displacements in hydrated lysozyme, native and denatured.* J Biol Phys, 2010. **36**(3): p. 291-7.
17. Nakagawa, H., H. Kamikubo, and M. Kataoka, *Effect of conformational states on protein dynamical transition.* Biochim Biophys Acta, 2010. **1804**(1): p. 27-33.
18. Flanagan, J.M., et al., *Truncated staphylococcal nuclease is compact but disordered.* Proc Natl Acad Sci U S A, 1992. **89**(2): p. 748-52.

CHAPTER 4

Using the design principle of mutually exclusive folding to introduce novel allosteric control of enzymatic activity

Work initiated by Tracy Young

4.1 Abstract

Using the principle of mutually exclusive folding, we have successfully engineered a modular ligand-based allosteric switch. The switch is a chimeric fusion of two proteins: a ligand binding protein, the switch's regulatory domain; and one of two model enzymes, staphylococcal nuclease and ribonuclease H. The domains are fused such that only one domain can fold into its native conformation at any given time. Discrepancy in end-to-end distance at the attachment point results in one domain's folding geometrically precluding the other domain's folding. Thermodynamics governs which protein is folded under any specific conditions. This gross allostery is regulated by the addition of ligand, which preferentially stabilizes the state containing folded regulatory domain and unfolded, inactive enzyme.

4.2 Introduction

Naturally occurring proteins are capable of executing complex functions that are difficult to design *de novo*. In addition to catalyzing reactions up to 10^{17}-fold faster than the uncatalyzed rate [1], proteins act as sensors, allowing the cell to respond to environmental cues, and participate in the elaborate regulatory networks involved in orchestrating such responses. For the last fifty years, biologists have largely focused on elucidating underlying mechanisms, but the emerging field of synthetic biology seeks to capitalize on this vast knowledge base by engineering novel cellular functions. The ability to design novel protein switches, in particular, is central to synthetic biology's goal of "hacking" and rewiring cellular machinery. For instance, turning microbes into efficient "chemical factories" that churn out desirable chemicals, such as pharmaceuticals or energy-dense hydrocarbons, is often hindered by the production of cytotoxic intermediates in the synthetic pathway. If, however, a critical enzyme could be turned ON or OFF at specific times during growth, accumulation of toxic compounds might be minimized.

One way to construct protein switches is through engineering allostery. Allostery is a form of regulation where binding at one site within a protein's structure affects change at a distal site. Typically this change manifests as modulated enzymatic activity or binding affinity. Coupling between the distant sites is mediated by conformational and/or energetic modulation. Some evidence suggests a physically connected network of interactions is needed to facilitate communication between sites [2], but other studies suggest energetic perturbations of a protein's conformational ensemble is sufficient for allostery to occur [3]. As an extension of the latter view, it has been proposed that maximal coupling between sites is achieved via the disorder-to-order transitions characteristic of intrinsically disordered proteins [4]. Mutually exclusive folding, wherein only one domain of a two-domain protein can be folded at a time, represents one generalizable strategy for designing an enzyme system that maximizes its allosteric response to ligand.

Radley *et al*. first demonstrated the principal of mutually exclusive folding by creating a chimeric fusion of two single domain proteins, ubiquitin and barnase [5]. In this construct, ubiquitin was inserted into a surface loop of barnase. Because ubiquitin has a longer end-to-end distance in its folded structure than is permitted within the context of barnase's loop, the two domains are sterically prohibited from being folded simultaneously. The resulting thermodynamic tug-of-war means that the most stable domain folds, causing the other domain to

be unfolded and, thus, inactive. In order to switch between folded states, domain stability can be modulated through mutation or, more practically, addition of ligand.

The goal of this project is to determine whether mutually exclusive folding represents a general strategy for engineering ON/OFF enzyme regulation. In our switch design, we develop maltose-binding protein (MBP) as a regulatory platform into which an enzyme of choice can be inserted. MBP is an ideal system for a number of reasons. First, it has successfully been exploited for other types of protein design, resulting in variants that bind diverse, unnatural ligands and with tunable affinities. Second, at equilibrium it exists in only two states, folded and unfolded, which is a necessary component of our design strategy. Lastly, MBP is quite stable, but many destabilizing mutations have been identified, allowing us to tune the stability of this domain to create a modular allosteric switch able to accommodate enzymes with various stabilities. Addition of maltose to the system will preferentially stabilize the MBP domain, leading to enzyme unfolding and inactivation.

Here we describe construction and characterization of MBP fusions with one of two model enzymes: staphylococcal nuclease (SNase) or cysteine-free *E. coli* ribonuclease H (RNase H*). Optimizing the location for insertion and individual domain stabilities are discussed at length (the details of some of these variants are unlikely to be published or documented elsewhere). In the end, only two switches out of the 28 tested were found to be inhibited by maltose. In the successful designs, variants of RNase H* replace a β-hairpin loop in MBP's structure. While the enzymes in these switches do demonstrate maltose-induced inhibition, it remains unclear whether they function using the intended mechanism of mutually exclusive folding. Furthermore, our lack of success with other switches suggests that our guiding principles for rational design need improvement.

4.3 Results

4.3.1 Design of mutually exclusive folding allosteric switches

A mutually exclusive folding switch is designed by fusing two proteins in such that only one can fold into its native conformation at a time (**Figure 1**). In the design, a protein with a long inter-termini distance is inserted into the sequence of another "host" protein, which sterically prohibits it from folding unless the host protein itself unfolds. At equilibrium, the more stable domain will be folded and the less stable domain will be unfolded. This gross allostery can be regulated with ligand, which will selectively bind to and stabilize only one of the domains. If the more stable domain is also the one that binds ligand, then no switching will occur. Therefore, it is essential that the less stable domain be the one that binds ligand. We refer to this protein as the regulatory domain. Furthermore, the ligand-induced stabilization must be sufficient to overcome the intrinsic stability of the other domain, which we refer to as the active domain. When these criteria are met, ligand will alter the distribution of accessible states such that most stable conformation switches between the active being folded and the regulatory domain being folded (**Figure 2**).

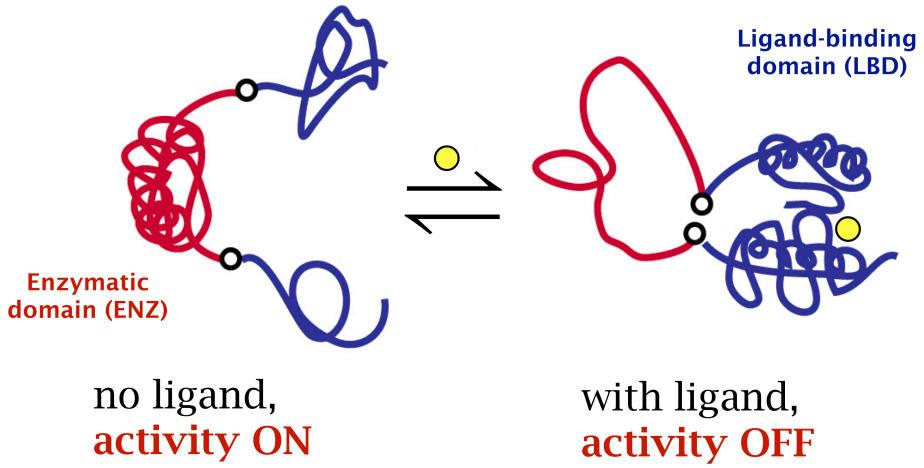


Figure 1. Mutually-exclusive folding switch. Only one domain can be folded at a time due to steric constraints, and ligand controls which state is preferentially populated. In the absence of ligand, the enzymatic domain is more stable, but in the presence of ligand, the ligand-binding domain is more stable.

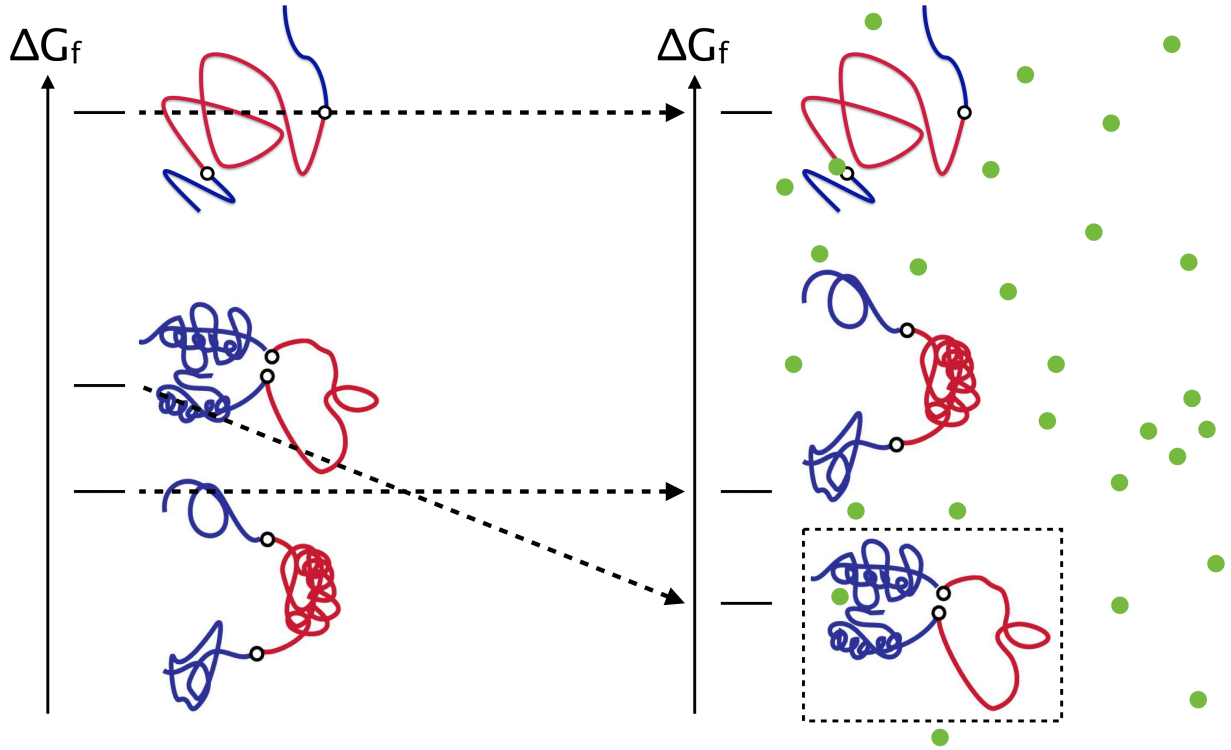


Figure 2. Boltzmann diagrams in the absence (left) and presence (right) of ligand. Selective stabilization leads to switching behavior, because the more stable, and therefore the most populate, state differs under the two sets of conditions.

Our guiding principles in the design of mutually exclusive folding allosteric switches are as follows:

The regulatory domain must:

1. be derived from a ligand-binding protein.
2. demonstrate two-state folding.
3. be functionally tolerant of a large insertion.

The active domain must

1. be derived from an enzyme.
2. demonstrate two-state folding.
3. have a large end-to-end distance in its folded state.
4. have a global stability that falls in between that of the regulatory domain with and without its ligand (**Figure 2**).

MBP satisfies all requirements for the regulatory domain. It binds maltose and other ligands [6], folds and unfolds in a two-state manner [7] and can tolerate amino acid insertions in several locations [8]. MBP binds maltose with K_d = 1 μM, which means 10 mM maltose will cause a stabilization of 5.5 kcal/mol, according to the relationship:

$$K_{f,app} = K_f \left(\frac{1+[\mathrm{L}]}{K_d}\right) \quad (1)$$

where $K_{f,\,app}$ is the apparent equilibrium constant for folding in the presence of ligand, K_f is the equilibrium constant for folding in the absence of ligand, and ligand is assumed to bind only the native state. The equilibrium constants relate to free energies by:

$$\Delta G_f = RT\ln K_f \quad (2)$$

where ΔG_f is the folding free energy, T is temperature and R is the universal gas constant.

Both staphylococcal nuclease (SNase) and the cysteine-free variant of *E. coli* ribonuclease H (RNase H*) satisfy the requirements for the active domain. They are enzymes that cleave single-stranded nucleic acids [9] and RNA-DNA hybrids [10], respectively. At equilibrium, they both show two-state folding [11, 12], and, based on crystal structures, have inter-termini distances of 37 Å and 40 Å, respectively [13-15].

Two locations within MBP's sequence were chosen for domain insertion. Both were chosen based on studies demonstrating their functional tolerance to amino acid insertion [8] and the ability to be reconstituted from the two fragments generated from cleavage at these locations [16]. One location is between residues 286 and 287 in a loop region of MBP. In the second location, the active domain replaces residues 169-181, which comprise a β-hairpin that spans 4 Å in MBP.

For clarity, results for model proteins meant to represent domains within the context of the chimeras are presented first in section 4.3.2. Then, results for all MBP-SNase chimeras are

presented in section 4.3.3. Lastly, results for all MBP-RNase H* chimeras are presented in section 4.3.4.

4.3.2 Models for the regulatory and active domains

4.3.2.1 Circular dichroism of model domains

4.3.2.1.1 Active domains

One of our design criteria is that the active domain has a global stability that falls in between that of the regulatory domain with and without its ligand. To aid in the construction of chimeras that satisfy our stability requirements, models for the regulatory and active domains were made and studied. Stabilities were either extracted from circular dichroism (CD) experiments measuring urea-dependent signals at 222 nm (**Figures 3** and **4**) or calculated based on published $\Delta\Delta G$ values for full-length variants of MBP, SNase and RNase H* are collected in Table 1.

Table 1. Measured and calculated stabilities for model switch domains

MBPs	ΔG_f (kcal mol^{-1})	**SNases**	ΔG_f (kcal mol^{-1})	**RNases H***	ΔG_f (kcal mol^{-1})
Wild type	-15.1	Wild type	-6.1	Wild type	-9.4
I226A	-10.7‡	P117G/H124 L/S128A	-5.6 [17]-8.9	I53D	-5.6 [17]
I226A/I161A	-7.7‡	H124L	-4.6 [18]-6.7	I25A	-4.6 [18]
I226A/L147A/ L115A	-4.8‡	I18G	-3.6†		
I226A/L147A/ I108A	-4.6‡	T62G	-2.6†		
I226A/I161A/ L147A/I108A	-1.8‡	G107A	-1.6†		
		V23G	-0.5†		
		H124L/V66K	-0.5†		
		L103G	+0.5†		

‡ Derived from $\Delta\Delta G$ values published in [19] and assuming additivity.
† Derived from $\Delta\Delta G$ values published in [20-22] and assuming additivity.

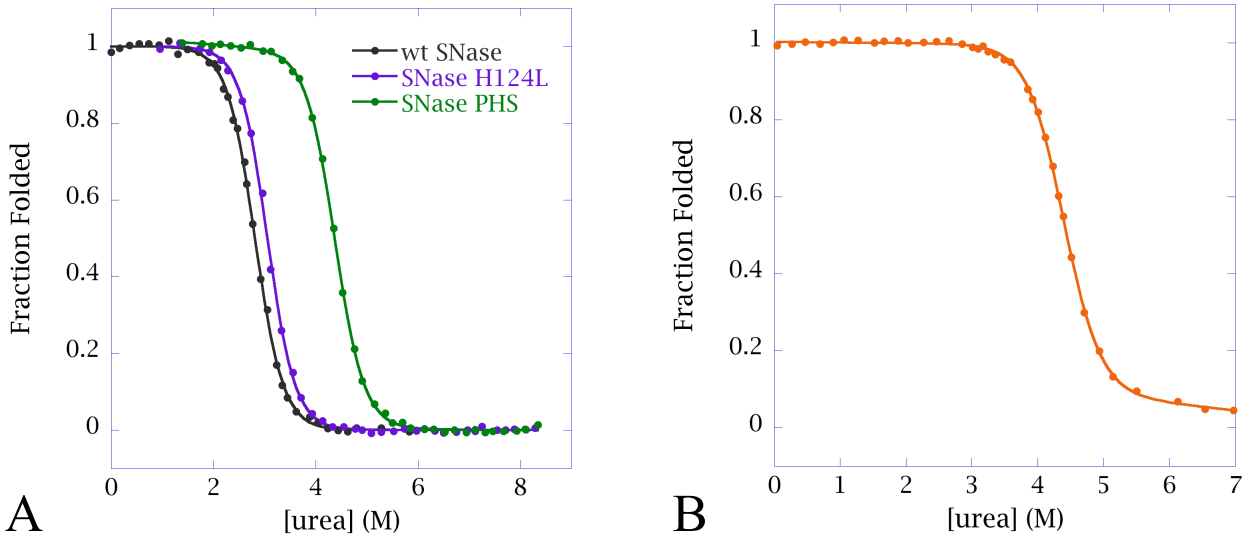


Figure 3. (A) Equilibrium urea denaturation melts of wt SNase (black), SNase H124L (purple) and SNase P117G/H124L/S128A (green) in 300 mM NaCl, 50 mM Tris-HCl (pH 7.5), 10 mM $CaCl_2$ and 100 mM maltose. (B) Equilibrium urea denaturation melt of RNase H* (orange) in 300 mM NaCl, 50 mM Tris-HCl (pH 7.5), 10 mM $MgCl_2$ and 100 mM maltose. Stabilities extracted from fits are described in Table 1.

4.3.2.1.2 Regulatory domains

Two models for the MBP domain in the context of the chimeras were constructed. The first MBP variant with deleted residues 169-180, referred to $MBP^{\Delta\beta}$, was constructed as a model for chimeras where the inserted enzyme domain replaces the β-hairpin. Data collected by Tracy Young (data not shown) reveal it is destabilized relative to wt MBP by 3 kcal/mol (**Table 2**).

The second MBP variant contains a seven amino acid insertion between residues 286 and 287 and serves as a model for chimeras with domain insertions at that same location. The variant, referred to here at MBP^{+7}, is destabilized relative to wild-type MBP (wt MBP) as evidenced by its lower C_m (**Figure 4**). It is difficult to extract an exact stability for MBP^{+7}, because its low *m*-value is highly suggestive of non-two-state behavior [23]. When the *m*-value is fixed to that of wt MBP, the stability of MBP^{+7} is 7.5 kcal/mol (**Table 2**). The addition of 10 mM maltose results in a stabilization by 6.2 kcal/mol (*m*-value = 6.1 kcal/mol·M), which is consistent with the value predicted based on K_d = 1 μM (**Equation 1**).

Table 2. Measured stabilities of insertion/deletion models for the regulatory domain.

	ΔG_f (kcal mol^{-1})	*m*-value (kcal mol^{-1} M^{-1})
wt MBP	-15.1 ± 0.5*	5.1 ± 0.1*
MBP^{+7}	-7.3†	ND
MBP^{+7} with 10 mM maltose	-13.5	6.1
$MBP^{\Delta\beta}$	-11.9‡	ND
$MBP^{\Delta\beta}$ with 10 mM maltose	-19.8‡	ND

* Values represent averages and standard deviations from three independent experiments.
† Stability derived from data fit with an *m*-value fixed to 5.1 kcal/mol*M (fit not shown).
‡ Data from Tracy Young.

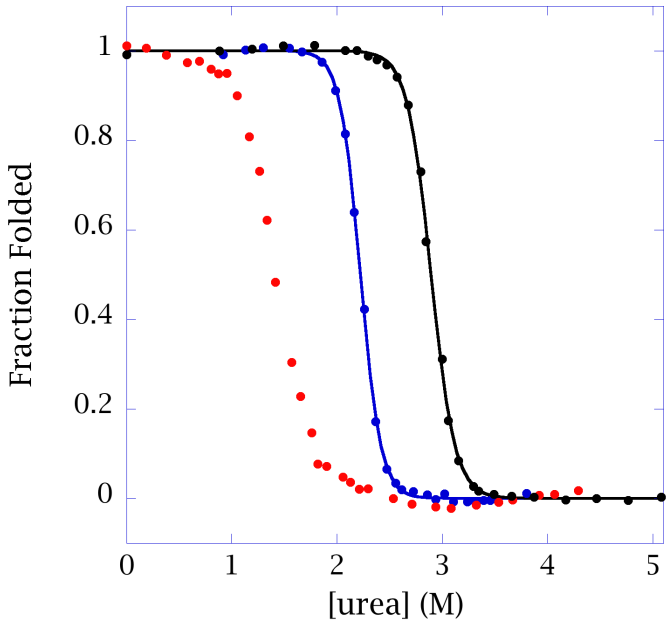


Figure 4. Equilibrium urea denaturation melts of wt MBP (black), MBP^{+7} (blue) and MBP^{+7} with 10 mM maltose (red). Stabilities extracted from fits are described in Table 1.

4.3.2.2 Activity of model domains

4.3.2.2.1 SNase activity

SNase hydrolyzes 5'-phosphodiester bonds of single-stranded DNA in a Ca^{2+}-dependent manner. It exhibits both endo- and exonuclease actvities, though the major products are mono- and dinucleotides [9]. Activity is monitored using the hyperchromic effect, as liberated bases absorb more strongly at 260 nm than single-stranded DNA. To prepare substrate, salmon sperm DNA is sonicated, boiled and quenched on ice. Substrate concentrations are given in molarity of internucleotide bonds and assume complete denaturation. Michaelis-Menten analysis of wt SNase finds $k_{cat} = 2 \times 10^4\ s^{-1}$, $K_m = 83\ \mu M$, in agreement with published values [24] (**Figure 5**).

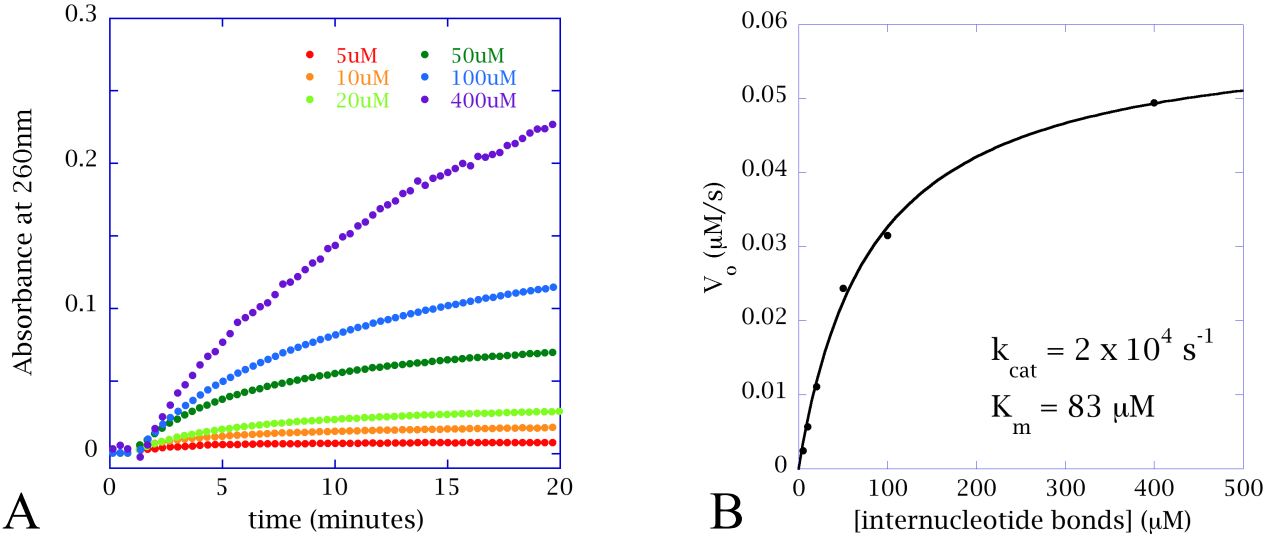


Figure 5. (A) Activity traces for 3 nM wild-type SNase with variable substrate concentrations. (B) Michaelis-Menten analysis of wild-type SNase.

To facilitate higher throughput analysis, an agar plate-based assay was also used to assess activity in the chimeras. Substrate and the pH indicator toluidine blue are dissolved in an agar matrix, and either purified protein or cell lysate can be spotted on the agar [25, 26]. After developing at 37 °C for several hours, a pink halo appears in response to SNase activity. This is due to the release of protons concomitant with DNA cleavage, which causes a local drop in pH. Though less quantitative then the spectroscopic assay, the blue-plate assay allows many constructs to be studied simultaneously without the need for protein purification. The spot size depends on the amount of protein spotted, the length and temperature of the experiment and the activity of the construct. SNase variants produce distinct pink halos after 3 hours with 0.2 pmol enzyme (**Figure 6**), and there is little variation observed between wt SNase and the stabilized variant SNase PHS (data not shown). Chimeras were less active, and 10 pmol was required to visualize activity (**Figures 13** and **17**). To test the effects of maltose on activity, the same amount of protein was co-spotted on two plates, one of which contained maltose, and the spot size was compared after equivalent time points. Successful switches should have smaller spots on the maltose-containing plates.

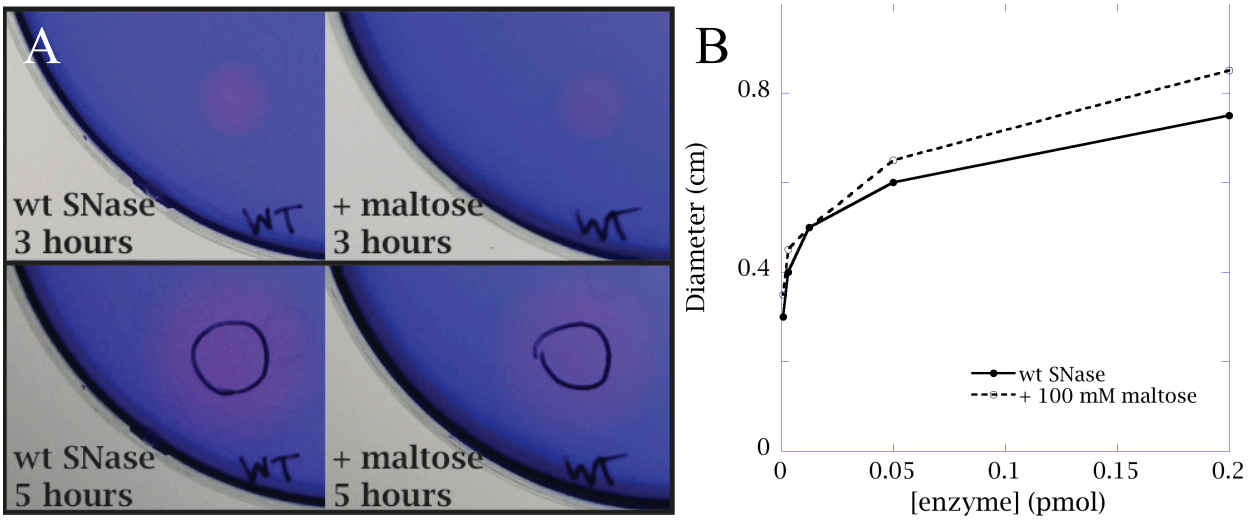


Figure 6. (A) 0.2 pmol wt SNase spotted on TB-agar substrate after 3 and 5 hours, with 100 mM maltose and without. Spot size after 3 hours is indicated in the 5-hour images with black circles. (B) Spot diameter after 3 hours as a function of enzyme concentration.

4.3.2.2.2 RNase H activity

RNase H hydrolyzes 3'-phosphodiester bonds of RNA in RNA-DNA hybrids in a Mg^{2+}-dependent manner [10]. Activity is monitored using the hyperchromic effect, as liberated bases absorb more strongly at 260 nm than either single- or double-stranded nucleic acids. Substrate is prepared by annealing dT_{20} oligomers to poly-rA strands. Substrate concentrations assume complete hybridization, but some variability is observed between substrate stocks prepared on different days. Where possible, RNase H* activity measured on the same day with the same substrate stock is shown on the same plot for comparison. RNase H* activity traces appears on the same plots as MBP-RNase H* chimeras in section 4.3.3 for the purposes of direct comparison. Unless otherwise noted, reactions were performed with 5 nM enzyme in 10 mM Tris (pH 8), 50 mM NaCl, 10 mM $MgCl_2$ and either 8.3 µg/mL or 16.7 µg/mL substrate.

4.3.3 MBP-SNase chimeras

4.3.3.1 MBP-SNase chimeras, insertion at residue 286

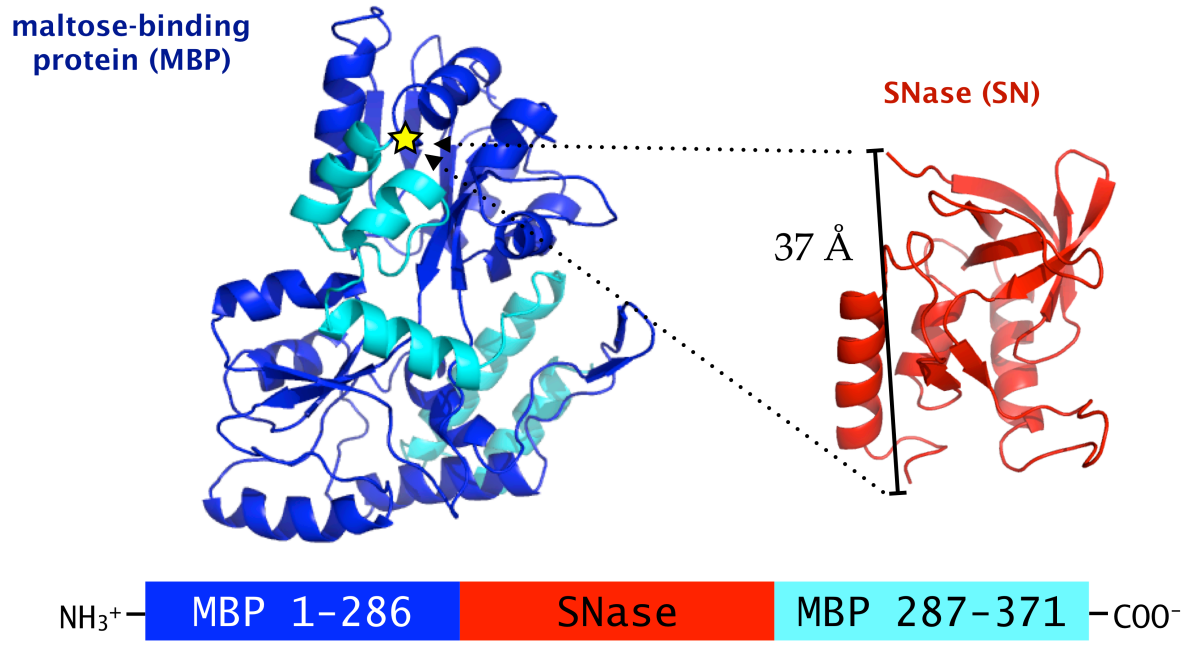


Figure 7. MBP-SNase chimera with insertion between residues 286 and 287. A star indicates the location in MBP's structure where two truncated versions of SNase are inserted to generate MS long and MS short. Below the structures is a schematic of the gene construct.

Two truncated variants of SNase were inserted between residues 286 and 287 of MBP to generate the first class of MBP-SNase chimeras (**Figure 7**). Full-length SNase is 149 residues long, but not all of its N- and C-terminal residues were included in the constructs. Linkers at the site of domain insertion can be used to adjust coupling between domains [27]. Shorter linkers create more tension, thus maximizing structural and energetic coupling. Rather than introducing synthetic linkers in our chimera designs, we opted to treat the N- and C-terminal residues of SNase, the inserted domain, as intrinsic linkers. This is a reasonable approach, because the first six and last eight residues of SNase are not visible in the crystal structure [21], which implies these regions are flexible and unstructured.

The two chimeras in this class are "$MBP_{1\text{-}286}$ – $SN_{10\text{-}136}$ – $MBP_{287\text{-}371}$" (MS short), which has a SNase containing residues 10-136, and "$MBP_{1\text{-}286}$ – $SN_{7\text{-}141}$ – $MBP_{287\text{-}371}$" (MS long), which has a SNase containing residues 7-141. MS long contains only the residues of SNase visible in its crystal structure, while MS short contains a slightly shorter SNase domain that begins and ends with residues involved in secondary structure elements. Results from structural and activity studies are presented first for MS short and then for MS long.

"$MBP_{1\text{-}286}$ – $SN_{10\text{-}136}$ – $MBP_{287\text{-}371}$" (MS short)

Versions of MS short containing SNase domains with varying stabilities were constructed [20]. Arranged from most to least stable, they include: MS short P117G/H124L/S128A (PHS); MS

short H124L; MS short; MS short I18A; MS short T62G; MS short G107A; MS short V23G; MS short V66K/H124L; MS short L103G. Many of these chimeras contain an MBP domain with the substitution I329Y, which increases the affinity for maltose tenfold [6]. These constructs are denoted M_YS short.

The CD spectrum for MS short is consistent with a large portion of the polypeptide being unfolded (**Figure 8**). Upon addition of maltose, a gross conformational change results in a spectrum that more closely resembles a folded protein. This is consistent with our design. In the absence of maltose, we expect the SNase domain to be folded and the MBP domain to be unfolded; however, because MBP is more than twice the size of SNase, the unfolded protein dominates the CD signal, which reflects average properties of the system. When maltose is added, MBP folds, causing SNase to unfold and giving rise to a more folded spectrum. Importantly, the addition of a non-hydrolyzable DNA analog, phosphorothioate (ptDNA), which is expected to bind to the SNase domain, does not impede the switching behavior. This is a non-trivial result, because we expect SNase to be stabilized by substrate, and our switch will not function if maltose binding cannot overcome this additional stabilization.

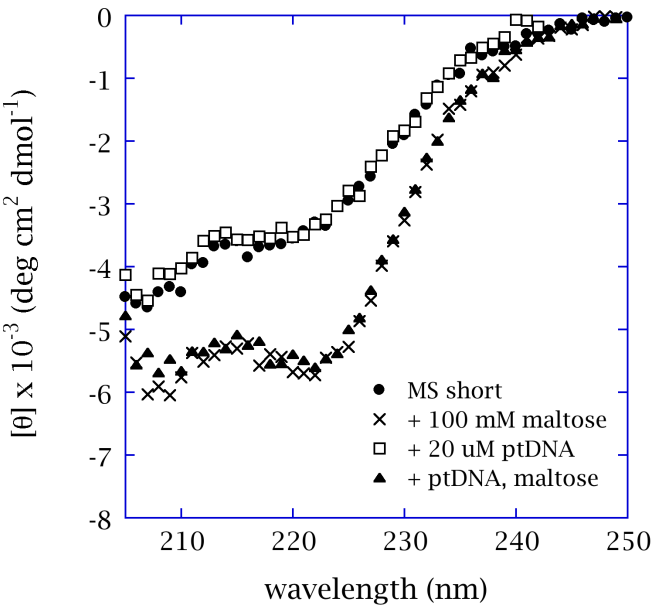


Figure 8. CD spectra of MS short with maltose, non-hydrolyzable DNA analog (ptDNA) or both.

When the SNase domain is stabilized, as with MS short PHS, or destabilized, as with MS short H124L/V66K, the same switching behaviors are observed (**Figures 9** and **10**). In the destabilized variant, however, the conformational change is less dramatic. This is consistent with the design, as we expect MBP to be more folded in the absence of maltose when SNase is more destabilized and the free energy gap is larger.

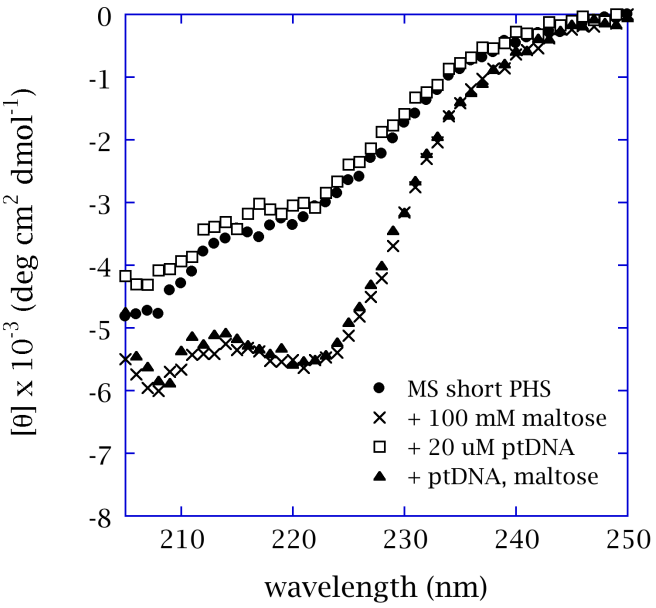


Figure 9. CD spectra of MS short P117G/H124L/S128A with maltose, non-hydrolyzable DNA analog (ptDNA) or both.

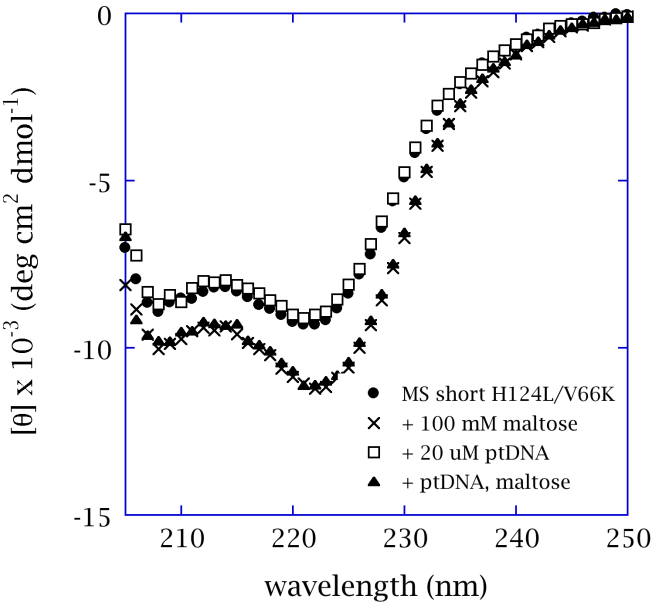
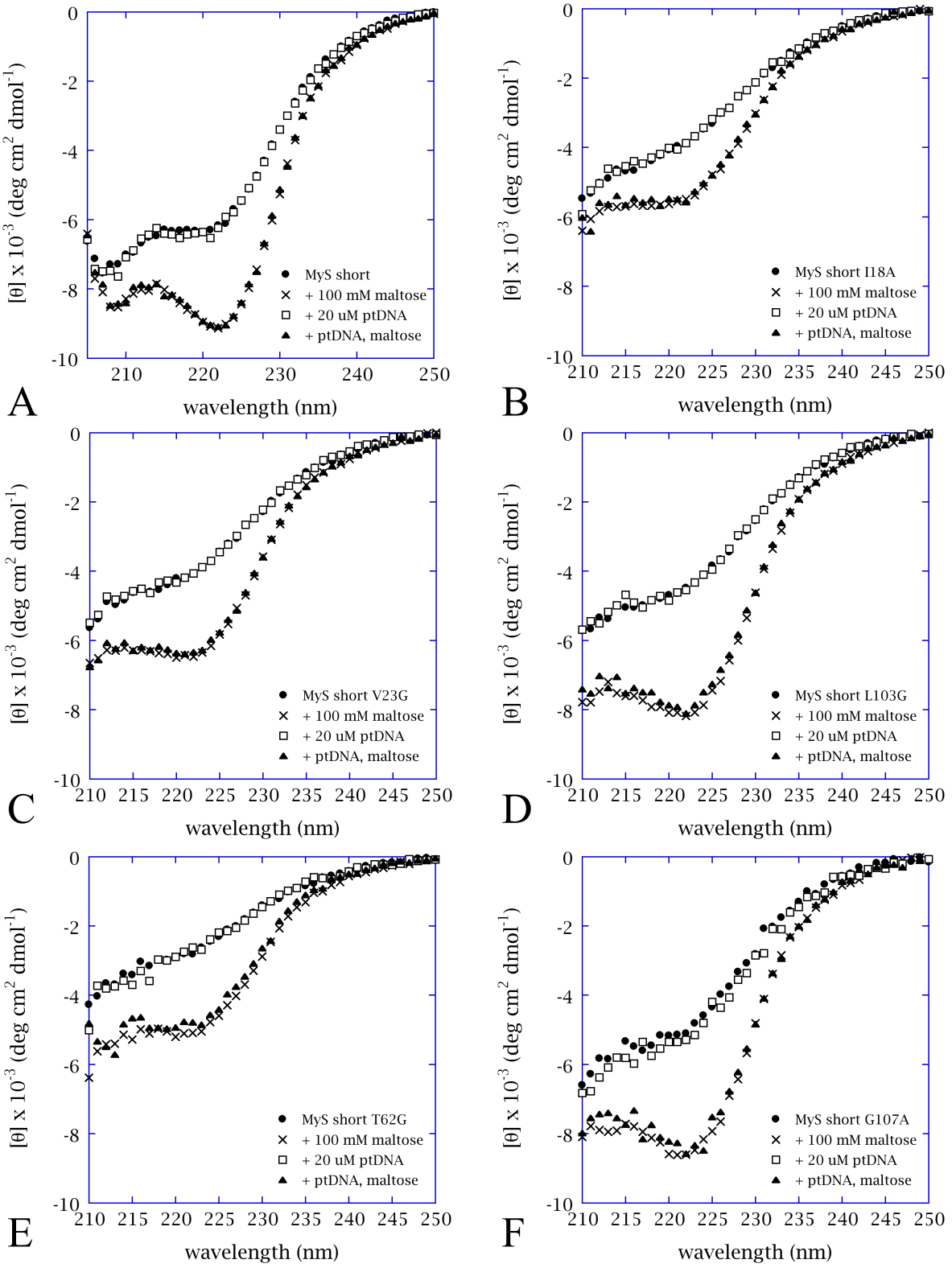


Figure 10. CD spectra of MS short H124L/V66K with maltose, non-hydrolyzable DNA analog (ptDNA) or both.

In order to maximize the potential stabilizing effect of maltose, constructs were made containing the high-affinity MBP variant I329Y. CD spectra of these M_YS short constructs, which have varying stabilities in the SNase domain, reveal maltose-induced switching behavior that is unimpeded by the presence of ptDNA (**Figure 11**). All of the M_YS short constructs share similar spectra in the absence of maltose but change to varying degrees upon the addition of maltose. It is difficult to rationalize the maltose-induced spectral shifts based on differences in SNase destabilization. Small variations in the absolute mean residue ellipticities are probably due to variability in concentration determination rather than differences in secondary structure. All of these constructs demonstrated a propensity for precipitation, making quantifying protein concentration challenging.

Figure 11. CD spectra of M_YS short chimeras containing destabilized SNase domains with maltose, non-hydrolyzable DNA analog (ptDNA) or both. Ordered by the theoretical stability of the SNase domain from most to least stable, they are (A) M_YS short (B) M_YS short I18A (C) M_YS short T62G (D) M_YS short G107A (E) M_YS short V23G (F) M_YS short L103G.



MBP-SNase chimeras were assayed for activity in the presence and absence of maltose; successful switches will be inhibited by maltose. The activities of MS short and MS short PHS were measured both spectroscopically (**Figure 12**) and using the blue-plate assay (**Figure 13**). MS short and MS short PHS are active but retain full activity in the presence of 100 mM maltose. This might indicate that substrate stabilization of the SNase domains in these constructs is preventing the maltose-induced switching behavior observed by CD (**Figures 8 and 9**). Destabilization of the SNase domain will also destabilize the SNase-substrate complex and perhaps enable the MBP-maltose interaction to win the thermodynamic tug-of-war. MS short V66K/H124L was made to test this hypothesis. It contains a very destabilized variant of SNase and is inactive under all conditions (data not shown). This suggests that our destabilization strategy was overly aggressive and that an intermediate stability is needed to ensure that SNase is folded and active unless maltose is present to stabilize MBP.

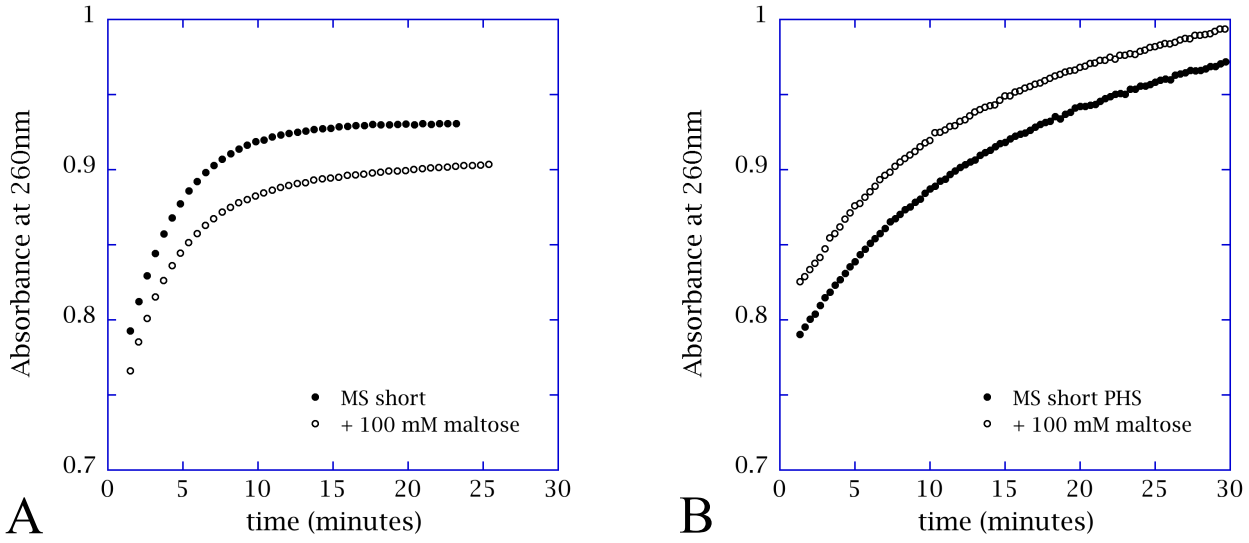


Figure 12. (A) Activity traces for 3 nM MS short with 100 μM substrate with and without maltose. (B) Activity traces for 3 nM MS short PHS with 100 μM substrate with and without maltose.

While MS short V66K/H124L has clearly gone too far in terms of destabilizing the SNase domain, we do appear to have bracketed the relevant stabilities. A series of chimeras containing SNase domains with intermediate stabilities was constructed and tested using the blue-plate assay. The resulting M_YS short chimeras also contain a high-affinity MBP variant in order to maximize stabilization of the MBP domain by maltose. The M_YS short chimeras are either active under all conditions or inactive under all conditions, but activity does not trend with stability. Furthermore, even the active constructs produced very faint spots that are much more difficult to visualize than those observed for MS short PHS (data not shown). In order to reconcile these confusing results, we revisited an important control experiment. We had interpreted inactivity of MS short V66K/H124L as an indication of a very destabilized domain, because SNase V66K/H124L on it own is active; however, the version of SNase in our chimeras is actually a truncated variant. The appropriate control, SNase short V66K/H124L, is inactive under all conditions (**Figure 14**), indicating that we have not, in fact, bracketed the relevant stabilities for the SNase domain. Moreover, there is evidence to suggest that the activity that is observed in

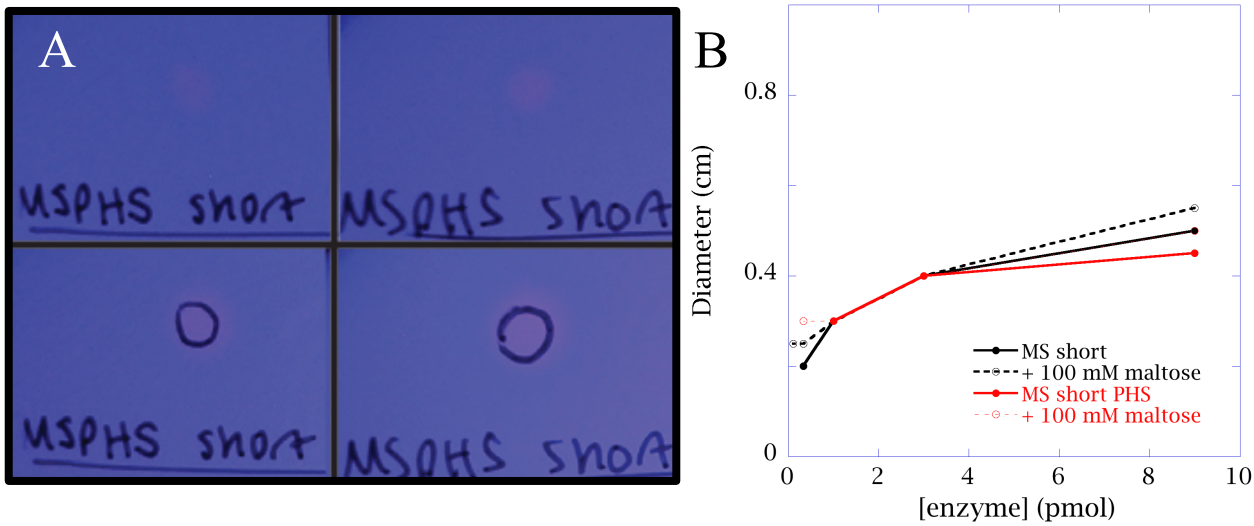


Figure 13. (A) 10 pmol MS short PHS spotted on TB-agar substrate after 3 (top) and 5 (bottom) hours, with 100 mM maltose (right) and without (left). Spot size after 3 hours is indicated in the 5-hour images with black circles. (B) Spot diameter after 3 hours as a function of enzyme concentration.

chimeras with stable SNase domains is due to the remarkable robustness of SNase rather than to hyperstabilization. For instance, one study showed that a C-terminally truncated SNase, lacking residues 137-149 like our short constructs, exists as a compact but unstructured conformational ensemble, but still retains some activity in low salt [28]. Another study showed that a slightly longer truncated SNase, lacking 140-149, is capable of binding substrate before folding [29]. Together, these data suggest that activity observed in some of the MS short constructs is very minimal and does not correlate with the thermodynamic stability of the SNase domain. This makes it difficult to interpret subtle difference between the variants and virtually impossible to rationally engineer activity based on stability.

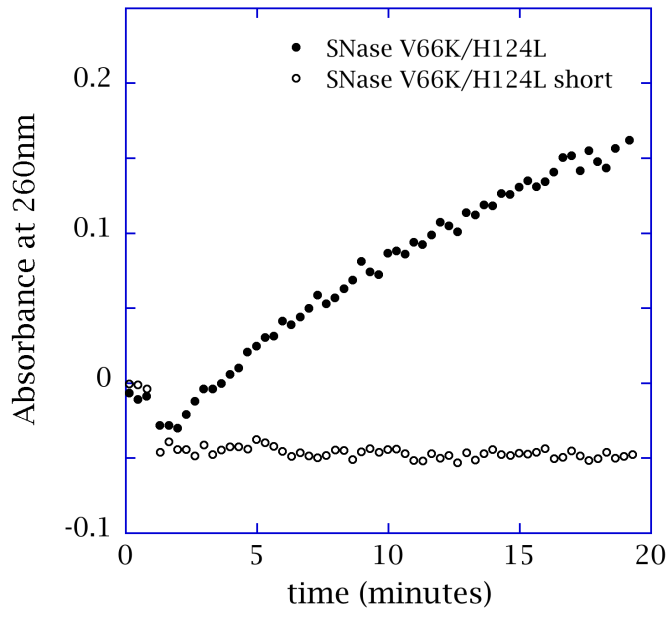


Figure 14. Activity traces for SNase short and SNase short V66K/H124L.

"MBP_{1-286} – SN_{7-141} – $MBP_{287-371}$" (MS long)

Versions of MS long with varying stabilities in the SNase domain were also constructed. Arranged from most to least stable, they include: MS long P117G/H124L/S128A (PHS); MS long; MS long I18A; MS long G107A; MS short V23G; MS long V66K.

The CD spectrum for MS long suggests that this construct is more folded in the absence of maltose than MS short (**Figure 15**). The spectrum changes by only a small amount upon addition of maltose. This is consistent with our design. In the absence of maltose, we expect the SNase domain to be folded and the MBP domain to be unfolded; however, the effective linker lengths are longer here than in MS short, perhaps causing MBP to remain folded to some extent even while SNase is folded. Furthermore, the addition of a non-hydrolyzable DNA analog (ptDNA) impedes the switching behavior. One explanation is that ptDNA binds to the SNase domain, and MBP, even bound to maltose, cannot overcome this additional stabilization in the thermodynamic tug-of-war. Stabilizing the SNase domain, as in MS long PHS, results in a more dramatic maltose-induced conformational change (**Figure 16**). Substrate analog, however, prevents switching in this construct, as well. Because maltose cannot induce switching in the presence of substrate analog, it is doubtful these chimeras will be inhibited by maltose.

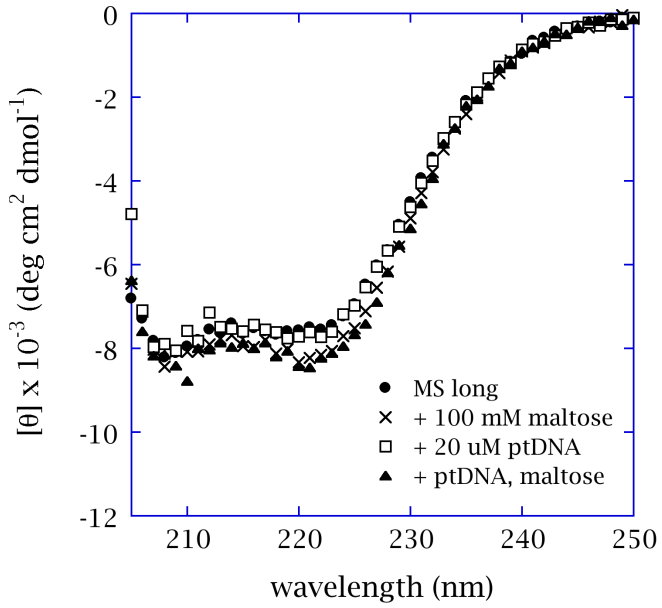


Figure 15. CD spectra of MS long P117G/H124L/S128A with maltose, non-hydrolyzable DNA analog (ptDNA) or both.

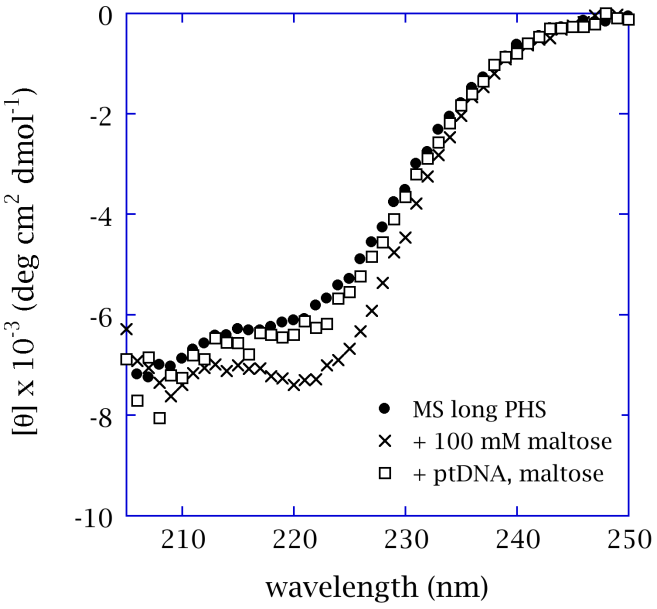


Figure 16. CD spectra of MS long P117G/H124L/S128A with maltose, non-hydrolyzable DNA analog (ptDNA) or both.

The activity MS long PHS was measured using the blue-plate assay (**Figure 17**). It is active but retains its full activity in the presence of 100 mM maltose. When compared with analogous data from MS short PHS (**Figure 13**), it is clear that the long construct, which lacks 8 residues at the C terminus rather than 13, is far more active than the short construct. MS long is also active, but MS long V66K is not (**Figure 18**). The appropriate truncated SNase controls, SNase long and SNase long V66K, are also both active. Again it seems as though we have bracketed the relevant stabilities for the SNase domain, so a series of chimeras with varying stabilities and their truncated SNase controls were constructed. Ordered from most to least stable, the chimeras are: MS long PHS; MS long; MS long I18A; MS long G107A; MS long V23G; MS long V66K. All truncated SNase variants are active. MS long PHS, MS long and MS long I18A show strong activity, and MS long G107A, and to a lesser extent MS long V23G, show weak activity (**Figure 19**). Although the activities do seem to trend with the expected stabilities of the SNase domains,

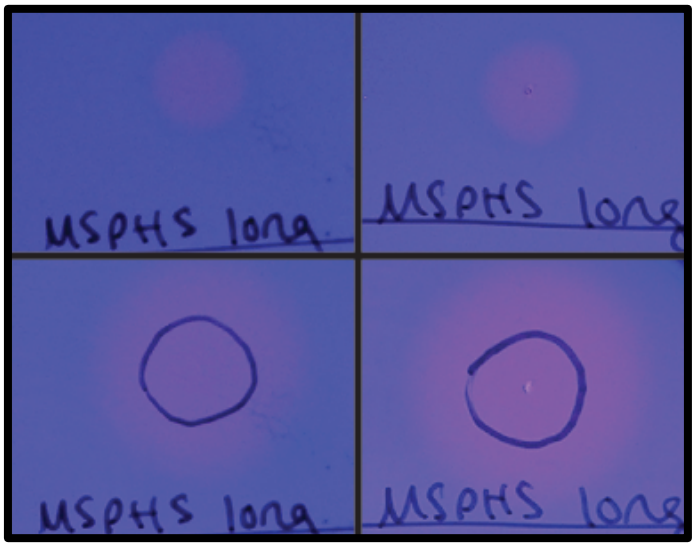


Figure 17. (A) 10 pmol MS long PHS spotted on TB-agar substrate after 3 (top) and 5 (bottom) hours, with 100 mM maltose (right) and without (left). Spot size after 3 hours is indicated in the 5-hour images with black circles.

none of the chimeras are effected by the presence of maltose, even after pre-equilibration with maltose for 60 hours before spotting (**Figure 19**).

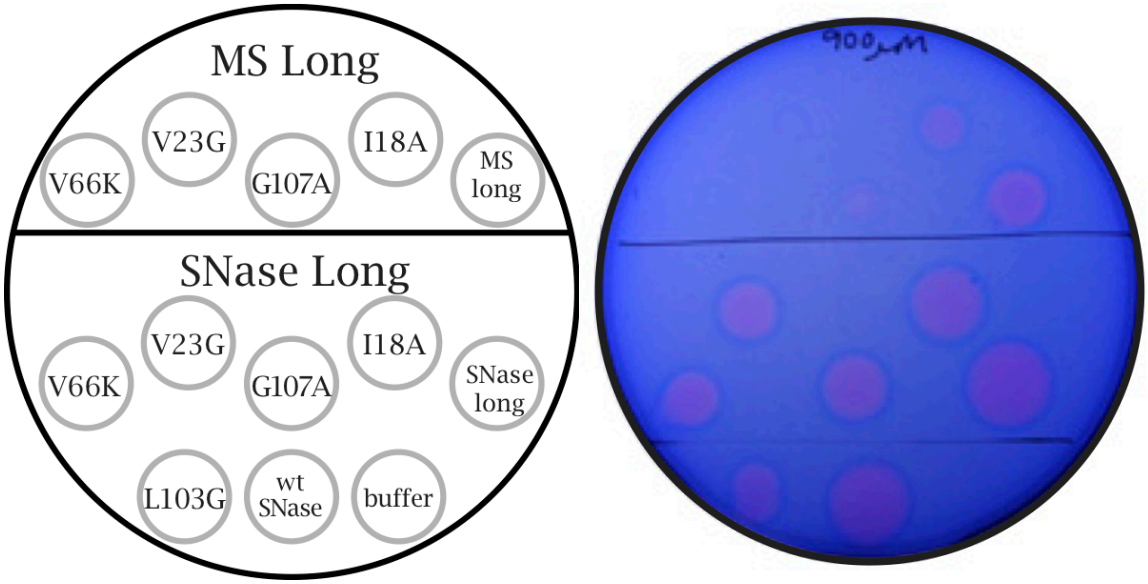


Figure 18. 10 pmol MS long and 0.2 pmol SNase long variants spotted on TB-agar substrate after 3 hours.

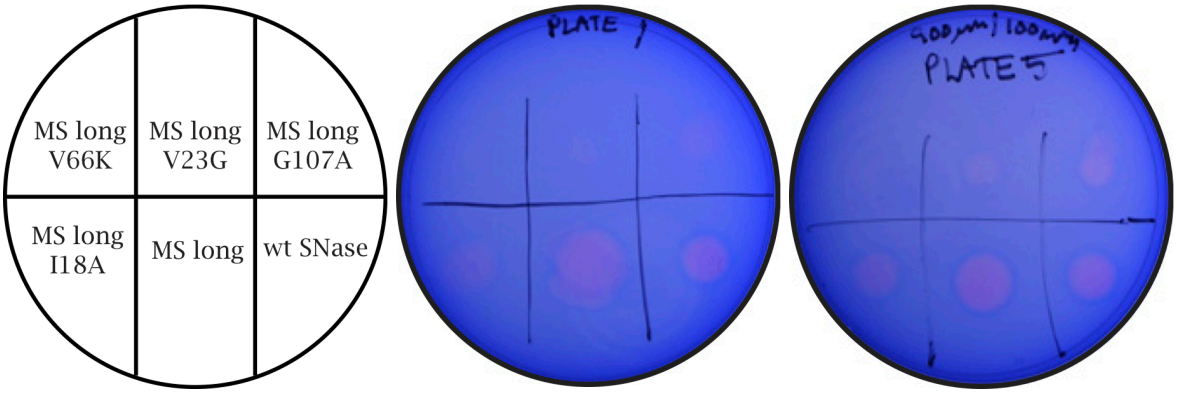


Figure 19. 10 pmol MS long variants spotted on TB-agar substrate with 100 mM maltose (far right) and without (middle) after 3 hours.

4.3.3.2 MBP-SNase chimeras, insertion at residue 169

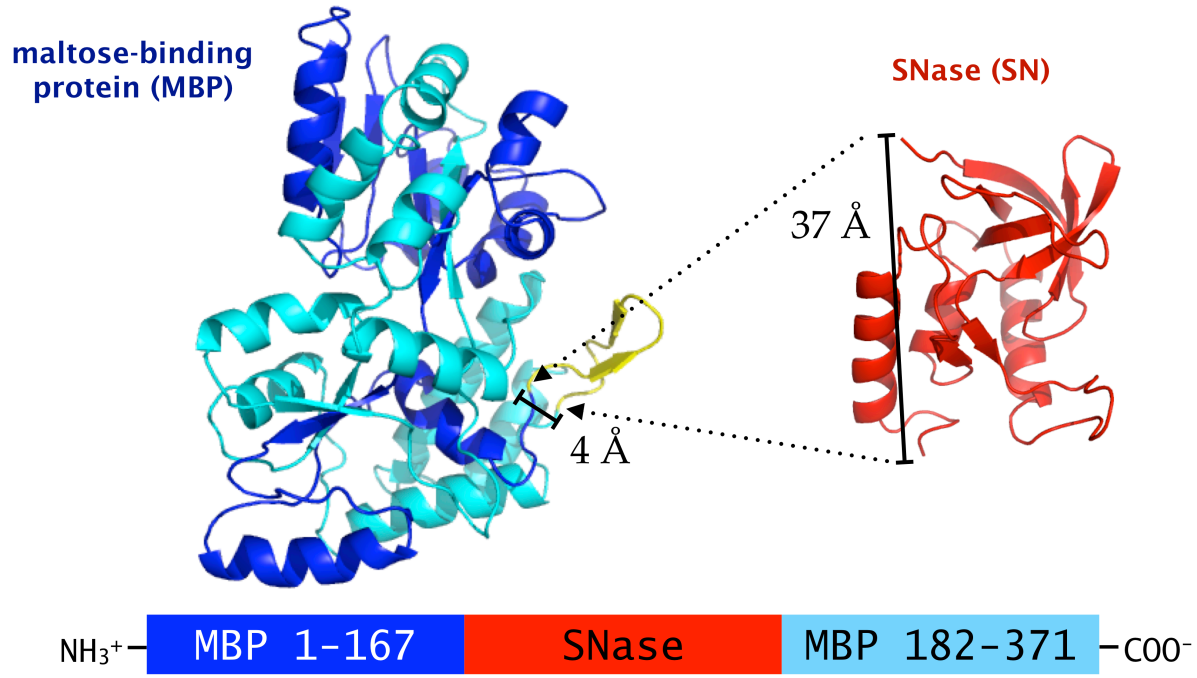


Figure 20. MBP-SNase chimera with insertion replacing residues 169-180. The yellow hairpin indicates the location in MBP's structure where two truncated versions of SNase are inserted to generate $M_{\Delta loop}S$ long and $M_{\Delta loop}S$ short. Below the structures is a schematic of the gene construct.

To test an alternate insertion strategy, a β-hairpin on the surface of MBP was deleted and replaced by the two truncated variants of SNase (**Figure 20**). Previous work showed that deleting one strand of this hairpin has only minor effects on MBP's binding affinity [8], and we surmised that removing the entire hairpin might minimize potentially deleterious structural defects.

The two chimeras in this class are "$MBP_{1\text{-}168} - SN_{10\text{-}136} - MBP_{181\text{-}371}$" ($M_{\Delta loop}S$ short), which has a SNase containing residues 10-136 and "$MBP_{1\text{-}168} - SN_{7\text{-}141} - MBP_{181\text{-}371}$" ($M_{\Delta loop}S$ long), which has a SNase containing residues 7-141. $M_{\Delta loop}S$ long contains only the residues of SNase visible in its crystal structure, while $M_{\Delta loop}S$ short contains a slightly shorter SNase domain that begins and ends with residues involved in secondary structure elements.

The chimeras $M_{\Delta loop}S$ short and $M_{\Delta loop}S$ long both demonstrated maltose-induced switching by CD, but were just as active in the presence as in the absence of maltose (data not shown, personal communication with Tracy Young).

4.3.4 MBP-RNase H chimeras*

4.3.4.1 MBP-RNase H* chimeras, insertion at residue 286

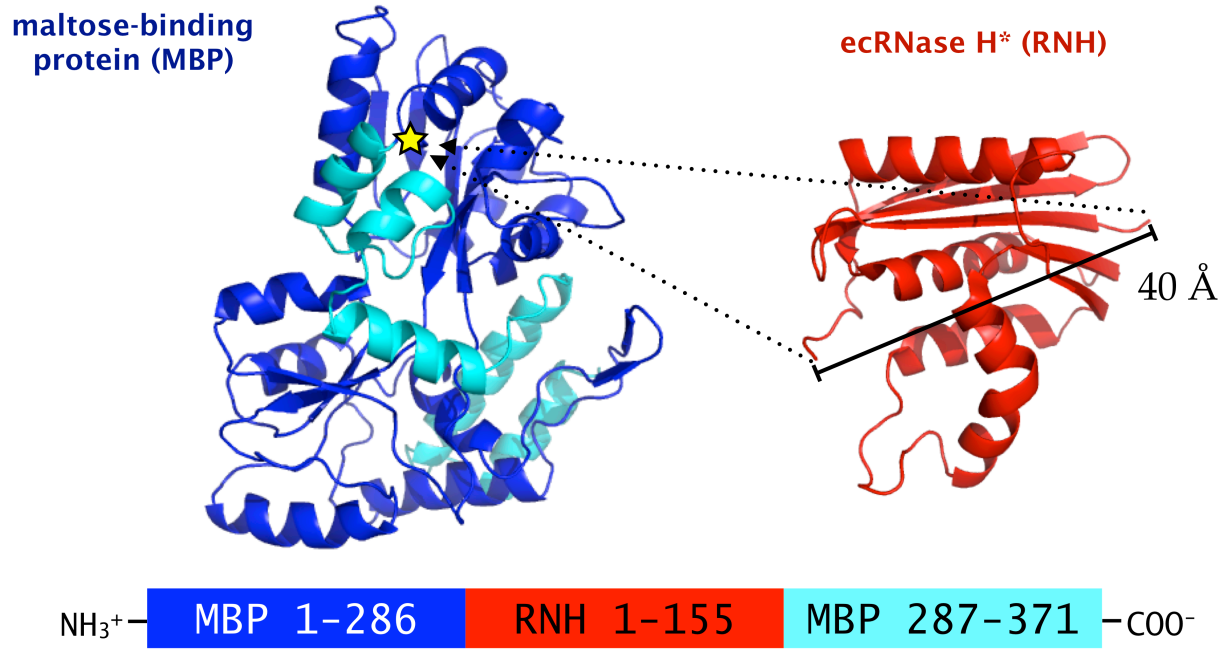


Figure 21. MBP-RNase chimera with insertion between residues 286 and 287. A star indicates the location in MBP's structure where RNase H* is inserted to generate MR. Below the structures is a schematic of the gene construct.

Full-length RNase H* was inserted between residues 286 and 287 of MBP to generate the first class of MBP-RNase H* chimeras called MBP$_{1\text{-}286}$ – RNH*$_{1\text{-}155}$ – MBP$_{287\text{-}371}$" (MR) (**Figure 21**). Versions of MR containing site-specific variants of RNase H* were constructed. The amino acid substitutions I53D and I25A destabilize the native state of RNase H* by a similar degree, 5.6 and 4.6 kcal/mol respectively, but each substitution has a different effect on the high-energy intermediate state. I53D effectively eliminates the intermediate, which is only significantly populated during refolding and not under equilibrium conditions [17]. I25A, on the other hand, selective destabilizes the native state without affecting the intermediate, which results in approximately 15% of the population existing as the partially folded intermediate [18].

Versions of MR I53D (MRD) with varying stabilities in the MBP domain were also constructed and named according to their expected ΔΔG effects [19]. For instance MRD 4 is destabilized by 4 kcal/mol in the MBP domain. Arranged from most to least stable, they include: MRD; MRD I226A (MRD 4); MRD I226A/I161A (MRD 7); MRD I226A/L147A/L115A (MRD 10a); MRD I226A/L147A/I108A (MRD 10b); MRD I226A/I161A/L147A/I108A (MRD 13)

"MBP$_{1\text{-}286}$ – RNH*$_{1\text{-}155}$ – MBP$_{287\text{-}371}$" (MR)

The CD spectrum for MR suggests that a large portion of the protein is unfolded (**Figure 22A**). Upon addition of maltose, a gross conformational change results in a spectrum that more closely resembles a folded protein. This is consistent with our design. In the absence of maltose, we

expect the RNase H* domain to be folded and the MBP domain to be unfolded; however, because MBP is more than twice the size of RNase H*, the unfolded protein dominates the CD signal, which reflects average properties of the system. When maltose is added, MBP folds, causing SNase to unfold and giving rise to a more folded spectrum. When a destabilizing amino acid substitution is introduced into the RNase H* domain, the resulting construct MR I53D exhibits similar switching behavior (**Figure 22B**).

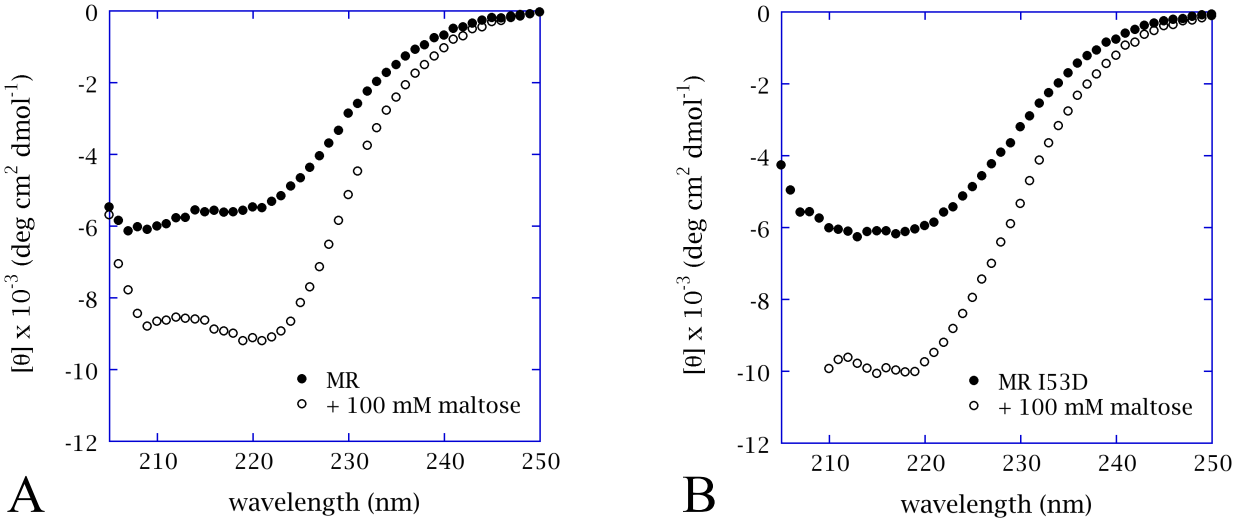


Figure 22. CD spectra with and without maltose for (A) MR and (B) MR I53D.

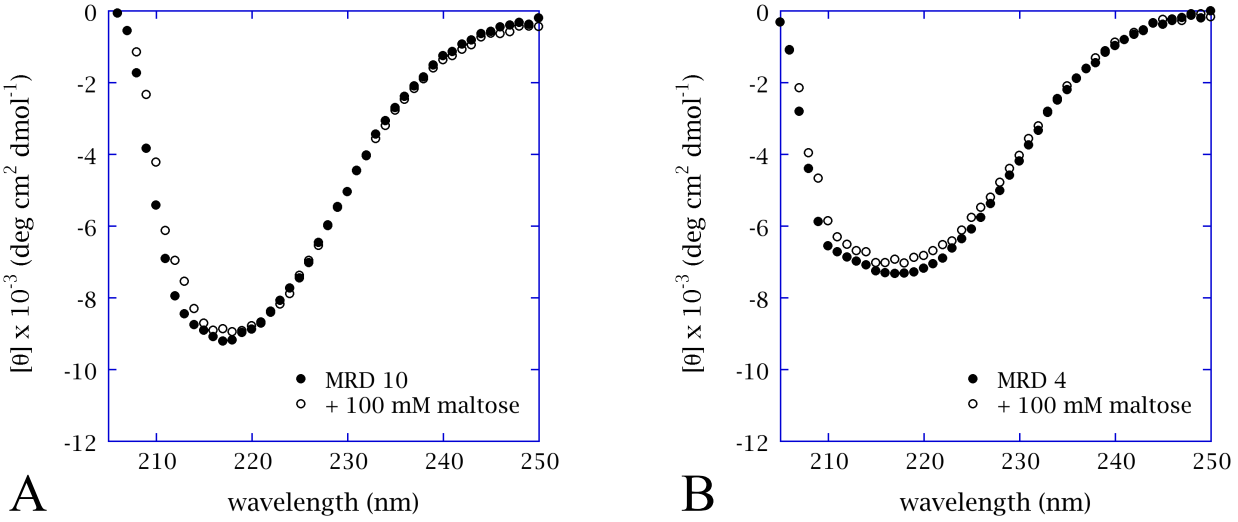


Figure 23. CD spectra with and without maltose for (A) MRD 10 and (B) MRD 4.

MBP-RNase H* chimeras were assayed for activity in the presence and absence of maltose; successful switches should be inhibited by maltose. MR is active under all conditions, and MR I53D is inactive (**Figure 25**). Thus several constructs were made with destabilizing substitutions in the MBP domain of MR I53D. Destabilizing the MBP domain should result in a relative stabilization of the RNase H* domain, enabling it to fold and be active. MRD 4 and MRD 10 contain multiple substitutions which, assuming additivity, cause destabilization in the MBP

domain of 4 and 10 kcal/mol, respectively. The expectation for these constructs is that the RNase H domain will be more stable than the MBP domain and therefore be folded and active until maltose is added. Addition of maltose, however, did not alter the CD spectra, suggesting that maltose does not cause large structural changes in these variants (**Figure 23)**. Other variants behaved in a similar manner (data not shown).

MR demonstrates RNase H activity but is a slower than the wild type enzyme (**Figure 24**). Maltose causes a slight lag in the initial velocity, but this might be due to increased viscosity of the 100 mM maltose solution. Overall, maltose does not seem to inhibit MR's activity, despite the fact that its CD spectrum changes with maltose (**Figure 22**). Substrate binds to the RNase H* domain, so it is possible that under activity conditions, maltose does not stabilize the MBP domain enough to overcome substrate-based stabilization and unfold the enzyme.

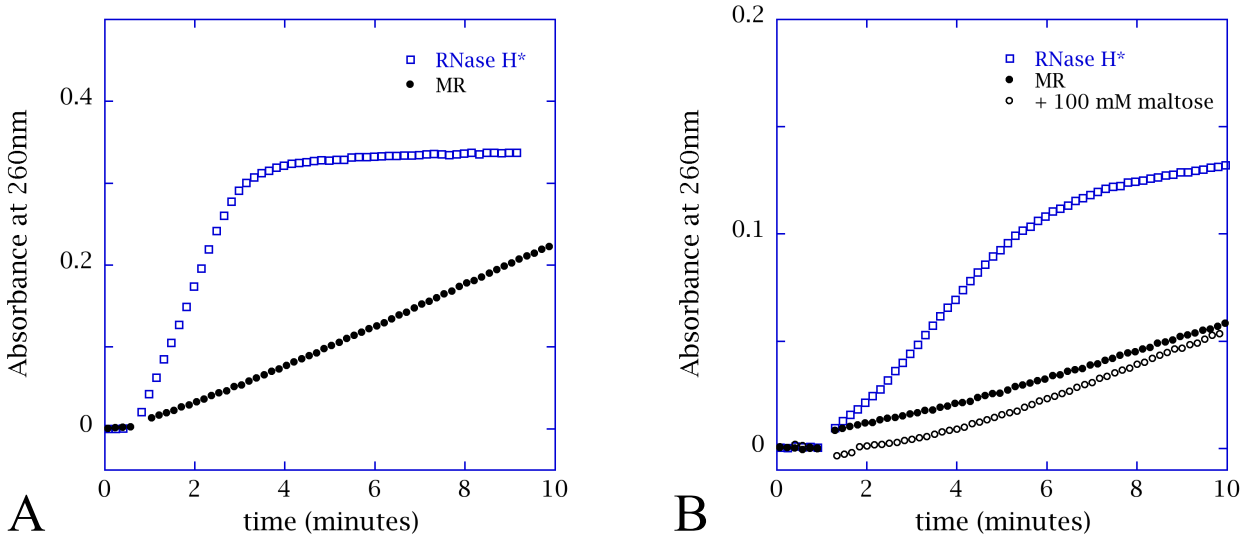


Figure 24. (A) Activity traces of RNase H* and MR with 50 μM substrate. (B) Activity traces of RNase H* and MR with 25 μM substrate compared with and without 100 mM maltose.

To counter the stabilizing effects of substrate, the destabilizing substitution I53D was introduced to the RNase H* domain. Although RNase H* I53D alone is active, MR I53D is inactive under all conditions (**Figure 25**). Believing we had now tipped the scales too much in the opposite direction, five variants of MR I53D were made that contained severely destabilized MBP domains (see Section 4.3.1.3). These variants contain considerable secondary structure, but do not exhibit maltose-induced switching by CD (**Figure 23**). None were active (data not shown).

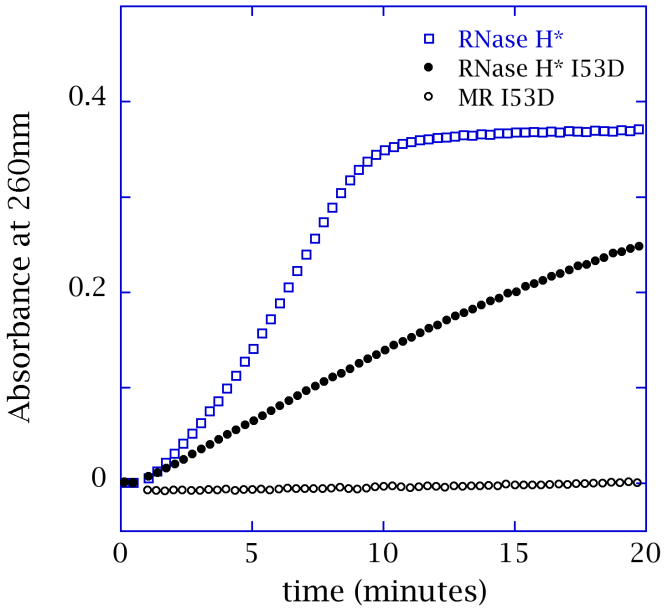


Figure 25. Activity traces of RNase H* and MR I53D with 50 μM substrate. Traces for variants of MR I53D are not shown, but they all overlay with MR I53D's trace.

4.3.4.2 MBP-RNase H chimeras, insertion at residue 169*

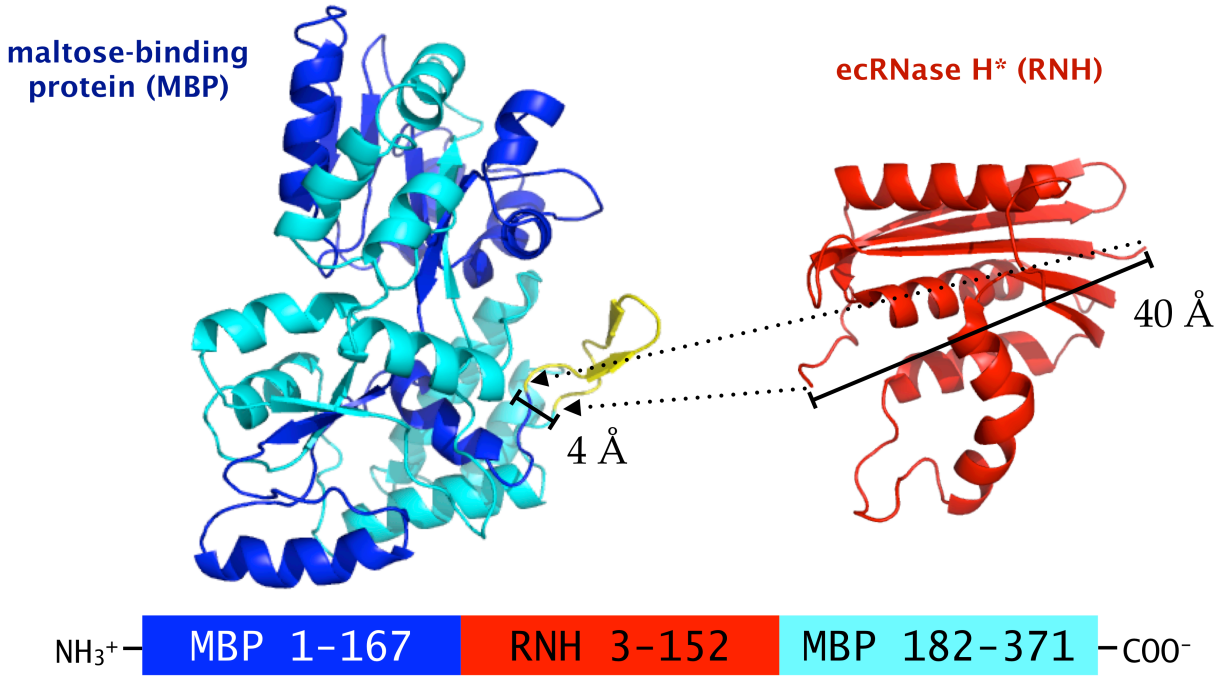


Figure 26. MBP-RNase chimera with insertion replacing residues 169-181. The yellow hairpin indicates the location in MBP's structure where a truncated versions of RNase H* is inserted to generate $M_{\Delta loop}R$. Below the structures is a schematic of the gene construct.

To test an alternate insertion strategy, a β-hairpin on the surface of MBP was deleted and replaced by a slightly truncated variant of RNase H*, which includes only residues visible in the crystal structure [15]. Previous work showed that deleting one strand of this hairpin has only

minor effects on MBP's binding affinity [8], and we surmised that removing the entire hairpin might minimize potentially deleterious structural defects.

The resulting chimera is called MBP_{1-168} – RNH^*_{3-152} – $MBP_{181-371}$" ($M_{\Delta loop}R$), and it contains a truncated version of RNase H* that is comprised of residues 3-152 (**Figure 26**). Similar to the rationale for the MBP-SNase chimeras, the RNase H* domain is shortened to ensure maximal coupling between the two domains. Versions of $M_{\Delta loop}R$ with several variants of RNase H* were constructed. See MR for a description of these variants.

"MBP_{1-168} – RNH^*_{3-152} – $MBP_{181-371}$" ($M_{\Delta loop}R$)

The CD spectrum for $M_{\Delta loop}R$ changes in response to maltose, as evidenced by a downward shift at the 222 nm minimum (**Figure 27**). A variant of this chimera, which contains the destabilizing substitution I25A in the RNase H* domain, does not exhibit this behavior (**Figure 28**). The absence of CD switching in $M_{\Delta loop}R$ I25A may indicated that the MBP domain is folded, and the RNase H* domain unfolded, under all conditions. Because substrate will bind to and stabilize the RNase H* domain, it is possible that a spectral shift under these conditions is not a requirement for a functional switch. Thus while we have used CD to probe maltose-induced structural changes, the true test for our switches is whether they are active enzymes and how the activity depends upon maltose.

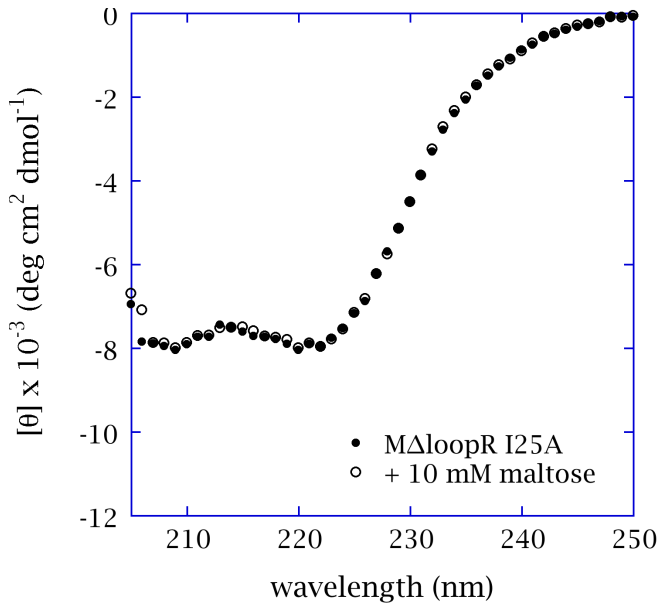


Figure 27. CD spectra of $M_{\Delta loop}R$ with and without maltose.

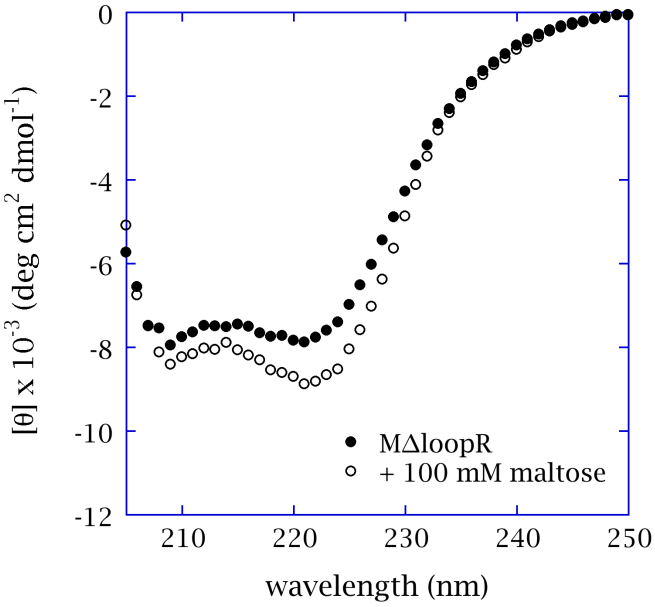


Figure 28. CD spectra of $M_{\Delta loop}R$ I25A with and without maltose.

$M_{\Delta loop}R$ demonstrates RNase H activity but is slower than the wild type enzyme (**Figure 29**). Maltose decreases the initial velocity of $M_{\Delta loop}R$ activity in a concentration-dependent manner. In 10 mM maltose, $M_{\Delta loop}R$ hydrolyzes at roughly half its rate in buffer, and 100 mM maltose slows the rate by over fourfold. Michaelis-Menten analysis of $M_{\Delta loop}R$ in the absence of maltose finds k_{cat} = 1.5 x 10^4 s^{-1}, K_m = 4.2 µM, yielding an overall catalytic efficiency that is near diffusion-limited at 3.8 x 10^9 $M^{-1}s^{-1}$ (**Figure 30A**). In the presence of 100 mM maltose, $M_{\Delta loop}R$ is 5 orders of magnitude less efficient with k_{cat} = 1.9 x 10^3 s^{-1}, K_m = 25 µM and k_{cat}/K_m = 3.8 x 10^9 $M^{-1}s^{-1}$ (**Figure 30B**). The reduction in efficiency is due both to a decrease in turnover and binding affinity, which suggests that maltose fundamentally changes the reaction mechanism rather than simply depleting enzyme via unfolding. If the reduction in activity were due strictly to a lower effective concentration of enzyme, then we would expect the K_m to remain unchanged.

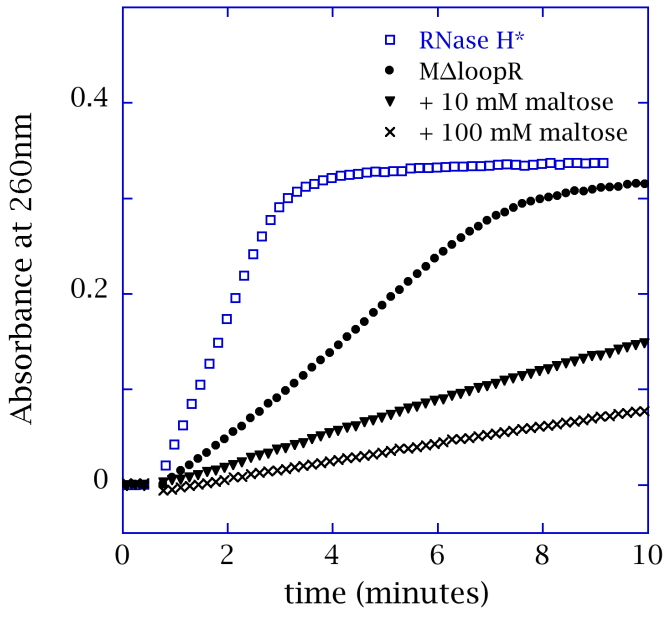


Figure 29. Activity traces of 5 nM RNase H* and 5 nM $M_{\Delta loop}R$ as a function of maltose with 50 µM substrate.

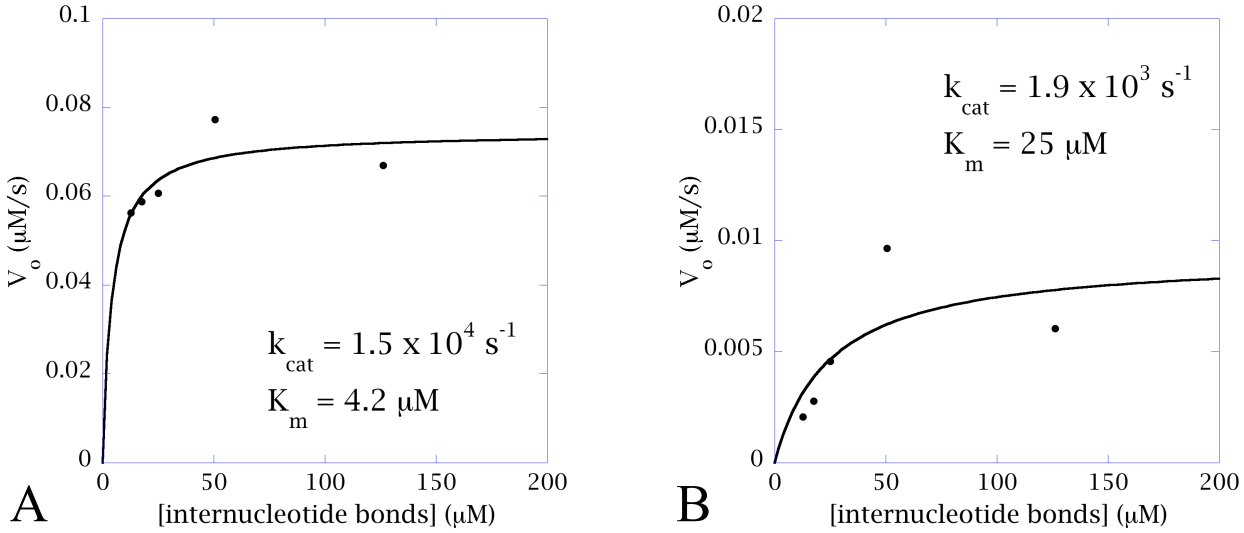


Figure 30. (A) Michaelis-Menten analysis of $M_{\Delta loop}R$ using 5 nM enzyme. (B) Michaelis-Menten analysis of $M_{\Delta loop}R$ using 5 nM enzyme and 100 mM maltose.

To test the limits of this design, we wanted to see if we could construct a variant that demonstrated complete inhibition by maltose. Our first strategy was the same one used in the MR chimera, where the RNase H* domain is selectively destabilized with an I53D substitution. The resulting construct is inactive under all conditions (data not shown). Our next strategy was to introduce a different amino substitution, I25A. While this substitution has nearly the same $\Delta\Delta G$ effect on the native state, it differs in its effect on the intermediate. As a result, I53D causes RNase H* to fold slowly in a two-state manner. RNase H* I25A, on the other hand, populates a partially folded intermediate state at equilibrium [18]. Thus $M_{\Delta loop}R$ I25A represents a departure from our original design criteria, which required enzymes existing in only two states, folded and unfolded.

Hydrolysis catalyzed by $M_{\Delta loop}R$ I25A is slower than that of $M_{\Delta loop}R$, but its activity is completely inhibited by 1 mM maltose (**Figure 31**). This represents our most successful design. One model to explain its exquisite maltose sensitivity is that the RNase H* domain exists in a partially folded, inactive state when MBP is bound to maltose. In the absence of maltose, the RNase H* I25A domain is more poised to bind substrate, because the energetic gap between its intermediate and native states is only 1.1 kcal/mol [18]. $M_{\Delta loop}R$ is less sensitive to maltose, because there is a much larger gap between its intermediate and native states, in the range of 6 - 7 kcal/mol [30, 31], and MBP binding maltose can only partially overcome it. This does not explain, however, why $M_{\Delta loop}R$ is active and $M_{\Delta loop}R$ I53D is not. The energetic gap between the native and unfolded states for RNase H* I53D is 5.7 kcal/mol [17], which is very close to the gap between the native and intermediate states of RNase H*. Based strictly on thermodynamics, it seems like if RNase H* can fold and become active in the presence of substrate, then RNase H* I53D should be capable, as well. The fact that $M_{\Delta loop}R$ I53D is inactive under all conditions suggests that the ability to populate a partially folded state is critical for function. It may, for instance, allow relief of conformational strain when MBP is folded while still preserving enough preformed structure to promote substrate-induced folding of the RNase H* domain. Perhaps folding from a fully unfolded state simply takes too long to be observed under these conditions,

even though thermodynamically it should be more stable than MBP bound to maltose. This might also explain why no MBP-SNase chimeras exhibited maltose inhibition, as SNase is strictly two-state.

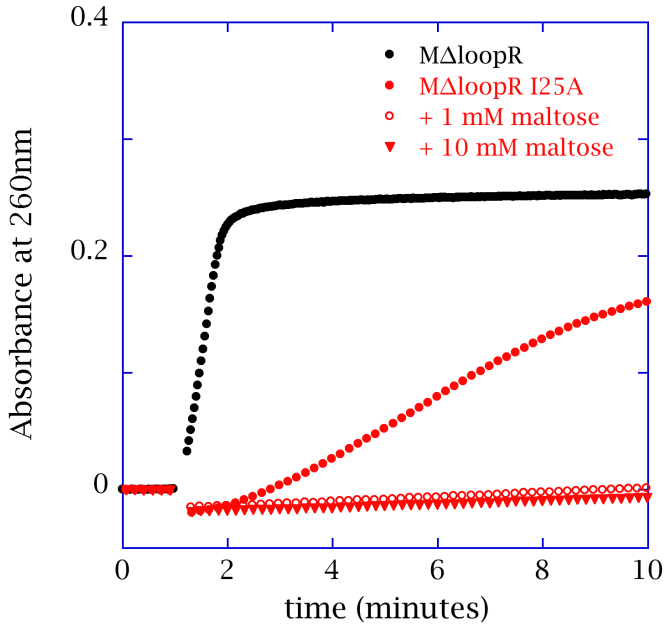


Figure 31. Activity traces of 50 nM $M_{\Delta loop}R$ and 50 nM $M_{\Delta loop}R$ I25A as a function of maltose with 50 μM substrate.

4.3.4.3 MBP-RNase H flip-flopped chimera*

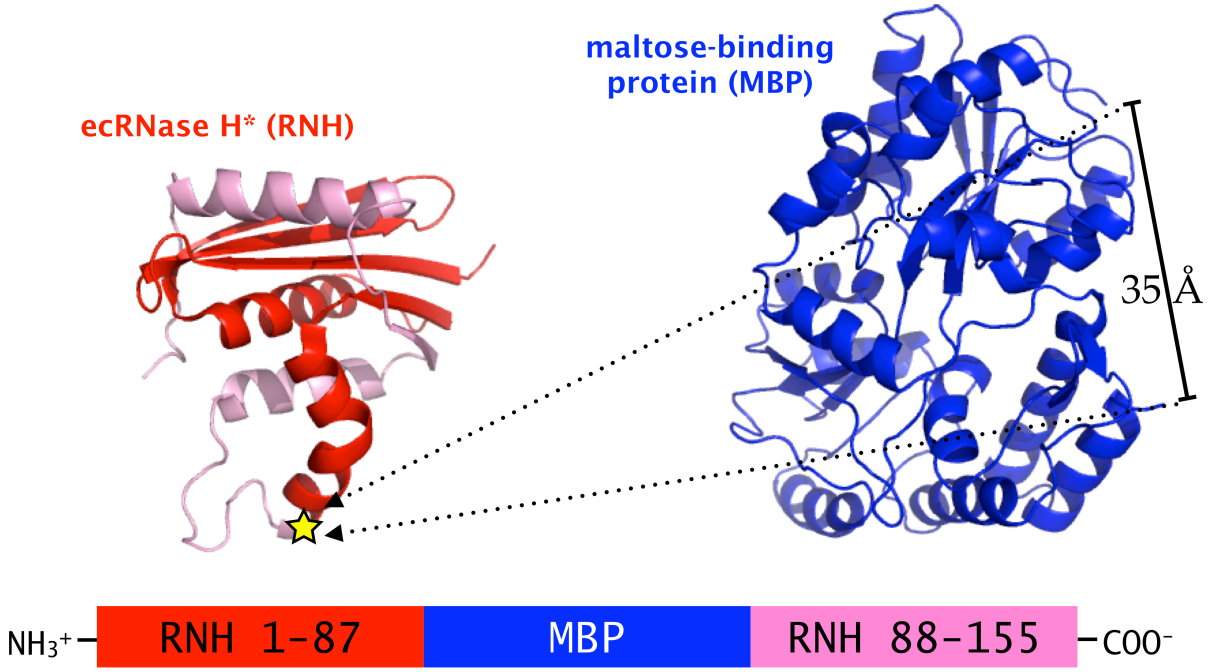


Figure 32. MBP-RNase chimera with "flip-flopped" topology. The yellow star indicates the location in RNase H*'s structure where full-length MBP is inserted to generate RM. Below the structures is a schematic of the gene construct.

To assess the effects of splitting the enzyme domain rather than the regulatory domain, an MBP-RNase H* chimera with opposite topology to MR was constructed. The resulting chimera, called "RNase H*$_{1\text{-}87}$ – MBP$_{1\text{-}371}$ – RNH*$_{88\text{-}155}$" (RM), is a flip-flopped version of MR, where MBP is inserted between consecutive residues in the basic loop of RNase H* (**Figure 32**). While this alternate topology represents a departure from our original design, it still satisfies our major guiding principles. MBP has an end-to-end distance of 35 Å in its folded structure [32], which is consistent the requirements for the inserted domain. And proteolytic studies of RNases H* suggest that cleavage in the basic loop does not compromise structure and stability in the rest of the protein [33], making it an ideal location for domain insertion.

"RNH*$_{1\text{-}87}$ – MBP$_{2\text{-}370}$ – RNH*$_{88\text{-}155}$" (RM)

Only one version of the "flip-flopped" chimera was made. The MBP domain starts with the second residue of the sequence and ends with the penultimate residue. RM represents a flip-flopped chimera, where MBP is inserted into the basic loop of RNase H*. It was constructed to test the hypothesis that in active chimeras insensitive to maltose, MBP folding might not generate enough force to fully unfold the enzyme. Indeed, it has been shown in circular permutants of barnase that forcing termini closer together creates strain but does not destroy activity ([34]. Splitting the active domain is expected to slow its refolding and might also make it easier to turn off activity, because unfolded enzyme will be ripped apart. Unfortunately, this extreme topological constraint results in complete inactivity. RM is inactive under all conditions (**Figure 33**), and no CD experiments were performed with this construct. The model for the RNase H domain, which contains a ten-glycine insertion in place of MBP, is active but slower than RNase H*.

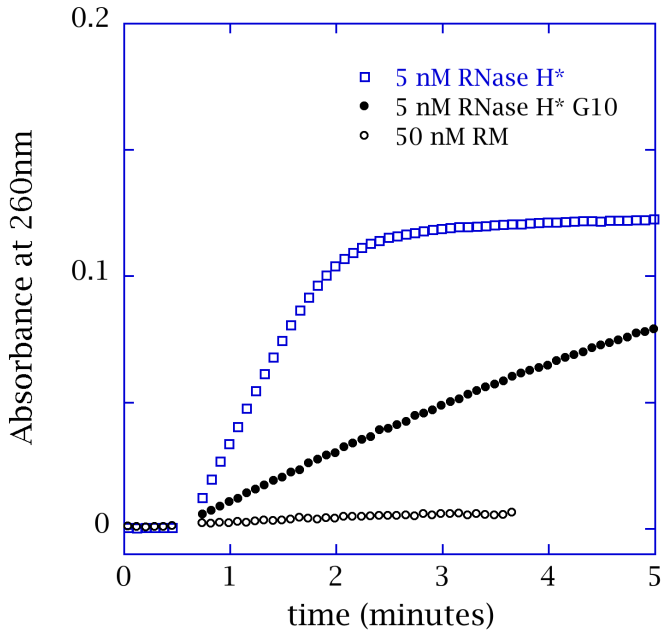


Figure 33. Activity traces of 5 nM RNase H* and 5 nM RNase H* G10 compared with the activity of 50 nM RM in 1 mM $MnCl_2$ and with 25 μM substrate.

4.4 Materials and Methods

4.4.1 Construction, expression and purification of switches

4.4.1.1 MBP-SNase chimeras

MBP-SNase chimeras were cloned by Tracy Young into a pET28 vector with a thrombin-labile N-terminal 6xHis tag. The thrombin site was replaced by a TEV cleavage site in a single step via site-directed mutagenesis to create a modified pET28 vector. Other site-specific variants were constructed via site-directed mutagenesis and verified by sequencing.

Plasmids were transformed into BL21(DE3)pLysS cells for expression under T7 promoter control. Cells were induced with 1 mM IPTG at OD = 0.6 and grown at 37 °C for 3 hours before harvesting. Cells were lysed in buffer via sonication; then inclusion bodies were isolated and washed with non-ionic detergent. Inclusion bodies were solubilized in 6 M urea, and the protein was run over a column containing Ni-NTA agarose resin (GE Healthcare) and eluted with 250 mM imidazole. Protein was diluted and dialyzed against buffer before cleaving with TEV protease overnight at 4 °C to remove the his-tag. The cleavage reaction was then run over the Ni-NTA column again to remove the tag and uncleaved protein. Protein was then concentrated, minimally, and dialyzed against either ammonium bicarbonate for subsequent freeze-drying and storage or appropriate buffer conditions for immediate use. Each protein's purity and molecular weight were confirmed by SDS-PAGE and electrospray mass spectrometry.

4.4.1.2 MBP-RNase H chimeras*

$MBP_{1\text{-}168}$ – $RNH^*_{3\text{-}152}$ – $MBP_{181\text{-}371}$ ($M_{\Delta loop}R$) was cloned by Tracy Young. $MBP_{1\text{-}286}$ – $RNH^*_{1\text{-}155}$ – $MBP_{287\text{-}371}$ (MR) was constructed using a protocol developed by Tracy Young. Blunt end PCR was used to amplify DNA encoding residues 1-286 and 287-371 of *E. coli* maltose binding protein (MBP) and residues 1-155 of *E. coli* RNase H. The 5' end of MBP was ligated to the RNase H gene and run on a gel. The appropriately sized band was isolated from the gel, purified and amplified by PCR. This product was ligated to the 3' end of MBP and run on a gel. The appropriate band was isolated from the gel, purified and amplified with PCR. The resulting fusion was ligated into the *Nde*I and *Hind*III sites of modified pET28 cloning and expression vector from Novagen. Other site-specific variants were constructed via site-directed mutagenesis and verified by sequencing.

Both $RNH^*_{1\text{-}87}$ – $MBP_{2\text{-}370}$ – $RNH^*_{88\text{-}155}$ (RM) and RNase H* G10 were constructed via a modified site-directed mutagenesis method optimized for large insertions. First mega-primers are generated using Phusion High-Fidelity DNA polymerase (NEB) to amplify the inserted gene, either encoding $(Gly)_{10}$ or MBP residues 2-370. The primers used in this initial PCR step add flanking regions that will anneal to the middle of the RNase H* gene between sequence encoding residues 87 and 88. The resulting PCR product was run on a gel and purified. It was then used to prime pSM101 in a modified cycling program where the annealing temperature is gradually increased after several initial rounds and using prolonged extension times. The entire gene was then PCR amplified again with flanking restriction sites for sub-cloning into the modified pET28 vector to add a TEV-labile N-terminal 6xHis-tag.

MBP-RNase H* chimeras were expressed and purified as described for the MBP-SNase chimeras.

4.4.2 Circular dichroism spectroscopy

CD measurements were collected on an AVIV 410 spectrophotometer. Spectra were taken with protein samples at 0.5 mg/ml in a 0.1 cm quartz cuvette at 25 °C. Many of the chimeras were not soluble at this concentration, so their spectra were taken in longer pathlength cuvettes with correspondingly lower concentrations (0.25 mg/mL in a 0.2 cm cuvette, 0.1 mg/mL in a 0.5 cm cuvette). Data points were collected from 250-200 nm at 1-nm intervals, and each data point represents signal averaged over 5 seconds. Data for which the dynode voltage exceeded 500 V were discarded. Buffer conditions for SNases and SNase chimeras were 300 mM NaCl, 10 mM Tris (pH 7.5), 10 mM $CaCl_2$ and varying amounts of maltose and/or a 6-mer oligophosphorothioate (IDT). Buffer conditions for RNases H* and RNases H* chimeras were 50 mM NaCl, 50 mM Tris (pH 8), 10 mM $MgCl_2$ and varying amounts of maltose. Buffer conditions for MBP and its variants were 100 mM NaCl, 10 mM Tris (pH 8) and varying amounts of maltose.

Urea denaturations were performed in a 1-cm pathlength quartz cuvette at 25°C by monitoring the CD signal at 222 nm and averaging the signal over 60 seconds for each data point. Individual samples containing 0.5 mg/mL of protein in the same buffer used for spectra and varying concentrations of urea were equilibrated at 25°C overnight. Urea concentrations were verified using a refractometer.

4.4.3 Activity assays

4.3.1 MBP-RNase H* chimeras

RNase H activity was assayed in 50 mM NaCl, 50 mM Tris (pH 8.0), 10 mM $MgCl_2$ at 25 °C. Substrate was prepared by mixing equal parts dT_{20} oligomers (IDT) and poly-rA (Sigma), heating to 95C for 5 min, then slowly cooling to room temperature for one hour before storing at 4 °C. The reaction was monitored at 260 nm using a Cary UV spectrophotometer. Increasing absorbance at 260 nm indicates the release of nucleotides as they are hydrolyzed. Initial velocities were measured in the 40-60 second range, which represents the first 10-30 seconds post-initiation, and fit to the Michaelis-Menten equation using KaleidaGraph (version 4.1.2):

$$v = \frac{k_{cat}[\mathrm{E}][\mathrm{S}]}{K_m + [\mathrm{S}]} \quad (3)$$

4.4.3.2 MBP-SNase chimeras

Spectroscopic assay:

SNase activity was assayed spectroscopically in solution with 300 mM NaCl, 50 mM Tris (pH 7.5) and 10 mM $CaCl_2$ at 25 °C. Substrate was prepared by sonicating salmon sperm DNA (Sigma) followed by boiling for 10 minutes before quenching on ice. Substrate concentration is given in internucletotide bonds, due to the heterogeous nature of the substrate, using ε_{260} = 8250 $M^{-1}cm^{-1}$ and 330 g/mol for the average nucleotide molecular weight. The reaction was monitored at 260 nm using a Cary UV spectrophotometer. Increasing absorbance at 260 nm indicates the release of nucleotides as they are hydrolyzed.

Blue plate assay:

SNase activity was assayed colorimetrically on 1% agar plates containing 300 mM NaCl, 50 mM Tris (pH 7.5), 10 mM $CaCl_2$, 300 μM toluidine blue and 900 μM sonicated and denatured salmon sperm DNA. The method was adapted from Lachica *et al.* [25, 26]. Plates were prepared by boiling to bring all components into solution followed by cooling in petri dishes. Either purified protein or cell lysate, discussed below, was spotted on plates in 2 μL volumes and allowed to develop for 2-5 hours at 37 °C. For purified proteins, 0.2 pmol of wt SNase and 10 pmol of each chimera were spotted. Activity is evidence by the appearance of a pink halo in the blue substrate. Toluidine blue is a pH indicator, and the pink color change occurs as protons are produced during DNA hydrolysis.

Lysate preparation for blue plate assay:

SNase activity can be measured in cell lysate using the blue plate assay. Cultures containg the expression plasmid are grown until OD = 0.6 and then induced with 1 mM IPTG for 2 hours. Cells from either 1 mL of culture, for SNase variants, or 25 mL, for chimeras, are harvested and resuspended in 2 mL of lysis buffer, which contains 300 mM NaCl, 50 mM Tris (pH 7.5), 10 mM $CaCl_2$, 10% sucrose, 1% octylglucoside and 1% Triton X-100. Lysate is then vortexed and spun down. Variants of SNase all express solubly and are spotted directly from the lysis supernatant. Because the chimeras all express insolubly, however, several steps are required post-lysis to prepare the samples for analysis. First, lysis supernatant is decanted. Then the pellet is resuspended in 2 mL of non-ionic detergent buffer (50 mM Tris pH 8.0, 1% Nonidet-P40, 1% deoxycholic acid), vortexed and spun down. Detergent wash is repeated once more. Then, the pellet is resuspended in 6 M urea, 300 mM NaCl, 50 mM Tris (pH 8.0) and 10 mM $CaCl_2$ and spun down to remove insoluble material. The supernatant is either diluted to 1 M urea or dialyzed against buffer overnight, which requires another spin before spotting. For samples with maltose, 100 mM maltose is included at every step, from the LB media during expression to the final dialysis buffer.

4.5 Discussion

Several chimeras demonstrate maltose-induced secondary structure changes, as reported by far-UV CD. Shifts in the CD signal are most likely due to the MBP domain folding upon binding maltose. In the absence of maltose, the active domain is folded, but because both SNase and RNase H* are small relative to MBP, the unfolded protein dominates the signal. CD spectra of an MBP variant designed to be unfolded under native conditions lends evidence for this model (**Figure 34A**). When maltose is added, the minimum at 222 nm shifts downward, more closely resembling the fully folded wt MBP (**Figure 34B**). That it never fully recapitulates the native signal suggests that either the maltose added isn't adequate to overcome the destabilization or that the folded spectrum for this variant differs fundamentally from wt MBP due to its amino acid substitutions.

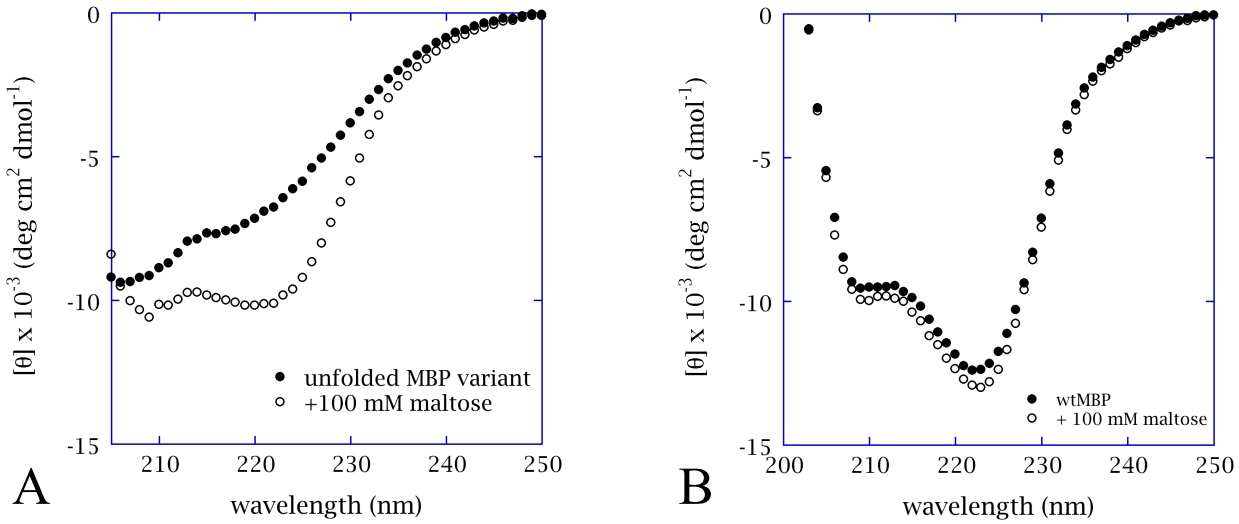


Figure 34. (A) CD spectra for an unfolded MBP variant with 100 mM maltose and without. This destabilized variant has a theoretical stability of ΔG = 0 kcal/mol and contains the substitutions I59A/L115A/L147A/I161A/I226A. (B) CD spectra for wt MBP variant with 100 mM maltose.

Evidence that MBP is unfolded in the absence of maltose does not necessarily mean that the active domain is folded; however, the fact that several of the constructs demonstrating CD switching behavior, such as MS short PHS and MS long PHS, are also active enzymes suggests that they might be folding in a mutually exclusive manner**.**

Maltose-induced structural switching is somewhat predictive of maltose-inhibition. MS short PHS exhibits the most promising behavior as probed by CD, as it has a maltose-induced structural change that persists in the presence of non-hydrolyzable substrate. This construct's activity is unaffected by maltose, but this is due to the face that the highly truncated SNase is only minimally active, possibly due to substrate-induced folding. When compared with other chimeras, it is effectively inactive under all conditions, which makes the inconsistencies with the CD data less meaningful. MS long PHS also has a maltose-induced structural change but not when non-hydrolyzable substrate is present. This is consistent with its activity being unaffected by maltose. $M_{\Delta loop}R$ has a maltose-induced spectral shift by CD and is also inhibited by maltose. $M_{\Delta loop}R$ I25A demonstrates the most maltose-dependent remediation, but shows no structural

change by CD. Likely this is because the RNase H* domain never fully unfolds but rather assumes a partially folded state that can accommodate maltose-bound MBP and has a similar CD signal to the fully folded state.

4.6 References

1. Lad, C., N.H. Williams, and R. Wolfenden, *The rate of hydrolysis of phosphomonoester dianions and the exceptional catalytic proficiencies of protein and inositol phosphatases.* Proc Natl Acad Sci U S A, 2003. **100**(10): p. 5607-10.
2. Lockless, S.W. and R. Ranganathan, *Evolutionarily conserved pathways of energetic connectivity in protein families.* Science, 1999. **286**(5438): p. 295-9.
3. Hilser, V.J., J.O. Wrabl, and H.N. Motlagh, *Structural and energetic basis of allostery.* Annu Rev Biophys, 2012. **41**: p. 585-609.
4. Hilser, V.J. and E.B. Thompson, *Intrinsic disorder as a mechanism to optimize allosteric coupling in proteins.* Proc Natl Acad Sci U S A, 2007. **104**(20): p. 8311-5.
5. Radley, T.L., et al., *Allosteric switching by mutually exclusive folding of protein domains.* J Mol Biol, 2003. **332**(3): p. 529-36.
6. Marvin, J.S. and H.W. Hellinga, *Manipulation of ligand binding affinity by exploitation of conformational coupling.* Nat Struct Biol, 2001. **8**(9): p. 795-8.
7. Ganesh, C., et al., *Thermodynamic characterization of the reversible, two-state unfolding of maltose binding protein, a large two-domain protein.* Biochemistry, 1997. **36**(16): p. 5020-8.
8. Betton, J.M., et al., *Location of tolerated insertions/deletions in the structure of the maltose binding protein.* FEBS Lett, 1993. **325**(1-2): p. 34-8.
9. Tucker, P.W., E.E. Hazen, Jr., and F.A. Cotton, *Staphylococcal nuclease reviewed: a prototypic study in contemporary enzymology. I. Isolation; physical and enzymatic properties.* Mol Cell Biochem, 1978. **22**(2-3): p. 67-77.
10. Keck, J.L., E.R. Goedken, and S. Marqusee, *Activation/attenuation model for RNase H. A one-metal mechanism with second-metal inhibition.* J Biol Chem, 1998. **273**(51): p. 34128-33.
11. Shortle, D. and A.K. Meeker, *Mutant forms of staphylococcal nuclease with altered patterns of guanidine hydrochloride and urea denaturation.* Proteins, 1986. **1**(1): p. 81-9.
12. Dabora, J.M. and S. Marqusee, *Equilibrium unfolding of Escherichia coli ribonuclease H: characterization of a partially folded state.* Protein Sci, 1994. **3**(9): p. 1401-8.
13. Loll, P.J. and E.E. Lattman, *The crystal structure of the ternary complex of staphylococcal nuclease, Ca2+, and the inhibitor pdTp, refined at 1.65 A.* Proteins, 1989. **5**(3): p. 183-201.
14. Yang, W., et al., *Structure of ribonuclease H phased at 2 A resolution by MAD analysis of the selenomethionyl protein.* Science, 1990. **249**(4975): p. 1398-405.
15. Goedken, E.R., et al., *Divalent metal cofactor binding in the kinetic folding trajectory of Escherichia coli ribonuclease HI.* Protein Sci, 2000. **9**(10): p. 1914-21.
16. Betton, J.M. and M. Hofnung, *In vivo assembly of active maltose binding protein from independently exported protein fragments.* EMBO J, 1994. **13**(5): p. 1226-34.
17. Spudich, G.M., E.J. Miller, and S. Marqusee, *Destabilization of the Escherichia coli RNase H kinetic intermediate: switching between a two-state and three-state folding mechanism.* J Mol Biol, 2004. **335**(2): p. 609-18.
18. Connell, K.B., G.A. Horner, and S. Marqusee, *A single mutation at residue 25 populates the folding intermediate of E. coli RNase H and reveals a highly dynamic partially folded ensemble.* J Mol Biol, 2009. **391**(2): p. 461-70.

19. Chang, Y. and C. Park, *Mapping transient partial unfolding by protein engineering and native-state proteolysis.* J Mol Biol, 2009. **393**(2): p. 543-56.
20. Shortle, D., W.E. Stites, and A.K. Meeker, *Contributions of the large hydrophobic amino acids to the stability of staphylococcal nuclease.* Biochemistry, 1990. **29**(35): p. 8033-41.
21. Truckses, D.M., et al., *Coupling between trans/cis proline isomerization and protein stability in staphylococcal nuclease.* Protein Science, 1996. **5**(9): p. 1907-1916.
22. Stites, W.E., et al., *In a staphylococcal nuclease mutant the side-chain of a lysine replacing valine 66 is fully buried in the hydrophobic core.* J Mol Biol, 1991. **221**(1): p. 7-14.
23. Spudich, G. and S. Marqusee, *A change in the apparent m value reveals a populated intermediate under equilibrium conditions in Escherichia coli ribonuclease HI.* Biochemistry, 2000. **39**(38): p. 11677-83.
24. Hale, S.P., L.B. Poole, and J.A. Gerlt, *Mechanism of the reaction catalyzed by staphylococcal nuclease: identification of the rate-determining step.* Biochemistry, 1993. **32**(29): p. 7479-87.
25. Lachica, R.V., P.D. Hoeprich, and C.E. Franti, *Convenient assay for staphylococcal nuclease by the metachromatic well-agar-diffusion technique.* Appl Microbiol, 1972. **24**(6): p. 920-3.
26. Lachica, R.V., C. Genigeorgis, and P.D. Hoeprich, *Metachromatic agar-diffusion methods for detecting staphylococcal nuclease activity.* Appl Microbiol, 1971. **21**(4): p. 585-7.
27. Cutler, T.A., et al., *Effect of interdomain linker length on an antagonistic folding-unfolding equilibrium between two protein domains.* J Mol Biol, 2009. **386**(3): p. 854-68.
28. Shortle, D. and A.K. Meeker, *Residual structure in large fragments of staphylococcal nuclease: effects of amino acid substitutions.* Biochemistry, 1989. **28**(3): p. 936-44.
29. Onitsuka, M., et al., *Mechanism of induced folding: Both folding before binding and binding before folding can be realized in staphylococcal nuclease mutants.* Proteins, 2008. **72**(3): p. 837-47.
30. Chamberlain, A.K., T.M. Handel, and S. Marqusee, *Detection of rare partially folded molecules in equilibrium with the native conformation of RNaseH.* Nat Struct Biol, 1996. **3**(9): p. 782-7.
31. Raschke, T.M. and S. Marqusee, *The kinetic folding intermediate of ribonuclease H resembles the acid molten globule and partially unfolded molecules detected under native conditions.* Nat Struct Biol, 1997. **4**(4): p. 298-304.
32. Quiocho, F.A., J.C. Spurlino, and L.E. Rodseth, *Extensive features of tight oligosaccharide binding revealed in high-resolution structures of the maltodextrin transport/chemosensory receptor.* Structure, 1997. **5**(8): p. 997-1015.
33. Park, C. and S. Marqusee, *Probing the high energy states in proteins by proteolysis.* J Mol Biol, 2004. **343**(5): p. 1467-76.
34. Butler, J.S., et al., *Structural and thermodynamic analysis of a conformationally strained circular permutant of barnase.* Biochemistry, 2009. **48**(15): p. 3497-507.

APPENDIX

Multiple sequence alignment of RNases H

A1.1 FASTA alignment of RNases H used to generate tree and reconstruct ancestors

>Chloroflexi | Herpetosiphon au | YP_001547468

----------KVVLFSDGGSDG-NP-GP----GGYGVVLR---S-G-S--------
EMR-----ELTGGFAR-TTNNRMELMGVITGLQAL-------SQP------
SKVVVYSDSAYVINGMHKGWAERWSKNGWRTTTG----PVKNPDLWQQLLELAQG--H-
TI-EWVQVPGHAGVKDNERCDRLAVQAA-----HQPNL-----PIDQGYRD----------

>Firmicutes | Halothermothrix | YP_002508646

----------REMTKMEPIKVYTDGACSG-NP-GP----GGYAAVIL---NQG------
-----QER-----VVAGYEDE-TTNNRMELRAVIEALKEI-------KEG------
REVHVYSDSSYIINGMK-SWIDDWKKRGWKTSSNK---PVSNKDLWLKLDNLSSK--F-NI-
KFKKVKGHSGDEYNEKADSLARKQI-----EENS------PE----------

>Deinococcus-Thermus | Deinococcus radi | NP_294623

------MTRPGRPSARKKPDTSRDLLPIRAGIQPEVPVGGQVVELYSDGACDT-TK-GH----
GGWATILR---Y-G-E--------REL-----VLSGNEEN-TTNNRMELRGLLEGLRTL-------RRP------
CQVKVITDSQYLRKAFTDGWILNWQRNGWKTASKE---PVKNQDLWEELIELAKV--H-
AL-TFLWVKGHAGHGENERVDELAVLER-----KKLRK----------

>Deinococcus-Thermus | Deinococcus geot | YP_001527665

----------MSPPLTAVRLVTDGACSG-NP-GP----GGWACILS---S-G-A------
---STR-----ELSGGEAQ-TTNNRMELTALLEGLRAL-------KRP------
CQVHVVSDSRYIIDAFEQGWLAGWQAKGWK--------KVKNPDLWQAIAEAARG--H-TL-
TFEWVQGHAGHPENERADQLAVQAR-----EQAARQPPAPPSGPAGGLF----------

>Firmicutes | Clostridium botu | YP_001780057

----------MKKVIIYTDGACRG-N--GQENTIGAYGIVLM---Y-G-E------
--HKK-----EIKKAFRD-TTNNIMELSAVVEALSLL-------KKP------
CSIELYSDSAYVINAINQKWLDNWKKNNWKTASKS---PVKNKELWEKLDELLKK--H-
SV-KFIKVKGHSDNEYNNRCDKLANEAM-----DEFNV----------

>Bacteroidetes | Salinibacter rub | YP_445310

----------CFFGFMNEVTIYTDGACSG-NP-GP----GGWAAILLPDDD-S-D--
------ATD-----PLTGGEPH-TTNNRMELTAALEALRAL-------DDR------
SRVALHTDSEYLSKAFNEGWLDSWQDNNWQTSSND---DVKNQDLWKALLEEADR--H-
EV-DWVWVKGHADDELNIMADELAVAAM-----EQYK----------

>Bacteroidetes | Rhodothermus mar | YP_003291257

----------MSTPRKHVVIYTDGACSG-NP-GP----GGWAAILR---Y-N-Q------
----HEK-----VLTGAAPH-TTNNRMELTAVIEALRAL-------KEP------
CRVDVYTDSNYIVRAFQEGWVDRWQRNGWRTASKK---PVENQDLWRALLELTRR--H-

DV-RFLKVKGHADDALNNRVDRLAVEAM-----RRGQ------TKAAGSAVND----------
--

>Tenericutes | Candidatus Desul | YP_001717962

----------MGEQPTMSEVVIYTDGACSG-NP-GP----GGWAAVIL---D-G-V--
------ARR-----ELTGSDPK-TTNQRMELLAAIRSLQAL-------GEE-----
PRRVTLYSDSAYLVNCFRDRWYERWEQNGWVNAKKQ---PVQNRDLWEELLRLARR--
H-RV-TFRKIKGHGSNELNNRADALARGAL----PPGAR----------

>Firmicutes | Syntrophomonas w | YP_754778

----------MKEIIIYTDGACSG-NP-GP----GGWGAVLA---Y-G-E--------
HQK-----EIAGAEAD-TTNQRMELMAVIEALKAI-------KGS-----
GWEIRVYSDSAYFINAIQKGWLENWQRNGWKNSKKE---DVANQDLWKALIPLLRK--N-
RV-RVEKVKGHSGDRWNERCDQLARNAI-----KSLG----------

>Firmicutes | Natranaerobius t | YP_001917343

----------MTNTNEDKKRVTIYTDGACSG-NP-GP----GGWGAILL---F-N-E--
------HKK-----ELSGSAEN-TTNQRMELYAAVQALKAL-------KYP------
CNVELCSDSAYLVNCFQQGWWKKWQRNNWLTKSKK---KVDNQDLWRELIELNDY--H-
SI-QWIKVKGHSDDELNNRADQLATEAI-----PDKS----------

>Firmicutes | Alicyclobacillus | YP_003183894

----------MSDETVILYTDGACSG-NP-GP----GGWAAILQ---W-N-G------
--HVK-----ELSGGERE-TTNQRMELKAVIEGLKAL-------KRP------
CDVIVHSDSAYVVNCFKQRWYVNWRKNGWINSKGE---PVQNRDLWEQLLEAIDG--H-
RV-RFEKVKGHAGVKWNERCDELARSAI-------PR----------

>Chlorobi | Chloroherpeton t | YP_001996233

----------MKQVVIYTDGACSG-NP-GK----GGWGAVLI---F-G-E--------
KRR-----EISGYEAQ-TTNNRMEMMAAIQALEQL-------KEP------
CAVDLYSDSSYLVNAFNEGWIDGWLRRGWKTAGKK---PVLNQDLWQRLLGLTSS--H-
QV-TFHKVKGHSTDELNNRCDFLATEAI----KTEGA----------

>Chlorobi | Pelodictyon phae | YP_002017618

----------METKITIYTGRRIAA-QI-RA----LLSGCAVD---V-R-L--------
HDS-----SIAGYSPA-TTNNRMELSAAIEALEAL-------KEP------
CRVDLYSDSSYLVNAINEGWLKRWTINNWKTSTKK---NVENIDLWKKILKLLTL--H-TI-
TFHKVKGHSDNPYNNRCDTLAREAI-----KKKS----------

>Chlorobi | Chlorobium chlor | YP_380038

----------MKKQVTIYTDGACSG-NP-GP----GGWGALLM---F-G-S------
--ITR-----EVSGSSPA-TTNNRMELGAAIEALALL-------KEP------

CLVDLYSDSSYLVNAINNGWLQRWQRNSWQTAAKK---SVENIDLWQKLIKLLKV--H-
EV-RFHKVKGHSDNAYNNRCDQLAREAI-----KKTS----------

>Chlorobi | Chlorobium limic | YP_001942694

----------MKKRVTIYTDGACSG-NP-GR----GGWGALMM---Y-G-T----
----VNR-----ELSGYEPA-TTNNRMELTAAIEGLDAL-------KEP------
CVVDLYSDSAYLVNALNQGWLKRWTTNNWTTSAKK---SVENIDLWKKILKLVTL--H-
QV-TFHKVKGHSDNPFNNRCDELARQAI-----KNNS----------

>Chlorobi | Chlorobaculum pa | YP_001998194

----------MEKTITIYTDGACSG-NP-GK----GGWGALLM---Y-G-N------
--TRK-----EISGYDPA-TTNNRMEMMAAIRALEAL-------KEP------
CRVELYSDSAYLVNAMNQGWLKRWLKNGWKTASKK---PVENIDLWQEIVKLTTL--H-
RV-TFHKVKGHSDNQYNNRCDELARLAI-----KEQS----------

>Chlorobi | Chlorobium tepid | NP_662495

----------MEKTITIYTDGACSG-NP-GK----GGWGALLM---Y-G-S-------
-SRK-----EISGYDPA-TTNNRMELMAAIKGLEAL-------KEP------
CRVQLYSDSAYLVNAMNEGWLKRWVKNGWKTAAKK---PVENIDLWQEILKLTTL--H-
RV-TFHKVKGHSDNPYNSRCDELARLAI-----KENS----------

>Chlorobi | Chlorobium phaeo | YP_001130910

----------MEKKVTIYTDGACSG-NP-GP----GGWGAMLM---Y-G-K-----
---TVR-----EISGGAPA-TTNNRMELSAAIEALQAL-------KEP------
CTVDLYSDSSYLVNAINEGWLKRWTANRWKTAAKK---TVENIDLWQKILELTDR--H-
RV-RFHKVKGHSDNPYNNRCDELARLAV-----RKKP----------

>Chlorobi | Chlorobium luteo | YP_375504

----------MVEPFLSMQKKITIYTDGACSG-NP-GK----GGWGAMLM---Y-G-
D--------AVR-----ELSGYSPA-TTNNRMELTAAIEALRAL-------KEP------
CSVALYSDSSYVVNAFREGWLDRWTRNNWKTAAKK---NVENTDLWKQILELTAR--H-
TV-TFHKVKGHSDNPYNNRCDELARQAI-----QKKP----------

>Chlorobi | Chlorobium ferro | ZP_01385334

----------MQKKLIIYTDGACSG-NP-GP----GGWGALLM---Y-G-P-------
-STR-----ELSGYSPA-TTNNRMELTAAIEALEAL-------KEP------
CRVDLYSDSSYLVNAINEGWLKRWVVNNWKTAAKK---NVENPDLWQKILKLIRL--H-
EV-TFHKVKGHSDNPYNNRCDVLAREAI-----KKHP----------

>Chlorobi | Chlorobium phaeo | YP_911061

----------MQKKIIVYTDGACSG-NP-GK----GGWGALLM---Y-G-A------
---STR-----EISGYSPA-TTNNRMELSAAIEALETL-------KEP------

CIVHLYSDSSYLVNAINEGWLKRWTANNWKTAAKK---SVENIDLWQKILTLIKL--H-DV-
TFHKVKGHSDNPYNNRCDELARQAI-----KNNR----------

>Chlorobi | Prosthecochloris | YP_002015246

----------MRKKIIIYTDGACSG-NP-GK----GGWGALLM---F-G-E--------
LNR-----EISGYSPA-TTNNRMELMAAIQALEAL-------KEP------
CDVDLYSDSSYLVNAIKLGWLKKWSSGGWTTASRK---PVENQDLWKKILQLIKL--H-
NV-TFHKVKGHSDNEYNNRCDYLARQAI-----KNNR----------

>Chlorobi | Chlorobium phaeo | YP_001959040

----------MQKKVTIYTDGACSG-NP-GK----GGWGALLM---F-G-S-------
--VKR-----ELSGYSPA-TTNNRMELMAAIQALEAL-------KEP------
CEVALYSDSSYLVNAINKGWLKRWTSNNWKTAAKK---PVENIDLWKMILELIRL--H-
SV-TFHKVKGHSDNEFNNRCDYLATQAI-----KNNR----------

>Firmicutes | Clostridium ther | YP_001037101

----------MKKVSIYTDGACSG-NP-GD----GGWGAILI---Y-G-N--------
HEK-----EVSGFEKD-TTNNRMELVAAINALKML-------KEP------
CEVDLYSDSAYLVNGFLQNWVEKWKKNGWKTSNKE---EVKNMELWQELDRLSNI--H-
KI-RWIKVKGHSDNEYNNRCDKLATDEI-----KKNS------KK----------

>Firmicutes | Clostridium papy | ZP_05494226

----------MKQVEIYTDGACSG-NP-GA----GGWGAVLM---Y-G-E-------
--HKV-----EISGFEKS-TTNNKMELTAAFEALKRL-------KEP------
CKVNLYSDSAYLVNAFLQGWLDKWIKNGWKRNKNE---EVKNIELWKELVRLADI--H-
EI-KWIKVKGHADNVYNNRCDKLATDEI-----KKNC----------

>Firmicutes | Clostridium cell | YP_002506085

----------MKQIEIYTDGACSG-NP-GA----GGWGAVLM---Y-G-E--------
-HKI-----EISGFEKS-TTNNKMELTAAYEALKRL-------KEP------
CRVNLYSDSAYLVNAFLQGWLDKWIKNGWKRNKNE---EVKNVDLWKELVKLADI--H-
EI-KWIKVKGHADNEYNNRCDKLATDEI-----KKNS----------

>Firmicutes | Brevibacillus br | YP_002770554

----------MTMREVEIYTDGACSG-NP-GP----GGWGAVLM---Y-G-Q-------
----HIK-----EMSGAEPH-TTNNRMELMAAIKALSTL-------KEP------
CKVTLSSDSAYLVNCFKQGWYKGWLKNGWKNSKGQ---QVENQDLWKELLQLMDT--
H-KV-EYVKVKGHADNKWNNRCDELATGAI-----KQL----------

>Firmicutes | Paenibacillus sp | ZP_04854933

----------MKEVTIYTDGACSG-NP-GP----GGWGAVLM---F-N-G--------
-HRK-----DLSGGEKM-TTNNRMEIQAVISALSQL-------KEP------

CQVKVYSDSAYVVNCFQQNWIRGWLKNGWKNSKNQ---PVENRDLWEELWRLMGI--H-
KV-EYIKVKGHSDNELNNYCDQLAREAI-----KRLSS----------

>Firmicutes | Paenibacillus sp | YP_003009571

----------MKEVTIYTDGACSG-NP-GP----GGWGAVLF---Y-G-V-------
-HRK-----ELSGGEKH-STNNRMEIQAVIEALNLL-------KEP------
CKAKIYSDSAYVVNCFQKGWIHGWLRNGWKNSKKE---PVENQDLWKTLWDLMKR--H-
QV-EYIKVKGHSDNEWNNRCDELAREAI-----KRL----------

>Firmicutes | Dethiobacter alk | ZP_03729095

----------MKDVIIYTDGACSG-NP-GP----GGWGAVLR---Y-G-S--------
HEK-----EISGGDEK-TTNQRMELQAAISALELL-------KEP------
CKVKLHSDSAYLVNAFKQRWFDKWQKNGWVNSKKE---PVVNRDLWERLLELDRK--H-
DI-EWVKVKGHADDELNNRCDQLARDAV-------PR----------

>Proteobacteria beta | Moorella thermoa | YP_429492

----------MKEVTIYTDGACSG-NP-GP----GGWGAVLI---Y-G-D--------
KRK-----ELSGAEPS-TTNQRMEITAAIAALRVL-------KEP------
CRVHLYSDSAYLVNAFRQGWLARWERNGWLTVKKQ---PVENQDLWRELLQVASR--H-
QV-EWLKVKGHSDNPENNRCDELARAAI-----AALR------RQEIPSS----------

>Firmicutes | Symbiobacterium | YP_076749

----------MREVIIYTDGACSG-NP-GP----GGWGAVLL---Y-G-S--------
HRK-----ELSGFHPH-TTNNRMEIQAAIEALRAL-------KYP------
CKVKLYSDSAYLVNAFRQNWLRTWQRNGWVNSRKQ---PVENQDLWQELLEAARP--H-
QV-EWLKVQGHADVAENNRCDELARAAI-----AAGT------QG----------

>Firmicutes | Anaerocellum the | YP_002573213

----------MKEVTIYTDGACSG-NP-GP----GGWCAILI---Y-K-G--------
IKK-----VLKGFERY-TTNNRMELKAVVEALKAL-------KEP------
CKVVIYSDSAYIVNAVNQNWIEKWQKNGWKTSEKE---EVKNIDLWNELVELMKI--H-
KV-TFEKVKGHADNELNNLCDRIARSMI-----KGEQ----------

>Firmicutes | Caldicellulosiru | YP_001180679

----------MKEVVIYTDGACSK-NP-GP----GGWCAILI---Y-K-G--------
IKK-----VLKGFEEN-TTNNRMELKAIIEGLKAL-------KEP------
CKVTVYTDSAYIVNAINQNWIGKWQKNNWKTSEKE---EVKNIDLWQELLEFLKV--H-
NV-KFEKVKGHSTDTLNNMCDEIARSMI-----KEMR----------

>Firmicutes | Thermoanaerobact | ZP_05334886

----------MANIPEIDIYTDGACSG-NP-GP----GGWGAVLI---Y-N-G--------
-IKK-----EISGYEEN-TTNNRMELTAVIKALSLL-------KRS------

CKINIYSDSSYLINAFNQKWIENWQKRGWLKSDKT---PVENKDLWLKLLDLSSC--H-DI-
KWIKVKGHSDNEYNNRCDKLATDEI-----RKHSI----------

>Firmicutes | Thermoanaerobact | NP_622980

----------MKNNNEIVEIYTDGACSG-NP-GP----GGWAAVLI---Y-K-G----
----IKK-----EISGFEEN-TTNNRMELKAAIEGLKAL-------KRP------
CKVNLYSDSSYLINAFNEGWIEKWQKNNWLKSDKT---PVENQDLWKELLEVSKP--H-
QI-NWIKVKGHSDNEYNNLCDRLATEQI-----KKHI------KENP----------

>Firmicutes | Thermoanaerobact | ZP_05379402

----------MENNIDIVEIYTDGACSG-NP-GP----GGWAAVLL---Y-K-E-----
---ARK-----EISGFEEN-TTNNRMELKAVIEALKAL-------KRP------
CKVNLYSDSSYVINAFKEGWLEKWQKNNWLKSDKT---PVENQELWKELLEVSKR--H-
QI-NWIKVKGHADDEFNNLCDRLATEQI-----KRNT------KKL----------

>Firmicutes | Thermoanaerobact | YP_001665169

----------MSNNIDVVEIYTDGACSG-NP-GP----GGWAAVLL---Y-K-G----
----TKK-----EISGFEEN-TTNNRMELKAVIEGLKAL-------KRP------
CKVNLYSDSSYVINAFKEGWLEKWQKNNWLKSDKT---PVENQDLWKELLEISKN--H-
QV-NWIKVKGHADNEYNNLCDRLATEQI-----KRNT------RQNPKE----------

>Chloroflexi | Roseiflexus cast | YP_001431304

----------LLNSGKVVMFTDGCFDS-ES-GS----GGYGVILK---H-R-D------
--RTK-----EISGGFRE-TTNNRMEIRACIEGLRAL-------KRP------
SEVVIFSDSKYVVDSMSKGWVQRWKDQGWMRNEKD---QAENSDLWEQLLELCNQ--H-
RV-EFRWVKGHNHTKENERCDQLASEAA-----KRSD------LPIDRRSP----------

>Thermobaculum | Thermobaculum te | YP_003323183

----------MKKVIIHTDGGCEP-NP-GP----GGWAAVIR---Y-N-S--------
EVQ-----EISGGEEN-TTNNRMEMTAVIKALEAL-------HEP------
HEVELYTDSEYLCKGMM-EWLPMWKAKGRLQKG-----SVKNADLWQRIDELMSR--H-
LV-KCYWVKGHAGNTDNERCDKLAYEEI-----KKIYKQKGQTPPPPIQMRLIS----------
----

>Proteobacteria epsilon | Campylobacter ho | YP_001407135

----------MKSVKLFSDGSCLG-NP-GI----GGWAYILE---F-N-G--------
HEK-----CECGGEML-TTNNKMELRAAIEGLKAL-------KEP------CEVKIFTDSSYVTNSIN-
GWLEKWVAKNFK--------GKQNVELWREFLRVSAM--H-KI-
SAFWVKGHAGHPQNERCDEMARNFA-----QNLKGA----------

>Proteobacteria epsilon | Campylobacter gr | ZP_05625344

----------MKSVKLFSDGSCLG-NP-GA----GGWAYILQ---Y-G-D-------
-AIK-----KASGAEAM-TTNNQMELTAAIMGLSAL-------KQP------

CRVELFTDSEYVVKAIS-SWLAKWVATDFK--------GKKNADLWRRYLAAAAP--H-EI-
KASWVKGHAGHPQNEECDAMARAAA-----EAIKG----------

>Proteobacteria epsilon | Campylobacter re | ZP_03611429

----------MKTVCLFSDGSCLD-NP-GP----GGWAYILE---Y-G-E--------
HKK-----TASGGEAH-TTNNQMELRAAIEGLKAL-------KQP------
CRVKLYTDSSYVANAVN-AWLEGWVKKNFK--------NVKNVPLWQEYLAASEP--H-EV-
EAIWVKGHAGHPQNELCDEMAREQA-----VKIK------NSLKGE----------

>Proteobacteria epsilon | Campylobacter cu | YP_001409227

----------MKTVTLFSDGSCLN-NP-GA----GGWAYILE---F-N-G--------
AVK-----KDSGGAAM-TTNNQMELTAVIEGLKAL-------KEP------
CEVRLFTDSSYVANAVN-SWLDGWVKKNFIGSDKK---PVKNIELWQEYLRVSRP--H-KV-
TASWIKAHNGHPQNEECDTMAREKA-----TKFQ------NEADI----------

>Aquificae | Persephonella ma | YP_002730213

----------MKKVEIFTDGSSLG-NP-GA----GGWCAILR---Y-N-K--------
HEK-----MIKGGKEN-TTNNEMEIKAVLEALKIL-------KEP------CEIDLYSDSEYVVKAMK-
EWIHNWAKNNWKTSKKK---DVAHKDMWQEIYRLMQI--H-RI-
NPIWVKAHAGHRENEICDRIAKKEA-----EKFRR----------

>Proteobacteria epsilon | Helicobacter cin | ZP_03659319

----------MKQVTLYCDGSSLG-NP-GA----GGWCGILC---F-K-D--------
KQK-----ILSGGEPY-TTNNRMELLAVIESLKAL-------KEP------CVVDLYSDSKYVCDGIN-
SWLKNWVAKDFK--------NVKNVDLWQSYLQVSSL--H-SV-
TAHWVKGHAGHPQNELCDSLAKQAA-----KDVM------AKDEAVRF----------

>Proteobacteria epsilon | Helicobacter hep | NP_860229

----------MIMKQVTLYCDGSALG-NP-GA----GGWCGILS---F-G-D-------
--KQK-----ILTGGETY-TTNNRMELLAVIESLKAL-------NQP------
CIVNVYSDSRYVCNGIN-LWLKSWISKQFK--------NVKNPDLWQLYLQVSSP--H-QV-
IAHWVKGHAGVAQNELCDKLAKESA-----QFYL------NKGISDE----------

>Proteobacteria epsilon | Campylobacterale | ZP_05070331

----------MKKITLFSDGSALG-NP-GP----GGYGVILR---Y-D-D--------
KER-----EIVGSEVH-TTNNRMELLGVIEGLRAL-------SEK------CEVDIISDSSYVVKGIN-
EWLANWIKKDFK--------KVKNPDLWRDYIEVSQG--H-KI-
NAIWVRGHDGHEENERCDKLARDEA-----EKIKASL----------

>Proteobacteria epsilon | Caminibacter med | ZP_01872048

----------MKKIEIYTDGSSLG-NP-GP----GGWCAILR---Y-K-G--------
KEK-----IISGGEEY-TTNNRMELKAVIESLKIL-------KEP------CEIELYADSTYVLKGIN-

EWLSNWVRKNFK--------NVKNEDLWREFLRYSKP--H-KI-
NVNWIKGHSGHIENERCDKIAKDEA-----LRRK------SVSK-------------------------

>Proteobacteria epsilon | Nitratiruptor sp | YP_001355685

-------------------------------------MKKVSLFSDGSSLG-NP-GP----GGYCAILR---Y-K-D--------
NEK-----IIKGGEPH-TTNNRMELKAVIEGLKAL-------KEP------CIVTVYSDSNYVVQAIN-
SWLSGWIKKDFK--------NVKNPDLWKEFIEVAKP--H-RI-
KAVWVKGHSGHEENERCDKIAKEMA-----KEAGIG---------------------------------

>Planctomycetes | Rhodopirellula b | NP_869340

-----------------------------MTDSKTEAAFKPVELYTDGACSG-NP-GP----GGWAFVLR---
CPRTL--------KEI-----QRSGGQPH-TTNNQMELMAVIRGLEAL-------KEP------
CAVDLYSDSKYVGQGMS-
SWMAGWKSRGWKRKDGSKLVPVKNVELWQELDQQMQA--H-RV-
TYHHVKGHAGHTENELCDKLAVAAY-----QQYL-----------------------------------

>Proteobacteria alpha | Ehrlichia rumina | YP_196682

----------------------------MSSFICCMKDELNKVVVYTDGACSG-NP-GP----GGWGAVLL---F-
D-N--------GEK-----TICGGHPN-TTNNRMELTAVVQALKFL-------DVT------
YVIDLYTDSVYVKSGIT-SWIKKWKINGWRTADKL---PVKNLELWLELDKIVKY--H-KI-
TWYWVKAHSGNLYNEKADMLARSQI-------VK----------------------------------

>Proteobacteria alpha | Ehrlichia chaffe | ZP_00544752

---------------------------------MKDELNKVVIYTDGACSG-NP-GP----GGWAAVLL---F-D-D---
-----NEK-----TICGNDSD-TTNNRMELTAVIEALKLL-------KVA------
YNVDLYTDSVYVKDGIT-LWIRKWKVNGWKTANKM---PVKNLELWLELDSLANF--H-
KV-TWYWVRAHVGDLYNQKADMLARSQI-------VR----------------------------------

>Proteobacteria alpha | Ehrlichia canis | YP_303391

---------------------------------MKDELNKVVIYTDGACSG-NP-GP----GGWGAILL---F-D-K----
----NER-----TICGNNPD-TTNNRMELTAVIEALKFL-------KVA------
YNVDLYTDSIYVKDGIT-LWIEKWKINGWRTASKL---PVKNLELWLELDSLASF--H-NV-
TWYWVKAHAGNLYNQKADILARSQI-----SK----------------------------------

>Proteobacteria alpha | Wolbachia endosy | ZP_03335577

----------------------------MLNLYPVNMDKKKVIIYTDGACSG-NP-GP----GGWAAVVM---Y--
-E---NKSVFIKK-----RISGGEEN-TTNNKMELKAVINGLKML-------KIS------
CKVIVHTDSQYIKQGIT-EWINKWKTNGWKTADKK---PVKNRELWQELDEVALQ--H-DI-
NWKWVRAHNGNMYNEEADRLARKES-----
KNLKYRDCEVKKSPKNRGNSKFHRLGGVLWQ----------

>Proteobacteria alpha | Wolbachia endosy | YP_001974908

------------------------------------MKKKEVTIYTDGACSG-NP-GT----GGWAAIIL---F-Q-N-------HRK-----NICGREEN-TTNNKMELTAVINGLKVL-------KFP------CNISLYTDSLYIKYGIT-EWINKWKMNGWKTSNKK---SVKNIELWKELDNAALQ--H-EI-NWNWVKAHNGDKYNEEADILARKAI-----INA----------

>Proteobacteria alpha | Anaplasma centra | YP_003328923

----------MSLYYVRYWNTIKNDGRMVLMGKSRVAIYTDGACSG-NP-GP----GGWGAVLR---F-GDG--------EER-----RISGGSDD-TTNNRMELTAVIMALAAL-------SGP-------CSVCVNTDSTYVKNGIT-EWIRKWKLNGWRTSSKS---AVKNVDLWMELERLTLL--H-SI-EWRWVKAHAGDEYNEKADMLARGEA-----ERRM------VAPK----------

>Verrucomicrobia | Verrucomicrobium | ZP_02925152

----------MESVLPQVIIHTDGGCLG-NP-GV----GGWAAVLE---SCG----------RRK-----EISGGEPA-TTNNRMELRAAIEALSHL-------KKT------CAVEMHTDSQYVRNGIT-KWLAGWKKNGWKTASKQ---RVKNEDLWSTLDAAAQR--H-QV-SWHWVKGHAGHDDNERCDQLCGEAM-----EAVKKQHTRQQLAAALVAFKDTGR----------

>Tenericutes | Candidatus Liber | YP_003065209

----------MDSKHLREVHAYTDGACSG-NP-GP----GGWGVLLR---Y-K-G-------KEK-----IISGGEKE-TTNNRMELMAAIKALTAL-------KYP------CKVLLYTDSSYVHKGFS-QWIKKWQQNGWKTSDKK---TVKNIDLWMKFVEASAQ--H-KV-DLYWIKGHAGNQENEKVDRIARNAA-----VSFKNKI----------

>Tenericutes | Candidatus Solib | YP_827887

------------------------------------MKKVQLITDGACLG-NP-GP----GGWSAILR---F-E-E--------QKK-----ELWGCEKQ-TTNNRMELTAAIEGLRAL-------REK------CQVEVVTDSEYVLKGIT-TWIDGWKRKGWMTAAKK---PVINQDLWKLLDEQVNR--H-QA-TWTWTKGHASHADNNRCDELATRAA-----REQS------KS----------

>Firmicutes | Faecalibacterium | ZP_05614459

----------MIRFPLDRTGKIALNRGIQKFKFREKQMKQVEVYTDGACSG-NP-GP----GGWGAVLR---Y-RFN-----GKVYEK-----ELSGGDAS-TTNNRMELTAFIEALRQL-------KEP------CEVRLCSDSQYVINGLEKGWARGWKRRGWKKSDGS---PALNPDLWEQALEQEAR--H-KI-TYVWVKGHAGHPENERCDQLAVAQSQAHGGRQGR----------

>Deinococcus-Thermus | Thermus aquaticu | ZP_03496041

----------MSLPLKRVDLFTDGACLG-NP-GP----GGWAALLR---Y-G-S----------QEK-----LLSGGEPC-TTNNRMELRAALEGLLAL-------REP------CQVHLHTDSQYLKRAFAEGWVERWQRNGWRTAEGK---PVKNQDLWQALLKAMEG--H-EV-AFHFVEGHSGHPENERVDREARRQA-----KAQPQVPCPPKEATLF----------

>Deinococcus-Thermus | Thermus thermoph | YP_144822

---------MNPSPRKRVALFTDGACLG-NP-GP----GGWAALLR---F-H-A---------HEK-----LLSGGEAC-TTNNRMELKAAIEGLKAL-------KEP------
CEVDLYTDSHYLKKAFTEGWLEGWRKRGWRTAEGK---PVKNRDLWEALLLAMAP--H-RV-RFHFVKGHTGHPENERVDREARRQA-----QSQAKTPCPPRAPTLFHEEA----------

>Actinobacteria | Gordonia bronchi | YP_003275599

---------MTESDSAGAPVVEISTDGACLG-NP-GP----GGWGAVLR---Y-R-G--------TEK-----RISGGEPN-STNNKMELTAAIEGLAAL-------TRP------
STVILYTDSTYVRNGIT-KWVKGWQRNGWKTADKK---PVKNADLWRRLVEEEKV--H-TV-EWRWVKGHAGDQYNEIADELATTAA-----RQIA--------DSGKVAG----------

>Proteobacteria delta | Syntrophobacter | YP_844984

----------MRGRRRMPETPAIRKHVEIFADGACRG-NP-GP----GGWGAVLR---YHG----------KEK-----ELSGYAEY-TTNNQMELAAVIQALRAL-------KEP------
CRVTITTDSRYLRDGIS-LWIHKWKQNGWKTRVKT---DVRNKELWIALDEACLP--H-EI-DWQWVKGHSGHPENERCDALARAAI-----DRHL------REAATEE----------

>Proteobacteria delta | Desulfohalobium | YP_003197359

---------MSETSVVRLYTDGACLG-NP-GP----GGWAAVLL---YGG-E---------ARK-----ELSGGYAK-TTNNRMEMLALIEGLKVL-------KRP------
CRVKVWTDSRYLHDGLTKGWLQKWQKNGWKTAAKK---PVKNKDLWQELAALTSR--H-QL-ELHWVRGHSGDPENERCDVLAKAAA-----NQPG-----LAKDPGHE----------

>Proteobacteria gamma | Xylella fastidio | ZP_00683618

----------YEIDHAYTDGSCLG-NP-GP----GGWAVLLR---Y-K-N--------NEK-----ELVGGELD-TTNNRMELMAAIMALERL-------SEP------CQIKLHTDSQYVRQGIT-EWMSGWVRRGWKTAAGD---PVKNRDLWERLCAATQR--H-MV-EWCWVKAHNGDSDNERVDVLARGQA-----MAQR------STVASR----------

>Proteobacteria gamma | Stenotrophomonas | YP_002027232

----------MKTIEIHTDGSCLG-NP-GP----GGWAALLR---Y-K-G--------HER-----ELSGGEAH-TTNNRMELMAAISGLETL-------TEP------CDIVLYTDSQYVRQGLT-QWMPGWIRKNWKTAGGD---PVKNRELWERLHAATLR--H-QI-DWRWVKGHSGDPDNERVDTLARNAA-----IQIR------DSSPVN----------

>Proteobacteria gamma | Xanthomonas oryz | YP_199680

----------MKSIEVHTDGSCLG-NP-GP----GGWAALLR---Y-N-G--------REK-----ELAGGEAV-STNNRMELMAAIMALETL-------TEP------CEIVLHTDSQYVRQGIT-EWMPGWVRRNWKTAGGD---PVKNRELWERLHAATQR--H-RI-DWRWVKGHNGDPDNERVDVLARNQA-----TAQR---------DGRATS----------

>Proteobacteria gamma | Xanthomonas camp | NP_636365

------------------------------------MKSIEVHTDGSCLG-NP-GP----GGWAALLR---Y-N-G--------
REK-----ELAGGEAN-STNNRMELMAAIMALETL-------TEP------CQILLHTDSQYVRQGIT-
EWMPGWVRRGWKTSGGD---PVKNRELWERLHAATQR--H-SI-
EWRWVKGHNGDPDNERVDVLARNQA-----IAQR------GGLATS-----------------------

>Verrucomicrobia | Chthoniobacter f | ZP_03131837

------------MEIAASSGRAKRNILPVAPPSRYNDIILKKVTIHTDGACEG-NP-GP----
GGWAAILE---Y-G-A--------VRK-----EISGGVIA-TTNNRMELTAALEALNRL-------KER-----
-CAVDLFTDSEYLRNGIT-KWIFGWKAKGWK---KG---TIKNIDLWQALDAAASR--H-KV-
EWHWVRGHAGHPLNERCDVLAVQET-----
QKFRQSHTNAERKAARAAFLAERVGVPEQPELSSSLLK---

>Actinobacteria | Streptomyces sp. | ZP_05480970

---------------------------------MAEQTEEAVEIYTDGACSG-NP-GP----GGWGALLR---Y-G-K--
------HER-----ELYGAEDTVTTNNRMELMAPIRALESL-------TRA------
SVVRIYTDSTYVRNGIL-QWMPRWKKNGWQTQAKQ---PVKNADLWQRLDTACRQ--H-
EV-EWLWVKGHAGLPENERADKLAVKGS-----QEAA------AAGVRRARG----------------
-

>Proteobacteria zeta | Mariprofundus fe | ZP_01453598

-----------------------------------MTEKPVVLAFTDGACSG-NP-GP----GGWGVLLR---M-G-K---
-----HEK-----EIYGGEAE-TTNQQMELQAAVEALKAL-------KQP------
CKITVISDSKYVVQGMN-EWIHNWKKKGWKTVGKK---PVSNLERWQELDTLAAR--H-
EV-QWQWVKGHAGHVENERADELARRGI-------PA---------------------------------

>Proteobacteria gamma | Francisella phil | YP_001678245

--------------------------------MGIFTKKNNVIAYTDGACKG-NP-GI----GGWGAILS---Y-N-G---
-----VDK-----EISGAEKD-TTNNRMELMAAIKTLQAL-------KRK------
CDITIYTDSKYLQNGIN-QWLANWKANGWKTAAKK---EVKNKDLWQELDSLTTK--H-
NV-TWSWVKGHSGNQGNEKADELANKAI-----AELT------GK---------------------------

>Proteobacteria beta | Nitrosomonas eut | YP_748363

--------------------------------MQLKSDMKRVEIFTDGACKG-NP-GP----GGWGVCLH---F-N-G-
-------ETR-----EFFGGEPV-TTNNRMELLAAIRALQELESLEDNGQQH------
LQVQLHTDSQYVQKGIS-EWIHGWKKRGWRTADKK---PVKNEALWRELDDLSQR--H-
QV-EWFWVRGHNGHAGNERADRLANQGV-----ESVL------SKKAD-----------------------

>Proteobacteria beta | Nitrosospira mul | YP_412312

---------------------------------MKAKLAEVVEIFTDGACKG-NP-GV----GGWGALLQ---Y-N-G-
-------HRR-----ELFGGEKM-TTNNRMELLAVIRALEAL-------TKP------
CEVRLHTDSLYVQKGIS-EWIHAWKKRDWRTADKK---PVKNDDLWRELDLLTQR--H-
KI-EWLWVRGHSGHDGNEYADMLANRGV-----QTAL------RGVSAN-----------------------

>Proteobacteria beta | Nitrosomonas sp. | ZP_05315423

----------MIGNISKVVEIYTDGACKG-NP-GI----GGWGALLR---Y-G-D----
----HER-----EIFGGEKL-TTNNRMELLAAIRALESL-------KRP------
CKIHLHTDSQYLQKGIS-EWLDSWKARNWCTADKK---PVKNEDLWKLLDQLTQQ--H-
EI-EWCWVRGHSGHIDNERADQLANRGV-----EMII------SE----------

>Proteobacteria alpha | Erythrobacter li | YP_457199

----------MKKVEIFTDGACKG-NP-GP----GGWGVLLR---M-G-K------
--HEK-----ELSGGEPE-TTNNRMELRAAIEGLNAL-------IEP------CEVELYTDSKYVVDGIT-
KWVHGWKKRGWVNASKK---PVRNDDLWHDLIEAELR--H-KV-
TWHWVKGHNGHAENERADRLASEAA-----DLQS----------

>Proteobacteria alpha | Erythrobacter sp | ZP_01864562

----------MKKVEIFTDGACKG-NP-GP----GGWGALLR---M-G-R-------
-HEK-----ELSGGEPD-TTNNRMEMTAAIRALSAL-------IEP------CEVALHTDSKYLIDGIT-
KWVHGWKKRGWVNASKK---PVRNADLWHELIELTAR--H-KV-
DWFWVKGHSGHPENDRVDQLASDAA-----ERIA---------AGEAL----------

>Proteobacteria alpha | Erythrobacter sp | ZP_01038775

----------MIGQSRPMKHVEIFTDGACKG-NP-GP----GGWGALLR---L-G-K-
-------HEK-----ELSGGEAD-TTNNRMELTAAIEGLRAL-------IEP------
CKVDLYSDSKYVIDGIT-KWVHGWKKRGWVNASKK---PVRNSDLWHDLIDVTSR--H-
EV-SWHWVKGHSGHTENERVDQLASDEA-----DRVA------RGE----------

>Proteobacteria alpha | Zymomonas mobili | YP_163336

----------MPDSSTQDKIVMIATDGACKG-NP-GF----GGWGALLR---Y-Q-
G--------HEK-----AISGSENP-TTNNRMELQAVIEALSCL-------KKP------
CQIELSTDSKYVMDGLT-RWIHGWQKNGWLTAAKK---PVKNADLWKQLLALTRQ--H-
DI-AWKWVKGHAGHPDNERADQLASDAA-----IALM--QQE-KA----------

>Proteobacteria alpha | Sphingopyxis ala | YP_616183

----------MSERRTVIVATDGACKG-NP-GP----GGWGAVLR---W-G-E---
-----VVK-----TLSGGEAD-TTNNRMELMAAIEALAAL-------KRP------
CNVELSTDSVYVRDGIT-KWIFGWQKNGWKTAAKK---PVANADLWQRLIKEAAR--H-
KV-EWLWVKGHAGHGDNELADQLASDAA--LKMARAR----------

>Proteobacteria alpha | Novosphingobium | YP_496364

----------MKHVEIFTDGACKG-NP-GK----GGWGALLR---M-G-E------
--HEK-----EMAGSEKE-TTNNRMELMAAIRALEAL-------KQP------
CRVTLHTDSKYVLDGIT-KWIFGWQKKGWKTADNK---PVKNEDLWRALVDAVRP--H-
KV-EWVWVKGHDGHPENERVDKLASDAA-----LAA----------

>Proteobacteria alpha | Sphingomonas wit | YP_001263184

----------MAELPLVEIATDGACKG-NP-GR----GGWGALLR---F-G-A-----
---TEK-----EMSGAENP-STNNRMELMAAIRALEAL-------KKP------
CRVKLSTDSRYVMDGLT-KWIHGWRKNGWKTADKK---PVKNAELWQRLLDAAAP--H-
RI-EWIWVKGHAGHPDNERADKLASDAA-----LGL----------

>Actinobacteria | Frankia alni ACN | YP_714042

----------MAQRDGRVAVDIHTDGACSG-NP-GP----GGWGAVLR---Y-G-
E--------HER-----ELHGGEPARTTNNRMELTAAIMALEAL-------TRP------
SVVRLHTDSTYLRSGIT-TWIAGWRRNGWLTKDRT---PVRNADLWQRLEAAVAR--H-
EV-EWLWVRGHAGDPGNERADALAARGL-----QEAR------QTPPPA----------

>Proteobacteria gamma | Rickettsiella gr | ZP_02061806

----------MLKIPKIEIFTDGACRG-NP-GP----GAWAALLR---FQG---------
-KEK-----TLSGTEAS-TTNNRMELMAAIQALIAV-------KKP------CRIILSTDSKYVQKGIT-
EWLPQWKRRAWLTANKK---PVKNSDLWKELALQAER--H-QI-
SWEWVKGHSGHPENDRVDYLANVAL-----DKLL--GSF----------

>Proteobacteria gamma | Coxiella burneti | ZP_01947020

----------MAKQEQNIVYLYCDGACRG-NP-GP----GGWGVLLR---Y-N-Q-
-------HER-----QLHGGVAN-TTNNQMELTAAIEGLKSL-------KKP------
CQVVVTTDSQYLRRGIT-EWLPVWKRRGWRTSNKK---PVKNQPLWETLEREVER--H-TI-
VWHWVKGHSGHAENEIADELANRGI-----DEVL------KRGAR----------

>Proteobacteria alpha | Rhodospirillum r | YP_428136

----------MSAAAGDEIKRVRVDMFTDGACSG-NP-GP----GGWGTILR---W-
G-D--------TEK-----ELWGGETP-TTNNRMELMAVIRGLEAL-------RRP------
VTVTIHTDSRYVHDGIT-GWIHGWKRNGWKTAAKK---PVKNEDLWRRLDAALGT--H-
DI-SWQWVRGHSGHVENERADELARRGT-----SEAR--QGK-VDGQSSTIL----------
-

>Actinobacteria | Jonesia denitrif | YP_003160869

----------MNNQSTSDDSDRATVTMWTDGACKG-NP-GV----GGWGVWMT---
S-G-A--------HTK-----ELFGGENH-TTNNRMELMAVIEGLRAL-------KRP------
CDVNLHVDSTYVMKGIT-SWIHGWKRNGWRTADKK---PVKNAELWRELDDQVTR--H-
RV-TWTWVKGHSGDVGNDKADELANKGV-----ELVR--STT-SSTPTEPPSRHMPATKE-----
----

>Actinobacteria | Cellulomonas fla | ZP_04368013

----------MSPVSKTDDLPVVEMWTDGACKG-NP-GV----GGWGAWMR---F-
G-D--------QER-----ELWGGEAA-TTNNRMELSAVIEGLRAL-------KRP------
CRVTLHVDSTYVMNGLQ-KWLPNWKRNGWRTGDKK---PVKNQELWQALDTEVQR--H-
HV-TWVWVKGHAGDPGNERADALANRGV-----DDVR------AGVR----------

>Actinobacteria | Sanguibacter ked | YP_003315226

----------MTQHSPTTTEPSTEPSTDPSTEADSAVEIWTDGACKG-NP-GV----
GGWGAWLR---A-G-G--------HER-----ELFGGETV-TTNNRMELTAVIEALRAL-------KRP---
---CVVNLHVDSTYVMNGMS-KWIAGWKRNGWRTGDKK---PVKNVDLWQALDEQVAR-
-H-TI-TWTWVKGHSGDVGNEKADELANRGV-----AEVR--ARG------------------------------

>Proteobacteria beta | Thiomonas interm | ZP_05498609

----------MTTESDNEIIIYTDGACKG-NP-GP----GGWGVVLR---S-G-A-----
---HEK-----TLHGGEPQ-TTNNRMELMAAIMALEAL-------KRP------
SRVLLHTDSQYVLKGMT-EWIVGWKRRGWTTADKK---PVKNVDLWQRLEKAAAP--H-
TL-RWVWVRGHTGDPGNEQADALANQGV-----EAAG------RG----------------------------

>Proteobacteria beta | Methylobacillus | YP_545588

----------MSSNVIEIYADGACKG-NP-GP----GGWGAWLS---F-A-G------
--HEK-----ELWGGELV-TTNNRMELTAVIRALEAL-------KRQ------
CSVRIYTDSVYVQKGIT-EWVHSWKARNWLTSDRK---PVKNVDLWKALDSLVQQ--H-
QV-EWVWVKGHAGNVGNERADALANKGV-----DQVL------GREVV-------------------------

>Proteobacteria beta | Methylovorus sp. | YP_003051221

----------MAVEEGCVVIYADGACKG-NP-GP----GGWGAWLA---M-G-
G--------HEK-----EMCGGELL-TTNNRMELTAVIRALQAL-------KRP------
CQVKIYTDSVYVQKGIT-EWMTGWKARNWRTSDKK---PVKNEDLWRELDQTVQP--H-
NI-EWLWVKGHAGNAGNERADALANQGV-----LQAL------EAKERA-------------------------

>Proteobacteria beta | Bordetella avium | YP_787429

----------MSTANPPADLVEMWTDGACKG-NP-GP----GGWGVLMR---Y-
G-S--------HEK-----TFFGGDPQ-TTNNRMEILAVVEGLRAL-------KRA------
CTVVIHTDSQYVMKGMT-EWLPNWKRRGWLTADKK---PVKNAELWQLLDAQVAR--H-
EV-RWQWVRGHNGDPGNEMADMLANQGV-----ASVA------RN----------------------------

>Proteobacteria beta | Bordetella petri | YP_001629234

----------MMQTDSNEDGPQVEMWTDGACKG-NP-GP----GGWGVLMR---
A-G-A--------HEK-----TLHGGEAG-TTNNRMELLAVIEGLRTL-------KRP------
CQVVIHTDSQYVMKGMT-EWLANWKRRGWLTADKK---PVKNAELWQALDEQVAR--
H-KV-SWRWVRGHAGDPGNERADALANLGV-----ESLR------KRRAGA-----------------------
-

>Proteobacteria beta | Bordetella parap | NP_885986

----------MQNLEGSGDGQQVEMWTDGACKG-NP-GP----GGWGVLMR---
A-G-Q--------HEK-----TMHGGERQ-TTNNRMELMAVIEGLRAL-------KRP------
CRVTIHTDSQYVMKGMT-EWLANWKRRGWRTADKK---PVKNVELWQALDEQVGR--H-
QV-QWRWVRGHAGDPGNERADALANQGV-----EAAR------GR----------------------------

>Proteobacteria beta | Polynucleobacter | YP_001797659

------------------------------MLHTKSSHHQPHIVIYTDGACKG-NP-GP----GGWGAVLR---S-G-G--------HEK-----HIHGGEKL-TTNNRMEICAVIFALKAL-------KQS------
STVELWTDSQYVQKGVT-EWLEGWKKRGWKTASKD---PVKNADLWQELDTLIPD--H-DI-SWHWVRGHDGHPGNELADQLANKGV-----EEFL--P--------------------------------

>Proteobacteria beta | Polynucleobacter | YP_001155804

------------------------------MPHSKHPSSHPHIIIYTDGACKG-NP-GP----GGWGAVLR---S-G-S--------HEK-----HIHGGEKL-TTNNRMEICAVIFALKAL-------KQR------
SSVELWTDSQYVQKGVT-EWLEGWKKRGWKTASKD---PVKNADLWQELDTLLPD--H-DI-SWHWVRGHNGHPGNELADALANKGV-----EEFL--P--------------------------------

>Tenericutes | Candidatus Accum | YP_003169109

-----------------------------------MTDEVVITIFADGGCRG-NP-GP----GGWGVVLQ---A-G-E--------HEK-----ELWGGEPD-TTNNRMEMTAAIRALEAL-------KRP------
ASVRLHTDSQYLQKGIS-EWIHNWKRNGWRTADKK---PVKNADLWQRLDELAGE--H-RI-QWCWVKGHAGHSGNERADALANRGM-----DELQ--RTA-NRRPLAGDGS----------

>Proteobacteria beta | Neisseria mening | YP_002343102

-------------------------------------MNQTVYLYTDGACKG-NP-GA----GGWGVLMR---Y-G-S--------HEK-----ELFGGEAQ-TTNNRMELTAVIEGLKSL-------KRR------
CTVIICTDSQYVKNGME-NWIHGWKRNGWKTAAKQ---PVKNDDLWKELDALVGR--H-QV-SWTWVKGHAGHAENERADDLANRGA-----AQFS---------------------------------

>Proteobacteria beta | Neisseria mucosa | ZP_05977102

-------------------------------------MDDTVYLYTDGACKG-NP-GA----GGWGVLMR---Y-R-N--------HEK-----ELCGGEAE-TTNNRMELTAVIEGLKAL-------KRP------
CRVVICTDSQYVKNGME-GWIHGWKKNGWKTAAKK---PVKNDDLWKELDALSHK--H-EL-QWTWVKGHAGHSENEKADALANQGA-----ARFL------QSSS--------------------------

>Proteobacteria beta | Neisseria flaves | ZP_04758492

------------------MPIFQTALCYDTALFDKDMPMDKPVYLYTDGACKG-NP-GA----GGWGVFMR---Y-G-T--------HEK-----ELFGGEAE-TTNNRMELTAVIEGLKSL-------KRR--------CQVVICTDSQYVKNGME-SWIHGWKKNGWKTAAKK---PVKNDDLWKELDSLVQQ--H-DV-RWTWVKGHAGHPENEKADELANQGA-----AKFA---------------------------------

>Tenericutes | Candidatus Nitro | ACE75583

---------------------------------------MIEIYTDGACSG-NP-GP----GGWGALLR---I-D-N--------AET-----EMCGGDPA-TTNNRMELLAVIEALQSL-------TQP------
VEARVYTDSQYVQKGIS-EWIHSWKRRGWKTAGKE---PVKNEDLWRRLDTLASG--H-KL-EWHWVRGHNGHPENERVDALARAGL-----EQSR--RAG-KTVGGTRQSLF----------

>Proteobacteria beta | Thiobacillus den | YP_315421

----------MTADIIYIYSDGACKG-NP-GA----GGWGALLV---A-G-G-------HRK-----EISGGEPN-TTNNRMEMTAVIRALELL-------KRP------
STVEVHTDSQYVQKGVS-EWLPGWKRRNWRTADGK---PVKNQDLWQQLDALSQQ--H-RI-VWKWVRGHAGHPENERADVLANQGV----------LQARQY----------

>Proteobacteria beta | Rhodoferax ferri | YP_522728

----------MNAVEIYTDGACKG-NP-GP----GGWGAFLK---S-A-D-------SQK-----ELFGGELG-TTNNRMEMTAVIEALAAL-------KRP------
CQVTLHVDSQYVLKGMT-EWLAGWKARGWKTAAKQ---
PVKNVDLWQRLDELVSTSGH-RI-DWRWVRGHNGDPGNEHADMLANRGV-----ELAL-------RQR----------

>Proteobacteria beta | Curvibacter puta | CBA29144

----------MTEVLTKVVVYTDGACKG-NP-GP----GGWGVLLR---SAD-G-------TEK-----ELFGGELG-TTNNRMEMMAVIEALSAL-------KRP------
CQITLHIDSQYVLKGIT-EWLQGWKAKGWKTASKQ---PVKNVDLWQRLDALVSGAGH-TI-DWRWVKGHAGDPGNERADGLANRGVSWHCVNAPE------CCPAAHHPPM----------

>Proteobacteria beta | Variovorax parad | YP_002944233

----------MNEVVIYTDGACKG-NP-GP----GGWGAWLK---S-G-A-------TEK-----ELFGGELN-TTNNRMELTAVIEGLAAL-------KRP------
CKVILYLDSQYVRMGIT-EWIRGWKAKGWRTSTKQ---PVKNVELWQKLDKLVAEGGH-VI-EWRWVKGHSGDVGNERADMLANKGV-----DKAL------GRI----------

>Proteobacteria beta | Polaromonas sp. | YP_549102

----------MTDTQAGTTTQTQVVIYTDGACKG-NP-GP----GGWGVLLA---M-G-D--------TEK-----ELFGGEPV-TTNNRMEMTAVIEALAAL-------KRP------
CRVTLYLDSEYVRKGIT-EWIHGWKARGWRTAAKA---PVKNVDLWQRLDALVTSSGH-KI-DWRWVKGHNGDPGNERADALANQGV-----ERAL------GRR----------

>Proteobacteria beta | Polaromonas naph | YP_981914

----------MTDSAAEPTISQPQHVVIYTDGACKG-NP-GP----GGWGALLA---S-G-G--------TEK-----EIFGGEMG-TTNNRMEMTAVIEALAAL-------KKP------
CTVTLYLDSQYVLKGIT-EWIHGWKARGWRTAAKA---PVKNVDLWQRLDALLVSSGH-SI-DWRWVRGHNGDPGNERADALANKGV-----ERAL------GRL----------

>Proteobacteria beta | Acidovorax delaf | ZP_04763296

----------MNQIEIYTDGACKG-NP-GP----GGWGALLR---A-G-A-------TEK-----ELFGGELG-TTNNRMELMAVIEALSAL-------KRP------CAVTLYLDSEYVRKGIT-EWIHGWKARGWRTAAKQ---PVKNVELWQRLDALVTTAGH-RI-DWRWVRGHSGDPGNERADALANRGV-----DKAL------GRG----------

>Proteobacteria beta | Verminephrobacte | YP_995394

---------------------------------------MNQVEIYTDGACKG-NP-GP----GGWGVLLR---S-G-P-------TEK-----ALFGGALG-TTNNRMELMAVIEALSAL-------QRP------
CAVTLYLDSEYVRKGIT-EWIHGWKAKGWRTAARQ---PVKNVDLWQRLDALVSTGGH-RI-EWRWVKGHSGDPGNERADALANRGV-----DQAL------GRGALASAE----------

>Proteobacteria beta | Comamonas testos | YP_003277678

---------------------------------------MNQVVIYTDGACKG-NP-GP----GGWGALLQ---A-G-S-------AQK-----ELFGGELG-TTNNRMELKAVIEALSAL-------KRP------
CDVVLYLDSQYVRKGIT-EWIQGWKAKGWVTASKE---PVKNVELWKQLDALVQGSGH-RI-DWRWVKGHAGDPGNERADALANKGV-----ELAL------KKG----------

>Proteobacteria beta | Delftia acidovor | YP_001565909

---------------------------------------MNQVVIYTDGACKG-NP-GP----GGWGVVLE---S-G-S-------ARK-----ELFGGELN-TTNNRMEMMAVIEALSAL-------RRP------
CDVVLYIDSQYVLKGIT-EWIHGWKAKGWKTASKE---PVKNVELWQRLDALVQGGGH-RI-DWRWVKGHAGDPGNERADALANKGV-----DQAL------GR----------

>Proteobacteria beta | Acidovorax sp. J | YP_986009

---------------------------------------MNQVVIYTDGACKG-NP-GP----GGWGAVLR---S-G-T-------LEK-----ELFGGELG-TTNNRMELMAVIQALGAL-------KRP------
CQVALYLDSQYVRQGIT-EWIHGWKKKGWRTAAGQ---
PVKNVELWQRLDELAHQAGH-RI-EWHWVRGHAGDPGNERADMLANKGV-----EQVL-------GR----------

>Proteobacteria beta | Acidovorax citru | YP_970999

---------------------------------------MNQVVIYTDGACKG-NP-GP----GGWGVVLR---S-G-A-------LEK-----ELFGGELG-TTNNRMELLAVIEALGAL-------KRP------
CAVTLYLDSQYVRKGIT-EWIQGWKKKGWRTASGQ---PVKNVELWKRLDDLVAGGGH-VI-DWRWVKGHAGDPGNERADALANKGV-----DKAL------GRA----------

>Proteobacteria beta | Leptothrix cholo | YP_001791002

------------------------MSIESTAGPQAVIAKPEVVIYTDGACKG-NP-GP----GGWGAWLV---S-G-G--------HEK-----ELCGGEAN-TTNNRMEMMAVIEALASL-------KRS------
CRITVYTDSAYVQNGIS-SWIHGWKRRGWKTADNK---PVKNVDLWQRLDALSTL--H-QI-EWRWVKGHAGDPGNERADALANRGV---EVARAR----------

>Proteobacteria beta | Aromatoleum arom | YP_160719

-----------------------------------MTDQIEIFTDGACSG-NP-GP----GGWGAILR---S-G-A--------HEK-----EIWGGEPH-TTNNRMELLAVIRALELL-------KRP------VVARVHTDSQYVQKGIS-EWIHGWKARGWKTAAKA---PVKNEDLWRALDEAASR--H-QV-
QWVWVRGHAGHVENERADELARRGV-----DAVR------RQGAAVAG----------

>Proteobacteria beta | Azoarcus sp. BH7 | YP_933559

---------------------------------------MEEVDIYTDGACSG-NP-GP----GGWGAILR---S-N-G--------
HEK-----EIWGGEPQ-TTNNRMELIAVIRALEAL-------KRP------VAARVHTDSQYVQKGIS-
EWIHGWKARGWKTASKE---PVKNADLWRTLDEVAGR--H-QV-
KWLWVRGHAGHVENERADALARRGA-----EAAR------KQGTVVTN----------

>Proteobacteria beta | Thauera sp. MZ1T | YP_002355924

---------------------------------------MEEVDIYTDGACSG-NP-GP----GGWGAILR---S-G-S--------
HEK-----EIWGGEPA-TTNNRMELLAVIRALDAL-------KRP------
VAARVHTDSQYVQKGIS-EWIHGWKARGWKTASKE---PVKNADLWRALDDAASR--H-
QV-KWLWVRGHNGHPENERADALARRGV-----DAVR------KSGAAVQC----------
-

>Proteobacteria beta | Oxalobacter form | ZP_04579209

---------------------------------------MNEVEIYTDGACRG-NP-GP----GGWGVWMI---A-G-G-------
--HEK-----ELFGGDAD-TTNNRMELMAVIEALRAL-------KRP------
CKVVLHTDSQYVQKGIS-EWIHKWKARGWRTADKK---LVKNVDLWMELDQARAQ--H-
DI-DWRWIKGHAGHEGNEKADQLANKGV-----DSVL----------

>Proteobacteria beta | Oxalobacter form | ZP_04577078

---------------------------------MDSEKMSEVEIYTDGACRG-NP-GP----GGWGVWLR---A-N-G-
-------HEK-----ELFGGDAD-TTNNRMELTAVIEALRVL-------KRP------
CRVVLHTDSQYVQKGIT-EWIHKWKERGWRTSDRK---LVKNVDLWMELDEATRR--H-
DI-RWRWVKGHAGHEGNEKADQLANRGV-----DSVL----------

>Proteobacteria beta | Burkholderia phy | YP_001857103

-----------------------------------MSSDLIEIFTDGACKG-NP-GP----GGWGALLR---Y-G-T-------
-QEK-----ELFGGEAN-TTNNRMELMAVIAALEAL-------KRP------
CKAVVHTDSQYVQKGIS-EWIHGWKKKGWVTAARA---PVKNADLWKRLDALTQQ--H-
QL-EWRWVKGHAGHPENERADALANRGV-----ASLA--DL----------

>Proteobacteria beta | Burkholderia gra | ZP_02886594

-----------------------------------MTANIIDIYTDGACKG-NP-GP----GGWGALLR---F-G-D-------
-QEK-----ELFGGEAN-TTNNRMELMGVISALEAL-------KRP------
CKAVVHTDSQYVQKGIS-EWIHGWKKKGWVTAAKQ---PVKNADLWKRLDALVAQ--H-
EI-EWRWVRGHNGHPENERADQLANRGV-----ASLA--EL----------

>Proteobacteria beta | Burkholderia glu | YP_002911003

-----------------------------------MTLQLIDIYTDGACKG-NP-GP----GGWGALLR---F-G-D-------
--QEK-----ELFGGEAG-TTNNRMELLAVIRALEAL-------KRP------
CRVIVHTDSQYVQKGIS-EWIHGWKKKGWVTAAKT---PVKNADLWKQLDALVGQ--H-
EI-EWRWVKGHAGHAENERADALANRGV-----ESLS------QRA----------

>Proteobacteria beta | Burkholderia tha | ZP_02463089

-----------------------------------MTLQTIDIYTDGACKG-NP-GP----GGWGALLR---Y-G-T------
--QEK-----ELFGGEAG-TTNNRMELTAVIAALAAL-------KRP------
CKVVVHTDSQYVQKGIS-EWIHGWKKKGWVTAAKT---PVKNADLWQRLDALVAQ--H-
DV-EWRWVKGHAGHPENERADALANRGV-----ESLA------QA----------

>Proteobacteria beta | Burkholderia amb | ZP_02889023

-----------------------------------MTTDTIDIYTDGACKG-NP-GP----GGWGALLR---Y-G-D------
--REK-----ELFGGEPN-TTNNRMELMGVIGALEAL-------KRP------
CRVIVHTDSQYVQKGIS-EWIHGWKKKGWVTAAKT---PVKNADLWKRLDALVAQ--H-
EI-EWRWVKGHAGHPENERADALANRGV-----ESLV--A----------

>Proteobacteria beta | Burkholderia ubo | ZP_02379873

-----------------------------------MTTDTIDIYTDGACKG-NP-GP----GGWGALLR---Y-G-D------
--REK-----EMFGGEPN-TTNNRMELMAVIASLEAL-------KRE------
CRVVVHTDSQYVQKGIS-EWIHGWKKKGWVTAAKT---PVKNADLWKRLDALVAQ--H-
QV-EWRWVKGHAGHPENERADALANRGV-----ESLA--A----------

>Proteobacteria beta | Lutiella nitrofe | ZP_03697564

-----------------------------------MTQDIVEIYPDGACKG-NP-GP----GGWGVLLR---F-K-G------
--REK-----ELFGGEQG-TTNNRMELTAVIEGLAQL-------KRP------
CKVAVYTDSQYVQKGIS-EWIHGWKKRGWKTAAKE---PVKNADLWQKLDALQAG--H-
QI-SWHWVKGHAGHEFNERADQLANRGV-----ETLS--A----------

>Proteobacteria gamma | Chromobacterium | NP_900926

-----------------------------------MTTEDRVEIYTDGACKG-NP-GP----GGWGALMR---Y-K-G---
-----KEK-----ELFGGERG-TTNNRMEIMAVIRALAAL-------NRP------
CKVVVYTDSQYVQKGIS-EWIHGWKARGWKTAAKE---PVKNADLWQQLDAERNR--
HLDV-EWRWVKGHAGHEFNERADQLANKGV-----ESV----------

>Proteobacteria beta | Janthinobacteriu | YP_001352901

-------------------------------------MDKIDIYSDGACKG-NP-GR----GGWGALLV---M-G-E-------
-REK-----EIFGGELD-TTNNRMELKAVIEALNLL-------TRP------CEVVVHTDSQYVQKGIS-
EWIHGWKARGWKTAAKA---PVKNVDLWQALDAAQAR--H-KI-
EWRWVRGHNGHAGNERADALANRGV-----EVAA----------

>Proteobacteria beta | Herminiimonas ar | YP_001100572

-------------------------------------MEKIDIFTDGACKG-NP-GR----GGWGALLV---M-G-E-------
-REK-----ELFGGEPG-TTNNRMELKAVIEALNAL-------TRP------CEVIVHTDSQYVQKGIS-
EWIHGWKARGWKTAARA---PVKNVDLWQALDAAQAR--H-QI-
EWRWVRGHNGHVGNERADALANRGV-----ETVN--SN----------

>Proteobacteria beta | Ralstonia picket | YP_002981688

---------------------------------------MQEVTVYSDGACKG-NP-GL----GGWGTVLV---S-G-S------
--HEK-----ELFGGEAL-TTNNRMELMAVIEAFRAL-------KRP------
CRVQVYTDSQYVQKGIS-EWLAGWKARGWKTADKK---PVKNDDLWRTLDELVAG--H-
EV-SWHWVKGHAGHPGNERADALANKGV-----EMAR------QAKA-------------------------

>Proteobacteria beta | Ralstonia metall | YP_584356

---------------------------------------MQEVTIYSDGACKG-NP-GP----GGWGAVLV---A-G-G------
--HEK-----ELFGGESP-TTNNRMELMAVIEALRAL-------KRP------CIVNIYTDSQYVQKGIS-
EWIHGWKARGWKTADKK---PVKNADLWQALDEAQKP--H-QI-
TWHWVRGHNGHPGNERADALANRGV-----ASIN--T---------------------------------

>Proteobacteria beta | Ralstonia eutrop | YP_296396

---------------------------------------MQEVIIYSDGACKG-NP-GR----GGWGAVLV---A-G-T-------
-NEK-----ELFGGEAN-TTNNRMEMTAVIEALRAL-------KRP------
CTVQVYTDSQYVQKGIS-EWLPGWKARGWKTADKK---PVKNADLWQELDTLVQP--H-
KI-TWHWVRGHNGHPGNERADALANRGV-----ASLA--S---------------------------------

>Proteobacteria beta | Cupriavidus taiw | YP_002006001

---------------------------------------MQEVTIYSDGACKG-NP-GR----GGWGAVLV---A-G-T------
--SEK-----ELFGGEPN-TTNNRMEMTAVIEALRAL-------KRP------
CVVRVYTDSQYVQKGIS-EWLPGWKARGWKTADKK---PVKNADLWQALDTLAQA--H-
QI-SWHWVRGHNGHPGNERADALANRGV-----ESIG--R---------------------------------

>Proteobacteria gamma | Cardiobacterium | ZP_05706412

-----------------------------------MSTPLLIYTDGACKG-NP-GI----GGWGVLMC---Y-G-E-------
-HRK-----TLNGAEAM-TTNNRMELTAAIEALRAV-------KRA------
CPIVLTTDSSYVKNGIT-QWLAGWKRNGWKTADKK---AVKNVDLWQALDALVAQ--H-
QI-EWQWIKGHSGHPGNEMADQLANEAI-----AELR------AKG--------------------------

>Actinobacteria | Tsukamurella pau | ZP_04025800

----------------------------------MIIVADEIVIYTDGACLG-NP-GP----GGWGAVLR---F-G-E------
--HTK-----ELYGAEKD-TTNNRMELMGAISALEAI-------TKP------
FPVVLYTDSSYVKNGIT-KWVEGWKRNGWKTANKQ---PVKNVELWQRLDEVAAR--Y-
EI-DWRWVKGHAGNEGNELADQLASRGA-----AEAR--DS-------------------------------

>Proteobacteria gamma | Beggiatoa sp. PS | ZP_02002985

-----------------------------------MNESIVEAFTDGACRG-NP-GP----GGWGVLLR---C-Q-N------
--EEK-----QLYGGELN-TTNNRMELMAAIMALESL-------TRS------
NHIRLTTDSEYVKKGIT-EWIENWIKRGWKRANNE---PVKNIDLWQRLHAVTQK--H-QV-
DWQWIKGHSGHSENEQADSLANQGI-----DSVV------QS--------------------------

>Proteobacteria beta | Limnobacter sp. | ZP_01915744

----------MYADGACKG-NP-GP----GGWGVFLQ---S-G-D--------
HAK-----ELCGGELN-TTNNRMELTAVIEGLNAL-------KKR------
CSIDVYTDSQYVRKGVL-EWMPKWKMNGWKTSDKK---PVKNADLWQILDEASVR--H-
LV-RWHWVKGHSGNPGNEKADALANLGV-----EKAM------KQ----------

>Proteobacteria gamma | Reinekea blanden | ZP_01115650

----------MKTVTLYTDGGCRG-NP-GP----GGWGAVLI---Y-G-D--------
-HEK-----KLKGSEPE-TTNNRMELLAAIEGLEAL-------KQA------
VTVDLYTDSKYVQQGIT-QWIHNWKKNGWKTAGKK---PVKNQDLWQRLDSLMSK--H-
EV-NWHWVKGHAGHKYNEIADELANQAM-----DEMV------RQ----------

>Proteobacteria alpha | Oceanicaulis ale | ZP_00957665

----------MAENTIVIHTDGACSG-NP-GP----GGWGAILH---W-K-G--------
--HEK-----ELSGAEAE-TTNNRMELMAAIAALEAL-------KRR------
STVRLVTDSTYVRDGVT-KWIHGWKRNGWKTAAKK---PVKNDDLWKRLDAIASK--H-
DV-TWEWVKGHAGHPENERADQLARDAI-----ATLS------KG----------

>Proteobacteria alpha | BAL199 | ZP_02189003

----------MGSDVAAERVAIFTDGACSG-NP-GP----GGWGAVMC---W-R-
G--------TEK-----ELSGAEPL-TTNNRMELMAAIASLEAL-------SRR------
VPVDLTTDSTYVRDGIT-KWMKAWKARGWKTADKK---PVKNQDLWERLDAAAKA--H-
DV-AWHWVKGHAGHPENERADELARMAI-----AAMR------EAAR----------

>Proteobacteria alpha | Maricaulis maris | YP_757447

----------MSTITIHTDGACSG-NP-GP----GGWGAILE---W-N-G--------
HRK-----ELKGGEAD-TTNNRMEMMAAIQALEAL-------RKA-----
DRSVILITDSVYLRDGIT-KWIHGWKKRGWKTADKK---PVKNVDLWQRLDELTRS--H-
TI-DWRWVKGHAGDPGNERADELAREGL-----AEAR------GRQP----------

>Actinobacteria | Nocardiopsis das | ZP_04333609

----------MRNGDEVGQEPTQRVVIYTDGACSG-NP-GP----GGWGVWLR---Y-
G-G--------HEK-----ELYGGEAQ-TTNNRMELMAAIRALESL-------RQP------
LPVLVHTDSSYVRNGIT-SWLHGWKRRGWRTADKK---PVKNVDLWQRLDEVASR--Y-
EV-EWRWVRGHSGDEGNERADALARRGR-----DEAA------GV----------

>Proteobacteria beta | Gallionella ferr | ZP_04831427

----------MNSEIVEIFTDGACKG-NP-GV----GGWGALLR---S-K-G--------
-VQR-----ELFGGEAH-TTNNRMELMGAISALEAL-------TRR------
CQVKLHTDSKYVLQGIT-TWLAGWKRAGWKTSSRQ---PVKNEDLWRRLDALVIQ--H-
EI-EWVWVKGHSGHAGNEHADELANRGV-----AMIQ------EQANGML----------

>Proteobacteria gamma | Halothiobacillus | YP_003263698

----------MTIELKKGSADQQNADAQPATQTVHIWTDGACKG-NP-GP----
GGWGALLR---Y-G-D--------TER-----ELCGGEAH-TTNNRMELLAAISALEAL-------KRP----
--CTVHLTTDSQYVRQGML-EWLPNWRKKNWRRADGQ---PVKNADLWARLDEAAQR--
H-DM-HWHWIKGHAGHPENERADQLANQGT------PKG----------

>Proteobacteria alpha | Bartonella bacil | YP_988727

----------MLSETKVIEIYTDGACSG-NP-GL----GGWGAILR---W-N-S-----
---HER-----ELYGGKEY-TTNNQMELMAAICALNAL-------KES------
CSIDLYTDSVYVRNGIS-LWLENWKKNNWRTASKS---PVKNMELWQALDGACAR--H-
NV-RWHWVKGHAGHPDNERADALARKAI-----TEYR--------QNGYFKG----------

>Proteobacteria alpha | Bartonella graha | YP_002971483

----------MATQQKVVEIYTDGACSG-NP-GI----GGWGAILR---W-N-G---
-----HER-----ELYGGKVH-TTNNQMELMAAICALKAL-------KEP------
CLVDLYTDSVYVRNGIS-KWIEDWKKNNWRTASKN---PVKNMELWQALEDACSC--H-
TV-RWHWVKGHAGHPENERADALARKAI-----SQYR-------ENGRFPA----------

>Proteobacteria alpha | Bartonella tribo | YP_001609086

----------MASQQKVVEIYTDGACSG-NP-GV----GGWGAILR---W-N-G--
------HER-----ELYGGNAH-TTNNQMELMAAICALKAL-------KEP------
CLVDLYTDSVYVRNGIS-KWIEGWKKNNWRTASKS---PVKNMELWQTLEDACSC--H-
AV-RWHWVKGHAGHPENERADALARKAI-----AQYR-------ENGRFPT----------

>Proteobacteria alpha | Bartonella hense | YP_033281

----------MLHQQKVVEIYTDGACSG-NP-GV----GGWGAILR---W-N-G--
------HER-----ELYGGEVQ-TTNNQMELMAALCALKAL-------KES------
CSVDLYTDSVYVRNGIS-LWLKGWKKNNWQTVSKK---PVKNKELWQALEGVCSF--H-
TI-RWHWIKGHTGHPDNERADALARKAI-----AEYR-------ENGCFSA----------

>Proteobacteria alpha | Bartonella quint | YP_032046

----------MLNQQKVVEIYTDGACSG-NP-GV----GGWGAILR---W-N-G--
------HER-----ELYSGEVQ-TTNNRMELMAAICALKVL-------KEA------
CSVDLYTDSVYVRNGIS-LWLERWKMNNWRTTSKK---TVKNIELWKALEDVCSL--H-TI-
RWHWVKGHAGHPDNERADALARKAI-----TEYR--------KNGYFSA----------

>Proteobacteria gamma | Hahella chejuens | YP_433754

----------MKTVEIYTDGACKK-NP-GP----GGWGAILI---Y-G-K--------
NEK-----EIYGGELD-TTNNRMELMAAIEALRAL-------KQG------
CKVELYTDSQYVRKGIT-EWMQNWIKKGWRTSGGD---PVKNVDLWQALDKERNK--H-
DI-SWRWVKGHSGHPLNERADELANLGV-----KEAL-------GETG----------

>Proteobacteria alpha | Rhodospirillum c | YP_002299808

----------MSAEKLLVDIYTDGACSG-NP-GP----GGWGAILR---W-K-G----
----TEK-----ELKGGERL-TTNNRMELMAAIQALEAL-------KRP------
VTVRLHTDSQYVKNGIT-TWIHGWKKNGWKTAGRD---PVKNADLWQRLDELVGR--H-
TV-EFHWVKGHAGHPENERADQLAREGM-----RDTL------AAPAA----------

>Proteobacteria gamma | Sideroxydans lit | ZP_05337703

----------MSDVVEIFTDGACKG-NP-GL----GGWGALLR---V-K-G------
--KEL-----ELCGGEAH-TTNNRMELLAAISALEAL-------KRQ------
CRVRLHTDSKYVQQGIS-EWVHNWKLRGWKTADKK---PVKNEDLWRRLDTLAEQ--H-
HV-EWVWVKGHAGHDGNERADALANRGC-----ADVE------KHLKKS----------

>Proteobacteria alpha | Methylocella sil | YP_002361589

----------MPKPVVIFTDGACSG-NP-GP----GGWGAVMT---F-G-D------
--HLK-----ELCGGEAA-TTNNRMELMAAIMALEAL-------TRP------
CAVQLVTDSNYVKGGVT-TWLAGWKRNGWRTADKK---PVKNVDLWQRLEAAAEA--
H-AI-EWRWVKGHAGDELNERADALARLGM-----APFL------SARKAATP----------
-

>Proteobacteria alpha | Parvibaculum lav | YP_001411964

----------MSGEDIVEIYTDGACSG-NP-GP----GGWGVLMI---Y-K-D------
--REK-----ELCGGEQA-TTNNRMELMAAIQALEAL-------KRD------
AHVRIHTDSNYVKDGIT-KWIHGWKKNGWKNAAKQ---PVKNAELWRRLEAAIST--H-
QV-SWHWVKGHSDHPENDRADALARQGM-----APYL------PSK----------

>Proteobacteria gamma | HTCC5015 | ZP_05062252

----------MTDVTLYTDGACKG-NP-GP----GGWGVLLI---Y-G-G-------
-HEK-----ELCGGEAE-TTNNRMELMAAIEGLNAL-------KRS------
CRVALYTDSNYVRQGMT-QWLANWKKNGWRTAAKK---PVKNDDLWQALDAACER--
H-EI-EWHWVKGHSGDPGNERADELANRGV----LSAQA----------

>Proteobacteria alpha | Hyphomicrobium d | ZP_05377708

----------MTAEAPKILIYSDGACSG-NP-GP----GGWGAVLI---S-G-K------
--HRK-----EISGGEVL-TTNNRMELLAAISALEAL-------KKR------
SEVALYTDSAYVKNGIT-GWVHGWKKNGWRTADKK---PVKNVELWQALDALRNK--H-
DV-EWHWLKGHAGHPENERADELARQAM-----APFK--LAP-RPDASTKI----------
-

>Proteobacteria alpha | Asticcacaulis ex | ZP_04769350

----------MKKVTIYTDGACKG-NP-GK----GGWGAILT---F-G-P--------
HEK-----ELYGFEAE-TTNNRMELMAVIMALEAL-------KEP------
CEIDVHADSQYVLKGIK-EWIHGWKARGWKTADKK---PVKNDDLWIRLDAARQR--H-
KI-HWHWVKGHAGHEMNERADGLANKAI-----TEAA------AAF----------

>Proteobacteria alpha | Caulobacter sp. | YP_001686013

------------------------------------MTPKLVIYTDGACRG-NP-GP----GGWGALLM---Y-G-D------
--KKK-----EIMGGDLA-TTNNRMELMAAIQALEAL-------NKP------
TKAELHTDSQYVMKGVT-QWIHGWKAKGWKTADKS---PVKNVDLWQRLDAARAR--H-
EV-DWRWVKGHAGHVHNERADELARLGM-----LKTL------GERGSGKAV----------
--

>Proteobacteria alpha | Phenylobacterium | YP_002131727

------------------------------------MTPEVVIYTDGACSG-NP-GP----GGWGAILI---H-G-E--------
REK-----ELCGGEAA-TTNNRMELMAAIQALEAL-------KRP------
CRVELHTDSQYVQKGIH-EWIHGWKKRGWLTADKK---PVKNDDLWKRLDAARLR--H-
HV-DWRWVKGHAGHELNERADALARKGL-----SEAA------AARAAGGA----------
-

>Proteobacteria alpha | Caulobacter segn | ZP_06122411

------------------------------------MTPKVTIYTDGACKG-NP-GP----GGWGAILF---Y-G-D-------
-KKK-----EICGGEPG-TTNNRMELMAAIQALELL-------NRP------
CKVELHTDSQYVMKGIQ-EWIRGWKARGWKTADKS---PVKNDDLWKRLDAARAR--H-
DV-DWRWVKGHAGHPLNERADALANEGL-----RQAN------PRFG----------

>Proteobacteria alpha | Brevundimonas su | ZP_06170982

-------------------------------------MSHVIIHTDGACKG-NP-GP----GGWGAIIQ---F-G-E--------
KAK-----EMSGGEPL-TTNNRMELTAAIMALEAL-------TRP------
CKIDLHTDSKYVMDGIT-GWIHGWKARGWKTADKK---PVKNDDLWKRLDVARTR--H-
EV-KWHWVKGHAGHALNERADQLANRGI-----EEMR------AAKAKA----------

>Proteobacteria alpha | Brevundimonas sp | ZP_05034100

-----------------------------------MSPTDHVIIHTDGACKG-NP-GP----GGWGALLQ---TGG-G-----
----HEK-----ELWGGEPN-TTNNRMELMAAIMALEAL-------KRP------
CRVELHTDSKYVMQGIT-EWMRGWKARGWLTADKK---PVKNADLWQRLDAARLK--H-
DV-KWRWVKGHAGHELNERADQLANRGV-----ADLR------RV----------

>Proteobacteria alpha | Rhodobacterales | ZP_01741948

-------------------------------------MPDLFAYTDGACSG-NP-GP----GGWGVLLI---A-K-N---
ADKVLREK-----ELCGGEQE-TTNNRMELMAAISALENL-------SRP------
STLTIITDSVYVKNGVT-QWVHGWKRNGWKTASKK---PVKNEELWKRIDEAQAR--H-
QV-TWKWIKGHAGHEENERADELARRGM-----APFK------K----------

>Proteobacteria alpha | Rhodobacterales | ZP_01011925

-------------------------------------MVDLIAHTDGACSG-NP-GP----GGWGVLMQ---A-K-D---
GGTVVKER-----TLSGGEPA-TTNNRMELMAAIMALETL-------ERA------
SKITIVTDSAYVKNGVT-GWIHGWKRNGWRTANKK---PVKNVELWQRLDEAAKR--H-
DV-EWRWIKGHAGHEENERADELAREGM-----APFK----------

>Proteobacteria alpha | Rhodobacter sp. | ZP_05842972

---------------------------------------MTDLFAYTDGACSG-NP-GP----GGWGVLMQ---A-R-D---
GAVVVKER-----TLSGGEAD-TTNNRMELMAAISALEAL-------KRD------
AGIVIVTDSAYVKNGVT-TWMTGWKRNGWKTADRK---PVKNVDLWLRLDEAQAR--H-
KV-EWRWIKGHAGHEENERADELARAGM-----APFK------KPKG--------------------------

>Proteobacteria alpha | Rhodobacter spha | YP_001044401

---------------------------------------MPDLYAYTDGACSG-NP-GP----GGWGVLML---A-R-E---
GEAVVKER-----TLQGGEVL-TTNNRMELMAAISALEAL-------TRP------
TEITIVTDSAYVKNGVT-TWIHGWKRNGWKTADRK---PVKNAELWERLDAAQQR--H-
KV-VWRWIKGHAGHAENERADELARAGM-----APFK------TR----------------------------

>Proteobacteria alpha | Oceanicola granu | ZP_01157443

---------------------------------------MAELFAYTDGACSG-NP-GP----GGWGAVLI---A-R-E---
GGAVLKER-----ELSGGEAR-TTNNRMELMAAISALEAL-------ERP------
SRLTMVTDSNYVKDGIT-SWIAGWKRRGWKTAAKK---PVKNEDLWRRLDEAAAR--H-
QV-TWEWVKGHAGHPENERADELARAGM-----APFK------R----------------------------

>Proteobacteria alpha | Thalassiobium sp | ZP_05343008

---------------------------------------MPDLIAYTDGACSG-NP-GP----GGWGALMI---A-R-D---
GDTVLKKR-----ELKGGEAH-TTNNRMELLGAINVLETL-------AKP------
SVITIVTDSAYVKGGIT-EWIFGWKRRGWKTSTKK---PVKNEDLWKRLDEVTQR--H-TV-
TWEWVKGHAGHPENERADELARAGM-----EPFK------PAK--------------------------

>Proteobacteria alpha | Loktanella vestf | ZP_01002514

---------------------------------------MADLYAYTDGACSG-NP-GP----GGWGALLI---A-R-D---
GDKVVKER-----ALSGGEAD-TTNNRMELLAAISALETL-------GRA------
TAITIVTDSAYVKDGIT-SWIHGWKRRGWKTSANK---PVKNEDLWRRLDSAVAQ--H-
QV-RWEWVKGHAGHVENERADELARAGM-----APYK------P----------------------------

>Proteobacteria alpha | Roseobacter sp. | ZP_01749501

---------------------------------------MPDLFAYTDGACSG-NP-GP----GGWGALLV---A-R-D---
GDKVLKER-----ELCGGEAD-TTNNRMELLAAISALETL-------DRS------
TALTIVTDSSYVKDGIT-QWIHGWKARGWKTAAKK---PVKNEDLWKRLDEVTAR--H-
DV-TWEWVKGHAGHPENEKADELARAGM-----EPFK------P----------------------------

>Proteobacteria alpha | Jannaschia sp. C | YP_508444

---------------------------------------MPDLVAYTDGACSG-NP-GP----GGWGALMR---A-K-D---
GDTILKER-----ELKGGEAD-TTNNRMELLAAISALEAL-------DRP------
STLTIITDSAYVKNGIT-GWMHGWKRNGWKTSTRK---PVKNVDLWQRLDEAQSR--H-
TV-TWEWIKGHAGHEGNEKADELARAGM-----APFK------TGKRGKDG----------------------

>Proteobacteria alpha | Paracoccus denit | YP_917380

---------------------------------------MNALFAWTDGACSG-NP-GP----GGWGVLMR---A-M-D---
GDRMLKER-----ELSGGEAE-TTNNRMELMAAISALEAL-------TRP------
SEITVTTDSAYVKNGVT-QWIHGWKKNGWRTADRK---PVKNADLWQRLDAAQAR--H-
QV-RWEWIKGHAGHPENERADELARAGM-----APFK------PARVSG----------

>Proteobacteria alpha | Dinoroseobacter | YP_001531532

---------------------------------------MPELFAYTDGACSG-NP-GP----GGWGALLI---A-R-D---
GDTVVKER-----ALKGGEAE-TTNNRMELLAAIHALEAL-------ERP------
ARLTVVTDSAYVKGGVT-GWIHGWKRNGWKTSTKK---PVKNEDLWRRLDAAQAR--H-
EV-QWEWVKGHAGHPENERADALAREGM-----APFK------PGKSKAGR----------

>Proteobacteria alpha | Rhodobacterales | ZP_05076900

---------------------------------------MAKLLAYTDGACSG-NP-GP----GGWGVLMR---A-M-D---
GDEIVKHR-----ELSGGAEL-TTNNQMELMAAISALEVL-------ERA------
SELTIITDSTYVKNGVT-GWIHGWKKNGWKTSAKK---PVKNVELWQRLDAAQAR--H-
QV-TWEWVKGHAGHPENERADELARAGM-----APFK------KTA----------

>Proteobacteria alpha | Roseobacter sp. | ZP_05102097

---------------------------------------MAKLIAYTDGACSG-NP-GP----GGWGALMR---A-M-E---
DGKIVKER-----ELKGGEAA-TTNNRMELMAAISALEAL-------ARP------
TEITIVTDSNYVKNGIT-NWIHGWKKNGWKNAAKK---PVKNAELWQRLDAANAR--H-
SV-TWKWVKGHAGHPENERADELARAGM-----APFK------GK----------

>Proteobacteria alpha | Oceanibulbus ind | ZP_02152227

--------------------MTATRRAHKPRWKLPETTMPDLYAYTDGACSG-NP-GP----
GGWGVLMR---A-M-D---GDKIVKER-----ELKGGEGQ-TTNNRMELMAAISALESL-------
SRT------TEITIVTDSNYVKNGIT-GWIFGWKKNGWKNAAKK---
PVKNAELWQRLDAANAR--H-NV-TWKWVKGHAGHPENERADELARAGM-----APFK----
--PGGKK----------

>Proteobacteria alpha | Roseovarius nubi | ZP_00960363

---------------------------------------MVDLVAYTDGACSG-NP-GP----GGWGVLMQ---A-K-R---
GAEVIKQR-----ELSGGEAL-TTNNQMELMAAITALETL-------EKP------
STITIVTDSQYVKNGVT-GWIFGWKKNGWKTSAKK---PVKNVELWQRLDAAQAR--H-
KV-TWEWVKGHAGHPENERADELAREGM-----APYK------PKAAK----------

>Proteobacteria alpha | Roseobacter lito | ZP_02141318

---------------------------------------MPELFAYTDGACSG-NP-GP----GGWGVLLQ---A-K-E---
GDRLVKER-----ALKGGEAH-TTNNRMELLAAINALESL-------SRA------
STITIVTDSNYVKNGIT-GWIHGWKRNGWKNAAKK---PVANAELWQRLDEANAR--H-
DV-TWKWVKGHAGHAENERADELARAGM-----APFK------P----------

>Proteobacteria alpha | Roseobacter sp. | ZP_01055839

---------------------------------------MADLYAYTDGACSG-NP-GP----GGWGALLQ---A-K-D---
GGSVIKEK-----ELKGGEAN-TTNNRMELLAAINALESL-------DRP------
SALTVVTDSNYVKNGIT-GWIFGWKKNGWKNAAKK---PVKNAELWQRLDAAQSR--H-
QV-TWEWVKGHAGHPENERADELARAGM-----APFK------KSKSKA----------

>Proteobacteria alpha | Roseovarius sp. | ZP_01878558

---------------------------------------MPALFAYTDGACSG-NP-GP----GGWGVLMR---A-M-D---
GDAILKER-----ELSGGEAD-TTNNRMELWAAIAALEAL-------SRP------
STITIVTDSAYVKNGVT-GWMHGWKRNGWRTADKK---PVKNVELWQRLDEAQKR--H-
TV-TWEWVKGHAGHPENERADELARAGM-----APYK------PSKAKG----------

>Proteobacteria alpha | Sagittula stella | ZP_01744336

---------------------------------------MPDLYAYTDGACSG-NP-GP----GGWGVLLR---A-M-E---
GDEVVKQR-----ELKGGERV-TTNNQMELMAAISALESL-------TKP------
SRITVITDSQYVKNGVT-GWIFGWKKNGWKTAAKK---PVKNVELWQRLDAAQAR--H-
DV-VWEWVKGHAGHPENERADELARAGM-----APFK------S----------

>Proteobacteria alpha | Oceanicola batse | ZP_00998823

---------------------------------------MPDYFAYTDGACSG-NP-GP----GGWGVLLQ---A-K-D---
GETVLKER-----DLKGGEAA-TTNNRMELLAAINALEAL-------GRS------
TAITIVTDSAYVKNGVT-GWIHGWKRNGWKTAAKK---PVKNADLWQRLDEAQAR--H-
DV-TWQWVKGHAGHPENERADELARAGM-----APFK----------

>Proteobacteria alpha | Ruegeria pomeroy | YP_168415

---------------------------------------MPELFAYTDGACSG-NP-GP----GGWGVLLR---A-I-E---
GETVLKER-----ELCGGEAE-TTNNRMELLAAINALETL-------ERP------
SKITVVTDSAYVKNGVT-GWIFGWKRNGWKTAGKK---PVKNVELWQRLDLAQAR--H-
DV-TWKWVKGHAGHPENERADELARAGM-----KPFK------PKKARA----------

>Proteobacteria alpha | Roseobacter sp. | ZP_01901987

---------------------------------------MPDLFAYTDGACSG-NP-GP----GGWGVLLQ---A-I-E---
GDTVLKER-----ELSGGEAE-TTNNRMELLAAINALETL-------AKP------
SKITIVTDSAYVKNGVT-GWIHGWKRNGWKTAARK---PVKNVELWQRLDEAQAR--H-
DV-TWEWVKGHAGHPENERADALARAGM-----APFK------PSA----------

>Proteobacteria alpha | Rhodobacteraceae | ZP_05122674

---------------------------------------MPDLFAYTDGACSG-NP-GP----GGWGVLLR---A-M-D---
GETVLKER-----ELKGGEAE-TTNNRMELLAAISALETL-------ERA------
SDITIVTDSAYVKNGVT-GWIFGWKRNGWKTSNKK---PVKNVDLWQRLDEAQAR--H-
QV-TWEWVKGHAGHPENERADELARAGM-----APFK------PGKAQA----------

>Proteobacteria alpha | Silicibacter lac | ZP_05786713

---------------------------------------MPDLFAYTDGACSG-NP-GP----GGWGVLLR---A-V-D---
GETVLKER-----ELNGGEAE-TTNNRMELLAAISALEAL-------ERP------
SKITIVTDSAYVKNGVT-GWIHGWKRNGWKTASRK---PVKNVDLWQRLDEAQQR--H-
DV-TWEWVKGHAGHPENERADELARAGM-----APFK------PKKART----------

>Proteobacteria alpha | Ruegeria sp. TM1 | YP_614567

---------------------------------------MPDLFAYTDGACSG-NP-GP----GGWGALLR---A-M-D---
GETVLKER-----ELKGGEKE-TTNNRMELLAAIHALESL-------ARP------
SKITVVTDSAYVKNGVT-GWIFGWKKNGWKTSAKK---PVKNVELWQRLDAAQSR--H-
DV-TWEWVKGHAGHPENERADELARAGM-----APFK------SSGKSSKG----------

>Proteobacteria alpha | Citreicella sp. | ZP_05782407

---------------------------------------MPELFAYTDGACSG-NP-GP----GGWGVLMR---A-M-N---
GEDIVKER-----ELKGGEAD-TTNNRMELLAAINALESL-------TRP------
TTITVVTDSAYVKNGVT-GWIHGWKRNGWNTAAKK---PVKNAELWQRLDEAQRM--H-
SV-TWKWVKGHAGHPENERADELARAGM-----APFK------PGGK----------

>Proteobacteria alpha | Ruegeria sp. R11 | ZP_05089769

---------------------------------------MAELFAYTDGACSG-NP-GP----GGWGALLR---A-M-D---
GDTVIKEK-----ELKGGEAE-TTNNRMELLAAIHALESL-------ARP------
STITVVTDSAYVKNGVT-GWIHGWKRNGWKTASKK---PVKNVELWQRLDEAQRR--H-
TV-TWEWVKGHAGHPENERADELARAGM-----APFK------QGKTKA----------

>Proteobacteria alpha | Phaeobacter gall | ZP_02145641

---------------------------------------MAELFAYTDGACSG-NP-GP----GGWGVLLR---A-M-D---
GETIVKEK-----ELSGGEAE-TTNNRMELLAAINALENL-------ARP------
STLTVVTDSAYVKNGVT-GWIHGWKRNGWKTASKK---PVKNVELWQRLDEAQRR--H-
TV-TWEWVKGHAGHPENERADELARAGM-----APFK------PGKAKA----------

>Proteobacteria alpha | Rhodobacterales | ZP_05077818

---------------------------------------MPELFAYTDGACSG-NP-GP----GGWGVLLR---A-M-D---
GDSIVKEK-----ELSGGEAE-TTNNRMELLAAINALESL-------ARP------
STITVVTDSAYVKNGVT-GWIFGWKKNGWKTSNKK---PVKNVELWQRLDEAQRR--H-
KV-TWEWVKGHAGHPENERADELARAGM-----APFK------KKKGA----------

>Proteobacteria alpha | Roseobacter sp. | ZP_01753735

---------------------------------------MPDLYAYTDGACSG-NP-GP----GGWGVLLR---A-M-D---
GEAIIKEK-----ELQGGEAE-TTNNRMELLAAINALESL-------ARS------
STITVVTDSAYVKNGVT-GWIFGWKKNGWKTAAKK---PVKNVELWQRLDEAQSR--H-
RV-TWEWVKGHAGHPENERADELARAGM-----APYK------KSKG----------

>Proteobacteria alpha | Rhodopseudomonas | YP_484963

----------MSGALSEAGAGPRPVVIHTDGACSG-NP-GP----GGWGAILK---F-G-D--------TEK-----ELKGGEAH-TTNNRMELLAAISALEAL-------TRP------
CTVDLYTDSQYVKNGIG-SWIHNWKRNGWKTADKK---PVKNVDLWQRLDAALKS--H-QV-RWHWVKGHAGHDENERADQLARDGL-----TENR-----MKSRIG----------

>Proteobacteria alpha | Rhodopseudomonas | NP_949605

----------MSEADQKPVIIHTDGACSG-NP-GP----GGWGAILK---F-G-D----------VEK-----ELKGGEPH-TTNNRMELLAAISALEAL-------TRP------
CSVDLYTDSQYVKNGIG-SWIHNWKRNGWKTADKK---PVKNVDLWQRLDAALKT--H-SI-RWHWVKGHAGHAENERADQLARDGL-----TENR-----MKSRVK----------

>Proteobacteria alpha | Rhodopseudomonas | YP_783034

----------MSTLPAVLIHTDGACSG-NP-GP----GGWGAILK---F-G-E--------REK-----ELKGGESH-TTNNRMELMAAISALEAL-------TKP------
CSVDLHTDSQYVRNGIS-SWIHGWKKNGWKTADKK---PVKNVDLWQRLDAALKQ--H-EV-RWHWVKGHAGHAENERADQLARDGL-----SENR-----LKSRIG----------

>Proteobacteria alpha | Rhodopseudomonas | YP_533921

----------MSALPAVRVHTDGACSG-NP-GP----GGWGAILK---F-G-E----------IEK-----QLKGGETH-TTNNRMELLAAISALEAL-------TKP------
CTVDLYTDSQYVRQGIT-AWIHNWKRNGWKTADKK---PVKNVDLWQRLDAALKQ--H-DL-RWHWVKGHAGHDENERADQLARDGL-----IEHK-----LKSKIG----------

>Proteobacteria alpha | Bradyrhizobium j | NP_767956

----------MSELPVVTIYTDGACSG-NP-GP----GGWGAILK---F-G-D----------KEK-----ELNGGERH-TTNNQMELMAAISALEAL-------KKP------
CTVDLYTDSQYVRQGIT-GWIHGWKRNGWRTADKK---PVKNVELWQRLDAALKA--H-QV-RWHWVKGHAGHPENERADQLARDGI-----VKAR------LQQRVAE----------

>Proteobacteria alpha | Bradyrhizobium s | YP_001237112

----------MSELPTVSIYTDGACSG-NP-GP----GGWGAILR---F-G-D----------KEK-----ELKGGEPH-TTNNRMELMAAISALEAL-------KKS------
CQVELYTDSQYVRQGIT-GWIHGWKRNGWKTADKK---PVKNAELWQRLDAALKP--H-KV-NWHWVKGHAGHAENERADQLARDGV-----AMAR------LQKNVRG----------

>Proteobacteria alpha | Nitrobacter sp. | ZP_01045333

----------MNSPALPHVTIFTDGACSG-NP-GP----GGWGAILR---F-G-D----------IEK-----ELKGGEPH-TTNNRMELLAAISALEAL-------KRP------
ALVDLTTDSQYVRQGIM-SWIHNWKRNGWRTADKK---PVKNADLWQRLDAALQP--H-QV-RWHWIKGHDGHSENERADQLAREGV---AIARLK----------

>Proteobacteria alpha | Nitrobacter hamb | YP_578927

----------MNSSALPHVTIFTDGACSG-NP-GP----GGWGAILR---F-G-E----------IEK-----ELKGGEPH-TTNNRMELLAAISALEAL-------KKA------
ASVDLTTDSQYVRQGIT-SWIHNWKRNGWRTADKK---PVKNADLWQRLDTALQP--H-
QV-RWHWIKGHAGHDENERADQLAREGV---ALARLK----------

>Proteobacteria alpha | Oligotropha carb | YP_002290298

----------MSEPQRPHVVIFTDGACSG-NP-GP----GGWGAILR---F-G-E----------IEK-----ELKGGENP-TTNNRMELLAAISALEAL-------KRS------
AIVDLTTDSQYVRQGIT-SWIFNWKKNGWRTSDKK---PVKNVDLWQRLDAALKP--H-
EV-RWHWIKGHAGHAENERADELAREGL-----AENR----------

>Proteobacteria alpha | Beijerinckia ind | YP_001833435

----------MSGRVTIYTDGACSG-NP-GP----GGWGAILM---F-G-Q----------HEK-----ELSGGEAQ-TTNNRMELTAAIRALEAL-------TRP------
CAVDLHTDSNYLRGGVT-SWIKGWKKNGWRTADKK---PVKNVELWQELDQLAAS--H-
EI-AWHWVKGHAGHPLNERADALARQGM-----APFR------GGARIHPH----------

>Proteobacteria alpha | Azorhizobium cau | YP_001524382

----------MSRVEIWTDGACSG-NP-GP----GGWGAILR---S-G-P----------HEK-----ELKGGEAL-TTNNRMELMAAISALEAL-------KKP------
CGVDLHTDSEYLRNGIT-KWMFGWKRNGWRTADKK---PVKNQDLWERLDAALHS--H-
DI-AWHWVKGHAGNELNERADQLARDGM-----APFK------MGGRVA----------

>Proteobacteria alpha | Xanthobacter aut | YP_001417257

----------MKEVAVFTDGACSG-NP-GP----GGWGAILR---F-G-A----------HEK-----ELSGGEAL-TTNNRMELMGAIAALEAL-------KEP------
CTVDLHTDSNYLKDGVT-KWMHGWKRNGWRTADKK---PVKNQDLWERLDAALKR--
H-TL-RWHWVKGHAGHAENERADELARAGM-----APFK------LKGRLSG----------
--

>Proteobacteria alpha | Sphingomonas sp. | ZP_01303532

----------MSDLPQVEIFTDGACKG-NP-GP----GGWGAVLR---F-G-D----------TEK-----EISGGEAQ-TTNNRMEMTAALEALNLL-------KKP------
CAVTLYTDSKYVMDGIT-KWVFGWQKKGWRTADNK---PVKNVEIWQNLVKAAAR--H-
QM-TWKWVKGHAGHPENERADQLASAAA-----ETFR------R----------

>Proteobacteria alpha | Fulvimarina pela | ZP_01439595

----------MSKENRVEIYTDGACSG-NP-GP----GGWGVLLR---F-G-E----------HSK-----ELKGGEAN-TTNNRMELLAAIEALSAL-------KRP------
CAIDLHSDSSYMRDGIM-KWIHGWKKNGWKTADKK---PVKNAELWQRLDEERSR--H-
DV-TFHWVKGHAGHEGNERADQLANDGM-----EPFK------KKARSA----------

>Proteobacteria alpha | Aurantimonas man | ZP_01228016

----------MSAEGRVEIHTDGACSG-NP-GP----GGWGAILR---F-N-G------
--NEK-----ELKGGEEH-TTNNRMELLAVIEALTAL-------KRS------
CPVDIYSDSQYMRDGIT-KWIHGWKRNGWKTADKK---PVKNAELWQKLEEEKGR--H-
DV-TFHWVKGHAGDEMNERADQLARDGM-----EPFK------RRKRSA----------

>Proteobacteria alpha | Labrenzia alexan | ZP_05115201

----------MSQENRVTIYTDGACSG-NP-GP----GGWGVIMR---F-G-E------
---HER-----ELKGGEVE-TTNNRMELTAAIEALNAL-------KRP------
CVVDLYTDSTYVRSGIS-EWMYGWKRKNWKTAANK---PVKNADLWQALDAARER--H-
DV-TWHWVKGHAGHPDNERADELARGGM-----EPFK------KNS----------

>Proteobacteria alpha | Labrenzia aggreg | ZP_01548145

----------MTENTNRVTIYTDGACSG-NP-GP----GGWGVILR---F-G-E------
---HEK-----ELCGGEAE-TTNNRMELMAAIEALNAL-------KRP------
CAVDLYTDSTYVRSGIK-EWMYGWKRKNWRTAANK---PVKNADLWQALDAAKER--H-
DV-TWHWVKGHAGHPDNERADELARGGM-----APYK------NGEKTNPV----------
-

>Proteobacteria alpha | Chelativorans sp | YP_673315

----------MKRIEIFTDGACSG-NP-GP----GGWGAILR---Y-N-G--------
TEK-----ELYGGEAD-TTNNRMELTAAIEALEAL-------KEP------CEVDLHTDSNYLRDGIS-
GWIEGWKRNGWRTADRK---PVKNAELWQALDEARRR--H-KV-
HWHWVRGHAGHPENERADALARAGM-----APFK------KKKGGDTASSEEGSARRR---------
-----

>Proteobacteria alpha | Ochrobactrum ant | YP_001369151

----------MKRIEAYTDGACSG-NP-GP----GGWGAILR---W-N-D--------
-NVK-----ELKGGEAD-TTNNRMELMAAISALSAL-------KEP------
CEVDLYTDSVYVRDGIS-GWIEGWKRNGWKTAAKK---PVKNAELWQALDEARKP--H-
KV-NWHWVKGHAGHPENERADELAREGM-----EPFK------YGGRKSLKVQ----------
--

>Proteobacteria alpha | Brucella meliten | NP_540374

----------MKRIEAYTDGACSG-NP-GP----GGWGALLR---W-N-G------
--NEK-----ELKGGEAE-TTNNRMELMAAISALSAL-------KEP------
CEVDLYTDSVYVRDGIS-GWIEGWKRNGWKTAAKK---PVKNAELWQALDEARKA--H-
KV-TWHWIKGHAGHPENERADELARAGM-----EPFK------YAGHRTLKVK----------
-

>Proteobacteria alpha | Mesorhizobium op | ZP_05808546

----------MSKNVEIFTDGACSG-NP-GP----GGWGAILR---F-N-G--------
ATK-----ELSGGEAE-TTNNRMELLAAISALNAL-------KEP------CTVELHTDSKYVMDGIS-
KWIHGWKKNGWKTADKK---PVKNGELWQALDEANRR--H-KV-

TWNWVKGHDGHVENERADELARQGM-----APFK------
KGPFKPAAPAKPSAPAKQPAATKARRSTQSY

>Proteobacteria alpha | Hoeflea phototro | ZP_02167424

------------------------------------MKKVEVFTDGACSG-NP-GP----GGWGAILR---Y-G-E-------IEK-----EMSGGEAA-TTNNRMELLAAINALNAL-------KGA------
CEVELHTDSKYVMDGIS-KWIHGWKKNGWKTAAKK---PVKNAELWQALEEARKP--H-
KV-NWHWVKGHAGHPENERADELARFGM-----EPYK------NKTAPRRSAG----------

>Proteobacteria alpha | Agrobacterium vi | YP_002548759

------------------------------------MKHVDIFTDGACSG-NP-GP----GGWGAVLR---Y-G-E-------VEK-----DLCGGEAD-TTNNRMELLAAITALNTL-------KTP------
CEVDLHTDSKYVMDGIS-KWIFGWKKNGWKTADKK---PVKNGELWQQLDAANQR--H-
KV-TWHWVKGHAGHPENERADELARKGM-----EPFK------KGGAGSKPIL----------

>Proteobacteria alpha | Rhizobium sp. NG | YP_002825084

------------------------------------MKHVDIFTDGACSG-NP-GP----GGWGAVLR---Y-G-E-------VEK-----EMFGGEAE-TTNNRMELMAAISALNAL-------KQP------
CEVDLHTDSKYVMDGIS-KWIHGWKRNGWKTGDRK---PVKNGELWQALDEARDR--H-
QV-TWHWVKGHAGHPENERADELARKGM-----EPFK------KVRRTDAVK----------

>Proteobacteria alpha | Sinorhizobium me | NP_385020

------------------------------------MKHVHIFTDGACSG-NP-GP----GGWGAVLR---Y-G-D-------VEK-----EMSGGEAE-TTNNRMELLAAISALNAL-------RQP------
CEVDLHTDSKYVMDGIS-KWIHGWKRNGWKTGDRK---PVKNGELWQALDEARNR--H-
NV-TWHWVKGHAGHPENERADELARKGM-----EPFK------KARRADAVK----------

>Proteobacteria alpha | Agrobacterium tu | NP_353800

------------------------------------MKHVDIFTDGACSG-NP-GP----GGWGAVLR---Y-G-E-------TEK-----ELSGGEAD-TTNNRMELLAAISALNAL-------KSP------
CEVDLYTDSAYVKDGIT-KWIFGWKKKGWKTADNK---PVKNVELWQALEAAQER--H-
KV-TLHWVKGHAGHPENERADELARKGM-----EPFK------RR----------

>Proteobacteria alpha | Rhizobium etli K | ZP_03502453

------------------------------------MKHVDIFTDGACSG-NP-GP----GGWGAVLR---Y-G-D-------VEK-----ELCGGEAD-TTNNRMELMAAISALQAL-------KTP------
CEVDLYTDSAYVKDGIS-KWIFGWKKNGWKTSDKK---PVKNAELWQALEEARNR--H-
KV-TLHWVKGHAGHPENERADELARRGM-----EPFK------KGKAVSV----------

>Proteobacteria alpha | Agrobacterium ra | YP_002543663

---------------------------------------MKQVDIFTDGACSG-NP-GP----GGWGAVLR---Y-G-D-------KEK-----ELFGGEAE-TTNNRMELMAAISSLNAL-------KSP------
CEVNLYTDSKYVMDGIS-KWIFGWKKNGWKTADKK---PVKNAELWQALEEARNR--H-QV-TLHWVKGHAGHPENERADELARKGM-----EPFK------RK----------

>Proteobacteria gamma | Chromohalobacter | YP_573993

-----------------------------MTDSHATGDMPRVTIYTDGACRG-NP-GP----GGWGAVLR---Y-G-Q--------HEK-----TLKGGEAV-TTNNRMELMAAIQALRTL-------TRA------
CDVALWTDSEYLRKGIT-EWIHGWVKRGWKTAAKQ---PVKNAELWRELLAETQR--H-RI-EWHWVKGHSGHEGNELADTLANAAT-----DEIQ------AAKRQAMAGEQ----------

>Proteobacteria gamma | Allochromatium v | ZP_04772617

-------------------------------------MSDCVEAFTDGACKG-NP-GP----GGWGVLLR---W-G-E---------VEK-----ELHGGERE-TTNNRMELMAVIMALEAL-------KRP------
TPIRITTDSQYVKRGVG-EWMPRWKRNGWRTADRQ---PVKNRDLWERLDRALGQ--H-EV-SWRWVKGHAGHAENERADWLANLGV----------PTTGGR----------

>Proteobacteria gamma | Legionella longb | EEZ95027

---------------------------------------MIVEIYTDGACKG-NP-GP----GGWGVLLR---C-K-G--------QEK-----TLHGGEAH-TTNNRMELMAAIKGLEAL-------KRS------
CIVDLYTDSQYLRQGML-DWLPNWKMKGWRNSKKE---PVKNADLWMMLDELASR--H-QI-NWHWVKGHSGHLENELVDALANQAI-----EELR------E----------

>Proteobacteria gamma | Nitrosococcus oc | YP_344792

-------------------------------------MTEIVEIFTDGACRG-NP-GP----GGWGALLC---Y-Q-G--------REK-----TLSGAESK-TTNNRMELMAAIRALETL-------KRP------CRVHLTTDSQYLRQGIT-CWLSNWKRRGWKTANRQ---PVKNIDLWQRLDQVAAQ--H-RI-EWFWVRGHEGHPGNERADALARSAI---TNGEEK----------

>Proteobacteria gamma | Bermanella maris | ZP_01308670

-------------------------------------MSQIVEMFTDGACKG-NP-GP----GGWGVLLR---Y-G-A---------HEK-----ELFGGELE-TTNNRMELMAAIRGLEAL-------TKP------
CKVRLTTDSQYVRQGIT-QWLSGWKKKNWMTSSRQ---PVKNKELWQRLDSAVSK--H-DV-EWHWVKGHSGHIENERADDLANKGV-----EQVT------KS----------

>Proteobacteria gamma | Methylococcus ca | YP_113229

----------------------------------MSETEPTVYAYTDGACRG-NP-GP----GGWGVLLR---Y-G-S---------KTR-----EIYGGERE-TTNNRMELMAAIRALETL-------SRP------
CKVKIVTDSQYVKKGIT-EWVAQWEKRGWKTAGRS---PVKNIDLWQRLIQAEQR--H-QV-SWGWIKGHSGHPENEAADRLANRGI-----DELL------QSDKIPA----------

>Proteobacteria alpha | Marinomonas sp. | ZP_01075303

--------------------------VFLKKANRKRDEIVKHVIIYTDGACKG-NP-GP----GGWGAWIT---F-
G-E--------HEK-----RLCGGEND-TTNNRMELSGAIEGLKAL-------TEP------
CKVTLYTDSSYVQKGIT-QWLAGWKKKGWKTASKQ---PVKNKDLWQALDEECQR--H-
DI-EWKWVKGHAGIKGNEIADELANLGI-----EKIR------D----------

>Proteobacteria alpha | Marinomonas sp. | YP_001340565

---MYTDGACKG-NP-GI----GGWGAWLT---F-G-E--------
HEK-----HLCGGEHD-TTNNRMELMGAIEGLRAL-------KEK------
CSVTLYTDSSYVQKGIT-EWLAGWKRKGWMTASKQ---PVKNKDLWQALDEQCQY--H-
EV-TWKWVKGHAGIEGNEIADQLANKGI-----DELR------AAS----------

>Proteobacteria beta | Dechloromonas ar | YP_284812

----------------------------------MTAEETVEIFTDGACKG-NP-GP----GGWGAILR---L-G-P------
--HEK-----ELWGGEKE-TTNNRMELTAAIRAIEAL-------KRP------
IGGKIYTDSQYVMKGIN-EWIHGWKKNGWKTSDKK---PVKNADLWQLLDAQVKL--H-
KL-EWIWVRGHSGHPENERADALANRGI-----EELK------G----------

>Proteobacteria gamma | Marinobacter sp. | ZP_01738333

-------------------------------------MSGNVIMYTDGACKG-NP-GR----GGWGVVLR---W-G-E-----
----VCK-----TLHGGEQH-TTNNRMELMAAIEGLKAL-------KRD------
CDVELYTDSQYVRKGIT-EWLAGWKRNGWKTAAKK---PVKNDDLWKALDEQSER--H-
RV-NWHWVKGHAGVPDNELADQLANQGV-----EELT------G----------

>Proteobacteria gamma | Marinobacter alg | ZP_01895194

-------------------------------------MAGKVVLYTDGACKG-NP-GP----GGWGVVLR---Y-G-D-----
----ANK-----MLHGGEAN-TTNNRMELMAAIQGLKAL-------RRT------
CDVELYTDSQYVRKGIT-EWMTGWKRNGWKTSAKK---PVKNEDLWRELDNEVAR--H-
KV-NWHWVKGHSGNPDNELADELANRGV-----EELA------NAG----------

>Proteobacteria gamma | Marinobacter aqu | YP_958805

-------------------------------------MAGKVVMYTDGACKG-NP-GP----GGWGVVLR---Y-G-D-----
-----ACK-----TMHGGELQ-TTNNRMELMAAIRGLREL-------KRA------
CQVELYTDSQYVRKGIT-EWMSGWKRNGWKTSAKK---PVKNADLWQELDAETAR--H-
TV-NWHWVKGHSGHPDNELADELANRGV-----RELN------GA----------

>Proteobacteria gamma | Thioalkalivibrio | ZP_03690355

-----------------------------------MTDERVYIYTDGACRG-NP-GP----GGWGAILR---Y-R-G------
--TER-----ELYGAEAE-TTNNRMELTAAIRALETL-------KRP------
CKVELVTDSKYVKQGLT-EWLPGWKRRNWKSASGS---PVKNRDLWEALDAEAAR--H-
DI-EWYWVRGHSGHPENERADELANAGI-----DALL------AGQPIEN----------

>Proteobacteria gamma | Alcanivorax bork | YP_692944

---------------------------------------MKNVIIYTDGACRG-NP-GP----GGWGAILL---Y-G-D--------
KEK-----ELFGGEPE-TTNNRMELMAAIVALETL-------NTP------CQVVLTTDSKYVMDGIT-
QWMANWKKRGWKTASKQ---PVKNVDLWQRLDAAVQR--H-DI-
DWQWVKGHSGHPGNERADALANRGI-----DEMK------HKQGQAS----------

>Proteobacteria gamma | Kangiella koreen | YP_003147108

---------------------------------------MKKIEIYTDGACKG-NP-GP----GGWGALLR---Y-N-K--------
HEK-----HLFGGELN-TTNNRMELMAAIEALKAL-------KDK------
CQVDLTTDSVYVKNGIN-QWLENWKAKGWKTANRK---PVKNQDLWQQLDQQVAR--H-
NV-TWHWVKGHSGHPENDIADELANKGV-----EKVL------QSSGV----------

>Proteobacteria gamma | Thioalkalivibrio | YP_002513030

-----------------------------------MAESQRVEIYTDGACRG-NP-GP----GGWGAVLR---F-K-G-----
----RER-----TLKGAEAE-TTNNRMELTAAIMALETL-------TRP------
CAVDLTTDSQYVKQGLT-QWIHGWKRKGWRTADGK---PVKNQDLWMRLDAAAAR--
H-EV-AWHWVRGHTGHPENELADQLANEAI-----DEML------AGA----------

>Proteobacteria alpha | Magnetospirillum | YP_420125

-------------------------------MSETAPKPETVEIYTDGACSG-NP-GP----GGWGAILR---F-K-G-----
-----IEK-----ELKGGESP-TTNNRMEMMAVLVALNTL-------TRS------
CAVDVYTDSEYVKKGMT-EWLRGWKARGWKTADKK---PVKNDDLWKALDEAAAR--
H-KV-SWHWVKGHAGHPENERADALAREGI-----ADLR------ART----------

>Proteobacteria alpha | Magnetospirillum | CAM75322

-------------------------------------MTERVEIFTDGACSG-NP-GP----GGWGAILR---Y-K-G--------
VEK-----ELCGGENP-TTNNRMEMMAAIMALETL-------SRS------
CPVTLYTDSQYVMKGMT-EWLKGWKARGWKTADKK---PVKNDDLWQRLDAACAR--
H-QI-TWQWVKGHAGHPENERADQLARDGI-----KVVL------GK----------

>Proteobacteria gamma | HTCC2080 | ZP_01625362

---------------------------------------MKVVEAFTDGACKG-NP-GP----GGWGVLLR---M-D-N-----
---QSR-----EIFGGDGA-TTNNRMELTAAIEALAAL-------KEA------
CTVELTTDSTYVKDGVT-RWMENWERNGWRTAAKK---PVKNQDLWQALKAQVAR--H-
EV-NWHWVKGHSGHPENELADMLANKGI-----AELQ------R----------

>Proteobacteria gamma | NOR5-3 | ZP_05126429

---------------------------------------MKNVELFTDGACRG-NP-GP----GGWGALLC---Y-A-G------
--KER-----EVYGAEPN-TTNNRMELSAAIEGLAAL-------SEP------
CAVRLVTDSTYVMKGIT-EWLPNWKRRGWKTSAKK---PVANADLWQLLEVQNQR--H-
KV-SWEWVKGHSGHPGNERADALANRAI-----DEML------S----------

>Proteobacteria gamma | Congregibacter l | ZP_01101009

------------------------------------MKRVDLFTDGACRG-NP-GP----GGWGALLV---Y-G-S------
--KER-----ELYGGAAD-TTNNRMELSAAIEGLAAI-------SEP------
CAVKLVTDSTYVMKGIT-EWLPNWKRRGWKTAAKK---PVANADLWQRLEAECER--H-
SI-EWEWVKGHSGHPGNERADALANKAI-----DEML------A----------

>Proteobacteria gamma | Thiomicrospira c | YP_391198

------------------------------------MQEVELFTDGGCRG-NP-GP----GGWGALLR---F-G-G-------
-VEK-----ELKGAELD-TTNNRMELTAAIEGLKAL-------KRP------
CKVTLTTDSQYVKNGIT-QWMTNWKKNNWKTAAKK---PVKNKDLWQALDEALQP--H-
DV-TWAWVKGHSGHDENERVDELANQAM-----DELT------G----------

>Proteobacteria gamma | Pseudomonas aeru | NP_250506

-----------------------------------MTDKEQVVIYTDGACKG-NP-GR----GGWGALLL---Y-K-G---
-----AER-----ELWGGEPD-TTNNRMELMAAIQALAAL-------KRS------
CPIRLITDSEYVMRGIT-EWLPNWKKRGWKTASKQ---PVKNADLWQALDEQVAR--H-
QV-EWQWVRGHTGDPGNERADQLANRGV----AELPR----------

>Proteobacteria gamma | Pseudomonas stut | YP_001172753

------------------------------------MSDDWVEIYTDGACKG-NP-GP----GGWGALLI---Y-K-G-----
---VKR-----ELWGGEPD-TTNNRMELMAAIRALAEL-------KRP------
CKVRLVTDSQYVMQGIN-DWMPNWKKRGWKTASKQ---PVKNADLWQQLDEQVNR--
H-EV-SWQWVRGHTGHPGNEQADLLANRGV-------VQ------AKRQPIV----------

>Proteobacteria gamma | Azotobacter vine | YP_002800101

------------------------------------MSDKVEIYTDGACKG-NP-GP----GGWGALLV---C-Q-G------
--VER-----ELWGGEAE-TTNNRMELTAAIRALAEL-------KRP------
CEVHLTTDSEYVMRGIL-EWLPNWKKRGWKTAARQ---PVKNADLWQQLDEQVGR--H-
RV-TWGWVRGHTGHPGNERADLLANRGV-----AAAR-----AQKKHGV----------

>Proteobacteria gamma | Pseudomonas mend | YP_001187559

------------------------------------MSETDEVVIYTDGACKG-NP-GP----GGWGALLV---Y-K-G----
----VEK-----ELWGGDPS-TTNNRMELMAAIAGLIAL-------TRP------
CSVKLVTDSQYVMKGIQ-EWLPNWKKRGWKTASKE---PVKNADLWQKLDEEVNR--H-
QV-SWQWVRGHTGHPGNERADQLANRGV-----DEVR------SAR----------

>Proteobacteria gamma | Pseudomonas ento | YP_609085

------------------------------------MSDSVEIYTDGACKG-NP-GP----GGWGVLMV---F-K-G------
--VEK-----ELWGGERE-TTNNRMELMAAIEGLKAL-------KRE------
CEVVLTTDSQYVMKGIN-EWMVNWKKRGWKTAAKE---PVKNADLWMALDEQVNR--
H-KV-TWKWVRGHIGHPGNERADQLANRGV-----DEVR------AQR----------

>Proteobacteria gamma | Pseudomonas puti | YP_001750319

------------------------------------MSDSVELYTDGACKG-NP-GP----GGWGVLLI---Y-K-G------
--VEK-----ELWGGERE-TTNNRMELMAAIQGLMAL-------KRE------
CEVVLTTDSQYVMKGIN-EWMVNWKKRGWKTAAKE---PVKNADLWQQLDEQVNR--
H-KV-TWKWVRGHIGHPGNERADQLANRGV-----DEVR------AKR--------------------------

>Proteobacteria gamma | Pseudomonas syri | ZP_05637838

------------------------------------MSDSVELFTDGACKG-NP-GP----GGWGALLV---C-K-G------
--VEK-----ELWGGEAN-TTNNRMELTGAIRGLEEL-------KRP------
CEVTLVTDSQYVMKGIT-EWMVNWKKRGWKTAAKE---PVKNADLWQLLDEQVSR--H-
NV-KWQWVRGHIGHPGNERADQLANRGV-----DEVR------GIKS-------------------------

>Proteobacteria gamma | Pseudomonas fluo | YP_002872231

------------------------------------MTDSVELFTDGACKG-NP-GP----GGWGALLV---C-K-G------
--VEK-----ELWGGEAN-TTNNRMELMGAIRGLEEL-------KRR------
CNVLLVTDSQYVMKGIN-EWMVNWKKRGWKTAAKE---PVKNADLWQLLDEQCNR--
H-DI-TWKWVRGHIGHPGNERADQLANRGV-----DEVR----------GYKQS----------------------

>Proteobacteria gamma | Pseudomonas fluo | YP_347918

----------------------------------MSESVDSVELFTDGACKG-NP-GP----GGWGALLV---C-K-G---
-----VEK-----ELWGGEAN-TTNNRMELLGAIRGLEAL-------KRP------
CEVLLVTDSQYVMKGIN-EWMANWKKRGWKTAAKE---PVKNADLWKALDEQVNR--
H-KV-TWKWVRGHIGHHGNERADQLANRGV-----DEVR----------GYKQS----------------------

>Proteobacteria gamma | Pseudomonas fluo | YP_260402

------------------------------------MSDSVEIFTDGACKG-NP-GP----GGWGALLV---C-K-G-------
-VEK-----ELWGGEAN-TTNNRMELMAAIRGLEEL-------KRQ------
CDVQLVTDSQYVMKGIN-EWMANWKKRGWKTAAKE---PVKNADLWQQLDEQVNR--
H-NV-TWKWVRGHTGHHGNERADQLANRGV-----DEVR----------GYKQP----------------------
-

>Proteobacteria gamma | HTCC2207 | ZP_01224052

------------------------------------MKAVELFTDGACRG-NP-GP----GGWGVLMR---Y-G-D------
---KEK-----TLCGGEAE-TTNNRMELTAVIEGIAAL-------SEP------
CKVSVTSDSTYVLKGIQ-EWMPAWKKRNWKTASKK---PVKNVDLWQKLDAVIKH--H-
DI-DWHWVKGHSGHAENEIADQLANRGI-----DEL-----------------------------------

>Proteobacteria gamma | NOR51-B | ZP_04958093

------------------------------------MTTSVEAFTDGACRG-NP-GP----GGWGVLLR---R-G-D------
--RER-----ELWGGEAQ-TTNNRMELTAAIEALRSL-------KDG------
STVDLTTDSVYVRDGIT-RWVAGWKRNGWRTAARK---PVKNQDLWQSLDEQCAR--H-
EV-RWHWVKGHSGHAENDRADALANRGI-----DELK------G----------------------------

>Proteobacteria gamma | Neptuniibacter c | ZP_01167471

---------------------------------------MKTVEIFTDGACKG-NP-GP----GGWGAVLR---Y-G-D-------
-AEK-----QMHGGEND-TTNNRMELMAAIVALETL-------NRP------
CEVILTTDSQYVRQGIT-EWIEGWKRKGWKNSQKK---PVKNADLWQRLDAARQP--H-
KI-DWRWVKGHSGHPENELADQLANKGV-----EELG------RA----------

>Proteobacteria gamma | Nitrococcus mobi | ZP_01128586

---------------------------------------MAPVEIYTDGACRG-NP-GP----GGWGAVLR---Y-G-G-------
-HEK-----SLCGGATQ-TTNNRMELTAAIQALESL-------KRP------CRVVLTTDSQYLRRGIT-
EWLPNWKRRGWRTAERK---PVKNADLWQRLDMLAAR--H-EV-
DWRWVRGHNGHPGNEQADRLANQGI-----DEML------ARR----------

>Proteobacteria gamma | Alkalilimnicola | YP_742824

--MYAWTDGACRG-NP-GP----GGWGVVLR---Y-R-G--------
HER-----TLHGGEPH-TTNNRMELTAAIQALEAL-------DRP------CVVHLTTDSQYVRKGIT-
EWMAGWKRRGWRTAARK---PVLNEDLWRRLDALNQR--H-EV-
HWHWVRGHSGHAENEQADALANRGI-----DEMQ------EAGAT----------

>Proteobacteria gamma | Halorhodospira h | YP_001003151

----------------------------------MTEQRGVVEAFTDGACRG-NP-GP----GGWGVLLR---Y-G-E--
------HER-----ELYGGEPE-TTNNRMELTAAIRALEAL-------DRP------
CRVVLTTDSQYVRRGIT-EWLEGWKRRGWRTASRK---PVLNQDLWQRLDELAAY--H-
QV-DWHWVRGHAGHAENERADALANQGI-----DELV------A----------

>Proteobacteria gamma | Teredinibacter t | YP_003073577

---------------------------------------MKKIEIFTDGACRG-NP-GP----GGWGVLLR---Y-G-D--------
KEK-----TLHGGERD-TTNNRMELRAAIEGLSAL-------KEP------CEVRLVTDSQYVRKGIT-
EWIANWKKRGWRTAAKK---PVMNVDLWQALDTACDQ--H-QI-
TWEWVKGHSGHRENEIADELANLGI-----DELN------VRR----------

>Proteobacteria gamma | HTCC2148 | ZP_05093760

---------------------------------------MKNIEIFTDGACRG-NP-GP----GGWGALLR---F-Q-G--------
KEK-----SLYGGEAQ-TTNNRMELQAAIEGLKAL-------KEP------CVVALTTDSIYVKNGIT-
SWLPGWKKKGWKTSNKK---PVKNVDLWQSLDEQNQR--H-QV-
DWHWVKGHSGHRENEIADQLANRGI-----DELT----------

>Proteobacteria gamma | Cellvibrio japon | YP_001982514

---------------------------------------MKTVEIFTDGACKG-NP-GP----GGWGALLR---Y-G-Q-------
-VEK-----SLYGGEPE-TTNNRMELMAAIAALSAL-------KEP------CAVVITTDSQYVRKGIT-
EWMPGWKRNGWRTAAKE---PVKNADLWQRLDEQNQR--H-QV-
TWKWVKGHSGHRENELADALANRGI-----DELR----------

>Proteobacteria delta | Stigmatella aura | ZP_01463545

----------MTLPLVHIYCDGACSP-NP-GI----GGWGSILVSPAH-G-H-----
---ARK-----ELSGAEPG-TTNNRMELTAALMALRAL-------KSP------
CQVQLFTDSQYVRNAFQEKWLDKWQRTGWKTAGRQ---PVQNADLWQALLEQTRV--H-
QV-SWNWVRGHSGHVENERADAMAVAAR-----LA----L---AAKLGR----------

>Proteobacteria gamma | Buchnera aphidic | NP_777852

----------MLKTIKIFSDGSCLG-NP-GP----GGYSFIIQ---H-L-E--------
YEN-----ISSSGFYL-TTNNRMELMGIIVATESL-------KQP------CCITISTDSQYVQKGIL-
YWIKNWKTKGWKTSRKT---YVKNVDLWLRLEKSLNL--H-QV-
TWKWIKSHSGNKKNEQCDHLARESA-----KFPT------LKDFGYIL----------

>Proteobacteria delta | Desulfonatronosp | ZP_03736090

----------MSRNNNTQVVDIYTDGACLG-NP-GP----GGYAAILK---W-G-D-
-------LEK-----EISQGTPG-TTNNRMELMAVLEGLKAL-------KYP------
CRVRIHTDSQYIARAINEKWLEKWQRNGWKTAQKE---DVKNRDLWEELAALLQE--H-
KV-EFKWVRGHSGHEYNERCDSLARQAA-----QAVDD----------

>Proteobacteria delta | Desulfomicrobium | YP_003156805

----------MTDTSVTIYTDGSSLG-NP-GP----GGWGAVLI---W-A-D------
--SKK-----ELSRGYIE-TTNNRMEIRGVLHALEHL-------KRP------
CTVHVHSDSRYVCDAISKKWIQSWLKNGWLTSAKK---PVKNRDLWEQLLSLLQK--H-
KV-IFHWVKAHDGHPENERCDELAKNAA-----KARE-----REIDEGYLRNT----------

>Proteobacteria gamma | Buchnera aphidic | NP_660587

----------MLKLVKMFSDGSCLG-NP-GS----GGYGTILR---Y-K-L--------
HEK-----ILTSGFFL-TTNNRMELMGVICGLESL-------KES------CIVEITIDSQYVKQGIT-
NWIATWEKKKWKTTKKK---LIKNLDLWLRINAVIKN--H-HI-
TWFWVKAHMGHLENERCDKIARQSA-----QSPS------VKDFFYENNFYQNKNL----------
-

>Actinobacteria | Atopobium vagina | ZP_03946132

----------MENSSSIKAKAMDAARTNSTKPQVIIWTDGSSRG-NP-GP----
GGYGAVML---F-Y-D---SAGREHKR-----ELSCGYRQ-TTNNRMELLAPIIALEEL-------KYP-
-----CKVELHSDSQYVIHAFQQHWIDGWQKRGWKTANKQ---
PVKNVDLWKRLLRAMQP--H-EM-SFVWVKGHAGTELNERCDELATTAA-----DADV---
SLLQVDEGFEALS----------

>Actinobacteria | Atopobium rimae | ZP_03568735

----------MYEDVYTQSETIGSSSKRAQVVAYTDGASRG-NP-GP----GGYGAVLV--
-Y-V-D---ASGKRHTR-----EFSQGYRL-TTNNRMELLGVIVALEAL-------TQP------
CVVEVHSDSKYVVDAFNQGWIFGWMRRGWKTSNKQ---NVKNIDLWKRLLVASSS--H-
EV-HYVWVKGHAGEELNERCDELATTAA-----DGQD-----LLEDVGFQGE----------
-

>Firmicutes | Butyrivibrio cro | ZP_05792240

------------------------------------MNNVIIYTDGAARG-NPNGP----GGYGVVLE---Y-T-D---
KNGIVHHK-----ELSQGYKK-TTNNRMELMAAIAGLEAL-------KAP------
CNVTLYSDSKYLVDAFRQKWIDSWIAKDFKRGKNE---PVKNPDLWKRLLKAKEN--H-
NV-EFVWVKGHAGHAMNEKCDMLATSAA-----DGDN--L---ADDVTLENIV----------
---

>Actinobacteria | Atopobium parvul | YP_003180122

--------------------------------MEAATSYMHVTVYTDGASRG-NP-GP----GGYGAVLL---Y-T-
D---PSGQQHTK-----EFSQGYKT-TTNNRMELLGVIVALEAL-------KRP------
CQVELYSDSKYVVDAFNQKWVSGWVRKGWKTASKE---PVKNVDLWKRLLAAMED--
H-EV-SFKWVKGHAGHPLNERCDQLATEAA-----DGSN-----LLDDQGFFADQSLL----------
----

>Firmicutes | Clostridium hath | ZP_06117226

-------------------------------------MGKVLLFTDGAARG-NPDGP----GGYGAVLQ---F-T-D---
SKGQLHEK-----TLSAGYVR-TTNNRMELMAAIAGLEAL-------NRP------
CEVELYSDSKYVTDAFNQHWIDNWVKNNWKRGKSG---PVKNIDLWKRLLKAMEP--H-
RV-TFCWVKGHAGHPENERCDQLATTAA-----DGDS-----LLVDEGL----------

>Firmicutes | Clostridium phyt | YP_001559482

-------------------------------------MKVTIYTDGAARG-NPDGP----GGYGTILS---Y-I-D---
STGVEHIR-----EYSGGYKK-TTNNRMELMAAIVGLEAL-------TKP------
CVVTLYSDSQYVVKAFNEHWLDGWIKKGWKRGKNE---PVKNVDLWKRLLAAKNQ--
H-DV-TFCWVKGHDGHPQNERCDVLATTAA-----DGGN------LADDNVVE----------
---

>Firmicutes | Bryantella forma | ZP_05348692

-------------------------------------MLVKIYTDGAARG-NPDGP----GGYGTILH---Y-T-D---
TKGVLHER-----TFSQGYEK-TTNNRMELMAAIIGLEAL-------NRP------
CQVELYSDSKYLTDAFNRHWIDSWQRKGWKRGKNE---PVKNVELWKRLLAAMEP--H-
EV-SFIWVKGHDGHELNERCDRLATSAA-----DGEE----LAVDDGGEMR----------

>Actinobacteria | Eggerthella lent | YP_003181454

-------------------------------------MHVDIYTDGAARG-NP-GP----GGYGTVLR---F-V-D---
SKGAVHEK-----ELSQGYER-TTNNRMELMAVVAGLEVL-------KRP------
CSITLYSDSQYVVNAFNQHWVDGWLKRGWKNAQKQ---PVKNDDLWKRLLAAKEP--H-
DV-TFVWVKGHAGHPENERCDELATTAA-----DGAG-----RIRDEGFNG----------

>Actinobacteria | Slackia heliotri | YP_003143435

-------------------------------------MHVEIYSDGSSRG-NP-GP----GGYGSVLH---Y-T-D---
AQGQLHVK-----ELSQGFVT-TTNNRMELLGVIVALEAL-------KRP------

CSVDVYSDSQYVVKAFNDHWIDGWLKRGWKNSKKE---RVKNQDLWRRLLAAKAP--H-
QV-SFHWVKGHAGHPENERCDQLATEAA-----DGSG-----LILDEGFTSEDV----------

>Actinobacteria | Slackia exigua A | ZP_06159453

----------MHVDIYSDGSSRG-NP-GP----GGYGAILR---F-V-D---
PSGGVHQK-----ELSGGFER-TTNNRMELLGVIVALEAL-------KAP------
CEVAVYTDSQYVVKAFTDRWVDGWKRRGWKNAKKE---PVKNQDLWMRLIAALEG--
H-RV-SFHWVKGHAGHPENERCDELATTAA-----DGAD-----RLIDEGFTD----------
-

>Nitrospirae | Leptospirillum f | EES51503

----------MESDEVELFADGACSG-NP-GP----GGWGVLLR---C-R-G------
---HVR-----EISGGEFQ-TTNNRMELSGVIAGLSAL-------KKP------
CRVMVTTDSQYVKNGMT-TWIRSWKKNGFRTSSGQ---PVKNEDLWRELDRLAAL--H-
EI-TWHWVRGHDGHPENERVDLLAREAI-----SSVR--K---GNSQGA----------

>Proteobacteria delta | Desulfuromonas a | ZP_01311301

----------MSQKQHVEIYSDGACRG-NP-GP----GGYGTLLR---C-G-S------
---HIK-----ELSGYEAQ-TTNNRMELLGAIAGLEAL-------KKP------
CIVTLTTDSQYVYKGMT-QWLSGWKKKGWKNSQKK---DVLNRDLWERLERAAQD--H-
EV-TWQWVKGHAGHEENERCDELARTAI-----DLAE----------

>Proteobacteria delta | Pelobacter carbi | YP_355635

----------MSESRSMVEIFSDGACSG-NP-GP----GGFGTLLR---C-G-E-------
--RVR-----ELSGFDPE-TTNNRMELLGAIAGLEAL-------TRP------
CRVRLTTDSQYVCKGMT-EWIHGWQKKGWKNSKKE---DVANRDLWERLLVLVSK--H-
EV-SWHWVRGHAGHAENERCDELARQAI-----ADGC--S---SVV----------

>Proteobacteria delta | Pelobacter propi | YP_902594

----------MSARPDQSRTAAPSATTSPVEIYCDGACSG-NP-GP----GGYGAILR---Y-
N-G--------HEK-----EIRGSEAH-TTNNRMELTAAMEALRLL-------TRP------
CRITIVTDSQYLVKGMT-EWIQGWQRRGWQNSKKE---PVLNRDLWEELLKLSAH--H-
DV-SWQWIRGHAGHAENERCDSLARQAI-----TEMR------GAP----------

>Proteobacteria delta | Geobacter lovley | YP_001953011

----------MKQQNSLTEVEIFCDGACSG-NP-GP----GGYGTILR---C-R-G------
----KEK-----ELSGAATE-TTNNRMELTAALEGLRQL-------TRS------
CRVTITTDSQYLVKGMT-EWLPGWQRNGWKNSKKE---PVLNRDLWEALVEASKP--H-
QV-AWQWVRGHAGHAENERCDTLAREAI-----SAMQ------VRN----------

>Proteobacteria delta | Geobacter sulfur | NP_953120

----------MSAEVEVFCDGACSG-NP-GV----GGYGAILR---Y-G-S-------
-AEK-----ELSGADGD-TTNNRMELTAAIRALEAL-------SRP------

CAVTITTDSQYLVKGMT-EWLSGWVRRGWVNSKKE---PVLNRDLWERLRELTGK--H-
QV-RWVWVRGHNGHPENERCDALARRAI-----DAYR------NERR--------------------------

>Proteobacteria delta | Geobacter sp. FR | YP_002538914

---------------------------------------MKVEIFCDGACSG-NP-GV----GGWGCILR---Y-G-D--------
NVK-----EMSGADGN-TTNNRMEMTAAIEALASL-------KRP------
CEVHLTTDSQYLVKGMT-EWIGGWVRKGWVNSKKE---PVLNRELWERLMELSRL--H-
TI-HWLWVRGHNGHPENERCDELARAAI-----ETFR------RSC--------------------------

>Proteobacteria delta | Geobacter uranii | YP_001231597

---------------------------------------MKVEIFCDGACSG-NP-GV----GGWGSILR---Y-G-D--------
TVK-----ELSGADGD-TTNNRMEMTAAIEALASL-------KRP------
CEVVLTTDSQYLVKGMT-EWMSGWIRKGWVNSKKE---PVLNRELWERLLALSKI--H-KI-
RWAWVRGHNGHPENERCDELARAAI-----EVFK------GRKP--------------------------

>Proteobacteria delta | Geobacter sp. M1 | ZP_05310359

---------------------------------------MQVEIFCDGACSG-NP-GV----GGWGSILR---Y-G-D--------
KVK-----ELSGAEGE-TTNNRMEMSAAIGALEAL-------TRP------
CEVVVTTDSQYLAKGMT-EWVAGWIRKGWVNSKKE---PVVNRDLWERLVALARV--H-
RI-KWVWVRGHNGHVENERCDELARAAI-----DSYRAARRQETVSPPNPPAADTRL--------
----

>Proteobacteria delta | Geobacter sp. M2 | YP_003021576

---------------------------------------MQVEIFCDGACSG-NP-GV----GGYGSILR---C-G-E--------
TVK-----EISGADGD-TTNNRMEMSAAIAALEAL-------KRP------
CQVVVTTDSQYLAKGMT-EWLSGWVKRGWVNSKKE---PVLNRDLWERLLELSKV--H-
QI-RWVWVRGHNGHVENERCDELARAAI-----DSYR------AGNGR--------------------------

>Proteobacteria delta | Desulfovibrio sa | YP_002989784

-----------------------------------MSKKKLTIYTDGSCLG-NP-GK----GGYGAVLL---F-N-E-------
-HRK-----ELSQGYKK-TTNNRMEMRAVIAALTEL-------KEP------
CEVTLYTDSQYVKNAFTKKWIDNWQKNGWKTAAKK---PVKNKDLWLQFIPLLEK--H-
DV-TFRWVKGHAGDPENERCDDLARTAA-----SSGD-----LIVDEGA--------------------------

>Proteobacteria gamma | Baumannia cicade | YP_588912

-----------------------------------MRKKIEIFTDGSCFG-NP-GP----GGYGAILR---Y-K-K--------
YEK-----EHSAGFLL-TTNNRMELMAAIIALEFL-------RDP------CEAIVYIDSKYVHQGVI-
QWIYNWKKHNWKNSAKK---IIKNLDLWQRLDVVSNL--H-VI-
HWRWVKSHTGHPENERCDELARIAA-----EHPK------FEDIGYKR----------------------

>Proteobacteria gamma | Pseudoalteromona | ZP_01132449

-----------------------------------MRKSVAIYTDGSCLG-NP-GP----GGYGVVMR---Y-N-E-------
--HLK-----ELSQGFEL-TTNNRMELLAAIVGLESL-------KQA------

CDVVLTTDSQYVKQGIE-SWLQGWKKRNWLTANKQ---PVKNIDLWQRLDIINQT--H-
HV-QWRWVKGHSGHFENERCDVLARSAA-----ESST-----LLPDEGYRATE----------
-

>Proteobacteria gamma | Haemophilus para | ZP_02478158

---------------------------------------MKLVDIFTDGSCLG-NP-GK----GGIGILLR---Y-Q-G--------
KEK-----RISQGYYL-TTNNRMELLAVITALNAL-------KEP------
CNVHLHSDSQYMQNGIQ-KWIFNWKKNNWKTSNNT---PVKNQDLWIALDKAITR--H-
QV-EWQWVKGHSGHTENEICDQLAKEGA-----NNPT------LEDVGYPPS----------

>Proteobacteria gamma | Haemophilus ducr | NP_873664

---------------------------------------MKSVNIFTDGSCLG-NP-GP----GGIGVVLR---Y-N-Q--------
HQK-----KVSQGYFQ-TTNNRMELRAVIEGLSML-------KEA------
CNVTLYSDSQYMKNGIT-KWIFKWKKSNWKTANGK---AVKNKDLWLLLDEKIQI--H-
YI-EWKWVKGHSGHYENEICDELAKLGA-----NNPT------LEDVGYQPA----------

>Proteobacteria gamma | Mannheimia haemo | ZP_04977670

---------------------------------------MMKLVEIFTDGSCLG-NP-GK----GGIGILLR---Y-N-G--------
YEK-----TVSKGYFQ-TTNNRMELRAVIEALAML-------KEP------
CKVQLNSDSQYMKNGIQ-KWIFNWKKNDWKTSDKK---PVKNKDLWVALDQEIQR--H-
QI-EWSWVKGHSGHRENEICDELAKQGA-----NNPT------LDDVGYIAD----------

>Proteobacteria gamma | Actinobacillus m | ZP_04752263

---------------------------------------MKQVEIFTDGSCLG-NP-GK----GGIGILLR---Y-N-Q--------
HEK-----TVSQGYFQ-TTNNRMELRAVIEALAML-------KEP------
CQVHLHSDSQYMKDGIT-KWIFNWKRNNWKTANGK---AVKNQDLWIALDMEIQR--H-
KM-EWHWVKGHAGHRENEICDELAKAGA-----NNPT------LEDIGYNAE----------

>Proteobacteria gamma | Actinobacillus p | YP_001651475

---------------------------------------MKLVEIFTDGSCLG-NP-GK----GGIGIVLR---Y-N-G--------
HEK-----QVSKGYLQ-TTNNRMELRAVIEALAML-------KEP------
CQVQLNSDSQYMKDGIT-KWIFNWKKNNWKTANGK---PVKNKELWIALDQEIQR--H-
KI-EWTWVKGHSGHRENEICDELAKAGA-----NNPT------LEDIGYNAD----------

>Proteobacteria gamma | Haemophilus somn | YP_718816

---------------------------------------MLKKIEIFTDGSCLG-NP-GA----GGIGILLR---Y-K-Q--------
HEK-----KLYQGFFQ-TTNNRMELRAVIVALNSL-------KEP------CSVILYSDSQYMKNGIT-
KWIFNWKKNNWKTSSGN---AVKNQDLWCSLDQAIQK--H-QI-
EWRWVKGHNGHRENEICDQLAKQGA-----ENPT------LEDVGYRAE----------

>Proteobacteria gamma | Mannheimia succi | YP_088763

-----------------------------------MYQIMRKQIEIFTDGSCLG-NP-GA----GGIGVVLR---Y-K-Q-----
---HEK-----TLSQGYFK-TTNNRMELRAVIEALNLL-------KEP------

CAVTLHSDSQYMKNGIT-QWIFNWKKKNWKASNGK---PVKNQDLWMALDNAVQA--H-
TI-DWRWVKGHSGHRENELCDQLAKQGA-----ENPT------LEDIGYQPD----------

>Proteobacteria gamma | Actinobacillus s | YP_001344181

------------------------------------MRKQIEIFTDGSCLG-NP-GV----GGIGVVLR---Y-K-Q--------
HEK-----TLSKGYFQ-TTNNRMELRAVIEALNLL-------KEP------CEIILHSDSQYMKNGIT-
QWIFNWKKNNWRASTGK---PVKNQDLWIALDSAIQP--H-TI-
HWRWVKGHSGHRENEMCDELAKQGA-----ENPT------LEDTGYRQD----------

>Proteobacteria gamma | Aggregatibacter | YP_003256376

------------------------------------MRKQIEIFTDGSCLG-NP-GA----GGIGILLR---Y-K-Q--------
HEK-----KLSKGFFL-TTNNRMELLAVVEALNSL-------KEP------CDIHLYSDSQYMKNGIT-
QWIFNWKKNHWKASSGK---PVKNQDLWMALDQAIAR--H-KV-
DWRWVKGHAGHRENEICDQLAKQGA-----ENPT------LNDEGYQAEA----------

>Proteobacteria gamma | Aggregatibacter | YP_003006768

------------------------------------MRKQIEIFTDGSCLG-NP-GA----GGIGVLLR---Y-K-Q--------
HEK-----SLSKGYFL-TTNNRMELLAVIEALNSL-------KEP------CDIHLYSDSQYMKNGIT-
QWIFNWKKNNWKASSGK---PVKNQDLWIALDQAIAR--H-KV-
DWRWVKGHTGHRENEICDQLAKQGA-----ENPT------LHDEGYQGE----------

>Proteobacteria gamma | Haemophilus infl | ZP_05851053

-----------------MFNLSLSIKIPAILHNNLFVMQKQIEIFTDGSCLG-NP-GA----GGIGAVLR---
Y-K-Q--------HEK-----MLSKGYFK-TTNNRMELRAVIEALNTL-------KEP------
CLITLYSDSQYMKNGIT-KWIFNWKKNNWKASSGK---PVKNQDLWIALDESIQR--H-KI-
NWQWVKGHAGHRENEICDELAKKGA-----ENPT------LEDMGYFEE----------

>Proteobacteria gamma | Haemophilus infl | ZP_04465570

------------------------------------MQKQIEIFTDGSCLG-NP-GA----GGIGAVLR---Y-K-Q--------
HEK-----TLSKGYFK-TTNNRMELRAVIEALNTL-------KEP------CLITLYSDSQYMKNGIT-
KWIFNWKKNNWKASSGK---PVKNQDLWKALDESIQR--H-KI-
NWQWVKGHAGHRENEICDELAKKGA-----ENPT------LEDMGYIKE----------

>Proteobacteria gamma | Pasteurella dagm | ZP_05920208

------------------------------------MQKQIEIFTDGSCLG-NP-GP----GGIGILLR---Y-K-Q--------
HEK-----QISKGYIQ-TTNNRMELRAVVEALNAL-------KEP------CTVTLHSDSQYMKNGIT-
KWIFNWKKNNWKASTGK---PVKNQDLWIALDQAIQR--H-NI-
NWQWVKGHSGHVENEICDELAKAGA-----EKPT------LEDVGYQPE----------

>Proteobacteria gamma | Pasteurella mult | NP_245044

------------------------------------MQKQIEIFTDGSCLG-NP-GP----GGIGVLLR---Y-K-Q--------
HEK-----QISAGYFL-TTNNRMELRAVIEALNTL-------KEP------CSVTLHSDSQYMKNGIT-

KWIFNWKKNNWKASTGK---PVKNQDLWIQLDQAIQR--H-HI-
NWQWVKGHSGHIENEICDQLAKAGA-----ENPT------LQDVGYQPE----------

>Proteobacteria gamma | Idiomarina loihi | YP_156076

----------MSNSKTVHLYTDGSCLG-NP-GP----GGYGAVLE---Y-G-K----
----HHK-----ELSQGYRL-TTNNRMEMLATIAGLREL-------KRS------
CHVILTTDSQYVKQGVE-QWMHRWKQNGWRTSARK---AVKNKDLWQQLDEEVNR--H-
KV-EWKWIKGHSGHKQNERCDELARDAA----TREPM------LEDEGFGGE----------

>Proteobacteria gamma | Idiomarina balti | ZP_01044146

----------MAKVHIFTDGSCLG-NP-GP----GGYGVVLE---Y-G-Q--------
HHK-----ELSGGYQC-TTNNRMELLACIKGLQVL-------NRA------CDVILTTDSQYVKQGIE-
QWIHNWKRNGWRTSNKK---AVKNVDLWQQLDQAIAA--H-KV-
TWEWVKGHAGHPQNERCDELARAAA-----EQNP-----TQVDTGFSTE----------

>Tenericutes | Candidatus Bloch | YP_277736

----------MYKKIEIFTDGSCLG-NP-GP----GGCAAILR---Y-K-Q--------
HKK-----EFSIGYRL-TTNNRMELMAAIIALESL-------KNP------CQIILNTDSQYLLHGIT-
QWIHIWKKHHWKTSEEK---LVKNIDLWQRLDVAIQI--H-
SIIHWNWLKSHTGHPDNERCDQLARLAAKCPINEDFY----------

>Tenericutes | Candidatus Bloch | NP_878522

----------MYKQIEIFTDGSCLG-NP-GP----GGCAGILR---Y-R-Q--------
YKK-----EFSAGYHI-TTNNRMELMAAIIALESL-------KNS------CQIILYSDSQYLLTGIT-
QWIQIWKKHHWKTADSK---LVKNIDLWRRLDIAIQP--H-
NIKDWRWLKSHTGHPDNERCDQLARKAA-----KYPLNKDFDNNPVVLYNDNDLMKID---
----

>Proteobacteria gamma | Tolumonas auensi | YP_002892125

----------MLKQITLYTDGSCLG-NP-GP----GGYAAVLI---Y-K-Q--------
HRK-----ELAQGYEL-TTNNRMELMAAIAGLQSL-------SEP------CQVRLTTDSQYVRQGIT-
QWIHGWKKKGWKTANRE---PVKNVDLWLLLDSEIQR--H-DV-
EWFWVKGHSGHPENERCDELARNAA----LADSR------LIDSGYPS----------

>Proteobacteria gamma | Aeromonas hydrop | YP_856105

----------MLKKIDLYTDGSCLG-NP-GP----GGYGAVMV---Y-G-K------
--HRK-----ELAGGFRL-TTNNRMELMAAIMGLRTL-------NEP------
CQVRLTTDSQYVRQGIT-QWIIGWKKKGWVTASRQ---PVKNVDLWQALDAEVAR--H-
QI-EWLWVKGHSGHPENERCDELAREAA-----SGKQ-----LAEDTGYQP----------

>Proteobacteria gamma | Aeromonas salmon | YP_001142554

----------MLKHIDLYTDGSCLG-NP-GP----GGYGAVLV---Y-G-D-------
-HRK-----EISGGFRL-TTNNRMELMAAIMGLRTL-------NAA------

CQVRLTTDSQYVRQGIT-QWIIGWKKKGWMTSNRQ---PVKNVDLWKELDAEVAR--H-
QI-EWLWVKGHSGHPENERCDELARDAA-----SGKE-----LAEDTGYQP----------

>Proteobacteria gamma | Alteromonadales | ZP_01614115

------------------------------------MQKTVEIYTDGSCLG-NP-GP----GGYGIFMI---Y-D-A--------
HEK-----KLSQGYKL-TTNNRMEMLAAIVALESL-------NRA------
CVVNLTTDSQYVKQGIE-SWISNWKKRGWITSAKK---PVKNVDLWKRLDAACSK--H-
TV-NWKWVKGHSGHKYNEIVDDLARDAA-----GSTD-----LLEDVGYQP----------

>Proteobacteria gamma | Pseudoalteromona | YP_340464

------------------------------------MEKTVEIYTDGSCLG-NP-GP----GGYGIFMI---Y-N-E--------
HEK-----KLSQGYKL-TTNNRMEMLGAIVALEVL-------TRP------CVINITTDSQYVKQGIE-
SWITNWKKRGWLTSAKK---PVKNVDLWKRLDLACAK--H-TV-
TWKWVKGHSGHKYNEIVDDLARDAA-----GSKD-----LLDDVGYQP----------

>Proteobacteria gamma | Shewanella denit | YP_563025

----------------------------------MMTTHKQVNIYTDGSCLG-NP-GP----GGYGIVMQ---Y-K-Q---
-----HSK-----EIADGFAL-TTNNRMELLAPIIALEAL-------MEP------
CIVTLTSDSQYMRQGIT-QWIHGWKKKGWMTSNKQ---AVKNVDLWKRLDSVSQR--H-
NI-DWRWVKGHTGHKQNERCDKLARDAA----EAKPK------QIDTGYQESL----------
-

>Proteobacteria gamma | Shewanella oneid | NP_718146

-----------------------------------MTELKLIHIFTDGSCLG-NP-GP----GGYGIVMN---Y-K-G-------
-HTK-----EMSDGFSL-TTNNRMELLAPIVALEAL-------KEP------CKIILTSDSQYMRQGIM-
TWIHGWKKKGWMTSNRT---PVKNVDLWKRLDKAAQL--H-QI-
DWRWVKGHAGHAENERCDQLARAAA----EANPT------QIDTGYQAES----------

>Proteobacteria gamma | Shewanella sp. W | YP_963624

-----------------------------------MTERKLIHIFTDGSCLG-NP-GP----GGYGIVMN---Y-K-G-------
-HTK-----EMSDGFAL-TTNNRMELLAPIIALESL-------KEP------CRVVLTSDSQYMRQGIM-
TWIHGWKKKGWMTSNRT---PVKNVDLWKRLDKVSQM--H-TI-
DWQWVKGHAGHAENERCDILARSAA----EANPT------QIDEGYQP----------

>Proteobacteria gamma | Shewanella balti | YP_001050372

-----------------------------------MTELKLIHIFTDGSCLG-NP-GP----GGYGIVMN---Y-K-G-------
-HTK-----EMSDGFAL-TTNNRMELLAPIIALESL-------KEP------CQVVLTSDSQYMRQGIM-
TWIHGWKKKGWMTSNRT---PVKNVDLWKRLDKASQM--H-TI-
DWQWVKGHAGHAENERCDVLARTAA----ESKPT------QPDLGYQP----------

>Proteobacteria gamma | Shewanella halif | YP_001674098

-----------------------------------MTGLKQISIYTDGSCLG-NP-GP----GGYGIVLK---Y-K-K-------
-QTK-----ELADGFAL-TTNNRMELLAPIVALEVL-------KVP------CQVILTSDSQYMRQGIT-

QWIHGWKRKGWLTSAGQ---PVKNVDLWKRLDTVSQR--H-QI-
DWRWVKGHAGHTENERCDDLARQAA----EAKPS------QEDSGYINQQAQA----------
-

>Proteobacteria gamma | Shewanella peale | YP_001502258

----------MNLYISQLNFLGILARNYSYGSNLMTGLKQISIYTDGSCLG-NP-GP----
GGYGIVLK---Y-K-K--------RTK-----ELADGFAL-TTNNRMEMLAPIIALEAL-------KVP------
CEVILTSDSQYMRQGIT-QWIHGWKRKGWMTSTNQ---PVKNVDLWKRLDTVSQR--H-
QV-EWRWVKGHAGHSENERCDDLARQAA----EAKPT------QEDNGYLAQQKQD----------
----

>Proteobacteria gamma | Shewanella amazo | YP_927756

----------MSELKQIRIYTDGSCLG-NP-GP----GGYGVVMI---Y-K-Q------
--HRK-----ELADGFAL-TTNNRMELLAPIVALESL-------KEP------CDVILTSDSQYMRQGIT-
EWIHGWKKKGWVTASKT---PVKNVDLWQRLDAAAAK--H-KV-
DWRWVKGHAGHAENERCDTLAREAA----EAKPT------QIDKGYQP----------

>Proteobacteria gamma | Shewanella frigi | YP_750889

----------MAELKQLYIFTDGSCLG-NP-GP----GGYGVVMK---Y-K-H----
----QQH-----EIADGFSL-TTNNRMELLAPIIALETL-------YEP------CNIILTSDSQYMRQGIM-
TWIHGWKKKGWITSTKQ---PVKNVDLWKRLDAVSQL--H-KI-
DWHWVKGHAGHIENERCDVLARKAA----EAKPQ------QVDTGYNPE----------

>Proteobacteria gamma | Shewanella loihi | YP_001094237

----------MHGLKQLLIFTDGSCLG-NP-GP----GGYGVVMK---Y-K-A----
----HVK-----ELSGGFAL-TTNNRMELLAPIMALEAL-------KEP------
CQIILTSDSQYMRQGIT-QWIHGWKKRGWLTAAKE---PVKNVDLWQRLDAATST--H-KI-
DWRWVKGHAGHIENERCDTLAREAA----EAGPS------EVDTGYQAKG----------

>Proteobacteria gamma | Shewanella benth | ZP_02159445

----------MLKPLSIFTDGSCLG-NP-GP----GGYGVVMQ---Y-K-S--------
RIK-----ELSDGFLL-TTNNRMELLAPIIALEAL-------KVP------CKIVLTSDSQYMRQGIT-
QWIHAWKKKGWQTAAKQ---PVKNVDLWKRLDAATAS--H-EI-
EWRWVKGHAGHVENERCDTLARVAA----EAKPT------QEDIGYPV----------

>Proteobacteria gamma | Shewanella sedim | YP_001473729

----------MMGMKQLSIFTDGSCLG-NP-GP----GGYGVVMK---Y-K-Q---
-----HTK-----EIADGFLL-TTNNRMELLAPIIALEAL-------KVP------
CKIVLTSDSQYMRQGIT-QWIHGWKKKGWITSSKQ---PVKNVDLWKRLDLASKG--H-EI-
DWRWVKGHAGHVENERCDTLAREAA----EAKPK------QEDIGYQA----------

>Proteobacteria gamma | Shewanella woody | YP_001760980

-----------------------------------MSVLKQLSIFTDGSCLG-NP-GP----GGYGVVMK---Y-K-A-----
---HTK-----ELSDGFAL-TTNNRMELLAPIIALEAL-------KVP------CKIILTSDSQYMRQGIT-
QWIHGWKKKNWITSTKQ---PVKNVDLWKRLDAATQS--H-EI-
DWHWVKGHAGHVENERCDTLARVAA----EAKPT------QEDLGYQPSVSSS----------

>Proteobacteria gamma | Psychromonas ing | YP_941950

------------------MQVLHNEVGGFIIINQFGIRMKKIQLFTDGSCLG-NP-GP----GGYGAVMI--
-Y-N-E--------HCK-----ELSEGFLL-TTNNRMEMLACIKALQSL-------TEP------
CEVELTTDSQYVRQGIT-LWIHNWKKRGWKTAAKA---PVKNVDLWKALDAAQEK--H-
KV-AWHWVKGHSGHPENERCDDLARRAA----ENNPT------QEDIGYEG----------

>Tenericutes | Candidatus Hamil | YP_002923568

-------------------------------------MKQVEIFTDGSCLG-NP-GA----GGYASILR---Y-Q-Q--------
HEK-----IFSQGYRL-TTNNRMELMASIVALQAL-------KSP------CTVTLFTDSQYVRQGIT-
QWVVHWKKRGWKTSERK---EVKNIDLWKALDAEIQK--H-QI-
NWQWVKGHAGHPENERCDKLARLAA-----SSPT------QEDRGYQGSC----------

>Proteobacteria gamma | Edwardsiella tar | YP_003296868

-------------------------------------MLKQVEIFTDGSCLG-NP-GP----GGYGAILR---Y-R-Q--------
HEK-----ALSAGYRL-TTNNRMELMAAIVALETL-------TSA------CQVTLFSDSQYVRQGIT-
QWIHGWKRRGWKTADKK---PVKNVDLWQRLDQAIGP--H-QV-
EWIWIKGHAGHPENERCDELARSAA-----GAPS------LEDSGYNPE----------

>Proteobacteria gamma | Sodalis glossini | YP_454271

-------------------------------------MRKQVAIFTDGSCLG-NP-GP----GGYGAILR---Y-K-Q--------
HEK-----TFSAGYRL-TTNNRMELMAAIVALEAL-------TDA------
CEVVLSTDSQYVRQGIT-QWIHNWKKRGWKTAYKK---PVKNVDLWQRLDAAIQP--H-
TL-RWDWVKGHSGHPENERCDELARTAA-----CHPA------LEDIGYRVEAQTSGGRAD-----
----

>Proteobacteria gamma | Erwinia pyrifoli | YP_002649744

-------------------------------------MLKQVEIFTDGSCLG-NP-GP----GGYGAIMR---Y-G-K--------
HEK-----IFSAGFHL-TTNNRMEMMAAIVALEAL-------TQP------CAVVLSTDSQYVRQGIT-
SWIHNWKKRGWKTADKK---PVKNVDLWKRLDAALSH--H-DI-
NWKWVKGHAGHVENERCDVLARTAA-----GCPT------FDDVGYQA----------

>Proteobacteria gamma | Yersinia ruckeri | ZP_04616430

-------------------------------------MTKQVEIFTDGSCLG-NP-GP----GGYGAILR---Y-K-Q--------
HEK-----TFSAGYRL-TTNNRMEMMAAIVALEAL-------TSP------CEITLSTDSQYVRQGIT-
QWIHNWKKRGWKTSDRK---PVRNVDLWQRLDAAILG--H-NV-
QWEWVKGHAGHPENERCDVLARDAA-----NAPT------LEDTGYNPD----------

>Proteobacteria gamma | Yersinia frederi | ZP_04633911

----------MSLPEMTKQVEIFTDGSCLG-NP-GP----GGYGAILR---Y-K-Q----
----HEK-----TFSAGYFL-TTNNRMELMAAIVALEAL-------TSP------
CKVTLSTDSQYVRQGIT-QWIHNWKKRGWKTTDRK---PVRNVDLWQRLDLAIQT--H-
TV-QWEWVKGHAGHPENERCDELARQGA-----NSPT------LEDSGYNPD----------

>Proteobacteria gamma | Yersinia aldovae | ZP_04620327

----------MTKQVEIFTDGSCLG-NP-GP----GGYGAILR---Y-K-Q--------
HEK-----MFSAGYYL-TTNNRMELMAAIVALEAL-------TSP------CEVTISTDSQYVRQGIT-
QWIHNWKKRGWKTTDRK---PVRNMDLWQRLDLAIQT--H-TI-
QWEWVKGHAGHPENERCDELARLAA-----NSPT------QDDVGYNPD----------

>Proteobacteria gamma | Serratia proteam | YP_001477145

----------MLKQVEIFTDGSCLG-NP-GP----GGYGAILR---Y-K-Q--------
TEK-----TFSAGFRL-TTNNRMEMMAAIVALEAL-------TTP------CEVTLSTDSQYVRQGIT-
TWIHNWKKRGWKTADKK---PVKNVDLWQRLDLAIQR--H-TV-
KWEWVKGHAGHPENERCDVLARDAA-----SNPT------QDDVGYKPES----------

>Proteobacteria gamma | Serratia odorife | EFA15383

----------MLTALVYVRLGFYKTDRKSLPEMLKQVEIFTDGSCLG-NP-GP----
GGYGAILR---Y-K-Q--------VEK-----TFSAGYRL-TTNNRMELMAAIVALEAL-------TAP-----
-CEVTLSTDSQYVRQGIT-SWIHNWKKRGWKTADKK---PVKNVDLWQRLDLAIQT--H-
TI-KWEWVKGHAGHPENERCDVLARDAA-----GNPT------QDDVGYKPEN----------
-

>Proteobacteria gamma | Pantoea sp. At-9 | ZP_05732207

----------MRKQVEIFTDGSCLG-NP-GP----GGYGAILR---Y-R-Q--------
HEK-----TFSAGYRL-TTNNRMELMAAIVALEAL-------TQP------CEVVISTDSQYVRQGIT-
SWIHNWKKRGWKTADKK---PVKNVDLWQRLDLALSS--H-QI-
VWEWVKGHAGHPENERCDELARSAA-----SQPT------QDDVGYQPES----------

>Proteobacteria gamma | Cronobacter turi | YP_003209201

MWFIPVVSVSRTICRVIAVLIALVYVRLGFLLTGSLPEMRKQVEIFTDGSCLG-NP-GP----
GGYGAILR---Y-K-Q--------HER-----TFSAGYRL-TTNNRMELMAAIVSLEAL-------REH-----
-CIVTLSTDSQYVRQGIT-QWIHNWKKRGWKTAEKK---PVKNVDLWQRLDAALSQ--H-
EI-KWEWVKGHAGHPENERCDELARAAA-----MAPT------LEDTGYQPEATAS----------
---

>Proteobacteria gamma | Enterobacter sp. | YP_001175485

----------MTKQVEIFTDGSCLG-NP-GP----GGYGAILR---Y-R-G--------
HEK-----TFNEGYHL-TTNNRMELMAAIVALEAL-------KED------CDVVISTDSQYVRQGIT-
QWIHNWKKRGWKTADKK---PVKNVDLWKRLDAALSH--H-TI-
KWEWVKGHAGHPENERCDELARAAA-----MNPI------QEDVGYQPGS----------

>Proteobacteria gamma | Klebsiella pneum | YP_001333913

------------------------------------MLKQVEIFTDGSCLG-NP-GP----GGYGAIMR---Y-R-Q--------
HEK-----TFSAGYRL-TTNNRMELMAAIVALEAL-------KEH------
CEVVLSTDSQYVRQGIT-QWIHNWKKRGWKTAEKK---PVKNVDLWQRLDAALGQ--H-
KI-KWEWVKGHAGHPENERCDELARAAA-----SHPT------LDDVGYLPES----------

>Proteobacteria gamma | Citrobacter sp. | ZP_04560679

------------------------------------MLKQVEIFTDGSCLG-NP-GP----GGYGAILR---Y-R-G--------
REK-----TFSEGYNL-TTNNRMELMAAIVALEAL-------KEQ------CEVILSTDSQYVRQGIT-
QWIHNWKKRGWKTADKK---PVKNVDLWKRLDAALGP--H-QI-
KWEWVKGHAGHPENERCDELARTAA-----MSPT------QDDIGYQTEA----------

>Proteobacteria gamma | Salmonella enter | ZP_03355493

------------------------------------MLKQVEIFTDGSCLG-NP-GP----GGYGAILR---Y-R-G--------
HEK-----TFSEGYTL-TTNNRMELMAAIVALEAL-------KEH------CEVTLSTDSQYVRQGIT-
QWIHNWKKRGWKTAEKK---PVKNVDLWKRLDAALGQ--H-QI-
KWVWVKGHAGHPENERCDELARAAA-----MNPT----Q------------------------------

>Proteobacteria gamma | Escherichia coli | NP_285902

------------------------------------MLKQVEIFTDGSCLG-NP-GP----GGYGAILR---Y-R-G--------
REK-----TFSAGYTR-TTNNRMELMAAIVALEAL-------KEH------CEVILSTDSQYVRQGIT-
QWIHNWKKRGWKTADKK---PVKNVDLWQRLDAALGQ--H-QI-
KWEWVKGHAGHPENERCDELARAAA-----MNPT------LEDTGYQVEV----------

>Proteobacteria gamma | Pectobacterium w | YP_003258564

------------------------------------MRKQVEIFTDGSCLG-NP-GP----GGYGALLR---Y-K-Q-------
-HEK-----ALSAGYRL-TTNNRMELMAAIAALETL-------TTD------
CDVVLSTDSQYVRQGIT-SWIHNWKKRGWKTADKK---PVKNVDLWKRLDTAIQR--H-
SV-RWEWVKGHAGHPENERCDELARAAA-----SAPT------LDDTGYQAE----------

>Proteobacteria gamma | Dickeya dadantii | YP_002988438

------------------------------------MLKQVEIFTDGSCLG-NP-GP----GGYGALLR---Y-K-Q--------
HEK-----TLSGGYRL-TTNNRMELMAAIAALETL-------TTE------CEVTLSTDSQYVRQGIT-
QWIHNWKKRGWKTTEKK---PVKNADLWQRLDTAVQR--H-HL-
HWKWIKGHSGHPENERCDVLAKQAA-----NNPT------QEDTGYQPD----------

>Proteobacteria gamma | Dickeya dadantii | ZP_05725899

------------------------------------MLKQVEIFTDGSCLG-NP-GP----GGYGALLR---Y-K-Q--------
HEK-----TLSAGYRL-TTNNRMELMAAIVALESL-------TSP------CEVTLSTDSQYVRQGIT-
SWIHNWKKRGWKTAEKK---PVKNIDLWQRLDVAIQR--H-TL-
HWMWVKGHAGHPENERCDELARQAA-----NMPT------LDDTGYQPE----------

>Proteobacteria gamma | Providencia rett | ZP_06126237

------------------------------------MTKQVEIFTDGSCLG-NP-GP----GGYGVVLR---Y-Q-Q-------
-HEK-----TLSEGYFL-TTNNRMELLAAIKALESL-------TRP------CDIILTTDSQYVRQGIT-
QWIHGWKRKQWRKADKS---PVVNVDLWKRLDDAIQR--H-TI-
DWRWVKGHAGHPENEKCDELARAAA-----SAPT------KEDTGYQPAQN----------

>Proteobacteria gamma | Providencia rust | ZP_05973989

------------------------------------MTKQVEIFTDGSCLG-NP-GP----GGYGIVLR---Y-Q-Q--------
HEK-----TLSDGFFL-TTNNRMELLAAIIALESL-------TQP------CDVILTTDSQYVRQGIT-
QWIHNWKRRQWKKADKS---PVVNVDLWKRLDQAITR--H-TI-
DWRWVKGHAGHAENEKCDELARAAA-----NSPT------KEDTGYQPAQE----------

>Proteobacteria gamma | Photorhabdus lum | NP_928278

------------------------------------MGKQVEIFTDGSCLG-NP-GP----GGYGVLLR---Y-Q-Q-------
-HEK-----TLSEGFYH-TTNNRMELMAAIIGLETL-------TRP------CKIVLTTDSQYVRQGIT-
QWIHNWKKRGWRKADKS---PVSNVDLWQRLDQAISR--H-NI-
DWQWVKGHAGHDENERCDELARAAA-----NSPT------ETDTGYLENRD----------

>Proteobacteria gamma | Proteus mirabili | YP_002150008

------------------------------------MHKQVEIFTDGSCLG-NP-GP----GGYGAILR---Y-Q-Q--------
HEK-----TLSEGFFM-TTNNRMELLAAIVALEAL-------KFP------CKITLTTDSQYVRQGIT-
KWIHSWKKRQWRKADKS---PVLNVDLWKRLDKAIER--H-EI-
EWHWVKGHAGHDENERCDELAKAAA-----QSPT------KEDTGYLESQQDKT----------

>Proteobacteria gamma | Grimontia hollis | ZP_06054290

--MR---Y-K-Q--------HEK-----
ELSEGFSL-TTNNRMELLAAIVGLASL-------KES------CNVTLTTDSQYVRQGIT-
QWIHNWKKRDWKTADKK---PVKNADLWQRLDSETQR--H-TV-
DWQWVKGHAGHPENERCDELARTAA-----ENPT------SPDTGYQPDA----------

>Proteobacteria gamma | Vibrio fischeri | YP_205319

---------------------------------MITEIMKQVEIFTDGSCLG-NP-GP----GGYGIVMR---Y-K-G-----
---TEK-----TFSEGFNK-TTNNRMEMLAAVVALRKL-------KEP------
CSVILTTDSQYVRQGIT-QWIHGWKKRDWKKADKK---PVVNADLWKQLDAESER--H-
KI-DWRWVKGHAGHRENEMCDELARTAA-----ENPT------QDDTGYPG----------

>Proteobacteria gamma | Aliivibrio salmo | YP_002263754

---------------------------------MITEIMKQVEIFTDGSCLG-NP-GP----GGYGIVMR---Y-K-G-----
---TEK-----TFSGGFNQ-TTNNRMEMLAAVVALRNL-------KEP------
CIVVLTTDSQYVRQGIT-QWIHGWKKRGWKKADKK---PVVNADLWKQLDAEAER--H-
TV-DWRWVKGHAGHRENEMCDDLARTAA-----ENPT------QDDTGYPG----------

>Proteobacteria gamma | Vibrio furnissii | ZP_05876735

------------------------------------MKKQVEIFTDGSCLG-NP-GP----GGYGVVMR---Y-K-Q-------VEK-----TLAKGYRL-TTNNRMEMMAAVVALKTL-------KEP------
CHVSLTTDSQYVRQGIT-QWIHNWKKRGWKTADKK---PVKNADLWQALDAETAR--H-QV-EWHWVKGHAGHRENEMCDELARSAA-----ENPT------EDDVGYQPEK----------

>Proteobacteria gamma | Vibrio cholerae | NP_231865

------------------------------------MNKQVEIFTDGSCLG-NP-GP----GGYGIVMR---Y-K-Q-------VEK-----TLARGYRL-TTNNRMEMLAAVMALQAL-------KEP------
CRVILTTDSQYVRQGIT-QWIHNWKLRGWKTADKK---PVKNADLWQALDKETAR--H-QV-EWRWVKGHAGHRENEMCDELARQAA-----ENPT------EDDIGYQPEPQ----------

>Proteobacteria gamma | Vibrio metschnik | ZP_05881164

---M----------
EMLAAVIALQSL-------KEP------CDVILTTDSQYVRQGIT-QWIHNWKQRGWKTADKK---PVKNADLWQALEKETAR--H-QV-DWRWVKGHAGHRENEMCDQLARSAA-----ENPT------EDDVGYQP----------

>Proteobacteria gamma | Vibrio vulnificu | NP_760761

------------------------------------MTKQVEIFTDGSCLG-NP-GP----GGYGVVLR---Y-K-Q-------VEK-----TLAQGYRL-TTNNRMEMMATIVALQAL-------KEP------
CNVILTTDSQYVRQGIT-QWIHNWKKRGWKTADKK---PVKNADLWQALDKETTR--H-TI-DWRWVKGHAGHRENEMCDELARAAA-----ENPT------LDDTGYQPAE----------

>Proteobacteria gamma | Vibrio coralliil | ZP_05884158

------------------------------------MTKQVEIFTDGSCLG-NP-GP----GGYGIVLR---Y-K-Q--------VEK-----TLAKGYTL-TTNNRMEMMATIVALQAL-------KEP------
CDVILTTDSQYVRQGIT-QWIHNWKKRDWKTSDKK---PVKNADLWKALDAETGR--H-KI-DWRWVKGHAGHRENEMCDELARAAA-----ENPT------DEDTGYQPS----------

>Proteobacteria gamma | Vibrio orientali | ZP_05946276

------------------------------------MTKQVEIFTDGSCLG-NP-GP----GGYGVVLR---Y-K-Q-------TEK-----TLAKGYTM-TTNNRMEMLATIVALQAL-------KES------
CDVILTTDSQYVRQGIT-QWIHNWKKRGWKTADKK---PVKNADLWKALDQETER--H-TV-DWRWVKGHAGHRENEMCDELARGAA-----ENPT------EEDTGYIPN----------

>Proteobacteria gamma | Vibrio parahaemo | ZP_05120769

------------------------------------MTKQVEIFTDGSCLG-NP-GP----GGYGIVLR---Y-K-Q--------TEK-----TLAKGYTL-TTNNRMEMLATIVALQAL-------KEP------CDVILTTDSQYVRQGIT-QWIHNWKKRGWKTADKK---PVKNADLWKALDAESER--H-NI-DWRWVKGHAGHRENEMCDELARTAA-----ENPT------EEDTGYIPN----------

>Proteobacteria gamma | Vibrio sp. MED22 | ZP_01065183

------------------------------------MTKQVEIFTDGSCLG-NP-GP----GGYGIVLR---Y-K-K--------
VEK-----TLAEGFTL-TTNNRMEMLAAVVALQAL-------KEP------CSVILTTDSQYVRQGIT-
QWIHNWKKRDWKTADKK---PVKNADLWQRLDKETAR--H-SV-
DWRWVKGHAGHRENEMCDDLARSAA-----ENPT------QEDTGYQPS----------

>Proteobacteria gamma | Vibrio harveyi 1 | ZP_06174988

------------------------------------MTKHVEIFTDGSCLG-NP-GP----GGYGIVLR---Y-K-Q--------
TEK-----TLAKGYTL-TTNNRMEMLAAVVALQTL-------KEP------
CQVTLTTDSQYVRQGIT-QWIHNWKKRGWKTADKK---PVKNADLWQALDKETAR--H-
QV-DWHWVKGHAGHRENEICDELARTAA-----ENPT------EEDTGYQAS----------

>Proteobacteria gamma | Glaciecola sp. H | ZP_03560606

-------------------------------------MQEVQIFTDGSCLG-NP-GP----GGYGAIMV---Y-G-K--------
HRK-----EIAEGYFA-TTNNRMELLAPIKALSLL-------KKP------CRVILTTDSQYVKNGIN-
QWIHNWRKNGWKTSNKQ---PVKNADLWMALDEAVKG--H-HI-
DWRWVKGHSGHPENERCDELARHAA-----EAAA--KGSGQDDNGYQPA----------

>Proteobacteria gamma | Pseudoalteromona | YP_661935

-------------------------------------MKHIEIYTDGSCLG-NP-GP----GGYGAVLL---F-N-Q--------
HSK-----ELSQGFVH-TTNNRMELLATIEALASL-------TET------CKVDLTTDSQYVKNGIN-
QWIKNWRKNGWRTSDKK---PVKNVDLWKRLDEQVGR--H-DV-
KWHWVKGHSGHPMNERCDVLARDAA-----SGKS-----LLPDEGFQG----------

>Proteobacteria gamma | Alteromonas macl | ZP_04714008

----------------------------------NAVAQKTIHIYTDGSCLG-NP-GP----GGYGAVLI---Y-K-Q------
--HRK-----ELSDGFAH-TTNNRMELLAPIEALNSL-------NEP------
CNVELTTDSQYVKNGIN-QWIHNWRKNGWRTADKK---PVKNADLWQRLDEAVKK--H-
KI-NWHWVKGHSGHPENERCDDLARGAA----EANPT------KPDEGFVGK----------

>Proteobacteria gamma | Alteromonas macl | YP_002126679

----------------------------------MAQKTIHIYTDGSCLG-NP-GP----GGYGAVLI---Y-K-Q-------
-HKK-----ELSDGFAH-TTNNRMELLAPIEALNSL-------TEP------CAVELTTDSQYVKNGIN-
QWIHNWRKNGWRTSDKK---PVKNADLWQRLDEAVKK--H-QV-
NWHWVKGHSGHPENERCDELARGAA----EAKPT------QIDEGFVGN----------

>Proteobacteria delta | Desulfovibrio ma | YP_002952549

-----------------------------MTEETKAPQQNVIIFTDGACLG-NP-GP----GGYGAVLL---R-G-D-
-------ERR-----EFSGGRKL-TTNNRMELLACIVALEEL-------VEP------
SVVSITTDSRYVHDAIEKRWLASWQKKGWVNSEKK---PVKNQDLWLRLLPLLSR--H-
KV-KFSWVRGHTGHPENERCDVLARQAA-----NSRG-----LEADAGYPG----------

>Proteobacteria delta | Lawsonia intrace | YP_595131

----------MRSNYLKSVEVFTDGSCLG-NP-GA----GGWAAILR---Y-G-D---
-----YEQ-----EISGGFSY-TTNNRMEMIAAIYALEKL-------KES------
CLVMLYTDSQYLRNAVEKQWLVFWEKNNWKTASKK---PVKNQDLWKRLQRQLER--
H-NV-IFTWVRGHSGHFENERCDNLARMEA-----SRSN-----LPKDCGFINEG----------
--

>Proteobacteria delta | Desulfovibrio de | YP_002478635

----------MQNVTIHTDGSCLG-NP-GP----GGWAAILR---LDEGD-------
--HRK-----EFSGGYAL-TTNNRMEMLAVIEALALL-------KSP------
CTVDLYTDSRYVCDSVSKGWLWGWVKKNWIKSDKK---PVLNVDLWQRMLPLLRQ--H-
KV-NFHWLKGHAGHPENERCDVLARAQA-----SRRD-----LPPDTGYKP----------

>Proteobacteria delta | Desulfovibrio de | YP_389430

----------MKQVDIFTDGSCLG-NP-GP----GGWAAVLR---Y-A-G-------
-TQK-----ELGGGFSG-TTNNRMEILAVIEGLEAL-------QEP------
CTVNLYTDSQYVRNAVEKKWLDSWQRNGWKTAARK---PVKNKDLWLRLLPLLAR--H-
TV-KFHWVRGHSGHPENELCDTIARGHA-----SRGG-----LPPDTQAAG----------

>Proteobacteria delta | Desulfovibrio vu | YP_009911

----------MSQFDVTVFTDGSCLG-NP-GP----GGWAAIMR---C-N-G------
---CEK-----ELSGGFAL-TTNNRMEILAVLEALEAL-------RDP------
CKVTLFTDSQYVRNAVEKKWLAGWQRNGWKTADKK---PVKNRDLWERLVPLLAK--H-
SV-SFRWVRGHSGHPENERCDVLARAQA-----SRRG-----LPEDPGFTA----------

>Proteobacteria delta | Desulfovibrio vu | YP_002437076

----------MTMKNVQAFTDGSCLG-NP-GP----GGWAAVLR---C-N-G-----
----SER-----ELSGGFAL-TTNNRMEILAVIEALALL-------KEP------
CGVDLYTDSQYVRNAVEKKWLAGWRRNGWKTSDKK---PVKNRDLWERLQPLLDL--H-
QV-RFHWVRGHSGHPENERCDVLARTQA-----SSRG-----LPPDTGYRE----------

>Lentisphaerae | Victivallis vade | ZP_01924103

----------MQGPHSPKKETCQIVKSVQIYTDGACKG-NP-GP----GGYGAVLL---Y-
K-T--------YRR-----ELSGGFRH-TTNNRMEIFAAIAAVELL-------NEP------
CEITLYSDSSYLVNAVTKRWLYNWKRSGWVKRDGQ---PVNNIDLWKRFLAAVEP--H-
KL-HMVWVKGHADNVENSRCDALAVAAA-----ARRN----ALPPDTGFR----------
-

>Chloroflexi | Chloroflexus agg | YP_002463237

----------AAVSPDTVVMYTDGSALG-NP-GP----GGYGVVLR---Y-N-Q---
-----HYK-----ELSGGFRR-TTNNRMELMACIAGLRAL-------KRP------
MRVVIYSDSKYVVDAVQEGWVQRWQAKNWMRTSTE---PAQNADLWAELVQLCTI--H-
QV-QFVWVPGHSGVPDNERCHQLATAAA-----QQPNLPPDIGFEQADEQKP----------
----

>Lentisphaerae | Lentisphaera ara | ZP_01873307

-----MKKEVLLATDGACKG-NP-GP----GGYGTILI---F-N-Q--------
YRK-----EFAEGFRL-TTNNRMEMLAVIKGLEAL-------KES------
CKVKVLSDSKYIVDNVKGGHPWKWQARGWVLASKK---PAKNSDLWEDLLNLLAK--H-
EV-EFEWVKGHSGHELNDRADELATGAA-----EQGT-----LLEDYGFEK----------

>Cyanobacteria | Synechocystis sp | NP_442483

----------MASTPNSVTLYTDGACSM-NP-GP----GGYGAVIL---Y-G-D---
G----RRE-----ELSAGYKM-TTNNRMEIMGAIAALSHL-------QEP------
SQVLLYTDSRYMVDAMSKGWAKKWKANGWQRNAKE---KAKNPDLWETMLTLCEK--
H-QV-TFQWVKAHAGNKENERCDRLAVAAY----QNNPN------LVDEGFGKF----------
-----

>Firmicutes | Desulfotomaculum | YP_003190872

----------MSQVEIYTDGACSG-NP-GP----GGYGVVLK---Y-G-D--------
KIK-----ELSAAYRK-TTNNRMEILAAIIGLEAL-------RRP------
CTVTLYSDSQYLVNAMTKGWVKRWKANNWMRNKQE---AAKNIDLWERMLPLLEQ--
H-QV-DWVWVKGHADNYYNNRCDFLAVRAI-----KEQA-----LLEDEGFKK----------
----

>Cyanobacteria | Trichodesmium er | YP_721337

----------MTEKRTEITIYTDGACSG-NP-GP----GGYGIIIL---S-E-K--------
KRQ-----ELSGGYKL-TTNNRMELMAVIVGLEQL-------EIP------
SIVNLYTDSKYIVDAVTKGWAKRWRANSWKRNKKD---KAMNPDLWGKLLDLCSK--H-
QV-EFSWVRGHSGNIENERCDKLAVKAS-----QKLD-----LPSDLGYQ----------

>Cyanobacteria | Lyngbya sp. PCC | ZP_01620565

----------NSSKLQEVILYTDGACQG-NP-GP----GGYGIVLI---R-G-D-------
-HRE-----ELSGGFQF-TTNNRMEMMAAIVGLEVL-------DKK------
SKVKLYSDSKYVVDAIEKGWAERWQANGWKRNKKE---LAMNPDLWEQLLKLCSQ--H-
QV-KFVWVKGHAGNRENECCDRLAVQGC-----QQQN-----LLQDVGYENPEMQQISLF-----
----

>Proteobacteria delta | Syntrophus acidi | YP_462765

----------MKATSKAKTHPPGATAAKDPQKQVIIYTDGACLG-NP-GP----
GGYGVVLL---Y-G-E--------HRK-----ELSGGYRL-TTNNRMEILAAIKGLEAL-------KSA-----
-CSVTLYSDSQYLVNAINKGWAQRWKANGWKRNARE---KALNPDLWERLLELCSR--H-
DI-TFVWVRGHANNKENERCDVLSKEAA-----GRAD-----LKADPGYP----------

>Cyanobacteria | Arthrospira maxi | ZP_03275282

----------MNMGITKVTIYTDGACSG-NP-GK----GGYGAVLM---C-G-S---
-----HRK-----EISGGFRL-TTNNRMEMMAAIAALRAL-------KFP------
CSVTLYSDSKYLVDAMTLGWAKRWQKNGWRRNQKE---WAKNPDLWAQLLGLCEE--

H-QV-RFVWVKGHAGDRENEICDRLAVEAT-----HRDS-----LPPDAGYENPPQPQDIDSMS-
----

>Cyanobacteria | Cyanothece sp. P | YP_002374450

----------MNDSPKKVLIYTDGACSG-NP-GS----GGYGTVLI---Y-N-N------
---HRK-----ELSGGFRL-TTNNRMEMMAAIVGLETL-------TIK------
CAVTLYTDSRYLVDAITKGWAKKWKANGWKRNAKE---NAKNPDLWEKLLDLCSQ--H-
EV-DFVWVKGHAGHQENEYCDRLAVRAS-----QQTNLPSDEVYENKGIET----------
----

>Cyanobacteria | Crocosphaera wat | ZP_00515066

----------MNKVQIYTDGACSG-NP-GK----GGYGIILA---Y-N-E--------
HRK-----ELSGGYRL-TTNNRMEMMAAIIALEAL-------NKP------
CDVILYTDSRYVVDAITKGWAKKWQANDWQRNKKE---QAKNPDLWQRLLDLCEQ--H-
QV-EFVWVKGHAGHPENEQCDRLAVAAC-----QEVELSIDAVYEEQK----------
-

>Cyanobacteria | Cyanothece sp. C | ZP_01729710

----------MKKVQIYTDGACSG-NP-GK----GGYGIILV---Y-N-E--------
HRK-----ELSGGYRL-TTNNRMEMMAAIIGLEAL-------KTP------
CEVTLYTDSRYLVDAITKGWAKKWQANGWKRNNKE---AAKNPDLWQKLLDLCKK--H-
EV-KFVWVKGHAGHPENEQCDRLAVTAT-----QQLT-----LAIDEVYEL----------

>Cyanobacteria | Cyanothece sp. A | YP_001805804

----------MKKVQIYTDGACSG-NP-GK----GGYGIILV---H-N-E--------
HRK-----ELSGGYRL-TTNNRMEMMAAIIGLEAL-------KMP------
CDVTLYTDSRYLVDAITKGWAKKWQGNGWKRNKKE---TAKNPDLWQKLLDLCEE--H-
EV-EFVWVKGHAGHPENEQCDRLAVTAA-----QQSELAIDEVYEEY----------

>Firmicutes | Desulfotomaculum | YP_001113791

----------MNTNQNTNLKEITMYTDGACSG-NP-GP----GGYGVVML---Y-K-
G--------HRK-----ELSAGFRD-TTNNRMELLATIVGLETL-------KEK------
CNVNLYTDSQYVVNAIEKGWAKKWRANGWMRNKKE---PALNPDLWERLLKLCEF--H-
NV-KFNWVKGHAGHPENERCDQLAVAAA-----KQPN-----LPLDVR----------

>Firmicutes | Heliobacterium m | YP_001681348

----------MTQAKRKEVTIYTDGACLG-NP-GP----GGYGAVLI---Y-G-E---
-----HRK-----ELSEGFRD-TTNNRMEMLAAIKALEAL-------KEP------
CQVVLYSDSRYLVDAVTQGWARRWKANGWMRNKKD---PALNVDLWERLLQLLER--
H-QV-EFRWVKGHAGNPENERCDKLATAAA-----ARPD------LPLDGRC----------

>Proteobacteria epsilon | Heliobacillus mo | AAN87534

----------MIHNDKKRSGPLYRWVHCSG-NP-GP----GGYGVVLI---Y-G-E--
------HRK-----EMSGGYQD-TTNNRMEMLAAIRGLEAL-------KEP------
CRVTLYSDSRYLVDAVKQGWARRWKANNWMRNKKD---PALNVDLWKKLLDLLDK--
H-DV-DFQWVKGHAGHPENERCDVLATSAA-----AKGD-----LPPDIRG----------

>Bacteroidetes | Blattabacterium | YP_003284123

----------MNQKIHIYTDGSSKG-NP-GP----GGYGIFIE-TTI-G-N---SY---
NRK-----IISEGFRY-TTNNRMELLAVIVGLEKI-------EKR-----
KQNIVVFTDSKYIVNTIQNNWIHQWKKNNFFQKK--------NVDLWKRFLKIYNK--N-II-
DFQWIKSHNNHYINDYCDRLSVEAS-----KRKILKIDYIYEKQNKSL----------

>Bacteroidetes | Chryseobacterium | ZP_03854256

----------MRIEIYTDGACSG-NP-GK----GGYGILMR---V-P-E---KN---
YQK-----TFSRGFRK-TTNNRMELLAVITALEKL-------KST-----
ENEIHIYTDSKYVSDAINQNWIAGWIKRGWK--------NVKNPDLWKKFVELYNK--H-NP-
KMHWIKGHAGHFENELCDKLAVAAA-----NSSDLEIDTYFENLDNNSLF----------

>Bacteroidetes | Flavobacteriacea | YP_003096853

----------MSLRIEIYTDGACSG-NP-GK----GGYGIVMK---V-P-E---KN-
--YEK-----HFSKGFRL-TTNNRMELLAVIVALEKL-------KSP-----
DNDIHIYTDSKYVSDAINKKWLLGWIKKGYK--------NVKNPDLWRRMVPLLAT--H-KT-
TFHWIKGHAGHPENEICDQLAVKAA-----QSGKLETDQYFEDQKNGGLF----------

>Bacteroidetes | Pedobacter sp. B | ZP_01884505

----------MIEIYTDGAASG-NP-GP----GGYGVILR---S-G-N--------
HYK-----ELSGGFRM-TTNNRMELLAVIVGLNAL-------KTP-----
GQEVMIFSDSKYVVDSVEKKWVFGWVKKGFK--------DKKNKDLWLRFLEVYKL--H-
QV-RFTWIKGHNAHPENERCDVLAVAAS-----KNKAA-L---AIDAPFEAEKNSQRLL---------
------

>Bacteroidetes | Pedobacter hepar | YP_003093229

----------MIEIYTDGAASG-NP-GP----GGYGVILR---S-G-N--------
HYK-----ELSAGFRL-TTNNRMELMAVIVGLNAL-------KTP-----
GQEVTVFSDSKYVIDSVEKKWVFGWVKTGFK--------GKKNKDLWMQFLNSYKL--H-
HV-KFVWIKGHNNHPENERCDQLAVAAS-----KNRAA-L---AIDGPFEAEKNSASLL---------
------

>Bacteroidetes | Sphingobacterium | ZP_04781167

----------MIELYTDGASSG-NP-GP----GGYGTILRTR-YSG-
ENEAFKGKLIEK-----TFSEGFRR-TTNNRMELMAVIIGLEAL-------KSP-----
QQQVTIYSDSKYVIDAIDKKWVYGWIQKGFQ--------GKKNKDLWIRLMKSYKL--H-QV-
RLVWVKGHAGHPDNERCDQLAVAAS-----KDKAN-W---KIDAVFEQEEKALG----------
---

>Bacteroidetes | Flavobacteria ba | ZP_03702482

----------MKSKPVYLYTDGSSLG-NP-GP----GGYGLRLE---W-A-E---
MS---YVK-----EFSQGFVR-TTNNRMELLAVIVGLELL-------KKQ-----
PLEVVVFSDSKYVIDSVDKKWVFGWEKKAFK--------DKKNSDLWKRFLKIYRK--H-NV-
NFQWIKGHNQHPQNERCDELAVIAA-----KGKNLIPDVFFEQIEKENSKA----------

>Bacteroidetes | Flavobacteria ba | ZP_03700731

----------MHKADVHVYTDGAASG-NP-GP----GGYGIVME---W-V-G---
TP---YKK-----EFSQGFTH-TTNNRMELLAVIEALRKL-------KKA-----
PLKVLVFTDSKYVVDAVEKKWLQRWVKTNFK--DKK------NVDLWKAFLKEYPK--H-
EV-RFQWIKGHNNHPQNERCDVLAVAAS-----KGKD-----LYIDSGFVKTT----------

>Bacteroidetes | Gramella forseti | YP_863387

----------MQTPKVHIYTDGAARG-NP-GP----GGFGVVME---W-V-G---
KP---YKK-----EYAQGFKL-TTNNRMELMAVIVAISKL-------KNP-----
GTPAKVFTDSKYVADAVNKGWVFNWEKKNFV--NRK------NTDLWKAFLKVFRR--H-
EV-QFQWIKGHNDHPQNERCDALAVMAS-----KGKD-----LLEDTGYKA----------

>Bacteroidetes | Kordia algicida | ZP_02161748

----------MVDVHIYTDGSSRG-NP-GP----GGYGIVME---W-V-G---
KP---YHK-----EFSEGYRK-TTNNRMELLAVIVALEKL-------KFM-----
HTEAKVFTDSKYVVDSVEKKWVFGWEKKGFS--GKK------NADLWMRFLKIYRK--H-IV-
HFQWIKGHNNHPQNERCDFLAVEAS-----KKEKLKIDTFYESESNRLF----------

>Bacteroidetes | Flavobacteria ba | ZP_01733613

----------MSHEVHIYTDGAAKG-NP-GP----AGYGVVME---M-V-G---
TP---YKK-----EFYEGFRL-STNNRMELLAVIVGLEKL-------KNP-----
KTKVLVVSDSKYVVDSVEKRWVFQWEKINFK--AKK------NPDLWMRFLKIYRQ--H-QV-
DFQWVKGHNSHPQNERCDELAVMAS-----QQEKLSIDEFYEREEEKLL----------

>Bacteroidetes | Flavobacterium p | YP_001295134

----------MNYQVHIYTDGAAKG-NP-GP----GGYGVVME---L-V-G---
TA---FKK-----EFYEGFRH-TTNNRMELLAVIVGLEKL-------KNP-----
NMKVLVVSDSKYVVDSVEKKWVLGWEKKGFK--DRK------NSDLWKRLLIIYRK--H-
QV-DFKWIKGHNSHPQNERCDQLAVFAS-----NQKTLSVDAFYEKEEAKLL----------
----

>Bacteroidetes | Flavobacterium j | YP_001193487

----------MSHEVHIYTDGAAKG-NP-GN----GGYGVVME---L-V-G---
TP---YKK-----EFYEGFRL-TTNNRMELLAVIVGLEKL-------KNP-----
NMKVLVISDSKYVVDSVEKKWVFGWEKKGYT--------GKKNPDLWKRFLIAYRK--H-
KV-DFKWIKGHNNHPQNERCDQLAVMAS-----MQPKLSVDVYYETIGSKE----------
----

>Bacteroidetes | Cytophaga hutchi | YP_679299

----------MITLYTDGSSRG-NP-GP----GGFGVVLL---Y-K-Q--------
HRK-----EISGGFRM-TTNNRMELLAVITGLEAL-------KDP-----
GHDVLIYSDSKYVIDSVEKGWLMGWVKKNFK--DKK------NEDLWRRYLYVSSK--H-KI-
RFQWVRGHAGNIENERCDVLATQAA-----DGPNKQIDFGYETENGMLNKNHLS----------
-----

>Bacteroidetes | Algoriphagus sp. | ZP_01719037

----------MISIYTDGAAKG-NP-GP----GGYGAVLL---F-N-N---KGSI-
LRK-----ELSEGYRL-TTNNRMELLAVIRALQAL-------KVT-----
GIPVQIYSDSKYVVDAIEKGWLWGWQKKGFK--DKK------NPDLWLRYIPLHLK--Y-KP-
KFIWVKGHAGNPENERCDQLAVEAA-----EGRN-----LPADVGYEDSQK----------

>Bacteroidetes | Chitinophaga pin | YP_003123318

----------MSEVIIYTDGSSRG-NP-GP----GGYGVVLM---W-N-S--------
VRK-----ELSQGYRL-TTNNRMELMAVIVALEAL-------KRD-----
GLQVKIFTDSQYVVNSVEKGWLWGWVKTGFKDK--K------NKDLWQRFIPAFKK--H-
QV-KFNWVKGHSTNPLNNRCDELATQAA-----DSGN-----WLDDVGFEGE----------
-

>Thermoprotei | Metallosphaera s | YP_001192308

----------MKALGRFDGLCEPKNP-GG---IATFGYVIY---I-NGN--------
VIEGMGLASE-PWSVN-STNNVAEYTGLICLLKKM-------LTLG----
VTEARVEGDSQLVIRQLKGEYSVKSK----------RIIPLYEKAKELLAK--FSSV-
EIEWIPR--EENK--EADRITRIAFKKVLNGELK----------

>Methanomicrobia | Methanoculleus m | YP_001046256

----------TDAVTLYTDGASRG-NP-GD----AAWAYVI----VRDGS------
--VVA-----GRSGYIGT-ATNNVAEYHAVINGLDAA-------REFT----
GGRLEVRSDSELVVRQLTGRYRITKE----------HLAGLAEEVRRRMRH--FAEV-
RFESVPR--EHPCIQVADRLCNETLDAERRGRR----------

>environmental samples | uncultured archa | AAU83668

----------MKKLIIYTDGACRG-NP-GP----AGIGIVIC---NESGK--------
KIK-----EDKEFIGD-ATNNIAEYRALIKALELA-------SDFS----
VTRVECFSDSELMVRQLNGAYRVKDE----------KLGELFLQVKEKERL--FEEV-
TYSHVPR--KNNLIKRADSLANLGIDDKEPKETT----------

>Halobacteria | Haloquadratum wa | YP_657009

----------GGRAHVYFDGACRG-NP-GP----AAIGWVLV---TNEG--------
--IIA-----DGGEEIGK-TTNNRAEYAALERAIEMA-------RQYG----
FTEIDIRGDSQLIIRQVTGEYDTNEP----------TLREYRVRVRELLQT--FDRW-
SIEHVPR--DVNS--HADKLANEAFDHG----------