EVOLUTION AND ASSEMBLY OF THE RIBOSOME

BY

KE CHEN

DISSERTATION

Submitted in partial fulfillment of the requirements
for the degree of Doctor of Philosophy in Biophysics & Computational Biology
in the Graduate College of the
University of Illinois at Urbana-Champaign, 2013

Urbana, Illinois

Doctoral Committee:

Professor Martin Gruebele, Chair
Professor Zaida (Zan) Luthey-Schulten, Director of Research
Professor Taekjip Ha
Professor Klaus Schulten
Evolution and Assembly of the Ribosome
Chen, Ke. 2013.

Abstract
The ribosome is a large macromolecular assembly responsible for protein synthesis in all living cells. A typical bacterial ribosome consists of three ribosomal RNA (rRNA) molecules and approximately fifty ribosomal proteins (r-proteins), which are arranged into two subunits of unequal size and distinct function. The large subunit promotes formation of the peptide bond, and the small subunit enforces the recognition between the mRNA codons and the tRNA anticodons. With the availability of the ribosome crystal structure, it becomes clear that the two major functions, peptide bond formation and decoding, are performed within an entire RNA environment devoid of proteins. Combining with the fact that the majority of the ribosomal components are conserved across all three domains of life, it is believed that the ribosome has its origin deep in the RNA world before the last universal common ancestor (LUCA). Further evolutionary studies lead to the hypothesis that the evolution of the ribosome begins with a prototype ribozyme that catalyzes peptidyl transferase reaction.

Structural and sequence analysis suggests that the small ribozyme capable of catalyzing formation of short peptides may still exist in the core of the modern ribosome. Accordingly, a proto-ribosome model is constructed computationally using RNA fragments near the peptidyl-transferase center (PTC), and is proven to be stable throughout the micro-second molecular dynamics (MD) simulations. The model is capable of incorporating freely diffusing substrates spontaneously into its binding site, and holds them in both pockets long enough to reach a transition intermediate favorable for peptide bond formation. This in silico designed proto-ribosome is then subjected to experimental investigations to test its ability to assemble and bind potential substrates in solution. The successful design of the proto-ribosome presents a possible scenario for the initial development of the early translation apparatus. The proto-ribosome coupled with the probable parallel evolution of ancient tRNAs might have driven the emergence of the oldest coded protein shortly afterwards.

The universally conserved r-protein S4 is likely an ancient protein due to its role in the initiation of the 30S assembly, control of the translational accuracy, and regulation of the conserved α-operon. However, the N-terminal domain of S4 is identified as a “molecular signature” that distinguishes between Bacteria and Archaea, and hence might be a newer addition to the protein. The presence of both an old and a new component in the same protein makes it an extremely interesting case to study for the ribosomal evolution. Therefore, we perform phylogenetic analysis of S4 in relation to a broad sharing of zinc/non-zinc binding sequence in the N-terminal domain of the protein, and study the scope of horizontal gene transfer (HGT) of S4 during bacterial evolution. The complex history presented for “core” protein S4 suggests the existence of a gene pool before the emergence of bacterial lineages and reflects the pervasive nature of HGT in subsequent bacterial evolution. We then continue the study to understand the molecular driving force for such differentiated evolutionary history concerning one single protein. Consistency between experimental measurements and all-atom MD simulations indicates that the addition of the disordered N-terminal domain of S4 coevolved with a molecular signature in the rRNA helix h16, to couple the folding and binding process...
and accelerate the protein:RNA recognition.

The coupling between protein and RNA molecules in both evolution and modern dynamics inspired further study of the protein:RNA interactions during early assembly of the bacterial 30S small subunit. A practical protocol of all-atom MD simulation combined with RNA conformation clustering is developed to probe the folding landscape of the RNA molecules. In addition, the structure-based Gō-potential is developed within the framework of the all-atom molecular dynamics CHARMM force field, with which hundreds of simultaneous folding and binding events between the rRNA and r-protein are captured. Comparison between these simulations with the smFRET experiments reveals folding pathways constructed upon distinct structural intermediates of the RNA molecule. Our studies illustrate the complex nature of RNA folding in the presence of a protein binding partner, and provide insight into the role of population shift and the induced fit in the protein:RNA recognition process. The methodology developed therein will facilitate future studies of the ribosomal assembly on increasingly larger scale.
To my grandfather Duo Chen, who had encouraged my pursuit in science.
Acknowledgements

First and foremost, I am deeply indebted to my advisor Professor Zan Luthey-Schulten. Her patience and generosity in teaching a stranger both to the field of computational biology and to the country have greatly encouraged me and led me onto the right way. It is not only knowledge that I have learnt from her, but also the perspective in choosing the right scientific problem, the courage to rise to a cutting-edge project however risky it is, and the fortitude to reach high standard in each step of the research. While offering so much, she has also granted me the freedom in shaping my career the way I wish.

I want to thank all of the members of the Luthey-Schulten group, past and present, whom I have enjoyed working alongside. A few people who I worked with on various projects are Johnathan Lai, Tyler Harpole, John Cole, Zhaleh Ghaemi and Trang Thi Diem. Discussion with the other group members, Taras Pogorelov, Anurag Sethi, Jonathan Montoya, Dan Wright, Damien Matthews, Alexis Black Pyrkoz, Leonardo Sepulveda, Li Li, Andrew Magis, Piyush Labhsetwar, Tyler Earnest and Yuhang Tang, kept me in an intellectually stimulating and challenging environment. In particular, I want to thank Elijah Roberts and John Eargle, for guiding me hand in hand on my first few projects. They, as excellent computer scientist and biologist, have taught me how to approach a scientific problem logically and computationally. I also relied heavily on Mike Hallock, who has maintained the computational resources available through the U of I School of Chemical Sciences.

I would like to thank my dissertation committee members Professor Martin Gruebele, Taekjip Ha, and Klaus Schulten for providing me with their insight and expertise in protein folding, single molecule physics, and computational biology. They have exerted positive influences throughout my graduate career starting from the required course such as rotation and tutorial, and continuously shared their knowledge and opinions with me during collaboration. Working with their group members, Krishnanjun Sarkar (Gruebele Lab), Hajin Kim, Jingyi Fei (Ha lab), has also been extremely enlightening and fruitful. I would also like to thank Carl Woese, for teaching me the proper way to think about biology and evolution; and Nathan Price, for broadening my knowledge in systems biology.

Collaboration with scientists outside Illinois has always been a pleasant and exciting experience for me during my Ph.D studies. I want to thank Professor Sarah Woodson (John Hopkins University), as well as her students Magan Mayerle and Sanjaya Abeyesirigunawardena, for sharing their expertise in RNA chemistry. I would also like to thank Professor Ada Yonath (Weizmann Institute of Science, Israel) for inspiration, encouragement and support for my most exciting project.

Finally, I owe a special amount of thanks to my family for their understanding and support for me to continue on my scientific pursuit. My husband, Bo Wang, has been my spiritual anchor during my six
years in Illinois, not only in daily life, but also in my studies and scientific research. He gave me important suggestions when I was lost and almost wanted to give up. Without him, I could not have survived my difficult time to finish this dissertation.

Needless to say, numerous scientific funding agencies such as NSF, DOE and NIH provide support for the researches during the course of my Ph.D studies at various times. Specific grants are acknowledged in the appropriate publications.
Table of Contents

List of Tables ....................................................................................................................... x

List of Figures ......................................................................................................................... xi

List of Abbreviations .............................................................................................................. xiii

Chapter 1 Introduction ............................................................................................................. 1
  1.1 Ribosomal Components and Structure ............................................................................ 1
  1.2 Working Cycle of the Ribosome and the Essential Players in Translation ....................... 2
  1.3 Timeline of Ribosomal Evolution .................................................................................... 3
  1.4 Organization of the Dissertation ..................................................................................... 6

Chapter 2 In silico and In vitro Design of a Prebiotic Proto-Ribosome ........................................ 9
  2.1 Background ....................................................................................................................... 9
  2.2 Structure and Sequence Symmetry of the Ribosomal Peptidyl-transferase Center (PTC) .... 10
      2.2.1 Characterization of the structural conservation and symmetry ................................. 10
      2.2.2 Conserved sequence symmetry in the PTC .............................................................. 12
  2.3 Design of the Proto-Ribosome Model .............................................................................. 13
      2.3.1 Model construction based on tertiary interactions ................................................... 13
      2.3.2 Secondary structure stability of the best designed models ....................................... 15
      2.3.3 Experimental investigation of the proto-ribosome model ......................................... 16
  2.4 Function of the Proto-Ribosome ....................................................................................... 17
      2.4.1 Substrate binding propensity of the proto-ribosome ................................................ 17
      2.4.2 Correct binding of one freely diffusing substrate into the P-site ................................. 19
      2.4.3 Water-mediated dynamical pathway towards peptide bond formation ....................... 20
      2.4.4 Discussion ................................................................................................................. 22
  2.5 Methods .......................................................................................................................... 24
      2.5.1 Structural and sequence conservation ....................................................................... 24
      2.5.2 Model construction .................................................................................................... 24
      2.5.3 Molecular dynamics simulations ............................................................................... 25
      2.5.4 Analysis of molecular dynamics trajectories ............................................................ 26
      2.5.5 RNA preparation and native gel electrophoresis ...................................................... 26
### Chapter 3  Communal Sharing of Ribosomal Proteins Among the Early Bacteria

<table>
<thead>
<tr>
<th>Section</th>
<th>Page</th>
</tr>
</thead>
<tbody>
<tr>
<td>3.1 Background</td>
<td>28</td>
</tr>
<tr>
<td>3.2 Classification and Phylogeny of Ribosomal Protein S4</td>
<td>30</td>
</tr>
<tr>
<td>3.2.1 Sequence alignment and classification of bacterial r-protein S4</td>
<td>30</td>
</tr>
<tr>
<td>3.2.2 Phylogenetic reconstruction of S4’s evolutionary history</td>
<td>34</td>
</tr>
<tr>
<td>3.3 Horizontal Gene Transfer of S4</td>
<td>36</td>
</tr>
<tr>
<td>3.3.1 Identification of paralogous and duplicated S4 genes</td>
<td>36</td>
</tr>
<tr>
<td>3.3.2 Regulation of C- S4 paralogs</td>
<td>40</td>
</tr>
<tr>
<td>3.3.3 Comparison of genome content near S4</td>
<td>40</td>
</tr>
<tr>
<td>3.4 Interpretation of the Evolutionary Patterns of S4</td>
<td>42</td>
</tr>
<tr>
<td>3.4.1 The ancestral form of S4 in the bacteria</td>
<td>42</td>
</tr>
<tr>
<td>3.4.2 Origin of S4 outside the α-operon</td>
<td>43</td>
</tr>
<tr>
<td>3.4.3 Evolutionary pressure and the loss of zinc binding in ribosomal proteins</td>
<td>46</td>
</tr>
<tr>
<td>3.4.4 Conclusions</td>
<td>46</td>
</tr>
<tr>
<td>3.5 Methods</td>
<td>47</td>
</tr>
<tr>
<td>3.5.1 S4 sequence acquisition and alignment</td>
<td>47</td>
</tr>
<tr>
<td>3.5.2 Phylogenetic reconstructions</td>
<td>47</td>
</tr>
<tr>
<td>3.5.3 Zinc regulatory motifs</td>
<td>48</td>
</tr>
</tbody>
</table>

### Chapter 4  Functional Role of Ribosomal Signatures

<table>
<thead>
<tr>
<th>Section</th>
<th>Page</th>
</tr>
</thead>
<tbody>
<tr>
<td>4.1 Background</td>
<td>49</td>
</tr>
<tr>
<td>4.2 Characterization of the Coevolving Signature Interactions in the Crystal Structure</td>
<td>51</td>
</tr>
<tr>
<td>4.3 Intrinsically Disordered Protein Signature: S4 N-terminal Domain</td>
<td>53</td>
</tr>
<tr>
<td>4.3.1 S4N flexibility in simulation</td>
<td>53</td>
</tr>
<tr>
<td>4.3.2 Temperature unfolding studies of S4N</td>
<td>55</td>
</tr>
<tr>
<td>4.4 Dynamics of the RNA Signature h16</td>
<td>55</td>
</tr>
<tr>
<td>4.4.1 Stability and secondary structure fluctuation of h16</td>
<td>55</td>
</tr>
<tr>
<td>4.4.2 Temperature melt experiment of h16</td>
<td>58</td>
</tr>
<tr>
<td>4.5 Dynamical Protein:RNA Signature Interactions and Their Function</td>
<td>59</td>
</tr>
<tr>
<td>4.5.1 Experimental measurement of protein:RNA interactions</td>
<td>59</td>
</tr>
<tr>
<td>4.5.2 Side-chain interactions between S4N and h16</td>
<td>61</td>
</tr>
<tr>
<td>4.5.3 Salt bridge interactions between RNA backbone and protein side chain</td>
<td>61</td>
</tr>
<tr>
<td>4.5.4 Role of S4N signature in the refolding of the five-way junction</td>
<td>63</td>
</tr>
<tr>
<td>4.5.5 Conclusions</td>
<td>64</td>
</tr>
<tr>
<td>4.6 Methods</td>
<td>65</td>
</tr>
<tr>
<td>4.6.1 Protein and RNA design and sample preparation</td>
<td>65</td>
</tr>
<tr>
<td>4.6.2 Experimental measurements</td>
<td>65</td>
</tr>
<tr>
<td>4.6.3 Molecular modelling and simulation</td>
<td>66</td>
</tr>
<tr>
<td>4.6.4 Base pairing and base stacking interaction determination</td>
<td>67</td>
</tr>
<tr>
<td>4.6.5 Salt bridge detection</td>
<td>67</td>
</tr>
</tbody>
</table>
Chapter 5  Protein-assisted Assembly Landscape of the Five-way Junction in the Ribosomal
Small Subunit ............................................................. 68
  5.1 Background .......................................................... 68
  5.2 Equilibrium Dynamics of the Five-way Junction ....................... 70
    5.2.1 Nucleotide flexibility in the five-way junction ...................... 70
    5.2.2 SHAPE mechanism revealed by comparison between experiment and simulation  71
    5.2.3 Consistency of standard MD and Gō potential simulations ............... 72
  5.3 Multiple Conformations of the Five-way Junction Near Native State .......... 72
    5.3.1 Conformational heterogeneity of the five-way junction observed with smFRET ... 72
    5.3.2 Identifying the metastable states in simulations ...................... 75
    5.3.3 Correlating atomic structures of the five-way junction conformations with the FRET states ...... 76
  5.4 Folding Landscape of the Five-way Junction .......................... 76
    5.4.1 Simultaneous folding and binding of the five-way junction and S4 .......... 76
    5.4.2 Binding pathway of S4 ............................................ 77
    5.4.3 Binding mechanism of the signature S4 N-terminus and the five-way junction ...... 79
    5.4.4 Inferring the folding landscape of the five-way junction ................. 80
    5.4.5 Concluding remarks ............................................... 83
  5.5 Future Studies ....................................................... 84
    5.5.1 S4 modulated 5′ domain rRNA dynamics ............................ 84
    5.5.2 Atomic resolution characterization of various components in the kinetic model of the S4:rRNA binding pathway .............. 85
  5.6 Methods ............................................................. 87
    5.6.1 RNA preparation .................................................. 87
    5.6.2 FRET measurement ............................................... 88
    5.6.3 Molecular dynamics simulations .................................. 88
    5.6.4 The structure-based Gō model .................................... 89
    5.6.5 Hybrid MD-Gō folding simulations ............................... 90
    5.6.6 Analysis of the simulated trajectories ............................. 90

Bibliography .............................................................. 92

Appendix A Secondary Structure Diagrams of the Escherichia coli 16S rRNA ............ 107

Appendix B Additional Details for Implementation of Hybrid MD-Gō Potential in NAMD .... 111

Appendix C Supporting Information ....................................... 116
**List of Tables**

2.1 Sequence symmetry of the PTC based on STAMP structural alignment .......................... 12
2.2 Summary of all proto-ribosome models ................................................................. 25

3.1 Genomes containing multiple copies of the S4 gene .............................................. 37
3.2 Taxonomic distributions of the C+/− ribosomal proteins ......................................... 45

4.1 Salt bridges within ribosomal protein S4 N-terminus during simulations ................... 56
4.2 Salt bridges between ribosomal protein S4 N-terminus and backbone phosphate of h16 ... 62

B.1 Improper dihedrals used to enforce chirality in the hybrid MD-Gō model ................. 111
## List of Figures

1.1 Vertical evolution of species after LUCA .................................................. 3  
1.2 Speculative timeline for translational evolution before LUCA ......................... 5  

2.1 Symmetry conservation of PTC ................................................................. 11  
2.2 Stability of the designed proto-ribosome constructs .................................... 14  
2.3 Comparison of the secondary structure stability of the two best proto-ribosome constructs ................................................................. 16  
2.4 Formation of the proto-ribosome model in solution .................................... 17  
2.5 Substrate binding propensity of the proto-ribosome .................................... 18  
2.6 Successful binding of the substrate into the P-site ..................................... 20  
2.7 Peptidyl-transferase reaction catalyzed by the designed proto-ribosome ............ 21  

3.1 Multiple sequence alignment of ribosomal protein S4 .................................. 31  
3.2 Consensus phylogenetic tree of S4 in *Proteobacteria* .................................. 33  
3.3 Consensus unrooted phylogenetic tree of S4 ............................................. 35  
3.4 Expansion of *Clostridia* branches of the consensus S4 phylogenetic tree ........ 38  
3.5 Expansion of *Actinobacteria* branches of the consensus S4 phylogenetic tree .... 39  
3.6 Genomic content near S4 and the $\alpha$-operon .......................................... 41  
3.7 Occurrence of different types of S4 throughout the bacterial phylogeny ............... 43  
3.8 Possible evolutionary pathways of S4 ..................................................... 44  

4.1 Visualization of the five-way junction and S4 on the ribosome ....................... 50  
4.2 Contacts between h16 and S4N ............................................................... 52  
4.3 Backbone and secondary structure fluctuations of S4N ................................ 54  
4.4 Backbone and secondary structure fluctuations of h16 ................................ 57  
4.5 Base pairing and base stacking interactions during simulations ...................... 59  
4.6 Binding of S4N and h16 .............................................................. 60
4.7 Unfolding and refolding of the five-way junction ....................................... 64  

5.1 Comparison of simulated and experimental SHAPE data ................................ 71  
5.2 Correlations between MD and hybrid MD-Gō simulations of native state fluctuations in the five-way junction:S4 complex ............................................. 73  
5.3 Multiple states are observed in sm-FRET experiment and MD simulations of the five-way junction ................................................................. 74  
5.4 S4 binds to the five-way junction in multiple steps ..................................... 77
List of Abbreviations

CD – circular dichroism
DGL – differential gene loss
EC – encounter complex
EF-G – elongation factor G
EF-Tu – elongation factor Tu
FI – flipped intermediate
FRET – Förster resonance energy transfer
HGT – horizontal gene transfer
IF – initiation factor
indel – insertion or deletion
ML – maximum-likelihood
mRNA – messenger RNA
MSA – multiple sequence alignment
NC – native complex
LSU – large subunit
LUCA – last universal common ancestor
nm – nanometers
ns – nanoseconds
RA motif – right angle motif
RF – release factor
RMSD – root mean square displacement
RNAP – RNA polymerase
rRNA – ribosomal RNA
r-protein – ribosomal protein
RRF – ribosomal recycling factor
SHAPE – selective 2′-hydroxyl acylation analyzed by primer extension
smFRET – single-molecule FRET
SSU – small subunit
SI – supporting information
TOC – table of contents
UPT – universal phylogenetic tree
| A. metalliredigens – Alkaliphilus metalliredigens |
| A. oremlandii – Alkaliphilus oremlandii |
| B. subtilis – Bacillus subtilis |
| B. stearothermophilus – Bacillus stearothermophilus |
| C. acetobutylicum – Clostridium acetobutylicum |
| C. novyi – Clostridium novyi |
| D. radiodurans – Deinococcus radiodurans |
| E. coli – Escherichia coli |
| F. magna – Finegoldia magna |
| H. marismortui – Haloarcula marismortui |
| L. borgpetersenii – Leptospira borgpetersenii |
| M. flagellatus – Methylobacillus flagellatus |
| M. tuberculosis – Mycobacterium tuberculosis |
| P. ingrahamii – Psychromonas ingrahamii |
| S. arenicola – Salinispora arenicola |
| S. coelicolor – Streptomyces coelicolor |
| S. tropica – Salinispora tropica |
| T. aquaticus – Thermus aquaticus |
| T. brucei – Trypanosoma brucei |
| T. thermophilus – Thermus thermophilus |
| T. thermophila – Tetrahymena thermophila |
Chapter 1

Introduction

1.1 Ribosomal Components and Structure

The ribosome is a large macromolecular assembly responsible for protein synthesis in all living cells. A ribosome consists of two subunits of unequal size and distinct function. In a typical bacterial 70S ribosome, the large subunit (LSU), called 50S, contains about 35 proteins (designated L1, L2, etc, according to their sedimentation coefficients) and two ribosomal RNA chains (23S and 5S). The small subunit (SSU), with a molecule weight half of that of the LSU, contains one RNA chain (16S), and approximately 21 proteins (S1, S2, etc.). Each subunit has three binding sites for the tRNA, designated the A (aminoacyl), P (peptidyl), and E (exit) sites. Enforcement of recognition between the mRNA codons and the tRNA anticodons at the A-site on the SSU promotes formation of the peptide bond between amino acids attached to the 3′ CCA ends of tRNAs docked into adjacent A- and P- cites located within the LSU. The cooperative effort of both subunits ensures the precise translation from the genetic code carried by the mRNA into the functional protein repertoire of a cell.

Our knowledge of the ribosome gradually accumulates until the advent of crystal structures of the ribosome in 2000 [1–4]. Seeing is believing. We have since then experienced a dramatic expansion in understanding of the ribosomal structure and function. For example, it is long known that the ribosomal RNAs (rRNAs) are capable of self-folding into the shape of each ribosomal subunit, and scaffolding for property spatial arrangement of the ribosomal proteins (r-proteins). However, the notion that rRNA may also participate in the catalytic function is only slowly being accepted 40 years after the discovery of ribosome. Once the structure of the ribosome is resolved, it immediately becomes clear that the ribosome is a “ribozyme”, because all functional cites, including mRNA-tRNA recognition and transpeptidation, are formed solely by the rRNA [5; 6]. Rapid advances in crystallization techniques, as well as the successful combination of computational techniques with cryo-electron microscopy [7; 8], provide us with three-dimensional structures of the ribosomes from ~10 different species covering all three domains life. The eukaryotic ribosomes are shown to vary in size due to different number of eukaryote-specific r-proteins and additional rRNA insertions, called expansion segments [9–12]. Nevertheless, they share with the bacterial and archaeal ribosome approximately 30 universal r-proteins, and the core rRNA structure is highly conserved compared to their bacterial and archaeal counterpart. Now, with the prosperous repository of high-resolution three-dimensional crystal structures of the ribosome capturing various functional states upon association with
different translational co-factors and regulators [13–25], or bound with numerous antibiotics [26–33], we
are facing the challenge of reconstructing the evolutionary history of the ribosome and translation, as well
as comprehending the dynamics of the ribosome during functioning.

1.2 Working Cycle of the Ribosome and the Essential Players in Translation

As early as in the 1950s, biophysicists working on the ribosome noticed that protein synthesis did not require
the integrity of the cell and can be performed after cell disruption. This understanding laid the basis for the
creation of the cell-free translation systems, which enabled the researchers to map out a general picture for
the molecular mechanisms of translation long before the crystal structure of the ribosome became available.
As we now know, translation consists of three basic working cycles [34], namely the initiation, elongation
and termination. Each cycle is facilitated by a group of special protein enzymes — called the initiation
factors, elongation factors, and release factors, correspondingly — which will be introduced briefly in the
following paragraphs based on the bacterial translation system.

Three initiation factors, IF1, IF2, and IF3, are typically present in bacterial cells [35]. The initiation
phase starts with the binding of IF3 to the 30S to promote dissociation of the ribosome into subunits. Ini-
tiation factor IF1 stimulates IF3 activities and in addition, binds specifically to block the A-site of the 30S
ribosomal subunit so that the initiator methionyl-tRNA (fMet-tRNA$_{Met}$) could be directed to the P-site.
Following subunit dissociation, IF2, mRNA, and fMet-tRNA$_{Met}$ associate with the 30S ribosomal subunit
in an order still open to discussion. Next, the initiation factors together assist the exact placement of the
fMet-tRNA$_{Met}$ against the initiation codon on the mRNA by monitoring the codon-anticondon interaction
at the P-site on the 30S ribosomal subunit. Once such stable initiation complex is formed, IF1 and IF3 are
ejected from the complex. Finally, IF2 leaves the 70S ribosome after stimulating the correct association of
the large subunit upon GTP hydrolysis.

The initiation phase produces ribosomes with an occupied P-site and a vacant A-site, as well as the
mRNA to be translated. At this point, the ribosome is ready to begin its elongation cycle that involves three
principal steps: codon-dependent accommodation of aminoacyl-tRNA, transpeptidation, and translocation.
Accommodation of the aminoacyl-tRNA is facilitated by the elongation factor EF-Tu, which is a member of
the GTPase family. EF-Tu in its GTP-bound conformation, carries the aminoacyl-tRNA to the ribosome and
binds at the T-site on the large subunit. Upon correct recognition between the codon and anti-codon at the
A-site on the small subunit, EF-Tu becomes activated and undergoes a conformational change triggered by
hydrolysis of GTP. The GDP-bound conformation of EF-Tu has a reduced affinity for tRNA, and, therefore,
is able to release the tRNA and leave the ribosome. Once the aminoacyl-tRNA is fully accommodated into
the A-site, transpeptidation is catalyzed by the peptidyl-transferase center (PTC) formed by the 23S rRNA,
resulting in the peptidyl-tRNA (with the peptide chain elongated by one amino acid) occupying the A-site
and the deacylated-tRNA occupying the P-site. At the last stage, elongation factor G (EF-G) comes to bind
the ribosomal A-site, forcing the peptidyl-tRNA to assume an A/P hybrid state in which it occupies the A-
site of the SSU but P-site of the LSU. Similarly, the deacylated tRNA is shifted to a P/E hybrid state. In such
a pre-translocation state, the GTPase activity of EF-G is induced. At the expense of GTP hydrolysis, the
30S ribosomal subunit undergoes a ratcheting motion that moves the mRNA forward by three nucleotides
and the peptidyl-tRNA into the classical P-site position. Meanwhile, the deacylated tRNA is turned into the E-site and ready to leave the ribosome when the next aminoacyl-tRNA is accommodated. Now, the ribosome is once again ready to read the next codon.

The elongation cycle is repeated for every codon on the mRNA template until one of the three stop codons (UAA, UAG, or UGA) is met. The positioning of the stop codon in the ribosomal A-site signals the beginning of the termination cycle, which works in a mechanistic similar way as the elongation phase [36]. Analogous to the aminoacyl-tRNA, class I release factors (RFs) bind to the ribosome and catalyze the hydrolysis of the peptidyl-tRNA to release the peptide chain upon their recognition of the stop codon. Particularly in bacteria, RF1 recognizes a UAG or UAA stop codon, while RF2 mediates termination in response to a UGA or UAA codon. In addition, class II release factor RF3 promotes termination by binding to the RF1 or RF2 in a GTP-dependent manner, and facilitates the dissociation of class I release factors by stabilizing the hybrid state of the ribosome [24]. Finally, the ribosomal recycling factor (RRF), together with the previously introduced EF-G and IF3, cooperatively trigger the dissociation of ribosomal subunits and return the ribosome into a state capable of initiation.

The scheme sketched above outlines only the most fundamental steps during translation. Although many intermediate states have been omitted in the description, we have listed the most essential players in the process. In summary, translation could be reconstituted in a cell-free environment with purified components including the ribosomes (rRNA and r-proteins), template mRNAs, a set of aminoacyl-tRNAs (or a system of tRNA aminoacylation), and protein co-factors introduced in this section. With additional energy source provided by GTP, translation should occur spontaneously. Most importantly, the majority of these essential components are universally conserved among all three domains of life, which leads to the hypothesis that evolution of the translation is largely completed before the last universal common ancestor (LUCA).

### 1.3 Timeline of Ribosomal Evolution

![Vertical evolution of species after LUCA](image)

**Figure 1.1: Vertical evolution of species after LUCA.** (A) The universal phylogenetic tree based on 16S rRNA. (B) Location of sequence signatures (red) and structural signatures (blue) in the aligned 16S rRNA from two bacteria and one archaea. (*Figure courtesy of J Eargle & E Roberts*)

Life as we see today evolves through a generally vertical path [37; 38] that is usually depicted by the universal phylogenetic tree (UPT). Conserved evolutionary history seen in all housekeeping translational genes, including rRNAs, tRNAs, universal r-proteins, and the elongation and initiation factors, strongly supports the canonical pattern among the three domains of life: **Bacteria, Archaea, and Eucarya** (Figure 1.1 A). Particularly, a thorough comparative study of the 16S and 23S rRNA genes [39] identifies sequence and structure signatures which are clearly characteristic for and distributed through each individual domain of life.
life (Figure 1.1 B). These signatures preserve the idiosyncratic modifications in each of the three major cell types, and in the same time indicate an evolutionary phase transition at the root of the UPT.

Carl Woese proposed that such transition occurs as the loosely organized aboriginal cells cross a “Darwinian threshold”, upon communal interactions driven by massive horizontal gene transfers (HGT) [38]. It is the community as a whole that evolves, against environmental bottleneck, into what we now know as the last universal common ancestor, or LUCA [40]. Comparative genomics and deep evolutionary reconstructions from Koonin’s lab suggest that ancestors of hundreds of extant genes were already present in LUCA, among which 90% of the protein genes and 70% of the RNA genes are translation system components [41; 42]. These studies inspired further comparative analysis on the structure of the ribosome, leading to the “onion model” for the 23S rRNA evolution before LUCA [43–46]. The model theorizes that the peptidyl-transferase center emerges as the functional core of the ribosome in the RNA world, upon which other RNA structural elements could be built layer by layer to gain complexity and functionality during ribosomal development [45]. Interestingly, the structural and sequence signatures on the 23S rRNA, which are unique to only one branch of the tree of life, appears predominantly on the outside layers (newer additions) according to this onion model. Such consistency in evolutionary inference demonstrates how the rich information hidden in the primary sequences, three-dimensional folding and functional interactions of the essential translational genes [47; 48] would help the understanding of the pre-LUCA evolution of translation and ribosome. A few more examples will be given in the following paragraphs.

Transfer RNAs, for example, are thought to be made of two separate domains. One domain is the acceptor stem with the terminal CCA sequence to which the incoming amino acid is attached; the second contains the anticodon stem that interacts with the mRNA. It is argued that the CCA domain is older, because the CCA domain itself may form a minihelix structure that can be aminoacylated by the modern tRNA synthetase or evolved ribozymes [49], and incorporated into the ribosome for peptide elongation [50]. On the other hand, the anticodon domain is not required for recognition by many class II tRNA synthetase, indicating it being less ancient.

Another example takes advantage of the idea of molecular mimicry, from which we deduce timing information by assuming the newer gene as an imitation of the original copy. As we know, the shape of EF-G closely mimics that of the ternary complex of EF-Tu-GTP-tRNA. Hence the two elongation factors, EF-Tu and EF-G, compete for an overlapping binding site on the ribosome to facilitate tRNA accommodation and translocation. In this case, the EF-G is likely newer, because i) the tRNA, as a central component of translation system, is supposed to be older; ii) EF-G is shown to be dispensable for translocation for certain systems [51; 52]. The mimicry principle can also be applied to the ribosomal proteins. The r-protein S1 contains six copies of the RNA binding domain (OB fold), which together make what is known as the S1 domain. This motif is found in a variety of proteins, including initiation factor IF1/εIF1α, α subunit of eIF2, ribonuclease E, polynucleotide phosphorylase, etc. Among all known proteins encompassing a S1 motif, the initiation factor IF1 is most likely the original copy because it is the only one that is universally conserved across all domains of life. S1, on the other hand, is not an integral part of a working ribosome and therefore appears to be a latter addition during evolution.

The assembly maps for both the small and large subunits [53; 54] describe the dependencies of the
r-proteins upon binding to the rRNA during in vitro ribosomal assembly. These maps provide not only understanding of the modern ribosomal biogenesis [55], but also rich information of the ribosomal evolution. One would hypothesize that the more ancient r-proteins would i) bind directly to the rRNA and be incorporated into the ribosomal subunits earlier during assembly; ii) be more interconnected on the assembly map in terms of protein:protein interactions and binding dependencies; and iii) be more prevalently distributed across species. In fact, a strong correlation is observed between the distribution of r-proteins and their assembly order in the large subunit [47]. Furthermore, older r-proteins tend to form conserved clusters on the bacterial genomes called operons, and the oldest of them likely regulates expression of all proteins located within the same operon. As a result, it is suggested that L2, L3, and L4 are among the oldest large subunit r-proteins, when assembly is taking into account together with the other criteria. Interestingly, these most ancient proteins are in close proximity to the peptidyl-transferase center in the 23S rRNA, which is thought to be the oldest portion of the rRNA according to several studies using different evolutionary analysis [43–46].

A surprising number of individual series of timing events can be inferred similarly using the sequence and structure information, and conserved reactions, pathways and regulations. Although it is much more difficult to relate one sequence of events to another, Fox and Naik [47; 48] made the attempt to organize them in a general way and suggested a tentative timeline for the pre-LUCA development of the ribosome (Figure 1.2). The order of events can be greatly complicated by the parallel nature of prebiotic maturation of the translational machinery. Therefore, the speculation presented in Figure 1.2 should not be viewed as facts, but rather inspiration for future discussion of the ribosomal/translational evolution and the origin of
1.4 Organization of the Dissertation

An evolutionary perspective provides us with a natural way to section the ribosome into functional parts and study the way it works. In this dissertation, we employ dynamical characterizations in addition to information obtained from the static signatures, sequences and structures to explore possible pathways and intermediates for ribosomal evolution and assembly. To start, Chapter 2 describes the rational design of a prebiotic ribozyme that could potentially make statistical peptide chains based on our understanding of the ribosome structure as well as dynamical properties of the RNA molecules. This study lends evidence to the existence of such ribozyme, which might have triggered the transition from the RNA world to the modern cell, and stayed in the core of the translational machinery since then [56].

Chapter 2 – The fundamental role of RNA in the universal processes of transcription and translation lead to the hypothesis of an era where RNAs could serve as both genetic material and metabolic enzymes. In accordance to the RNA world hypothesis and evolutionary analysis of the ribosome structures and sequences, it is suggested that a small ribozyme capable of catalyzing formation of small peptides still exist in the core of the modern ribosome. Accordingly, a model of such proto-ribosome is constructed computationally using RNA fragments near the peptidyl-transferase center. With only 4% of the nucleotides in a modern bacterial ribosome, our model is proven to be stable throughout micro-second molecular dynamics simulations, and able to incorporate freely diffusing substrates spontaneously into the peptidyl-tRNA binding site. Moreover, the model binds the charged tRNA mimics in both the aminoacyl- and peptidyl-tRNA binding sites and reorients the substrates into a transition intermediate favorable for peptide bond formation. The \textit{in silico} designed proto-ribosome is then put under experimental investigations with native polyacrylamide gel electrophoresis (native PAGE), which indicate that the symmetric fragments assembly, and bind the potential substrates. This study reveals a possible pathway for the ancient RNA world to acquire proteins that could assist the buildup of a more complex translation apparatus.

The next two chapters deal with a universally conserved small subunit r-protein S4. Unlike the large subunit, the 30S assembly map does not tell an obvious story of its history. However, the r-protein S4 is very likely an ancient protein for the following reasons: i) it is distributed across all domains of life; ii) it is both a primary binding and an initiator protein of the 30S assembly; iii) it regulates the gene expression for the conserved \(\alpha\)-operon. On the other hand, the N-terminal domain of S4 appears to be a “molecular signature” that distinguishes between \textit{Bacteria} and \textit{Archaea} [39], and hence might be a newer addition to the protein. The presence of both an old and a new component in the same protein makes it an extremely interesting case to study for the ribosomal evolution. In Chapter 3, we examine the evolutionary history of S4 through phylogenetic analysis; and in Chapter 4, we investigate the dynamical function of the protein, especially its N-terminal signature, coupled with its coevolving rRNA signature.
Chapter 3 – The universal ribosomal protein S4 is essential for the initiation of small subunit ribosomal assembly and translational accuracy. Being part of the information processing machinery of the cell, the gene for S4 is generally thought of as being inherited vertically and has been used in concatenated gene phylogenies. Here we report the evolution of ribosomal protein S4 in relation to a broad sharing of zinc/non-zinc binding sequence in the N-terminal domain of the protein, and study the scope of horizontal gene transfer (HGT) of S4 during bacterial evolution. The complex history presented for “core” protein S4 suggests the existence of a gene pool before the emergence of bacterial lineages and reflects the pervasive nature of HGT in subsequent bacterial evolution. This has implications for both theoretical models of evolution and practical applications of phylogenetic reconstruction as well as the control of zinc economy in bacterial cells.

Chapter 4 – Inspired by the discovery that the N- and C-terminal domains of the r-protein S4 experienced different evolutionary histories, we continued the study hoping to understand the functional driving force involved in its evolution. In this chapter, we investigate, both experimentally and computationally through all-atom molecular dynamics simulations, the dynamics of the coevolving signature complex including the N-terminal fragment of the ribosomal protein S4, and the helix 16 (h16) which is part of the five-way junction in 16S rRNA. Our results show that the S4 N-terminus signature is intrinsically disordered in solution while h16 is relatively stable by itself. The dynamic disordered property of the protein fragment is exploited to couple the folding and binding process and accelerate the protein:RNA recognition through a “fly-casting” mechanism. This study provides insight into the mechanism for the early and fast binding of S4 in the assembly of the ribosomal small subunit, as well as the interplay between RNA and protein molecules during ribosomal evolution.

In the final chapter, we continue to explore the modern aspect of the assembly process, concerning the protein-guided dynamics of rRNA folding. A practical simulation protocol is developed to probe the folding landscape of the RNA molecules, which will facilitate future studies of the ribosomal assembly on increasingly larger scale.

Chapter 5 – The assembly of the bacterial ribosomal small subunit (SSU) begins with the folding of the five-way junction upon interaction with the primary binding protein S4. In this chapter, we expand the conformational space observable in MD simulations by varying the ion concentrations at different stages in the folding of RNA molecules. This simulation protocol, together with appropriate clustering techniques, leads to the identification of multiple metastable conformations of the naked five-way junction without the presence of S4. Furthermore, to reach the timescale of biologically relevant in vitro RNA:protein recognition, we develop the structure-based Gō-potential implemented within the framework of the all-atom molecular dynamics CHARMM force field, with which hundreds of simultaneous folding and binding events between the five-way junction and S4 are captured. Comparison between both all-atom and hybrid MD-Gō simulations with the smFRET experiments reveals folding pathways constructed
upon distinct structural intermediates of the RNA molecule. Our studies illustrate the complex
nature of RNA folding in the presence of a protein binding partner and provide insight into the
role of population shift and the induced fit mechanisms in the protein:RNA folding and binding
process.
Chapter 2

*In silico* and *In vitro* Design of a Prebiotic Proto-Ribosome

2.1 Background

The RNA molecules participate in fundamental and conserved processes in the modern cell, including carrying genetic information from DNA to the translational system (mRNAs), regulating transcription and translation (riboswitches, small RNAs, etc.), and catalyzing a broad range of chemical reactions (RNase P, self-splicing introns, signal recognition particles, ribosomes, etc.). The critical and versatile role of the RNA molecules leads to the RNA world hypothesis that an RNA-base era precedes the contemporary cellular life, where RNA provides both the genetic material and enzymatic functions [57]. Although the extent of the RNA world is still subject to dispute, it is generally agreed that translation is essentially completed before the last universal common ancestor (LUCA) — a compendium of universally conserved genes among which ~80% come from the translation system [41]. Success of *in vitro* selection of small ribozymes that are able to catalyze principal chemical reactions of protein synthesis now performed by various translational components, such as RNA aminoacylation by the tRNA synthetase [49; 58], and peptidyl/amine-acid transfer reactions by the ribosome [59; 60], further indicates that the complex translation machinery has evolved from an RNA world.

Particularly, the ribosome is believed to have an origin in the RNA world once its crystal structure has been solved [48]. The two major functions of the ribosome, peptide bond formation and decoding, are each performed within an RNA environment (peptidyl-transferase center in the large subunit, and decoding center in the small subunit) devoid of proteins and stabilized by magnesium ions [3; 5; 61; 62]. This fact relieves the demand for proteins to participate in the peptide making process, and proposes an answer to the chicken-or-egg problem that the functional units of the ribosome should have been created in the form of ribozymes before the availability of proteins. Yet, a prototype ribozyme that catalyzes peptidyl transferase reaction, and therefore begins the evolution of the ribosome and triggers the transition of the RNA world into the modern cellular life has not been fully settled.

As envisioned in [56], such primitive translation apparatus should be constructed from RNA:RNA com-

---

The contents of this chapter are based in part on work in preparation as Ke Chen, Bo Wang, Jingyi Fei, Ada Yonath, and Zaida Luthey-Schulten. *In silico and In vitro Design of a Prebiotic Proto-Ribosome.*
ponents that are still present in the core of the modern translational machinery. Comparative sequence and structural analysis of the ribosome suggests that the peptidyl-transfer center (PTC) is the oldest component in the ribosome [43–46] and is potentially selected over the other peptidyl-transferase ribozymes present in the RNA world to be fixed at the heart of the contemporary ribosome. More detailed examination on the PTC reveals two structurally symmetric RNA fragments located at the A- and P- sites of the ribosomal large subunit (LSU), each consists of ∼ 60 nucleotides and adopts a tRNA-like L-shape [28; 63]. It is hypothesized that the two RNA molecules may each fold into its native conformation and further dimerize under prebiotic conditions. The resulting complex, which hosts a binding pocket for the early tRNA acceptor stem mimics, is able to make statistical peptide bonds in a non-coded fashion [64; 65].

Based on increasing understanding of the RNA structure-dynamics relations, it is known that small RNAs have a tendency to retain their secondary structures when excised from larger structures [66]. Moreover, RNA molecules are able to associate with each other noncovalently to form complexes through base pairing, base stacking and tertiary interactions such as the A-minor motifs; those complexes can then be ligated together to form covalently stable RNAs of increasing size. Therefore, it is possible to build larger functional RNA out of pre-existing small ones while preserving structural and functional properties of the latter. Taking advantage of these designing principles, we presented a proto-ribosome model constructed from fragments of the ribosomal PTC and pieced together by frequently occurring tetra-loops and A-minor interactions. The model was shown to be stable over micro-second molecular dynamics simulation, and able to bind an adenosine-alanine ligand into its peptidyl-tRNA binding site using the same interactions as seen in the ribosome. Most importantly, the proto-ribosome demonstrated a rigorous pathway mediated by water molecules that would orient the two stably bound charged tRNA mimics into a transition state favorable for peptide bond formation. The results provided evidence for the RNA world origin of the ribosome, and illustrated the possible evolutionary pathway towards a more complex translation machinery.

2.2 Structure and Sequence Symmetry of the Ribosomal Peptidyl-transferase Center (PTC)

2.2.1 Characterization of the structural conservation and symmetry

Previous studies in Yonath’s lab identified a conserved region (SymR) deep in the PTC of the ribosome, of which the two sub-regions (A- and P-regions) are related with a two-fold symmetry such that they overlap each other after a 180° rotation. It is hypothesized that this symmetrical arrangement of the PTC ensures the two charged tRNA substrates to be positioned with a favorable sterochemistry that enables peptide bond formation as well as elongation. Therefore, a detailed understanding of the symmetric properties in the PTC would render a good approximation for the design of a proto-ribosome.

The SymR region contains approximately 180 nucleotides in total, encompassing helices H74, H75, H80, H89 (P-region), H90, H91, H92, H93 (A-region) radiating from the central loop in domain V of the 23S rRNA. Alignment of the SymR region from six crystal structures (Figure 2.1A) shows structure conservation over all three domains of life, with the largest backbone RMSD being ∼ 1.6 Å between those of the Deinococcus radiodurans and Saccharomyces cerevisiae structures. Most nucleotides share the same orientation in all six crystal structures (Q_{res} > 0.8, Figure 2.1B), except a few (e.g. A2602 and U2585, E. coli
numbering) that probably represent different functional states of the ribosome. The only insertions/deletions are found flanking the gap in H80 (P-loop), which is filled by ~170 nucleotides in the modern ribosome. The symmetric eukaryotic insertion on the 5′ and deletion on the 3′ of the gap maintain the length of H80 with respect to the bacterial counterpart, indicating that the additional nucleotides might be a later invention for more complex functions.

Figure 2.1: Symmetry conservation of PTC. (A) Structural alignment of the PTC from crystal structures of six organisms colored by $Q_{res}$. (B) $Q_{res}$ for all nucleotides in the PTC. (C) Structural (colored by $Q_{res}$, left panels) and sequence symmetry (right panels) between the A- and P- cores (top panels) or A- and P- loops (bottom panels) of the PTC.

The rotational and secondary structure symmetry of the A- and P-regions were mapped out rigorously in [63]. However, a close examination of the SymR crystal structure reveals a lot more base-base interactions than predicted in the conventional secondary structure diagram, many of which are non-canonical, long-range base-pairs or base-triplets (Figure 2.2A&B). These additional interactions are especially important in stabilizing the central loop junction in both regions. Furthermore, the secondary structure can be highly dynamic due to the constantly occurring registered shift of base-pair pattern in the RNA stem [67]. Taking into account these complications, we reassign symmetry between the two regions by the overall backbone conformation using STAMP structural alignment [68]. In this way, we allow more flexibility in the inference of the symmetry breaking insertion/deletion.

The results of the structural alignment is shown in Figure 2.1C, colored by structural and sequence similarity, respectively. The short arm (H93/H89) and central loop enjoy the most structure symmetry ($Q_{res} > 0.5$). On the other hand, the less degree of structural symmetry in the long arm of the A- and P-cores (H90 and H74) is compensated by compatibility in sequence (identical or complementary nucleotides). The least structural and sequence agreement between the A- and P-loops is likely an indication that these two components were later additions to the system in response to the need of accommodating larger substrates. It
Table 2.1: Sequence symmetry of the PTC based on STAMP structural alignment

<table>
<thead>
<tr>
<th></th>
<th>Total (A/G/U/C)</th>
<th>Identical (A/G/U/C)</th>
<th>Complementary (GC/AU)</th>
<th>Transition (AG/UC)</th>
<th>Transversion (AC/GU)</th>
<th>Insertion (A) (A/G/U/C)</th>
<th>Insertion (P) (A/G/U/C)</th>
</tr>
</thead>
<tbody>
<tr>
<td>All Positions</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Core</td>
<td>67</td>
<td>18 (2/7/5/4)</td>
<td>18 (13/5)</td>
<td>14 (6/8)</td>
<td>9 (4/5)</td>
<td>4 (3/0/0/1)</td>
<td>4 (3/0/1/0)</td>
</tr>
<tr>
<td>Loop</td>
<td>22</td>
<td>8 (1/3/2/2)</td>
<td>2 (2/0)</td>
<td>4 (2/2)</td>
<td>5 (0/5)</td>
<td>1 (0/1/0/0)</td>
<td>2 (1/0/1/0)</td>
</tr>
<tr>
<td>Core Conserved</td>
<td>45</td>
<td>16 (2/6/5/3)</td>
<td>10 (6/4)</td>
<td>8 (3/5)</td>
<td>4 (1/3)</td>
<td>4 (3/0/0/1)</td>
<td>3 (2/0/1/0)</td>
</tr>
<tr>
<td>Loop Conserved</td>
<td>16</td>
<td>5 (0/3/1/1)</td>
<td>2 (2/0)</td>
<td>2 (1/1)</td>
<td>4 (0/4)</td>
<td>1 (0/1/0/0)</td>
<td>2 (1/0/1/0)</td>
</tr>
</tbody>
</table>

is also interesting to notice that 7 out of all 11 symmetry-breaking insertions come from adenosine (Table 1). Structure investigation shows that most these insertions are located on the interface between the A- and P-cores, and they very often feature sharp backbone curvatures. Considering adenosine’s propensity to make tertiary contacts (for example, A-minor interactions), it is highly possible that the adenosine dominance of insertion is a structural requirement for the proto-ribosome to gain more stability and maintain evolvability in complexity.

2.2.2 Conserved sequence symmetry in the PTC

Sequence identities between the A- and P-regions based on such structural alignment are summarized in Table 2.1. Excluding the insertions, we see a slightly biased distribution towards identical nucleotide composition between the symmetric A- and P-regions, which raises the question whether the two regions come from two independent random sequences (null hypothesis $H_0$) or they come from a self-replication event. Assuming that the structure of the proto-ribosome is captured by the modern ribosome in a quick transition and since then it is buried deep inside and subjected to little evolutionary changes, we can estimate the significance of our observation against the outcome of a Bernoulli experiment. The probability of observing $k$ identical compositions between two sequences of length $n$ under $H_0$ can be expressed as $Prob(k) = \binom{n}{k} p^k (1-p)^{n-k}$, with $p$ being the probability of two positions being identical ($p = \frac{1}{4}$). The p-value of an outcome possessing at least $K$ identical matches is the accumulative probability for all $k \geq K$: $P_K = \sum_{k \geq K} Prob(k)$. The calculation suggests a small chance (6%) that we see 26 identical nucleotides between the 78 aligned positions of the A- and P- regions. This probability decreases to 0.8% if we consider only the 51 conserved nucleotide pairs. Therefore, we might reject $H_0$ with a 94% confidence level, and conclude that the two symmetric regions in the PTC originated from an ancient duplication of the same RNA sequence.

This optimistic estimation definitely needs more evolutionary scrutiny. The existence of ribosome signatures [39] in the PTC indicates active evolutionary changes after the LUCA even at the heart of the ribosome, which might further increase the confidence of our statistical inference. On the other hand, the critical 6-nucleotide sequence (AUAXA) in the central loop of the P-region shares no sequence similarity with its counterpart sequence in the A-region. This sequence is the binding site for the peptidyl-tRNA in
the ribosome, and it is shown to be present in the in vitro selected ribozymes that binds CCdApPuro [69] or makes peptide bond [70]. This fact appears inconsistent with the assumption that a prebiotic self-duplication event created the two symmetric units of the PTC. The discrepancy may be resolved by investigating the possibility of either lost of functional sequence in the A-region or gain of functional sequence in the P-region. In either case, the two symmetric units should merge under a pressure for greater structural stability or some function other than transpeptidation, which is consistent with the current view of ribosomal evolution [71]. However, this discussion is beyond the scope of this dissertation.

2.3 Design of the Proto-Ribosome Model

2.3.1 Model construction based on tertiary interactions

In the crystal structure of the *T. thermophilus* ribosome, the A- and P-regions interact with each other mainly at the entrance to the exit tunnel through covalent bond linking the backbone of C2051 and G2502, and the non-canonical base-pair formed by U2504 and C2452. In addition, the A-minor interaction between H93 (A2598) and H74 (G2436), and hydrogen bond interaction between U2596 and U2075 increase the binding affinity between the two regions. Considering the small number of interactions, we started the design of the proto-ribosome with RNA helices that can be symmetrically aligned between the two regions (Table S1), plus nucleotides 2502 to 2505, 2056 to 2060, and 2611 to 2612 (see Methods). These additional nucleotides are incorporated into all models to complete the central loop that provides backbone connection, stabilizing base-pairs, and cross-strand base-stacking interactions, even though they don’t share spatial symmetry.

Taking into account the observation that the A- and P-loops are less symmetrically conserved than the cores, the initial models are realized in two ways: 1, a small model that includes only the core helices (H90, H93, H74 and H89) and the central loop (see Methods, model I); 2, a larger model that includes the A- and P-loops (H91, H92 and H75, H80) in addition (model II). Both models were prepared with and without substrates (CpCpA-Phe) bound in the aminoacyl- and peptidyl-tRNA binding sites, and their stability explored by molecular dynamics simulations. It turned out that model I could only keep the aminoayl-substrate for less than 90 ns, and model II held both substrates in the binding pockets for no more than 200 ns. To see whether this is a result of model instability, we calculated the number of hydrogen bonds between A- and P-regions as a measure for the degree of contacts between them. For model I, an average of 4.83 (without substrates) and 3.35 (with substrates) hydrogen bonds between the two regions at any given time were observed, among which most of them are located at the entrance to the exit tunnel. The A-minor interaction between H93 and H74 was lost shortly after the start of the simulation, which leads to large separations over 50 Å between the end loops of these two helices (Figure 2.2C). This separation decreases to a maximum of ∼30 Å for model II. However, the absence of hydrogen bond interactions (4.61 and 6.15 in simulations without and with substrates, respectively) and the loss of A-minor interaction between H93 and H74 disrupted the tertiary structure organization, and lead to unfolding of the proto-ribosome model (Figure 2.2D, yellow box).

Meanwhile, a third model was developed to investigate the hypothesis that the two regions may originate from self-replication of the same RNA sequence. The A-region sequence was chosen as the template due
Figure 2.2: Stability of the designed proto-ribosome constructs. (A) Secondary structure of Construct I: an early model that consists of only the core helices of the PTC. (B) Secondary and 3D structure of Construct II: extended model that includes the A- and P-loops. Tetraloops, linkers and A-minor interactions that are modeled in silico are highlighted in red. (C) Tertiary structure stability is shown using distances measured between the outer-edges of H93 and H74, where the A-minor interaction are kept, in each model trajectories. (D) Representative snapshots are presented to demonstrate the rearrangements of tertiary interactions among helices for each model.

to the end-loop adenosine present in a conformation that provides potential A-minor tertiary interaction. A second copy of the A-region structure was overlapped onto the crystal structure of the P-region to create a complete proto-ribosome (model III). Surprisingly, the core helices stayed in place with respect to each
other for a significant amount of time (∼0.77 µs) in this strictly symmetric model (Figure 2.2D, pink box). Although the average number of hydrogen bonds between the two A-region copies improved only slightly (6.29), one of the two potential A-minor interactions was quickly found within the first 30 ns and stayed for the rest of the simulation. Similar interactions on its symmetric counterpart were established several times during the simulated time.

The sub-microsecond simulation of the fully symmetric model demonstrated the mutual effect of multiple A-minor interactions in stabilizing the overall structure of the proto-ribosome model, and therefore inspired further improvements to our in silico design. In the following step, H89 was shortened to the appropriate length so that the tetra-loop from H93 could be symmetrically fitted in to cap the end of H89 truncation. Such remodeling was made to both model I and II, respectively. The final constructs (model IV & V, Figure 2.2A & B) of the proto-ribosome maintained most their tertiary contacts throughout the 0.6 µs simulation (Figure 2.2D, black and red boxes). Irrespective of the inclusion of A- and P-loops, distance between the end loops of H93 and H74 fluctuated around 11.5 Å without the presence of substrates in the binding pockets. With substrates bound, the fluctuation of distance increased, while the average distance remained even closer to the crystal structure value (Figure 2.2C, cyan). The binding affinity between the A- and P-regions of these two final proto-ribosome models were shown to be augmented by the increase of average number of hydrogen bond interactions between them in model IV (9.87) and model V (8.36), respectively.

2.3.2 Secondary structure stability of the best designed models

Secondary structure for both model IV and V also stayed stable over the total 0.6 µs simulation, with RMSD for each individual helix fluctuating between 3 and 5 Å (Figure 2.3A). Comparing the two models, it is noticed that the RMSD of H93 and H89 were reduced by the addition of A- and P-loops, although the two loops extended from H90 and H74, respectively. A quantitative analysis of the fraction of time each base pair was present in the trajectories (Figure 2.3B,C) revealed that the addition of A- and P-loops stabilized mainly the base pair interactions in the central loop where the amino acid substrates are expected to bind. The central loop involves many non-canonical base pairs reaching across large sequence separations that extend the shorter arm of H93 and H89, and hence exerts a larger affect on their dynamics. Furthermore, base pairing fluctuations displayed symmetrical patterns in accordance with the structural symmetry, particularly in the more stable H90 and H74. The dynamical symmetry is also seen in the A- and P-loops. For example, the loss of base pairing in the H91 (H75) because of backbone widening by the insertion from A2564 & A2565 (A2433 & A2434), was compensated by A-minor interactions made between the inserted adenosines and the minor groove of H91 (H75). The structure of the designed proto-ribosome model maintains a sophisticated balance between dynamical flexibility and the overall structural stability.
Figure 2.3: Comparison of the secondary structure stability of the two best proto-ribosome constructs. (A) Secondary structure stability is shown with the RMSD of individual helices for both models, gray for model IV and black for model V. Fraction of time each crystal structure base pair is present in the simulations of model IV (A) and model V (B). Secondary structure diagram is adjusted according to base pairs inferred by x3DNA [72].

2.3.3 Experimental investigation of the proto-ribosome model

To investigate the ability of the computationally designed proto-ribosome to assemble in solution, RNA molecules correspond to the A- and P-region sequences, as well as the full proto-ribosome were prepared based on model V (see Methods). 10 µM solutions of A-region RNA, P-region RNA, mixture of A- and P-regions, and the AP full model were loaded onto the polyacrylamide gel. The results showed a strong electrostatic dependence of dimer formation for the P-region RNA molecules. Although magnesium ions seemed to eliminate various monomers possibly existed in mis-folded states, they were not required for the association of P-region monomers. On the other hand, the A-region RNA existed dominantly as monomers.
Regardless of the ionic strength in solution. The mixture of the A- and P-region RNAs showed gel bands at positions of both the P-region dimer and A-region monomer in 10 mM [Mg^{2+}], indicating that both RNA sequences folded and interacted independently of the presence of the other type of molecules. These observations agreed with results for the similar but smaller RNA (A- and P-cores) constructs presented in [73].

![Figure 2.4: Formation of the proto-ribosome model in solution.](image)

Moreover, the RNA molecules of the full model prevailed in two populations without magnesium. One of them located at the same position as the P-region dimer, and likely represented a loosely folded conformation as predicted in simulation (see Supporting Information). The other one migrated faster, and therefore characterized a more compactly folded structure, which is greatly favored upon addition of magnesium into the solution. The gel mobility assay is consistent with our micro-second simulations, showing that association between the totally symmetric sequences are not strong enough, but a stable proto-ribosome could be produced under high [Mg^{2+}] with the A- and P-region covalently linked.

### 2.4 Function of the Proto-Ribosome

A functional proto-ribosome must be able to capture the appropriate substrates that are diffusing around, bind them tightly enough so that there would be enough time for the ribozyme to rearrange its own conformation according to the substrate(s) and orient the substrate(s) for peptide bond formation. With the stable model designs described in the previous section, it is now possible to investigate both the binding and the catalyzed reaction fulfilled by the proto-ribosome.

#### 2.4.1 Substrate binding propensity of the proto-ribosome

In the 0.6 μs simulation of model IV with substrates (adenosine-phenylalanine) placed in the aminoacyl- and peptidyl-tRNA binding sites, the proto-ribosome structure stayed stable while the two substrates fell out of the pockets shortly into the simulation. Interestingly, one of the substrate diffused around the proto-ribosome, bound to several locations briefly through base-stacking or A-minor interaction using its adenosine base, and was captured by the end loop of H74 around 300 ns (Figure 2.5A & D). More intriguingly, the other diffusing substrate was finally bound independently to the end loop of H93, brought to close proximity to the first bound substrate by the A-minor interaction made between H93 and H74 (Figure 2.5D). This sce-
nario illustrates the possibility of the designed proto-ribosome model to incorporate appropriate substrates for its catalytic function.

Figure 2.5: Substrate binding propensity of the proto-ribosome. Fraction of time each nucleotide is in contact with any substrates during simulations: (A) Model IV with two substrates A-Phe; (B) Model IV with six substrates A-Ala; (C) Model V with six substrates A-Ala and one CpCpA-Phe occupying the P-site initially. Representative snapshots of the substrates are shown binding in the end loops near the A-minor interaction between H93 and H74 (D); in the peptidyl-tRNA binding site (E), and in the entrance of the exit tunnel (F).

To testify whether the proto-ribosome may incorporate substrates into the designed pocket, additional simulations were performed starting with six adenosine-alanine molecules put at 65 Å away from the center-of-mass (CoM) of the proto-ribosome. All substrates were allowed to diffuse freely with no explicit constraints, and the potential of each nucleotide in the model to bind substrates was recorded and shown in
Figure 2.5 A-C. For proto-ribosome models with and without the A- and P-loops, the more flexible shorter helices H93 and H89 appeared more potent in binding substrates. However, these two helices bound substrates mainly through base-stacking interactions made between nucleotides in their stems and the substrate adenosine. The nature of the interaction determined the non-specificity of these binding sites, and thus explains why they were frequently revisited by different diffusing substrates instead of holding one for a long time. In addition, model IV bound substrates near the A-minor interactions made between H74 and H93, with an extensive interaction network involving both A-minor and base-stacking interactions. The increase in specificity of the interactions led to tight binding between the proto-ribosome and two of the substrates, which stayed for more than 0.5 µs in the simulation. This specificity may also help to select amino acid substrate attached to adenosine over the other nucleotides. The incorporation of A- and P-loops in model V rigidified structure of the end loops, making them less flexible for substrate binding. On the other hand, the stability of the model provided a well-defined primitive “exit tunnel” embanked with central-loop nucleotides. The tunnel appeared as a deep pocket that was capable of fitting in one substrate, moving it deeper into the tunnel and holding it for microsecond (Figure 2.5F). The direction of the substrate movement suggested that the short opening of the exit tunnel in the prebiotic proto-ribosome might serve as a two-way gate, capable of taking new substrate in and expelling used substrate out.

2.4.2 Correct binding of one freely diffusing substrate into the P-site

One of the six substrate molecules bound to the peptidyl-tRNA binding site in model IV transiently (∼5.5 ns), with a smallest RMSD of 1.67 Å with respect to the corresponding substrate in the crystal structure (Figure 2.5D). The situation was significantly improved with model V, which was able to hold a substrate in the peptidyl-tRNA binding site for up to 0.3 µs. This successfully bound substrate entered the proto-ribosome model V through contacts with the end loop of H89, then was quickly handed over to the bulged nucleotide C2573 facilitated by a stable base-stacking interaction (Figure 2.6A). However, in the crystal structure, C2573 is supposed to hydrogen-bond to the 2′ hydroxyl of the aminoacyl-tRNA C75. This difference in interaction prevented the substrate to be oriented correctly for reaction in the aminoacyl-tRNA binding site. Nevertheless, the adjacency of the two binding sites increased the probability of precise substrate binding if it can be kept close to either site. In fact, a transition induced by substrate’s interaction with nucleotides A2602 and A2451 occurred shortly afterwards, which rotated the substrate into the peptidyl-tRNA binding site (Figure 2.6B). The substrate remained bound in the peptidyl-tRNA binding site for the rest of the simulation, with a smallest normalized RMSD of 0.45 Å and a CoM distance of 0.08 Å with respect to the corresponding substrate in the crystal structure.

In the crystal structure, binding of peptidyl-tRNA to the ribosome is kept mainly through the A-minor interaction between A76 and nucleotide A2450, in which the 2′ hydroxyl of A2450 probably serve as both a hydrogen acceptor from N6 atom of A76, and donor to the N1 atom of A76. Two additional hydrogen bonds further assist the binding interaction, one between the 2′ hydroxyls of A2541 and A76, and the other between A2439 (N6 atom) and the A76 phosphate (Figure 2.6C). Encouragingly, a very similar interaction network is observed between the spontaneously bound substrate and the proto-ribosome. Not only the A-minor interaction was observed between the substrate adenosine and A2450, the two auxiliary hydrogen
Figure 2.6: Successful binding of the substrate into the P-site. (A) Binding trajectory of one of the six A-Ala molecules into the catalytic pocket of the proto-ribosome model, purple showing the beginning and yellow showing the end of the trajectory. (B) Normalized RMSD of the binding substrate with respect to crystal structure of the A-site substrate (purple) and P-site substrate (yellow). Snapshots of the substrate are shown in their peptidyl-tRNA binding conformation in the crystal structure (C) and during simulation (D).

bonds also alternated to reinforce substrate binding (Figure 2.6D). Therefore, the microsecond simulation proved that our design of the proto-ribosome is able to bind amino-acylated adenosine with interactions established through the same nucleotides as in the modern ribosome.

2.4.3 Water-mediated dynamical pathway towards peptide bond formation

The last crucial step towards the design of a functional proto-ribosome is to verify that the model may hold the two substrates simultaneously in the pocket and orient them into a conformation favorable for peptide bond formation. This was done by carrying out simulation of model V, bearing the two tRNA mimics (CpCpA-Phe) in their crystal structure conformations. Both substrates stayed tightly bound to their pockets, with all base pairing/stacking interactions between the proto-ribosome and the substrates maintained within the 0.6 µs of simulation. Relative motion between the two substrates were described using two reaction coordinates: 1, distance of the N and C atoms between which the peptide bond forms ($r_{N-C}$); and 2, distance between the O2’ atom of peptidyl-tRNA A76 and the carbonyl oxygen atom of the phenylalanine charged onto the aminoacyl-tRNA ($r_{catalytic}$). The second distance was chosen because of the critical role of the vicinal 2’ hydroxyl group in catalyzing the peptidyl-transfer reaction in the modern ribosome [74–77].

The trajectories of the two distances classified the substrate conformation into three major populations (Figure 2.7A). The last population appeared in the trajectory represented a very stable conformation (state
III), characterized by a small catalytic bond distance ($r_{\text{catalytic}} \sim 2.7\text{Å}$) and a $r_{N-C}$ around 4.5 Å (Figure 2.7A & B). It specifies the same transition state (TS) configuration for peptide-bond formation, obtained using quantum chemistry calculation done independently by Gindulyte et al. [75]. With a hydrogen bond locking the vicinal 2' hydroxyl with the carbonyl oxygen of the aminoacyl-substrate, the amino group was able to approach its destination with the correct S-chirality, although the distance between the bond-forming atoms may not get under the vdW radii in a standard MD simulation.

![Figure 2.7: Peptidyl-transferase reaction catalyzed by the designed proto-ribosome. (A) Distance between the reacting atoms (green) and distance of the catalytic hydrogen bond (orange) over time. Bars on the bottom show the presence of water molecules that receive hydrogen from (red) or donate hydrogen to (blue) the 2' hydroxyl of the peptidyl-tRNA A76. (B) Proposed transition state conformation for the peptidyl-transferase reaction observed in the micro-second simulation. (C) Dwell time of water molecules acting as hydrogen acceptor (donor) to the 2' hydroxyl of the peptidyl-tRNA A76 can be fitted to double-exponential decay (single-exponential decay) process. (D) Hydrogen bond network formed between a water molecule (red) that accepts hydrogen from the 2' hydroxyl of the peptidyl-tRNA A76, substrates (black), and the surrounding residues in the PTC (blue). This network stabilizes an intermediate state before the formation of the catalytic bond for reaction.](image)

Furthermore, a pathway towards the TS conformation for peptide bond formation (state III) can be identified by analyzing the other two populations. The first one populated the beginning of the simulation, centered around $r_{N-C} = 4.3\text{Å}$ and $r_{\text{catalytic}} = 9.7\text{Å}$. In this state, the vicinal hydroxyl group pointed away from the reaction center, and it interacted with proto-ribosome nucleotide C2063 (or C2064, A2439) through a water molecule that may donate hydrogen to both entities. However, each hydrogen-donor-water stayed for only a short average lifetime $\sim 24$ ps (Figure 2.7C), so the interactions remained highly fluctuant. In the second state, the vicinal hydroxyl group turned back to face the reaction center, yet the catalytic bond
may not be formed due to the presence of a water molecule that accepts hydrogen from the 2′ hydroxyl and in the same time donates hydrogen to the amino group of the amino-acylated substrate (Figure 2.7D). This water molecule, unlike the ones bridging the proto-ribosome and the vicinal hydroxyl group in state I, has a much longer life time of 90 ps and thus stabilizes intermediate state II. Comparing to the one in the modern ribosome LSU, the active site of the proto-ribosome is much more water accessible. Therefore, the water-mediated pathway defined by transitions from state I to state II and finally towards the TS conformation, envisions a reasonable reaction mechanism for peptide bond formation catalyzed by the designed prebiotic ribozyme.

2.4.4 Discussion

Evolution of the LSU rRNA has been studied extensively to map out the relative age of its subdomains. The oldest regions on the 23S deduced by comparative sequence analysis [43] largely overlap with the one obtained under the connectivity argument, which looks for the older regions based on their larger number of long-range base-base interactions [44]. Both findings are surprisingly captured in the mitochondrial large subunit rRNAs of Trypanosoma brucei [78]. Analysis of the crystal structures of the 23S rRNA improves on the connectivity argument by showing that the A-minor interactions would infer a similar hierarchical timing on the 23S rRNA evolution [45], such that PTC in domain V and part of domain II remain the oldest among all. Furthermore, comparison of structures from the two domains of life, Bacteria and Archaea, identifies four magnesium clusters, one of which centered on the PTC and another two connecting domain II with PTC and domain IV. All studies agreed on the fact that the PTC region originated in the RNA world as the core, with subsequent addition of RNA and protein components over time developed into the modern ribosome. A very recent effort proved successful to make an ancestral ribosome with the PTC and fragments from domain II an IV. This model, which incorporates ~20% of the LSU nucleotides, assumes a 23S rRNA like secondary and tertiary structure, and forms native complexes with ribosomal protein fragments [79].

Considering the difficulty of prebiotic synthesis of large RNA molecules, we employed a bottom-up scheme and started the design of a functional proto-ribosome from the very small core fragments of the PTC. Based on analysis of the structural and sequence symmetry, we remodeled the A- and P-core interactions with an additional A-minor motif, which turned out to increase the stability of our construct significantly. Further incorporation of the A- and P-loops results in a functional proto-ribosome, which is capable of binding one A-Ala substrate into its P-site and orienting two properly bound substrates into a transition state favorable for peptide bond formation. The final design is 191 nucleotides in length, in surprising agreement with the in vitro selected ribozymes (187 to 196 nt) that catalyze a similar peptidyltransferase reaction [60; 70]. Although the secondary structures differ between the in vitro selected and in silico designed ribozymes, they both share a 6-nucleotide sequence (AUAAXA) corresponding to the P-core central loop sequence in the modern ribosome. It is highly unlikely that this consensus is only a coincidence, because the same sequence has also been identified in the ribozyme evolved in vitro to bind CCdApPuro, a compound mimics the aminoacyl-RNA substrates and inhibits the ribosome [69]. Furthermore, the characteristic sequence carries two As that play critical role in binding the substrate (Figure 2.6D) and forming an intermediate towards the TS for peptide bond formation (Figure 2.7D). The consistency of such a binding sequence
speaks for its own essentiality in the evolution of the RNA machinery of translation, and also demonstrates the feasibility of the designed proto-ribosome model being the origin of the ribosome in the RNA world.

With a delicately designed proto-ribosome model, conjectures regarding the prebiotic translational machinery can be tested. For example, one may inquire the origin of the chirality preference of L-amino acids in modern organisms. It has been shown that D-amino acids can be incorporated into the A- and P-sites of the ribosome [80], and initiate protein synthesis under certain conditions [81]. Indeed, we see binding of the A-Ala molecule maintained by interaction only between the adenosine and proto-ribosome nucleotides, indicating that the preference in the L-chirality is more likely due to selection than being an intrinsic property of the proto-ribosome. Moreover, with the model, one may address how this selection could have occurred by studying how partially chiral peptide chains could improve the stability of the proto-ribosome and in turn exert further chirality.

The model proto-ribosome also provides us with an opportunity to investigate the mechanisms of modern ribosome functioning. Starting from its minimal size, we might first be able to resolve the controversy of how the ribosome catalyze the peptidyltransferase reaction. Two main opinions exist regarding the catalytic role of the 2′ vicinal hydroxyl on the peptidyl-tRNA A76. The proton-shuttle mechanism is greatly favored over hydrogen-bonding stabilization of the TS conformation, based on experimental measurements showing that removal of the 2′-OH results in a $10^6$ fold reduction in the rate of peptide bond formation [74]. It is argued that a few hydrogen bonds are not enough to provide that acceleration in reaction rates energetically. However, a more carefully designed study corrected the reduction in reactivity to be $\sim 100$ to 2000 fold, of which 10 fold came from intrinsic chemical difference between rA76 and dA76 [77]. This new estimation puts the two hypothesis again on the same scale. Meanwhile, quantum chemistry (QM) calculations were applied to propose TS conformations with the lowest energy barrier to form the reaction products, yet they hardly agreed on how the proton-transfer mechanism could have occurred. For example, studies in the Åqvist lab introduce three proton-shuttle transition intermediates (4-membered, 6-membered and 8-membered ring) involving different number of water molecules, based on various computational and experimental techniques [82–84]. They also present a different view about the rate-limiting step than inferred from kinetic isotope effect analysis [85]. Others [86–88] arrive at even more TS conformations using different reactant molecules mimicking the reaction or including different number of atoms in the QM calculation. Interestingly, the hydrogen-bond stabilization of the reactant is always seen as a local minimum, sometimes even presented with a comparable barrier height to the proton-shuttle approach [86]. In our micro-second simulation, a stable conformation involving the 2′-OH hydrogen bond is seen in accordance with the QM simulation presented in [75], supporting the hydrogen-bonding stabilization mechanism. On the other hand, angles of the hydrogen bonds in the 4- or 6-membered ring transition states are inconsistent with their existence under the current non-polarizable classic force field. No water molecule is observed in stable interaction with the substrates as suggested in other QM TS conformations. These observations, although not in favor of, do not prove the proton-shuttle to be wrong either, due to difference in water accessibility of the active pocket in the proto-ribosome and in the ribosome. Nevertheless, the proto-ribosome presented in this paper reduces the size of the functional molecular unit capable of catalyzing the reaction, so that a QM-MM calculation might be possible to explore the peptidyl-transferase process including all con-
tributors and the exact substrates. In prospect, gradual addition of ribosomal components to the available proto-ribosome model will offer a dynamic view on the evolution of ribosomal functions.

2.5 Methods

2.5.1 Structural and sequence conservation

Structural alignment were performed using STAMP structure alignment [68] implemented in VMD [89]. Crystal structures of the ribosome from six organisms were used for our structural conservation study: *Thermus thermophilus* [90] (PDB code: 2WDL), *Escherichia coli* [91] (PDB code: 2I2T), *Deinococcus radiodurans* [28] (PDB code: 1NJP), *Haloarcula marismortui* [92] (PDB code: 1KQS), *Saccharomyces cerevisiae* [9] (PDB code: 3U5D), and *Tetrahymena thermophila* [10] (PDB code: 4A18). Structural alignment between the A- and P- regions are done using *T. thermophilus* structure alone.

Multiple alignment of the 23S rRNA sequences was downloaded from Comparative RNA Web Site (CRW, Gutell Lab). To remove bias in sampling, we constructed the non-redundant set of domain V with a 95% sequence identity cutoff, which results in 106 23S rRNA sequences in total (78 bacterial and 28 archaeal sequences). The resulting alignment was used to get conservation statistics between A- and P-regions.

2.5.2 Model construction

Crystal structure of the *T. thermophilus* ribosome [90] (PDB code: 2WDL/2WDK) was used for constructing the proto-ribosome models. Gaps in the helix resulting from truncation of the proto-ribosome model from the ribosomal LSU are closed by the frequently occurring tetraloop GCAA, of which the coordinates are taken from crystal structure of the *E. coli* ribosomal LSU (PDB code: 2I2T) nucleotide (nt) 1869 to 1872. Mg$^{2+}$ ions within 3.2 Å of the proto-ribosome model are considered to be directly bound and therefore are included in the model. One Mg$^{2+}$ ion is seen in multiple crystal structures to bind tightly to the backbone phosphate of the P-site tRNA C75 and A76 as well as nucleotide A2602. In order to stabilize the A-site substrate in the final model, an extra Mg$^{2+}$ was modeled by symmetry, and bridges backbone phosphate of the A-site tRNA C75 and nucleotide U2493. The natural substrate CpCpA-Phe is obtained by replacing the amide linkage between A76 and the amino acid in the crystal structure of the modified Phe-tRNA$^{Phe}$ by an ester bond. When present, the peptidyl-tRNA binding site (P-site) substrate is acetylated, and the aminoacyl-tRNA binding site (A-site) substrate is neutralized with -NH$_2$, to best mimic the pre-peptidyl-transfer state of the proto-ribosome.

A summary of all models studied in the paper is provided in Table 2.2. To be specific, model I are constructed using nt 2056 – 2078, 2434 – 2463, 2487 – 2521, 2565 – 2612 of the 23S rRNA. Gaps between strands are closed by single nucleotide C, tetra-loop GCAA, single nucleotide C, respectively. Model II are constructed using nt 2056 – 2079, 2241 – 2258, 2430 – 2463, 2487 – 2522, 2543 – 2612 of the 23S rRNA. Gaps between strands are closed by tetra-loop GCAA, dinucleotide CA, GCAA, GCAA, respectively. Model III is constructed from two duplicated A-regions, including two copies of nt 2502 – 2522, 2543 – 2610 of the 23S rRNA. In addition to the GCAA tetra-loop that caps the gap within each region, the two A-regions
Table 2.2: Summary of all proto-ribosome models

<table>
<thead>
<tr>
<th>Model</th>
<th>Sim. ID</th>
<th>NT* #</th>
<th>NT§ #</th>
<th>Mg²⁺ Cutoff (Å)</th>
<th>Duration (µs)</th>
<th>Performance (µs/day)</th>
<th>Ligands</th>
</tr>
</thead>
<tbody>
<tr>
<td>I</td>
<td>1</td>
<td>142</td>
<td>6</td>
<td>25</td>
<td>9.098</td>
<td>0.186</td>
<td>-</td>
</tr>
<tr>
<td></td>
<td>2</td>
<td>142</td>
<td>6</td>
<td>25</td>
<td>9.055</td>
<td>0.126</td>
<td>CpCpA-Phe (A &amp; P)</td>
</tr>
<tr>
<td>II</td>
<td>3</td>
<td>196</td>
<td>14</td>
<td>29</td>
<td>9.382</td>
<td>0.210</td>
<td>-</td>
</tr>
<tr>
<td></td>
<td>4</td>
<td>196</td>
<td>14</td>
<td>29</td>
<td>10.179</td>
<td>0.403</td>
<td>CpCpA-Phe (A &amp; P)</td>
</tr>
<tr>
<td>III</td>
<td>5</td>
<td>186</td>
<td>8</td>
<td>30</td>
<td>10.496</td>
<td>0.768</td>
<td>-</td>
</tr>
<tr>
<td>IV</td>
<td>6</td>
<td>137</td>
<td>6</td>
<td>25</td>
<td>9.448</td>
<td>0.605</td>
<td>A-Phe (A &amp; P)</td>
</tr>
<tr>
<td></td>
<td>7</td>
<td>137</td>
<td>6</td>
<td>25</td>
<td>9.362</td>
<td>0.649</td>
<td>A-Ala (A &amp; P &amp; 4 diffusing)</td>
</tr>
<tr>
<td>V</td>
<td>8</td>
<td>191</td>
<td>14</td>
<td>30</td>
<td>9.809</td>
<td>0.605</td>
<td>-</td>
</tr>
<tr>
<td></td>
<td>9</td>
<td>191</td>
<td>14</td>
<td>30</td>
<td>9.932</td>
<td>0.605</td>
<td>CpCpA-Phe (A &amp; P)</td>
</tr>
<tr>
<td></td>
<td>10</td>
<td>191</td>
<td>14</td>
<td>29</td>
<td>10.157</td>
<td>1.205</td>
<td>CpCpA-Phe (P), A-Ala (6 diffusing)</td>
</tr>
</tbody>
</table>

* Total number of nucleotides of the designed system;  
§ Number of modeled nucleotides in the system.

are connected through backbone of nt 2502 and 2610 to create a continuous molecule as the other models. Model IV are constructed using nt 2056 – 2078, 2434 – 2461, 2490 – 2521, 2565 – 2612 of the 23S rRNA. These two models are different from model I in that H89 is shortened and capped with the tetra-loop GUGA taken symmetrically from H93. This modification offers an additional A-minor interaction between the two regions that stabilized the overall structure. The same improvement in model structure are made for model V with respect to model II.

### 2.5.3 Molecular dynamics simulations

The parameters for the substrates were derived by analogy to adenosine and phenylalanine/alanine in the CHARMM27 force field. The parameters for the ester bond connecting the charged amino acids to the 2′ hydroxyl group of A76 were taken from those developed previously from the ester bond in fatty acids [93]. All systems were neutralized with K⁺ placed by Ionize [94], in addition to the Mg²⁺ ions kept from the crystal structure (see previous section). Water molecules were placed to complete the primary solvation shells for all Mg²⁺ atoms, and hydrogens were added to the system by psfgen in VMD [89]. 150 mM KCl was added to the final solution. Solvent was prepared in two phases: first Solvate [95] was run to place the first five solvation layers and then the solvate plugin to VMD was used to complete the water box.

5-ns equilibrium simulations were performed using NAMD [96] with the CHARMM27 force field [97] in order to provide the Anton machine with equilibrated systems. Stepwise minimization and equilibration were performed according to the protein:RNA simulation protocol established in Eargle et al [93]. Equilibration runs used periodic boundary conditions and the NPT ensemble with pressure set to 1 atmosphere and temperature set to 298 K. Multiple time-stepping was used to calculate bonded interactions at 1 fs, van der Waals (vdW) interactions every 2 fs, and electrostatic forces every 4 fs. Particle mesh ewald summation
was used to evaluate electrostatic interactions, and the vdW force calculations used a cutoff of 12 Å and a switching distance of 10 Å.

To maintain consistency and stable nucleic acid simulations, we performed NPT ensemble simulation on Anton at 298 K and 1 atmosphere pressure with periodic boundary conditions. Multiple time stepping was employed, with an integration timestep of 1.0 fs. Short-range forces and long-range electrostatics were evaluated every timestep and every four timesteps, respectively. Short-range non-bonded interactions were calculated with cutoffs shown in Table 2.2; long-range electrostatics was calculated using the Gaussian Split Ewald method with a \( 64 \times 64 \times 64 \) grid. The performance on Anton for our systems using the above parameters range from 0.95 to 2.0 \( \mu \)s/day depending on the models (Table 2.2).

### 2.5.4 Analysis of molecular dynamics trajectories

**Distances** between the end of helices H93 and H74 were calculated between the P atoms of nucleotide A2598 and U2075. These two atoms were chosen due to their close adjacency in the crystal structure (\( \sim 7.3 \) Å). Because of the participation of nucleotide A2598 in the critical A-minor interaction that brings together H93 and H74, this distance indicates degree of separation between the two helices.

**Hydrogen bonds** between the A- and P-regions are determined for all nitrogen and oxygen atoms with a distance cutoff of 3.5 Å and an angle cutoff of 30°. The same criteria are used for calculating hydrogen bonds between the 2′ hydroxyl of the P-site A76 and water molecules.

**Base pairs interactions** were determined using 3DNA [72], which reports both canonical and non-canonical base pairs.

**Root mean square deviation (RMSD)** of the substrate with respect to crystal structure is calculated using nucleotides that are: 1, in direct contact with the substrate; 2, stable during simulation. Therefore, the peptidyl-substrate is aligned using 3 nucleotides 2450, 2451, and 2585. The aminoacyl-substrate is aligned by 8 nucleotides 2451, 2452, 2506, 2507, 2573, 2583, 2584, and 2585. The RMSD values are then further normalized by subtracting RMSD of the substrate itself with respect to its crystal structure. The resulting RMSD represents best how much the substrate deviates from its binding position.

### 2.5.5 RNA preparation and native gel electrophoresis

DNA sequences of the A-region (GATGTCGGCTCGATCCTGCAAGGGTTGGGCTTGTTCCCAT TAAAGCCGACCGCAGCTGGTTCAGAAGCTGTAGACAGTTTCGTCTC), P-region (GAAAAAG ACCCCGTGGAGCTTTACTGCAAAGTTTGACTGGGGCGGTCCAAATAAAGTTACCCCGGGGATAAA CAGGCTGATCGTGAGTTTGGCACCTC) and the combined A- and P-regions (GAAAAAGACCCCGTGAGCTTTACTGCAAAGTTTGACTGGGGCGGTCCAAATAAAGTTACCCCGGGGATAAA CAGGCTGATCGTGAGTTTGGCACCTC) were designed according to the stable model V, and inserted into the pUC57 vector. To generate RNA sequences, the plasmids were
digested with BgIII and HindIII, and then the templates were amplified from the digested fragments through PCR. In vitro transcriptions with T7 RNA polymerase were performed using standard approach. Purified RNA at 10µM concentration was annealed in 20 mM K-HEPES, 10 mM MgCl₂ and 330 mM KCl, at 72°C and 50°C each for 5 minutes, then cooled at room temperature, and run on 10% polyacrylamide gel in TAE buffer supplemented with 10 mM MgCl₂.
Chapter 3

Communal Sharing of Ribosomal Proteins Among the Early Bacteria

3.1 Background

The ribosome is an elaborate ribonucleoprotein complex whose evolution is intrinsically linked with that of the cell. It has been recognized since the 1970’s that the molecular core of the ribosome was in place before the divergence of the three primary organismal lineages, Bacteria, Archaea, and Eucarya (the domains of life). The history of these lineages, as inferred from the ribosomal RNA (rRNA) and represented by the universal phylogenetic tree (UPT) [37], provides an organismal reference by which the evolutionary history of a gene can be studied. Despite conservation of a large portion of the ribosomal structure among the lineages, the ribosomes of each domain of life contain certain sequence and structural signatures that are unique to and constant within the domain. Such signatures have been identified in both the rRNA and ribosomal proteins (r-proteins), including many r-proteins that are specific to one of the primary lineages. This suggests that both large and small scale changes in the ribosome were still evolving after the domains diverged and then spreading among all of a domain’s developing sub-branches [39]. The exact mechanism by which homogenization of the branches might have occurred is unclear and certainly a matter of some debate, but pervasive horizontal gene transfer (HGT) among aboriginal cellular life [38; 98] is one possible mechanism. Studying the pattern by which such a signature spread among the evolving lineages can help resolve the dynamics of the evolutionary process at the time.

Horizontal gene transfer, the acquisition of non-inherited genetic material, is widely regarded as a common and important evolutionary phenomenon [99–104]. It is now understood that HGT allows microorganisms to break out of strictly clonal, bifurcating lineages in their search for genetic innovation [105]. Despite the complexity of the ribosome and the potential for malfunction from acquiring a new version of a single ribosomal component, r-protein genes are known to have been horizontally transferred within a domain of life, although no inter-domain HGT has been identified. The first instance of HGT of an r-protein gene was reported by Brochier et al. [106] for S14. In their study, they classified the bacterial S14 sequences into distinct groups based on characteristic insertions or deletions (indels) and presented phylogenetic evidence

---

The contents of this chapter are based in part on work previously published as Ke Chen†, Elijah Roberts†, and Zaida Luthey-Schulten. “Horizontal gene transfer of zinc and non-zinc forms of bacterial ribosomal protein S4,” BMC Evol. Biol., 9:179 (2009). † denotes equal contributions.
that, in some cases, the groups were at odds with the classical bacterial phylogeny. They argued that these discrepancies, as well as unusual gene ordering and duplications in the affected lineages, were the result of ancient HGT events. Thus, they proposed that there must have been some evolutionary pressure favoring the fixation of the transferred r-protein gene, in accordance with the “complexity hypothesis” of Jain et al. [107] regarding the lower probability of HGT for informational genes.

A later bioinformatics study further extended analysis of HGT and gene duplication in the r-proteins. Using genomes of thirty bacteria and genomic data for r-proteins of mitochondria and chloroplasts from seven eukaryotic organisms, Makarova et al. [108] found six additional ribosomal proteins (S18, L28, L31, L32, L33, L36) that shared similar evolutionary patterns to S14 within the bacterial lineage, including discrepancies in genome organization and gene copy number. Furthermore, they showed that the phylogenetic patterns were related to the zinc binding ability of the r-proteins. Two variants were found of each r-protein: one containing a zinc finger motif with four conserved cysteine residues (or occasionally three cysteine and one histidine residue) and another with a complete or partial disruption of the motif. The two variants were referred to as C+ and C−, respectively. Their data suggested that in each case the C+ variant was the ancestral form and that ancient gene duplication followed by disruption of the zinc finger in the paralog and later loss of the original C+ gene in some lineages (differential gene loss; DGL) was the major evolutionary pattern with HGT also occasionally occurring.

Following initial identification of C− variants of bacterial zinc-binding r-proteins, other laboratories began investigating their regulation in organisms with both C+ and C− genes to better understand the evolutionary pressures giving rise to the C− forms. It was predicted theoretically [109] and then found experimentally, first in Bacillus subtilis for S14 and L31 [110; 111] and then in Streptomyces coelicolor for S14, L28, L31, L32, L33, and L36 [112] and in Mycobacterium tuberculosis for S14, S18, L28, and L33 [113], that the paralogous C− versions of some r-protein genes were up-regulated under conditions of low zinc. These groups proposed that the C− paralogs served two possible functions, release of free zinc into the cell in low zinc environments (by ribosomal exchange with the endogenous C+ protein) and/or continued production of ribosomes under zinc-limiting conditions. Whether these are the only pressures that gave rise to the C− forms is unknown, but it is clear that some ribosomal proteins have a unique and interesting evolutionary history related to zinc binding.

All seven zinc binding r-proteins discussed above, except for S14, are unique to Bacteria. Such domain specific r-proteins are signatures of the bacterial ribosome. Roberts et al. [39] showed that the signatures are not limited to just entire domain specific ribosomal proteins, but can also take the form of domain specific insertions in the r-proteins that are universally distributed among all three domains of life. Such a case is found in the universal r-protein S4, a two domain protein ∼200 amino acids in length that is essential for the initiation of small subunit (SSU) ribosomal assembly and translational accuracy. The C-terminal domain of S4 (residues 46–206; all residues given in terms of Escherichia coli numbering) is known to be an RNA binding domain, binding to both rRNA and messenger RNA (mRNA) [114–116], and is homologous between Bacteria and Archaea. While the overall sequence identity for the C-terminal domain is only 36% among bacteria and 32% across all domains of life, the region making contact with the ribosomal RNA is conserved with an average sequence identity of 46% and 40% respectively. The
N-terminal domain, in contrast, appears to be non-homologous between Bacteria and Archaea and was identified in [39] as a bacterial-specific structural signature that coevolved with a bacterial specific extension of an RNA helix (helix h16) on the 16S rRNA. Alone, the bacterial S4 structure has been determined only without the unstructured N-terminus [117], but when complexed with the ribosome the structure of the full protein has been determined. The crystal structure of the Thermus thermophilus ribosome [118] shows that the N-terminal domain of S4 contains a zinc finger motif ligated to a zinc atom and the sequence analysis presented here shows conservation of the four cysteine residues in the zinc-finger motif only in a subset of the bacterial lineages. This variation in zinc binding ability within the bacterial lineages of S4 was overlooked in previous studies of the evolutionary history of zinc-binding r-proteins.

The recent growth in the number of available bacterial genome sequences allows a broad evolutionary history of a gene to be reconstructed, especially in regard to HGT [119; 120]. Besides sequence data for phylogenetic reconstructions, full genomes provide data on genome organization and gene distribution, which are particularly useful in aiding interpretation of possible HGT events. In this study, we use 660 available bacterial genomes to study the evolution of ribosomal protein S4 in the bacteria. We find that S4 can be classified into C+ and C- variants (zinc binding and non-zinc binding, respectively), with multiple independent origins of the C- form. A maximum likelihood tree of S4 shows disagreement with the standard bacterial phylogeny, indicating a more complex evolutionary history than previously known. Considering the fact that the S4 gene is part of a highly conserved gene cluster in bacteria consisting of the S10–spc–α operons [121], we see surprising evidence for the endogenous origin of the C- form in some phyla and hypothesize that both the C+ and C- forms may have been present before the bacterial phyla diverged with different lineages sampling from the variants according to the local environment. In accordance with this hypothesis, we also present evidence that C- paralogous copies in genomes containing both variants of S4, as well as all S4 genes outside the α-operon, are results of HGT events. Regulation of the paralogous S4 genes seems to differ from the zinc-binding r-proteins previously identified, and the expanded distribution of the C+/C- variants in all the zinc-binding r-proteins we present may provide insight on the evolution of zinc usage in bacterial lineages.

### 3.2 Classification and Phylogeny of Ribosomal Protein S4

#### 3.2.1 Sequence alignment and classification of bacterial r-protein S4

To study the history of S4 in bacteria, we first extracted 688 sequences of S4 and paralogs from 660 complete and draft bacterial genomes. We then constructed a multiple sequence alignment (MSA) of the sequences using automated alignment tools followed by manual correction (see Methods). Any evolutionary study of a large set of diverse organisms is likely to reveal a complex history, so to aid further analysis of the relationships between the sequences we classified them into six types using key sequence and structural signatures that define apparently monophyletic groups (such features are also known as synapomorphies). These types classify the sequences according to the presence of or disruption pattern in the zinc finger motif. We find one C+ version (with four conserved cysteines) and five C- subtypes (with various patterns of loss of the zinc finger). C-(I), C-(II) and C-(III) sequences possess a seven residue indel present in the C+ type but
Figure 3.1: Multiple sequence alignment of ribosomal protein S4. Shown is a representative sample of the full sequence alignment. Sequences are grouped according to specific sequence characteristics (see text) and positions are colored by conservation within the group at (blue) 95% and (red) 70%. The arrows above the alignment indicate positions of the two pairs of cysteine residues. The three-letter abbreviations indicate the phylum or class that the organisms belong to: ACB (Acidobacteria), ACT (Actinobacteria), ALP (Alphaproteobacteria), AQF (Aquificae), BAT (Bacteroidetes), BET (Betaproteobacteria), CHF (Chloroflexi), CHL (Chlamydiae), CHR (Chlorobi), CLT (Clostridia), CYN (Cyanobacteria), DEL (Deltaproteobacteria), DTH (Deinococcus-Thermus), EPS (Epsilonproteobacteria), FUS (Fusobacteria), GAM (Gammaproteobacteria), MAG (Magnetococcus), MOL (Mollicutes), PLN (Planctomycetes), SPR (Spirochaetes), VER (Verrucomicrobia).

show gradual loss of the four cysteines (from two to one to zero). C-(IV) and C-(V) subtypes are missing the indel characteristic of the C+ type as well as all four cysteines. Further distinctions between the C- subtypes are based on sequence signatures in the N-terminus. Figure 3.1 shows a sample of the N-terminal portion of the alignment from all of the major bacterial phyla grouped according to these classifications.

From a conservation analysis of the MSA, it is apparent that the S4 sequences can be broadly classified into C+ and C- variants (following the notation introduced by Makarova et al. [108]) based on the conservation of four cysteine residues in the N-terminal domain. C+ type sequences contain two conserved pairs of cysteine residues in a “CXXX...CXXX” motif. The first pair appears near the beginning of the sequence (at residues 9 and 12) and the second pair in a seven residue segment that is an insertion relative to most of the C- sequences (the first red block in Figure 3.1). As shown in the T. thermophilus ribosome structure,
these four cysteine residues bind a zinc ion. Within the C+ group, the N-terminal domain is highly conserved with an average percent sequence identity of 65%. The C+ group includes sequences from diverse bacteria groups: Acidobacteria, Actinobacteria, Chloroflexi, Clostridia, Cyanobacteria, Deinococcus-Thermus, Planctomycetes, Proteobacteria (Beta and Delta classes) and Thermotogae.

The C- variants of S4 show less homogeneity than their C+ counterparts. By definition, they all lack the four cysteine residues, but other characteristic features in the N-terminus allow them to be further classified according to their likely evolutionary origin. The major distinguishing feature between the C- forms is the presence or absence of the seven residue indel that contains the second pair of cysteine residues in the C+ form. Three C- types, C-(I), C-(II), and C-(III), possess the indel, but have disruption of the zinc binding motif. Each of these types is confined to a small portion of the bacterial tree, while sequences containing the indel with the conserved cysteine residues are seen in a wide variety of bacteria. This difference suggests that each of these three groups may have been formed by relatively recent, independent mutations of an ancestral C+ form.

To test this hypothesis, we performed a phylogenetic analysis of the sequences in the C-(I), C-(II), and C-(III) groups relative to the C+ sequences. The first group, C-(I), includes the S4 sequences from most of Betaproteobacteria and all of Gammaproteobacteria. The remaining Betaproteobacteria are all of the C+ type, and, interestingly, the C-(I) Betaproteobacteria show a gradual loss of the four cysteine residues from two to one and, finally, to zero. The Gammaproteobacteria also show a distribution of two, one, or zero cysteine residues. The C-(I) sequences have also lost a three residue turn (the second red block in Figure 3.1) compared to all of the other variants of S4. Figure 3.2 shows a maximum-likelihood (ML) reconstruction of the phylogenetic history of S4 in Proteobacteria except Alphaproteobacteria, which lack the seven residue indel. In the tree, Betaproteobacteria and Gammaproteobacteria share a common ancestor containing the C+ variant of S4 and the root of the Proteobacteria also appears to have been a C+ type S4. The most parsimonious explanation for the origin of the C-(I) form appears to be that a single evolutionary event, characterized by the deletion of the three residue turn and loss of the zinc-binding motif, occurred in the Betaproteobacteria lineage and was inherited monophyletically by the descendant Betaproteobacteria and the Gammaproteobacteria.

The C-(II) group contains all and exclusively sequences from the Epsilonproteobacteria. These sequences do contain the three residue turn that the C-(I) Beta- and Gammaproteobacteria sequences are missing. Additionally, the pattern of disruption in the zinc-binding motif is markedly different from the C-(I) group. In C-(II) sequences, the four cysteine residues are consistently replaced by two glutamic acid residues, one arginine residue, and one serine residue, as opposed to the variety of residues seen in C-(I). Differences in both the pattern of indels and of motif disruption suggest an independent origin for the C-(II) form and phylogenetic analysis supports this interpretation. In the tree shown in Figure 3.2, Epsilonproteobacteria branches outside of the C-(I) group, appearing to diverge near the root of Proteobacteria. The low bootstrap values at higher branch points do cast uncertainty as to whether Epsilonproteobacteria diverged from a common Proteobacteria ancestor or directly from the bacterial root. In either case, however, the C-(II) sequences would be a result of an independent mutation event in an ancestral C+ form that occurred after the divergence of the Epsilonproteobacteria lineage. Conservation of the “EXXE...RXXXXXS”
motif suggests that a salt bridge may have replaced the zinc finger as a structural element in the C-(II) S4 sequences, and threading of an epsilonproteobacterial sequence onto the T. thermophilus crystal structure of S4 confirms that the residues would be properly oriented.

The final C- type containing the seven residue indel, C-(III), is made up of S4 sequences from a subset of Spirochetes: the genus Leptospira. All other Spirochetes currently sequenced lack the indel in r-protein S4. C-(III) sequences have a zinc disruption pattern of “VXXM...LXXXXS” or “VXXM...FXXXXF” and do have the three residue turn missing in C-(I). Additionally, there are numerous sequences signatures separating the C-(III) sequences from those in either group C-(I) or C-(II). Phylogenetically, these sequences appear to branch directly from the root of the C+ form, no further relationships can be resolved. Since the C-(III) group appears to monophyletically descend from an ancestral C+ form, we consider that it too was
an independent evolution of zinc disruption in S4.

All of the remaining S4 C- sequences lack the seven residue indel and both pairs of cysteine residues. C-(IV), the largest C- group, consists of r-protein S4 sequences from a wide variety of bacteria: Actinobacteria, Alphaproteobacteria, Aquificae, Bacilli, Bacteroidetes, Betaproteobacteria, Chlamydiae, Chlorobia, Chloroflexi, Cyanobacteria, Deinococcus-Thermus, Deltaproteobacteria, Fusobacteria, Mollicutes, Planctomycetes, Spirochaetes, and Verrucomicrobia. The N-terminal domains of the C-(IV) sequence are much less conserved than the C+ form, having an average percent identity of 36%, and do not contain any characteristic sequence or structural signatures by which they could be further classified.

A small number of Clostridia sequences (17) constitutes the last defined type, C-(V). These C- sequences lack the seven residue indel, but are different from C-(IV) sequences (and all other S4 sequences) in that they are missing a “PGXHG” motif starting at residue 38. This motif is highly conserved in the other S4 sequences and is unambiguously alignable across all other groups. In C-(V) Clostridia sequences, this region is 2–4 residues shorter and can not be reliably aligned to the other types. All but one of the S4 sequences in this group are from genomes that also contain a C+ type S4.

### 3.2.2 Phylogenetic reconstruction of S4’s evolutionary history

The evolutionary history of S4 was analyzed using a ML phylogenetic reconstruction of all the sequences of r-protein S4 and its paralogs obtained from the 660 genomes, as described in Methods. Figure 3.3 shows an unrooted phylogenetic tree obtained from a consensus of 1000 ML trees. Like many phylogenetic reconstructions using a large number of sequences, branch points above the bacteria phyla level are difficult or impossible to reliably determine [122] and most branches appear to radiate from a few ancestral points in our consensus tree. Trees of the C-terminal RNA binding domain and the N-terminal bacterial specific domain were also generated separately using the same method (data not shown). The C-terminal tree had similar branchings as the tree shown in Figure 3.3, but with fewer well-supported branches near the bacterial root. The N-terminal domain, however, is too short to draw any reliable conclusions regarding its relative contribution to the phylogenetic signal.

The consensus phylogenetic tree of the entire protein shows good agreement with the classifications of S4 that we introduced earlier. It is roughly divided into two central foci, one representing the C+ form (white) and the other the C-(IV) form (blue), although a few C-(IV) lineages branch within the C+ half. C-(I), C-(II), and C-(III) (gray) are recent, independent mutations of an original C+ form (discussed above), and we treat them as C+ for the remainder of the discussion. The C-(V) form (yellow) is a monophyletic branch descending from the C-(IV) root.

Within the C+ branch of the tree, three bacteria phyla are monophyletic with high support values and yet contain both C+ and C-(IV) forms: Actinobacteria, Chloroflexi and Deinococcus-Thermus. In each case, there are two branches descending from the phylum that are monophyletic, one with respect to C+ and one to C-(IV). Despite the sequences in each C+ branch containing the seven residue indel characteristic of the C+ group and sequences in each C-(IV) branch lacking the indel, the branches have a higher average sequence identity (56%, 46%, 60%, respectively) than in general would be expected for a C+ and a C-(IV) group (~40%). We therefore consider it likely that these are real phylogenetic branches and not artifacts,
Figure 3.3: Consensus unrooted phylogenetic tree of S4. The tree was constructed from 1000 maximum-likelihood inferences. Node labels are bootstrap proportions estimated from 5000 replicates. Branches that are monophyletic with respect to a phylum or class and also with respect to a sequence classification have been collapsed, parentheses give the number of sequences in the branch. Colors indicate the sequence classifications: (white) C+, (gray) C-(I) – C-(III), (blue) C-(IV), and (yellow) C-(V). Sequences from genomes with multiple divergent copies of an S4 gene are marked with a (†) dagger. The two sequences marked with an (*) asterisk are identical copies resulting from large-scale genome duplication.

particularly for Actinobacteria and Deinococcus-Thermus.

Also within the C+ branch, the C-(IV) Chlorobi and C+ Epsilonproteobacteria group together, with an average percent identity of 50%. However, the Chlorobi sequences have nearly as high average percent identity with C-(IV) groups (46%) and the support value of 52 is fairly low, so this grouping may be a reconstruction artifact.

Comparing the S4 phylogenetic tree with the classical bacterial phylogeny, many bacterial groups show good agreement with the tree at the phyla level: The phyla Aquificae, Bacteroidetes, Chlamydiae, Tenericutes (Mollicutes) and Verrucomicrobia in the C-(IV) branch and Acidobacteria, Chlorobi, Chloroflexi,
Deinococcus-Thermus, Fusobacteria, and Thermotogae in the C+ branch are all monophyletic with high support values in the tree. A few other groups, most notably the Proteobacteria and the Firmicutes, are monophyletically supported at the class level. Proteobacteria classes Beta/Gamma-, Delta-, and Epsilon-proteobacteria independently meet at the root of the C+ branch, while Alphaproteobacteria originates in the C-(IV) branch. In the Firmicutes, Clostridia (non-paralogs) branches from C+ while Bacilli branches from C-(IV). The remaining bacteria phyla, Actinobacteria, Cyanobacteria, Planctomycetes, and Spirochaetes, have more convoluted branching patterns, with members branching in either the C+ or C-(IV) group with little regard for classical phylogeny. Disagreement with the classical bacterial phylogeny is an indication that a process more complex than standard vertical inheritance occurred with r-protein S4 during bacterial evolution.

3.3 Horizontal Gene Transfer of S4

3.3.1 Identification of paralogous and duplicated S4 genes

A key element that led to the identification of HGT and gene duplication with DGL in previous bacterial r-proteins was the analysis of genomes containing multiple copies of the r-protein genes [108]. Among the 660 bacterial genomes in our study, 26 organisms from the groups Clostridia, Betaproteobacteria, Deltaproteobacteria, Gammaproteobacteria, Actinobacteria, Spirochaetes and Planctomycetes possess more than one copy of the S4 gene (marked with a dagger symbol on the tree in Figure 3.3). Most have two copies, as shown in Table 3.1, and normally one copy is a C+ form and the other a C-(IV) or C-(V). Two Clostridia, A. metalliredigens and C. acetobutylicum, have three copies of the S4 gene, both have one C+ and two C-(V) variants. The genomes of Methylobacillus flagellatus, Psychromonas ingrahamii, and Leptospira borgpetersenii do not match the above pattern in that both genes are of the same type. However, in each of these three cases the sequence identities of the two copies are extremely high, 100%, 99%, and 100%, respectively. These cases are undoubtedly recent gene duplication events. In fact, P. ingrahamii and L. borgpetersenii have duplicated a large segment of their conserved operon cluster. M. flagellatus is known to have a large 140 kbp repeat in its genome [123], this repeated region contains the S4 gene.

Usually in a case of two divergent copies of a gene in a genome, one copy is the original and the other a paralog, either from an ancient gene duplication or from an HGT event. Without experimental evidence of activity, determining which is the active gene and which the paralog can often be problematic. In the case of r-protein S4, however, the genome content can provide evidence to make a determination: many ribosomal protein genes in bacteria are known to be located in conserved gene clusters. The gene for S4 is usually located in a cluster along with the genes for ribosomal proteins S13, S11, and L17 and the gene for the RNA polymerase alpha subunit, which together are known as the \( \alpha \)-operon because they are co-regulated in E. coli. If two copies of the S4 gene are present in a genome with one copy inside the \( \alpha \)-operon and the other outside it, we assume the copy inside the \( \alpha \)-operon is the original form and the other the paralog. In every genome containing two divergent S4 genes, the C+ form is located in the \( \alpha \)-operon and the C-(IV) or C-(V) form outside. Using the above criteria, we conclude that the C+ form is the original S4 sequence and the C-form the paralog in these genomes. It then remains to determine the origin of the paralogous C- sequences,
Table 3.1: Genomes containing multiple copies of the S4 gene

<table>
<thead>
<tr>
<th>Organism Name</th>
<th>Taxon</th>
<th># copies</th>
<th>PID(%)</th>
<th>Classification</th>
<th>Other Zn-ribbon duplicates&lt;sup&gt;a&lt;/sup&gt;</th>
</tr>
</thead>
<tbody>
<tr>
<td><em>Frankia</em> sp. EAN1pec</td>
<td>Actinobacteria</td>
<td>2</td>
<td>40.57</td>
<td>C+&lt;sup&gt;b&lt;/sup&gt;; C-(IV)</td>
<td></td>
</tr>
<tr>
<td><em>Salinispora arenicola</em> CNS-205</td>
<td></td>
<td>2</td>
<td>36.79</td>
<td>C+; C-(IV)</td>
<td>S14, L33, L28&lt;sup&gt;c&lt;/sup&gt;, L31, L32</td>
</tr>
<tr>
<td><em>Salinispora tropica</em> CNB-440</td>
<td></td>
<td>2</td>
<td>36.32</td>
<td>C+; C-(IV)</td>
<td>S14, L33, L28&lt;sup&gt;c&lt;/sup&gt;, L31, L32</td>
</tr>
<tr>
<td><em>Methyllobacillus flagellatus</em> KT</td>
<td>β-proteobacteria</td>
<td>2</td>
<td>100.0</td>
<td>both C-(IV)</td>
<td>L36</td>
</tr>
<tr>
<td><em>Nitrosomonas europaea</em> ATCC 19718</td>
<td></td>
<td>2</td>
<td>36.62</td>
<td>C+; C-(IV)</td>
<td></td>
</tr>
<tr>
<td><em>Psychromonas ingrahamii</em> 37</td>
<td>γ-proteobacteria</td>
<td>2</td>
<td>99.51</td>
<td>both C-(I)</td>
<td>L36&lt;sup&gt;d&lt;/sup&gt;</td>
</tr>
<tr>
<td><em>Bdellovibrio bacteriovorus</em> HD100</td>
<td>δ-proteobacteria</td>
<td>2</td>
<td>39.35</td>
<td>C+; C-(IV)</td>
<td></td>
</tr>
<tr>
<td><em>Myxococcus xanthus</em> DK 1622</td>
<td></td>
<td>2</td>
<td>36.41</td>
<td>C+; C-(IV)</td>
<td>S14, L33, L28&lt;sup&gt;c&lt;/sup&gt;</td>
</tr>
<tr>
<td><em>Sorangium cellulosum</em> ‘So ce 56’</td>
<td></td>
<td>2</td>
<td>42.45</td>
<td>C+; C-(IV)</td>
<td></td>
</tr>
<tr>
<td><em>Alkaliphilus metalliredigens</em> QYMF</td>
<td>Clostridia</td>
<td>3</td>
<td>38.53&lt;sup&gt;e&lt;/sup&gt;</td>
<td>C+; C-(V)&lt;sup&gt;f&lt;/sup&gt;</td>
<td></td>
</tr>
<tr>
<td><em>Alkaliphilus oremlandii</em> OhILAs</td>
<td>Clostridia</td>
<td>2</td>
<td>40.00</td>
<td>C+; C-(V)</td>
<td></td>
</tr>
<tr>
<td><em>Clostridium acetobutylicum</em> ATCC 824</td>
<td></td>
<td>3</td>
<td>38.61&lt;sup&gt;e&lt;/sup&gt;</td>
<td>C+; C-(V)&lt;sup&gt;f&lt;/sup&gt;</td>
<td></td>
</tr>
<tr>
<td><em>Clostridium botulinum</em> A str. ATCC 3502</td>
<td></td>
<td>2</td>
<td>38.79</td>
<td>C+; C-(V)</td>
<td></td>
</tr>
<tr>
<td><em>Clostridium botulinum</em> A3 str. Loch Maree</td>
<td></td>
<td>2</td>
<td>37.38</td>
<td>C+; C-(V)</td>
<td></td>
</tr>
<tr>
<td><em>Clostridium botulinum</em> A str. ATCC 19397</td>
<td></td>
<td>2</td>
<td>38.79</td>
<td>C+; C-(V)</td>
<td></td>
</tr>
<tr>
<td><em>Clostridium botulinum</em> A str. Hall</td>
<td></td>
<td>2</td>
<td>38.79</td>
<td>C+; C-(V)</td>
<td></td>
</tr>
<tr>
<td><em>Clostridium botulinum</em> B1 str. Okra</td>
<td></td>
<td>2</td>
<td>38.79</td>
<td>C+; C-(V)</td>
<td></td>
</tr>
<tr>
<td><em>Clostridium botulinum</em> F str. Langeland</td>
<td></td>
<td>2</td>
<td>39.25</td>
<td>C+; C-(V)</td>
<td></td>
</tr>
<tr>
<td><em>Clostridium kluyveri</em> DSM 555</td>
<td></td>
<td>2</td>
<td>39.25</td>
<td>C+; C-(V)</td>
<td></td>
</tr>
<tr>
<td><em>Clostridium novyi</em> NT</td>
<td></td>
<td>2</td>
<td>40.38</td>
<td>C+; C-(V)</td>
<td></td>
</tr>
<tr>
<td><em>Clostridium perfringens</em> str. 13</td>
<td></td>
<td>2</td>
<td>38.79</td>
<td>C+; C-(V)</td>
<td></td>
</tr>
<tr>
<td><em>Clostridium perfringens</em> ATCC 13124</td>
<td></td>
<td>2</td>
<td>38.79</td>
<td>C+; C-(V)</td>
<td></td>
</tr>
<tr>
<td><em>Clostridium perfringens</em> SM101</td>
<td></td>
<td>2</td>
<td>38.79</td>
<td>C+; C-(V)</td>
<td></td>
</tr>
<tr>
<td><em>Leptospira borgpetersenii</em> serovar</td>
<td>Spirochaetes</td>
<td>2</td>
<td>100.0</td>
<td>both C-(III)</td>
<td>S14, L36&lt;sup&gt;g&lt;/sup&gt;</td>
</tr>
<tr>
<td>Hardjo-bovis L550</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td><em>Gemmata obscuriglobus</em> UQM 2246</td>
<td>Planctomycetes</td>
<td>2</td>
<td>40.95</td>
<td>C+; C-(IV)</td>
<td></td>
</tr>
</tbody>
</table>

<sup>a</sup>Observations based on annotation.

<sup>b</sup>Classification written in bold indicates that it is inside the α-operon.

<sup>c</sup>The duplicates of the zinc-binding r-proteins S14, L28, L33 sit together in the genome.

<sup>d</sup>Resulting from the duplication of the entire s10-spc and α-operon.

<sup>e</sup>Average value of the percent identities of the two pairs of in-operon and out-of-operon copies.

<sup>f</sup>*Alkaliphilus metalliredigens* and *Clostridium acetobutylicum* have 2 copies of the C-(V) form.

<sup>g</sup>Resulting from the duplication of the entire s10-spc and α-operon.
Figure 3.4: Expansion of Clostridia branches of the consensus S4 phylogenetic tree. The C+ and C-(V) Clostridia branches are highlighted white and yellow, respectively. Sequences from organisms with multiple S4 genes are in bold underline. Sequences from the two genomes with three S4 genes are additionally marked with an (*) asterisk. (Figure courtesy of E Roberts)

Whether by HGT or gene duplication.

One clear-cut case of HGT appears to have occurred in the Proteobacteria. One Beta- and three Deltaproteobacteria have S4 paralogs that group within the C-(IV) branch of the phylogenetic tree in Figure 3.3 and C+ genes in the α-operon. Since the vast majority of Beta- and Deltaproteobacteria have only a single C+ copy of S4, we consider it unlikely that this pattern resulted from an ancient gene duplication that was lost in all Beta- and Deltaproteobacteria except these four organisms. Given the high support values near the branch with Cyanobacteria, we find it more likely that these organisms obtained the gene through a horizontal transfer from Cyanobacteria, although the S4 sequence is not similar enough to any available sequences for the specific source organism to be determined. Betaproteobacteria species M. flagellatus, which possesses only two copies of C-(IV) outside of the α-operon, also groups nearby in the tree and also likely received its C-(IV) S4 gene from Cyanobacteria before its large-scale genome duplication occurred. It must have lost its original C+ gene subsequent to the HGT, as it is no longer present in the α-operon.

Another example of probable HGT, albeit with a more complex pattern, is seen in the Clostridia. All Clostridia except one, Finegoldia magna, contain the gene for the C+ form of S4 in the α-operon. Fifteen Clostridia also contain a gene for the C-(V) form of the S4 sequence. Figure 3.4 shows an expansion of
Figure 3.5: Expansion of Actinobacteria branches of the consensus S4 phylogenetic tree. The C+ and C-(IV) Actinobacteria branches of S4 are highlighted white and blue, respectively. Sequences from organisms with multiple S4 genes are in bold underline. (Figure courtesy of E Roberts)

these two Clostridia branches from the consensus phylogenetic tree. In the C+ branch, all of the organisms with multiple copies of S4 are descended from a single branch. If we assume that the C+ form represents the vertical phylogeny in this group, then the pattern is consistent with either a single HGT event in the ancestor of the Clostridium and Alkaliphilus genera with later differential loss in a few branches; or with three later HGT events, one for Alkaliphilus, a second for Clostridium perfringens, and a third for the branch containing the organisms Clostridium botulinum/kluyveri/acetobutylicum/novyi. However, two organisms possess three copies of the S4 gene, and the percent identities between the two C-(V) copies are 60.3% and 54.3% for A. metalliredigens and C. acetobutylicum, respectively. So there may have been an additional recent HGT of the transferred C-(V) gene. The support values are too low to allow a determination the source of this recent HGT, but additional genomes of related organisms could shed light on the history of the C-(V) form of S4.

In the two remaining lineages with genomes containing multiple S4 genes, multiple occurrences are relatively rare. Of the four Planctomycetes genomes available, one contains both the C+ gene in the α-operon and C-(IV) out of it while the other three contain only C-(IV) out of the operon. The low number of available Planctomycetes genomes sequenced makes it impossible to reconcile the origin of the paralog using a parsimony argument. In the Actinobacteria, the three genomes with two copies of the S4 gene can be accounted for by two recent HGT events (see Figure 3.5), one in the Salinispora genus and the other in the species Frankia sp. EAN1pec. Both of these paralogous genes appear to have originated in a Streptomyces source. The two Salinispora species also have acquired paralogs of five other zinc-binding r-proteins.
3.3.2 Regulation of C- S4 paralogs

To fully understand the evolutionary pressure giving rise to paralogous genes, it is helpful to know their regulation mechanism, especially when the two copies have similar functions. For the previously known zinc-binding ribosomal proteins, experiments on specific bacteria have shown that the paralogs can be turned on and off in response to zinc conditions. In *B. subtilis* (S14, L31) [111], *M. tuberculosis* (S14, S18, L28, L33) [113], and *S. coelicolor* (S14, L28, L31, L32, L33A, L33B, L36) [112], the C- paralogs were found to be expressed only under low zinc conditions. Their regulation was controlled by the zinc uptake regulator (Zur) transcription factor, except for L33B and L36 in *S. coelicolor*, the regulation of which was controlled by a sigma factor (σ^R^).

The established Zur binding-sites from *Actinobacteria*, *Bacilli* and *Proteobacteria* are AT-rich palindromes found upstream from the genes being regulated [109; 112; 113]. Using a profile of Zur binding motifs from these bacterial groups, we searched the 26 genomes containing multiple copies of the S4 gene (from *Actinobacteria*, *Clostridia* and *Beta/Deltaproteobacteria*) for candidate Zur binding sites (see Methods). We were able to identify Zur binding sites upstream of the gene cluster of r-proteins L33, S14 and L28 and upstream of the paralogous genes of both L31 and L32 in *Salinispora arenicola* and *Salinispora tropica*, but no binding sites were found near the paralogous S4 genes. Neither were Zur binding sites found near ribosomal protein paralogs in the remaining genomes. Unfortunately, a Zur binding motif has not yet been reported for *Clostridia*, which comprises most of the genomes with paralogous copies of the S4 gene. Therefore, we can not exclude the possibility that the paralogous copies of S4 in *Clostridia* are regulated by Zur binding to a motif different from any in our profile.

However, according to gene expression data from two separate genomic-scale gene expression experiments in *C. acetobutylicum* [124] and *Clostridium novyi* [125], the paralogous C-(V) genes are not expressed under normal growth conditions but are up-regulated during sporulation. This leaves open the possibility that the C-(V) genes are related to some aspect of ribosomal function during sporulation and not used to regulate the zinc environment in *Clostridia*. If the C-(V) S4 proteins are indeed incorporated into ribosomes in clostridial spores, it would be interesting to examine any changes to these ribosomes, such as altered structure or changes in the assembly process.

3.3.3 Comparison of genome content near S4

Having used genome context in the analysis of several cases of horizontal transfer, we next examined the genome regions near S4 and the α-operon in the genomes of the bacteria without multiple copies looking for conserved patterns. Overall, the organization of the α-operon and nearby genes is highly conserved across a large number of bacterial groups. Many of the genomes have the conserved consensus gene cluster shown in Figure 3.6A, containing genes for initiation factor A (infA), L36, S13, S11, S4, RNA polymerase subunit A (rpoA) and L17. Variations are mainly seen in *Gammaproteobacteria* and *Magnetococcus*, which do not have infA near the cluster. Intriguingly, genes for both the C+ and C-(IV) forms of S4 can be found in the α-operon (green background in Figure 3.3). In fact, eight phyla have the gene for the C-(IV) form located in the α-operon, including all three of the phyla containing closely branching C+ and C-(IV)
Figure 3.6: Genomic content near S4 and the α-operon. Shown are a representative sample of the genomes. A) Consensus genome context of the S4 gene in the α-operon across most bacterial groups. B & C) Context of the α-operon and S4, respectively, in bacterial groups where an S4 gene is located outside the α-operon. Phyla abbreviations are given in parentheses. The lengths of the genes are to scale and gene are color-coded according to COG (clusters of orthologous groups) functional categories.

forms (Actinobacteria, Chloroflexi and Deinococcus-Thermus). Five other phyla, Aquificae, Bacteroidetes, Chlorobi, Fusobacteria, and Verrucomicrobia, contain only the C-(IV) form in the operon.

The remaining bacterial genomes, still covering a diverse set of bacteria, contain only an S4 gene of the C-(IV) form that is not located within the α-operon (red background in Figure 3.3). Figures 3.6B and C shows examples of the genomic context of the α-operon and the S4 gene, respectively, in these genomes. The organization of the genes remaining in the α-operon is unperturbed, but the context around the S4 gene is variable. Conservation of organization near the S4 gene can only be seen at the level of order or family; no correlations with the organization of any other genes could be detected at higher levels of taxonomy. When not located in the α-operon, the gene for S4 appears to be quite mobile.
3.4 Interpretation of the Evolutionary Patterns of S4

3.4.1 The ancestral form of S4 in the bacteria

Given the widespread occurrence of C-(IV) genes within the α-operon (see Figure 3.7), one must question the hypothesis that the C+ zinc-binding form of S4 is ancestral in the bacteria. If the C-(IV) form were a result of a single ancient gene duplication of a C+ gene, one would have expected to find nearly all of C-(IV) genes located outside of the α-operon. Instead, five classical bacteria phyla contain exclusively the C-(IV) gene in the α-operon. Moreover, three bacteria phyla contain monophyletic branches of both the C+ and C-(IV) genes, each organized in the typical α-operon style. Although it is known that horizontally transferred genes can replace their native copies in the genome, so called in situ gene displacement [126], such occurrences are still thought of as exceptions rather than the rule. The number of in situ displacements required to achieve the current distribution of C-(IV) genes in the α-operon would require replacement events of a much higher frequency or different character than that previously reported.

We propose instead that neither C+ nor C-(IV) is the sole ancestral form of S4 and interpret the data as implying the presence of both forms during the time when the bacterial lineages were diverging. The developing bacterial lineages would have sampled S4 genes from the bacterial pool according to some unknown criteria, perhaps related to the local environment (e.g., thermophilic organisms acquiring the zinc-binding form for added stability). While this sampling would have been functionally equivalent to HGT with in situ gene displacement, in that the gene order would be maintained, it would not have necessarily been mechanistically related to the process by which HGT occurs today.

Additional support for the existence of innovation sharing within gene pools comes from signatures in the S4 protein that were reported by Roberts et al. [127] to distinguish the bacterial and archaeal/eukaryal lineages. S4 proteins from both archaea and bacteria possess the RNA binding C-terminal domain, but have an N-terminal architecture distinct to each domain of life. Furthermore, the archaeal version of the α-operon is organized with S4 preceding S11 (S13-S4-S11), as opposed to S11 proceeding S4 as in bacteria (S13-S11-S4). Clearly, large-scale evolutionary changes occurred in S4 after (or at) the Bacteria and Archaea divergence, and yet the signatures are unvarying within each domain. Excluding the possibility that all extant bacteria can trace their vertical ancestry to a single individual cell and all extant archaea to another single cell, the respective organism pools at the time must have been able to efficiently share genes in an in situ manner that allowed the homogenization of the bacterial pool. This is the same evolutionary process required to support both a C+ and a C-(IV) form of the S4 gene in the bacterial pool.
3.4.2 Origin of S4 outside the α-operon

If, as suggested above, a bacterial pool allowed both the C+ and C-(IV) forms of the S4 gene to be brought into the genome in situ as needed, the question arises as to the origin of the C-(IV) gene outside of the α-operon in genomes where it is the sole copy. We propose that this organization is the result of HGT of the C-(IV) gene into C+ genomes after the phyla had diverged from the bacterial gene pool and the in situ evolutionary dynamic had slowed. Loss of the original C+ gene would have then allowed a reduction in zinc use without perturbation to growth of the organisms.

Figure 3.8 depicts the four possible evolutionary paths (labeled A–D) starting from either a C+ or C-(IV) gene inside the α-operon and ending with a single C-(IV) type gene outside the α-operon. Path A involves
Figure 3.8: Four possible evolutionary pathways for S4 (A–D) that result in the observed pattern of a single C-(IV) gene outside the α-operon.

A gene duplication of a C+ type, mutation of the C+ type into a C-(IV) type, and finally loss of the original C+ gene. This path is ruled out for two reasons: first, no duplications of C+ S4 genes were observed in any of the 660 genomes studied, and second, the path depends on an unlikely set of mutation events. The C-(IV) genes outside the α-operon are indistinguishable in sequence from the C-(IV) genes inside, including the loss of a characteristic seven residue indel. The probability of an independent mutational deletion of seven residues from a C+ gene leading to the exact same indel pattern as in the pre-existing C-(IV) gene is low. Additionally, there are other sequences signatures, such as residue 15 (in the RRXG motif) being glutamic acid in C+ and leucine/phenylalanine in C-(IV) and residue 21 being leucine in C+ and glycine/proline in C-(IV), that support a common origin for all of the C-(IV) sequences.

Path B specifies HGT of a C-(IV) gene into a genome containing the C+ type in the α-operon followed by loss of the C+ gene. Evidence supporting path B comes from the pattern of S4 HGT events presented in Results. The evolutionary history of S4 contains several relatively recent horizontal transfers, as supported by our analysis of genomes containing multiple copies of the S4 gene. In each of these cases, a C-(IV) gene was transferred into a genome with the C+ form of the S4 gene in the α-operon. Additionally, in two instances there was loss of the original C+ gene following the HGT of a C-(IV) gene, the clostridium F. magna and the betaproteobacterium M. flagellatus, exactly as prescribed in path B.

Both of the remaining paths, C and D, start with a C-(IV) gene and involve later acquisition of an additional C-(IV) gene either through duplication of the original or HGT, respectively. Our analysis found neither duplications nor horizontal transfers in any genome with the C-(IV) gene in the α-operon, although the sample size of known HGT events is low. Furthermore, there is phylogenetic evidence that some of the groups now containing only a C-(IV) gene outside of the α-operon are descended from lineages originally containing the C+ gene, which would preclude paths C and D. For example, all Alphaproteobacteria contain only the C-(IV) gene outside of the α-operon, while all other Proteobacteria have the C+ form (or recent variations thereof) in the α-operon. Even the genome of the unclassified proteobacterium Magnetococcus
Table 3.2: Taxonomic distributions of the C+/− ribosomal proteins

<table>
<thead>
<tr>
<th></th>
<th>L32</th>
<th>L36</th>
<th>L31</th>
<th>S14</th>
<th>L33</th>
<th>S18</th>
<th>L28</th>
<th>S4</th>
</tr>
</thead>
<tbody>
<tr>
<td>Verrucomicrobia</td>
<td>-</td>
<td>-</td>
<td>-</td>
<td>-</td>
<td>-</td>
<td>-</td>
<td>-</td>
<td>-</td>
</tr>
<tr>
<td>α-proteobacteria</td>
<td>+,-</td>
<td>-</td>
<td>-</td>
<td>-</td>
<td>-</td>
<td>-</td>
<td>-</td>
<td>-</td>
</tr>
<tr>
<td>Bacteroidetes</td>
<td>+,-</td>
<td>nd</td>
<td>-</td>
<td>-</td>
<td>-</td>
<td>-</td>
<td>-</td>
<td>-</td>
</tr>
<tr>
<td>Chlamydiae</td>
<td>+</td>
<td>-</td>
<td>-</td>
<td>-</td>
<td>-</td>
<td>-</td>
<td>-</td>
<td>-</td>
</tr>
<tr>
<td>Fusobacteria</td>
<td>+</td>
<td>nd</td>
<td>-</td>
<td>-</td>
<td>-</td>
<td>-</td>
<td>-</td>
<td>-</td>
</tr>
<tr>
<td>β-proteobacteria</td>
<td>-</td>
<td>+,+/-</td>
<td>+,-,+/-</td>
<td>-</td>
<td>-</td>
<td>-</td>
<td>-</td>
<td>+,-</td>
</tr>
<tr>
<td>γ-proteobacteria</td>
<td>-</td>
<td>+,-,+/-</td>
<td>+,-,+/-</td>
<td>-</td>
<td>-</td>
<td>-</td>
<td>-</td>
<td>-</td>
</tr>
<tr>
<td>Chlorobi</td>
<td>+</td>
<td>+</td>
<td>+</td>
<td>-</td>
<td>-</td>
<td>-</td>
<td>-</td>
<td>-</td>
</tr>
<tr>
<td>Planctomycetes</td>
<td>nd</td>
<td>-</td>
<td>+</td>
<td>+</td>
<td>-</td>
<td>-</td>
<td>-</td>
<td>-</td>
</tr>
<tr>
<td>Deinococcus</td>
<td>+</td>
<td>+</td>
<td>-</td>
<td>+,-</td>
<td>-</td>
<td>-</td>
<td>-</td>
<td>+,-</td>
</tr>
<tr>
<td>Cyanobacteria</td>
<td>-</td>
<td>+,-</td>
<td>+,-</td>
<td>+,-</td>
<td>+,-</td>
<td>-</td>
<td>-</td>
<td>+,-</td>
</tr>
<tr>
<td>Bacilli</td>
<td>+,-</td>
<td>+</td>
<td>+,-</td>
<td>+,+/-</td>
<td>+,-,+/-</td>
<td>-</td>
<td>-</td>
<td>+,-</td>
</tr>
<tr>
<td>Tenericutes (Mollicutes)</td>
<td>+,-</td>
<td>+,-</td>
<td>+</td>
<td>+,-</td>
<td>+,+/-</td>
<td>+,-,+/-</td>
<td>-</td>
<td>+,-</td>
</tr>
<tr>
<td>s-proteobacteria</td>
<td>-</td>
<td>+</td>
<td>+</td>
<td>+,+/-</td>
<td>+,+/-</td>
<td>+,+/-</td>
<td>+,-,+/-</td>
<td>+,-</td>
</tr>
<tr>
<td>Actinobacteria</td>
<td>+,-</td>
<td>+</td>
<td>+,-</td>
<td>+,-</td>
<td>+,-</td>
<td>+,-</td>
<td>-</td>
<td>-</td>
</tr>
<tr>
<td>Chloroflexi</td>
<td>+</td>
<td>+</td>
<td>+,-</td>
<td>+,-</td>
<td>+,-</td>
<td>+,-</td>
<td>+,-</td>
<td>-</td>
</tr>
<tr>
<td>Spirochaetes</td>
<td>+</td>
<td>+</td>
<td>+,-</td>
<td>+</td>
<td>-</td>
<td>+,-</td>
<td>+,-</td>
<td>+</td>
</tr>
<tr>
<td>Thermus</td>
<td>+</td>
<td>+</td>
<td>+</td>
<td>+</td>
<td>-</td>
<td>-</td>
<td>-</td>
<td>+</td>
</tr>
<tr>
<td>Magnetococcus</td>
<td>+</td>
<td>+</td>
<td>+</td>
<td>+</td>
<td>+</td>
<td>+</td>
<td>-</td>
<td>-</td>
</tr>
<tr>
<td>Aquificae</td>
<td>+</td>
<td>+</td>
<td>+</td>
<td>+</td>
<td>+</td>
<td>+</td>
<td>+</td>
<td>-</td>
</tr>
<tr>
<td>δ-proteobacteria</td>
<td>+</td>
<td>+</td>
<td>+</td>
<td>+,-</td>
<td>+</td>
<td>+</td>
<td>+</td>
<td>+</td>
</tr>
<tr>
<td>Clostridia</td>
<td>+</td>
<td>+</td>
<td>+</td>
<td>+</td>
<td>+</td>
<td>+</td>
<td>+</td>
<td>+,-</td>
</tr>
<tr>
<td>Acidobacteria</td>
<td>+</td>
<td>+</td>
<td>+</td>
<td>+</td>
<td>+</td>
<td>+</td>
<td>+</td>
<td>+</td>
</tr>
<tr>
<td>Thermogae</td>
<td>+</td>
<td>+</td>
<td>+</td>
<td>+</td>
<td>+</td>
<td>+</td>
<td>+</td>
<td>+</td>
</tr>
</tbody>
</table>

+,- indicates each form comprises at least 10% of the group, but occurs only once in each genome.
+/− indicates at least 10% of the genomes in the group contain both forms. This cut-off removes all minor cases of HGT.
nd indicates that the r-protein was not detected.

*MC-1*, which is phylogenetically closest to the *Alphaproteobacteria* [128; 129], contains the gene for the C+ form of S4 in the α- operon. Thus, the *Proteobacteria* phylum likely contained the C+ gene originally.

The above arguments provide support for our hypothesis that C+ was the original form of the S4 gene in the branch of the tree containing C-(IV) outside the α-operon (red background in Figure 3.3) and that these branches received the C-(IV) gene through HGT. The original source of the C-(IV) S4 gene must have been one of the phyla containing C-(IV) natively, i.e., one with the C-(IV) gene in the α-operon, but once the S4 gene made the transition from an operon gene to a standalone gene it may have become more readily transferable. Later HGT events may therefore have originated from organisms having already received prior transfers. From the phylogenetic tree in Figure 3.3 the best candidate phyla for the original source are *Aquificae, Bacteroidetes* and *Verrucomicrobia*, but the low support values near the radiating points in the tree leave a great deal of ambiguity as to the exact source. Given the strict presence of the C-(IV) form and absence of an S4 gene in the α-operon in the groups, it appears that both the HGT events and native gene losses are likely ancient.
3.4.3 Evolutionary pressure and the loss of zinc binding in ribosomal proteins

Insight into an evolutionary process comes from not only describing the mechanism of change, but also the pressures behind the change. As discussed earlier, seven other r-proteins have been reported to bind zinc and to have evolutionary histories disrupted in a similar pattern to what we have reported for S4. Table 3.2 shows the occurrence of C+, C-, or both C+ and C- genes of these r-proteins in the major bacterial groups. C+ was reported as being the ancestral form of these r-proteins [108] and, if that is indeed the case, it is clear from the distribution that large groups have developed either the ability to do without zinc for specific r-proteins or to switch to C- paralogs under low zinc conditions. Specifically, we see characteristic divisions below the phyla level, e.g. *Alphaproteobacteria* have replaced the C+ genes for almost all zinc-binding r-proteins with C- genes, *Epsilonproteobacteria* has exclusively C- forms of three, and *Magnetococcus sp. MC-1* encodes the C- gene for only one.

These observations point to the conclusion that some bacteria evolved to use C- variants of the zinc-binding ribosomal proteins (including S4) to regulate the zinc economy of the cell. Whether this lower zinc usage was a response to a change in the zinc conditions in the environment or whether some other change in the environment (such as lower temperature) caused the decreased need for zinc is still unclear.

3.4.4 Conclusions

The cellular information processing system is generally believed to be much less subject to the influences of HGT than other genetic systems. While recent metagenomic studies have not reported any reliable HGT events for the ribosome among the three domains of life, examples of disagreement with the UPT among the bacterial versions of seven zinc-binding r-proteins S14, S18, L28, L31, L32, L33 and L36 have been well documented [106; 108]. According to our study of 660 bacterial genomes, the bacterial version of the universal r-protein S4, shares similarities with these seven proteins, namely they all have two different versions of the sequence, zinc-binding (C+) and non-zinc-binding (C-), and their evolutionary histories all show patterns of disagreement with the standard UPT.

The evolutionary history of r-protein S4 reconstructed here shows that S4 was subject to horizontal transfer throughout the history of the bacterial lineages. Recent HGT of the standard character was observed along with other less well-defined evolutionary dynamics of ancient origin. We propose wide-spread sampling of ancestral C+ and C- forms of the S4 gene from a bacterial gene pool as a possible explanation, but definitive proof of such an ancient event cannot be easily obtained. The present study was only possible given the large number of available bacterial genomes, and perhaps additional genomes of other diverse bacterial lineages would provide additional evidence for or against this proposition. Experiments detailing the purpose and regulation of paralogous S4 genes in *Clostridia* also may shed light on the differences between the C+ and C- forms.

In more practical terms, it should now be understood that even “core” proteins can have a more complex evolutionary history than can be explained by vertical inheritance. One recent study attempting to reconstruct an organismal tree of life included S4 in a concatenated gene tree [130]. Although the authors did attempt to remove genes subject to HGT, none was detected in the case of S4. It is clear from the present
study that doing so is not always a simple proposition. Accurate evolutionary relationships for S4 were only uncovered with extensive coverage of the bacterial tree along with heavy use of genome content. Others have shown that concatenated genes trees may lack resolution [131], and this may be a direct result of mixing genes with different complex relationships, like the one reported here for S4.

3.5 Methods

3.5.1 S4 sequence acquisition and alignment

The analysis was based all of the complete bacterial genomes available at the time in NCBI GenBank. Additionally, in order to provide further data for a few poorly represented phyla, draft genomes from three Planctomycetes, five Fusobacteria, and six Verrucomicrobia were obtained from the Joint Genome Institute, as identified through the Genome OnLine Database (GOLD) [132].

To find S4 sequences and paralogs in the genomes, a non-redundant sequence profile was constructed as described in Sethi et al. [133] starting with annotated S4 sequences from the Swiss-Prot database [134]. This profile was used to do a BLAST search [135] on each genome with a cutoff of $10^{-7}$. Fragments containing only the C-terminal RNA binding domain were removed. Sequence were classified as C+ or C- by comparison to annotated sequences and then all sequences of each type were aligned using the ClustalW [136] multiple alignment function. The C+ and C- multiple alignments were combined using the ClustalW profile alignment function and the resulting alignment was hand edited to correct poorly aligned regions. All operations were performed within the MultiSeq [137] bioinformatics analysis environment.

3.5.2 Phylogenetic reconstructions

Maximum likelihood (ML) trees were reconstructed using RAxML version 7.0.4 [138]. A value of 10 was used for the maximum initial rearrangement distance (-i 10) and a value of 25 for the number of rate categories (-c 25). The tree for Proteobacteria was calculated using the JTT amino acid model [139] (-m PROTMIXJTT) and the tree for Bacteria using the WAG model [140] (-m PROTMIXWAG), as these models gave the best likelihood scores for a given maximum-parsimony tree of the respective alignments. A total of 1000 likelihood searches were performed for each alignment starting from unique, random maximum-parsimony trees (-f d -# 1000). The tree with the highest likelihood score was taken to be the ML tree. A consensus tree was constructed from the ML tree by removing bipartitions found in fewer than 50% of the other most likely trees. Following, 5000 non-parametric bootstrap runs were performed starting with the topology of the ML tree (-b -t ml.tre -# 5000) to determine support values for the bipartitions. Support values were mapped onto their corresponding branches in the consensus tree. Sequences from a few genomes (Candidatus Carsonella ruddii PV, Sorangium cellulosum ‘So ce 56’, Symbiobacterium thermophilum IAM 14863, Petrotoga mobilis SJ95, Rubrobacter xylanophilus DSM 9941, Myxococcus xanthus DK 1622, Clostridium phytofermentans ISDg) were highly mobile during ML reconstruction (likely long-branch artifacts) and so were excluded from the reconstruction and added afterwards using stepwise maximum-parsimony addition (-f p -t ml.tre).
3.5.3 Zinc regulatory motifs

Zinc regulation protein binding motifs, which are AT rich palindromes on the intergenic region of DNA strand were searched using MEME/MAST [141]. MEME was used to make a position specific substitution matrix (PSSM) based on input palindromes. The input profile of Actinobacteria was taken from experimentally determined Zur binding sites in Mycobacterium tuberculosis and Streptomyces coelicolor [112; 113] and the profiles for Bacillus group and γ-proteobacteria were taken from [109]. Then the resulting matrices were used as input of MAST to search for other binding sites in the whole genomes. Only those genomes that have paralogs of S4 genes were subjected to this analysis.
Chapter 4

Functional Role of Ribosomal Signatures

4.1 Background

Ribosomal signatures, idiosyncrasies in the ribosomal RNA (rRNA) and protein (r-protein) sequences, are characteristic of the individual domains of life: Bacteria, Archaea, and Eucarya. First identified by Carl Woese more than 20 years ago [37; 142], the positions of signatures in the rRNA of the ribosomal small subunit (SSU), were instrumental in constructing the universal phylogenetic tree which is still used today in the classification of biological organisms [143]. As such, these molecular fossils offer insight into the early evolution of the translational apparatus. With the rapid growth of genomic and structural data, the definition of signatures has been extended to include structure motifs in the rRNA and r-proteins and arrangements of genomic content that are unique to one domain of life. In a study that includes over 90,000 16S and 23S sequences, Roberts et al. [39] demonstrated that the sequence and structure signatures of rRNA account for 50% of the phylogenetic separation between Bacteria and Archaea. Correlations between the rRNA and r-protein signatures show that the rRNA signatures coevolved with both domain-specific r-proteins and inserts in universal r-proteins. The largest continuous bacterial rRNA structure signature in the SSU with such a coevolutionary protein partner is found in helix 16 (h16) of the 16S rRNA 5′ domain, which is held together by the universal r-protein S4. As seen in Figure 4.1, the N-terminal bacterial signature of S4 (S4N) interacts predominantly with the rRNA structure signature h16. S4 also makes a few contacts with h18, but with the exception of base pair C511-G540, none of these contacts are rRNA structure signatures.

The Nomura map developed in the 1970’s [53] showed that the assembly of the SSU in bacteria was dependent on the presence of S4 and five other primary binding proteins that are required to be in place before the remaining r-proteins can be incorporated. S4, S17, and S20 bind directly to the 5′ domain of the 16S rRNA, and the other primary binders, S15/S8 and S7, bind to the central and 3′ domains, respectively. Two of them, S4 and S7, were later identified as the only assembly-initiator proteins based on their noncooperative binding during the onset of assembly [144], and the effects of S7 binding on ribosomal stability were subsequently studied in silico through coarse-grained molecular dynamics (MD) simulations [145].

Figure 4.1: A) cartoon representation and B) crystal structure of the ribosome. Both plots show the five-way junction colored dark blue and S4 colored yellow. C) Five-way junction and S4 as well as the signature regions S4N (red) and h16 (green) on the ribosomal small subunit (SSU). This figure has been rotated by 90° around the horizontal axis with respect to A) and B). D) A blow-up of the 3D structure for the system studied in this paper with the same coloring. In the secondary structure diagram E), bases colored orange and yellow are in contact with the S4N and S4C, respectively. F) S4N and h16 sequences in the wild type *E. coli* ribosome and the modified versions. Residues that have been mutated in this study are highlighted.

More recently, the Nomura dependency map has been extended to include information about the kinetics and folding pathways for the assembly of the 30S ribosome. Using pulse-chase experiment monitored by
quantitative mass spectrometry, Williamson and coworkers determined the binding rates of each r-protein in the SSU and found that the 5′ domain proteins, especially S4, bind more quickly than the proteins in the central or 3′ domain [146; 147], indicating a 5′ to 3′ directionality in the assembly process.

Folding of the 5′ domain of 16S rRNA was studied with time-resolved hydroxyl radical footprinting which established the time dependence for the formation of rRNA tertiary contacts [148]. Without any proteins, the structure signature h16, the binding site of the N-terminal S4 bacterial signature, folds earliest under a wide range of ion concentrations. The “minimal” rRNA binding site for the complete S4 [149] was established from deletion and mutation studies on the RNA. The measured binding free energies show that S4 binds tightly to the five-way junction formed by h3, h4, h16, h17, and h18 and that truncations in the first three helices give the largest variations in binding free energies.

The critical role of S4 in the early assembly of the SSU and the existence of RAM mutations on the S4 [150] suggest that the interactions between the signatures in this region are functionally important for ribosomal assembly and fidelity of protein synthesis in bacteria, but further experiments and simulations are required to characterize these interactions. Both computational and experimental approaches are needed to elucidate the function of protein:RNA systems [151] so a combination of in silico and in vitro methods was employed in this work. This paper presents a detailed study of dynamics of the five-way junction and S4 using all-atom MD simulation with a particular emphasis on the signatures h16 and S4N. Circular dichroism (CD) and fluorescence spectroscopy experiments were performed in conjunction with the computational work, and qualitative agreement between computational and experimental results gives insight into the intrinsic disorder of the signature on S4 and the flexible nature of the interactions between the RNA:protein signatures. Our results are consistent with a fly-casting mechanism in which folding of S4N and partial refolding of the five-way junction are induced by S4N binding to h16 and suggest that this signature region on the ribosome was a domain specific invention in evolution aimed at speeding up the molecular recognition between the rRNA and the early binding r-protein S4.

4.2 Characterization of the Coevolving Signature Interactions in the Crystal Structure

In the crystal structure, the 31-nucleotide RNA helix h16 (nucleotide 406 to 436, E. coli numbering) contains thirteen native base pairs (with a total of 29 hydrogen bonds) and 22 native base stacking pairs (See Methods). The complexity of its secondary structure is seen in the diversity of the base pairs, including five Watson-Crick base pairs (three GC bp, two AU bp), three GU wobble pairs, one sheared GA pair, and one AU Hoogsteen pair. The remaining three base pairs (U420-G423, A411-A430, and A415-G428) are less common in the RNA structure. Two stems are established through these base pairs (labeled “stems” in Figure 4.2A) and stabilized by the base stacking interactions. The two stems of h16 are connected by an internal loop which itself is stabilized by interstrand stacking of three adjacent base pairs, G410-A432, U429-A431 and A430-A411 [152]. Other intrastrand base-stacking interactions, such as G428/A430, help keep the backbone continuity despite the large curvature created by the intrastrand base pair U429-A431 and the nearby interstrand stacking.

Positions of the protein:RNA contacts in the crystal structure are presented in the secondary structure diagram in Figure 4.1E, and the contact map and 3D structure in Figure 4.2. Contact distances in the map
are the shortest distance between any pair of the heavy atoms for each pair of the residues. h16 is located on the periphery of the SSU and has few interactions with the other parts of the RNA except the stem on h18 and the linkage into h17. However, it interacts extensively with the protein S4N. The backbone of h16 fits into the grooves defined on the protein:RNA interaction surface of S4N outlined by positively charged amino acids (shown in blue). The two main binding sites for S4N are on either side of h16 from residue U407 to G413 and from G425 to A430 with slightly more interactions existing with the 3′ side of the internal loop. Phosphate atoms in these two binding sites make hydrogen bonds with either the side-chain amines of lysine/arginine or with the backbone amines from the nearby amino acids. G413, one of the two bases in h16 that has neither base-pairing nor base-stacking interaction within the RNA molecule, has the interaction between the O6 atom on the base with the backbone nitrogen atoms of Lys30 and Ser31. From the S4 side, binding to the RNA is driven mainly by non-specific electrostatic interactions. In the MD simulations the positively charged residue may interact with phosphate atoms from nearby nucleotides, while all four negatively charged residues (shown in red) remain positioned away from the RNA. Contacts between the

Figure 4.2: Contacts between h16 and S4N. A) Contact distances (Å) within the RNA molecule (top-left), within the protein molecule (bottom-right), and between protein and RNA molecules (bottom-left). Residue numbers for the arginines/lysines are colored blue and for the aspartic acids/glutamic acids colored red. Helix position in the protein is indicated by a purple bar beside the residue numbers. B) Secondary structure diagram of h16 adapted from [153] (top) and generated by software jViz [154] (bottom). The bottom one was modified as follows: bonds in orange indicate canonical base pairing interactions, base pairs drawn as dots are noncanonical pairs that have two hydrogen bonds, base pairs drawn as open circles are noncanonical pairs that have only one hydrogen bond, and bonds in yellow indicate base stacking interactions. Nucleotides within 4.5 Å of S4 N-terminus are shown in red. C) Three-dimensional visualization of the h16 backbone sitting on S4N interface (top) and S4N backbone sitting on h16 interface (bottom). Protein is colored by residue type: positively charged (blue), negatively charged (red), polar (green), and hydrophobic (white).
RNA and protein residues 13 to 19 are prevented by the rigidity of the helix, which is oriented at an angle of $\sim 46^\circ$ with respect to the helical axis of h16.

4.3 Intrinsically Disordered Protein Signature: S4 N-terminal Domain

4.3.1 S4N flexibility in simulation

The interactions between S4N and h16 in the crystal structure from E. coli will serve as a reference for the simulations and experiments on the signature complex. The 39-residue S4N consists of a short $\alpha$-helix (Lys7–Glu14) followed by an unstructured loop with two helical turns. In the crystal structure of the T. thermophilus 30S subunit (PDB ID 2J00 [155]) and sequences of some other bacteria, this region of S4 contains a zinc-finger motif in which a zinc ion stabilizes the short helix and the two helical turns. While this motif is partially missing in E. coli, the eight-residue helix is stable without the presence of metal ions. Similar to the other r-proteins, S4 is highly charged. In E. coli, there are nine positively-charged amino-acids (arginine/lysine) and four negatively-charged ones (aspartate/glutamate), comprising one-third of the total number of residues in this segment. In the crystal structure of the complex, these charged residues sit on opposite sides of the S4N — with the positively (negatively) charged ones oriented toward (away from) the S4N:h16 contact interface (Figure 4.2C). In the crystal structure, no salt-bridges are present within the S4N, but salt bridges do exist between Glu14 and Arg55 on the first helix of the C-terminal domain of S4, between Arg2 and the region 5′ to h16, and between Arg13 and h18. In addition, two pairs of residues, Arg13-Glu34 and Asp28-Lys30, are positioned relatively close to each other so that transient short-lived salt-bridges form during the simulations as shown in Table 4.1.

The MD simulations reveal distinct behaviors in the two domains of unbound S4: the stability of its C-terminal domain and the disorder of its N-terminal fragment. As shown in Figure 4.3A, residues in the C-terminal domain have small RMSD values irrespective of the alignment method. The shape of the curve correlates well with the secondary structure of the protein, with the flat regions corresponding to helices and the peaks corresponding to loops. In contrast, the RMSD per residue in the S4N region shows large fluctuations.

When focusing on the behavior of S4N, its flexibility makes the structural alignment of conformations along the trajectories difficult to interpret when the alignment is based on the backbone of all 39 residues. A more informative comparison is achieved by aligning the relatively stable eight-residue $\alpha$-helix (Figure 4.3B). This procedure results in a backbone RMSD of less than 1.5 Å for the $\alpha$-helix in all the simulated trajectories, clarifying that the overall RMSD value, though exceptionally large, is a measure of how much the coils and loops have moved away from their original orientations. Another measure of similarity to the native structure is Q, which is based on residue-residue pairwise distances and does not depend on structure alignment.

Time averaged RMSD and Q were plotted for the last 45 ns from each of the 23 trajectories involving S4N (Figure 4.3C). Unbound S4N is much more flexible than S4N bound to h16. All simulations of the S4N:h16 complex (blue squares) have average S4N RMSD values smaller than 8 Å while all simulations of the unbound protein (red dots) have RMSD larger than 8 Å. Similar results are observed for the mean.
Figure 4.3: Backbone and secondary structure fluctuations of S4N. A) RMSD per residue shown in the plot is an average of two independent simulations of the full-length S4. B) RMSD per residue for S4N was calculated from selected representative replicates. C) Time-averaged Q plotted against backbone RMSD for all simulations involving S4N. Each marker represents one particular replicate. Inset figures are representative conformations sampled from complex (left) and unbound (right) simulations and colored by RMSD per residue. D) CD signal at 222 nm measured for S4N alone when temperature is increased gradually (left). The right plot shows how secondary structure changes over time for an unfolding simulation of S4N at 85°C.

Q values, with the complex formed from either the full system or just the signature regions all having values greater than 0.4. From calculations carried out for protein folding and structural phylogenetic studies, values of Q below 0.30 represent unfolded states or structurally unrelated proteins [156]. Furthermore, the disordered structure of unbound S4N is consistent with an NMR study where the N-terminal fragment of S4 could not be resolved together with the globular C-terminal domain [157].

The backbone motions arise mostly from the coil and loop region between residues 23 and 30, as seen in the RMSD per residue plots in Figure 4.3B as well as the inset representative conformations colored by
RMSD per residue in Figure 4.3C. When S4N is bound to h16, positively charged residues Lys21, Arg25, Lys30 and Lys32 interact with the RNA backbone strongly so that the coils vary little from their binding positions. Though the range of movements is limited, the side-chains of these lysines and arginines can interact with the phosphates of nucleotides neighboring the contact sites in the crystal structure as discussed later.

The high fraction of charged residues within S4N as well as the dominant random coil structure facilitates the formation of salt bridges when S4N is not bound to h16. Though none is observed in the crystal structure, 23 different salt bridges are recorded in the unbound runs (Table 4.1), while only nine are observed during the simulations of the signature complex. Without the presence of h16, the charged amino acids are free to interact with each other, and the only constraints on salt bridge formation are geometric ones enforced by the stability of the eight-residue helix. Of the 23 salt bridges reported in the 500-ns simulations of unbound S4N, six (Glu14-Arg2, Glu14-Arg13, Glu34-Arg13, Glu34-Arg2, Glu34-Arg12 and Asp17-Arg25) have both substantial occupation (>10%) and duration (> 2 ns). Formation of these many internal salt bridges gives rise to the large fluctuations reported in the RMSDs and the disordered property of S4N.

4.3.2 Temperature unfolding studies of S4N

During the temperature unfolding simulation (see Methods), the stability of the α-helix seen in both the unbound and complex simulations of S4N was maintained. The helix persisted across the 100-ns simulation with two smaller helices of length 4-5 transiently appearing over 40 ns after the temperature jump (Figure 4.3D). The additional helical turns are seen around the same positions in the crystal structure of *T. thermophilus* 30S subunit and also occur occasionally during the room temperature simulations (data not shown). However, it is clear that they persist more extensively when temperature is raised moderately, probably due to the rapid rearrangement of backbone φ, ψ values under such temperatures.

CD measurements taken during temperature melt experiments on unbound S4N also showed increasing α-helical content. The CD signal at 222 nm, the wavelength characteristic of α-helices, decreased linearly with increasing temperatures (Figure 4.3D). Furthermore, the CD temperature melt also supports the disordered structure of S4N at room temperature, since no cooperative structural transition of S4N was observed.

4.4 Dynamics of the RNA Signature h16

4.4.1 Stability and secondary structure fluctuation of h16

In the MD simulations, h16 in both the unbound and bound systems shows a much greater stability than S4N at room temperature. According to Figure 4.4A, the RMSDs of h16 in the bound forms are only slightly smaller than in the unbound forms. Similar trends are also observed for the time-averaged Q values. Fluctuations in h16 further decrease when it is part of the five-way junction complexed with the full-length S4, with an average RMSD near 2.3 Å and a Q value over 0.6. The calculated stability is in agreement with thermal denaturation experiments on the isolated, unbound h16, in which absorbance at 260 nm monitored global unfolding of the RNA molecule as a function of temperature. The result shows that h16 is stable at room temperature, and a gradual unfolding transition occurs with a melting midpoint of 41±5°C (Figure 4.4B).
The temperature dependence of 2AP fluorescence provides a more local measure of the base unstacking and decreased quenching by the Iowa Black label. It shows that the ends start fraying and the loss of base stacking occurs at 60 °C (Figure 4.4C), again confirming the stability of h16. The MD simulations at elevated temperatures show that without S4N to stabilize the internal bulge conformation, it expands allowing greater conformational flexibility. Base stacking along the frayed ends remains intact longer and only slowly becomes disordered. The difference in the melting behavior from the two probes indicates that the folding transition of h16 is more complicated than a simple two-state model.

Fluctuations in the structure of unbound h16 arise primarily from shifts in base pairing and base stacking interactions. In the crystal structure, h16 consists of two short stems connected by an internal loop and capped with a UUCG tetraloop at the end (Figure 4.1). Of the 13 base pairs and 22 base stacking interactions

---

**Table 4.1: Salt bridges within ribosomal protein S4 N-terminus during simulations**

<table>
<thead>
<tr>
<th></th>
<th>Unbound</th>
<th></th>
<th></th>
<th>Complex</th>
<th></th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Negative A.A.</td>
<td>Positive A.A.</td>
<td>Occu. * (%)</td>
<td>Dur. † (ns)</td>
<td>Replicate #‡</td>
<td>Occu. (%)</td>
</tr>
<tr>
<td>Glu14</td>
<td>Arg2</td>
<td>21.98 (19) 5.17</td>
<td>1,2,3,4,7,8,9,10</td>
<td>19.21 (31) 2.88</td>
<td>5,6,7,8,9,10</td>
<td></td>
</tr>
<tr>
<td></td>
<td>Arg13</td>
<td>26.73 (37) 3.27</td>
<td>2,3,4,5,6,7,8,(9)</td>
<td>29.29 (56) 2.37</td>
<td>2,3,4,(5),8,9,10</td>
<td></td>
</tr>
<tr>
<td></td>
<td>Lys32</td>
<td>0.34 (2) 0.60</td>
<td>10</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>Arg25</td>
<td>0.03 (4,9)</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>Lys30</td>
<td>0.01 (10)</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>Lys7</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Glu34</td>
<td>Arg13</td>
<td>17.48 (35) 2.35</td>
<td>2,4,(6,7),9,10</td>
<td>0.04 (1)</td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>Arg2</td>
<td>12.31 (6) 9.19</td>
<td>4,7,9,10</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>Arg12</td>
<td>10.93 (17) 2.91</td>
<td>1.2,9</td>
<td>0.16 (2)</td>
<td>0.18 (8),9</td>
<td></td>
</tr>
<tr>
<td></td>
<td>Lys9</td>
<td>6.73 (19) 1.65</td>
<td>2.9</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>Arg25</td>
<td>0.24 (2) 0.53</td>
<td>1</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>Lys30</td>
<td>0.01 (1)</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>Lys21</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>Lys32</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Asp17</td>
<td>Arg25</td>
<td>17.29 (23) 3.42</td>
<td>3,6,7,8,9,10</td>
<td>0.23 (1)</td>
<td>1.20 (7)</td>
<td>7</td>
</tr>
<tr>
<td></td>
<td>Arg12</td>
<td>8.31 (16) 2.33</td>
<td>1,2,6,(8),9</td>
<td>0.02 (1)</td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>Lys21</td>
<td>1.81 (16) 0.49</td>
<td>(1,2),7,(9),10</td>
<td>0.21 (2)</td>
<td>0.45 (5,7)</td>
<td></td>
</tr>
<tr>
<td></td>
<td>Lys32</td>
<td>2.90 (2) 6.78</td>
<td>10</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>Lys30</td>
<td>0.53 (2) 1.20</td>
<td>3.6</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>Lys7</td>
<td>0.44 (2) 1.05</td>
<td>8</td>
<td>0.09 (1)</td>
<td>0.20 (5)</td>
<td></td>
</tr>
<tr>
<td>Asp28</td>
<td>Arg25</td>
<td>7.39 (21) 1.58</td>
<td>1.6,7,9,10</td>
<td>0.37 (2)</td>
<td>0.98 (9)</td>
<td></td>
</tr>
<tr>
<td></td>
<td>Lys30</td>
<td>2.13 (17) 0.48</td>
<td>1,4,(5),6,8,10</td>
<td>0.09 (1)</td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>Arg12</td>
<td>1.28 (3) 1.92</td>
<td>1,(9)</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>Lys32</td>
<td>0.09 (1) 0.35</td>
<td>(3), 4,(6)</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>Lys21</td>
<td>0.04 (3,9)</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

* Averaged percentage of occupancy over ten replicate runs.
† Averaged residence time, measured in nano-seconds. Number in parenthesis is the number of times that the salt bridges formed in all replicates.
‡ Numbers in this column indicate which replicate runs have the specific salt bridges. Replicate number in parenthesis shows that this replicate run does not possess long stretches of salt bridges that are counted into the duration calculation.
seen in the crystal structure of h16, several were lost in the hairpin loop, the 5'/3' ends, and the internal loop where S4N binds (Figure 4.5). In both unbound and complex simulations, the most abundant non-native base pairs (not present in the crystal structure) are A414·A430, G413·G428, U429·A432 and U407·C436, and the first three of these base pairs are in the internal loop. Unbound replicates 5 and 9 have lost the largest number of native base pairs (Figure 4.4C). Particularly, in replicate 9, base pairing interactions in the internal loop have been greatly rearranged, and the two end bases (U421 and C422) in the tetraloop form an additional base pair that deforms the RNA backbone. In replicate 5, each base in A415 – C419 switched
base pairing partners with its downstream neighboring base resulting in a register shift of one base. This relatively stable large scale shift is unique, but smaller, brief register shifts of one or two base pairs are common in all the simulations.

The average number of base stacking interactions remained stable in both unbound and complex runs. However, several pairs of stacked bases that were absent from the crystal structure appeared during simulations. For example, U407/A408, C418/C419, U426/U427 and U434/2AP435, which kept the continuity of the stem; and A411/G413, which is the counterpart of the stacked bases G428/A430 on the 5′ strand, and sits on top of the three inter-strand stacking base pairs in the internal loop. Other cross-strand base stacking interactions can be attributed to a more twisted helical backbone.

Secondary structure interactions in h16 are approximately the same whether it is isolated or included in the five-way junction as seen in Figure 4.5. The most noticeable difference is that base G423 flips out of the h16 hairpin loop. Base pair U420·G423 and base stacking between G423/G424 were rarely seen in the five-way junction simulations, whereas in the unbound and complexed h16 runs occupancies of base pair U420·G423 are 56.2% and 38.8%; and stacked bases G423/G424 49.3% and 36.0%, respectively. The fact that the h16 hairpin loop is indeed a tetraloop in the five-way junction (but not in h16 alone) as seen in the traditional secondary structure diagram, can be attributed to the interaction between h16 and h18. The backbone fluctuations of h16 are slightly less when it is part of the five-way junction as seen in the RMSD shown in Figures 4.4. The collective motion of the five-way junction is mainly demonstrated in the relative positioning among the helices, especially the distances between interacting sites on h16 and h18 (will be discussed later). Even without S4, secondary structure of the individual helices h4, h16-h18 in the five-way junction remains stable over the 50 ns simulation, but h16 and h18 begin to separate from each other near the end of the simulation (Figure 4.7A).

In general, replicates with more flexibility in base pairing or base stacking interactions have larger overall RMSD and smaller Q values. Furthermore, the register shift of base pairs in an RNA stem should be able to maintain most of its original base stacking interactions. A more twisted backbone, which can shift base stacking from within one strand to across the strand, does not necessarily break the native base pairs. The rearrangement of base pairs in the internal loop, register shift of base pairs in the stems, and twisting of the helical backbone contribute to the subtle intrinsic fluctuations in h16 structure.

4.4.2 Temperature melt experiment of h16

The fluorescence melt curve for h16 alone is shown in Fig 4.4C. The native baseline has been subtracted from the data to make it easier to visualize the unfolding transition. The curve shows that the fraying starts at 60°C and proceeds to unfold up to the experimentally measured range of 90°C. The unfolded baseline is not observed and would be attained at an even higher temperature not accessible under the measurement conditions. The difference in the melting transition between the absorbance and fluorescence data for the h16 fragment is common among biopolymers (RNA and protein) having a rugged energy landscape. Different probes show slightly different unfolding temperatures depending on the kind of local environment they access as the molecule unfolds [158].
Figure 4.5: Base pairing and base stacking interactions during simulations. Each base pairing interaction and base stacking interaction is colored by occupancy during the last 45 ns of all relevant replicates. Black boxes indicate native interactions present in the crystal structure. Top two plots show the interactions in the unbound h16 and the S4N:h16 complex, and the bottom two plots show the interactions of h16 in the five-way junction with and without S4.

4.5 Dynamical Protein:RNA Signature Interactions and Their Function

4.5.1 Experimental measurement of protein:RNA interactions

How strongly the two molecules bind can be determined by measuring the fluorescence of the S4N:h16 complex during progressive dilution (Figure 4.6A). Tryptophan fluorescence from S4N is quenched in the complex due to the proximity of Iowa Black FQ which is attached to the 5′ end of h16, so higher fluorescence intensity indicates a larger fraction of unbound S4N in solution. The general trend of the increasing normalized fluorescence intensity clearly shows that as the solution is diluted, the binding equilibrium shifts to more unbound S4N and h16. The two-component thermodynamic model allows us to obtain a binding constant of 0.91 μM which is not surprising given the fact that only fragments of the rRNA and the protein S4 are used in this experiment.

CD measurements were taken to investigate any possible conformational changes for either S4N or h16 upon binding. At 20 μM concentration, the sum of the individual spectra of h16 and S4N were subtracted from that of the 1:1 S4N:h16 mixture to obtain the difference spectrum (Figure 4.6B red curve). This spectrum, together with the individual curves, indicates how the CD signal changes upon binding. At shorter wavelength (190 – 215 nm), CD signal changes are mainly due to conformational changes in the protein. Increase in this region is generally indicative of a decrease in random coil characteristics in S4N upon binding. However, the small decrease observed near 222 nm is inconclusive as to the change in helical content, because the amplitude of the signal is almost comparable to the noise. Most interesting is the drop
in CD signal at longer wavelengths (260-280 nm). There are two possible explanations for this drop: either h16 changes conformation upon binding S4 or the difference is caused by interactions between h16 and S4N, as nothing is expected from the protein in this region.

Figure 4.6: Binding of S4N and h16. A) Dilution measurement shows increasing signal as the concentration of S4N:h16 is lowered. Inset plot shows the fluorescence quenching spectrum of S4N, h16 and 1:1 mixture of S4N:h16 at the highest concentration of 7.5 μM. B) Experimentally measured and theoretically calculated (inset plot) CD signals.

To determine which scenario was occurring, CD spectra were calculated based on MD trajectory data. Snapshots were taken from every replicate, and coordinates therein were used to calculate the CD signals using the online software DichroCalc (Figure 4.6B inset, values have been scaled by 1200 times to enable direct comparison with experimental data). In the long wavelength region (260-280 nm) where unexpected decrease of CD signal was seen upon binding, the averaged calculated CD signals peaked at the same wavelength and spread out in the same order as in experimental measurements with an average diagonal correlation coefficient of 0.94. Furthermore, the mean CD spectra calculated using h16 and S4N coordinates
taken from both complex and unbound runs showed that neither h16 or S4N has a conformational change that causes the CD signal to change at the peak wavelength ~268 nm (data not shown). Therefore, the decrease of CD signal at the long wavelength range upon binding of h16 and S4N is a consequence of interactions between the two molecules. As further discussed in the following section, conformational changes due to the existence of several cation-π interactions between the bases on h16 and the positively charged side-chain S4N might give rise to the CD signal change, but it is not clear how they are included in the calculated spectra.

4.5.2 Side-chain interactions between S4N and h16

In order to characterize the possible interactions that cause the CD signal change around the 270-nm region, we looked into the trajectories and searched for specific interactions between protein and RNA molecules. First, salt bridge interactions are discussed in the paper. Second, backbone hydrogen bonding interactions between S4N and h16, were scanned through the simulations. The results showed that the interaction between the two molecules were very dynamic, and an average of 88 hydrogen bonds between different pairs of residues was estimated for each replicate. Approximately half of those hydrogen bonds were transient (with an occupancy less than 1% during simulations), while the majority of the other half were formed through positively charged residues. The backbones of Leu8 and A430 formed the most common hydrogen bond involving protein residues other than arginine or lysine, and this interaction should be important to draw the α-helix close to the RNA molecule. Other highly populated hydrogen bonds were mostly seen between glutamine-serine and h16.

Third, three regions on h16 are identified to be rich in cation-π interactions during simulations. First, G413 and A412, the two bases which are neither base-paired or stacked to any of the other bases in the crystal structure. Second, U409 and G410, where the interacting protein residues can reach into the major groove of the RNA helix. Third, U409 and G410, which is the opening end of the molecule. All three places are on the 5′ strand of h16 and they form cation-π interactions with Lys32/Lys30/Arg25, Lys21/Lys7 and Lys21/Arg2 respectively. The first two regions show such interactions in every complex replicate, and the interactions are stabilized by the nearby salt bridges (will be discussed later) or hydrogen bonding to the functional group on the bases. Cation-π interactions with G406 and U407 depend on the flexibility of the two bases and are mostly transient in the simulations. Once the end base pair opens, other interactions, such as the stacking between RNA base G406 and aromatic protein side-chain from Phe19 or Tyr3, may also occur. Theoretical calculation of the CD signal using DichroCalc includes the transition dipole moments of the RNA bases, peptide bonds, aromatic and negatively-charged side-chains of the protein. At this point, it is not clear what interaction contributes to the CD signal change most.

4.5.3 Salt bridge interactions between RNA backbone and protein side chain

A detailed description of the salt bridges formed between protein residues arginine/lysine and phosphate oxygens of the RNA (Table 4.2) is given to illustrate the variety of possible amino acid:nucleotide charge-charge interactions in S4N:h16 complex. Among the fifty salt bridges identified in simulations, only three (Arg12-U429, Lys9-G428 and Lys32-U426) appeared in every replicate run. Even for these three salt
Table 4.2: Salt bridges between ribosomal protein S4 N-terminus and backbone phosphate of h16

<table>
<thead>
<tr>
<th>ARG Res.</th>
<th>Occu. (%)</th>
<th>Dur. (ns)</th>
<th>Replicate #</th>
<th>LYS Res.</th>
<th>Occu. (%)</th>
<th>Dur. (ns)</th>
<th>Replicate #</th>
</tr>
</thead>
<tbody>
<tr>
<td>Arg2</td>
<td>G406</td>
<td>8.23</td>
<td>(33) 1.19</td>
<td>1,3,4,5,8,9,10</td>
<td>Ly57</td>
<td>9.22</td>
<td>(24) 1.78</td>
</tr>
<tr>
<td>U407</td>
<td>12.94</td>
<td>(25) 2.38</td>
<td>1,2,3,4,7,8,9</td>
<td>U407</td>
<td>24.33</td>
<td>(55) 2.03</td>
<td>1,2,3,4,5,6,7,9,10</td>
</tr>
<tr>
<td>A408</td>
<td>1.18</td>
<td>(6) 0.92</td>
<td>2,3,9</td>
<td>A408</td>
<td>20.48</td>
<td>(78) 1.22</td>
<td>1,2,3,4,5,6,7,9,10</td>
</tr>
<tr>
<td>A431</td>
<td>0.13</td>
<td>(1) 0.80</td>
<td>9</td>
<td>U409</td>
<td>0.64</td>
<td>(6) 0.46</td>
<td>2,3,5,7</td>
</tr>
<tr>
<td>Arg12</td>
<td>U426</td>
<td>5.16</td>
<td>(29) 0.79</td>
<td>1,4,5,6,7,8,9,10</td>
<td>U427</td>
<td>54.24</td>
<td>(21) 11.72</td>
</tr>
<tr>
<td>U427</td>
<td>9.22</td>
<td>(24) 1.78</td>
<td>1,2,3,4,5,6,7,9,10</td>
<td>U427</td>
<td>14.41</td>
<td>(20) 3.22</td>
<td>1,2,3,4,5,6,8,10</td>
</tr>
<tr>
<td>G428</td>
<td>0.02</td>
<td>(1) 0.57</td>
<td>9</td>
<td>G428</td>
<td>11.78</td>
<td>(11) 4.85</td>
<td>1,3,6,10</td>
</tr>
<tr>
<td>Arg13</td>
<td>U426</td>
<td>0.27</td>
<td>(1) 1.10</td>
<td>1</td>
<td>A430</td>
<td>8.52</td>
<td>(29) 1.34</td>
</tr>
<tr>
<td>U427</td>
<td>11.73</td>
<td>(26) 2.05</td>
<td>1,4,5,7,9</td>
<td>Ly521</td>
<td>62.96</td>
<td>(50) 2.51</td>
<td>1,3,4,5,6,7,8,9,10</td>
</tr>
<tr>
<td>G428</td>
<td>2.67</td>
<td>(11) 1.06</td>
<td>5.7</td>
<td>U407</td>
<td>10.91</td>
<td>(4) 1.20</td>
<td>5,6,7</td>
</tr>
<tr>
<td>A408</td>
<td>0.07</td>
<td>(1) 0.15</td>
<td>5</td>
<td>A408</td>
<td>12.07</td>
<td>(16) 3.38</td>
<td>(1,3,5,6,7,9,10)</td>
</tr>
<tr>
<td>Arg25</td>
<td>A408</td>
<td>0.03</td>
<td>(1) 0.20</td>
<td>8</td>
<td>U409</td>
<td>7.48</td>
<td>(36) 0.94</td>
</tr>
<tr>
<td>U409</td>
<td>3.99</td>
<td>(13) 1.50</td>
<td>6.8</td>
<td>G410</td>
<td>23.72</td>
<td>(32) 3.30</td>
<td>1,2,3,4,7,8,9,10</td>
</tr>
<tr>
<td>G410</td>
<td>22.90</td>
<td>(49) 2.23</td>
<td>1,3,4,5,6,7,8,10</td>
<td>A411</td>
<td>0.09</td>
<td>(2) 0.25</td>
<td>7</td>
</tr>
<tr>
<td>A411</td>
<td>27.52</td>
<td>(77) 1.71</td>
<td>1,3,4,5,6,7,8,10</td>
<td>A430</td>
<td>15.71</td>
<td>(54) 1.38</td>
<td>1,2,3,4,6,9,10</td>
</tr>
<tr>
<td>A412</td>
<td>10.67</td>
<td>(23) 2.12</td>
<td>1,4,5,7,10</td>
<td>Ly530</td>
<td>2.17</td>
<td>(6) 1.61</td>
<td>5,6,7</td>
</tr>
<tr>
<td>G413</td>
<td>2.36</td>
<td>(16) 0.70</td>
<td>1.2</td>
<td>A412</td>
<td>9.24</td>
<td>(22) 1.98</td>
<td>4,5,6,7,10</td>
</tr>
<tr>
<td>G425</td>
<td>0.71</td>
<td>(3) 1.08</td>
<td>(2),10</td>
<td>G413</td>
<td>12.59</td>
<td>(36) 1.58</td>
<td>(2),3,4,5,6,8,9,10</td>
</tr>
<tr>
<td>U426</td>
<td>0.23</td>
<td>(4) 0.39</td>
<td>10</td>
<td>G424</td>
<td>12.08</td>
<td>(24) 2.31</td>
<td>2,3,4,6,8,9,10</td>
</tr>
<tr>
<td>U429</td>
<td>17.48</td>
<td>(19) 4.31</td>
<td>3,6,9</td>
<td>G425</td>
<td>26.57</td>
<td>(82) 1.55</td>
<td>1,2,3,4,6,8,10</td>
</tr>
<tr>
<td>A430</td>
<td>7.13</td>
<td>(11) 3.11</td>
<td>3</td>
<td>U426</td>
<td>11.78</td>
<td>(11) 4.85</td>
<td>1,3,6,10</td>
</tr>
<tr>
<td>Lys32</td>
<td>G424</td>
<td>1.46</td>
<td>(8) 0.86</td>
<td>5</td>
<td>G425</td>
<td>24.41</td>
<td>(41) 2.74</td>
</tr>
<tr>
<td>U426</td>
<td>63.00</td>
<td>(65) 4.43</td>
<td>all</td>
<td>U427</td>
<td>23.09</td>
<td>(34) 3.09</td>
<td>2,3,4,6,7,8,10</td>
</tr>
<tr>
<td>U429</td>
<td>20.21</td>
<td>(18) 5.10</td>
<td>2,3,4,6,7,8</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

*Calculation has included atoms O1P, O2P, O5’ from residue N and atom O3’ from residue N-1;
†Percentage values are obtained by averaging over the last 45 ns of all ten replicates.
‡Averaged residence time, measured in nano-seconds. Number in parenthesis is the number of times that the salt bridges formed in all replicates;
§Numbers in this column indicate which replicate runs have the specific salt bridges. Replicate number in parenthesis shows that this replicate run does not possess long stretches of salt bridges that are counted into the duration calculation;
¶Rows in bold indicate presence in crystal structure;
∥These residues have interaction with other helices in the crystal structure (Supporting Material).

bridges, the largest occupancy is only ~68% and the longest average duration ~7.2 ns. Considering the other salt bridges present in the crystal structure (shown in bold in Table 4.2), the average occupancy can be as low as ~7.5% and average duration shorter than 1 ns. Most of the salt bridges appearing during simulation had an occupancy under 20% and mean duration under 3 ns.

Four of the nine positively charged protein residues interact exclusively with the 3’ strand of h16, and two interact with the 5’ strand for the majority of time. Generally, the 3’ strand of h16 interacts much more strongly with the S4N than the 5’ strand does. However, Arg25, Lys21, and Lys30 interact with both strands and especially the internal loop. The number of salt bridges formed for these three residues was large, but
not one of them dominated in terms of occupancy or duration. These three residues either sit in or are close to the coil region on S4N where RMSD per residue is as high as the opening ends even in complex runs. These fluctuating contacts contribute to the high RMSD values on this particular region of the S4N.

### 4.5.4 Role of S4N signature in the refolding of the five-way junction

The average electrostatic potential map of S4 has a predominantly positive surface along the interface to the rRNA. Due to its large number of positively charged residues, S4N contributes disproportionally to this potential. As suggested in the fly-casting mechanism for molecular recognition [159–161], the docking of a protein to its nucleic acid target can be accelerated by electrostatic forces and an increased capture radius when the protein is unfolded. This suggests that guided by electrostatic potential, the disordered S4N searches for its binding site while undergoing conformational changes. The unstructured coil gives it a larger searching volume centered on the stable C-terminus, and the alternating salt bridges expose different charged residues to the target RNA. As there are so many acceptable interactions between h16 and the disordered S4N, the initial contact can be established with very few tries. These interactions might be weak with a binding constant much smaller than that of the binding of the full-length S4 to 16S rRNA, but they could start the coupled folding and binding and speed up the assembly process.

To demonstrate how quickly and effectively these interactions can be established, two additional simulations, in which either S4N or magnesium ions were added to the unfolded five-way junction, were performed for 50 ns each. From alignments of the ensemble of partially unfolded S4 and five-way junctions, an initial placement of S4N was selected such that only a single contact between Lys7 and A411 existed. Within 50 ns, additional salt bridges between h16/h18 and the lysines/arginines on S4N were made. A comparison of the interactions in Figure 4.7B with those observed during S4N:h16 simulation (provided in the Supporting Material), reveals that many interactions between S4N and h16 were re-established within 50 ns. Furthermore, interactions established between Arg2/Arg13 and h18 are spatially close to those in the crystal structure. As measured by the decrease in distance between the O2′ atoms in C418 (h16) and G540 (h18) in Figure 4.7B, the addition of S4N partially refolds the five-way junction. This distance between h16 and h18 further decreased to ~7 Å within the next 50 ns, but then the pseudoknot in h18 began to separate. This is consistent with the crystal structure contacts (Figure 4.1) and the experimental measurements in [149], which shows that stabilization of the pseudoknot requires the presence of S4C. For comparison, addition of magnesium folds the five-way junction by drawing nucleotides on both h16 and h18 close together [162; 163]. However, since the nucleotides interacting with the magnesium ions do not form native contacts in the crystal structure, the five-way junction is caught in a wrong conformation.

The S4 is unique in being both a fast primary binding protein and an initiator protein. However, the suggested binding mechanism can be applied to other signature regions on the ribosome. Most ribosomal structure signatures are interacting with one or two r-proteins (see Supporting Material). Some of these proteins are defined as intrinsically disordered in the Disprot Database [164], such as S18, S19 and L27; others interact with the RNA using its obvious unstructured loops. It is reasonable to suggest that structure signatures occurred later during cellular evolution, promoting faster assembly of the ribosome using a fly-casting strategy with the help of cooperatively designed, disordered r-proteins. Further validation is needed.
for the functional role of signatures other than h16 and S4N, since those proteins depend on the prior binding of primary binding proteins.

4.5.5 Conclusions

In this chapter, we investigated the functional role of RNA (h16) and protein (S4N) bacterial signatures in the ribosomal 30S assembly. The signature complex is the result of coevolution, and it is small enough so that multiple, long, all-atom MD simulations can be performed to study the dynamics of the complex as
well as the individual components. Agreement between analysis of the simulated trajectories and results of the experimental data obtained from fluorescence spectroscopy and CD at room temperature demonstrate clearly the stability of h16 and the intrinsic disorder of the unbound S4N with respect to the C-terminal portion of S4. In the simulations, the dynamic fluctuations in the highly charged S4N were exploited to initiate binding of the protein signature to h16 in the unfolded five-way junction which resulted in its partial refolding. Such a mechanism is consistent with S4’s known role as a primary binding and initiator protein for the assembly of the bacterial SSU and may be a consequence of evolutionary pressure to assure the rapid binding of S4 to the 16S rRNA.

4.6 Methods

4.6.1 Protein and RNA design and sample preparation

The evolutionary analysis [165] of S4 indicates that the flexible N-terminus signature extends from residue 1 to approximately 45 in Escherichia coli. However, in order to study the interaction with h16, the S4N fragment in this study has been chosen such that the last residue is within 5 Å of the h16 in the crystal structure. The fragment of the wild type E. coli S4 from positions 1 to 39 was ordered from Genscript (Piscataway, New Jersey). In order to have a fluorescence probe within the protein for bulk measurements, the isoleucine at position 33 was replaced by tryptophan. The cysteine at position 31 was replaced by serine while the alanine at position 1 was replaced by cysteine. These two cysteine mutations were introduced to facilitate the labeling of the N-terminus with an Alexa-488 fluorophore for future single molecule experiments. The mutated residues were carefully chosen according to the sequence alignment of S4 provided in [165], ensuring minimum perturbation from native behavior. The N-terminus of the fragment was acetylated, and the C-terminus was amidated.

The h16 RNA fragment from position 406 to 436 was ordered from Integrated DNA Technologies (Coralville, Iowa) with two modifications: the adenine at position 435 was changed to 2-aminopurine (2AP), a fluorescent analog of adenine shown to not perturb RNA folding [166], and a non-fluorescent quencher Iowa Black FQ was attached to the 5′ end of the molecule. The wild type and modified sequences of both S4N and h16 are shown in Fig. 4.1.

Standard PE buffer, 10 mM sodium phosphate and 0.1 mM EDTA-Na2 with pH 7.1, was used for all experiments performed in this paper.

4.6.2 Experimental measurements

Absorbance thermal melts of the 1µM h16 solution were performed in a 1 cm path length cell using spectropolarimeter (Jasco Inc., Easton, MD). Absorbance intensities were measured at 260 nm, and the resulting curve was fitted to a two-state thermodynamic heat capacity model [167]. The determination of the unfolding transition temperature is described in detail in the Supporting Material.

Fluorescence melts were performed using the Cary Eclipse Fluorescence Spectrophotometer (Varian Inc., Palo Alto, CA). An excitation wavelength of 280 nm was used and the spectra were scanned from 320 nm to 440 nm. The temperature was increased at 5°C intervals. The integrated spectrum was used for analysis,
and the linear native signal baseline was subtracted from the h16 fluorescence melt to make the unfolding transition more obvious.

**Dilution measurements** were performed using a Cary Eclipse fluorescence spectrophotometer (Varian Inc, Palo Alto, CA). Excited at 280 nm, the fluorescence spectrum was measured for 600 µL of 7.5 µM sample of the 1:1 mixture of S4N and h16 at room temperature (22°C). Thereafter, 100 µL of sample solution was removed from the cuvette and replaced with 100 µL of buffer, stirred, and the spectrum was measured again. This was repeated until a concentration of 0.34 µM was achieved. The spectrum was baseline subtracted and then fitted to a two-component thermodynamic model to obtain the binding constant K and dissociation constant K$^{-1}$. Details of the model and formula used to calculate K are provided in the Supporting Material.

**CD spectra** were measured for a 20 µM solution of the protein and RNA molecules separately as well as the 1:1 mixture of the two molecules in a 1 mm path length cuvette using a spectropolarimeter (Jasco Inc., Easton, MD). The CD spectra were scanned from 190-300 nm.

### 4.6.3 Molecular modelling and simulation

Coordinates for S4 and the five-way junction in 16S rRNA were taken from the crystal structure of *E. coli* ribosomal SSU at 3.22 Å resolution (PDB ID 2I2P [91]). The Ile33Trp, Cys31Ser, and Ala1Cys mutations on the protein and A435(2AP) mutation on the RNA sequence were made according to sequences chosen to perform the experiments (Fig. 4.1) in the signature system (S4N and h16). Parameters for 2AP were developed by analogy with the separate adenine and lysine parameters already present in the CHARMM27 force field. The five-way junction system, including h3, h4, h16, h18, a truncated h17, a five-membered loop CUCAA that caps h4, and a seven-membered loop UUUUGCU that caps the truncated h17, was assembled according to the minimal S4 binding model suggested by Bellur et al. [149]. The two additional loops were taken from the *E. coli* SSU (PDB ID 2I2P, residue 618 – 622) and glutamine tRNA (PDB ID 2RD2, residue 932 – 938) with two mutations (C934U and A937C), respectively.

All systems studied in this paper were neutralized with Na$^+$ or Cl$^-$ [168] and prepared in VMD [89] according to the protein:RNA simulation protocol in [93]. 10 mM NaCl was added to the final solution according to experimental conditions, and equilibration simulations were run using NAMD2 [96] with periodic boundary conditions and the NPT ensemble with pressure set to 1 atmosphere and temperature set to 298 K. Electrostatics were calculated with the particle mesh ewald method. The van der Waals interactions were calculated using a switching distance of 12 Å and a cutoff of 14 Å.

Ten, 50-ns MD simulations were run for each of S4N, h16, and the S4N:h16 complex, and they are referred to as unbound S4N runs, unbound h16 runs, and complex runs, respectively. Four additional runs were performed to establish reference points: two 50-ns runs for the full-length, unbound S4 (∼205 residues), one 50-ns run for the five-way junction, and one 50-ns run for the complex of the five-way junction and full-length S4. Additional details of the methodology and parameters are provided in the Supporting Material.

For comparison to the experimental melting data, unfolding of the S4N was also simulated. After 20 ns of equilibration, the temperature was raised linearly from 298K to 358K through one hundred 5-ps steps. Production runs at 358K were carried out for another 99.5 ns to achieve a total of 100 ns of unfolding simulations.
4.6.4 Base pairing and base stacking interaction determination

Base pairing interactions were determined using 3DNA [72], which reports both canonical and non-canonical base pairs. Hydrogen bond patterns were also recorded for each base pair, including those between atoms on the base and atoms on the sugar ring or backbone. As the base stacking information given in 3DNA is limited to the overlapping areas of successive base pairs, a base stacking detection program was developed to include the occurrence of stacking interactions between bases not involved in base pairing. The criteria were based on the geometric measures established in [169], with slightly relaxed cutoff values obtained from [170] to incorporate the intrinsic RNA structure fluctuations. The three criteria used were: 1) distance between the geometric centers of the two base rings should be smaller than 5.5 Å, 2) the angle between the base normal vectors of the two base rings should be smaller than 30°, and 3) the angle between one of the two base normal vectors and the vector connecting the two ring geometric centers should be less than 40°. Both rings in the purine bases were calculated individually, and if one of the rings met all the above criteria, the base was considered stacked.

To validate the algorithm, pairs of bases were placed in various conformations, and each conformation was determined to be stacked or not according to the algorithm. The non-bonded energies were then calculated in each conformation to show that the identified stacking interactions were energetically favored. The algorithm was further tested using three tRNA simulations with high, medium and low magnesium concentrations from a previous study on EF-Tu:tRNA which demonstrated dramatic changes in both base pairing and stacking with ion concentrations. Trends in the amount of base stacking in the anticodon arm were in agreement with those obtained through both visual inspection and 3DNA calculations.

4.6.5 Salt bridge detection

Salt bridges were identified over the trajectories using a cutoff of 4.0 Å between nitrogen atoms on Arg/Lys side-chains and the oxygen atoms on Glu/Asp side-chains or on nucleotide phosphates (including atoms O1P, O2P, O5′ from residue N and atom O3′ from residue N-1). If more than one pair of atoms on the same pair of residues came within the distance cutoff, the salt bridge was only counted once. Percentage of occupancy for one particular salt bridge was averaged over a total of 450 ns equilibration of ten replicates for each system. Averaged duration for the salt bridge was calculated such that a spike of salt bridge formation in one or two frames was not counted, and a loss of salt bridge in one or two frames along a contiguous stretch of salt bridge was ignored.
Chapter 5

Protein-assisted Assembly Landscape of the Five-way Junction in the Ribosomal Small Subunit

5.1 Background

The assembly of the ribosome from over 50 individual components is a highly coordinated process that is crucial for cell growth. The study of ribosomal assembly first drew attention 40 years ago, when Nomura et al. summarized the hierarchy of the binding of ribosomal proteins (r-proteins) to the ribosomal RNA (rRNA) using reconstitution experiments [53]. The results showed that the assembly of the E. coli ribosomal small subunit (SSU) was dependent on the presence of S4 and five other primary binding proteins, which are required to be in place before the remaining r-proteins can be incorporated. Two of the primary binding proteins, S4 and S7, were later identified as the only assembly initiator proteins based on their noncooperative binding during the onset of the assembly process [144]. More recently, pulse-chase and quantitative mass spectrometry measurements of the binding kinetics for all r-proteins showed that the 5' domain binding proteins, especially S4, bind more quickly than proteins in the central or 3' domain, indicating a 5' to 3' directionality in the assembly process [146; 147; 171].

Progress in biophysical approaches allows the study of ribosomal assembly to include both protein binding and rRNA folding in the absence or presence of r-proteins. Förster resonance energy transfer (FRET) has been used to probe large conformational changes of the three-way junction in the central domain of 16S rRNA that binds the primary binding protein S15 [172–174]. It was shown that the RNA molecule went through a broad distribution of conformations at intermediate Mg$^{2+}$ concentrations, however, the binding of S15 induced the three-way junction into its native folded state. Chemical footprinting of RNA further pushes the description of RNA-folding to nucleotide resolution. Particularly, time-resolved hydroxyl radical footprinting of the 5'-domain of the E. coli 16S rRNA identified two folding stages in the folding of this domain [148; 175]. The early stage proceeds through multiple parallel pathways while the later is cooperative and induced by protein binding. More recently, the use of selective 2'-hydroxyl acylation analyzed by

\[ \text{The contents of this chapter are based in part on work previously published as Ke Chen, John Eargle, Jonathan Lai, Hajin Kim, Sanjaya Abeysirigunawardana, Magan Mayerle, Sarah Woodson, Taekjip Ha, and Zaida Luthey-Schulten. “Assembly of the five-way junction in the ribosomal small subunit using hybrid MD-Gō simulations,” J. Phys. Chem. B, 116(23):6819 (2012).} \]
primer extension (SHAPE) confirmed the role of the primary binding protein S4 in inducing and rearranging the five-way junction conformation during folding by measuring local RNA structure flexibility at single nucleotide resolution [176].

Computational methods, in principle, may yield a more comprehensive view of the assembly process such that both protein and RNA behaviors can be taken into account simultaneously at a high structural and energetic resolution. Methods and force fields to simulate peptide and protein folding have made considerable advances since the pioneering studies of Scheraga and coworkers [177–179], which were among the first to capture structural transitions. Although the availability of ribosome crystal structures has allowed all-atom molecular dynamics (MD) simulations to investigate functions of the intact ribosomes, computational studies of ribosomal assembly have been mostly based on reduced representations due to the number of atoms as well as the time-scale of the assembly process. Studies of ribosome stability, including coarse-grained Monte Carlo simulation [180], implicit solvent model [181], and Langevin dynamics simulations [182], provide insights into the correlation between order of protein binding with electrostatic energy or flexibility of the rRNA binding site.

In this work, we used a structure-based hybrid MD-Gō model to estimate the assembly pathway of a rRNA:protein complex which occurs on a millisecond experimental timescale. Gō models were developed in the context of protein folding according to the energy landscape theory [183]. The theory states that evolution has shaped the folding landscape of proteins into a rugged funnel. Frustration and barriers to folding come from competing interactions in the energy function which one can minimize by including stabilizing forces from native contacts and thereby reducing the probability of being trapped in non-native structures. Early studies with simple Gō models showed that the topology of the native structure of a protein determines to a large extent its folding mechanism [184–187] as well as its binding to other proteins [188–190]. Simple Gō models for RNA have also been able to differentiate the folding mechanisms of relatively small RNA pseudoknots [191].

In order to incorporate sequence-specific energetic contributions of the side-chains and base pairing, all-atom Gō models have been developed for protein folding [192–194] and transitions in nucleic acids. Specifically with the RNA models, one has been able to study transitions in riboswitches [195] and accommodation of tRNA to the ribosome [196].

Using an all-atom structure-based Gō model, we study the simultaneous folding and binding of the initiator r-protein S4 and its minimal binding site on the five-way junction in 16S rRNA [149]. This system is essential to the assembly process not only because the five-way junction is the part of the 5′-domain that folds first during ribosomal SSU assembly, but also due to its critical role in the ribosomal evolution. Helix 16 (h16) of the five-way junction is the largest bacterial structure signature (idiosyncrasies in the rRNA that are characteristic of the individual domain of life [37; 142]) in the ribosomal SSU, and it is in contact with a bacterial protein signature in the S4 N-terminus (S4N) [39]. The stability of h16 and intrinsic disorder of S4N revealed by MD simulations suggested that the interactions between this pair of coevolving signatures were key to the rapid recognition between the protein and RNA [67]. Here, we showed results from new MD simulations of the five-way junction, and identify multiple stable states of the naked RNA highlighting the rugged nature of its folding landscape. Starting from selected stable states of the five-way junction, binding
of S4 was initiated from a large distance away using the structure-based Gō model. To our knowledge, this is the first example of simultaneous folding and binding of RNA and protein molecules together using all-atom structure-based Gō model. With statistics obtained from hundreds of replicate simulations, we were able to reconstruct the binding pathway of S4 to the five-way junction. In addition, it also provides molecular details on how the intrinsically disordered S4N accelerates the binding process.

5.2 Equilibrium Dynamics of the Five-way Junction

5.2.1 Nucleotide flexibility in the five-way junction

Selective 2'-hydroxyl acylation analyzed by primer extension (SHAPE) measures local RNA structure flexibility at single nucleotide resolution, and has been shown to successfully predict RNA secondary structure as well as some tertiary interactions [197–200]. Although dynamics information is embedded in the SHAPE chemistry, SHAPE reactivities have only been compared to structural characteristics of static crystal structures or theoretical models of the corresponding RNA molecules. Here, we report all-atom molecular dynamics simulations with explicit solvent and hybrid MD-Gō simulations of the five-way junction with S4 bound that directly visualize these motions. As shown in Figure 5.1, the dynamical flexibility of each nucleotide in the five-way junction extracted from simulations on the E. coli system compared very well with the SHAPE reactivity obtained for five-way junction bound to B. stearothermophilus S4 [176]. The slightly higher nucleotide flexibility in simulations is most likely due to the lower binding affinity of the E. coli S4 compared to that of the B. stearothermophilus S4 observed in earlier experiments [201].

The flexibility of each nucleotide was measured as the root mean square deviation (RMSD) of heavy atoms from their average positions during simulation. Two alignment schemes were employed to remove the translational and rotational degrees of freedom. Each frame in the trajectory was either aligned back to the whole five-way junction to account for tertiary interaction change, or aligned to each individual helix to account for secondary structure fluctuations. The overall Pearson correlation coefficient between SHAPE reactivity and nucleotide fluctuations are 0.505 and 0.468 for tertiary and secondary structure alignment in all-atom MD simulation; and 0.618 and 0.604 in the hybrid MD-Gō simulations respectively. The slightly higher correlation achieved when aligning the whole RNA molecule supports previous evidence that the SHAPE chemistry measures nucleotide stability not only constrained by base-pair interactions but also by long-range tertiary contacts [202].

Among the five helices of the five-way junction, h17 shows the largest deviation between simulated nucleotide flexibility and the SHAPE reactivity. However, this discrepancy is expected due to the large structural difference arising from the truncation of h17 in the simulations (see Methods) as compared to the native h17 used in the SHAPE experiments. Differences in the signal intensities for the 5’ and 3’ end of h3 can be attributed to the “end effects” that strongly affect RNA folding, as well as the use of an extended 3’ sequence in SHAPE studies. Despite these difficulties, we still observe the same trend in signal fluctuation (relative intensity among residues) between calculated flexibility and SHAPE reactivity.
5.2.2 SHAPE mechanism revealed by comparison between experiment and simulation

Two major reasons have been proposed to account for the high experimental SHAPE reactivity for rigid nucleotides in the structure. First, since the nucleophilicity of the ribose 2′ hydroxyl group is sensitive to the electronic influences from nearby chemical groups, a larger than average distance between 2′-OH and the 5′ phosphate oxygen atoms (O1P or O2P) of the next nucleotide, or the purine N3 atom or pyrimidine O2 atom may increase the SHAPE reactivity. Second, reactivity might be enhanced by the C2′-endo pucker conformation of the sugar ring instead of the C3′-endo commonly seen in the A-form RNA. These suggested mechanisms were examined and confirmed in our simulations. For example, high reactivities of G517, A520, and G530 can be explained by relatively large average distances (> 4.2 Å) between 2′-OH and the N3 atoms on their bases; and that of A520 and U552 should be a result of large average distances between 2′-OH and the phosphate oxygens. On the other hand, the low reactivity of U421 is likely due to close adjacency between 2′-OH and phosphate oxygens when its flexible base flipped out into the solvent. One single nucleotide, A533, is seen to adopt a C2′-endo pucker conformation in the crystal structure and through the hybrid MD-Gō simulations. Note that none of these proposed mechanisms is sufficient to account for all the difference between experimentally measured SHAPE reactivity and computationally calculated flexibility. While the distance argument is supplementary to nucleotide flexibility, sugar pucker conformation has only minimal contribution to the SHAPE reactivity. It is pointed out by Gherghe et al. that the C2′-endo pucker
5.2.3 Consistency of standard MD and Gō potential simulations

In general, nucleotide flexibility averaged over multiple hybrid MD-Gō simulations (Figure 5.1 bottom) gives a better prediction of the SHAPE reactivity compared to that from a single MD simulation (Figure 5.1 top). This is expected, since SHAPE chemistry measures averaged nucleotide flexibility in an ensemble of RNA conformations. If one particular conformation dominates the RNA sample, as in the case of five-way junction bound to r-protein S4, high correlation between measured reactivity and calculated nucleotide fluctuation is expected, and discrepancies may be explained by the dynamics of the structure. On the other hand, if multiple conformations exist in the sample, especially when the relative population of each conformation is hard to estimate, poor correlations will probably result.

Also seen in Figure 5.1, the dynamics of each nucleotide in MD and hybrid MD-Gō simulations share the same characteristics. Since simulation with the Gō-potential involves a bias towards the native/target structure, it is often used to explore large conformational changes during transitions between various states. RMSD per residue was compared between MD and hybrid MD-Gō simulations, and a linear relationship was observed with high confidence for both RNA ($R^2 = 0.71$) and protein ($R^2 = 0.66$) (Figure 5.2). Particularly, it has been shown that N-terminal domain of S4 is highly disordered in solution [67], and the same behavior is seen in the hybrid MD-Gō simulations (red dots in Figure 5.2, top panel). Thus, the consistency of nucleotide flexibility between simulations and SHAPE, as well as between two kinds of simulations suggests that structure-based simulation with Gō-potential is able to recapitulate dynamics of molecules at atomic resolution.

5.3 Multiple Conformations of the Five-way Junction Near Native State

5.3.1 Conformational heterogeneity of the five-way junction observed with smFRET

The dynamics of individual molecules observed in real time complements the structural information from footprinting assays like SHAPE, and thus provides a more detailed picture of molecular interaction. We directly measured the dynamics of the five-way junction by single molecule FRET (smFRET) [204]. Fluctuations in the distance between h16 and h3 were measured by hybridizing Cy3 and Cy5 labeled primers to extensions in the RNA sequence (Figure 5.10). These extensions do not change the RNA secondary structure nor S4 binding affinity (S.A. and M.M., unpublished). Figure 5.3 A shows the Mg$^{2+}$ dependence of the RNA dynamics obtained from smFRET observations. The complex behavior of the unliganded RNA is demonstrated by multiple FRET levels emerging at different Mg$^{2+}$ concentrations. At 1 mM Mg$^{2+}$, the majority of the molecules fluctuate rapidly between low FRET states ($0.1 \sim 0.35$). The fluctuation appears in the histogram as a broad peak instead of two well-defined separated FRET levels. The fluctuations slow down and resolve into low and medium-low states at higher Mg$^{2+}$, as seen in the example trace for 5 mM Mg$^{2+}$. At the same time, a fraction of the molecules temporarily exhibits a medium-high FRET state (0.55), implying that the RNA transiently folds into a structure with smaller distance between h16 and h3. At 20 mM Mg$^{2+}$, the population in this mid-FRET state increases, and excursions to higher FRET levels become
Figure 5.2: Correlations between MD and hybrid MD-Gō simulations of native state fluctuations in the five-way junction:S4 complex. RMSD per residue is linearly correlated between MD and hybrid MD-Gō simulations for both the protein (top) and the RNA (bottom). Red dots in the top panel are RMSD of the S4 N-terminal residues which show higher flexibility in both MD and hybrid MD-Gō simulations.

visible, as illustrated in the sample trace in Figure 5.3 A. At 100 mM Mg$^{2+}$, this transition becomes more frequent and thus the high-FRET population increases.

The behaviors described above are not uniformly exhibited by all molecules under the same conditions. There is a high degree of molecule-to-molecule variation. As an example, Figure S2 in the Supporting Material depicts additional traces at 20 mM Mg$^{2+}$ showing that some molecules stay mostly in the low-FRET state, a small fraction resided in the medium-high FRET state, and yet others fluctuated quickly between the states. On rare occasions, we find molecules switching between different behaviors. These observations suggest a complex folding energy landscape of the RNA, which bears similarity to the rugged multidimensional folding energy landscape with multiple distinct conformational states previously reported
Figure 5.3: Multiple states are observed in sm-FRET experiment and MD simulations of the five-way junction. (A) FRET signal distributions of the five-way junction under varying Mg$^{2+}$ concentrations. Representative time traces are shown for each concentration to illustrate signal fluctuations over time. (B) Multi-dimensional scaling (MDS) reduction of the ten-dimensional angle vector into a two-dimensional space for visualization of the unfolding MD trajectory simulated in NaCl solution. The solid line shows the time trace with a low sampling rate. On the path, the red (+) marks the starting frame and the red (○) marks the ending frame. (C) Representative snapshots of the five-way junction are shown together with a proposed transition map among different states: N (native structure), I (intermediate state), II (unfolded state), and III (misfolded state). (D) FRET distance measured over time between U422 on the tetraloop of h16 and A559 of h3 for the five-way junction in the unfolding simulation in NaCl solution. (E) FRET distances measured over time in the Mg$^{2+}$ quenching simulations starting from corresponding conformational states (color-coded the same as in B).
for the *Tetrahymena* ribozyme [205; 206]. In the five-way junction system, we were able to detect a larger number of conformational states describing more complex folding behavior of the ribosomal RNA. Even then, the observed states might still be degenerate due to different configurations of the bound Mg$^{2+}$ ions or different conformations of the whole RNA molecule giving rise to similar distances by chance.

5.3.2 Identifying the metastable states in simulations

All-atom MD simulations of the five-way junction without S4 protein can help to identify meta-stable states corresponding to high and low FRET signals. It was shown in our previous study that the five helices in the five-way junction start to separate after 40 ns of simulation in NaCl buffer, while maintaining their secondary structure (except h3) [67]. It follows that conformational states of the unfolding are best described by the relative orientations among the constituent helices, which give rise to a ten-member angle vector (see Methods).

Accordingly, four conformational states of the five-way junction have been identified from the MD unfolding trajectory: the native structure (N), the intermediate state (I), the extended state (II) and the misfolded state (III) (Figure 5.3 B & C). State I is the closest to the native structure, though h17 is stretched and separated from h4. Contacts between h16 and h18 fluctuate, but the two helices remain close to each other. h3 has the largest fluctuations in secondary structure but still maintains most of its tertiary contacts with h18. State II is the most extended conformation; however, the five helices of the molecule keep moving within a plane so that inter-conversion between states I and II may occur occasionally. State III distinguishes itself from the other states by the loss of the planar conformation adopted in the native structure. Particularly, h16 and h3 move out of the plane while h17 and h18 move into the plane in such a way that a right angle between h18 and the co-axially stacked h16 & h17 is made. Interestingly, the relative positioning of h16 and h18 in state III is very similar to the conformation adopted in the eukaryotic ribosome structure [207; 208], in which the S4 N-terminus is nonhomologous to that in the bacterial ribosome [165]. The main difference among various conformational states lies in the relative positioning of h16 to h18. It is shown in the Supporting Material that the distribution of angle between h16 and h18 has three peaks, corresponding to state I (and the native structure N), II and III, respectively.

To explore the stability of each identified conformation, further MD simulations were carried out. Representative conformations of states II, III, as well as the native structure were taken as the starting conformations, and Mg$^{2+}$ ions were placed using protocol developed in Eargle *et al.* [93]. The same clustering analysis was applied to these magnesium quenching simulations, and the results showed that both states N and III would stay in their local minima within the 100 nanoseconds of simulation. State II, however, was shown to be inter-changeable with state I during the simulation period. Importantly, clustering using either the ten-dimensional angle vector or the distribution of the angles between h16 and h18 showed similar results. This consistency granted us the simplicity to use angle between h16 and h18 as the single criterion in determining conformational states of the five-way junction (see Supporting Material).
5.3.3 Correlating atomic structures of the five-way junction conformations with the FRET states

Having clarified all major conformers of the five-way junction in MD simulations, we examined their correlations with the FRET states observed in experiments (Figure 5.3 A). Due to the uncertainty of both FRET labels on the five-way junction posed by large extension in RNA structure (especially for the position of Cy3, see Figure 5.10), relative distance fluctuations are considered more critical in correlating with the FRET states rather than the absolute distance values. In the unfolding simulation of the five-way junction, distance between the FRET labels on h16 and h3 oscillated over a large range (Figure 5.3 D), which would correspond to both high and low FRET efficiency values. However, once Mg\(^{2+}\) ions were added, the near-native structure and state III are both locked in the small-distance states (Figure 5.3 E). Distance between labeled positions on h3 and h16 occasionally reach lower than 50 Å, which could be consistent to the medium-high FRET signals seen in traces obtained under 20mM and 100mM of Mg\(^{2+}\) concentrations. The fact that distance between FRET labels fluctuated in the same range in simulations starting from both the native structure and the misfolded conformation, indicates that the medium-high FRET population, though very small, could still consist of a mixture of conformational states. On the other hand, distance between FRET labels in the simulation starting with the extended state II has fluctuations as much as 40 Å, and the labeled positions seldom get closer than 70 Å. The data demonstrates flexibility of the extended state and is consistent with the observation in experiment that different RNA molecules would oscillate between different low-FRET states. Since the average behavior of hundreds of molecules shows only one peak in the FRET efficiency histogram, the various FRET signal levels should be best described as sub-states within the extended state II of the five-way junction. Further experiments where the S4 protein is included are required to precisely correlate FRET efficiencies with conformations identified in simulations.

5.4 Folding Landscape of the Five-way Junction

5.4.1 Simultaneous folding and binding of the five-way junction and S4

We employed the structure-based Gō model to study the simultaneous folding and binding of the five-way junction and S4. Two conformations of the five-way junction, state II (extended) and III (mis-folded or non-planar) from MD simulation, were considered as initial conformations in the simulations. These two distinct starting conformations share some, though definitely not all, of the changes in structure upon S4 binding. We present here the shared features of the binding process starting from these two different RNA conformations, differences will be discussed in the next section. Results of simulations starting from state II were used to illustrate the common theme in the five-way junction:S4 binding process. Additional details of the choice of initial conformations are provided in the Methods section.

Within the designed number of steps, the five-way junction and S4 bound in 61 out of the 100 replicates starting from state II and in 72 simulations starting from conformation III. To describe the binding process, we first monitored the center of mass (CoM) distances between different domains of S4 (see Methods for domain definition) and their binding sites on the five-way junction. The results for all bound replicates from simulations starting from state II are shown in Figure 5.4 A. From the histogram, we see two phases of the binding process — a diffusing phase and a binding phase. In the diffusing phase, the two domains of
S4 randomly explore the configuration space, each trying to reach into contacts with the five-way junction independently. However, when the COM distance gets below 25 Å between S4N and h16 or below 30 Å between S4 C2-domain and the five-way junction, the binding phase begins. The binding phase consists of two major pathways. In one pathway, the N-terminal domain of S4 binds first, followed by the binding of the C2-domain. In the other, the order is reversed. While binding initiated from the C2-domain seems dominant from the histogram, its diffusion might be relatively free compared to the S4N. As long as the S4N remains close to the five-way junction, the C2-domain can diffuse as far away as 45 Å and the whole protein may still be able to bind.

Figure 5.4: S4 binds to the five-way junction in multiple steps. (A) Histogram of the center-of-mass distance between S4N and h16 versus that between S4-C2 domain and the five-way junction shows two distinctive phases in the binding process: the diffusing phase and the binding phase. (B) Histogram of the fraction of native interface interactions between S4N and h16 versus that between S4-C2 domain and the five-way junction. Three representative pathways are shown in yellow, red and green from left to right. (C) Correlation between the conformations of protein and RNA is visualized by histograms as a function of RMSD of the protein N-terminus (top) or C2-domain (bottom) and $R_G$ of the corresponding binding site on the RNA. (D)–(F) Representative time traces of the fraction of native contacts between h16 and h18 (shown in light blue), and the fraction of native interface contact between S4N and the five-way junction (shown in red).

5.4.2 Binding pathway of S4

To take a closer look at the “binding phase”, we calculated the fraction of native interface contacts (interface Q) between the N- or C2-domains of S4 protein and the five-way junction. When starting from state II of the five-way junction with the two domains of S4 put at equivalent distances from their binding sites, the first
contact to be made is, for the majority of the bound replicates (56 out of 61), between the S4 N-terminus and h16. However, this first contact between S4N and h16 does not always lead to complete binding. Starting from state II, for example, it takes another $5.5 \times 10^5$ time steps (one fourth of total simulation time) on average before the S4 C2-domain makes contact with the five-way junction. After both domains of S4 establish their initial interactions with the five-way junction, the binding phase starts from the left bottom corner of the histogram (Figure 5.4 B) where both domains have interface Q values close to zero. One dominant binding pathway can be seen clearly from the histogram, in which the C2-domain binds first, and S4 N-terminus follows with two distinct barriers (yellow). The barrier located at interface Q $\sim 0.15$ represents a binding intermediate where the disordered coil in the S4 N-terminus interacts with the 5′ strand of h16 (A408 – A412). This region faces the protein interface in the extended conformation state II, and it is close to the junction where the C2-domain binds. In order for the binding to continue, a slight rotation in h16 is needed then in order for the S4 N-terminus to make further contacts with both strands of the helix in the bulge region. A previous study [67] has shown that the h16 internal loop is relatively flexible compared to the helical regions. Therefore, the requirement for simultaneous rearrangements in the internal loop of h16 and the S4 N-terminal coil gives rise to the second barrier at Q $\sim 0.4$.

In addition to the dominant one, two other pathways exist in which the N-terminus of S4 initiates binding. In one path, S4 N-terminus binds completely before the C-terminal domain starts to make contacts (green). In the other path, only 30% of the native contacts between the S4 N-terminus and the five-way junction are made before the C-terminus begins to bind (pink). The ratio of the number of molecules going through each of the three pathways is about 3:1:1 in simulations starting from state II. The exact partition changes according to initial conformation (data not shown) while the positions of the three pathways as well as the energetic barriers remain the same regardless of the starting conformations.

In summary, the two domains of S4 behave differently. The C-terminal domain of S4 binds quickly and tightly once the right contacts are made. Conversely, complete binding of S4 N-terminus is the rate-limiting step in the binding phase due to two energetic barriers arising from structural rearrangements in both h16 and S4N. The reason for their different functions in binding lies in the distinct structural properties of both the protein and the RNA binding sites. The C-terminal domain of S4 binds to the central junction of the five helices, while the S4 N-terminus binds to both h16 and h18, which may separate widely (as in extended state II) when S4 is not bound. Figure 5.4C shows that the S4 C-terminal domain and its binding site are both stable, fluctuating independently through a relatively small range of conformations. However, conformations of the S4 N-terminus and its binding site are clearly correlated. In the diffusing phase, both of them are flexible, with RMSD of S4N fluctuating between 4 to 8 Å and $R_G$ of h16 and h18 fluctuating between 17 to 21 Å. Only after binding occurs are both constrained close to their native conformations. This suggests that the protein-RNA interactions have co-evolved such that their structural features match to facilitate binding.

Since the S4 C-terminal domain is stable throughout the diffusing and binding phases, and no apparent energy barrier is present for its binding to the five-way junction, we suggest that the binding of S4 C-terminus is mainly determined by the search in entropic space. Once it is oriented correctly with respect to its binding site, all the contacts are made at once and binding occurs within a short time. The binding of N-terminus
needs more investigation. It has been suggested in [67] that the main purpose of the S4 N-terminus is to bring the separated h16 and h18 back together into their native configuration. Interactions between h16 and h18 can be characterized by interface Q between the two helices. There is intrinsic fluctuation between h16 and h18, and they may make and lose contacts repetitively when the S4 N-terminus is not bound (Figure 5.4 D). This is consistent with the observation from MD simulations that states I and II interchange spontaneously. Once the S4 N-terminus is tightly bound, fluctuations between the two helices are suppressed, and their interactions remain until the end of the simulations (Figure 5.4 E). In general, the S4 N-terminus initiates the interaction between h16 and h18. However, there are also cases where it binds to one of the states while h16 and h18 are fluctuating (Figure 5.4 F). Supposedly, if the N-terminus of S4 catches the five-way junction in state I in which h16 and h18 are close together, the binding barriers should be eliminated. However, this circumstance is rarely seen in the simulations.

5.4.3 Binding mechanism of the signature S4 N-terminus and the five-way junction

We consider two possible scenarios for the mechanism of S4 N-terminus binding — population shift or induced fit. Results in the previous section suggest that conformations of the S4 N-terminus and its binding site on the five-way junction are closely related, and it is the interactions between them that correlate with the conformational changes. In order to determine the binding mechanism, we plot the RMSD of S4 N-terminus against its interface Q with the five-way junction. A population shift scheme (Figure 5.5A) results in two separate conformational populations existing regardless of whether the interactions between the S4N and the five-way junction are present or not. However, the increase in interface interactions should result in a change in the relative amount of molecules stay in each of the two conformational states. In contrast, binding achieved through induced fit (Figure 5.5B) predicts two distinct conformations populated under different conditions (no interaction, or bound).

![Figure 5.5: S4N binding is via induced fit. Schematic representations of changes in conformational densities results in (A) population shift or (B) induced fit mechanisms. (C) 2D Histogram of the folding of S4N (RMSD) as a function of its docking to the five-way junction as measured by the fraction of native interface contacts between them. The histogram resembles the induced fit schematic in that the minor component in the unfolded population is absent.](image)

Our simulations show that the S4 N-terminus clearly adopts an induced fit mechanism as it remains flexible before any interaction with the five-way junction is established (Figure 5.5C). Only upon binding to
the five-way junction (interface $Q > 0.5$) will the N-terminal coil fold into a stable conformation even when a bias toward the native state is present in the Gō-potential model.

The binding mechanism is more complicated for the five-way junction, and the interpretation strongly depends on the starting conformation. In simulations starting from the extended conformation of the five-way junction, state II, conformational states were measured using radius of gyration ($R_G$) of h16 and h18. Two populations existed in the absence of S4N interaction (Figure 5.6A), one with $R_G \sim 17$ Å close to that in the native structure (state N or I), the other with a higher and fluctuating $R_G$ (state II). This is consistent with the occasional interconversions between state I and II seen in the MD simulations of the five-way junction without S4, and with the hybrid MD-Gō simulations where h16 and h18 fluctuate between states with or without interactions between each other. After S4 binds, density for the conformation with large $R_G$ disappears completely. Binding of S4N is able to begin from either state, as shown with representative traces (yellow from state I, and orange from state II) on the histogram (Figure 5.6A). Interestingly, binding from state II occurs in a step-wise manner, alternately establishing interface interactions and increasing compactness of the h16-h18 entity. In this case, conformations of the five-way junction (particularly, h16 and h18) are distributed in two populations intrinsically, and the interactions from S4 eliminate one of them in an extreme example of population shift.

On the other hand, folding of the five-way junction starting from conformation III is a clear example of the induced fit mechanism (Figure 5.6B). Due to stability of conformation III seen in the MD simulations, it is expected that h16 and h18 should stay at a right angle without S4 bound. Therefore, we see only one population centered at $65^\circ$ when $Q < 0.05$. Folding of h16 and h18 back to the native structure will not happen before a substantial fraction of the S4 N-terminus interactions are established ($Q > 0.4$). This fact supports our previous finding that state III is a relatively stable conformation of the five-way junction that will only fold correctly with the help of S4.

5.4.4 Inferring the folding landscape of the five-way junction

Folding of an RNA molecule can be qualitatively summarized by its folding energy landscape, which shows the positions of alternative conformations (including the native state, unfolded state, and all intermediates) as well as the energetic connections between them [205]. In addition, structural features that define the intermediate states and the energy barriers between them will further characterize folding behaviors along the pathways. Folding of the RNA molecules, in particular, is known to be hierarchical [209; 210], such that secondary and tertiary structures fold sequentially due to the large difference in their energetic contributions. Folding of the secondary structure is characterized by large negative $\Delta G$ and is sensitive to kinetic traps [209], while forming tertiary contacts is highly cooperative. However, recent studies on the folding of riboswitches and ribozymes showed the non-hierarchical nature of RNA folding and the existence of multiple collapsed and stable non-native structures in the native basin on the energy landscape [206; 211; 212].

In this chapter, we assumed that the secondary structure of the five-way junction forms during the co-transcriptional self-assembly of the rRNA. Based on this assumption, we tried to obtain information about the topology of its folding landscape in the vicinity of the native basin. Since formation of RNA tertiary structure is sensitive to the amount of Mg$^{2+}$ in solution, we first depleted all Mg$^{2+}$ ions in the system and
Figure 5.6: Five-way junction binding mechanism depends on the starting conformation. Histograms showing conformational changes in the five-way junction upon docking with S4N. Conformations of the five-way junction are measured using radius of gyration ($R_G$) of h16 and h18 when simulations are started from state II (A) and using angles between h16 and h18 when initiated from the state III (B). In panel (A), a population shift mechanism is seen since one of the two existing populations is eliminated by the binding of S4N. In panel (B), conformational change of the five-way junction occurs only after the binding of S4N which induces a fit to the native conformation.

ran MD simulations which showed that the five-way junction unfolds in a way that tertiary contacts were disrupted while secondary structure was maintained (Figure 5.7, top left). The unfolding trajectory was analyzed, and conformations in the trajectory grouped into discrete states according to the angles between helices (a measure of tertiary interactions with respect to the native structure). In order to check whether these discrete clusters of the five-way junction states correspond to stable intermediates during the final folding process, Mg$^{2+}$ ions were put back into the simulations starting from representative conformations chosen from each of these groups (Figure 5.7, bottom left).

We found, similar to the presence of multiple native states of the *Tetrahymena* group I ribozyme [206], that multiple stable states existed for the five-way junction alone in solution in addition to its native state.
Two conformations (states I and II) are separated by a relatively small barrier, inferred from their rapid interconversion. Second, a third intermediate (state III) with a free energy even lower than that of the native state, is probably separated from the native state by a large barrier because of its apparent stability and the difficulty folding back to the native structure on its own. These observations from MD simulations, which were also seen in the FRET experiments, gave rise to the proposed energy landscape depicted in Figure 5.7 (bottom left). Furthermore, visualizing atomic details in MD simulations provided us with a straightforward way to recognize structural characteristics of the different states. Interestingly, identification of the multiple stable conformations can be simply reduced to a low dimensional metric that involves only the angle between h16 and h18.

In vitro binding of S4 to the five-way junction and rearrangement to the final stable complex takes a few seconds to minutes to complete [176], which is a time scale currently inaccessible to MD simulation. To understand what role the protein plays in the folding of the five-way junction, simultaneous binding and folding of the two molecules were performed using Gō-potentials developed in the framework of the NAMD package (see Methods). States II and III were chosen as the initial unfolded conformations of the five-way junction. Considering that states I and II were observed to interconvert spontaneously, we viewed the protein-assisted folding process to start from all possible stable points on the energy landscape (Figure 5.7 middle column). By adding in the r-protein S4, the folding landscape was altered in terms of relative stability among the different states and the barrier height between them. Apparently, the binding of S4 greatly stabilizes the native structure of the five-way junction, since no fluctuations back to the three non-native states were ever observed after S4 was correctly bound. Folding from state III was now possible.
with the help of S4, showed by a decreased barrier between it and the native structure. Most likely, this barrier is even smaller than that of the other folding pathway, due to the higher probability of successful binding events observed in multiple replicates of the binding simulations. The folding pathway that went through states I and II downhill to the native state had been smoothened, however, the shape of the landscape was preserved which is consistent with the presence of two barriers seen in the S4N binding process starting from state II.

S4N was shown to be intrinsically disordered by many experiments as well as MD simulations, and consequently it was suggested that the flexibility of S4N is utilized by the protein to speed up the molecular recognition between the five-way junction and S4 [67]. Here, we examined this possibility by visualizing the binding process using hybrid MD-Gō simulations. The flexibility of S4N facilitates the binding process in two ways. First, by closely working with the globular domain of S4C which serves as an anchoring point, it decreases the five-way junction folding barriers at the cost of its own entropy loss upon binding. Second, because of its flexibility, it is capable of binding to all possible conformers of the five-way junction, and, thus, may increase the probability of successful binding. After the initial binding, S4N alters its own conformation accordingly, and induces both the folding of five-way junction and the complete docking between the two molecules (Figure 5.7 right column). This mechanism is in agreement with the extended view of biomolecular recognition, in which binding of the biomolecules is initiated by selection among a dynamical conformational ensemble followed by subsequent adjustments by induced fit [213; 214].

5.4.5 Concluding remarks

Simultaneous folding and binding of the five-way junction in the 5′-domain of 16S rRNA and the primary binding protein S4 is essential for the assembly of the bacterial ribosomal SSU. This process is studied in this paper using the hybrid MD-Gō simulations developed within the framework of all-atom molecular dynamics CHARMM force field. With the support from all-atom MD simulations, single-molecule FRET experiment and SHAPE chemistry, we have identified multiple metastable states of the five-way junction near its native basin and estimated the shape of its folding landscape. Analysis of the five-way junction-S4 interactions from hundreds of hybrid MD-Gō replicate simulations shows that the intrinsic disorder of the N-terminus of S4 controls the rate of binding in two ways. First, it is capable of binding to all different conformations of the five-way junction during the self-assembly process of the RNA molecule. Second, it induces the folding of the five-way junction to the native state. Our simulations capture the cooperativity in the simultaneous folding and binding of the five-way junction and S4, and shed further light on the general principle of protein:RNA interactions and evolution.
5.5 Future Studies

5.5.1 S4 modulated 5′ domain rRNA dynamics

The assembly of 30S ribosomal subunit requires the coordinated addition of 20 proteins to the 16S ribosomal RNA. Although the hierarchy of protein addition is well established, how proteins induce structural changes in the rRNA to produce highly cooperative assembly is little understood. Here, we use single molecule fluorescence resonance energy transfer (FRET) to observe the real-time dynamics between a primary binding ribosomal protein S4 and the 5′ domain of 16S rRNA, at an early stage of 30S assembly. Using data obtained from the S4:rRNA binding complex at equilibrium and time-resolved FRET measurements upon addition of S4, a kinetic model of the S4:rRNA binding pathway was constructed (Figure 5.8 A). In this model, S4 and the 5′ domain RNA initially form encounter complexes (EC, rapidly fluctuating FRET) characterized by rapid fluctuations of h3 and possibly other RNA helices. The EC diffuses into the flipped intermediate (FI, low FRET), in which h3 is undocked from S4 and h18, consistent with previous footprinting data [215]. Finally, h3 docks against S4 and the h18 pseudoknot to form the native complex (NC, high FRET).

Three-color smFRET between h3, h16 and S4 was measured in addition to see how S4 modulates the motions of the RNA helices. As a result, a geometric model was constructed according to the observed dynamics (Figure 5.8 B). In the free RNA, the heterogeneous fluctuations of the distance between h3 and h16 suggest that the RNA helices may move in several directions with varied frequency. When S4 binds, the motion of h16 is restricted, consistent with the many direct contacts it makes with S4. By contrast, h3 continues to fluctuate between flipped and docked states in a plane such that its distance to h16 does not significantly change. To gain further insight into the molecular details of the S4-modulated RNA dynamics, we analyzed the all-atom MD trajectory of the minimal five-way junction RNA, and pictorialized the global motions of the helices by representing them with cylinders and tracing the movements of their tips. Without S4, the h16 trajectory explored two distinct regions, while h3 spanned a broad range in space without any specific direction (Figure 5.8 C), supporting the heterogeneous fluctuations between h3 and h16 of the free rRNA observed in FRET experiments. The binding of S4 quenched the movement of h16, allowing it to sample only a small region in space, consistent with the stable S4-h16 FRET signal (Figure 5.8 D).

More interestingly, S4 restrained h3 to an in-plane bending motion toward S4, maintaining a nearly constant distance between h3 and h16. These S4-induced changes in the helix motions agreed well with the experimentally observed dynamics of S4-h3-h16 distances. Therefore, both experiments and simulations showed that the RNA remains dynamic after S4 binding, but the large-scale motions of the helices are confined in their direction.

Beyond equilibrium dynamics, the simultaneous folding and binding between S4 and the five-way junction was successfully reproduced by hybrid MD-Gō simulations. Comparing with the FRET signal observed in the non-equilibrium FRET experiments (Figure 5.8 E), the simulations captured not only the transition from low to high FRET states of h3, but also the fluctuating encounter conformation at mid-FRET (Figure 5.8 F). In the majority of the successful binding trajectories, initial contact established with the S4...
C-terminal domain quickly reduced motions in the five-way junction, restricting the movement of h16 and allowing it to interact with the S4 N-terminal domain. It follows that the S4 N-terminal domain became increasingly ordered, while h16 adjusts its conformation toward h18 into a more native-like structure. It is also suggested that the two domains of S4 may each favor a different conformation of the five-way junction, and they work cooperatively together to induce efficient folding of the RNA molecule.

5.5.2 Atomic resolution characterization of various components in the kinetic model of the S4:rRNA binding pathway

Detailed analysis over the MD trajectories provides further insight into the characteristics that distinguish among the structural intermediates along the kinetic S4:rRNA binding pathway. Particularly, it is known that a right-angle (RA) motif is present in the junction linking h3, h4 and h18, which aligns the three helices at right angles with respect to each other (Figure 5.9 A top). In the crystal structure, the RA motif consists of base stacking between adenosines A547 and A499, both of which are orientated perpendicular to base-pairings in the h18 stem (Figure 5.9 A bottom). A couple of GA-minor interactions in this area are supposed
to help maintain the motif, but appear to be unstable in simulations of the five-way junction both with and without S4. Instead, the interactions between U405 and the adenosine-stacking are seen to play an important role in determining the stability of the RA motif. In the simulation without S4, the A547/A499 stacking was maintained perpendicular to h18 with additional Hoogsteen-edge AU base pair formed between A547 and U405 (Figure 5.9 B). This base pair stayed for over 100 ns, and once broken, disrupted the orientation of the RA motif. As a result, the base stacking across A547, A499, and A546 now appeared as a continuation of the h18 stem (Figure 5.9 C). In the simulation with S4, however, A547 flipped over to the other side of the backbone. The RA motif was then formed by the alternative adenosine stacking between A499 and A546 (Figure 5.9 D), which was kept again at a right angle to the base pairs in the h18 stem due to a similar Hoogsteen-edge base pair between U405 and A499. This conformation stayed until the end of the 150 ns simulation.

Figure 5.9: RA motif characterization of the 5′ domain structural intermediates. (A) Crystal structure RA motif in the h3-h4-h18 junction. (B) & (C) Snapshots taken from MD simulation of the naked five-way junction, showing representative structures of the tentative “FI” and “EC” intermediates. (D) Snapshots taken from MD simulation of the five-way junction with S4, showing alternative structure for the “FI” conformation observed in the simulation. (E) FRET histograms at different Mg^{2+} concentrations or mutant rRNAs.

We tentatively assign the conformations shown in Figure 5.9 B & D to the flipped intermediate (FI), because the right angle between base pairs in the h18 stem and the adenosine stacking potentially restricts the movement of h3 in a certain direction — likely perpendicular to h18 and h16 (Figure 5.8 D). Following the same argument, the conformation seen in Figure 5.9 C might be assigned to the encounter complex (EC), due to the loss of the right angle. In this unconstraint state, h3 should be able to move freely in any direction (Figure 5.8 C). This hypothesized assignment is tested with FRET measurements using different
mutant RNA molecules. According to the observations in MD simulations, one would expect the mutation of A499U to weaken the base-stacking propensity and lead to greater local flexibility in the RNA, and therefore shift the equilibrium away from the native complex. This is exactly what we see in the FRET histogram for mutant A499U, where the high FRET population (NC) is almost entirely eliminated (Figure 5.9 E bottom left). On the other hand, mutation of A547 would have less influence on the FRET histogram (Figure 5.9 E bottom right), because an alternative RA motif may be established between A499 and A546 to rescue the perpendicular motion of h3 as in the FI.

In summary, the all-atom and coarse-grained MD simulations combined with experimental measurements are able to explore the complex landscape of RNA folding under the influence of protein binding. The methodology established in this chapter will be applied to larger scale rRNA folding with multiple proteins binding simultaneously. Our goal is to reconstruct a full picture of the ribosomal assembly, which will guide not only researches in the ribosomal assembly and evolution, but also dynamics of the RNA:protein complex in general.

5.6 Methods

5.6.1 RNA preparation

In the fluorescence resonance energy transfer (FRET) experiments, rRNA containing the E. coli 16S rRNA (nts 21–46, 395–556) was used to construct the minimal five-way junction according to Bellur et al. [149], with the sequence CUCAA capping the gap between nucleotide G46 and C395. Two extensions were added to target h16 and h3 for fluorescent labeling. First, a 26-nucleotide insertion (5′-CGCGUCGCCAGACCAGACGCUCGCCGCG-3′) was placed between C419 and G424, which extends h16 by 4 GC pairs for higher stability. Then a 15-nucleotide RNA primer labeled with the Cy3 dye (5′-GAGCGUCUGGUCUGG-Cy3-3′) (Integrated DNA Technologies, Inc) was hybridized to the h16 insertion from the 8th to the 22nd position. The 3′ end of h3 (C556) was extended by 38 nucleotides (5′-AGGACGACACACUUUGGACAGGACACA-CAGGACACAGG-3′) with a 37-nucleotide complementary DNA strand labeled with Cy5 and biotin (5′-biotin-CCTGT GTCCCTGTGTGTCCTGTCCAAAGTGTGTCGTCC-Cy5-3′) (Integrated DNA Technologies, Inc).

Molecular models for the simulations were created using crystal structure of the five-way junction and S4 from the E. coli ribosomal SSU (PDB code: 2I2P). The five-way junction was modeled according to the minimal S4 binding system suggested by Bellur et al. [149]. A five-membered loop (CUCAA) was added to cap the end of h4 producing a single, continuous RNA molecule, and h17 which normally extends away from the S4 binding site was truncated and capped with a seven-membered loop (UUUUGCU). These two additional loops were taken from the E. coli SSU (PDB ID 2I2P, residues 618–622) and the anticodon loop of tRNA^Gln (PDB ID 2RD2, residue 932–938) with two mutations (C934U and A937C), respectively.

Results from the selective 2′-hydroxyl acylation analyzed by primer extension (SHAPE) experiment performed on the five-way junction RNA [176] were used to compare with simulation data in the paper. The 3′ end of the native five-way junction was extended with the sequence 5′-aaagcaatatagactcgggatgaggaCATCA-TGGCCCTTACGA-3′ for primer extension in the SHAPE experiment. The 3′ extension is shown not to
5.6.2 FRET measurement

The RNA was immobilized on the surface with the biotin on the h3 primer, using the same method as previously reported in Hohng et al [216]. Single molecule FRET traces were measured using a TIR-based fluorescence microscope and analyzed using conventional methods [216]. The imaging solution contained 80 mM K-HEPES at pH 7.6, 300 mM KCl, the desired amount of MgCl$_2$, saturated Trolox of 2 mM, 0.6% glucose, 0.1 mg/ml glucose oxidase (Sigma), and 0.02 mg/ml catalase (Sigma). For the data analysis, each FRET histogram was built by accumulating 100 ms data bins of 300–500 molecules.

5.6.3 Molecular dynamics simulations

The five-way junction and the r-protein S4 were neutralized with Na$^+$ and/or Mg$^{2+}$ placed by Ionize [94], and hydrogens were added to the system by psfgen in VMD [89]. Water molecules were placed to complete the primary solvation shells for all Mg$^{2+}$ atoms. 400 mM NaCl was added to the final solution to match the corresponding experimental conditions. Solvent was prepared in two phases: first Solvate [95] was run to place the first two solvation layers and then the solvate plugin to VMD was used to complete the water box.

MD simulations were run using NAMD2 [96] with the CHARMM27 force field [97]. Stepwise minimization and equilibration were performed according to the protein:RNA simulation protocol established in Eargle et al [93]. Production runs used periodic boundary conditions and the NPT ensemble with pressure set to 1 atmosphere and temperature set to 298 K. Multiple time-stepping was used to calculate bonded interactions at 1 fs, van der Waals (vdW) interactions every 2 fs, and electrostatic forces every 4 fs. Particle mesh ewald summation was used to evaluate electrostatic interactions, and the vdW force calculations used a cutoff of 12 Å and a switching distance of 10 Å.

A total of 400 nanoseconds of MD simulations on the naked five-way junction are presented in this chapter, including 100 ns of the unfolding simulations without Mg$^{2+}$ ion; 100 ns of equilibrium simulation
from the crystal structure with Mg$^{2+}$ ions; 100 ns refolding simulations with Mg$^{2+}$ from each of the two conformations selected from the unfolding trajectory (see Results and Discussion). In all three simulations with Mg$^{2+}$, 25 Mg$^{2+}$ ions were placed according to the electrostatics of the five-way junction starting conformation, among which only 4 were placed in direct contact with the RNA molecule. R-protein S4 was not included in these MD simulations. In addition, to compare nucleotide flexibility in simulation with the SHAPE reactivity, one 100 ns MD simulation of the five-way junction bound with S4 was performed.

5.6.4 The structure-based Gō model

In our heavy-atom hybrid MD-Gō models, all bonded terms (including bonds, angles, dihedrals and impropers) are described by the transferrable CHARMM27 force field [97] with additional improper angle potentials to maintain chirality (see Supporting Material). All non-bonded potentials (vdW and electrostatics) were replaced by the knowledge based Gō-potential. The Gō-potential is formulated with respect to a reference structure, which can either be the equilibrated native structure for the folding of RNA/protein or a target conformation that is important for the function of the biomolecules. For atom pairs closer than 4 Å within a molecular chain or pairs closer than 4.4 Å between the RNA and protein chains in the reference structure, the pair is defined as a native contact, and is subject to a Lennard-Jones-style potential:

$$V_{\text{native}}(\sigma_{ij}^{\text{native}}, r_{ij}) = 4\epsilon_{\text{native}} \left( \left( \frac{\sigma_{ij}^{\text{native}}}{r_{ij}} \right)^a - \left( \frac{\sigma_{ij}^{\text{native}}}{r_{ij}} \right)^b \right)$$

(5.1)

where $\sigma_{ij}^{\text{native}}$ is determined by the native pairwise distance, $r_{ij}^{\text{native}}$, between atoms $i$ and $j$ in the reference structure:

$$\sigma_{ij}^{\text{native}} = \left( \frac{b}{a} \right) \frac{1}{\epsilon_{\text{native}}} r_{ij}^{\text{native}}$$

(5.2)

such that the potential reaches the minimum at $\sigma_{ij}^{\text{native}}$. $r_{ij}$ is the instantaneous distance between atoms $i$ and $j$ in the simulation, $\epsilon_{\text{native}}$ specifies the depth of the potential, and the exponents $-a$, $b$ – determine the shape of the potential. All non-native atom pairs with a distance longer than 4 Å in the reference structure experience a repulsive potential of the form:

$$V_{\text{non-native}}(r_{ij}) = \epsilon_{\text{non-native}} \left( \frac{\sigma_{ij}^{\text{non-native}}}{r_{ij}} \right)^c$$

(5.3)

where $\epsilon_{\text{non-native}}$ scales the strength of the potential, $\sigma_{ij}^{\text{non-native}}$ determines where the potential equals $\epsilon_{\text{non-native}}$, and the exponent, $c$, determines the long range decay.

The above described hybrid MD-Gō model is implemented within NAMMD version 2.9, and our implementation allows for the study of complex systems in which molecules with different properties are present at the same time. It is designed so that each molecule may have its own parameters set independently of the others. In the particular system that we are studying in this chapter, we set $a = 12$ and $b = c = 6$ for the folding of both protein and nucleic acid. The interaction strengths $\epsilon_{\text{native}}$ are set to 0.1 kcal·mol$^{-1}$ and 0.23 kcal·mol$^{-1}$ for RNA and protein, respectively, to ensure that S4 and the five-way junction have similar
transition temperatures. All contacts within protein and RNA, irrespective of being involved in sidechain-sidechain, sidechain-backbone, backbone-backbone, sugar-sugar, sugar-base, or base-base interactions, are treated equally. To facilitate binding initiated from large separations, protein-nucleic interaction potential is set with exponents $a = 8$ and $b = 4$ and $\epsilon_{\text{native}} = 0.15$ kcal·mol$^{-1}$. All non-native parameters are set to $\epsilon_{\text{nonnative}} = 0.01$ kcal·mol$^{-1}$, and $\sigma_{\text{nonnative}} = 2.5$ Å in this study.

5.6.5 Hybrid MD-Gō folding simulations

The hybrid MD-Gō simulations were prepared without water molecules or ions, and only the heavy atoms of the five-way junction and S4 protein were used. Two representative states of the partially unfolded five-way junction (discussed in Results and Discussion) were chosen from the all-atom MD simulations as the starting conformations for the simultaneous folding and binding simulations. Given the high degree of secondary structure stability of S4 especially in its globular C-terminal domain [217; 218], we chose a slightly unfolded structure with an overall RMSD of $\sim$5.2 Å in which the disordered S4 N-terminus has a RMSD of 7.2 Å and the globular S4 C-terminal domain has a RMSD of 4.5 Å. The unfolded S4 and five-way junction were placed $\sim$38 Å (center-of-mass distance) away from each other, such that the closest contacts between them are just under the non-bonded interaction cutoff (12 Å). A 2 fs time step was used for all hybrid MD-Gō simulations. Each simulation was run for 2,250,000 steps, such that if the five-way junction and S4 were not bound within this time, they were likely to diffuse away from each other. 100 replicates starting from each chosen five-way junction conformation were performed with parameters introduced in the previous section, and statistics were generated. Supporting Material contains a table summarizing all the simulations presented in this chapter.

5.6.6 Analysis of the simulated trajectories

The angle between RNA helices is calculated as the angle between the principal axes of inertia extending through each helix, and the resulting angle value is translated into the range $[0^\circ, 180^\circ]$. With this definition, we measured angles between every pair of helices in the five-way junction such that a ten-member angle vector can be obtained for every frame to depict the five-way junction conformation captured during simulation. Pairwise distances between frames were then calculated based on these angle vectors to get a distance matrix. This matrix was processed using affinity propagation [219] for clustering of the five-way junction conformations. The same matrix was entered into the multi-dimensional scaling method [220] to reduce data dimensionality and to visualize all frames from the trajectories as coordinates in a two-dimensional space.

Three domains in the r-protein S4 are defined in order to characterize the binding behavior in detail. The S4 N-terminal domain (S4N), which is unresolved in solution NMR experiments [157] and fully structured [91] but still highly fluxional [67] upon interaction with the five-way junction in the crystal structure, is taken from residue 1 to 40. The C-terminal domain of S4 (S4C) is split into two sub-domains: the C1-domain (residue 41 to 96 and 185 to 205) interacts with the lower junction between h18 and h4, while the C2-domain (residue 97 to 184) has contacts with the upper junction between h16 and h17. We focus our analysis on the
S4N and S4 C2-domain since these two make the most interactions with the five-way junction. The five-way
ejunction binding sites for these three domains are defined to include nucleotides that are within 5 Å of each
domain, as well as any nucleotides base-pairing to them. According to this definition, the number of nu-
cleotides involved in the binding sites are 45, 20 and 28 for the S4N, S4 C1- and S4 C2-domain, respectively.

**Number of native interface contacts** is defined using a cutoff distance of 4.4 Å. This results in 660 native
interface contacts between S4 and the rRNA five-way junction, among which 320 are contacts between the
S4N and the five-way junction. The C1- and C2-domains of S4 have 98 and 242 native contacts with the
five-way junction, respectively. During simulation, if the distance between the two atoms forming a native
contact gets within 1.2 times the distance in the reference structure, the corresponding native contact is said
to be formed. Q is defined as the fraction of all native contacts that is formed in one particular frame.

**Radius of gyration** \((R_G)\) is a measure of how unfolded (or more precisely, extended) the molecule is.
Since it is independent of any pre-alignment of the molecule to remove the translational and rotational de-
grees of freedom, it can be useful to describe the conformational state of the molecule in addition to the
conventional RMSD value. Furthermore, in the case of h16 and h18 of the five-way junction, \(R_G\) gives a
more sensitive description of the degree of extension of the structure, while it is positively correlated with
the angle between the two stems so that the conformational assignment remains consistent.

**Trajectories of the RNA helix motions** were calculated as follows: first, the RNA helices were reduced to
cylinders and the end point of a vector along the center of each cylinder traced. Then we aligned the MD
trajectories to the crystal structure using heavy atoms in the stable helices 4, 17, and 18, so that the align-
ment remained mostly invariant among different simulations. For each frame in the aligned trajectories, we
defined the orientation of a helix by the principal axis of inertia for all heavy atoms on the helix (purple
arrows in Figure 5.8 C, D). The dimensions (length and radius) of each helix represent the largest projection
of the backbone atoms onto each of the three principal axes of inertia. Principal axes of momentum were
calculated similarly to the calculation of helix orientation.
Bibliography


[41] Koonin, EV (2003) Comparative genomics, minimal gene-sets and the last universal common ances-


457:977–980.


 enzymic”) and factor-dependent systems of translation of polyuridylic acid by Escherichia coli ribo-


[54] Röhl, R, Nierhaus, KH (1982) Assembly map of the large subunit (50S) of Escherichia coli ribo-


Davidovich, C (2011) (Springerverlag Berlin Heidelberg).


Appendix A

Secondary Structure Diagrams of the *Escherichia coli* 16S rRNA
Escherichia coli 16S rRNA
Primary binding protein contacts

Figure A.1: Secondary structure diagram of the E. coli 16S rRNA colored by signatures and contacts of the primary binding proteins.

Adapted from secondary structure diagram downloaded from http://www.rna.icmb.utexas.edu
Escherichia coli 16S rRNA
Secondary binding protein contacts

Adapted from secondary structure diagram downloaded from http://www.rna.icmb.utexas.edu

Figure A.2: Secondary structure diagram of 16S rRNA colored by signatures and contacts of the secondary binding proteins.
**Escherichia coli 16S rRNA**

**Tertiary binding protein contacts**

Figure A.3: Secondary structure diagram of 16S rRNA colored by signatures and contacts of the tertiary binding proteins.

Adapted from secondary structure diagram downloaded from http://www.rna.icmb.utexas.edu

Nucleotides in contact with the **tertiary** binding proteins are colored as follows:

- 5′ domain: S5
- central domain: S21
- 3′ domain: S2, S3, S10, S14
Appendix B

Additional Details for Implementation of Hybrid MD-Gō Potential in NAMD

Maintaining chirality using additional improper dihedrals

To prevent interconversion from L-amino acids/D-ribooses to D-amino acids/L-ribooses in the current implementation of the hybrid MD-Gō model, we explicitly restrained each chiral center with the following harmonic potential:

\[ U_\psi = k_\psi (\psi - \psi_0)^2 \]  \hspace{1cm} (B.1)

where \( \psi \) is the improper torsion (°) and \( k_\psi \) is the force constant (kcal-mol\(^{-1}\)-rad\(^{-1}\)). The CHARMM27 topology and parameter files were modified to include this potential (B.1).

<table>
<thead>
<tr>
<th>( i^* )</th>
<th>( j )</th>
<th>( k )</th>
<th>( l )</th>
<th>( k_\psi )</th>
<th>( \psi_0 )</th>
<th>Atom Names</th>
</tr>
</thead>
<tbody>
<tr>
<td>( X )</td>
<td>( X )</td>
<td>CC</td>
<td>NH1</td>
<td>8.0</td>
<td>120.0</td>
<td>( C_\beta \ C_\alpha \ C \ N ) for backbone</td>
</tr>
<tr>
<td>( X )</td>
<td>CT1</td>
<td>C</td>
<td>( X )</td>
<td>8.0</td>
<td>120.0</td>
<td>( C_\beta \ C_\alpha \ C \ N ) for backbone</td>
</tr>
<tr>
<td>( X )</td>
<td>CP1</td>
<td>C</td>
<td>N</td>
<td>8.0</td>
<td>120.0</td>
<td>( C_\beta \ C_\alpha \ C \ N ) for PRO</td>
</tr>
<tr>
<td>CT1</td>
<td>CT1</td>
<td>CT2</td>
<td>CT3</td>
<td>8.0</td>
<td>240.0</td>
<td>( C_\alpha \ C_\beta \ C_\gamma_1 \ C_\gamma_2 ) for ILE</td>
</tr>
<tr>
<td>CT1</td>
<td>CT1</td>
<td>OH1</td>
<td>CT3</td>
<td>8.0</td>
<td>240.0</td>
<td>( C_\alpha \ C_\beta \ O_\gamma_1 \ C_\gamma_2 ) for THR</td>
</tr>
</tbody>
</table>

Amino acids

Nucleic acids

<table>
<thead>
<tr>
<th>( i^* )</th>
<th>( j )</th>
<th>( k )</th>
<th>( l )</th>
<th>( k_\psi )</th>
<th>( \psi_0 )</th>
<th>Atom Names</th>
</tr>
</thead>
<tbody>
<tr>
<td>CN7B</td>
<td>CN7B</td>
<td>ON6B</td>
<td>NN2</td>
<td>8.0</td>
<td>120.0</td>
<td>( C_2' \ C_1' \ O_4' \ N_1 )</td>
</tr>
<tr>
<td>CN7B</td>
<td>CN7B</td>
<td>ON6B</td>
<td>NN2B</td>
<td>8.0</td>
<td>120.0</td>
<td>( C_2' \ C_1' \ O_4' \ N_9 )</td>
</tr>
<tr>
<td>ON5</td>
<td>CN7B</td>
<td>CN7B</td>
<td>CN7</td>
<td>8.0</td>
<td>120.0</td>
<td>( O_2' \ C_2' \ C_1' \ C_3' )</td>
</tr>
<tr>
<td>ON2</td>
<td>CN7</td>
<td>CN7</td>
<td>CN7B</td>
<td>8.0</td>
<td>240.0</td>
<td>( O_3' \ C_3' \ C_4' \ C_2' )</td>
</tr>
<tr>
<td>ON5</td>
<td>CN7</td>
<td>CN7</td>
<td>CN7B</td>
<td>8.0</td>
<td>240.0</td>
<td>( O_3' \ C_3' \ C_4' \ C_2' )</td>
</tr>
<tr>
<td>CN7</td>
<td>CN7</td>
<td>CN8B</td>
<td>ON6B</td>
<td>8.0</td>
<td>120.0</td>
<td>( C_3' \ C_4' \ C_5' \ O_4' )</td>
</tr>
</tbody>
</table>

* All dihedrals are listed in the i-j-k-l format where the angle is measured between the i-j-k and j-k-l planes;
† The i-j-k-l indices refer to their CHARMM27 atom types, with wildcard atom types designated with ‘X’;
‡ The force constant, \( k_\psi \), is given in kcal-mol\(^{-1}\)-rad\(^{-1}\); 
§ The angle value, \( \psi_0 \), is given in degree(°);
¶ The central atom of each chiral center has been bolded for clarity.

The above parameters have been tested in the hybrid MD-Gō simulations for both proteins (r-proteins S4 and L2) and nucleic acids (five-way junction and the RNA pseudoknot in human telomerase) at various
temperatures, and the resulting trajectories were checked against chirality errors. With a force constant of 8 kcal-mol$^{-1}$-rad$^{-1}$, no chiral errors occurred in all tested simulations.

**Running hybrid MD-Gō simulations with NAMD**

The following changes are required in addition to the standard NAMD input files in order to run the hybrid MD-Gō simulation with NAMD:

1. The PSF file which describes the structure of the system needs to be modified so that all partial charges are set to zero.

2. An extra PDB file to specify the coordinates of the native/target conformation of the system. It is possible in this file to define separate chain types (given as an integer) using the occupancy field, such that each molecular chain may be subjected to a different Gō potential.

3. An extra configuration file that sets all the non-bonded Gō potential parameters. This file may include multiple sections, each section assigns a complete set of the parameters within each molecular chain or between the chains according to the definition in the native PDB file.

An example of the configuration file specifying Gō parameters for the simultaneous binding and folding simulations of five-way junction and S4 presented in Chapter 5 is shown below:

```plaintext
chaintypes 1 1 # Gō potential for the protein molecule
epsilon 0.23
exp_a 12
exp_b 6
sigmaRep 2.5
epsilonRep 0.01
cutoff 4.0
restriction 0

chaintypes 1 2 # Gō potential between the RNA and protein molecules
epsilon 0.15
exp_a 8
exp_b 4
sigmaRep 3.5
epsilonRep 0.01
cutoff 4.4

chaintypes 2 2 # Gō potential for the RNA molecule
epsilon 0.1
exp_a 12
exp_b 6
sigmaRep 2.5
epsilonRep 0.01
cutoff 4.0
restriction 0
```

4. A specialized force field parameter file where all the vdW interactions are zeroed out and additional improper dihedrals to keep chirality are added.

After these files are prepared, one needs to turn off electrostatic and turn on the Gō potential calculation
in the NAMD standard configuration file to start the hybrid MD-Gō simulation.

**Validation of the Gō model for nucleic acids**

An eight-nucleotide RNA hairpin consisting of two canonical GC base-pairs and a UCCG tetraloop was used to test the Gō model implementation in NAMD. Starting from a single-stranded conformation selected from a long MD simulation, the hairpin was simulated with the Gō potential at various temperatures. Showing by the fraction of native contacts (Q), the molecule stay folded for the majority of time at temperatures lower than the transition temperature $T_F$. Around $T_F$, the hairpin is seen going between the folded hairpin and the unfolded conformation back and forth. Above, the hairpin is unfolded most of the time with only short excursion into the hairpin conformation (Figure B.1 B). Simulations from all temperatures were then combined using the weighted histogram analysis method (WHAM), and the density of states were estimated to obtain the potential of mean force (PMF) over the reaction coordinate Q. It is shown clearly on Figure B.1C that at $T < T_F$, the folded hairpin structure is favored; while at $T > T_F$ the unfolded conformation is favored.

![Figure B.1](image)

**Figure B.1:** (A) the hairpin is a two-state folder. (B) Fraction of native contacts (Q) over time at selected temperatures. (C) potential of mean force (PMF) over Q at $0.8T_F$ (green), $T_F$(red), and $1.2T_F$ (blue).

To check the kinetic behavior of the hairpin folding, fifty replicates of short Gō simulations starting from the unfolded conformation were performed at three chosen temperatures. These simulations can be modelled as a single transition from the unfolded to the hairpin conformation with a folding rate of $k$ ($U \xrightarrow{k} N$), and the transition of such simple process should be described well by single-exponential decay. Integrating the first passage time (FPT) from unfolded to folded state for all replicates, it is confirmed that the fraction of unfolded molecules decays exponentially over time at all three temperatures (Figure B.2A). Furthermore, the folding of the RNA hairpin occurs at a higher rate at lower temperatures. This trend can be explained by the energy landscape perspective of biomolecule folding, which is described by the balance between the need to cross the energy barrier and to search in the entropic space. At temperatures higher than the folding temperature, the folding transition is impeded mainly by the large entropic space to search in the landscape. The higher the temperature is, the larger the space to search for, therefore the slower the folding
is. This is consistent with the trend observed in the folding experiment using Gō potential. However, when the temperature is lower than the folding temperature, one should observe an opposite trend since the major obstacle becomes the energy barrier between the two states. Hence, the lower the temperature is, the harder to cross the barrier and the slower to reach the folded state. The absence of this trend is due to the fact that Gō potential is designed to remove the major energy barriers in the landscape so that the transitions can be studied in a reasonable computational time. The effect of the smoothened energy landscape is also depicted in Figure B.2B, where all frames from the trajectory are projected onto a 2D space based on their pairwise RMSD by multi-dimensional scaling (MDS). The result shows that except the folded state, no other obvious intermediate states appear populated during the simulation. The RNA molecule is searching randomly and evenly across the energy landscape towards its folded state.

Figure B.2: (A) Fraction of unfolded RNA over time at $0.67T_F$ (red square), $0.80T_F$ (blue circle) and $T_F$ (green triangle). Solid lines show the single-exponential fit of the simulation data. (B) 2-D folding landscape, where each dot represents one frame and the distance between each pair of dots reflect the pairwise RMSD between the two frames. Color are scaled by relative time in the trajectory.

Parameterization of the Gō potential calculation

Parameters that control the strength of the non-bonded contacts, $\epsilon_{native}$, are chosen so that the ratio between the total dihedral energy and the non-bonded Gō potential energy is balanced. Individual molecules in the system are parameterized separately, such that each molecular chain is expected to have a similar transition temperature. In this way, we hope to observe a more realistic scenario in the dynamics and interactions between the biomolecules that reflects what happens in the cell.

We tested our parameterization scheme on a pseudoknot structure found in the RNA components of the human telomerase (PDB code: 1YMO). For this smaller pseudoknot system (about one fourth of the size of the five-way junction), $\epsilon_{native}$ is set to 0.4 and all other parameters are kept the same as in the five-way junction simulations. Starting from an unfolded conformation generated during the high temperature MD simulation (with RMSD 28.2Å, $R_G$ 23.6Å, and 8% of the native contact remaining), we refolded the single-stranded RNA molecule into its native pseudoknot structure with the hybrid MD-Gō program. Results from multiple replicates run at $T \ll T_F$ show that the system collapse rapidly towards the native state, but can be caught in several intermediate misfolded states (B.3 A). In two out of the ten replicate trajectories, the pseudoknot successfully folded. It is seen from the folded trajectories that a successful folding event
starts from the rearrangement of the Loop 2 conformation, followed by simultaneous folding of Stem 2 and Loop 1 (B.3 C). Stem 1, on the other hand, serves as the rate limiting step in the complete folding of the pseudoknot (B.3 D). The observations in the folding pathway are consistent with what is reported in the recent paper\(^1\), in which a coarse-grained model of the nucleic acids were employed. It is interesting in our all-atom structure-based Gō model that misfolded states are also found, where Stem 1 folded over the Loop 2 region instead of folding under Loop 2 (B.3 E). For the structures to fold correctly, Stem 1 needs to unfold and fold again behind Loop 2. This strongly implies that Stem 1 is more easily trapped into misfolded conformations compared to the other secondary structure elements and that the rate of pseudoknot folding will be largely dominated by the correct folding of Stem 1.

![Figure B.3: Folding of the RNA pseudoknot in the human telomerase (A) Folding of the pseudoknot from four representative Gō trajectories, shown with RMSD (left), radius of gyration (middle), and fraction of native contacts (right). Each simulation was run for 10,000,000 time steps. (B) Secondary structure of the pseudoknot in its native conformation, with structural elements colored as follows: Stem 1 (red), Loop 1 (green), Loop 2 (yellow), and Stem 2 (blue). (C)–(E) Representative snapshots taken from the trajectories, with structures color coded according to the secondary structure diagram shown in (B). The native conformation is shown in background in gray. (Figure courtesy of J Lai)](image)

Therefore, our transferable structure-based Gō model is able to reproduce the hierarchical folding pathway for complex nucleic acids structure. Furthermore, with the all-atom implementation, it is possible to capture the dynamics of the side-chain which may leads to precise identification of distinct intermediate/trapped states along the pathway.

Appendix C

Supporting Information

Copies of all supporting information for this dissertation including, but not limited to, source code, data files, and laboratory notes have been deposited with Zan Luthey-Schulten, UIUC. Written laboratory notebooks were physically transferred and electronic materials archived to tape and then deposited.