**Progress report for Department of Energy grant
DE-FG02-05ER64144: The role of horizontal gene
transfer as a control on the coevolution of ribosomal
proteins and the genetic code**

Carl R. Woese, Nigel Goldenfeld and Zaida Luthey-Schulten
Institute for Genomic Biology,
University of Illinois at Urbana-Champaign

March 31, 2011

1 Introduction

Since the beginning of the DOE grant on “The role of horizontal gene transfer as a control on the coevolution of ribosomal proteins and the genetic code” in September 2005, we made steady and impressive progress in our goal to develop the conceptual and computational tools necessary to understand the evolution of the universal processes of translation and replication and to identify events of horizontal gene transfer that occurred within the components. The research resulted in the publication of ten articles acknowledging the grant [1, 2, 3, 4, 5, 6, 7, 8, 9, 10], and aspects of this research are still on-going. The major findings are summarized below:

2 Project Findings

2.1 On the coevolution of ribosomal proteins and RNA at the Darwinian threshold

We have postulated that the emergence of modern cells was from a communal state and not a unique, shared common ancestor. The evolutionary dynamic of the communal state involves innovation sharing among the communal cells until a degree of complexity is achieved that no longer tolerates ambiguity in the components of translation. Woese termed this point just prior to the emergence of vertical evolution and the three major lineages to be the Darwinian threshold. According to our theory of the evolution of translation, the innovation sharing occurring in early evolution was dominated by HGT. Traces of this HGT have been detailed in both sequence [11] and structure [12, 13, 14] phylogenetic analyses of the aminoacyl-tRNA synthetases. While one would still expect to find HGT in the more peripheral components of translation, once a requirement for increased translation accuracy had been established we would anticipate much reduced variation in the key components - namely the ribosomal proteins and RNA.

Since Woese et al. used the evolutionary relationship between the 16S ribosomal RNA (rRNA) of the small ribosomal subunit (SSU) to infer the universal phylogenetic tree (UPT) [15], the 16S rRNA has been used to classify all extant and newly discovered life forms. This classification is possible because of the appearance of rRNA signatures at the Darwinian threshold and at each subsequent transition in the vertical descent of organisms. The signature analysis for the 16S rRNA have been detailed in numerous studies [16, 17], and a similar analysis was extended by Woese and the Luthey-Schulten group to the 23S ribosomal RNA and the ribosomal proteins in the large subunit (LSU) in 2008 [4]. In the RNA there are in general two kinds of signatures, homologous (similar structure, but significant sequence variation) and non-homologous (containing significant secondary and tertiary structure variations). Signature analysis of all the protein:RNA complexes involved in translation motivated the development of the evolutionary analysis software *MultiSeq* [3] in the early phase of this grant which is described below.

The assembly and stability of the SSU and LSU is facilitated by the binding of ribosomal proteins to the rRNA. Since there are physical interactions between the ribosomal proteins and RNA, we proposed that changes in the rRNA must lead to compensating changes in the ribosomal proteins and *vice versa*. Such coevolution can manifest itself in numerous ways, including the replacement of rRNA structural regions with proteins and the lengthening of proteins to stabilize lengthened rRNA stem-loops. Associated with many of the signatures regions on both the 16S and 23S rRNA are distinct sets of ribosomal proteins which have coevolved with the rRNA signatures. One example of this coevolution is the ribosomal protein S4, which is critical to the assembly of

Figure 1: Overlap of the large subunit of the ribosome from bacteria (*E. coli*; blue) and archaea (*H. marismortui*; pink) using the *MultiSeq* extension to VMD. Also shown are the ribosomal proteins L2 and L3 colored by structural similarity (red/low to blue/high).

the SSU [18, 19] and functions by exploiting the intrinsic disorder of the N-terminal domain. Our coevolution analysis of S4 and the 16S rRNA [4, 5] shows that the N-terminal portion of this protein has evolved with a region of the 16S (400-430 in *E. coli* numbering) that is a signature region between the bacterial and archaeal types: the bacterial stem-loop is extended by about 5 base-pairs compared to the archaeal one and also contains a bulge loop. The N-terminal region of the bacterial S4 is around 10-12 residues longer than in the archaeal S4 and these residues make contact with the extended stem-loop, presumably stabilizing it. 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. In [5] we report the evolution of ribosomal protein S4 in relation to a broad sharing of zinc/non-zinc forms of the gene and study the scope of horizontal gene transfer (HGT) of S4 during bacterial evolution. We present the complex evolutionary history of ribosomal protein S4 using 660 bacterial genomes from 16 major bacterial phyla. According to conserved characteristics in the sequences, S4 can be classified into C+ (zinc-binding) and C- (zinc-free) variants, with 26 genomes (mainly from the class *Clostridia*) containing genes for both. A maximum likelihood phylogenetic tree of the S4 sequences was incongruent with the standard bacterial phylogeny, indicating a departure from

Figure 2: The regions of the 23S rRNA interacting with L15 (green) and L18e (red) are shown. L15 has a long tail that interacts with a number of helical loops in the 23S rRNA and a number of tertiary contacts between regions far off in sequence (Domain 1 of rRNA shown in Violet, Domain 2 of rRNA shown in Blue, and Domain 5 of rRNA shown in Orange) are formed closed to the tail. L18e is an archaeal specific protein that has coevolved to interact with an archaeal specific insert (shown in yellow).

strict vertical inheritance. Further analysis using the genome content near the S4 genes, which are usually located in a conserved gene cluster, showed not only that HGT of the C- gene had occurred at various stages of bacterial evolution, but also that both the C- and C+ genes were present before the individual phyla diverged. To explain the latter, we theorize that a gene pool existed early in bacterial evolution from which bacteria could sample S4 gene variants, according to environmental conditions. The distribution of the C+/- variants for seven other zinc-binding ribosomal proteins in these 660 bacterial genomes is consistent with that seen for S4 and may shed light on the evolutionary pressures involved. The complex history presented for the “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.

While the majority of the ribosomal proteins exhibit the full canonical pattern, there are approximately 12 proteins, so-called domain specific rproteins, found on the LSU that are only present in either the bacteria or the eukarya/archaea domains. This recent discovery raises the question whether these proteins are relatively recent inventions, as might be implied from their localization

in a particular domain, or whether they are more ancient, perhaps already in existence at the time of the last universal common ancestral state (LUCAS). In our study on the evolutionary history of Cys-tRNA^{Cys} formation [1], we developed a comparative phylogenetic approach, combining evolutionary information from protein sequences and structures, that takes both the signature of HGT and the recurrence of the full canonical phylogenetic pattern into account, in order to document the complete evolutionary history of cysteine coding and to understand the nature of this process in LUCAS. Similarly, using a combination of sequence and structure information, we were able to conclude that L18e, found only in archaea and eukarya, is homologous to the globular domain of ribosomal protein L15, which appears in all three domains of life. Phylogenetic analysis has been used to reveal that L18e displays the canonical archaeal/eukaryal pattern. In addition, the phylogenetic analysis also reveals that L18e had diverged from L15 via an ancient gene-duplication event followed by functional diversification leading to their specific binding sites on the 23S rRNA that occurred before LUCAS. The structure of the ribosome reveals that L18e interacts with an insertion in the rRNA that is found only in archaea and could explain the absence of L18e from the bacterial domain of life (Figure 2) indicating that at LUCAS, the proteins present in archaic bacterial ribosome was different from that in the archaic archaeal/eukaryotic ribosome.

2.2 Models of the evolution of translation

In [20] we presented essential concepts from our theory of evolution that pertain to the evolution of the genetic code. There we proposed that reticulate evolution provides a generic and natural explanation for the universality and optimality of the genetic code, and outlined a variety of collective mechanisms for the dynamics of the genetic code. We demonstrated these ideas in a simplified dynamical model of code evolution within the context of early communal life. We speculated that the corollary of the gradual optimization of the genetic code over evolutionary time is the relatively conserved nature of the cell's translational machines as compared to those that perform replication and transcription, and briefly discussed the relationship between the dynamics of the genetic code and the Darwinian transition. The journal *Nature* invited us to write an essay on the theory of evolution [21] that was the first article in a series on making connections between disciplines and explores the need to develop a new language to address the new concepts required to understand evolution.

The work of Goldenfeld and Woese on the modeling the translation apparatus and its evolution, was then quickly expanded to model collective and nonlinear dynamical effects, with a number of experimental predictions made. The main findings were as follows:

- * predicting the threshold behavior when a prophage is induced from a host cell by external stress, by modeling the collective behavior in the DNA damage repair mechanism. The predictions of this work were verified in experiments performed at Illinois by Ido Golding.

- * predicting the pattern of genome composition biases in microbes as arising from the interplay between mutation-drift-selection and tRNA population dynamics in the cell, thus explaining how biases in GC/AT, codon usage and lagging-strand asymmetry emerge from template-directed synthesis in replication, translation and transcription.

- * quantitative predictions of phenotype variation in biofilms arising from horizontal gene transfer, correctly predicting the rate of horizontal gene transfer from growth rate curves. This work predicted macroscopic biofilm phase variation in detail, with only one fitting parameter (rate of HGT). The value we inferred for this parameter was verified by direct visualization of HGT in a

paper by Miroslav Radmans group, published in Science as our work was in the process of submission/review/publication in PNAS.

* discussion of the conceptual relationship between molecular biology and templating, and the dynamical perspective on the evolution of the translation apparatus, emphasizing the role of collective effects.

* Detailed phylogenetic analysis of the evolution of the cells replication machinery across different domains of life. This work was undertaken in order to assess domain-wide trends in evolutionary history for the different template-directed synthesis mechanisms within the cell. The work on the replication machinery convinced us that it is premature to conduct this study across all the cellular machinery, due to the difficulty in the phylogenetic reconstruction. Ongoing efforts now being conducted by Woese and others at IGB are reviewing the phylogenetic history of the amino-acyl tRNA-synthetases with a view to reassessing this perspective for the translation machinery.

2.3 Evolutionary analysis software Multiseq/VMD

A considerable portion of our time was invested in developing the necessary software tools and protocols that combine structure and sequence information in order to characterize the evolution of the protein:RNA components involved in translation, specifically the aminoacyl-tRNA synthetases, the translation factors, and ribosomal proteins and RNA.

MultiSeq [3] is a new extension for VMD [22] for creating, editing, and analyzing protein and nucleic acid multiple sequence and structure alignments. VMD, which is publicly distributed by the NIH Resource for Macromolecular Modeling and Bioinformatics at the University of Illinois, was originally designed to analyze molecular dynamics trajectories of large systems. By adding the ability to carry out multiple sequence and structure analysis in the framework of evolutionary biology, we moved towards our goal to develop a more integrative environment in which to study the correlations and strategies of self-organization in the evolution of proteins and RNA involved in translation, transcription, and replication. This includes the ability to construct statistically well-balanced evolutionary profiles, identify HGT events, and carry out structural overlaps of entire ribosomes, identify the molecular signatures of ribosomal evolution, and represent the comparisons in both sequence and structure phylogenetic trees. The MultiSeq plugin has been especially designed through a collaboration with Carl Woese to address questions arising in evolutionary biology and incorporates state of the art multiple sequence and structure alignment programs such as MAFFT and STAMP as well as the RAxML phylogenetic tree algorithm. To be able to handle large systems like the ribosomes, much of the bioinformatic analysis is carried out through scripts running on clusters.

Some of MultiSeq's features are shown in Figure 1. The program calculates and displays sequence and structure similarity within and between homologous groups and performs a variety of evolutionary and bioinformatic tasks, including the construction of distance-based phylogenetic trees for both sequence and structure. Sensitive evolutionary profiles, such as those used in the work on the history of cysteine coding [1] to detect HGT events and to compare distant homologs and in the ongoing work on the coevolution of the ribosomal proteins and RNA, can be created only by combining non-redundant sequence and structure information. *MultiSeq* contains the multidimensional QR factorization algorithm for both sequence and structure that remove redundancy from a multiple alignment and creates a minimal basis set of proteins that best represents the topology of the phylogenetic tree at a given level of diversity. The sequence editor has been designed to include the basic and most useful features of the AE2 editor, available through the ribosomal database

project¹, and includes an electronic notebook for saving additional annotations. A simplified version of the program for working only with structures was first released in version 1.8.3 of VMD [2]. The combined sequence and structure evolutionary analysis tools were released with VMD version 1.8.5 in August 2006 along with the publication of the new algorithms and approaches used to study both proteins and RNA simultaneously [3]. To date, more than 10,000 researchers from around the world have downloaded and used the *MultiSeq* software. MultiSeq has been updated, extended and improved every year since it was introduced into VMD in 2006.

The programs are publicly available to the broader scientific community through Luthey-Schulten’s website² along with tutorials on the evolution of translation³ which are based on the publications related to this grant and demonstrate how to use the software. The software and tutorials are publicly distributed and have been updated, extended, and improved every year since it was introduced into VMD in 2006.

References

- [1] P. O’Donoghue, A. Sethi, C. R. Woese, and Z. A. Luthey-Schulten, “The evolutionary history of Cys-tRNA^{Cys} formation,” *Proc Natl Acad Sci U S A*, vol. 102, pp. 19003–19008, Dec 2005.
- [2] J. Eargle, D. Wright, and Z. Luthey-Schulten, “Multiple Alignment of protein structures and sequences for VMD,” *Bioinformatics*, vol. 22, pp. 504–506, Feb 2006.
- [3] E. Roberts, J. Eargle, D. Wright, and Z. Luthey-Schulten, “MultiSeq: Unifying sequence and structure data for evolutionary analysis,” *BMC Bioinformatics*, vol. 7, p. 382, Aug 2006.
- [4] E. Roberts, A. Sethi, J. Montoya, C. Woese, and Z. Luthey-Schulten, “Molecular signatures of ribosomal evolution,” *Proc.Natl.Acad.Sci. USA*, vol. 105, pp. 13953–13958, Sept 2008.
- [5] K. Chen, E. Roberts, and Z. Luthey-Schulten, “Horizontal gene transfer of zinc and non-zinc forms of bacterial ribosomal protein S4,” *BMC Evol. Biol.*, vol. 9, pp. 179–195, 2009.
- [6] N. Chia, C. R. Woese, and N. Goldenfeld, “A collective mechanism for phase variation in biofilms,” *Proc Natl Acad Sci U S A*, vol. 105, no. 38, pp. 14597–14602, 2008.
- [7] C. Woese and N. Goldenfeld, “How the microbial world saved evolution from the scylla of molecular biology and the charybdis of the modern synthesis,” *Microbiology and Molecular Biology Reviews*, vol. 73, no. 1, pp. 14–21, 2009.
- [8] N. Chia, I. Golding, and N. Goldenfeld, “Lambda-prophage induction modeled as a cooperative failure mode of lytic repression,” *Phys. Rev. E. Rapid Communications*, vol. 80, no. 030901, pp. 1–4, 2009.
- [9] K. Vetsigian and N. Goldenfeld, “Genome rhetoric and the emergence of compositional bias,” *Proc Natl Acad Sci U S A*, vol. 106, no. 1, pp. 215–220, 2009.
- [10] N. Chia, I. Cann, and G. Olsen, “Evolution of dna replication protein complexes in eukaryotes and archaea,” *PLoS One*, vol. 5, no. e10866, 2010.

¹http://rdp8.cme.msu.edu/docs/rdp_contrib.html

²<http://www.scs.illinois.edu/schulten/software/index.html>

³<http://www.scs.illinois.edu/schulten/tutorials/index.html>

- [11] C. R. Woese, G. J. Olsen, M. Ibba, and D. Soll, “Aminoacyl-tRNA synthetases, the genetic code, and the evolutionary process,” *Microbiol Mol Biol Rev*, vol. 64, pp. 202–236, Mar 2000.
- [12] P. O’Donoghue and Z. Luthey-Schulten, “On the evolution of structure in the aminocyl-tRNA synthetases,” *Microbiol. Mol. Bio. Rev.*, vol. 67, pp. 550–573, 2003.
- [13] A. Sethi, P. O’Donoghue, and Z. Luthey-Schulten, “Evolutionary profiles from the QR factorization of multiple sequence alignments,” *Proc Natl Acad Sci U S A*, vol. 102, pp. 4045–4050, Mar 2005.
- [14] P. O’Donoghue and Z. Luthey-Schulten, “Evolutionary profiles derived from the QR factorization of multiple structural alignments gives an economy of information,” *J Mol Biol*, vol. 346, pp. 875–894, Feb 2005.
- [15] C. R. Woese, O. Kandler, and M. L. Wheelis, “Towards a natural system of organisms: proposal for the domains Archaea, Bacteria, and Eucarya,” *Proc Natl Acad Sci U S A*, vol. 87, pp. 4576–4579, Jun 1990.
- [16] C. R. Woese, L. J. Magrum, R. Gupta, R. B. Siegel, D. A. Stahl, J. Kop, N. Crawford, J. Brosius, R. Gutell, J. J. Hogan, and H. F. Noller, “Secondary structure model for bacterial 16S ribosomal RNA: phylogenetic, enzymatic and chemical evidence,” *Nucleic Acids Res*, vol. 8, pp. 2275–2293, May 1980.
- [17] S. Winker and C. R. Woese, “A definition of the domains Archaea, Bacteria and Eucarya in terms of small subunit ribosomal RNA characteristics,” *Syst Appl Microbiol*, vol. 14, no. 4, pp. 305–310, 1991.
- [18] V. Nowotny and K. Nierhaus, “Assembly of the 30S subunit from *Escherichia coli* ribosomes occurs via two assembly domains which are initiated by S4 and S7,” *Biochemistry*, vol. 27, no. 18, pp. 7051–7055, 1988.
- [19] K. Chen, J. Eargle, K. Sarkar, M. Gruebele, and Z. Luthey-Schulten, “Functional role of ribosomal signatures,” *Biophysical Journal*, vol. 99, pp. 3930–3940, Dec 2010.
- [20] K. Vetsigian, C. Woese, and N. Goldenfeld, “Collective evolution and the genetic code,” *Proc Natl Acad Sci U S A*, vol. 103, pp. 10696–10701, Jul 2006.
- [21] N. Goldenfeld and C. Woese, “Biology’s next revolution,” *Nature*, vol. 445, no. 7126, pp. 369–369, 2007.
- [22] W. Humphrey, A. Dalke, and K. Schulten, “VMD – Visual Molecular Dynamics,” *J. Mol. Graphics*, vol. 14, pp. 33–38, 1996.