

UC San Diego

UC San Diego Electronic Theses and Dissertations

Title

The gospel according to LUCA (the last universal common ancestor)

Permalink

<https://escholarship.org/uc/item/6pw6n8xp>

Author

Valas, Ruben Eliezer Meyer

Publication Date

2010

Peer reviewed|Thesis/dissertation

UNIVERSITY OF CALIFORNIA, SAN DIEGO

The gospel according to LUCA (the last universal common ancestor)

A dissertation submitted in partial satisfaction of the
requirements for the degree Doctor of Philosophy

in

Bioinformatics

by

Ruben Eliezer Meyer Valas

Committee in charge:

Professor Philip E. Bourne, Chair
Professor William F. Loomis, Co-Chair
Professor Russell F. Doolittle
Professor Richard D. Norris
Professor Milton H. Saier Jr.

2010

The Dissertation of Ruben Eliezer Meyer Valas is approved, and it is acceptable
in quality and form for publication on microfilm and electronically:

Co-chair

Chair

University of California, San Diego

2010

DEDICATION

I dedicate this work to Alexander “Sasha” Shulgin. Sasha is probably the wisest person I’ve met. He refuses to allow anything other than his own imagination dictates the rules of what is possible. Many of us are just living inside his dreams. This thesis is my attempt at answering the great question of “why are we here?”, and Sasha is one of the few people I’d really listen to if he tried to answer it.

EPIGRAPH

"First they ignore you, then they ridicule you, then they fight you, then you win."
Mahatma Gandhi

TABLE OF CONTENTS

Signature Page	iii
Dedication	iv
Epigraph.....	v
Table of Contents.....	vi
List of Figures	xi
List of Tables	xiii
Acknowledgements.....	xv
Vita.....	xvii
Abstract of the Disseration	xix
Chapter 1: Introduction to the great tree of life	1
Introduction.....	1
Fossils and the origin of life	3
The biggest divides in life.....	5
LUCA or the root of the tree of life or the cenancestor or the progenote.....	7
Methods of rooting the tree.....	8
Conclusion of the introduction	12
Figures	13
Chapter 2: Save the tree of life or get lost in the woods.....	14
Abstract.....	14
Open peer review	15
Background.....	15

Results and Discussion	19
The Great Tree of Cellular Function.....	19
Placing sequence trees in their cellular context	25
Saving the tree of life	31
Reviewers' Comments	35
Reviewer's report 1	35
Reviewer's report 2	60
Figures	75
Tables.....	77
Chapter 3: Nothing about Protein Structure Classification Makes Sense Except in the Light of Evolution.....	78
Abstract.....	78
Introduction.....	78
Bottom-up Approaches to Protein Structure Classification.....	79
Top-down Approaches to Protein Structure Classification	80
The Protein Domain as Today's Unit of Structural Classification	81
Domain Distribution in Complete Genomes	82
Is the Domain the Correct Unit of Classification?.....	85
Subdomain Features.....	86
Macrodomain Features	87
Putting it all Together	88
Chapter 4: Critical evaluation of structure based phylogenies	92
Introduction.....	92

Problems with the models.....	93
Testing the trees	100
New Algorithms for Building Fake Trees	102
Potential new methods (that didn't work)	105
Directed 1.....	105
Directed 2.....	105
Hybrid of Song and Gustavo.....	106
Derived Domain Combinations	106
Results.....	108
Discussion: Why this project failed.....	108
Figures	110
Chapter 5: Rethinking proteasome evolution: Two novel bacterial proteasomes	115
Abstract.....	115
Introduction.....	116
Results.....	121
Phylogenetic Analysis.....	121
Structural Analysis.....	125
Function of Anbu and BPH.....	128
Discussion.....	131
Methods	133
Figures	135
Tables.....	140

Chapter 6: Structural analysis of polarizing indels: an emerging consensus on the root of the tree of life	141
Abstract.....	141
Background.....	142
Results.....	147
Indels in elongation factors place the root within Bacteria.....	147
HisA and HisF exclude the root from all Gram-negatives except the Eobacteria. .	149
Protein structure alignment renders the Hsp70/MreB indel inconclusive	150
Quaternary structure of PyrD excludes the root from the Archaea and Firmicutes	151
Ribosomal protein S12 and RpoC are probably not paralogs.....	154
Discussion	156
Methods	159
Reviewers' comments	161
Reviewer's report 1	161
Reviewer's report 2	166
Reviewer's report 3	170
Figures	175
Tables.....	181
Chapter 7: The origin of a derived superkingdom: how a gram-positive bacterium crossed the desert to become an archaeon	182
Abstract.....	182
Introduction.....	183
Three robust reasons the Archaea are derived	189

Ribosomal revolutions are historical fact, not science fiction	193
There truly is a great divide in DNA replication machinery, but it can be bridged ...	198
Are the Archaea Paraphyletic or Holophyletic? We're agnostic	202
Where are the bacterial roots of the Archaea?	210
Weakening the neomuran hypothesis	210
Evidence that supports a Firmicute ancestry	214
Peroxisomes: the red herring?	218
Viruses are the missing link between the prokaryotic superkingdoms.....	221
The greatest battle ever fought.....	226
Discussion	234
Figures	241
Tables.....	247
Chapter 8: The evolution of amino acids: better, faster, cheaper	260
Introduction.....	260
Old AAs conserved in extant proteins	261
Why this Alphabet?	264
Discussion	268
Figures	270
Tables.....	273
References	274

LIST OF FIGURES

Figure 1-1. A hypothetical example of indel polarization.	13
Figure 2-1. Genomic vs Cellular Abundance in <i>L. interrogans</i>	75
Figure 2-2. The yin and yang of evolution	76
Figure 4-1. A comparison of two species in the protein universe.	110
Figure 4-2. Scheme for encoding derived domain combinations.	111
Figure 4-3. Song's method vs Dollo parsimony.	112
Figure 4-4. Song's method vs Dollo parsimony in the presence of HGT.....	113
Figure 4-5. Comparison of effects of HGT on 6 different methods.	114
Figure 5-1. Maximum likelihood tree of proteasome subunits in the Superfamily database.....	135
Figure 5-2. Maximum likelihood tree from a structural alignment.	135
Figure 5-3. Comparison of catalytic triads in different proteasomes.....	136
Figure 5-4. Comparison of Anbu to crystal structures of known proteasomes.	137
Figure 5-5. Structural Comparison of BPH.	138
Figure 5-6. Summary of operons for Anbu and BPH.	139
Figure 6-1. EF-2 contains a derived insert.....	175
Figure 6-2. HisA does not exclude the root from the Eobacteria.	176
Figure 6-3. Structural Alignment of MreB/Hsp70.....	177
Figure 6-4. Quaternary Structure of PyrD	178
Figure 6-5. Maximum likelihood tree of PyrD 1B	179
Figure 6-6. Summary of data.	180
Figure 7-1. Two scenarios for interpreting the three polarizations.....	241

Figure 7-2. Alignment of RadA sequences from representative Archaea.	242
Figure 7-3. Maximum likelihood tree of GatD argues for multiple horizontal transfers.	242
Figure 7-4. Structural alignment of EF-1 and EF-Tu.	243
Figure 7-5. Sequence alignment of EF-1 from representative Archaea.....	243
Figure 7-6. Maximum likelihood tree of fungal type Fatty Acid Synthase complex.	244
Figure 7-7. Alignment of L7Ae paralogs in Archaea and Firmicutes.	245
Figure 7-8. 23s rRNA A2058 (<i>E. coli</i> numbering) is well conserved across bacterial hyperthermophiles.....	245
Figure 7-9. Summary of our hypothesis.	246
Figure 8-1. Normalized comparison of AA usage between old and young ribosomal proteins.....	270
Figure 8-2. Catalytic sites of lysine biosynthesis module.....	270
Figure 8-3. Plot of Van der Waals volume vs charge of standard and possible amino acids.	271
Figure 8-4. Plot of hydrophobicity vs charge of standard and possible amino acids.	271
Figure 8-5. Plot of hydrophobicity vs volume of standard and possible amino acids. ..	272

LIST OF TABLES

Table 2-1. Coverage in terms of cellular versus genomic abundance in <i>E. coli</i> 's cytosol.	77
Table 2-2. Coverage in terms of cellular versus genomic abundance in a <i>L. interrogans</i> cell.....	77
Table 2-3. Inconsistency scores (ISs) for the most abundant proteins in each data set. The most.....	77
Table 5-1. Combinations of proteasomes in bacteria.....	140
Table 5-2. Comparison of genomic occurrence of transglutaminase-like catalytic domains.	140
Table 6-1. Length between motifs in elongation factors.	181
Table 7-1. Essentiality of ribosomal proteins.	247
Table 7-2. Summary of differences in DNA replication machinery of Archaea and Bacteria.	248
Table 7-3. Analysis of potentially informative gene structures in Korarchaeota and Thaumarchaeota.	251
Table 7-4. Informative ribosomal proteins in Thaumarchaeota and Korarchaeota.	253
Table 7-5. Summary of data used to support Actinobacterial ancestry of the Archaea..	254
Table 7-6. Summary of data that supports Bacilli ancestry for the Archaea.	256
Table 7-7. Pfams that originated near LAECA and their distribution in the viral world.	257
Table 7-8. Drug targets found across Bacteria that were probably not in LACA.	258
Table 7-9. Examples of drugs targets sites with resistance in Archaea.	259

Table 8-1. Comparison of prebiotic amino acid use in old and young ribosomal proteins.	
.....	273
Table 8-2. Comparison of standard AA alphabet to 10000 random alphabets.....	273

ACKNOWLEDGEMENTS

I would like to acknowledge Professor Philip Bourne for his help as my committee chair. His assistance in developing my scientific skill set was priceless. All of my manuscripts tell better stories because of his help. Professor Russell Doolittle has always been a voice of encouragement when I needed it, and has given me much perspective on the scientific process. Professor William Loomis has provided me with many useful discussions.

I would also like to acknowledge the Bourne Lab. They have listened to me ramble about LUCA for half a decade, and have always provided encouraging feedback. I must especially thank Song Yang. Song's pioneering work on using protein structure to study evolution is the foundation of my work.

Chapter 2, in part, has been submitted for publication of the material as it may appear in *Biology Direct*, 2010, Valas, RE; Bourne, PE. The dissertation author was the primary investigator and author of this paper.

Chapter 3, in full, is a reprint of the material as it appears in *Current Opinions in Structural Biology*, 2009, Valas, RE; Yang, S; Bourne, PE. The dissertation author was the primary investigator and author of this paper.

Chapter 5, in full, is a reprint of the material as it appears in the *Journal of Molecular Evolution*, 2008, Valas, RE; Bourne, PE. The dissertation author was the primary investigator and author of this paper.

Chapter 6, in full, is a reprint of the material as it appears in *Biology Direct*, 2009, Valas, RE; Bourne, PE. The dissertation author was the primary investigator and author of this paper.

Chapter 7 in part is currently being prepared for submission for publication of the material Valas, RE; Bourne, PE. The dissertation author was the primary investigator and author of this material.

VITA

- 2005 Bachelor of Science, Carnegie Mellon University
- 2010 Doctor of Philosophy, University of California, San Diego

PUBLICATIONS

Valas RE, Bourne PE (2008) Rethinking proteasome evolution: two novel bacterial proteasomes. *J Mol Evol* 66(5): 494-504.

Yang S, Valas R, Bourne PE (2009) Evolution studied using protein structure. In: Gu J, Bourne PE, editors. *Structural bioinformatics*. 2nd ed. Hoboken, N.J.: John Wiley & Sons. pp. p.

Veretnik S, Wills C, Youkharibache P, Valas RE, Bourne PE (2009) Sm/Lsm genes provide a glimpse into the early evolution of the spliceosome. *PLoS Comput Biol* 5(3): e1000315.

Valas RE, Yang S, Bourne PE (2009) Nothing about protein structure classification makes sense except in the light of evolution. *Curr Opin Struct Biol* 19(3): 329-334.

Valas RE, Bourne PE (2009) Structural analysis of polarizing indels: an emerging consensus on the root of the tree of life. *Biol Direct* 4: 30.

Hodin J, Bishop CD, Sharpe FA, Valas RE (2009) A creative celebration of evolution. *Nature* 461(7265): 733.

Bourne PE, Briedis K, Dupont C, Valas R, Yang S (2010) Evolution of Molecular Repertoires: Genome Evolution. In: Caetano-Anollés G, editor. *Evolutionary genomics and systems biology*. Hoboken, N.J.: Wiley. pp. p.

Dupont C, Butcher A, Valas R, Bourne P, Caetano-Anolles G (2010a) The history of biological metal utilization inferred through phylogenomic analysis of protein structures. *Proceedings of the National Academy of Sciences*.

Valas RE, Bourne PE (2010a) Save the tree of life or get lost in the woods. *Biology Direct*. (under review)

Valas RE, Bourne PE (2010b) The origin of a derived superkingdom: how a gram-positive bacterium crossed the desert to become an archaeon. *Biology Direct*. (in preparation)

ABSTRACT OF THE DISSERTATION

The gospel according to LUCA (the last universal common ancestor)

by

Ruben Eliezer Meyer Valas

Doctor of Philosophy in Bioinformatics

University of California, San Diego, 2010

Professor Philip E. Bourne, Chair
Professor William F. Loomis, Co-Chair

One of the central ideas of Darwin's theory of descent with modification is that all life had a single origin, and all the various forms we see today are the result of variation and selection. 150 years later in the age of genomics we have realized there are three fundamentally distinct superkingdoms of life: the Archaea, Bacteria, and Eukaryotes. How did these three groups evolve from one common ancestor? This dissertation attempts to answer that question with a variety of bioinformatics tools. The standard approach to this problem is to use mutations in protein sequence as a guide to construct

phylogenetic trees. The failure of this approach has led many to argue there is no “tree of life”. I argue against that strategy and conclusion by looking at the limitations of protein sequence data, and instead supplement it with other data sources. I argue that the rRNA trees which have been a standard of phylogeny for 30 years are misleading. They imply the root of the tree of life lies between the Archaea and the Bacteria. Instead of constructing trees, I have focused on creating timelines by differentiating between ancestral and derived taxa. The divergences in question occurred billions of years ago, so much of the signal present in sequence has degraded. Protein structure is more conserved than protein sequence on these time scales, so I use it as my tool of choice wherever possible. I argue the root of the tree of life is actually within the Gram-negative Bacteria and may have had a U-DNA genome. I have developed three polarizations that imply the Archaea are derived, and refuted all arguments that exclude the root from the Gram-negative Bacteria. One of these transitions led to the discovery of two novel proteasome homologs in Bacteria. It is not sufficient to argue the Archaea are derived from the Bacteria, one must provide a detailed explanation of how the most fundamental biological systems could all go through such a dramatic change to give rise to a novel superkingdom. I present a novel hypothesis that is a combination of selective pressure from antibiotic warfare, the right preadaptation in a Gram-positive bacterium, and relaxation of constraints from a viral endosymbiosis starting a chain of events that led to the birth of the Archaea.

Chapter 1: Introduction to the great tree of life

Introduction

Erasmus Darwin (Charles's grandfather) is credited as the first to suggest that life had a monophyletic origin [1]. Charles Darwin included what he called a "historical sketch" with the *Origin of Species* [2] to give credit to others who proposed parts of his idea before. Others had already proposed that species changed over time, and taxonomies naturally formed branching structures. Evolutionary representations before and after Darwin were recently reviewed [3]. Darwin's great insight was not that species evolve, but rather the process by which they evolve. He tied many fragmented ideas about life into one coherent story.

Darwin's descriptions of natural selection are still valid today, however our knowledge of variation was unimaginable to him. Darwin had no knowledge of the nature of the genetic material or the mechanism by which it replicates, but heredity was still apparent to him. He admitted his ignorance on the subject, but correctly predicted these processes would be well understood some day. The world is quite different today, with knowledge of the double helix nature of DNA being a trivial fact.

However, the idea that species bifurcate into novel species under environmental pressures still holds. The bifurcation process gives rise to a tree-like structure, and hence Darwin's metaphor of the great tree of life. Darwin only included one illustration in

Origin of Species; a hypothetical tree that shows how species could adapt and diversify. And that was the first big picture of the history of life. Man for the first time was placed in his proper place in the universe where all life was connected by a common history and process. The greatest praise for Darwin came from Wallace, “Mr. Darwin has given the world a new science, and his name should, in my opinion, stand above that of every philosopher of ancient or modern times”.

Biology has spent the past 150 years trying to make the data observable in the world fit into a neat and ordered sketch in a similar matter. Darwin was wise enough to use a hypothetical example because otherwise critics would nitpick at the details instead of focusing on the big picture. And so today we are left arguing over the details of the history of life, and the correct way to represent and reconstruct it. In this chapter I hope to give a sketch of the state of confusion the community is in about evolution. There is no debate in the scientific community that evolution happened (despite the hopes of the intelligent design community), but there is still much debate on the history of the major events.

Lately there has been a lot of discussion on the nature of the tree of life and whether it is the correct metaphor for evolution [4, 5]. Genomic data indicates that horizontal gene transfer is common enough to invalidate the tree metaphor in many circumstances. I will discuss this subject at length in chapter 2, but for now keep in mind wherever I use the term tree some would prefer to build webs, and that not all genes in genome share the same history. That said, for the reason I will outline in that chapter I still prefer to use a tree as the main description for the history of life.

Fossils and the origin of life

My journey into biology began with dinosaur fossils. Like many young children I was fascinated by their size and ferociousness. As I've grown older they appeal to me because they represent a world I do not know; the world of the ancient past. Even though they have been extinct for 65 million years, we still know much about the world of the dinosaurs. One can learn what dinosaurs are the oldest by their location in the rock record. Evolutionary histories can be reconstructed by combining the timeline of fossils with morphological similarities between species. Even putative ecosystems can be reconstructed by looking at the context of the fossils. The fossil record of metazoan life is incomplete, but is detailed enough to reconstruct a fairly complete history. Molecular data has supplemented this, and several relationships that could not be determined by the fossil record alone are now well resolved (reviewed in [6]).

But what about prokaryotic fossils? Can the fossil record constrain prokaryotic life in the same way it does the metazoans? All of the earliest evidence of life comes from indirect fossils, as small single cells just do not leave as distinct of a mark as a *Tyrannosaurus rex*. All of the data for the oldest origins of life are controversial, but they at least give a plausible lower bound for when life started.

Methanogenesis and photosynthesis both leave a distinct isotopic mark in the carbon record. Some of the oldest evidence for life comes from grains of graphite that have low levels of $\delta^{13}\text{C}$ dated at 3.7 Ga (billion years ago), which would be consistent with the presence of photosynthesis [7]. This could also be due to abiotic processes such

as metamorphism. Another line of evidence for early life are microscopic tubular structures from 3.5 Ga formed in volcanic crusts [8]. This type of structure cannot presently be explained by abiotic processes, but similar structures could be made by more recent prokaryotes. Stromatolites form from layers of sediments left from bacterial mats, so they are another form of indirect fossils. There are several stromatolites sites older than 3 Ga. However, each of these is open to interpretation as they could be the result of an abiotic process [9, 10].

Each of these lines of evidence could be the result of some abiotic process. The key point is that early life is highly ambiguous in the fossil record. Even dating the origin of life requires a fudge factor of almost a billion years. Even if any single data set were conclusive it would still leave many questions unanswered. For instance, several different taxa of bacteria can form stromatolites. So the presence of stromatolites means any one of these taxa could have been alive at that point. Archaeal and bacterial cells are similar enough that they cannot be distinguished in the fossil record. To the best of my knowledge there are no fossils from any period that are unambiguously Archaea. Also, fossils always represent a lower bound. The fossils we have found could be relatively young compared to the taxa that left them. We simply do not know because the fossil record is so incomplete.

One ancient event that can be dated with accuracy is the great oxidation event. Approximately 2.3 Ga the atmospheric levels of O_2 began to rise, although oxygenic photosynthesis probably evolved hundreds of millions years earlier [11]. There is no doubt this was caused by Cyanobacteria, as no other prokaryotes produce O_2 . This can be used to constrain the origin of oxygen dependent processes, as well as investigate

adaptations in proteomes in response to long term environmental changes (discussed in chapters 5 and 6). The Cyanobacteria must have been alive and well at that point, but it tells us nothing about what came before them. On the other hand, taxa and metabolisms that are ancestrally aerobic can be constrained by that date [12].

Unfortunately the fossil record does not describe the origin of life at any level of detail. It is highly unlikely that a magic bullet find will ever change this, but the origin of some early groups may be clarified using novel biomarkers. If one wishes to have a detailed understanding of the origin of life they have to use other data sources besides the fossil record. Throughout the rest of this work I attempt to polarize events in evolution. But because of the uncertainty of the prokaryotic fossil record I am highly skeptical of attaching dates to any of these events besides the great oxidation.

The biggest divides in life

The simplest major divide in life was between that which was observable with the naked eye, and that which requires a microscope to see. Further observation led to realization that many microbes had nuclei, and a division was drawn between the eukaryotes and prokaryotes [13]. At that point taxonomy was all about morphology. Eukaryotic cells are much bigger than prokaryotes. Prokaryotic cells could be classified by their shape, or by the color they turned in the presence of Gram staining.

And morphology reigned king of phylogeny. Pauling and Zuckerkandl were the first to propose that mutations occurred in a clock manner and primary sequence could be used to build phylogenies [14]. This idea hit a new high (or low depending on your

perspective) in 1977, as Woese pioneered universal sequence trees [15]. He proposed that rRNA, which is present in every extant cell, could be used as a universal chronometer.

When the authors began sequencing rRNA, they were surprised to find a division within the prokaryotes. This was the first time anyone had noticed the Archaea were not like the other prokaryotes. And thus the three domain model of Eukarya, Bacteria, and Archaea was born. This brilliant insight was initially decried as blasphemy and nonsense[16], but is now universally accepted. I must note at this point that terminology about these groups has become somewhat muddled. Some use prokaryote and bacteria synonymously; while differentiating between eubacteria and archaeobacteria. I find Archaea and Bacteria to roll off the tongue better, so when I say Bacteria I mean eubacteria, not prokaryotes. There are certainly some political motivations behind these names. Woese, who roots the tree between the prokaryotic superkingdoms dropped the “bacteria” suffix. On the other hand, Cavalier-Smith stresses Archaea are not that different from bacteria; so he prefers to give them a name that implies their similarity.

There are several general characteristics we can say about each of these groups. The prokaryotes are generally much smaller in physical and genomic size than the eukaryotes. All eukaryotes have their DNA enclosed in membrane organelle, the nucleus. Some prokaryotes do technically have nuclei [17], but they are still obviously bacteria in all other regards. The major differences between the Archaea and Bacteria lie in their membrane structure, ribosome structure, and DNA replication machinery. It is often supposed that the Archaea and Bacteria are so distinct that they must represent a primordial bifurcation in the TOL. These differences will be discussed in detail in chapter 9, where I argue that none of them justify that position.

It is now known that all eukaryotes are the results of an endosymbiosis. All eukaryotes have a mitochondria (or a mitochondrial-derived organelle), that descended from a free living proteobacteria [18]. There are numerous hypotheses on the nature of that host, but they all agree an archaeon was involved in some way (reviewed in [19]). I will discuss ways of evaluating some of these in chapter 9. Most of them are genuinely creative hypotheses that can easily be dismissed with phylogenetic data, but seem to persist anyway.

The three domain model of life is generally accepted, but there is one alternative hypothesis. James Lake proposed that the eukaryotes are more closely associated with a specific group of archaea he calls the Eocytes (Crenarchaea) [20]. He based this idea on ribosome structure and a shared indel (ambiguous insertion deletion in gene or protein) between the eocytes and eukaryotes. This idea was never taken that seriously because it had little phylogenetic support, but recent work has argued that it may be an artifact [21]. I will also discuss this idea in chapter 9.

LUCA or the root of the tree of life or the cenancestor or the progenote

The last universal common ancestor goes by many names in the literature, but they all mean the same thing. There are significant similarities shared across the 3 superkingdoms that could not have arisen by convergent evolution. The core of the ribosome, the anti parallel double helical shape of DNA, the genetic code, and some membrane insertion proteins are universal in every extant cell [22]. This strongly implies

that at some point in life's history there was a bottleneck, and all extant life descends from there. LUCA is not the first life form, or even the first cell. A large portion of evolution occurred before LUCA.

There is very little agreement on the nature of LUCA. Woese initially proposed LUCA was a community with rampant HGT, but in my opinion this is just an *ad hoc* explanation of the three domain tree. Some have proposed LUCA was not cellular, and the Archaea and Bacteria invented cellularity independently[23]. However, the latest revision of this idea does have a cellular LUCA [24]. There is also a good deal of debate of whether LUCA had an RNA or DNA based genome (see [25]), and in chapter 8 I even propose LUCA had u-DNA genome.

The other issue is the location LUCA on the tree. If we assume the molecular clocks hypothesis to be true LUCA must be the midpoint on the universal tree. That would place LUCA halfway between the Archaea and Bacteria. Midpoint rooting can be used on any tree as long as some molecular clock assumption is made. I will refer to this as the canonical rooting, as it has dogmatically been accepted at this point by almost everyone who does not study the subject. Others have recently argued that LUCA was an archaean [26], a Gram-positive bacterium[27], a Gram-negative bacterium [28], and even eukaryote-like[29]. But how have so many different conclusions been reached? The answer is that they use very different criteria for rooting the tree.

Methods of rooting the tree

Some use the differences between the Archaea and Bacteria to support the canonical rooting. I will discuss these differences at length in chapter 9. But for now it shall suffice to say they are based on the logic of looking for ancestral branches of the tree. I shall call these as a general strategy an inclusive method, as opposed to exclusive methods that define derived branches. I will discuss below why these methods are different from each other.

The first type of inclusive argument we can consider is based on environment. Wong *et al.* list 20 reasons they believe the root is probably within Archaea, specifically near *Methanopyrus kandleri* [26]. This extremophile lives nears black smokers on the ocean floor. The authors make several assumptions about what a primitive ecosystem must have been like, and then argue that LUCA must be near *M. kandleri* in the phylogenetic tree since it lives in those conditions today. They offer several other lines of reasoning, but each case is an argument for why the Archaea represent a primitive state.

Another method for determining which branches are the most primitive is paralog rooting [30-32]. If genes mutated like clocks, two paralogs should create identical trees. If one aligned both sets of sequences and built a tree, it should contain the species tree twice, and each paralogs would be the out-group of the other tree. The TOL could potentially be rooted by examining paralogs that diverged before LUCA. If the subtrees are consistent, the position of the out-group represents the position of the root of the TOL. However, this method relies heavily on clock-like behavior in proteins. Of 154 usable paralogs sets, 137 were inconclusive, 9 supported the canonical rooting, 7 supported rootings within the bacteria, and 1 supported a rooting within the archaea [30-32]. These inconsistencies are the result of horizontal transfer as well as non-clock-like

behavior of the proteins. This method does not reach a strong consensus on the position of the root.

An alternative to inclusive rooting methods is what we call exclusive methods. Instead of looking for states that may be primitive, exclusive methods focus on robustly defining branches of the tree as derived. Any polarizable transition can be used to define a branch of the tree as derived. The best example of such a transition is the endosymbiosis of mitochondria. It is well agreed the mitochondria are descended from a free living proteobacterium and that all eukaryotes had a mitochondrion at one point in their history[19]. Therefore the last common ancestor of the eukaryotes must be younger than the last common ancestor of Proteobacteria. Of course, many species of both groups have evolved since then, but we can say with certainty the root of the TOL cannot be within the Eukaryotes. There are many traits that can be polarized in a similar manner.

The biggest divide within the bacteria is their membrane structure. The Gram-negatives have two cellular membranes, as opposed the Gram-positives that have only one (there are many species classified into these groups that don't actually stain that way; and there are some "other" outer membranes ([33]), but this is very generally a good division). Cavalier-Smith has argued that it would be far simpler to lose an outer membrane than to gain one [28]. He also argues that if a Gram-positive bacterium had a flagella it would be very difficult to preserve its function while adding a second membrane. Since Gram-positives and Gram-negatives have homologous flagella he argues the Gram-positives must be derived.

Indels are another character that have been polarized extensively to root the tree. As an example, consider the hypothetical set of two paralogous genes in Figure 1-1. For

simplicity lets assume these genes are universally distributed and the duplication that resulted in the paralogs occurred before LUCA. All the Eukaryotes have an indel in E'. It would normally be ambiguous whether this is the result of an insertion or deletion. However, we know all the paralogous and some of the orthologous sequences have that region, which implies the ancestor of both genes had it. So E' must be a derived form of the gene. There are six ways of rooting a tree with three taxa. The most parsimonious scenario for explaining the deletion of that region for each of these trees is presented below. A rooting within the Eukaryotes requires at least two losses, so it less parsimonious than the other five trees, and we can exclude the root from the Eukaryotes in this example. Lake *et al.* have published several papers on this method in the past few years [27, 34-37], and I have created a detailed rebuttal to this work [38]. I will discuss this in more detail in chapter 8.

Another source of polarizable data, which I will refer to often, is quaternary structure. Protein chains assemble themselves into higher units of structure, but that assembly reflects the evolutionary process itself [39]. If we can define which form of a structure is derived then we can exclude the root for the branches that have that structure. I will discuss the polarization of Abnu to the 20s proteasome in chapter 7 and the transition from PyrD 1A to PyrD 1B in chapter 8.

Conclusion of the introduction

Many methods have attempted to reconstruct and root the tree of life. Many of these have made naïve assumptions about the behavior of genetic material so they are not that effective in practice. However, I believe that as this field matures exclusive methods will gain more respect. Horizontal transfer cannot make a derived group appear ancient without a selective sweep. Right now our sampling of the tree is still pretty biased. Mapping deep unsampled taxa will provide support that the derived state was present in the last ancestor of a group. Therefore these methods will be robust (not immune) to a major source of noise in sequence analysis. We may never fully resolve the tree of life, but questions of which superkingdom came first should be reliably answered one day.

Figures

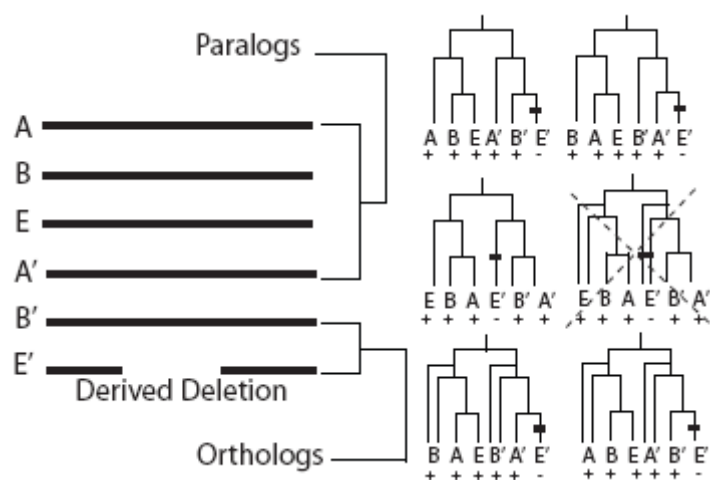


Figure 1-1. A hypothetical example of indel polarization.

The derived deletion in E' excludes the root from the eukaryotes.

Chapter 2: Save the tree of life or get lost in the woods

Ruben E. Valas and Philip E. Bourne

Abstract

Background

The wealth of prokaryotic genomic data available has revealed that the histories of many genes are inconsistent, leading some to question the value of the tree of life hypothesis. It has been argued that a tree-like representation requires suppressing too much information, and that a more pluralistic approach is necessary for understanding prokaryotic evolution. We argue that trees may still be useful representation for evolutionary histories in light of new data.

Results

Genomic data alone can be highly misleading when trying to resolve the tree of life. We present evidence from protein abundance data sets that genomic conservation greatly underestimates functional conservation. Function follows more of a tree-like structure than genetic material, even in the presence of horizontal transfer. We argue that the tree of cells must be incorporated into any new synthesis in order to place horizontal transfers into their proper selective context. We also discuss the role data sources other than primary sequence can play in resolving the tree of cells.

Conclusions

The tree of life is alive, but not well. Construction of the tree of cells has been viewed as the end goal of the study of evolution, where in reality we need to consider it more of a starting point. We propose a duality where we must consider variation of genetic material in terms of networks and selection of cellular function in terms of trees. Otherwise one gets lost in the woods of neutral evolution.

Open peer review

This article was reviewed by Dr. Eric Bapteste, Dr. Arcady Mushegian, and Dr. Celine Brochier. For the full reviews, please go to the Reviewers' comments section.

Background

The pendulum of scientific opinion often swings back and forth in the light of new data and hypotheses. 150 years ago Darwin's observations pushed opinion towards believing the universal tree of life (TOL) existed for the first time [2]. This view was pushed to an extreme 30 years ago as Woese pioneered the use of sequence data to build universal trees [15]. But the pendulum has begun to swing back the other way in the past decade, as a wealth of prokaryotic genomic data has demonstrated a higher than expected frequency of horizontal gene transfer (HGT).

Ford Doolittle and Eric Bapteste's arguments against the TOL hypothesis are quite compelling [5], and this view seems to be gaining support [4]. These authors argue that HGT is so rampant that tree-like representations of prokaryotic species contain too little information to capture evolutionary histories. Their work questions whether the

metaphor of the TOL is inspired from a historical bias from the taxonomy of eukaryotes, and therefore should not be applied to prokaryotes. This is an important and worthwhile question to ask. Resolving the eukaryotic tree is a distinct problem because there is much less horizontal transfer and a much better preserved fossil record. The conclusion of Doolittle and Bapteste is not so much that the inability to build the tree is the problem, rather it is forcing the data into a tree that needs to be questioned, and in a pluralistic framework, avoided, since this model does not allow a precise description of the evolutionary processes.

The TOL and tree of cells (TOC) should be one and the same. However, the meaning of the former has become the trees we can build, and the latter has become the hypothetical tree we cannot build. This difference was recently discussed in [40]. The reason the TOC is truly a tree is simple and has been stated by many before us. Every extant cell on this planet is the daughter of a cell that came before it [41]. Prokaryotic cells divide by binary fission. Therefore there must truly be a TOC in the prokaryotic superkingdoms. Nobody seems to dispute this. If every daughter cell's membrane kept track of who its parent was, reconstructing the evolution of cell divisions would be a trivial task. But since there is no selective pressure for cells to do that, so we are left with a more difficult task.

Since the membranes do not keep track of heredity we chose a different representation of ancestry, the genome. The genetic material of the cell does keep track of its parents in some sense as there is selective pressure to ensure fidelity of replication. All of the issues the community is currently having with the TOL hypothesis stem from the simple fact that genomes are not a perfect representation of membrane history.

Membrane heredity is a tree-like structure, but all of the recent work on the pervasiveness of HGT has shown that genome heredity is often more of a network than a tree. We are beginning to have enough technology to reconstruct genomic evolution, but we are only beginning to realize how vastly different that is from cellular evolution. However, even genomic evolution makes little sense without the light of cellular evolution.

Ernest Rutherford said, "Physics is the only real science. The rest are just stamp collecting". Some biologists have taken this as a challenge to create universal laws in biology on par with those in physics. This is a noble endeavor, and has produced many interesting results, but the goal should be to collect stamps in a way that is justified by the laws. The promise of the TOL did just that. It was a collection of every living thing as well the laws that organized that collection.

Instead of a consensus TOL emerging from the vast amount of genomic data available, the community was faced with the disappointment that very few genes are universally conserved. The universal sequence tree created from 31 concatenated proteins [42] has been criticized as "the tree of 1%" because the average prokaryote genome has about 3000 genes [43]. They argue that even if this gene set did produce a reliable tree it would only reflect a small portion of evolution, since this is such a tiny portion of the genome. The assumption that genomic histories were congruent with cellular histories hid the fact that much of the collection could not be explained by the TOL hypothesis. The community was lost in the woods without knowing it under tree monism.

We worry that the sound arguments against the TOL hypothesis will shift focus away from evolutionary histories. For instance Dagan *et al.* [44] have quantified rates of horizontal transfer for every gene family. This is a tour de force of quantifying a law in

biology. However, they do not give examples of the rates for any specific family, or cite any example they found where a horizontal transfer played a role in speciation. They also say their results are independent of the vertical tree used as input, which we find worrying. The overall rates of HGT may not change, but we assume the rates for each family almost certainly would. Not worrying about that difference is getting lost in the woods to us because the real history of the stamp collection is lost in search of a concise law.

Here we argue for the need to be cautious about how far away from the TOL hypothesis we swing, as novel sources of data already bring into question the conclusions supported by genomics. The arguments against the TOL are centered on the idea that the modern synthesis of biology from 50 years ago was too eukaryote-centric. We hope to offer a perspective that will spare this current synthesis from being labeled too genome-centric 50 years from now. We are not arguing for tree monism. Instead we are attempting to demonstrate that the TOC becomes even more important under a pluralistic approach. Not all genes contribute equally to the cell, and we will demonstrate that vertical inheritance of function has a more pronounced signal than vertical inheritance of genetic material. Using only universal sequences to attempt to resolve the TOC is a narrowing strategy, and we will discuss alternative sources of data that may still shed light on this problem.

Results and Discussion

The Great Tree of Cellular Function

The attempts at resolving the TOL using universal sequences only represent a small part of the history of genomes, but what portion of the history of cells does it represent? Genomic methods represent all genes equally. Subsequently, a gene that is only expressed under specific conditions can be just as useful as a housekeeping gene for building a species phylogeny if both are present across the same set of genomes. If we created a concatenated protein sequence, both proteins would be counted as equals proportional to their length despite vast differences in their actual abundance as proteins in the cell. Here lies the fundamental shortcoming of genomics; confusing the genome and the cell. If we wish to measure a gene's contribution to the cell there are many different metrics: essentiality, abundance of proteins, number of transcripts, and portion of total weight are just a few. Any of these will give dramatically different proportions than simply counting the copy number within the genome.

The abundance of many proteins present in *Escherichia coli*'s cytoplasm has recently been calculated experimentally [45] as well as for the entire cell of *Leptospira interrogans* [46]. For the first time data are available to measure what portion of a prokaryotic cell each protein comprises. All of these numbers should be taken with a grain of salt due to experimental noise, but the trend is clear; the core proteins make up a larger portion of the cell than the genome. The data used to calculate these values are available as supplemental material. We argue abundance is a good proxy for evolutionary importance because there is a correlation between the abundance of a protein and the energy a cell invests into producing it. It has been demonstrated that highly expressed

proteins have been optimized to use less energetically costly amino acids [47], and that highly abundant proteins are shorter on average [45]. The abundant proteins justify the use of a large portion of the cell's energy despite these optimizations, so they must be important. Proteins perform most of the functions in the cell. Comparing how many of the same functions two cells are doing at the same time is a good measure of similarity. The downside of abundance is it is dynamic during a single generation, while genomes are static. This makes direct comparison more difficult, but it still gives insight into the evolution of genomes. Our point is not that this data magically fixes all problems with the TOL hypothesis, but rather that many important details are left unresolved in our understanding of the big picture that still may come into focus.

Let us consider the so called “tree of 1%”. The authors list 36 genes that are universal but claim that only 31 have not been horizontally transferred [42], although later analysis claims the number is actually 22 [48]. However, there are arguments that a TOL is still meaningful despite a large incongruence between individual gene trees [49], but a detailed argument against that view is presented in [4]. 34 of these genes are present in the *E. coli* data set that measures abundance for 1103 proteins. For this argument let us consider the universal set because in this case the HGTs appear to be displacements of genes that were already present. That is to say the function of these genes was vertically inherited despite HGTs of the genetic material. This brings up the point that there are two distinct forms of HGT that we need to consider that are currently not distinguished enough; functional innovations (relative to the recipient genome) and displacements.

For example, the histories of the 20 aminoacyl-tRNA synthetases (AARSs) contain many inconsistencies, the result of HGTs [50]. However, it still appears that most of these enzymes date back to the last universal common ancestor (LUCA), with the possible exceptions of AsnRS and GlnRS. HGT makes many of these proteins unusable for reconstructing a universal TOC as their sequence trees would be inconsistent with the cellular history. However, the HGTs of this family would displace a copy of the gene that was already present. In that sense there is no functional innovation caused by the transfer. Even though the genetic material was inherited horizontally the functional content was still transmitted vertically, and would still be consistent with the TOC. We argue events like this are far more deleterious to tree reconstruction algorithms than they are to the recipient cells. Current methods for estimating HGT rely on measuring inconsistencies between sequence trees or looking for unusual compositional features [51], so there is no way for them to distinguish between innovations and displacements. We must also consider the role functional redundancy plays in prokaryotes. There are nearly two hundred known cases of non homologous enzymes performing analogous reactions [52]. HGT of such enzymes should be relatively easy since they can plug into existing metabolic pathways. Therefore any current measure of the vertical inheritance of genetic material is a lower bound on functional vertical inheritance. If one wishes to measure the size of the vertical component of evolution it must be done in terms of function as well as genetic material.

The universal genes are about 3% of the E. coli dataset if we count all genes equally. However, if we count proteins by abundance, the 34 from the universal core make up 6.6% of the data set (Table 2-1). That would double the thickness of the vertical

component! A tree of 2% may not seem dramatically better than a tree of 1% but the point is that the universal proteins make up a larger portion of the cell than the genome. This gets quite dramatic when one takes into consideration that 84.5% of the abundance in this dataset is made up of ribosomal protein L33. L33 is universal across the Bacteria, but absent in the Archaea. Although it is not universal it still must be very ancient. The fact that a large portion of the cytosolic proteome of an extant cell would be present in the cytosol of LUCA truly speaks to the fact that vertical inheritance can be a major force along long evolutionary time scales. This is consistent with previous work that showed older genes are expressed at higher levels than younger genes [53]. However, the authors did not quantify the contribution of the universal proteins in this manner.

The protein abundance from *L. interrogans* [46] provides us an opportunity to test the trends we see in *E. coli*. This data set was not dominated by a single protein like L33 in *E. coli*. The 36 universal proteins make up only about 1% of this genome, but 5.4% of the entire cell's abundance (Table 2-2). A tree of 5% starts to sound significant! This shows the vertical component of the TOL is five times thicker than genomics alone would lead us to believe. Again, regardless of the nature of LUCA, extant cells still have a significant amount of function in common with her.

The TOL hypothesis has also been challenged on the grounds that currently defined taxa may have a very small genomic core (the intersection of their gene sets) and very large pan genomes (the union of their gene sets) [4, 5]. We analyzed the genomic cores' abundances to explore their contribution to the cell. A subset of the proteins in any Enterobacteriaceae genome are conserved in enough genomes to be considered part of their genomic core, and even fewer are conserved across all nine genomes studied in [54].

This is a very diverse set of bacteria ranging from endosymbionts to free-living species. In each case the coverage in terms of genes is much smaller than the coverage in terms of protein abundance (Table 1). The most dramatic example is that the genes conserved in all nine genomes only account for about 61% of the genes in this dataset, but they account for 99.8% of the protein abundance! Since the ribosomal proteins are so dominant in the data set, we repeated the same measures excluding ribosomal proteins as was done in the initial study. The results are not as impressive, but in each case the core proteins are up to 20% more abundant in the cell than they are in the genome. Measuring the similarity of species based on shared gene content greatly underestimates their functional conservation.

We also considered a set of genes conserved across four Spirochete genomes [55] that includes both obligate and non-obligate parasites. 412 proteins that were conserved across four Spirochete species were mapped to the abundance data set. This accounts for only 11.3% of the *L. interrogans* genome, but makes up 31% of the cell when counted by abundance (Table 2-3). There is stronger conservation between these species than genomic data implies. Again, even though these species have vastly different lifestyles a large cellular core has remained conserved between them.

We assigned COGs (clusters of orthologous genes) to every gene in the *L. interrogans* dataset using the STRING database [56]. This allowed us to compare the relative age of a COG (the percentage of bacterial genomes that have a particular COG) to the genomic or cellular portion of this dataset that COG composes. A cumulative plot of genomic and cellular abundance reveals that at every level genomic abundance underestimates cellular abundance (Figure 1). In some cases the difference can be as high

as 20%. A similar plot for the *E. coli* dataset was not informative because the dominating feature is ribosomal protein L33 (data not shown).

We mapped the tree inconsistency scores (ISs) for all trees from the forest of life in [57] to these datasets to test our proposal that abundance is a barrier to HGT. IS is defined as the frequency the splits in a specific tree are found in all single gene phylogenies in the forest of life. Therefore it is an estimate of the horizontal transfer rate for each family. The authors of that work found that ISs have a bimodal distribution, with many families having very low IS (below .6 of the average IS) or very high IS (above 1.4 of the average IS). They noticed many ribosomal proteins had very low IS and proposed this is due to their numerous physical interactions. If highly abundant proteins are less likely to be transferred their trees should not have high ISs. However, there may still be proteins that are not abundant with very low ISs for reasons besides abundance, so we do not think a comparison of average ISs is informative. Instead we counted how many high ISs were found in the 100 most abundant proteins that had an IS. We repeated this measure excluding ribosomal proteins to ensure this result is due to abundance and not just physical constraints on protein complexes. P-values were estimated by taking 10,000 permutations of 100 ISs taken out of all possible ISs. In each case they were significantly less high ISs among the most abundant proteins than would be expected by chance (

Table 2-3). Abundance does not mean it is impossible for a transfer to occur, but it certainly seems to limit it. Even though the backbone of vertical inheritance may appear to be dwarfed by HGT on the genomic scale, it is clear that vertical inheritance plays a larger role in the cell than the genome.

One of the major criticisms of tree monism is the arbitrary methods used to deal with incongruent data and ambiguities in a single tree of life model [4]. Baptiste *et al.* cite the ambiguous placement of *Aquifex aeolicus* in a supertree as an example of conflicting signals; sometimes they are placed near the Proteobacteria and sometimes near *Thermotoga maritima* [58]. We propose that abundance could be used to consider the relative weights of two incongruent signals such as these. We predict the vertical signal would become clear if we could measure the abundance of proteins shared between *A. aeolicus*, *T. maritima*, and the Proteobacteria. It is not just the number of genes that are shared that matters; one needs to consider what their contribution to the cell is.

Placing sequence trees in their cellular context

So far we have argued the large role vertical inheritance of function has played in evolution is larger than the network of genomes implies. But how can we proceed forward if that tree is not the one reconstructed by gene sequence trees? We must first discuss the very reasons the TOL metaphor is appealing in the first place. Trees are the natural representation of replicating processes [59]. As discussed in [4] the tree representation has three strengths: 1) it provides a hierarchy for classification; 2) shared traits are implied by that hierarchy; and 3) ancestral traits are inferred from the branch order of the tree. We believe the strength of the hierarchy rests on the last two points so we will address those first.

The branching order of a true species tree can be read like a timeline. It implies the ancestral state for each group, as well as the transition that defined the groupings. Without a timeline much of evolution becomes gibberish. There are many transitions that can be polarized. The biggest transition in evolution is between the prokaryotes and eukaryotes. It is truly impossible to represent endosymbiosis in a tree representation, but an undirected network would just say that the eukaryotes are related to both the archaea and bacteria. However, it is not just the relationships we need to know to understand this event; the key point is the directionality of the transition. The eukaryotes came from the prokaryotes. That is to say there was a time where prokaryotes existed but eukaryotes did not. Trees have many shortcomings for representing prokaryotic evolution, but any data structure that lacks temporality is even worse.

Since there is no real branch order on any of the current universal sequence trees or networks, some take this as evidence that all the origin of the major prokaryotic taxa are contemporaneous (the condensed cladogenesis model) [60] or are the result of intense periods of HGT (the biological big bang model) [61]. If the major events in prokaryotic evolution all happen near the root of the tree, then it might be really be impossible to reconstruct life's history. However, there are no fossils that unambiguously mark the origin of any major prokaryotic taxa (reviewed in [62]). Therefore the justification for both of these models comes from a lack of resolution in both the fossil and sequence data. It is critical we make a distinction between the big picture of evolutionary history being hard to resolve and it not existing.

We argue against the expectation that every major prokaryotic taxa showed up at the same moment in evolutionary history. Not all innovations are possible immediately in

evolution; they need the right push in selective pressure from the environment. The most important examples are processes that require an oxygenic atmosphere such as sterol synthesis [63]. Oxygen dependent metabolism could not thrive before the great oxidation event that occurred about 2.3 billion years ago [64]. This can be used to constrain the ages of several branches of the TOL [12].

This begs the question of whether the major prokaryotic taxa are contemporaneous in origin or not. Despite the disagreement between the current macrophylogenies [27, 28, 65], they imply the major prokaryotic taxa did not appear at the same time. The disagreement between these phylogenies is not in terms of how to define the major taxa but rather in the proper way to polarize the data, especially the indels (insertion deletions) which we have discussed [38]. However, the distribution of these traits themselves implies specific taxa evolved before another, regardless of the direction of each polarization. For example, there is a large insert in HSP70 (heat shock protein 70) that is present across the Gram-negative bacteria, but absent in the Gram-positives. There is no reason to assume the insertion deletion event occurred early in evolution. That event would be contemporaneous to the change in membrane structure. There must be a branch order between the Gram-positives and Gram-negatives; even if it is not resolved in sequence trees. One would be very hard pressed to draw a detailed scenario of transfers that explains the distribution of fixed indels better than a more timeline like structure. There are numerous similar divides that can be drawn across the prokaryotes. Relatively stable traits like these must be used as guides to reconstruct the TOC. If the origins of the major clades really are very close to contemporaneous we should not expect it to be possible to reconstruct a macrophylogeny at all using such

traits. Despite their disagreement, we take the macrophylogenies as evidence that the origins of major prokaryotic taxa could not all be contemporaneous. If the clades arose out of intense periods of HGT the indels and other data points should have largely independent histories. Instead many indels appear on the same points on the tree, which act to independently verify each other.

Why would there be more signal left in the rare events instead of the sequence alone? We propose an alternative model where cladogenesis is primarily caused by revolutions in the “abundome.” Some proteins that were highly abundant in LUCA may be entirely absent in an extant cell. This would be a result of major events that led to dramatic changes in gene expression of even the most conserved genes. Such a change in the abundome could make some HGTs more deleterious since the protein would be plugged into a modified core. It is possible these events could actually be periods of reduced HGT. Abundance data may make it possible to quantify what Simpson coined “quantum evolution” when referring to the metazoan fossil record [66], the idea that changes in one part of an organism can trigger a domino effect of rapid evolution, on a molecular level in prokaryotes. There is an inverse relationship between population size and evolutionary rate [67]. Initially members of a novel niche could evolve rapidly. It is also possible that could lead to a population size that is large enough to cause extreme purifying selection, essentially freezing ribosomal sequences. It seems impossible to predict the effects cladogenesis will have on sequences without taking into account other sources of data. This view that revolutions in abundances play a role in cladogenesis is supported by the observation that there are major changes in gene regulation between the major prokaryotic taxa [68]. Since gene expression is a major driver of evolutionary rates

[69], there is no reason to expect protein sequence to be well behaved across these events even if cellular population size remained constant. That, taken with the new observation that evolutionary rates vary greatly between prokaryotic groups [70] implies our null hypothesis should be that sequence trees would not resolve the branch order of the major prokaryotic groups even in the complete absence of HGT. Therefore we are in complete agreement that a tree created by concatenating protein sequences together is not the TOL or the TOC. But we do not take that as evidence the TOC does not exist, is not resolvable, or is not useful as a concept for understanding prokaryotic evolution. It just means the community needs to move beyond primary sequence analysis.

It will be possible to look at the coevolution between the cellular and genomic cores as abundance data becomes available from more species. This will allow us to divide extant species into groups that have maintained a cellular backbone. Evolution within these groups should be well suited for study using sequence since there will be fewer confounding factors. This view is supported by the growing list of prokaryotic clades that form well defined sequence trees discussed in [40]. If we can identify the innovation between groups that leads to the differences in abundances it may be possible to polarize these transitions in the manner pioneered in [28]. This approach creates a timeline that is appropriate for classification purposes, and thus we approach the advantages of the TOL while reconstructing the TOC. Of course there will be traits that do not fit that timeline and we must consider them in a pluralistic fashion. However, the timeline will allow us to polarize many of the HGTs and place true innovations in their proper context. Current sequenced based methods could be made much less arbitrary by comparing them against these other lines of data. Combining the TOC with genomic

histories would capture all the positive aspects of the TOL hypothesis, while accommodating HGT.

It takes a universal sequence that has not been horizontally transferred and has evolved at a steady rate to build a universal tree. As discussed above that does not leave us with very much data. It takes two widely distributed paralogous proteins to polarize an indel [37], which leaves us with even less. Therefore we are very interested in non-ubiquitous traits that may be useful in resolving the branch order of the prokaryotic taxa. We have found protein structure to be a highly useful tool for studying evolution, but hopefully there are others as well. A transition within quaternary structure only requires a protein to be universal within a taxon of interest. We have presented two transitions in quaternary structure that exclude the root from the Archaea: Anbu evolving into the 20s proteasome [71] and PyrD 1A evolving into PyrD 1B [38]. Neither of these proteins is in Ciccarelli *et al.*'s dataset because they are not universal. But they are derived structures that are universal enough in the Archaea to provide compelling independent arguments that exclude the root of the TOL from within the Archaea. Even if the proteasome sequences have been horizontally transferred all over the Archaea it does not take away from the fact all Archaea have a proteasome (those would be horizontal displacements). Therefore a protein might be useful for resolving a branch order in the tree even if there is major incongruence between the cellular and genetic history. It is currently possible to predict the tertiary structure for about half of a prokaryotic proteome [72]. However, quaternary interactions are not being predicted from sequence in the same way. This gives us hope that there are still untapped sources of data that might resolve the branch order of the major prokaryotic taxa.

Saving the tree of life

We must keep in mind the humor of calling the central metaphor for evolution “the tree of life”. The phrase first appears in Genesis 2:9 : “And the LORD God made all kinds of trees grow out of the ground—trees that were pleasing to the eye and good for food. In the middle of the garden were the tree of life and the tree of the knowledge of good and evil“. There is irony in using the name of a tree central to the creation story to argue against that very myth. Therefore we doubt that any phrase will ever pack as much punch as the “tree of life”, even if the pattern of common descent is more of a web. It is very important that the community stops labeling any tree derived from a single data source the TOL. The recent attempts to resolve the TOC using primary sequence should be labeled “universal sequence trees”, a name that is grounded in the limitations of the data. The title TOC should be reserved for branch orders that are supported by several lines of independent evidence, and the TOL should be the synthesis of those branch orders and horizontal process.

Perhaps the most important line of reasoning that the TOL exists is the fact that HGT is so rampant. Why is HGT possible at all? The answer is obviously common descent. If it was not for common descent the genetic code would not be universal, and most HGTs would not even be translatable in their new host. Many biological parts are interchangeable because they have evolved in conjunction with the same systems. Therefore, we argue the very reason the TOC is so hard to reconstruct is because it exists!!

It has now become clear that many expectations about prokaryotic evolution were based too heavily on observations of eukaryotes. There is truly a fundamental divide in the way these two groups use the communal gene pool as a genetic memory [73]. However if we give up on the eukaryotes as a model, it is not clear what our expectation of prokaryotes should be. It does not make sense to us to criticize a tree as “the tree of 1%” without providing a justified cutoff of $n\%$ that would be enough for the vertical component of genomic evolution to be meaningful. We think the TOL crisis would be worse if it was the “tree of 99%”, as it would be quite difficult to explain the phenotypic differences between humans and *E. coli*. It is remarkable any genes are conserved since LUCA, and therefore the TOL still rings true to us.

Likewise, it is not clear what level of genome conservation between strains of prokaryotes would be satisfying enough to consider them evolving in a tree-like manner. It is true that two strains of the same species may have relatively few genes in common, but we have argued above this is probably an exaggeration of functional distance. Abundance data from different strains under similar growth condition will shed light on their true functional differences. We predict this gap will be much smaller than it appears from counting genes. Of course one reaches a point where two species live under different enough conditions that comparing their abundances is like comparing apples to oranges. But since expression is highly correlated with evolutionary rates [69] these are probably cases where primary sequence analysis would fail too. More abundance data may shed light on why some branches of the tree are so much harder to resolve than others. Fortunately species most likely to exchange genes horizontally live in similar conditions [58]. This means it will be possible to compare the relative contribution of

horizontal and vertical inheritance to the cell when protein abundance data are available from different taxa living in the same environment.

The landscape of genomes is rapidly being filled, and many higher level taxa are now well sampled. Despite this, there is no consensus on the TOL and many are ready to abandon the notion that we will ever reconstruct it. There is still plenty of data that needs to be generated to elucidate the history of cells. More information on protein abundance will shed light on the true revolutions in the history of cells and help prioritize conflicting signals in the genetic material. Protein and cellular structure will help us polarize the major events in evolution. It seems to us that genomes simply are not enough to study genomics. Of course it would be naïve to expect that some new data source will be a magic bullet that will resolve the TOC. Instead we must realize each data source has its shortcomings, many of which cannot be illuminated except in the context of other data. It is not just that we need more data, we need more details. Automated methods fall prey to numerous confounding factors but can still be highly informative. They must always be supplemented by experts whose intuitions have been tempered by careful examination of details from multiple data sources. Therefore the best way to move forward is to take sequence data off center stage and supplement it with these other data sources.

This view can be summarized by several dualities, best exemplified by the classic symbol of yin and yang (Figure 2). The basic lesson of this symbol is that one finds the darkest point in the center of the light, and the brightest light in the center of the darkness. It is only through understanding the interplay between the light and dark that one gains insight into their true nature. Neither can exist without the other. A new understanding of evolution comes from the study of interplay between a series of

dualities. It is now clear that there is large distinction between heredity of genetic material and of membranes. There is clearly a duality in Darwin's theory of descent with modification; the history of variation is well described by a network and the history of selection is well described by a tree. A web of life (WOL) may be more factual than a single TOC, but we argue it is a less accurate depiction of life's history. It is possible to precisely represent the relationship between most extant genetic material on this planet using a network. But without a tree (or time line) of life this undirected graph is mostly functional displacements and shifts in redundancy in our opinion. The "light" in the confounding "darkness" of horizontal transfer must be the TOC. Of course the "darkest" points are in the center of the "light" too. Endosymbiosis is clearly a non-vertical event that has profoundly influenced the structure of the tree of function. Likewise there are many horizontal innovations that were important for shaping the prokaryotic tree of function. We feel the most productive way to move forward is create a duality between the horizontal transfers that shape evolution and those that confound our tree building algorithms. The point is that neither of these extremes invalidates the other; they complement each other. Darwin wrote, "Thus, from the war of Nature, from famine and death, the most exalted object which we are capable of conceiving, namely, the production of the higher animals, directly follows." His understanding that death comes from life, and life comes from death fits perfectly with the symbol of yin and yang (the link between the quote and symbol of yin and yang was noted in [74]).

The increase in the size of the functional tree argued here may not be enough to persuade everyone saving the TOL is possible. Some prefer to be cautious and label this the "central trend in the forest of life" [57]. The fact that a large chunk of universal

cellular function has remained conserved and its sequence behaves in a mostly tree-like manner after billions of years makes the reconstruction of the TOC seem possible despite the issues of HGT. Resolving and rooting this tree are meaningful problems that are worth pursuing. That is not to say that the tree is resolved simply by taking a consensus or average of universal gene sequences. However, those who rather look at the forest life need to keep in mind that there are some trees in that forest that are much older and larger than the others. The central tree(s) must be the landmarks used to navigate the rest of the forest. HGT has clearly shaped the prokaryotic world, but if we do not keep in mind the histories of both genomes and cells we will end up lost in the woods.

Competing interest

There is no competing interest.

Authors' contributions

REV conceived the study and analyzed the data. PEB assisted in writing the manuscript.

Reviewers' Comments

Reviewer's report 1

Eric Bapteste, Université Pierre et Marie Curie, UMR CNRS 7138, 75005 Paris, France

This paper, in many respects well-balanced, proposes a strategy to reconstruct a tree of cells, and discusses questions regarding the TOL in ways, that in my opinion, should be significantly improved to be convincing.

The authors begin by acknowledging that genetic evolution is largely reticulated in prokaryotes, but that genetic and cellular evolution should be distinguished, since, importantly, cellular evolution could be accurately described by a tree.

They argue that the tree of cells could be reconstructed by (i) considering the distribution of functions (e.g. the repertoires of functions present in genomes) rather than the repertoires of genes ; and by (ii) giving a greater weight to genes that are abundantly expressed in cells (rather than giving similar weight to all genes) to define the branching pattern of cellular evolution. Genes with a greater abundance, they propose, would be less easy to transfer. In that regard, they introduce a novel (and possibly quite sound) ‘complexity hypothesis’ based on abundance. In the complexity hypothesis, genes with more interactions are expected to be less transferred, in the ‘abundant hypothesis’, genes with the most abundant expression are expected to be less transferred.

Whether the evolution of functions is tree-like as the authors repeatedly claim could be tested by reconstructing a phylogenetic network based on the functional content of genomes. The authors should do it in a revised version of the MS and add that analysis and a figure to their paper. This test of the tree-like evolution of functions would improve the paper, since in some parts of their manuscript, the authors mention the problematic

possibility that even functions evolution could be to some extent affected by HGT (e.g. p.19). In particular, the extent to which such repertoires of functions could be convergent to adapt to some environments (i.e. animal guts, or hypersaline environments) is probably partly an open question, that could complicate the interpretation of the branching pattern in such trees. Likewise, the authors mention that indels (another type of slowly evolving characters, in their view less affected by HGT) appear « on the same points on a tree ». It would be tempting as well to see how a phylogenetic network of these data on indels look (e.g. how tree-like the distribution of indels is), and how it matches with the tree of functions.

Author's response:

We are arguing that the tree of function requires a well developed tree of cells to define when lines of cells gained or lost function. Our other work focuses on reconstructing a large portion of the tree of cells, and hints of the tree of function can be seen in it (in preparation). In the tree of functions the source of the function does not matter. The ancestral line did not have the function and the derived group does, which is well described by a branch order on a tree. Whether other branches on the tree have the same function is irrelevant to those cells from a functional point of view unless they are in competition. To really build the tree of functions one would also need to represent the relationships between different functions. We think the history of cells, genomes, and functions are long term goals that cannot begin to be reconstructed in a single figure, but we hope we have argued they are goals worth pursuing still.

We agree it would be interesting to map characters such as indels onto networks.

Gene abundance, especially in extinct cells, may be quite difficult to quantify. The data are currently limited to decide which genes are abundant and which ones are not. Nonetheless, this limitation could be a chance, since it should be possible to test whether these abundant genes evolve vertically or not, by aligning these genes (and corresponding controls) and by searching possible traces of recombination in them, or evidence of inconsistent or odd branchings in their trees. Knowing whether these genes appear to recombine / transfer in proportions comparable with that of non abundant (control) genes would help evaluating the authors' claim that abundantly expressed genes are less affected by LGT. Such an analysis should also be added to a revised version of this paper. Indeed if molecular changes accumulate in these sequences largely due to non vertical processes, at some evolutionary depth, the proposition made by the authors that such genes would better describe vertical evolution than other markers (and thus should be preferred in case of conflict between markers) would simply be wrong.

Author's response:

We have added an analysis of the inconsistency scores for the highly abundant proteins. This data supports our proposal abundance is a barrier to horizontal transfer.

The authors' conclusions that the tree of functions (should it be consistent with the data, once reconstructed) is a good proxy of the tree of cells, itself a perfect match of the Tree of Life, is very arguable. The tree of functions, the tree of cells and the Tree of

Life can hardly be one same thing. They can hardly be considered isomorphic for a simple reason : they do not have the same explanatory powers, nor the same explanatory scopes. Evolution in general is much more than the evolution of cells, or the evolution of functions (even if these two aspects are very important to understand evolution). The problem is that biological diversity resulting from evolution by far exceeds these two aspects: many evolutionary units (recombined genes, operons, transferred genes, mosaic genomes, consortia, communities, ‘acellular’ and ‘intercellular evolution’ mediated by mobile elements such as phages and plasmids) cannot be exactly mapped onto a tree of functions or onto a tree of cells. The evolutionary fates of these objects are partly (and sometimes largely or totally) uncoupled with the ones described by the tree of functions or by the tree of cells. The tree of functions – if it can be reconstructed- would certainly be informative about the evolution of functions ; the tree of cells divisions – if it can be reconstructed- will be informative about a part of cellular evolution. However such a tree of cell divisions won’t inform us about most of what cannot be considered as mere details in evolution : the lifestyle, adaptation, processes creating and sustaining the genetic diversity, the selection pressures at play and the evolution of species (that is how remarkable groups of organisms emerged (or failed to emerge) from the interplay of evolutionary processes).

Author’s response:

You are absolutely right that each of these trees has different explanatory scopes. The tree of functions is certainly not the tree of cells. Our argument is that tree-like thinking is more useful when abstracting beyond the level of genetic material. One would

need to combine the tree of cells, tree of functions, and networks of genomes to get the explanatory power of the dreamt TOL. We are not pattern monists, so we have no problem with that.

We completely disagree that the trees of cells would not be informative to studying adaptation. It might not provide much explanatory power on its own, but how can one study adaptation without a history of the cell ? The tree of cells provides the snap shots of before and after the adaptive process. Mapping any other evolutionary data source back onto the tree of cells makes it more informative. Without the history you cannot say what any of the processes you listed actually changed in a cell. The bottom line is the contribution of these other factor is always on the tree on cells. That is why we argue the tree of cells becomes more important under a pluralistic model.

This limited explanatory power of such a tree is even clearly demonstrated in this manuscript : the ‘cellular core’ of four spirochetes is already uninformative about the Spirochetes lifestyles. Using these ‘abundant genes’ would not allow explaining much of the spirochetes biology, and thus of these species origin (how some become obligate parasites for instance). The duality between the tree of cells and the network of genes, well acknowledged by the authors, seems irreducible, because real and relevant to our understanding of evolution. In other words, while the authors rightly argue that the genome and the cell should not be confused with one another, they seem to be tempted to

approximate the entire biology and the entire evolution by the history of cell divisions. This confusion too should be discouraged.

Author's response:

The tree of function is not just the abundant genes; it would include the losses and gains that define the adaptations to these different environments. We are not arguing cell divisions are all that matter, but rather their history is necessary for a true understanding of these other processes.

It would certainly be interesting to polarize in time many evolutionary scenarios, but it does not follow that, based on the history of some genes with a slower dynamic and based on some 'frozen' features, we will be able to infer the independent histories of the other genes and of the other organismal properties evolving with a distinct (faster) pace. In that respect, knowing the tree of cell division might not help much in understanding precise gene histories (for instance). A tree of cells will have some useful explanatory power but not as much as a dreamt TOL.

Author's response:

We agree that some genes histories will still not make sense even in the framework of something that resembles the TOL. Our point is that without some branch order for the major prokaryotic groups it becomes difficult to come up a meaningful history for ANY gene. The tree would let us differentiate the slow and fast properties,

which would give great insight to evolutionary processes that are either tree-like or network-like.

The authors' choice « to keep using the phrase 'tree of life' when referring to the pattern of common descent even if it is more of a web », because this would somehow annoy creationists, is in my opinion not a good idea. I do not think creationists should dictate us any of our scientific agenda, or influence our wording, as they have no scientific competence to evaluate evolutionary studies. When phenomena are not tree-like, we should not call them a tree. When they are tree-like and are supported by several lines of independent evidence, we should call them the 'most corroborated evolutionary tree' or the 'best evolutionary tree' but not the tree of life, because maybe features did not evolve in a tree like fashion, and thus cannot be reduced to that scheme to be fully understood.

Author's response:

If this was our only reason for this title you would be absolutely correct. We have presented many reasons besides this why we think 'tree of life' is a worthy title for the combination of the histories of cells and genomes. It is not about annoying creationists, which comes quite naturally to us. We agree that they should not dictate our agenda, but clearly they have already shaped our wording. The point is 150 years ago the phrase 'tree of life' invoked a vision of a talking snake and a magic apple. Now it is a story that involves genomes, viruses, and algorithms. The meaning of the 'tree of life' will continue to evolve, but it will continue to provide an explanation of where life came from. We

don't think any other title could ever have quite the same aesthetic value, but beauty is in the eye of the beholder.

There is then a cost to do as if the Tree of Life existed (but not testing this scientific hypothesis) : it reifies parts of the tree, like the nodes and the branches.

Lawrence and Rechtless have masterfully shown that nodes, when conceived as points of speciations, are not 'real'. When prokaryotic species do not evolve by a series of dichotomies, it is a delusion to impose a dichotomy to describe a speciation.

Author's response:

Lawrence and Retchless [75] have demonstrated that nodes are fuzzy in terms of genetic material due to varying levels of recombination during the divergence process. If we consider a tree of functions then the nodes are real. Consider an ancestral state that lacks a function. The function is gained (through HGT or innovation). There is now a derived and ancestral node that can be described on a tree, but the history of their genetic material is no longer so well behaved. Their may be a fuzzy functional intermediate, but that would not be a stable state due to selective pressures. Again trees appear a better data structure if we abstract past the genome.

The root poses a comparable issue. The authors keep referring to LUCA, as if there were one one last universal common ancestor of all life that was a cell. The literature on early life is, to say the least, divided about this notion. Invoking LUCA to

prove that there is a tree, and the tree to prove that there is a LUCA, without any principled way (or any test) to refute that there is a LUCA or that there is a tree is unfortunately a circular argument. That all cells share a given gene/function does not mean that all cells evolved vertically. If LUCA ever existed (which I doubt, and most importantly which in my view could explain more than a part of early evolution), how big was its pangenome? What kind of mobile elements drove its evolution? We need to make sure that a prioris about the tree and about LUCA being real do not already bias our conclusions, if these a prioris cannot be tested. Otherwise, we might reinforce our habits of tree-thinkers, but not necessarily our knowledge of evolution per se.

Author's response:

The issue of LUCA is certainly muddled in the literature. Where direct the readers to a recent empirical argument that there must have been a LUCA, even if it was a community [76]. You are right in pointing out we are biased in our view of LUCA. Our other work has led us to support Cavalier-Smith's assertion the Chloroflexi are the earliest branching extant group. If one accepts the canonical rooting between the Archaea and Bacteria the idea of a large pan genomed LUCA is certainly more appealing because of the large differences between the prokaryotic superkingdoms. If LUCA was a large pan genome our focus should be on what genes could NOT have been in that community and must be younger than LUCA.

In that regards, I have a few detailed suggestions where some simplifications could maybe be corrected in the text.

p.3 : The authors wrote : « If every daughter cell's membrane kept track of who its parent was, reconstructing evolution would be a trivial task. » I feel that this is a bit misleading, it would inform us on a part of evolution : « If every daughter cell's membrane kept track of who its parent was, reconstructing evolution **of cell divisions** would be a trivial task. » It would not tell us anything on how species taxa, and genes, and phages, and communities, etc. evolved.

Author's response:

A more precise statement which we have adopted.

p.3 : It is very arguable that « even genomic evolution makes little sense without the light of cellular evolution ». There is certainly lots of knowledge to be gained from metagenomic analyses, from the study of mobile elements, from the study of gene evolution, lots of patterns and processes to explain, even without the light of cellular evolution. This is not to say that we would not benefit from that particular light. But this light will mostly make a 'genealogical' sense on evolution, and evolution is more than genealogy.

Author's response:

Studying evolution without genealogy makes little sense to us. What is the point of studying the process if we cannot point to examples where the process occurred? The study of the process must be grounded in the history.

p.4 : The TOL does not become even more important under a pluralistic approach, quite the opposite : it is regionalized under a pluralistic approach, as no single model can explain everything about evolution under that perspective. Finding the tree of cells for instance remains an important and ambitious goal, but not the alpha and omega of evolutionary research. The importance of the TOL thus decreases while the importance of additional interesting evolutionary questions increases.

Author's response:

We have changed this to TOC in the text. We are arguing building it is the alpha, but not the omega. The fact that the TOC will be used to formulate more questions than initially expected makes it more important.

p.5. It seems to me that proteins abundancy and core cellular features might be a basal make-up of cellular lineages on which further adaptations are adjusted. If there is some ratchet, abundant proteins can not be easily gotten rid of, but that does not mean that most of the evolutionary dynamic concerns these proteins and their coding genes.

Author's response:

Evolution is certainly not just about the abundant genes. We have included abundance data to add a dimension beyond the genome to study the big picture in evolution.

p.5. « Comparing how many of the same functions two cells are doing at the same time is a good measure of similarity », but is it a good proxy of the genealogy ? This depends on the amount of convergence and selective pressures on functions induced by the environment. Is not it possible that bacteria of the gut microbiomes (or of a salty environment) will perform the same functions at the same time even if they are not directly related ?

Author's response:

This is certainly a vital question to answer. Abundance data from similar environments will allow us to test this in the future, but for now we are left to speculation.

p. 6. For a detailed argument that a TOL is not as meaningful as claimed in [13], when there is a large incongruence between individual gene trees, see « Baptiste et al. Biol Direct. 2009 Sep 29;4:34. Prokaryotic evolution and the tree of life are two different things ».

Author's response:

We lean towards your arguments in this case unless one can deal with the incongruence in a non-arbitrarily manner as we have proposed here.

p.11 : I absolutely disagree with the following statement : « Trees have many shortcomings for representing prokaryotic evolution, but any data structure that lacks temporality is even worse. » Reconstructing a wrong tree (imposing an irrelevant structure to the data because of our a prioris) is the worst thing one can do. We can learn a lot from unrooted gene trees, on which by definition there is no temporality.

Author's response:

We mean that we need the benefits of the trees we have listed, while trying to accommodate the shortcomings of that representation. We are not forcing data to fit the tree. We are saying the data that does not fit the tree can only be noticed and studied once you have the tree. There are certainly other representations that are useful, but the TOC is necessary for weaving them into a coherent story.

p. 18 and p. 19 : When commenting on dualities, the text becomes pretty metaphysical in places. I do not see why 'the' light must be 'the TOL' : a light can be a tree of cells, another light a tree of functions, and so on. Why do you need only one light to explore the darkness of evolution anyway, while so many processes are occurring, creating a diversity of phenomena that calls for more than one explanation ?

Author's response:

We think a discussion of the metaphor used to represent the process of life should have a healthy dose of metaphysics. The nice thing about yin and yang is that you can

switch them and it still tells the same story. We have chosen light because in this case the answer the TOL provides (if it exists) is more directly readable than the WOL. There should be multiple lights, but we cannot see how anything could be more useful to understanding the history of genes than understanding the history of cells. These are the two primary replicative processes in evolution. The assumption they were the same process is the source of the problem. As we begin to separate them we must keep them connected. Therefore we feel the yin and yang is fitting. One could argue the confusion caused by forcing data to fit a tree is the darkness and the light is the realization of processes like HGT. The point is that HGT and the TOC are inseparable and cannot exist without the other, regardless of how we label each one.

A thinking about evolution in terms of yin and yang is possibly not entailed by the quote on (I believe) Darwin's malthusianism. I doubt that historians of sciences and / or philosophers of sciences would be convinced that this is a proper use of that particular quote. I do not think it is needed in the paper.

Author's response:

We probably did go a bit too far in our use of this quote. We have softened the implications, but have kept the quote. We see the yin and yang in that quote, as well as in the data, regardless of what Darwin was thinking when he wrote it.

p.18 : I disagree with this sentence : « A web of life may be more factual than a single TOL, but we argue it is a less accurate depiction of life's history. ». The authors possibly have in mind a fairly simple web of shared genes. But even these graphs can be further studied to gain knowledge on history. Dagan and Martin for instance have shown how such networks can be exploited to learn about life's history. And what about phylogenetic networks for taxa with a limited amount of HGT : are they worse than a tree to describe life's history? To me this kind of claim is counterproductive, as it fails to acknowledge that it might just be time to change our habits and our thinking about how evolution should be described.

Author's response:

We are making a similar point to the one you made above about explanatory power. The WOL does explain as much as the dreamt TOL. Therefore we would be settling for too little if we thought it was enough.

I also would like to make some further precisions :

p.2. The conclusion of Doolittle and Baptiste is not so much that the inability to build the tree is the problem, rather it is forcing the data into a tree that needs to be questioned, and in a pluralistic framework, avoided, since this model does not allow a precise description of the evolutionary processes.

Author's response:

Changed.

p.17 : The authors write that « it is not clear what level of genome conservation between strains of prokaryotes would be satisfying enough to consider them evolving in a tree-like manner. It is true that two strains of the same species may have relatively few genes in common, but we have argued above this is probably an exaggeration of functional distance. » Even if the second sentence might be correct, its association with the first one suggest that the authors tend to overlook the importance of recombination in prokaryotic genomes, a major process that is not tree-like. This non-tree like phenomenon can in part be masked by zooming out at a higher taxonomical level, still the real processes responsible for evolution are not tree-like. In that respect, a tree of cells or a tree of functions will fall short in explaining major evolutionary processes at play on genomes.

Author's response:

We are trying to emphasize the many tree-like patterns that could be masked by recombination of genetic material. Certainly a network is needed to understand the history of the genome. Our key point is that does not mean a network necessarily describes the evolution of cellular function better even in the presence of recombination.

p.17 : Just like « genomes simply are not enough to study genomics », cells (or functions) are not enough to study evolution : you need to include phages, plasmids, etc.

Author's response:

True, but in reality phages and plasmids only really affect evolution when they enter cells.

p.19 : The duality that the authors propose between « HGT that shape evolution » and « HGT that confound our tree building algorithms » seems a distinction between good and bad HGTs. This distinction (if it can be achieved, how ?) could help them building an evolutionary tree, but it would not make the processes of evolution and prokaryotic genetic evolution more tree-like in nature. Both good and bad HGTs are non strictly vertical processes.

Author's response:

This distinction can only be made by having a hypothetical ancestral genome before the transfer occurred, which is not a trivial task is given the amount of transfers that have occurred. If the function is novel to the recipient cells lets call it a good HGT. In both cases there is non-tree like evolution, but we argue that bad HGTs are just red herrings leading us away from the tree that does exist. Our point is that HGTs are not strictly non vertical from a functional perspective.

Overall, in many places of the MS the authors could advantageously replace TOL by Tree of Cells, which would address (simply) most of my concerns.

Author's response:

Calling the TOL a universal sequence tree in much of the current literature would have addressed many of our concerns as well, but your arguments have convinced us TOC is more precise in several places. We think a compromise is to use the term TOL to refer to the combination of the network of genomes and tree of cells. It would not be entirely a tree, but it would have nearly same explanatory power as the original TOL hypothesis. We hope this work adds precision to these terms instead of just muddying the waters.

To sum up, I feel that the current title of the manuscript is misleading, unnecessarily dramatic, and should be modified.

The title is meant to be dramatic. We have explained what mean by 'lost in the woods' a little better in the introduction. Now that we have changed TOL to TOC in many places we explain the title as rescuing the explanatory power of the TOL by remembering the WOL needs to be grounded in the TOC. We feel the need to be dramatic because many appear ready to abandon the TOC because it is confused with the TOL.

What this MS proposes is how a tree of cellular functions, equated with the tree of cells, could possibly be reconstructed by taking into account additional (novel) sources of data (such as the functional repertoire of genomes and the abundance of expressed genes in the cells) rather than by focusing on the mere gene content of genomes, and by giving comparable weights to the phylogenetic signal(s) of each individual gene.

The Tree of Life and the tree of cells are however two different things : in particular they do not offer similar explanations of evolution. The tree of cells is by definition more limited in its scope than the legendary Tree of Life. It is then important to stress that the tree that could be saved if the author's proposition hold is (and that is already quite good) either the tree of functions or the tree of cells.

The title also suggests that evolutionary biologists would be lost without this one tree. I think this claim is unduly pessimistic, and stems from our acquired habits to explain evolution with a tree model. Evolutionary biologists won't be lost without the tree of life : they will be challenged. They will need to reconsider their practices, their goals, and their explanatory toolkits to make sense of an evolution that is not just tree-like.

There are lots of fascinating researches to be done to learn about the evolutionary processes and mechanisms, that do not require the inference of a unique tree of life , i.e. to harvest the phylogenetic forest of unrooted trees (see Lapointe et al. Trends in Micro, in press), or to exploit genome networks (see Dagan and Martin's works ; Fani, Fondi et al.'s works, Lima-Mendez, Leplae et al.'s or Halary et al.'s works). Our explanations will be different, but evolutionary biologists won't be out of job or hopeless. Such a possibility could / should have been explored more by the authors, as they reckon that « it is not clear what our expectation of prokaryote should be ». Precisely, clarifying this expectation, with the least possible a prioris, is an exciting prospect for evolutionary science.

Author's response:

Processes and mechanisms are certainly important, but evolution is about history to us. We only care about the mechanisms and processes because they caused the history. You cannot understand the mechanisms or processes without the history. We certainly believe that networks have a lot to teach us, but they only are meaningful when grounded in a TOC. Therefore the TOC becomes more important as we try to understand the processes that do not fit into that scheme.

This is why I finally beg all authors who might be tempted to send me some more papers to review on themes such as ‘rescuing the TOL’ or ‘saving the TOL’ in a near future not to: I have definitely said all I had to say on that issue for a little while, and it is time for me to move on more exciting research topics ☺ .

Author's response:

We sincerely thank you for one last round on this subject. Unfortunately for you, your insightful review of this manuscript will probably make others want to continue this discussion with you. But we understand the need to move on from this topic.

Additional specific comments :**Further questions :**

p.8 : Do large pangenomes have larger repertoires of functions ? If so, won't that affect the reconstruction of a tree of functions ?

Author's response:

A large pangenome may have a large amount of functional redundancy. This will not be a problem if one has a good functional outgroup, but that requires a well defined TOC.

p. 9 : Why should we assume that COGs that are the most widely distributed in extant taxa are the most ancient ones? Why can not they be highly transferred ?

Author's response:

You are right that some of these could be the result of frequent transfers. To the best of our knowledge there is no case of a young protein being transferred to the majority of a superkingdom, but there are many proteins clearly in the ancestor of a superkingdom that have been retained. Therefore in general the most widely distributed proteins are the oldest. It seems we would need the tree of cells, tree of functions, and network of genomes to be certain though, so for now this is a reasonable estimate.

p. 19 : « The fact that a large chunk of universal cellular function has remained conserved and its sequence behaves in a mostly tree-like manner after billions of years earns it the title of TOL in our opinion. » How is this a fact ? How has this been tested ?

Author's response:

We have changed this one to TOCs to soften it. We are calling the 5% of the 'abundome' represented by the universal proteins a large chunk, which is certainly arguable. However, this number increase if one considers the functional content of the last common ancestor of each superkingdom instead. The nearly universal trees in [57] have a high level of consistency. Therefore, we think this statement is justified in its current form.

Unclear sentences :

p.1. 'results' : What do you mean by « **proper** selective context » ?

Author's response:

Differentiating between whether it is a good or bad horizontal transfer.

p.2. Why should the words be 'words of **neutral** evolution' ?

Author's response:

Because we believe most of the noise coming from HGT are actually just displacements.

p.5. What do you mean by : « The downside of abundance is it is dynamic, while genomes are static. » ? What time-line / evolutionary scale do you have in mind ? At the TOL level, genomes are very dynamic.

Author's response:

It is true on evolutionary time scales genomes are dynamic. We mean within a single cell.

p. 12 : This entire section: « The disagreement between these phylogenies is not in terms of how to define the major taxa but rather in the proper way to polarize the data, especially the indels (insertion deletions) which we have discussed [38]. However, the distribution of these traits themselves implies specific taxa evolved before another, regardless of the direction of each polarization. For example, there is a large insert in HSP70 (heat shock protein 70) that is present across the Gram-negative bacteria, but absent in the Gram-positives. One form of the protein must have predated the other. There is no reason to assume all the informative indels were fixed early in evolution, and one would be very hard pressed to draw a detailed scenario of transfers that explains their distribution better than a more timeline like structure. » is unclear, and should be somehow rewritten. If this is a philosophical point (rather than an empirical comment on the data distribution), I would say that the best explanations are not necessarily the ones

that match a tree, these latter are only the simplest explanations. When irrelevant, they do not help much.

Author's response:

This is an empirical comment. We arguing if HGT was truly so rampant as to annihilate any trace of the TOC it should not be possible to find independent traits that support these phylogenies. We have rewritten it try to make our point clearer. We see no need to invoke a more complicated explanation if the simple one works.

p.13 : « Abundance data may make it possible to quantify what Simpson coined “quantum evolution” when referring to the metazoan fossil record [66] on a molecular level in prokaryotes. » This sentence needs to be developped or better explained (as it is I do not recognize Simpson's theory – that gives a main role to the environmental selection in quantum evolution- if I recall correctly, as a particularly valid analogy here).

Author's response:

We have inherited our use of this term from Cavalier-Smith. To be precise we mean events where there is a domino effect across numerous proteins that results in rapid evolution. We are arguing that if some major change in abundance was tolerated by rapidly shifting the abundance of other proteins it would very difficult to resolve with sequence data regardless of HGT,

p. 16 : This highly finctional sentence makes no sense whatsoever to me : « We think the TOL crisis would be worse if it was the “tree of 99%”, as it would be quite difficult to explain the phenotypic differences between humans and *E. coli*. It is remarkable any genes are conserved since LUCA, and therefore the TOL still rings true to us. »

Author's response:

We are saying there tree of 1% argument makes no sense without a null hypothesis. This purely fictional sentence is an example of another tree we could be dealing with that would cause a different set of problems. Put another way, what % did the community expect to be conserved before genomic sequences were available and why?

Reviewer's report 2

Arcady Mushegian, Stowers Institute for Medical Research, Kansas City, Missouri, USA.

I have read the manuscript by Valas and Bourne with considerable interest, wholeheartedly agreeing with several ideas in it and disagreeing with some. The best home for this study is probably in the *Opinion* category within *Biology Direct* – this is not really a research paper.

Author's response:

We feel the paper is both research and opinion, and hopefully it will fit fine in either category.

There are two main themes, one of which is more of a research proposal, the other more of a philosophy-of-science talking point. The research proposal is essentially to enhance the utility of genic traits by assigning weights to them – the weights which, directly or indirectly, estimate relative contribution of each gene to the phenotype; if I understand the proposal correctly, the significance of the phylogenetics signal can therefore be reordered by the “functional rank” of the sequence from which this signal was obtained. I think this is a good proposal, and Adami/Wilke and Koonin’s groups, among others, have already said a lot about gene “relative importance”; important technical details of all that have not been worked out, however. The more methodological question, of what to make out of the purported lack of tree-like signal, or of the Doolittle and Baptiste’s “pattern pluralism” and other related proposals, is also of interest, and my intuition runs close to the author’s, but I still think that he is led astray by the setups of the problem in the literature.

In more detail, much of the “conceptual” literature on the HGT is hand waving about “rampant”, “massive” etc. aspects of horizontal transfer. This usually refers to the large number of events observed in a particular dataset, but generally fails to acknowledge that this high *number* of events usually accounts for a small *proportion* of the genes in the dataset and correspondingly relatively *low average ratio* of horizontal to

vertical branches in the trees. (Ninety-nine percent of the trees, for example, may show some evidence of HGT, but in the vast majority of these trees, there may be just one or very few HGT events, and so on; see, e.g., Pubmed 19077245, 18062816 and 15799709). Thus, instead of talking about the applicability of the TOL “metaphor”, perhaps we should be talking about TOL quantitative model, the alternatives to it, and which model or mixed model is best compatible with the data.

Author’s response:

We agree with your sentiment towards these results. We are all for a mixed model, but it needs to be a true duality where cells and genomes are treated as such, instead of just a reticulated network of genomes.

The author states about Doolittle and Baptiste’s proposal: “A key point of their work is that any data can be forced to fit a tree, even if that representation of the data makes no sense” – in fact, this has been known for a long time; the same can be said about any representation of the evolutionary process (e.g., alignment algorithm will align even unrelated sequences, and network algorithm will build a network even on a hierarchical set of OTUs); and finally, so what?

Author’s response:

We have changed this sentence to one suggested by Eric Baptiste in his review. The other two examples you bring up are valid. It is easy to forget the results of high

powered computing tools we have still are prone to 'garbage in garbage out'. Some alignments are probably forced to fit, but the authors are saying ALL universal trees are forced to fit. It took their persuasive argument to demonstrate how forced many of the genome trees are. It is the scale of the problem that makes their work worthwhile.

In other words, the author should stop fighting the windmills: the goal of phylogenetics should not be to build a tree, nor to build any graph with another kind of predetermined topology, but to improve our understanding of which evolutionary events actually happened and led to the observed data. I think this accommodates “pattern pluralism” naturally.

Author's response:

We agree, but one must have some data structure in mind when designing algorithms and strategies to reconstruct these histories. We feel trees capture the history of the events better, and that is needed to supplement the networks to get anywhere.

I would advise to the author to get all this out of the way early and to focus on the relatively independent proposal of including protein abundance and other information (such as perhaps correlated essentiality and degree of conservation) into the judgement of importance or relevance of any particular tree topology for phylogenetics. I would like to

see the discussion of several points in more detail.

Author's response:

We'd rather let the reader see our wild speculation and eastern symbolism after a little well grounded research.

1. "A cumulative plot of genomic and cellular abundance reveals that at every level genomic abundance underestimates cellular abundance" – so what does this tell us about phylogeny? Also, the datasets that are available to us are full of parasitic microorganisms whose genomes may be experiencing net gene loss, which contributes to the reduction of the "genomic abundance" of almost all categories of genes. Would the picture change if we focus on free-living/saprotrophic organisms?

Author's response:

This implies that when a large group of trees is in agreement about phylogeny that forest is large portion of those cells. That makes the phylogeny more historically real to us. It would certainly be interesting to focus on abundance in parasites and their free living relatives. We assume that most of genes they retain would be highly abundant proteins in free living cells, and they mostly streamline what is usually necessary to power that core. That would be consistent with higher levels of conservation when measured by protein versus genomic abundance in the spirochete data set, but more data here would certainly be informative.

2. The authors want to rescue the tree by bringing in the functional importance/protein abundance (phenotype), as discussed before. In this case, would not the change of function be equivalent to a HGT, and would this be less or more often than a true genetic HGT?

Author's response:

Functional changes are never neutral, and we argue HGTs are neutral most of the time so they are not equivalent. It is hard to speculate on the frequency of such events because so there are so many ORFans in sequenced genomes, and so many proteins beyond that which have not been functionally characterized. If that portion of genomic space represents nuanced novel functions they could be more frequent than HGT. We think this is exactly the sort of question that requires both a tree and a network to answer properly.

Reviewer's report 3

Celine Brochier, Laboratoire de Chimie Bactérienne, CNRS-Aix Marseille Université, 31 Chemin Joseph Aiguier, 13402 Marseille Cedex 20, France

In this paper Valas and Bourne propose an original approach to reconstruct the tree of life.

To my point of view this contribution is more an opinion than an experimental paper. This is at odds with the organisation of the paper that includes a large “results” section (11 pages), whereas the real experimental part of the work is represented by a single figure and one paragraph (1 page). By contrast, the discussion” section is rather

short (3 pages). I think it would be more appropriate to combine the results and the discussion sections into a single section, with subdivisions corresponding to the different points that are discussed. Finally, I think it is important to clearly classify this manuscript as an opinion and not as an experimental paper.

Author's response:

We have combined the results and discussion as suggested. We agree that this paper is not a traditional research paper, but we still feels it belongs in that category as it is a combination of novel research, opinion, and review..

The contribution of Valas and Bourne comes within the scope of the hot debate around the Tree Of Life (TOL). Indeed, based on genomic data the suitability of tree-like structures to represent the evolutionary history of all organisms has been highly debated [1-3]. The two main arguments are that in prokaryotes (1) the evolutionary history of genes is different from the evolutionary history of organisms because of horizontal gene transfers (HGT) [2], and (2) HGT may be so frequent that a substantial part of the genes in a genome have been affected by HGT. Then, jumping from genomes to organisms (perhaps because we have entered in a “*too genomic-centric*” area, as stated by the authors), this has led to the conclusion that, at least for prokaryotes, a tree-like structure does not reflect the evolution of genomes, which will be better represented by a network (NOG, Network Of Genes). However, if nobody can deny that HGT have played an

important role in evolution (and not only in prokaryotes), it is also undisputable that cell division in prokaryotes occurs by the division of a mother cell in two daughter cells. It is therefore theoretically possible to trace-back the history of cell lineages and to represent it with a tree-like structure, the TOL. However, the TOL and the NOG are often confounded, maybe because genes are the only informational entities that are transmitted from one generation to another, whereas TOL and NOG represent two different things that are equally interesting and highly complementary to understand the evolution of living organisms [4, 5]. The authors state these points well by writing: “*All of the issues the community is currently having with the TOL hypothesis stem from the simple fact that genomes are not perfect representation of membrane history. [...] even genomic evolution makes little sense without the light of cellular evolution*”.

The challenge is now to reconstruct the TOL in a NOG context. Classical approaches consist to identify (and to analyse) the sets of genes that may be used to reconstruct the different parts of TOL. This step is important because it appears a utopia to think that it is possible to fully resolve the TOL (from the root to the leaves) based on the analysis of a few universal genes. This would be the tree of 1%. In fact, it would be cleaver to divide the problem by looking at the set of genes suitable to reconstruct different parts of the TOL. For example, the set of genes suitable to resolve the phylogeny of animals will probably be different to the set of genes that may be used to trace back the relationships within Methanococcales (Archaea). This is well known by botanists and zoologists who used different sets of characters for different levels of their classifications. Ideally, the TOL should be a synthetic drawing showing the relationships

between organisms (not species, which are artificial entities, or genomes) by combining the results obtained by the phylogenetic analyses of different sets of genes. In this case, the TOL will not be the tree of 1% but the tree of dozens of percents, each gene contributing to resolve some parts of the TOL.

Here the authors propose a radically different approach based on the vertical inheritance of functions rather than on the vertical inheritance of genetic material. The approach is based on the assumption that all genes do not contribute equally to the cell: some are more important than others. The authors underline that the gene contribution to the cell should be an important criterion to take into account when reconstructing the TOL. As the authors point out there are different metrics to measure the gene contribution to a cell: “*essentiality, abundance of proteins, number of transcripts, portion of total weight, etc*”.

1) My first question is how to organize these factors in a hierarchy, i.e. which criterion is the more suitable to represent protein importance in a cell? And what do these factors exactly represent from an evolutionary point of view? Is it possible to develop evolutionary models for such data (that are mainly quantitative and not discrete characters)? Unfortunately the authors do not propose methodological approaches to analyse such data. I think this is important to discuss about their suitability to reconstruct the TOL.

Among these factors, the authors chose to study the abundance of proteins in cells (the “*adundome*”). Based on recently published data on the abundance of proteins present

in the cytoplasm of *Escherichia coli* cells (GammaProteobacteria) and of the complete proteome of *Leptospira interrogans* (Spirochaetes), the authors argue that “*abundance is a good proxy for evolutionary importance because there is a correlation between the abundance of a protein and the energy the cell invests into producing it*”.

Author’s response:

*We do not have precise answers for these questions, but they are certainly going to be important to answer. Our demonstration that abundance is a barrier to transfer supports the notion it is evolutionary important. But there are abundant genes that have been transferred. It would be naïve to say those are less important. It might not be possible to precisely quantify how important each gene is to the cell, but we have demonstrated that in general the important genes evolve in a more tree-like manner. We are hopeful it will be possible to develop evolutionary models for the evolution of “abundomes”, but we doubt they will behave well enough to resolve the TOL or TOC on their own. Rather we think they will be tools to help us understand how the TOC was shaped. In either case it would certainly be premature to begin developing these methods from the two datasets currently available as they are not directly comparable. A sampling of many strains of *E. coli* seems like a good place to start addressing these ideas.*

2) This raises my second question: what does the “*evolutionary importance*” of a protein mean from an evolutionary point of view? The word “*importance*” is a subjective

and indefinite criterion. The abundance is one side of the importance, essentiality is another. Indeed, a protein may be important even if it is not abundant in a cell (e.g. transcription regulators). The authors should discuss more this point.

Author's response:

Importance is certainly a subjective term. We are not arguing that abundance is a perfect representation of importance but it seems to be a straightforward and objective measure. We argue that as we measure the importance of function in some meaningful way the importance of HGT will begin to shrink, and the vertical component will grow in size.

3) More problematic, and this is partially raised by the authors, the abundance of a protein is a dynamic parameter that may vary across cells depending for example of their lifestyle. More importantly, for a given cell the relative abundance of its proteins may vary in time, depending for example on the developmental state reached (e.g. cells in exponential growth or in stationary phase, etc) or environmental conditions. How to take this difficulty into account?

Author's response:

There are certainly going to be many difficulties in using abundance data. We argue the proteins that remain abundant under a variety of conditions are probably the

most important, but surely there will be many interesting caveats to discover as more data become available. Again, we are not suggesting using this data to build phylogenetic trees, but rather as a tool to better understand the ones created from other data sources.

4) I think the large paragraph on indels and the timing of appearance of prokaryotic phyla should be removed because it is beyond the scope of the paper, and I am afraid that the reader will lose grasp on the logical succession of ideas. Same remark for the paragraph dealing with protein structures. On the contrary, I think the authors should rather focus on their proposal to use “abundome” to reconstruct the TOL and in particular on the methodological aspects.

Author’s response:

We are not arguing the “abundome” data can reconstruct the TOC, but we think the indels and quaternary structures can. In some sense we are using abundance data to show the phylogenies created using other data sources are meaningful despite the arguments against the TOL. Most of these arguments are against sequence based methods, so we think it is appropriate to include our other work as a demonstration the TOC is still evolutionary meaningful and can still be resolved.

5) Finally, I have a few comments regarding assumptions that are made on phylogenetic studies based on gene sequence analyses. The authors say that “*current methods for estimating HGT rely on measuring inconsistencies between sequence trees or looking for unusual compositional features, so there is no way for them to distinguish between innovations and displacements*”. I think this statement should be toned down because, in the case of phylogenetic studies, most of the time a careful examination of trees allows discriminating between gene acquisition and gene replacement. This is for example the case for aminoacyl-tRNA synthetases (that are discussed in the paper), where clear cases of gene replacements can be identified.

Author’s response:

The aminoacyl-tRNA synthetases are a special case because the combination of their trees and knowledge of the essentiality of their function implies these are displacements. It is much more difficult to conclude that from the trees alone. It is not impossible to discriminate between these scenarios, but it seems like many do not seem to worry about the difference when looking at forests of phylogenetic trees.

Very minor points:

I do not understand the sentence “*We argue events like this are far more deleterious to tree reconstruction algorithms than they are to the recipient cells.*”

Author's response:

We feel that sequence has persisted as the primary tool to study evolution because of the relative ease algorithms can represent it, as opposed to these other sources of data. Again we are trying to emphasize the difference between displacement and innovation needs to be made by including other data sources.

The legend of Figure 1 is poorly understandable.

I did not understand the last sentence of the abstract.

Author's response:

See our reply to Eric Baptiste.

I disagree with the allusions to Darwin in the discussion section “*There is clearly a duality in Darwin’s theory of descent with modification; the history of variation is well described by a network and the history of selection is well described by a tree*“. First, this sentence is not clear. Second the history of variation may be represented by a tree: for example, the evolutionary history of a gene (irrespective to HGT) may be depicted by a tree and it is possible to indicate on each branch the mutations that occurred, and therefore to follow the history of variation of this gene.

Author's response:

We are speaking in very general terms. Your example is correct, but the variation of organisms is in terms of their entire genomes. The gene cannot really be selected for independently of the rest of the cell and genome, so the variation is relative to them as well. It was assumed for a long time the history of these two processes is the same, and we think it has become time to explicitly separate them.

Acknowledgements

We would like to thank Russell Doolittle, William Loomis, Stella Veretnik, and the entire Bourne laboratory for useful discussions. We would like to thank Pere Puigbo for providing the inconsistency score data.

Figures

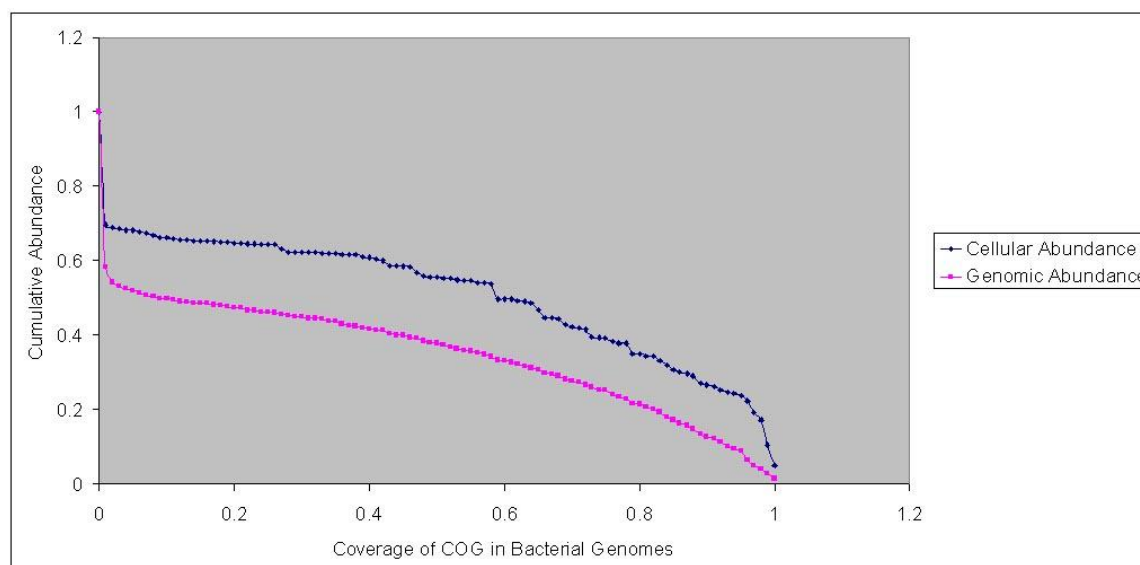


Figure 2-1. Genomic vs Cellular Abundance in *L. interrogans*.

Each COG's presence in bacterial genomes was plotted against the cumulative abundance of all COGs that are present in at least that many genomes. The conserved genomic core is always an underestimate of the conserved cellular core, in some places by as much as 20%.

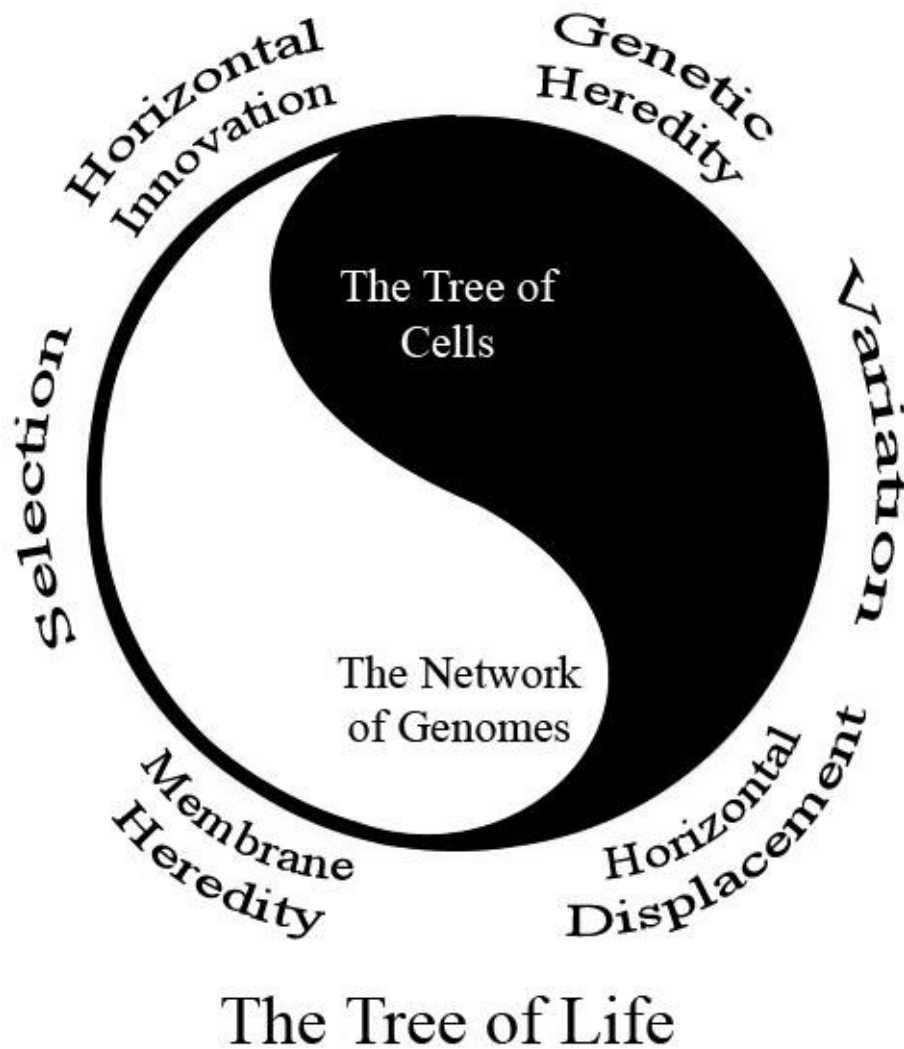


Figure 2-2. The yin and yang of evolution

Several key dualities in evolution are better understood when they are viewed as complements to each other under the framework of the classic symbol of yin and yang. Common descent is the prerequisite for HGT, but horizontal innovation shapes the pattern of descent. Inheritance of genetic material is often web-like, but membrane heredity is tree-like. Both polarities of each of these dualities exist because of the other. The existence of darkness does not invalidate the existence of light, just as the prevalence of HGT does not invalidate the TOC.

Tables

Table 2-1. Coverage in terms of cellular versus genomic abundance in *E. coli*'s cytosol.

Gene Set	Protein Abundance Coverage	Genomic Abundance Coverage
Universal 34	6.57%	3.08%
Core <i>Enterobacteriaceae</i>	99.96%	79.22%
All <i>Enterobacteriaceae</i>	99.80%	61.34%
Non-Ribosomal Core <i>Enterobacteriaceae</i>	85.14%	78.10%
Non-ribosomal All <i>Enterobacteriaceae</i>	81.70%	59.33%

Table 2-2. Coverage in terms of cellular versus genomic abundance in a *L. interrogans* cell.

Gene Set	Protein Abundance coverage	Protein Abundance Coverage
Universal 36	5.09%	0.96%
Spirochete Core	30.99%	11.28%

Table 2-3. Inconsistency scores (ISs) for the most abundant proteins in each data set. The most abundant proteins have fewer high ISs which implies abundance is a barrier to HGT.

	Proteins with high IS	P-value
100 most abundant proteins with IS in <i>L. interrogans</i>	7	.0067
100 most abundant non-ribosomal proteins with IS in <i>L. interrogans</i>	9	.033
100 most abundant non-ribosomal proteins with IS in <i>E. coli</i>	6	.0013
100 most abundant non-ribosomal proteins with IS in <i>E. coli</i>	9	.033

Chapter 3: Nothing about Protein Structure Classification Makes Sense Except in the Light of Evolution

Ruben Valas, Song Yang and Philip E. Bourne

Current Opinion in Structural Biology, Vol. 19, No. 3. (24 June 2009), pp. 329-334.

Abstract

In this the 200th anniversary of Charles Darwin's birth and the 150th anniversary of the publication of the *Origin of Species* it is fitting to revisit the classification of protein structures from an evolutionary perspective. Existing classifications use homologous sequence relationships, but knowing that structure is much more conserved than sequence creates an iterative loop from which structures can be further classified beyond that of the domain, thereby teasing out distant evolutionary relationships. The desired classification scheme is then one in which a fold is merely semantics and structure can be classified as either ancestral or derived.

Introduction

In 1980 the Protein Data Bank (PDB; [77]) contained less than 100 structures and structural biologists had studied and could name most if not all of them. Today the PDB contains approximately 55,000 macromolecular structures of proteins, DNA, RNA, and

complexes thereof, often combined with a variety of small molecules [78]. No human can assimilate such a breadth of information and so it is only natural, as has happened in so many areas of science with positive consequences, that we attempt an act of reductionism. Thus, the classification of protein structures is an attempt at reductionism from which biological function can be better interpreted. In its purest form reductionism would imply that the application of a simple theory could take a subset of structures, the unique set, and generate all others from it. Clearly this cannot be done completely, Nature is far too tricky, but the notion of generating all structures from a parts list [79] has persisted. Two parts are considered the same if they can be superimposed in 3-dimensions. This raises at least three issues. What constitutes a part; what metric defines two parts as the same, and most importantly, does that sameness convey any biological meaning? Stating the problem a different way, the parts list approach could be considered a bottom-up approach, whereas a consideration of the biological context a top-down approach. The issues then become how well do the two approaches mesh in the middle and what constitutes the biological context?

Already we have introduced a very significant set of issues, yet enormous scientific progress has been made through existing classification schemes. Let us briefly consider some of these schemes in the context of the bottom-up versus top-down approaches. This will serve as an introduction to why we believe the future calls for a more detailed classification which only makes sense in an evolutionary framework.

Bottom-up Approaches to Protein Structure Classification

A large variety of protein structure comparison algorithms have been developed over the past 20 years (see [80] for a review). While they use different methods of protein representation, different algorithms for comparison and different scoring functions, in the majority of cases the end result is a geometric comparison which results in a superposition of the structures according to a root means-square deviation (RMSD), length of alignment, number of gaps, and a score of the statistical significance. As was shown a number of years ago [81] and again more recently [82] there is rarely a unique answer and at a fine level of detail (the devil is often in the details) certainly leads to mis-alignments by failing to capture the biological relevance. Nevertheless, these methods lead to a reductionism which provides a non-redundant structural set as originally exemplified by Dali [83] and the FSSP database [84], with a number of other databases of classified protein structures following [80]. In the majority of cases the comparison is between protein domains and beyond that has little biological context.

Top-down Approaches to Protein Structure Classification

Top-down approaches are exemplified by CATH [85] and SCOP [86], today's gold standards for protein structure classification. While the sheer volume of data to classify requires automation (CATH more than SCOP), human expertise is still used since difficult cases require manual inspection. Much has been written about CATH and SCOP and comparisons have been made between these classification schemes [87] [88] and there is no need to go into further detail here. Both methods involve a consideration of protein domains and incorporate the biological context primarily through detecting homologous sequence relationships. This later point implies that evolution is already a

consideration in structure classification; here we suggest that this needs to be taken further. How extant proteins emerged from smaller building blocks, the role of gene duplication, convergence versus divergence, and co-evolution in a functional context are examples of evolutionary considerations that need to be incorporated into future protein structure classification schemes as we shall see subsequently. In this context we would argue that the end goal of protein classification is to describe the evolutionary pathways between all protein structures.

The Protein Domain as Today's Unit of Structural Classification

Protein domains, as independent folding units, are the modular building blocks of proteins and most current protein structure classification schemes, whether top-down or bottom-up, are based on domains. Protein domain definition from 3D structure is not a fully solved problem [89, 90] which explains some of the differences between existing classification schemes. Since many proteins are multi-domain proteins, and multi-domain proteins are more common in eukaryotes than prokaryotes, we already have a hint for the role evolution can play in an extended protein structure classification scheme. Some domains have high sequence similarity and are evolutionarily related; others are distantly related, sharing obvious structure similarity but not sequence similarity; others have similar topologies, but not to the point where there is clear evidence of common ancestry. Taking SCOP as an example, the first two groups are further classified into the family and superfamily levels, forming a hierarchical scheme. There lies a fundamental problem, a domain can be thought of as both an evolutionary and non-evolutionary unit. Difficulties with current schemes are further compounded by the notion of folds (all or

part of a domain) which are considered discreet components in current top-down classification schemes. Folds are not considered from an evolutionary perspective, but they may be related. Folds do change during evolution to give rise to new folds [91, 92]. Grishin proposed that it is possible for an all-alpha fold to evolve into an all-beta fold by sequential secondary structure flip-over [93]. Similarly, recent work attempted to create two short peptides with high sequence similarity but distinct folds [94]. They achieved this goal with two 50 amino acid peptides with 88% sequence identity, but totally different structure and function. Finally, another case which is difficult for the current classification schemes to embrace are chameleon sequences which can adopt multiple folds [95]. If one accepts the notion of gradual structural variation at the fold level, how can protein structures be classified this way? One notion is the use of smaller fragments [96], but as we shall propose subsequently, this too only makes sense in the light of evolution. In summary, whether or not two proteins are in the same fold is really semantics, whereas describing which is ancestral and which is derived truly captures their relationship. Unfortunately this is a harder problem than simply clustering similar structures. In part it is harder since first you need to identify that protein within extant species and second you need to know the relationship between those species and their ancestors. Ironically, the first problem is addressed well using existing classification schemes.

Domain Distribution in Complete Genomes

The recent accumulation of genomic and structural data as well as improvements in homology detection algorithms has led to the reliable prediction of the protein domain

content of all completed genomes using both SCOP and CATH domain definitions [72, 97]. These protein domain distributions are the starting point for the investigation of protein domain evolution in the genomic era [98-102].

The work of assessing the distribution of domain content across the tree of life began shortly after the completion of the first genomes from each of the three superkingdoms [103]. As the number of structures and the number of genomes accumulated a power law distribution of domains [104] and domain combinations [105] emerged. Several models have been proposed to explain this distribution [106, 107]. To illustrate this point, according to SCOP 1.73 which contains 1087 folds, 692 folds contain only one family (and hence one superfamily). Therefore, the majority of folds correspond to one homologous family that covers a very tiny portion of sequence space. Conversely, the ferredoxin-like fold (SCOP d.58) is found in 55 superfamilies, comprising 123 families. This imbalance is undoubtedly the result of evolution as can be seen by considering the power law relationship with respect to the complexity of the organism. Two independent groups compared domain abundance to features representing complexity, namely genome size [108] and numbers of cell types [109]. Ranea et al. [108] clustered domain families into three categories in terms of their relationship to genome size: unrelated (mainly translation and biosynthesis), linearly-related (mainly metabolism) and non-linearly-related (mainly involved in gene regulation). Vogel et al. [109] compared domain family abundance with cell type numbers in different eukaryote species. About 10% of domain families have a strong correlation with complexity. Half of these superfamilies are involved in extracellular processes and regulation. Such results

infer subtle structure-function relationships of protein domains during evolution leading to the current protein structure repertoire.

An important evolutionary consideration is not just the abundance of domains, but their organization. Over 70% of proteins in eukaryotes and over 50% of proteins in prokaryotes contain more than one domain [98]. These multi-domain proteins are represented by linear combinations of domains; the domain architecture [110]. Domain architectures arise through domain shuffling, domain duplication, and domain insertion and deletion (see [111, 112] for a review) leading to new functions [113]. Baus et al. [114] defined “promiscuous” domains as those that occur in diverse domain architectures. The authors provided a measurement of promiscuity of domains based on the frequency of their coexistence with different domain partners. A systematic comparative genomic analysis in 28 eukaryotes resulted in 215 strongly promiscuous domains. It is not surprising that most are involved in protein-protein interactions, especially in signal transduction pathways. Vogel et al. [115] observed an over-representation of some two-domain or three-domain combinations in complete genomes and termed them “supra-domains.” Those supra-domains (described here as macrodomains) have stable internal domain architectures that are conserved over long evolutionary distances, acting like a single domain in combination with other domain partners. About 1400 macrodomains have been identified with diverse functions, indicating that the preferred association of certain domains is universal and evolutionarily advantageous. These two examples show that domain combinations are determined by functional constraints and evolutionary selection, not just random processes [104]. As such, domain combinations are an important aspect of any protein classification scheme.

A logical extension of these findings is to map domain combinations to presumed phylogenetic relationships derived by other means as exemplified by Snell et al. [116]. Kummerfeld et al. [117] counted the distribution of various types of single domain and multi-domain proteins across the tree of life and predicted that fusion is four times more common than fission in domain combinations. Fong et al. [118] viewed the domain architecture in multi-domain proteins as the rearrangement of existing architectures, acquisition of new domains or deletion of old domains, and proposed a parsimony model to derive the evolutionary pathways by which extant domain architectures may have evolved. Guided by the evolutionary information in phylogenetic trees, Ekman et al. [119] studied the rate of multi-domain architecture formation across different branches of the phylogenetic tree and found that there are elevated rates of domain rearrangement in Metazoa, whereas creation of domains was more frequent in early evolution. Similarly, Itoh et al. [120] observed a large number of group-specific domain combinations in animals and investigated the difference in domain combinations among different phylogenetic groups. Yang et al. [121] aimed to derive the entire evolutionary history of each domain and domain combination throughout the tree of life by mapping current domain content onto the species trees. This approach reveals the origin of each protein domain as well as evolutionary processes such as horizontal gene transfer among more distant species.

Is the Domain the Correct Unit of Classification?

The discussion thus far has focused on the protein domain as the best single level for classifying protein structure, but it is by no means the only one. Just as Ford Doolittle

has argued the shortcomings of tree representations to illustrate the relationship between species [5], calling for a pluralistic approach where no one tree maps all species, we propose a pluralistic approach to protein structure classification incorporating domains, subdomains, macrodomains, and both convergent and divergent evolution.

Subdomain Features

There are currently several available tools for comparing proteins at the subdomain level. Fragnostik is a database that defines relationships in the PDB based on shared fragments between structures [122]. These fragments share both structural and sequence similarity. They can be varying sizes from 5 to 20 residues. Each of these edges is ambiguous (not defined as divergent or convergent evolution) and directionless. However, combining this information with other sources of information could polarize and test some of these edges as a hypothesis for structural evolution.

Another subdomain unit is the closed loop. Most protein structures are composed of loops that come back around on themselves every 25-30 residues [123]. Domain Hierarchy and closed Loops (DHcL) is a web server that decomposes protein structures into domains and closed loops based on van der Waals energies [124]. The protein modules that are the most conserved since the last universal common ancestor (LUCA) correspond to closed loops [125]. Recently all prokaryotic proteins were decomposed into 20 residue fragments (possible closed loops) and clustered based on an identity threshold [126]. The authors found that fragments that corresponded to closed loops were more likely to form large clusters. It is possible to walk between clusters because some have small connections. The authors propose this description is superior to a domain based one

because it represents a finer view of protein function. Closed loops of a common origin in different superfamilies could be evidence for a common ancestor between those superfamilies.

Functional sites are another subdomain feature that could be used for classification. Many distinct superfamilies bind the same ligand. It is possible that these superfamilies share a common ancestor that bound that ligand, but diverged in global structure while the site that binds the ligand is conserved. SMAP [127] finds such binding pockets with both sequence and structural conservation, so these are probably the result of divergent evolution. However, it is also possible that two superfamilies could converge on the same ligand. The PROCOGNATE database defines what superfamilies bind what ligand using structural information from the PDB [128]. A combination of these approaches could create a ligand based classification for domains that encompasses both convergent and divergent evolutionary events.

Macrodomain Features

A protein-protein interaction site is an example of a macro feature conserved from an evolutionary perspective. The interface is conserved while the composite proteins form new superfamilies. A comparison between all protein-protein interfaces in the PDB revealed several examples of highly similar interfaces between different pairs of superfamilies [129]. MAPPIS is a tool for aligning protein-protein binding sites [130]. This level of classification is best done using quaternary structure. 3D complex is a database that classifies protein structures by their quaternary structure [131]. Homomeric complexes evolve in a stepwise fashion from monomers to structures with cyclic

symmetry and then to structures with dihedral symmetry [39]. This information can be used to establish evolutionary relationships between homomers. As an example consider the SCOP family N-acylglucosamine (NAG) epimerase (48222). SCOP 1.73 has two structures in this family; N-acyl-D-glucosamine 2-epimerase(1fp3) and NAG isomerase (2afa). N-acyl-D-glucosamine 2-epimerase is a dimer with cyclic symmetry (C2) and NAG isomerase is a hexamer with dihedral symmetry (D3) according to the 3D complex database [131]. This implies that NAG isomerase must be derived from N-acyl-D-glucosamine 2-epimerase which evolved from one of the many monomeric structures found in this superfamily. It should be noted there may be structural intermediates that have not yet been solved.

Quaternary structure can also define the evolution of some heteromeric complexes. The simplest case is when a heteromer is composed of the same chains as a homomer. The heteromer is almost certainly derived via gene duplication. There are many examples in SCOP where proteins in the same family or superfamily have different quaternary structures. We propose that this information must be incorporated in a classification scheme. A domain based scheme would simply say these proteins share a common ancestor, while a system that includes quaternary structure defines them more explicitly. In summary a domain based classification implies common ancestry, but a macrodomain and subdomain analysis implies an evolutionary hypothesis.

Putting it all Together

We are proposing a pluralistic (some would say fuzzy) approach to protein structure classification that depends much more on evolution than simply defining

homologous relationships between sequences as used in current top-down approaches. Yet these existing schemes form the basis from which pluralism is possible. Pluralism still proposes the domain as a fundamental evolutionary unit, yet encompasses the notion of subdomains and macrodomains.

The scheme needs to be dynamic since many phylogenetic relationships upon which the classification is based will change. For example, there are currently several proposed branching orders for the major taxonomic groups [28, 65]. In the Cavalier-Smith scheme [28], archaea and eukaryotes are both derived superkingdoms, so if there is a link between a protein in bacteria and another found in only archaea, the archaeal protein must be derived. The tree of life infers polarity in the evolution of proteins, but the classification of proteins can also polarize the tree of life. Ideally the two would eventually converge to a solution that captures the history of species as well as proteins. Difficulties arise with our pluralistic scheme since convergence of structure reflects independent evolutionary invention of similar structural folds. Although convergent evolution of structure is rare, it does occur and thus can we really know if promiscuous folds, such as the TIM beta/alpha barrel fold, did not emerge several times independently in evolution? How many cases are there like this?

In our pluralistic scheme any relationship can be defined as divergent, convergent, or ambiguous. What would the map of protein classification/evolution look like when it is complete? It would likely consist of a series of views at different levels of structural granularity where each feature in a given structure could be mapped to equivalent features in other structures and mapped to its presence or absence in extant organisms and by inference common ancestors. The ancestry of modern proteins would reveal the

history of their domains and domain combinations as well as similar and dissimilar micro and macro features. The architecture of the classification scheme would depend on the level it was being explored. Domains would exist as part of a directed acyclic graph if their ancestry was established or as undirected graphs for convergent or ambiguous events.

If such an integrated scheme were in place, and it is a big if, we could contemplate protein evolution in and before LUCA. The superfamily content of the last universal common ancestor (LUCA) has been estimated to contain over 140 different superfamilies [132], although we argue this is an underestimate (in preparation)). It has also been proposed that the oldest fold is the P-loop containing nucleoside triphosphate hydrolase [133]. But how did this fold arise? If we are to root a classification based on evolution we need to explain how to get from that fold to 140 different superfamilies. This is not possible by simply comparing sequences or even structures of whole domains. Protein evolution probably began with structures smaller than what we would consider a domain. It has been proposed that the earliest proteins were created by trans-splicing RNAs that code for protein modules and the origin of genes is much later, independent in archaea and bacteria [134]. Understanding the relationship between the modules that composed LUCA is essential to testing this idea and other hypothesis' about LUCA. This will only be possible by classifying protein structures based on an evolutionary scheme at all levels of protein structure.

The possibility of a pluralistic scheme of protein structure classification is only possible by virtue of the foresight and hard work that has gone into creating our existing bottom-up and top down approaches. Notwithstanding, if improvements in important

areas such as functional annotation and structure prediction are to be made new insights are needed. Further use of what evolution can teach us would seem to be required. In so doing Nature's reductionism will become the reductionism that helps science advance.

Chapter 4: Critical evaluation of structure based phylogenies

Introduction

Since structure is more conserved than sequence several groups have proposed methods for reconstructing species trees using structural characters. Previous work from our group constructed global phylogenies based on the absence or presence of protein superfamilies in completed genomes [135]. A binary string is constructed for each genome that represents its superfamily content. A distance is calculated between each pair of species and that is used to build a neighbor joining tree. Other groups have proposed using dollo parsimony on structural families[136]. Another approach is to use overall abundance of a superfamily instead of its absence or presence [137, 138]. That group uses a more sophisticated parsimony method that I will be discuss below. Each of these methods reconstructs a three domain tree, which is in general agreement with sequence based methods.

These methods are similar to gene content methods in principle[139] but offer several advantages since they rely on structure. HMMs seeded from multiple structure detect more distant homologs than would be found with sequence alone. It also appears that superfamilies are reasonably discreet units, although some argue structure space is becoming increasingly contiguous[140]. It is usually clearer where to draw the line

between superfamilies in terms of structure as opposed to genes using e-values from BLAST hits. Horizontal transfer of superfamilies will have some effect on tree reconstruction, but this effect should be diminished by using structure instead of sequence data. Sometimes the transfers of superfamilies do not matter because the receiving species already had a copy of that superfamily in another gene. This becomes more of a problem when dealing with domain combinations, as it is less likely any particular domain combination is already present in the recipient genome.

In this chapter I present arguments against these methods. Each of them seems a reasonable model since they resolve a three domain tree, but I argue that is not a robust test. If a novel method agrees with the previous one in a very general way, was there any point in making it? If there are differences how can we decide which tree is right? Therefore I developed methods for directly and quantitatively comparing these algorithms, as well as some possible novel ways to incorporate additional data. However, my conclusion on this subject is these methods do not produce quantitatively better trees than sequence.

Problems with the models

The underlying assumption of all these models is that new superfamilies are generated in a somewhat clock-like manner. This is almost certainly not true as the major innovations appear to occur at the roots of the Bacteria and later at the root of the eukaryotes[141]. Changes in environment as well as relaxations in selective pressure

allowed the Eukaryotes to invent many novel superfamilies. Eukaryotes have a higher preference for protein structure than bind zinc [142] and contain more disulfide bonds[143] relative to the prokaryotes. These may make the Eukaryotes appear more divergent from the prokaryotes than they are in reality on trees constructed from structure content data.

It has been observed these methods have a lot of trouble dealing with parasites [135]. An *ad hoc* adjustment was made to account for this. The problem is that the parasites lose proteins so fast it completely violates any clock-like signal. These trees are going to have a tougher time resolving quantum evolutionary events as sequence trees would. Therefore they cannot be better at understanding the major transitions between the superkingdoms.

Caetano-Anolle's method[137, 138] relies on an additional assumption of how superfamilies grow within a genome; the birth death innovation model (BDIM). The BDIM model is a simple model for predicting the distribution of protein (gene) families in a genome[144]. Genes within a family can be duplicated (birth), deleted (death), or drift to the point where they would be considered a new family (innovation). It is possible to grow a genome that follows a power law distribution by specifying rates for each of these factors based on the size of the family. This model is also known as “rich get richer”. Larger families are more likely to duplicate, so they increase in size faster than smaller families. Since this model does not take selection into account there is a direct relationship between the size of a family and its age. The older gene families will have had more time to duplicate. It has been claimed that this model reflects the distribution of families within a genome without invoking selection[144].

I argue that although genomes currently follow a power law distribution in terms of gene families, the BDIM is actually describing something different. There are several clear examples where selection must be included to explain the distribution of genes in extant organisms. Consider the distribution of ribosomal protein S19. This protein is the only member of the fold and superfamily of the same name in SCOP[145], so we can easily to look at its distribution in all complete genomes using the Superfamily Database[72]. This domain is universal, so it absolutely must have been present in LUCA regardless of the order the superkingdoms diverged. It is usually present in only one copy in prokaryotes with the exception of *Bartonella bacilliformis* KC583 and *Leptospira borgpetersenii* ser. *Hardjo-bovis* L550 which each have two copies. It is usually present in two or three copies in Eukaryotes. The *Xenopus laevis* genome is the champion with six copies. The BDIM predicts that any superfamily that is present in more copies must be older than ribosomal protein S19. There are obviously many younger superfamilies that appear have more copies in eukaryotic genomes than this protein. Ribosomal proteins were noted as an example that deviates from the BDIM in [144], but the question is whether this is the exception or the rule. I argue that selective pressure is important in the evolution of nearly all superfamilies.

There are also specific examples of young superfamilies being present in very high copy numbers. The immunoglobulin superfamily is certainly a young superfamily; it is nearly metazoan specific. There is no doubt that the metazoan common ancestor is much younger than LUCA. A very conservative estimate would be a two billion year difference. The immunoglobulin superfamily is present 1221 times in 611 different proteins in *Xenopus laevis*, making it the second most populated superfamily in that

genome. There is no way to explain such a deviation from the model without invoking selection. It is more advantageous to duplicate the immunoglobulins because of their high level plasticity in terms of what they can bind. Ribosomal protein S19's overall fold has limited use because it is a structural protein that lies between other proteins in a complex. Unless another complex needs a piece to fit between proteins of similar folds, there is no advantage to duplicating that protein.

I looked at the distribution of FSFs in the human genome to test whether these two examples are oddball cases or part of a larger trend. The BDIM predicts that the oldest FSFs will be the most abundant. Prokaryotes predate the Eukaryotes by at least one billion years. Therefore any FSF that is Eukaryote-specific should be less abundant in the human genome. As a rough estimation of age I considered any superfamily that is present in less than 25 Prokaryotic genomes (out of 443) and present in the human genome to probably be an innovation that occurred after the birth of the eukaryotes. The BDIM predicts that none of these families should be among the most abundant in the human genome. I ranked each the abundance of each superfamily in the human genome (data not shown). Superfamily has two abundance numbers, one based on each possible transcript (alternative splicing) and the other based on the longest transcript for each gene for the human genome. 47 of the top 100 superfamilies from the all transcripts dataset are post-eukaryote innovations. 49 of the 100 are post-eukaryote innovations when ranked by abundance of the longest transcript. In general older protein superfamilies are not more abundant than younger superfamilies. The reason must be that these superfamilies are under positive selection to fix duplications more rapidly than other families. Even a slight

increase in the rate of duplication will allow a young superfamily to catch up in size to a larger one.

The converse is also true for the oldest superfamilies; some of them are present in very low copy number. Any superfamily that is present in almost all species was probably present in LUCA and is therefore very ancient. Of the 101 superfamilies present in at least 588 species (out of 600), 38 are present in 10 copies or less in all transcripts of the human genome. When only examining the longest transcripts, 53 superfamilies are present in 10 copies or less. These superfamilies must be under purifying selection to keep their copy number much lower than would be expected by the BDIM.

The problem is that the simulations of the BDIM involve only one genome. The real test of such a model is to create hypothetical ancestral genomes and include speciation events. The model works if the distribution in the leaves matches the distribution of extant organisms. The current BDIM would not be able to take proteins that are present in one copy in the ancestor, such as many ribosomal protein families, and keep them at a constant level in the leaves.

Caetano-Anolles has presented a model based on the BDIM for estimating the ages of every protein fold[137, 138]. This model also builds a phylogeny that describes the relationship between all protein folds. The age of each fold is estimated as the node distance to the first fold on that tree. This model has inherited the same shortcomings as the BDIM. Ribosomal protein S19 (fold d.28 in SCOP) has a node distance of about .4465 from the ancestral fold. Later work by this group places the split between the Archaea and Eukaryote-Bacteria clade at a node distance of about .2264[146]. The fact that a universally distributed protein is considered younger than the first split between the

superkingdoms is a clear indication that selection must be taken into account in these models.

It has been observed there is an overall law that describes the relationship between domains that bind metals and total genome size (Dupont in prep.). One of the key corollaries of this finding is that as a genome expands its use of one metal, usually in response to a change in the environment, it decreases the use of another metal. This is another case where the birth rate of one family affects the death rate of another family due to selection. The age of superfamilies that expand late in evolution due to a change in the environment will be overestimated.

Another situation where selection needs to be considered is innovations that are revolutionary improvements. The earliest enzymes evolved under a simplified genetic code, so the repertoire of possible catalytic residues would have been limited. The incorporation of new amino acids in the code would allow for innovations where a new family with a new set of catalytic residues outperforms the old one. It might not be possible to put a new catalytic site on the existing scaffolding, so this may be an unrelated structure. In this case the birth rate for the new family would probably be high while each enzyme in the old family is replaced with a functional homolog. This would also increase the death rate of the old family. There is no way this scenario could be modeled without considering selection and competition between these families.

It is possible that some of the largest enzymatic superfamilies we see today are not the earliest families. Examining the catalytic residues of apparently old families is a good step towards verifying their possible antiquity [133]. Signatures of the primitive genetic code may also have been left on 100% conserved sites within a family [147]. These

details will probably give us a higher resolution picture of early protein evolution than abundance alone.

It has been demonstrated that superfamilies scale to genome size at different rates[108]. Some superfamilies remain constant despite changes in genome size. Others scale with genome size; the relationship can be linear or non-linear. The difference between these trends must be due to selective pressures. The BDIM clearly breaks down when comparing constant size superfamilies to ones that scale with genome size due to selection. However, if we could understand the pressures that separate these classes of expansion it may be possible to use the BDIM to estimate the ages within a class. We could potentially create sets of proteins that would be expected to be under similar selective pressures that would each be well described by a BDIM. If two superfamilies relative rankings within a genome remain the same across the tree of life, it is probably safe to estimate the age of those superfamilies with this model. Since neither one overtakes the other in terms of size as genomes expand and contract, they must be growing at the same rate. Any time the relative ranking changes there must be some pressure causing the smaller superfamily to expand more rapidly. If we could locate such events on the tree of life it would probably indicate an event that caused a significant change in selection pressures.

The BDIM is an important step forward in our understanding of the processes that shape genome evolution. Clearly gene duplication is vital to this process. However the distribution of proteins in extant organisms cannot be explained without invoking selection despite the fact the BDIM can create a power law distribution without it. It is not safe to say that in general older protein families are more abundant than younger

ones. Any results that are based upon this model need to be treated with caution. It may still get some of the general trends correct, but one needs to consider the functional constraints any one family is under. The BDIM is certainly useful as a null hypothesis, if two superfamilies are under the same pressures the older superfamily will be bigger. Anytime we see a deviation from this rule we know there must be a shift in selective pressure. In this sense the BDIM can be used to detect evolutionary events that affect selective pressures.

Recent work has proposed the Chinese restaurant model as a follow up to the BDIM [148]. In this model a proteome is a restaurant and each table is a superfamily. When one enters the restaurant (a new protein) they can pull up a chair at a table that has people at it (duplication) or start a new table (innovation). This method differs from the BDIM by allowing each table to have a unique characteristic (table cloth). The authors gathered empirical data from the Superfamily Database to come up with some superfamily specific parameters. That is to say they implicitly consider selection. They note that superfamilies are invented to do novel functions, so new superfamilies can expand quickly under the right circumstances. Their model is a much better fit of the actual domain data. However, since it throws out the simplifying assumption that older proteins are more abundant it is not clear how this method could be used to build phylogenies or date superfamilies.

Testing the trees

Each of the methods that build a species trees using structure content show some agreement with trees based on sequence information, and has some nodes with high bootstrap values. However, there is currently no way of evaluating these different methods. This is due to the fact that there is no test for phylogeny since we do not know the true tree of life.

I propose that the solution to this problem is to create a synthetic tree of life. Several groups have proposed models for how structures evolve in a genome[106, 144]. These models implicitly create a directed graph that contains the relationship between all superfamilies. Some models have even taken speciation into account where individual species contain subgraphs of the protein universe [149]. This model is quite simple and only allows for birth and innovation of genes, death and horizontal transfer are not accounted for. This method creates a phylogeny where one would know the superfamily content of each species, as well as the phylogeny of the protein universe.

One could create a more complex model that produces complicated phylogenies. This model would need to incorporate issues that make phylogenies difficult to resolve such as horizontal transfer and severe genome reduction in parasites. The method would need to be run with a variety of parameters to see what methods perform the best under specific assumptions.

While objectively evaluating existing methods would be useful on its own, the real point of this exercise is to develop novel methods for reconstructing phylogeny. The hope would be that this test set could be used to develop a method that is robust under a variety of conditions and then use that method on the actual data. Since the directed graph of the relationship between superfamilies is known in this ideal case we can compare graph

based methods (both directed and undirected) to current methods. This would allow us determine whether using additional information such as the relationship between domain combinations could be useful for building phylogenies.

The basic idea is that we create a tree of species and proteins using our model. One then reconstructs that tree using the proteomic data that is available in each species. One can compare the distance of these trees using treeDist (packaged as part of Phylip[150]). For any model we can see which method produces closer trees on average to the true answer. One will need to vary several parameters to the model, but hopefully one method will be superior under a variety of circumstances. One will also need to evaluate which parameters make the model the most realistic.

New Algorithms for Building Fake Trees

The naïve model assumes that protein families only grow[106]. At each time step a protein is randomly chosen and duplicated. If a random value is above a threshold that protein is considered to have diverged enough to start a new superfamily, otherwise it increases the population of the superfamily it was duplicated from. They analyze this process on the domain level, but I'm lumping connected domains together into superfamilies.

The next addition to this model is to take loss into account. I give each genome an equilibrium size (in this case all genomes should be the same size). At each time step we randomly decide whether there will be a gain or loss with so that at the equilibrium point

gain and loss have equal probabilities. This represents a cost ceiling on genome size which is supported by the thin variation in genome sizes of free living prokaryotes[73].

$$P_{gain} = 1 - \frac{P}{2n_{ma}}$$

Equation 1. Probability of adding a new domain at each time step

This model can be made more realistic by changing the equilibrium point between speciation events. Some species should have their protein ceilings halved to represent parasites.

$$P_{gain} = 1 - \frac{P}{2n_{ma}}$$

Equation 2. Each species has a distinct genome size

As more and more superfamilies are seeded it should become harder and harder to innovate new folds. Less new folds means there is less information to build phylogenies with.

$$P_{innovation} = \frac{Wsf}{S_{st}}$$

Equation 3. Probability of inventing a new fold superfamily. W is a constant representing the maximum likelihood of innovation early in evolution.

After a certain point superfamily innovation stops, which I call the saturation point (measured in generations). If there is no room for innovation early in the evolution

of the tree it becomes very difficult to determine the relationship between the nodes using absence or presence alone, as too little is shared between all species. On the other hand if we cannot resolve any speciations after the saturation points because they will have too few differences.

In reality some types of speciation events will have a greater affect on superfamily content than others. Some folds could be lost due to hyperthermophily or gained due to the presence of oxygen or other metals. This was not implemented because it would require giving many specific parameters to each superfamily.

The model also needs to include horizontal transfer as this is the effect that really makes building phylogenies difficult using sequence data. There can be a parameter h that is the probability of a transfer occurring between a pair of species and all genes are equal candidate for transfer. One can also include a variable for how long it's been since the genomes diverged (similar organism should have an easier time accepting new genes).

This model will become really complicated if we include selection. None of the models at the time I did this work took selection into account. We would need to give every superfamily a fitness value and then each superfamily's growth rate could be varied. This should actually be two parameters, one that describes how fast a superfamily will grow (plasticity) and one that describes how hard it lose to a member of that superfamily (essentiality). Essentiality could also be described on a protein by protein basis instead of a superfamily. This may not be necessary and it will greatly increase the number of parameters we need to include.

Potential new methods (that didn't work)

Directed 1

These directed algorithms use knowledge of the structure of the protein universe to build the tree of species. Every pair of species subgraphs of the protein universe are compared to each other (Figure 4-1). The distance is derived from nodes that disagree; superfamilies that the species don't have in common. For each node that disagrees I calculate to a path to nearest node found in both species. If the path has a node that is not in both species on then we do not count that as a difference, because it is probably a transfer.

The motivation behind this is to get levels of relationship beyond that of a superfamily. When two species have proteins that have high structural similarity (same family) it should taken as evidence as more recent common ancestry that simply sharing the same SCOP fold. This implementation is currently unrealistic, but I thought it would be interesting to see how much a perfect classification system could help us.

Directed 2

This is the same as Directed 1 but with an extra step that tries to account for loss. When a node is present in one genome but not the other I check to see if that node's

children are present in both genomes. If the derived structure is present but not the ancestor I take that as evidence the ancestral structure was lost.

Hybrid of Song and Gustavo

It seems to me that some superfamilies behave better under Song's model than the abundance model and visa versa. Song considers superfamily present in one copy the same as if it was present in 100 copies. This means the method gets no information out of superfamilies that are always present in high abundance such as the p-loop hydrolase. Gustavo's method, on the other hand, encodes the intermediate abundances so some information could be gained from even universal superfamilies. However the superfamilies that are usually only present in 1 copy (ribosomal proteins) will have values of either 0 or max. Horizontal transfer or loss would look like big changes for these superfamilies. I propose using Song's method on small superfamilies (based on their maximum abundance in any genome) and Gustavo's method on larger superfamilies. I used 10, 20, 30 copies per genome as my abundance cutoffs to divide the line between these two methods.

Derived Domain Combinations

Since we do not know the true tree of proteins I attempted to use domain combination data instead. One can hypothesize what domain combinations are derived

assuming single superfamilies are ancestral and fusion is more likely than fission. PAUP allows one to provide a tree that describes how characters are related to each other for parsimony analysis[151]. That allows one to be explicit about how characters are derived from each other. This could allow one to get more signal out of the data. Consider the case of a combination of domains A and B (Figure 4-2). There are eight possible states for that combination: 1) neither A nor B is present in the genome 2) A is present, but not B 3) B is present but not A 4) A and B are both present but not AB 5) AB is present but not A 6) AB is present but not B 7) AB, A, and B are all present and 8) AB are present but not A and B.

Clearly state 7 must be derived as it must be the result of a fusion. State 4 is also derived since it involves the addition of additional superfamily. The motivation of this approach is that the ancestor of AB had A and B, so species that have either of superfamilies should be more closely related to the ones that have the derived combination.

There are some big problems with this approach. First of all HGT of AB can make a genome seem derived. Did a genome receive AB or lose A and B? The second issue is that loss can mask a derived state as ancestral. Either of these cases can be dealt with one at a time, but that prevents one from using this data in an unsupervised manner. Then there's the problem that combinations are difficult to assign and many contains gaps (unassigned region) that make their true domain structure ambiguous. This method failed to produce even some of the simplest results on test cases. This is certainly a situation where simpler is better, as additionally complexity in the algorithm is misled by complexity in the biology.

Results

I've compared Song's metric to dollo parsimony under a couple of models. This is data is from a model with 128 species that includes genes loss and parasites (severe reduction events). The parameter for the maximum number of superfamilies was varied and all others held constant. It's intuitive that these method will do better with more superfamilies as there is more phylogenetic signal and this is clearly the trend (lower is better). Dollo does a little better than Song because it distinguishes between loss and gain (and doesn't count loses in its distances), but with lots of superfamilies these methods give about the same result (Figure 4-3). If there is no loss then these methods should give identical distances, so loss helps Dollo (or hurts Song's metric).

Horizontal transfer should have the opposite effect. Dollo parsimony reconstructs ancestral states by assuming each trait was only invented once in the tree. A horizontal transfer across the tree results in that protein being considered present in LUCA with losses in multiple lineages. With additional HGT Dollo clearly begins to get misled (Figure 4-4). It is very difficult to model HGT better without invoking some sort of selection because some superfamilies are more easily transferred than others. On more complicated tests there was no clear winner between the different methods (Figure 4-5).

Discussion: Why this project failed

This model for generating protein universes and species trees has many free parameters. Every free parameter creates an order of magnitude of more testing that needs to be done. Coming up with a method for testing the fit of our model to real data would be extremely useful for saving time and making our results seem more realistic, but this would involve having the correct answer for the tree of life or proteins. It is certainly possible to improve upon the methods that are out there now, but I doubt they will be able to resolve the tree of life much better than Song's trees. Without a detailed understanding of the innovation of new folds we cannot truly judge how clock-like the data is.

Why do these methods work at all on the real data if their assumptions are so far off? The reason is that they only really show that there are three superkingdoms, and I argue this is like shooting fish in a barrel. I believe this because the Archaea underwent a massive reduction at their root. The Eukaryotes clearly had a massive expansion at their root. Regardless of the tree of life, there are major events that differentiate the superfamilies at the root of each of the superkingdoms. This is seen in the venn diagram's in [135]. Instead on analyzing these methods on how well they do on the easy case we should test those using independent lines of evidence, such as polarized transitions. However, now that the number of sequence genomes has topped 1000 it seems pointless to try to reconstruct trees that have at most 1600 useful characters (the size of the superfamily universe).

Figures

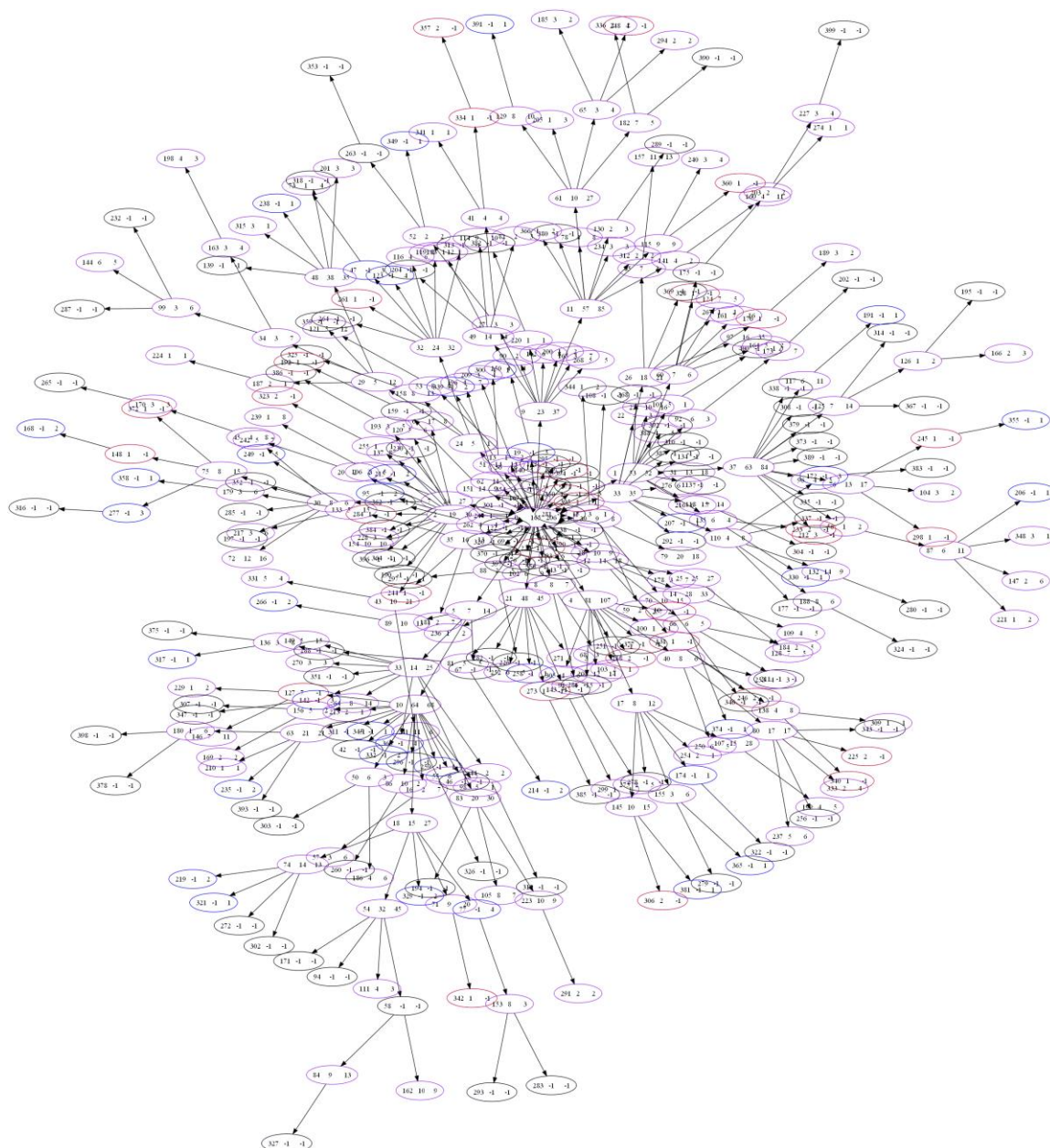


Figure 4-1. A comparison of two species in the protein universe.

Each node corresponds to a protein superfamily. Red nodes are in species A, blue in species B, purple in both, and black in neither.

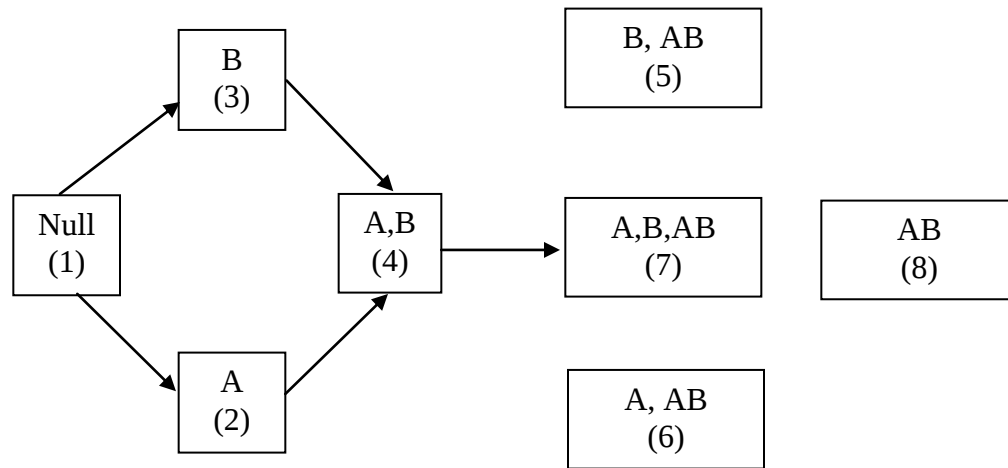


Figure 4-2. Scheme for encoding derived domain combinations.

This is opposed to standard methods of just using absence or presence data.

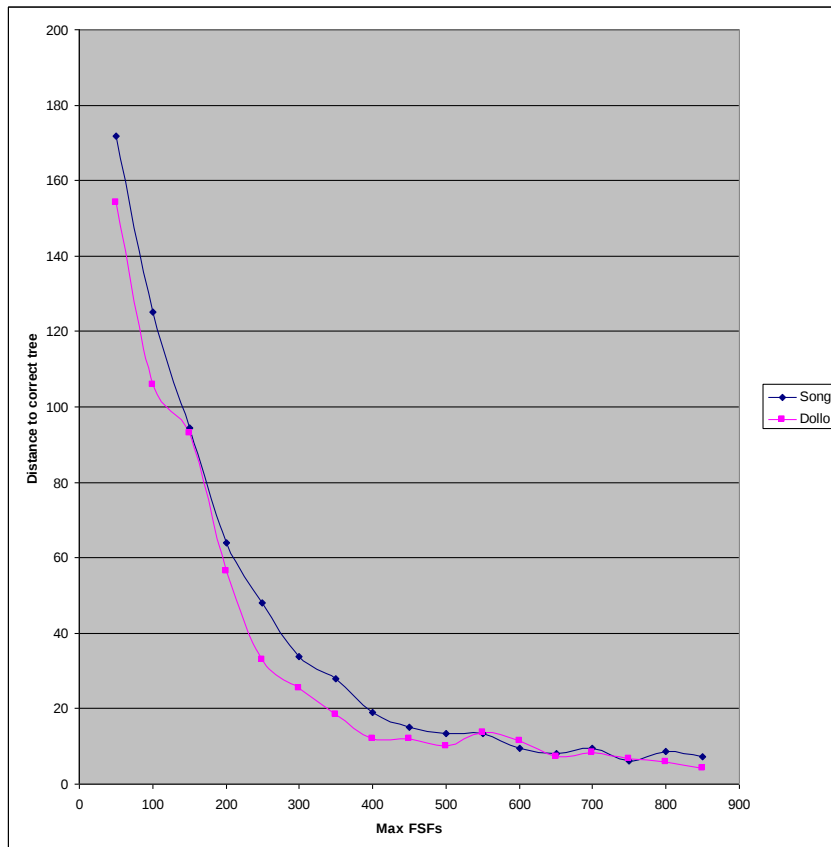


Figure 4-3. Song's method vs Dollo parsimony.

A synthetic tree of life of 128 species was used. Increasing the number of superfamilies in the universe helps both methods.

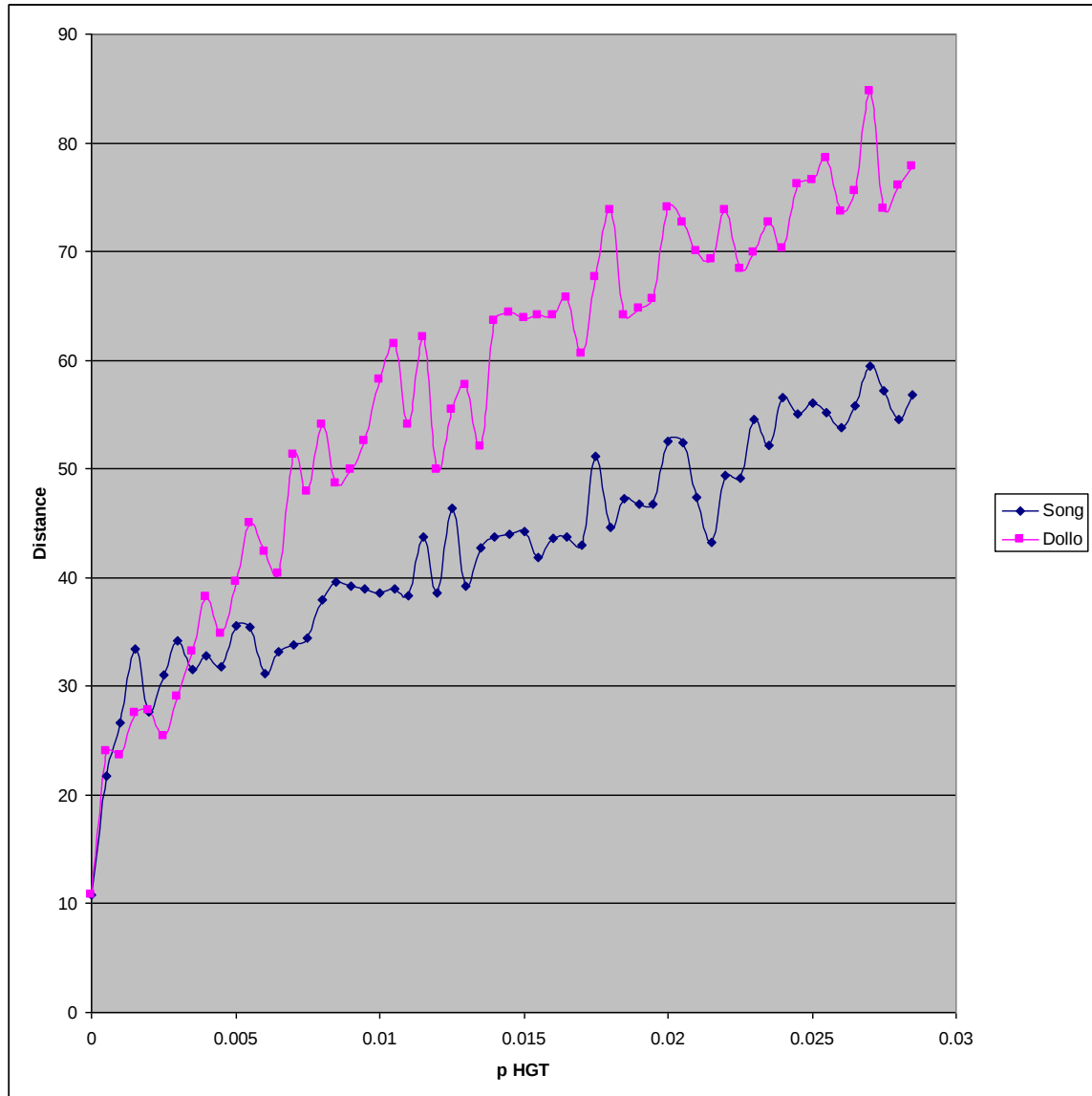


Figure 4-4. Song's method vs Dollo parsimony in the presence of HGT.

This graph was done with 64 species (Dollo is really slow compared to Song's method) with 10 replicates at each point.

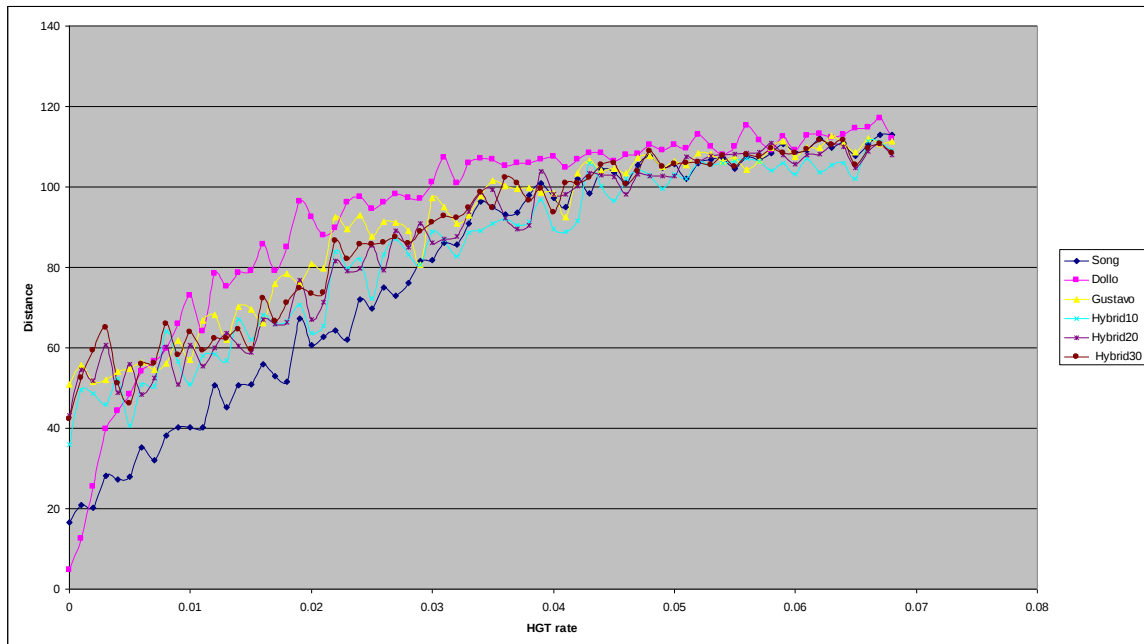


Figure 4-5. Comparison of effects of HGT on 6 different methods.

This universe had 600 superfamilies and a .01 chance of a species becoming a parasite. The hybrid 10 was slightly more tolerant to horizontal transfer than Song's method.

Chapter 5: Rethinking proteasome evolution: Two novel bacterial proteasomes

Ruben E. Valas and Philip E. Bourne

Journal of Molecular Evolution. 2008 May;66(5):494-504. Epub 2008 Apr 4.

Abbreviations: BPH, β -proteobacteria proteasome homolog; hsl, heat-shock locus; SCOP, Structural Classification of Proteins,; PDB, Protein Data Bank.

Abstract

The proteasome is a multi subunit structure that degrades proteins. Protein degradation is an essential component of regulation because proteins can become misfolded, damaged, or unnecessary. Proteasomes and their homologs vary greatly in complexity; from HslV (heat shock locus v) which is encoded by 1 gene in bacteria to the eukaryotic 20s proteasome which is encoded by over 20 genes. Despite this variation in complexity all the proteasomes are composed of homologous subunits. We searched 238 complete bacterial genomes for structures related to the proteasome, and found evidence of two novel groups of bacterial proteasomes. The first, which we name Anbu, is sparsely distributed among cyanobacteria and proteobacteria. We hypothesize that Anbu must be very ancient because of its distribution within the cyanobacteria, and that it has been lost in many more recent species. We also present evidence for a fourth type of bacterial proteasome found in a few β -proteobacteria, which we name β -proteobacteria proteasome

homolog (BPH). Sequence and structural analysis show that Anbu and BPH are both distinct from known bacterial proteasomes, but have homologous structures. Anbu is encoded by 1 gene, so we postulate a duplication of Anbu created the 20s proteasome. Anbu's function appears to be related to transglutaminase activity, not the general stress response associated with HslV. We have found different combinations of Anbu, BPH, and HslV within these bacterial genomes which raises questions about specialized protein degradation systems.

Introduction

The proteasome is a complex, multi subunit protein assembly which forms a barrel with multiple internal active sites that function together to recognize and degrade proteins (reviewed in [152]). All archaea and eukaryotes have a 20s proteasome as well as some actinobacteria, but most bacteria have a simpler homologous structure heat-shock locus v (HslV). Although proteasomes are found across the tree of life there are many bacterial species that lack them entirely. The 20s proteasome and its HslV homologs either function to degrade misfolded proteins [153], as occurs under conditions of heat shock, or as a precise regulatory mechanism by degrading proteins, usually defined by a ubiquitin tag [154].

There is an evolutionary progression in structural complexity of the proteasome. Several Protein Data Bank (PDB) structures [78] of the 20s proteasome from all 3 superkingdoms and HslV from bacteria all form barrels structures that have the active

sites on the inside of the barrel. All of these barrels consist of inner subunits responsible for cleavage and outer subunits responsible for protein recognition. The core of the 20s proteasome is a 4 layered barrel found in archaea, eukaryotes, and actinobacteria. Each layer comprises a heptameric ring. All 28 subunits are in the same structural family in the Structural Classification of Proteins (SCOP) database [145] implying they all share a single common ancestor gene.

In the actinobacteria and archaea the 20s proteasome is usually encoded by 2 genes, the α and β subunits. The β subunits are catalytically active and form the 2 middle layers. The α subunits that form the outer 2 layers are catalytically inactive and act as scaffolding for the β subunits. The α subunits form an antechamber that restricts access of the substrate to the proteolytic chamber [155]. Some actinobacteria and archaea contain more than one type of α and β subunit. The eukaryotes are more complicated still. The core of the yeast 20s proteasome is coded by 14 different genes (7 α and 7 β) , with only 3 of the β subunit genes having catalytic activity [156]. The 20s proteasome can be knocked out from archaeal cells under normal conditions, but it is essential for surviving heat shock conditions[157].

Many bacteria contain a simpler proteasome homolog, HslV [158]. HslV is a heat shock protein and is expressed as part of a general response to stress that causes proteins to misfold. Unlike the heptameric rings of the 20s proteasome, HslV is made up of 2 layers of hexameric rings which are encoded by only 1 gene. According to SCOP, the subunits of HslV are also in the same structural family as the subunits of the 20s proteasome. Because of its simpler structure HslV is a good model system for the 20s proteasome.

The 20s proteasome and HslV both associate with ATPases that use ATP to unfold proteins and translocate them into the proteasome or HslV structure, respectively (reviewed in [159]). There is a corresponding increase in complexity in the ATPases and other regulatory proteins associated with the proteasome in moving from the bacteria to the eukaryotes. A hexameric ring of ATPases known as heat-shock locus U (HslU) binds to each side of HslV. Likewise, hexameric ATPases bind to either side of the 20s proteasome interfacing with the α subunits. Unlike the subunits of HslV and the 20s proteasome, the subunits HslU and the ATPases associated with the 20s proteasome are in different structural superfamilies [160]. HslU is related to ClpX, the ATPase of the ClpP protease. The 20s proteasome ATPases are related to the ATPase domain of the protease FtsH. In HslU and the 20s proteasome of actinobacteria and archaea all 6 ATPases are encoded by a single gene [161, 162]. This structure is more complicated in eukaryotes, which encodes 6 different homologous ATPases and at least 11 other proteins in the PA700 complex (also known as the 19s regulatory cap) [163]. Eukaryotes also have 2 alternative caps, PA28 (also known as the 11s cap) [164] and PA200 [165], that do not use ATP or recognize ubiquitin. Both the 20s proteasome core, as well as different combinations of the 20s core and its caps are found within eukaryotic cells in significant numbers [166].

There is also a progression in complexity of the targeting systems (i.e., recognizing which proteins to degrade) in the various proteasomes. In eukaryotes most proteins are targeted for degradation by ubiquitin tagging [154, 167], although there is a growing number of proteins found to be degraded in a ubiquitin-independent manner [168]. No such tagging is known in prokaryotes, although it has recently been shown

many bacteria have some homologs to this tagging pathway [169]. Some proteins contain a tag in their N-terminus, such as the ARC protein, which targets them for degradation by HslV [170]. We speculate that similar targeting may be used with other proteasomes in species that lack ubiquitin.

Previous work has analyzed genomic data to study the evolution of the proteasome [171]. The authors looked at 61 complete and 60 incomplete genomes. They found that several protists contain both HslV and a 20s proteasome. Some actinobacteria also contained a 20s proteasome with distinct α and β subunits. The authors state this was probably due to horizontal transfer, but they could not identify the source. They found that several bacteria had no homolog to either proteasome or HslV, so they must use other proteases instead. They also noted that 2 bacteria, *Magnetospirillum magnetotacticum* and *Enterococcus faecium*, had 2 copies of HslV. They conclude that *Magnetospirillum magnetotacticum* had a recent duplication of HslV and *Enterococcus faecium* acquired a second copy through horizontal transfer. They found no bacteria that have both HslV and a 20s proteasome.

We are able to extend these results by analyzing many more complete genomes and using data derived from protein structure. Since structure is more conserved than sequence, this facilitates studies over long evolutionary time scales. Support for this approach comes from our recent work in constructing the tree of life based on the presence or absence of superfamilies derived from structure [135]. The structure data for our work comes from the Superfamily database which holds Hidden Markov Model (HMM) predictions of structural domains, families and superfamilies based on the SCOP classification scheme [145], for completed proteomes [172]. Superfamily was used to

determine which of the 238 completed bacterial genomes had multiple genes predicted to be proteasomal subunits. Most bacteria had 1 gene which hit the proteasome subunit family, which was usually HslV. The actinobacteria with known 20s proteasomes had 2 hits as one would expect. To our surprise there was also a large number of proteobacteria with 2 hits and several β -proteobacteria had 3 genes encoding proteasome subunits. There were also some genomes with no hits to this family as had been observed before.

Magnetospirillum magnetotacticum was one species that had 2 hits. We are able to analyze when these genes were duplicated by looking at how the additional hits cluster in a phylogenetic tree. The 2 proteins from *Magnetospirillum magnetotacticum* clustered on opposite ends on the tree, each with other sequences. This led us to the conclusion that this is not just a second copy of HslV, but rather a representative novel proteasome homolog which we named Anbu. We also found a distinct cluster of sequences in some β -proteobacteria, which we name β -proteobacteria proteasome homolog (BPH). No species containing BPH from this group was mentioned in [171], so this group is entirely novel. We found that our 2 novel clusters match 2 unannotated clusters in the NCBI Protein Clusters database: CLS882959 is Anbu; and CLS856934 is BPH.

Our trees show distinct clusters, but do not show an unambiguous history of the proteasome. Since these sequences diverged billions of years ago it is not surprising that it is difficult to get a clear phylogenetic signal. However, structural inference can be linked to sequence, so we can combine structural information with these trees to better recreate the evolutionary relationship of these families. Threaded structure predictions were created for 2 representative sequences from both Anbu and BPH. Anbu sequences were taken from *Rhodopseudomonas palustris* and *Hahella chejuensis* and BPH

sequences were taken from *Thiobacillus denitrificans* and *Ralstonia Metallidurans*. These are high quality predictions because each prediction was created from several known structures of HslV and the 20s proteasome. We compared these predictions to other known structures to determine the evolutionary history of the different proteasome homologs.

HslV and the 20s proteasome are clearly evolutionarily related from their common structures. HslV is a good model system for the 20s proteasome from that fact alone. However, the question of which proteasome came first has interesting implications for evolution. If HslV is ancestral to the 20s proteasome then the archaea must be younger than the bacteria, as all archaea have a 20s proteasome[28]. Since there were no other known simple proteasome homologs as potential predecessors this seemed reasonable. The introduction of Anbu changes this view. We show that Anbu is a more probable candidate as the ancestor of the actinobacterial 20s proteasome than HslV based on its position in the phylogenetic tree and its structural features. Further study of Anbu will shed more light on the function of the 20s proteasome rather than studying just HslV.

Results

Phylogenetic Analysis

We constructed a maximum likelihood tree from a multiple sequence alignment of sequences predicted to be in the proteasome subunit family by the Superfamily

database. This tree shows 5 distinct clusters. 3 of these clusters are known proteasome subunits; HslV, the 20s proteasome α subunit and the 20s proteasome β subunit. There are some low bootstrap values, but most of the critical edges have high values. The 2 novel clusters, Anbu and BPH, are both supported as true novel groups with bootstrap values of 100%. A 100% bootstrap value also separates this tree into 2 groups; BPH with HslV, and Anbu with the 20s proteasome. This tree strongly supports Anbu being ancestral to the 20s proteasome, not HslV which is the current view.

Thr1, Lys33, and Gly47 are all catalytic residues in *Thermoplasma Acidophilum*'s 20s proteasome [173-175]. A deprotonated Thr1 performs a nucleophilic attack on the substrate which is stabilized by Gly 47. Lys33 promotes the deprotonation of Thr1. The corresponding sites are universally conserved throughout Anbu and BPH with only 1 exception (Supplemental tables 1 and 2). This is evidence that these novel groups function like the known bacterial proteasomes. The distribution of Anbu and BPH on the tree of life has several interesting features (Supplemental Figure 1). Anbu is found in α -proteobacteria, β -proteobacteria, γ -proteobacteria, as well as cyanobacteria according to the Superfamily database. This is noteworthy since no cyanobacteria has HslV. Anbu is present in *Gloeobacter violaceus*, which is an early branching cyanobacteria [176]. It appears Anbu was present in the cyanobacterial ancestor so it must be very ancient. A BLAST search revealed that Anbu was also present in *Leptospirillum ferrooxidans* as well as *Solibacter usitatus*. *Cytophaga hutchinsonii*, a sphingobacteria, was found to have 2 copies of Anbu. A species with a duplication of Anbu could be the precursor to the 20s proteasome. Anbu's distribution is sparse but broad, which infers it is an ancient protein that has been lost many times. This repeated loss is not unrealistic given that

photosynthesis was also lost many times in the proteobacteria [177]. The BPH group only includes β -proteobacteria. This extremely narrow distribution implies that BPH is a relatively young proteasome. This, combined with BPH's position in the phylogenetic tree, implies that BPH evolved from HslV.

The current view is that bacteria either have HslV, a 20s proteasome, or no proteasome. There are no known cases of a bacteria having both HslV and a 20s proteasome. With the discovery of Anbu and BPH it is now clear that proteasome homologs occur in many combinations in bacterial genomes (Supplemental Figure 1, Table 5-1). Anbu, HslV, and BPH are present together in several genomes in different combinations, but none of them are ever found in the same genome as a 20s proteasome. However, both HslV and the 20s proteasome were found in a recent metagenomic study of *Leptospirillum* group II bacteria. The authors state that in this case 20s proteasome was probably horizontally transferred from the actinobacteria[178]. A BLAST search revealed this metagenome also contains Anbu. Although this metagenomic sample is dominated by *Leptospirillum* group II bacteria[179], this data is not from a single species. Therefore this is not evidence that a single genome contains HslV, Anbu, and the 20s proteasome. However, it is evidence that all 3 of these proteasomes can be useful in the same environment. The 3 *Ralstonia* species in our sample have Anbu, BPH, and HslV. We believe that these 3 proteasomes are functionally distinct (discussed below). This raises an important question of how bacteria target a protein to a specific proteasome to be degraded without using ubiquitin. BPH is never found as the sole proteasome homolog. It can be inferred that BPH degrades proteins that cannot be degraded by one of the other mechanisms, but it does not degrade a wide enough variety of substrates on its own to

replace HslV or Anbu. It would be interesting to create knockouts in these species to see how BPH functions and hence to compare the functions of BPH and HslV in these species. This would allow us to determine whether BPH's function is redundant or whether it degrades additional substrates. The 20s proteasome may be able to act on a wider variety of substrates than other homologs, so it can replace the function of different proteasome families. The idea that bacteria can only have HslV or the 20s proteasome exclusively is too simple. Instead we need to determine the specific functions of each family and how they interact in all of these different combinations.

These new proteasome families are good candidates for structure prediction using fold recognition (threading), because the PDB has several structures for the 20s proteasome from archaea, bacteria, and eukaryotes as well as HslV. We created 2 models from sequences of both Anbu and BPH using the Phyre web server, which is the successor to 3D-PSSM [180]. Anbu was modeled from archaeal and eukaryotic 20s proteasome structures. Anbu from *Rhodopseudomonas palustris* has 18% sequence identity with the structure of the archaeal 20s proteasome. BPH was modeled from structures of the eukaryotic proteasome and HslV. BPH from *Thiobacillus denitrificans* has 22% sequence identity with the structure of HslV. We built a multiple sequence alignment from a multiple structural alignment for each cluster using Combinatorial Extension [181], and built a tree using maximum likelihood (Figure 5-2). This was done to increase the quality of the alignment using structural information. Anbu again falls right between the α and β subunits and BPH clusters with HslV. This tree is in agreement with the one constructed from sequence alone. It supports Anbu being ancestral to the 20s proteasome, and HslV being ancestral to BPH.

Structural Analysis

The predicted structures of both Anbu and BPH align very well with known proteasome subunits, but each has unique structural features. The areas around the active sites align particularly well (Figure 5-3). This conserved catalytic triad is strong evidence that Anbu and BPH both function as proteasomes.

After a crystal structure of HslV from *E. coli* was determined it was compared to the β subunit from the archaea *Thermoplasma Acidophilum* [182]. The authors proposed several differences that could account for HslV forming a hexamer while the 20s proteasomes forms a heptamer. The first is that the β subunits may be forced by the α subunits to form a heptamer. Helix 1, which is in contact between the α and β subunits, is extended by 5 residues in the β subunit relative to HslV (highlighted in red in Figure 5-4A). The β subunit also has an extra C-terminal helix (highlighted in green in Figure 5-4A) which could affect the way the subunits pack together into rings. We compared our models of Anbu with known structures of HslV and the 20s proteasome. Helix 1 is extended in Anbu compared to HslV (highlighted in red in Figure Figure 5-4E). Anbu's C-terminal tail is also extended relative to HslV (highlighted in green in Figure 5-4E). The threaded models of Anbu cuts out about 30 C-terminal residues that do not hit known structures. The secondary structure of this region is predicted to be a sheet followed by a helix with possible loops between them. There are several highly conserved positions in the missing section of the tail. It is possible this region has a functional role that is not present in HslV or the 20s proteasome. Anbu has other features that are not shared by

any of the other proteasome families. Both turns 3 and 6 have significant extensions in Anbu that could affect packing in the biological unit (highlighted in yellow in Figure 5-4B and Figure 5-4C. These turns are colored orange in the biological unit of the 20s proteasome in Supplemental Figure 2). The extended loop 3 could act as a gate into the proteasome if Anbu forms 2 layers of rings. We cannot definitively conclude Anbu's biological unit from these features, but they do give a strong indication that the 20s proteasome evolved from Anbu. Both the helix extension and C-terminal tail discussed above are present in both the α and β subunits of the 20s proteasome (Figure 5-4B and Figure 5-4D). Both structural features were probably present in the ancestor of both subunits. A duplication of Anbu would be more likely to result in a 20s proteasome like structure than a duplication of HslV because Anbu already has both of these structural features. That taken with Anbu's position in our tree's indicate that the 20s proteasome evolved from Anbu, not HslV.

We also compared the predicted structure of BPH to HslV and Anbu. It is highly unlikely that the 20s proteasome evolved from BPH or visa versa based on their distributions in the bacteria. BPH has an extended loop 2 relative to both HslV and Anbu (highlighted in green in Figure 5-5) BPH's helix 1 is also extended relative to HslV, but it does not have a C-terminal extension. Structurally BPH shares similarities with both Anbu and HslV, but it is probably not an intermediate structure because of its narrow distribution within the β -proteobacteria.

The 20s proteasome and HslV both degrade proteins in an ATP-dependent manner. The ATPase binding surface in these complexes are very different because the 20s proteasome is 4 layers and HslV is 2 layers (Figure 4 in [28]). This means the

ATPases are binding to opposite faces of the proteasome subunit in 2 and 4 layered proteasomes. We could postulate as to whether Anbu forms a 2 or 4 layered biological unit by determining whether its ATPase is more like HslU or the ATPases associated with the 20s proteasome. A BLAST [183] search was run against cyanobacteria to find potential ATPases for Anbu. The distribution of ATPase homologs in cyanobacteria is informative since they do not have HslV and only some have Anbu. We were unable to locate a known proteasomal ATPase that matched the distribution of Anbu or BPH. This could mean that an ATPase is moonlighting or that Anbu or BPH is acting in an ATP-independent manner and only breaking peptides down. It's possible that one of the genes of unknown function associated with Anbu (discussed below) could be its ATPase.

It has been argued that HslV could not evolve from the 20s proteasome because the decrease in pore size from a heptamer to a hexamer would not be favorable [28]. Also, the loss of the inactive α subunits would be a major transition that would result in a proteasome with a large pore and no regulatory ATPase, which would not be favorable. By this same logic it is highly unlikely that Anbu or BPH could evolve from the 20s proteasome as they appear to only have active subunits.

Our structural predictions infer that Anbu is the ancestor of the 20s proteasome. Larger structural features such as whether Anbu's rings are heptameric or hexameric will make for stronger evolutionary arguments. It will be necessary to get a crystal or cryo-electron microscopy structure to understand the biological units of Anbu and BPH. If we are correct that Anbu is the ancestor of the 20s proteasome, a structure of the complex would provide an excellent opportunity for an improved understanding of the 20s proteasome.

Function of Anbu and BPH

Anbu is found in a very diverse set of bacteria, including both oxygenic and anoxygenic phototrophs. It is also present in many species that have unique phenotypes such as: *Ralstonia metallidurans* which can withstand high metal concentrations and plays a role in the formation of gold [184], *Rhodoferrax ferrireducens* which can reduce Fe(III) [185], and *Burkholderia xenovorans* which is capable of degrading polychlorinated biphenyl [186]. HslV expression is increased under heat shock and other stresses that cause proteins to misfold, so we searched the literature on microarray experiments to see if any of the stresses these bacteria face in these varied environments induced expression of Anbu. Anbu was not induced in *Synechocystis PCC 6803 sp.* in response to heat shock [187], UV-B light [188], salt stress, and hyperosmotic stress [189]. Anbu was also not differentially expressed under oxidative stress conditions (addition of H₂O₂) in *Synechocystis*[190] and *Rhodobacter sphaeroides*[191]. *Pseudomonas putida* KT2440 did not induce Anbu expression in the presence of any of several different aromatic compounds, although some triggered increased expression of HslV [192]. Although these experiments do not reveal Anbu's function they show that Anbu is not differentially expressed in several situations that HslV would be. This is functional evidence that Anbu is distinct from HslV. Future microarray experiments in these species may reveal when Anbu is induced. Unfortunately we could not find any microarray experiments with these kind of stresses for the few species that have BPH.

We compared the operons of HslV, Anbu, and BPH using the MicrobesOnline

[193] operon browser (Supplemental Figures 3-6). HslV almost always falls in the same predicted operon as HslU, and they are always predicted to be in the same regulon. We noticed Anbu is often expressed in an operon with genes labeled as COG2307, COG2308, and COG1305 (Figure 5-6a). When Anbu is not in the same operon as these 3 genes, they are almost always predicted to be in the same regulon. COG2307 and COG2308 are uncharacterized conserved proteins. COG2308 is predicted by Superfamily to have a glutathione synthetase ATP-binding domain. The hits to this superfamily are near the threshold of what is considered a significant hit in Superfamily. Understanding how COG2308 uses ATP may be key to understanding Anbu's function. It's possible this uncharacterized proteins interacts with Anbu, but it would have to have some other function as well since it appears in genomes that lack Anbu. COG1305 is a transglutaminase-like protein. Some bacterial transglutaminases act as proteases [194] while others selectively cross link proteins [195]. Either function could have interesting interactions with a proteasome. If this transglutaminase acts as a protease it could break down the peptides that come out of Anbu into even smaller pieces. If it acts at as a cross linker Anbu may degrade it to regulate the levels of cross linking in the cell. Either of these functions could also act to regulate Anbu. We compared the average number of predicted transglutaminases catalytic domains using Superfamily in genomes that have Anbu, the 20s proteasome, or neither (Table 5-2). Genomes that had either Anbu or the 20s proteasome both had a statistically significantly higher average occurrence of transglutaminases than genomes that had neither of these proteasomes. The species that have Anbu have over 5 times more transglutaminases on average than the species that lack both Anbu and the 20s proteasome. We observed the same result when we repeated

this measure in genomes from just the α -proteobacteria, β -proteobacteria, and γ -proteobacteria. There was no genome that had Anbu and completely lacked transglutaminase. It should be noted that a major exception to this trend is *Rhodopirellula baltica*. It has 11 transglutaminase catalytic domains, the most of any genome in this study, but has no proteasome homologs. These proteins are predicted to have a domain with similar structure to the transglutaminase associated with Anbu, but their functions could be very different. Transglutaminases do not strictly require Anbu, but there is a definite association between them. Understanding Anbu's function will require better characterization of the different functions of bacterial transglutaminases as well as COG2307 and COG2308.

The few samples of BPH showed 2 operon based patterns. In *Thiobacillus denitrificans* and *Chromobacterium violaceum*. BPH is in the same operon or regulon as ornithine carbamoyltransferase (argI or argF) and argininosuccinate synthase (argG) (Figure 5-6b). Both of these proteins are involved in arginine biosynthesis which is induced as part of the heat shock response in several species including *Bacillus subtilis* [196] and *Desulfovibrio vulgaris* [197]. HslV and HslU are in the same operon as argF, which is next to argG in *Desulfuromonas spp.* BPH is in the same operon as Heat Shock Protein 33 (HslO), a chaperone that is activated under oxidative stress[198], in *Chromobacterium violaceum*. In these species BPH appears to be acting as part of a heat shock response. This could be the result of functional conservation if we are correct that BPH evolved from HslV. Both *Thiobacillus denitrificans* and *Chromobacterium violaceum* have BPH and HslV but lack Anbu. Identifying the difference in conditions that induce expression of BPH and HslV will help explain BPH's function. However,

BPH seems to play a different role in the other species that have Anbu as well. In the *Ralstonias* and *Poloramonas* BPH was in an operon with the 3 genes encoding the pyruvate dehydrogenase complex. In *Escherichia coli* these genes are in the same operon as the auto regulator *pdhR*. *pdhR* represses transcription of that operon in the absence of pyruvate[199]. BPH may play a similar regulatory role, degrading the pyruvate dehydrogenase complex in the absence of pyruvate. It would be interesting if transcriptional regulation was replaced by regulation at the level of degradation. It is possible that BPH has been adapted to both regulatory and heat shock roles, but its hard to draw a conclusion on how conserved these operons are from a sample of only 6 species.

Discussion

Anbu's position in the trees and its hypothetical structure make a compelling case for it being ancestral to the 20s proteasome found in the actinobacteria. Sequence and functional data indicates that BPH evolved from HslV. Determining whether HslV or Anbu is older is a much more challenging problem. Cavalier-Smith argues that the oldest groups of bacteria are the Cyanobacteria, Hadobacteria, and Chlorobacteria (from youngest to oldest) [28]. Neither HslV nor Anbu has been found in any Chlorobacterial genome. Anbu is present in several Cyanobacteria. *Thermus Thermophilus*, a Hadobacteria, has HslV. Its sequence is related to HslV of other hyperthermophiles which may reflect a horizontal transfer. This makes it hard to say which proteasome is

older based on their distribution in these bacteria. The pattern of repeated loss of Anbu in genomes that have HslV infers HslV replaced Anbu. In this scenario Anbu would be the oldest proteasome. Solving the biological units of Anbu and probing its interactions may also help sort this out by showing which transitions between proteasomes are the most favorable.

It has been argued that the actinobacteria are ancestral to both the eukaryotes and archaea because they are the only group of bacteria with a 20s proteasome while the 20s proteasome is found in all eukaryotes and Archaea [28]. Although we have shown that Anbu is more likely to be the ancestor of the 20s proteasome than HslV, our data still supports the actinobacteria having the original 20s proteasome. A horizontal transfer of the 20s proteasome to the actinobacteria as proposed in [171] is unnecessary with the discovery of Anbu. Our work also shows that bacterial evolution has tinkered with the proteasome much more than previously thought. We have found bacteria that have many different combinations of the 20s proteasome, HslV, and Anbu. It is important to note there is evidence that any proteasome can and has been lost under the right circumstances. Many of these conclusions can only be drawn because of the large number of genomes we looked at in this study, but this number will be considered small in a few years. There may be many groups of species specific proteasomes like BPH in parts of the tree of life that have not been sampled. Finding a group of bacteria outside the actinobacteria with a true 20s proteasome would have major implications for the evolution of the eukaryotes and archaea, but until then the actinobacteria proteasomes seem the most plausible ancestor of eukaryotic and archaeal proteasomes.

Methods

The Superfamily[172] database was used to identify 216 bacterial proteins (Supplemental Table 3) in the SCOP[145] proteasome family. All hits had e-values less than .0001 at the superfamily level. We took all hits to the proteasome family regardless of e-value because we are interested in proteins that are not represented by known structures. The hit from *Deinococcus radiodurans* was not included because this sequence was a multidomain protein, while all other sequences only included a proteasome subunit domain. This protein may include N-terminal nucleophile aminohydrolase domain as it weakly hits that superfamily. Since it does not align with any other proteasome subunit, including those from *Thermus Thermophilus*, this sequence is probably not a proteasome subunit. *Frankia*, an actinobacteria, had a hit in addition to the 20s proteasome. This sequence did not align well with any of the 5 clusters, and is also probably an N-terminal nucleophile aminohydrolases, but not a proteasome subunit. Excluding this sequence increased the quality of the multiple alignment.

Sequences were aligned using MUSCLE[200], part of the STRAP (<http://www.charite.de/bioinf/strap/>) suite of programs. Multiple structural alignments were performed using Combinatorial Extension [181], also packaged in STRAP. All trees were built using PHYML[201] with the JTT model of evolution, estimated variance and gamma, and 4 substitution rate categories. PHMYL was packaged as part of Geneious [202] (<http://www.geneious.com/>). Each tree was bootstrapped from 100

replicates.

Representative proteasome subunits were taken from the PDB [78]. These structures came from various species in all 3 superkingdoms. 1JJW, 1E94, and 1M4Y are HslV structures. 1Q5Q and 2FHG are actinobacterial 20s proteasomes. 2 sequences from Anbu (from *Rhodopseudomonas palustris* and *Hahella chejuensis*) and 2 sequences from BPH (from *Thiobacillus denitrificans* and *Ralstonia metallidurans*) were threaded using the Phyre web server, which is the successor of 3D-PSSM [180]. Each predicted model was created from several known structures. All resulting structure predictions have very high structural similarity to known proteasome subunit structures. The predicted models were aligned to create a phylogenetic tree. All structural images were created in Protein Workshop, part of the Molecular Biology Tool kit [203].

BLAST [183] searches were performed using HslU from *Ralstonia solanacearum* (GI:17427050) and proteasome-associated ATPase from *Mycobacterium tuberculosis* (GI:113700393) against cyanobacteria and β -proteobacteria to find potential ATPases for Anbu and BPH.

Table 5-2 was created using Superfamily's predictions for the transglutaminase catalytic domain. The P-values in Table 5-2 were calculated using a one-tailed t-test without the assumption that the variances of the groups were equal.

Acknowledgements

We would like thank Russell Doolittle for his useful discussions and Stella Veretnik for reviewing the manuscript.

Figures

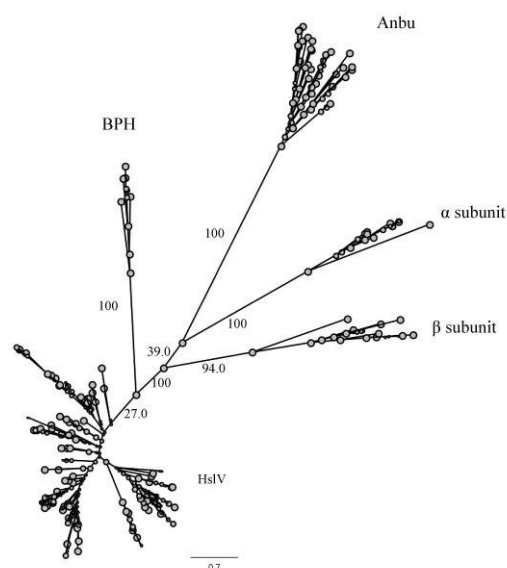


Figure 5-1. Maximum likelihood tree of proteasome subunits in the Superfamily database.

100 replicates were run to obtain bootstrap values. The Anbu and BPH clusters represent 2 novel proteasome homologs. Anbu's position near both subunits of the 20s proteasome implies it is ancestral to the 20s proteasome.

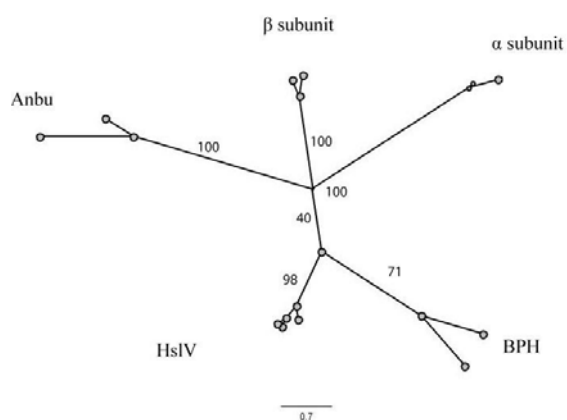


Figure 5-2. Maximum likelihood tree from a structural alignment.

7 proteasome subunits were taken from the PDB and 4 structural predictions (2 from Anbu and 2 from BPH). The placement of BPH and Anbu in this tree is in agreement with the tree in figure 1.

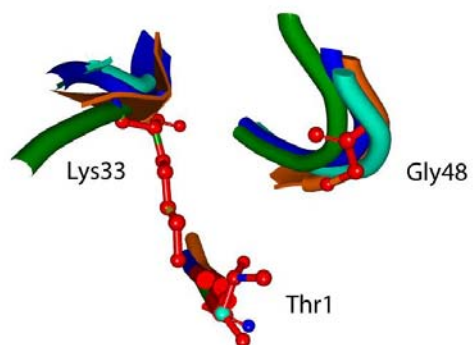


Figure 5-3. Comparison of catalytic triads in different proteasomes.

HslV (1E94) is green, β subunit (1Q5Q_H) is cyan, Anbu (predicted structure from *Rhodopseudomonas palustris*) is blue, and BPH (predicted structure from *Thiobacillus denitrificans*) is orange. The side chains of HslV are colored red. The corresponding backbone and neighboring residues are visible from each structure. All 3 sites are highly conserved in sequence as well as structure. The labels refer to the positions of these residues in 1E94.

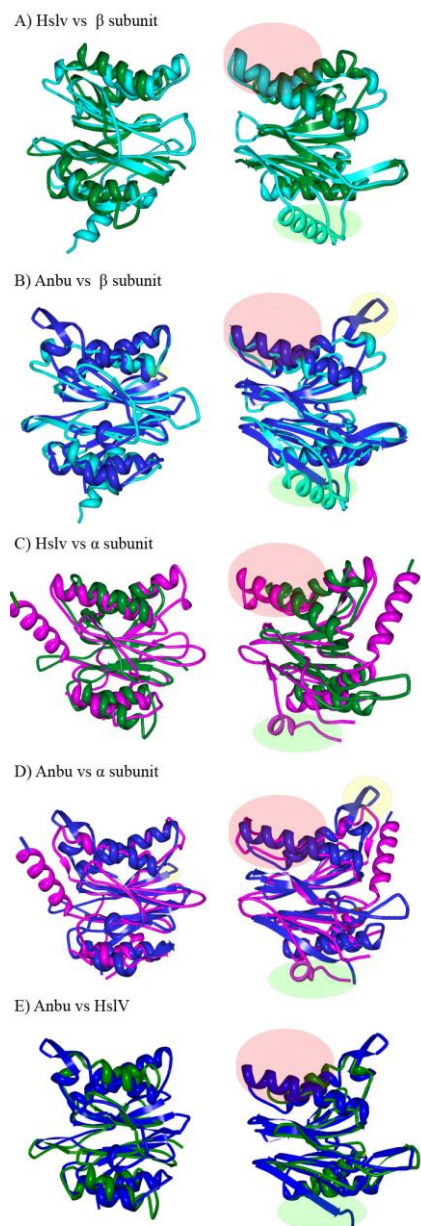


Figure 5-4. Comparison of Anbu to crystal structures of known proteasomes.

The image on the right is approximately a 180 degree rotation of the image on the left. HslV (1E94) is green, α subunit (1Q5Q_A) is magenta, β subunit (1Q5Q_H) is cyan, and Anbu (predicted structure from *Rhodopseudomonas palustris*) is blue. Red ovals highlight an extended helix shared between Anbu and both subunits of the 20s proteasome, but is absent in HslV. Green ovals highlight an extended c-terminal shared between Anbu and both subunits of the 20s proteasome, but is absent in HslV. The yellow ovals highlight an extended turn that is unique to Anbu.

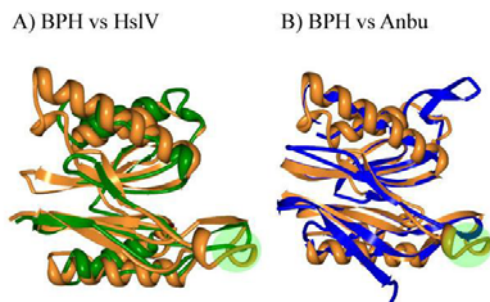


Figure 5-5. Structural Comparison of BPH.

BPH (predicted structure from *Thiobacillus denitrificans*) in orange against HslV (1E94) in green and Anbu (predicted structure from *Rhodopseudomonas palustris*) in blue. The green oval highlights an extension unique to BPH.

A) Anbu's Operon

Tables

Table 5-1. Combinations of proteasomes in bacteria.

Different combinations of proteasome homologs found in 238 bacterial genomes. Anbu, BPH, and HslV are found in several combinations, but never with the 20s proteasome.

Combination	Occurrence in Genomes	Example Species
No proteasome homolog	64	<i>Synechococcus</i> sp.
HslV	132	<i>Escherichia coli</i>
20s Proteasome	10	<i>Mycobacterium tuberculosis</i>
Anbu	6	<i>Synechocystis</i> sp. PCC 6803
Anbu + HslV	19	<i>Rhodobacter sphaeroides</i>
BPH	0	No occurrence
BPH + Anbu	1	<i>Polaromonas</i> sp.
BPH + HslV	2	<i>Thiobacillus denitrificans</i>
BPH + Anbu + HslV	3	<i>Ralstonia solanacearum</i>

Table 5-2. Comparison of genomic occurrence of transglutaminase-like catalytic domains.

The average number of transglutaminases per genome was computed for genomes with different proteasome content. Neither means genomes that have neither Anbu nor the 20s proteasome. The p-values are the probability of the mean being that much higher compared to the genomes that have neither proteasome. This calculation was repeated in just the genomes of α -proteobacteria, β -proteobacteria, γ -proteobacteria. Genomes that have Anbu or the 20s proteasome have significantly more transglutaminase domains than genomes that have neither Anbu nor the 20s proteasome.

	Neither	Anbu	20s proteasome	$\alpha\beta\gamma$ -Neither	$\alpha\beta\gamma$ -Anbu
Mean	0.752525253	3.17241379	2.454545455	0.722222222	3.24
Variance	1.831846383	3.14778325	3.072727273	1.596129838	3.606667
P-value		2.27E-08	0.004472988		3.42E-07

Chapter 6: Structural analysis of polarizing indels: an emerging consensus on the root of the tree of life

Ruben E. Valas and Philip E. Bourne

Biology Direct 2009, **4**:30doi:10.1186/1745-6150-4-30

Part of the series “Evolutionary Biology 150 years after the 'Origin': is a post-modern synthesis in sight?”

Abstract

Background

The root of the tree of life has been a holy grail ever since Darwin first used the tree as a metaphor for evolution. New methods seek to narrow down the location of the root by excluding it from branches of the tree of life. This is done by finding traits that must be derived, and excluding the root from the taxa those traits cover. However the two most comprehensive attempts at this strategy, performed by Cavalier-Smith and Lake *et al.*, have excluded each other's rootings.

Results

The indel polarizations of Lake *et al.* rely on high quality alignments between paralogs that diverged before the last universal common ancestor (LUCA). Therefore, sequence alignment artifacts may skew their conclusions. We have reviewed their data

using protein structure information where available. Several of the conclusions are quite different when viewed in the light of structure which is conserved over longer evolutionary time scales than sequence. We argue there is no polarization that excludes the root from all Gram-negatives, and that polarizations robustly exclude the root from the Archaea.

Conclusions

We conclude that there is no contradiction between the polarization datasets. The combination of these datasets excludes the root from every possible position except near the Chloroflexi.

Reviewers

This article was reviewed by Greg Fournier (nominated by J. Peter Gogarten), Purificación López-García, and Eugene Koonin.

Background

There are two basic strategies for rooting the tree of life and defining the nature of the last universal common ancestor (LUCA), inclusive and exclusive. The recent argument for an Archaeal rooting [26] is inclusive and relies on arguments of primacy to establish the root. We feel such arguments often lead to circular reasoning based on one's expectation of what primitive cellular life should look like. For instance if one assumes that cellular life began in hydrothermal vents systems (reviewed in [204]) then one could argue that any organisms living in or near hydrothermal vents could be LUCA-like. But this does not prove that cellular life started in that condition. And even if it did, it is

possible that later organisms invaded that niche, so the extant organisms there today are nothing like LUCA. Paralog rooting, where one uses paralogous sequences as an outgroup in a sequence tree [30, 31], is technically an inclusive method since it attempts to determine which groups of sequences are the most primitive. However, this method is not self consistent [32, 205] and technical objections have been raised [28].

Exclusive rooting defines branches as derived and thus they are omitted until only the root is left, thereby establishing LUCA. Ideally these two strategies would converge, but at this point there is no consensus even within one particular strategy as there are multiple ideas on the nature of the most primitive cellular systems [206, 207] as well as how to properly exclude the root from a particular branch [37, 208].

One method for arriving at an exclusive solution is top-down rooting using indels [34-37]. Usually an indel is ambiguous: it could be an insertion or a deletion. But if one knows the ancestral state the indel is polarized. That is, one can say which forms of the indel are the ancestral and derived states. One can then exclude the root from any branches where all the organisms have a derived form of the gene. One can infer the ancestral state of an indel by comparing a pair of paralogous genes that were duplicated before LUCA. Traditionally this technique would require a paralog set to be ubiquitous. Otherwise one could not be sure the paralogs diverged before LUCA. The advantage of top down rooting over traditional indel polarization is the ability to handle non-ubiquitous genes by considering gene loss and invention as well as insertion and deletion when analyzing the most parsimonious scenario for the history of a paralog set.

Lake *et al.* have presented 8 indel polarizations (summarized in [27, 209]). They conclude the root of the tree of life lies between two clades. The first is the

Actinobacteria (single membrane bacteria) and Gram-negatives (which Lake *et al.* refer to as double membrane bacteria). The second is the Firmicutes and Archaea (both of which contain single membranes). The authors have presented indels that apparently exclude the root from each of these clades so they conclude the root must lie between them. We present evidence using the addition of protein structure data that implies this conclusion is not supported.

There are several prerequisites for an indel argument to be correct in polarizing the tree of life. First, one needs a set of nearly universal paralogs (at least universal to the taxa being rooted). Second, one needs a quality alignment between those paralogs. This is often difficult as paralogs duplicated before LUCA have billions of years to drift and are under different selective pressures. The conclusions reached rely heavily on the alignment and this is the Achilles' heel of indel polarization. Where protein structures exist they offer the opportunity to get past the limitation of sequence drift, since structure is more conserved than sequence over long evolutionary time scales, and hence provides strong evidence when aligning proteins that diverged before LUCA. We introduce structural information into Lake *et al.*'s analysis where possible.

Cavalier-Smith has presented 13 polarizing transitions using a variety of data to reach an exclusive solution [208]. These transitions include information from indels, quaternary structures, as well as cellular organization. It is difficult to do his analysis justice in a short summary but here are the main points of his argument. He excludes the root from the Archaea and Eukaryotes based on proteasome evolution. He argues the transition from a Gram-positive membrane structure to a Gram-negative membrane would be much more difficult than the other direction, so this excludes the root from the

single membrane prokaryotes. The Chloroflexi have the simplest known outer membrane, lacking outer membrane protein 85 (OMP85). OMP85 is present across all other gram-negative taxa so Cavalier-Smith places the root within or next to the Chloroflexi. He presents other arguments as well that resolve the structure of the rest of the tree of life.

Despite reaching different conclusions about the root, both Lake *et al.* and Cavalier-Smith agree the root must lie within the Bacteria by excluding the root from the Archaea and Eukaryotes. Both agree that the Archaea must be derived from a Gram-positive bacterium, but Cavalier-Smith argues it was an Actinobacteria and Lake *et al.* argue it was a Firmicute. The arguments for each seem sound; both groups probably contributed genes to the Archaeal ancestor. The difference between which are the result of vertical versus horizontal transfer is not yet resolved in our opinion. The two methods also agree that the root is not within the Actinobacteria or the Firmicutes. It is important to realize that despite claims that all rootings of the tree are contradictory [210], the newest exclusive methods are converging on a Bacterial rooting. The most important thing these two bodies of work agree on is that there is a single backbone to the tree of life that can be resolved using rare events in evolution despite recent claims by the sequence tree community that no such tree exists or can be built [5, 57]. However, the apparent disagreement between these datasets weakens the position that a single tree of life exists. We believe this work supports the idea that a backbone to the tree of life can be resolved and rooted by showing their analyses converge on the same result.

This work will focus on the fundamentally different conclusions about where the root of the Bacteria lies. Cavalier-Smith argues for a Gram-negative root based on his ideas about cells having an inside out origin, obcells [211]. He argues it would be easier

for a Gram-positive cell to evolve by simply losing the outer membrane than it would be for a Gram-positive cell to gain an outer membrane and the cellular machinery needed to make it functional. The idea of a Gram-positive root is compatible with several scenarios for the origins of cells as well [24, 212]. This argument will probably not be resolved on the basis of which theory of the origin of cells is more elegant since most of the ideas of early cellular evolution are highly speculative. Instead continued polarization of the transition between Gram-negatives and Gram-positives will lead to an understanding of which of these scenarios is even plausible.

We believe the indel in GyrA robustly excludes the root from the Actinobacteria using sequence alone [36] and there is no need to invoke structural alignments. However, we will subsequently present structural alignments, as well as other data, that support the exclusion of the root from Archaea based on insertions in elongation factors [35] since objections to these conclusions have been raised [26, 213]. Lake *et al.* present 3 polarized indels that they claim exclude the root from the Gram-negatives: HisA (*P*-ribosylformimino-AICAR-*P*-isomerase), Hsp70 (heat shock protein 70 aka DnaK), and PyrD (dihydroorotate dehydrogenase). We will present evidence to suggest that none of these arguments truly excludes Cavalier-Smith's rooting. The Eobacteria (Cavalier-Smith's term for *Deinococcus-Thermus* and *Chloroflexi*) have the ancestral form of HisA despite being Gram-negatives. The conclusions about Hsp70 are based on a sequence alignment artifact, which is evident when a structural alignment is used instead. The arguments made by Lake *et al.* using PyrD are not self consistent, so we polarize this indel using quaternary structure. This excludes the root from the Archaea and Firmicutes, and probably from their last common ancestor as well. We also discuss the insert in

Ribosomal Protein S12, which would have the potential to exclude Cavalier-Smith's rooting, but does not.

Results

Indels in elongation factors place the root within Bacteria

Several objections have been raised against the exclusion of the root from Archaea based on indels in the paralogs of initiation factors (IF) and elongation factors (EF) [26, 213]. They claim the conclusions reached in [35] are based on alignment artifacts. These indels would be ideal to analyze using structure since they narrow the root to a single superkingdom. Di Giulio criticizes the alignment between EF-G and EF-Tu because there is a 4 residue stretch between the insertions that is more similar between some paralogs than between some orthologs [213]. Di Giulio is correct in raising a red flag here; there is probably an artifact in the sequence alignment. However, we argue the exclusion of the root from the Archaea is still valid in spite of that artifact.

Unfortunately the crystal structure of EF-2 (the Archaeal and Eukaryal orthologs of EF-G) from the Eukaryotes have a large disordered (unresolved) region near the indel of interest, hence these proteins are less suited for a structural alignment than the ones discussed below. The multiple structure alignments of these 3 regions is of poor quality due to the disordered region, which can be seen in the differing positions of the highly conserved residues on each end of this region (glycine colored green and aspartic acid colored magenta in Figure 6-1). However, the middle of the alignment seems reasonable and supports an insertion in EF-2 at the root of the Archaea.

We counted the distance between the well conserved RG(IV)T and PGH motifs in all elongation factors to reach a stronger conclusion (Table 6-1). Some sequences lack these motifs, but it is strongly implied they were present in the ancestral elongation factors since they are conserved across paralogs. The motifs are 20 residues apart in every EF-Tu and EF-1 sequence that have the motifs. The majority (55.95%) of EF-G sequences have the motifs 20 residues apart. This may actually be an underestimate because the next most populated length of 27 residues (32.16%) are mostly from β and γ -proteobacterial sequences. According to the Genomes Online Database [214], of the 1000 completed genomes published as of May 2009, 64 are β -proteobacterial and 215 are γ -proteobacterial genomes. These groups are over sampled relative to many others which would deflate the true proportion of the EF-G sequences that lack an insertion relative to EF-Tu. Even so, the most parsimonious ancestral elongation factor would have 20 residues between these motifs. Every Archaeal sequence that has the motifs in EF-2 has them 24 residues apart. Therefore, regardless of the actual alignment there must be a 4 residue insertion somewhere in EF-2 of the Eukaryotes and Archaea. Therefore the conclusion that the root can be excluded from the Archaea [35] is correct even though there is a sequence alignment artifact.

The region of the indel in IF-2 using EF-G/Ef-2 as an outgroup examined in [26] is also in a region that does not align well structurally. Its sequence anchors are also much less conserved than in the indel discussed above, so the critique of this indel may be correct. However, the strength of this indel polarization appears to be a moot point. To the best of our knowledge no one has argued against Cavalier-Smith's exclusion of the root from Archaea based on proteasome structure [62, 208], which is strongly

supported by our own conclusions on proteasome evolution [71]. That taken with the derived insertion in EF-G and the quaternary structure of PyrD (discussed below) there are 3 strongly polarized arguments that each place the Archaea as derived from the Bacteria. To the best of our knowledge there is not a single argument that excludes the root from all the Bacteria in the same way these 3 polarizations exclude the root from the Archaea. Therefore the goal of the rest of the analysis of the indel polarizations is to narrow the root within the Bacteria.

HisA and HisF exclude the root from all Gram-negatives except the Eobacteria.

HisA and HisF are an ideal paralog set because they are nearly ubiquitous and have a relatively high degree of sequence similarity among paralogs. A structural alignment of the 3 forms of this indel reveals that the conclusions based on sequence alignments are valid (data not shown). Lake *et al.* conclude this excludes the root from the Actinobacteria and Gram-negatives [27]. However, their own summary of the indel shows the insertion that is present in most Gram-negatives is apparently absent in a *Deinococcus* genome. A realignment of just a few species that have the insert with the Eobacteria shows that all 11 fully sequenced Eobacterial genomes have a deletion relative to the other Gram-negatives (Figure 6-2). This means that the indel in HisA actually excludes the root from the Actinobacteria and all Gram-negatives except the Eobacteria. Cavalier-Smith claims the Eobacteria are some of the most ancient bacteria because they lack lipopolysaccharide in their membranes. The fact that HisA does not exclude Cavalier-Smith's root would not matter on its own, because two other indels apparently exclude the root from the Eobacteria. But we will argue neither of these arguments holds

water, and that the results of Lake *et al.* and Cavalier-Smith converge on a rooting within the Eobacteria.

Protein structure alignment renders the Hsp70/MreB indel inconclusive

Lake *et al.* claim that the Hsp70/MreB indel excludes LUCA from the Gram-negative bacteria [37]. This is not a new idea, and was first proposed by Gupta 10 years ago [215]. Hsp70 contains a large indel between the Gram-positives and Gram-negatives. Since Hsp70 is nearly universally distributed, if one can deduce the ancestral state of Hsp70 it would reveal which group is ancestral and which is derived. There is no indel between MreB and Hsp70 from the Gram-positives in Gupta's alignment. He argues the Gram-negatives are derived since they have an apparently derived insertion in Hsp70. However, Philippe has made the argument that MreB and Hsp70 are very distant paralogs, so it is difficult to align them [216]. In his alignment it is not clear whether or not the Gram-positive form of Hsp70 has an insertion relative to MreB. He raises the possibility that there are actually two independent insertion-deletion events. The newer work on this indel has dealt with the issue of the gene being missing in some species, but has not significantly improved the quality of the alignment [37].

The recently solved crystal structure of a Gram-positive Hsp70 from *Geobacillus Kaustophilus* provides an opportunity to review the Hsp70/MreB situation using a structural alignment [217]. Structures of Hsp70 from both Gram-positive and Gram-negative bacteria were aligned with MreB using the CE-MC webserver [218]. These structures align well, which is expected since they are all in the same SCOP superfamily [145]. It is implied that Hsp70 from Gram-positives aligns perfectly to MreB in this

region [37, 215, 219]. A review of the structural alignment reveals this is not the case. Rather Hsp70 from Gram-positive bacteria have an insertion relative to Mreb (Figure 6-3). There has to be 2 independent insertion-deletions events here to account for the 3 different structures seen in this region. Therefore, it is impossible to determine the ancestral state of Hsp70. Every scenario requires two insertion-deletion events regardless of the root, and therefore this indel cannot be used to polarize the transition between Gram-positive and Gram-negative bacteria.

Quaternary structure of PyrD excludes the root from the Archaea and Firmicutes

Lake *et al.* have polarized an indel in PyrD using HemE (uroporphyrinogen decarboxylase) to exclude the root from the Archaea and Firmicutes [34]. Later they polarized the same indel in PyrD using HisA and HisF as outgroups to exclude the root from the Gram-negatives and the Actinobacteria [27]. With these conclusions one could root the universal tree of life by polarizing the PyrD indel alone. This appears to be supported by the indels in HisA and Ribosomal Protein S12, but as discussed above and below, respectively, the conclusions Lake *et al.* reach on these 2 indels are also in question. We argue there is a contradiction in the analysis of PyrD, and propose an argument based on quaternary structure to resolve this contradiction.

All of the most parsimonious rootings with a HisA (or HisF) outgroup have the ancestral state of PyrD being a deletion relative to the derived state [27]. The authors consider this result independently of their results with HemE outgroups. However, the ancestral state of PyrD should be the same regardless of the outgroup. All of the trees that are the most parsimonious with the HemE outgroup imply the ancestral state of PyrD

was an insertion relative to the derived state [37]. It is impossible for any one rooting to be the most parsimonious with both HemE and HisA as outgroups.

There are two possible sources of the contradiction. The first is an alignment artifact. Our structural alignment of PyrD, HisA, and HisF is in agreement with Lake *et al.*'s sequence alignment (data not shown). HemE appears to be more distant in structure and sequence to these other 3 proteins than they are between themselves. The structure alignments between PyrD and HemE are not consistent. They vary greatly depending on which structures are used. The alignment in [34] is between the 3rd beta sheet in HemE and the 7th beta sheet in PyrD. These regions are technically homologous because this fold arose through a series of internal duplications [220], but the fact these are different regions within paralogous structures indicates this alignment should not be used for polarizing the indels. The duplication between the paralogs is more recent than the duplication between the subbarrels, so one should be aligning the same region of the structures between paralogs. If we assume their sequence alignments to be correct then the other possible source for the contradictory conclusions is convergent evolution. There are several variants of PyrD and HemE at this site which include an additional 1 residue indel. This implies this region of PyrD is tolerant to small indels, so convergent evolution at this site is not out of the question. Top-down rooting excludes all trees that are not the most parsimonious, which assumes there was no convergent evolution. In this case there is evidence for convergent evolution so top-down rooting should not be applied to this indel set.

Since the indel arguments contradict themselves, it is worth considering another line of reasoning. Lake *et al.* do not consider the quaternary structure of PyrD. There are

3 families of PyrD, each with a different solved quaternary structure [221]. The distribution of each family was examined using the NCBI Protein Clusters Database [222]. PyrD 2 (PRK07565) is a membrane bound monomer and is found across the Gram-negatives and Actinobacteria. PyrD 1A (PRK02506) is a homodimer that is mainly found within the Lactobacillales. PyrD 1B (PRK07259) is a heterotetramer found across the Archaea and Firmicutes (except in *Staphylococcus* that have PyrD 2). It has an extra subunit, PyrK. The core of this enzyme is a homodimer that is similar to PyrD 1A [223] (Figure 6-4). PyrD 1B has a deletion relative to the 2 other subfamilies. This deletion is polarized as the ancestral state when HisA or HisF are used as an outgroup, but is derived when HemE is used as an outgroup.

We argue the most parsimonious route for quaternary structure evolution would be monomer -> homodimer -> heterotetramer. A new protein-protein interface evolves at each step in this scenario. One can imagine a scenario where protein-protein interfaces are lost at each step but this requires a heterotetrameric ancestor. None of the other known structures in PyrD's or PyrK's superfamilies bind each other, which means there is no outgroup that makes a heterotetrameric ancestral state seem plausible. HemE is a homodimer and HisA is a monomer, so if either of these are the true ancestor of PyrD it does not make sense for the homotetramer to be the ancestral state. HisF is a heterodimer, but the other subunit appears to be unrelated to PyrK. Without a heterotetrameric outgroup the only way PyrD 1B could be ancestral is to have gained PyrK at the root of the PyrD tree. However, this subunit would have to be lost in PyrD 1A, so this is obviously not the most parsimonious scenario. The most parsimonious scenario for quaternary structure evolution is the one described above, and that excludes the root from

Firmicutes and Archaea as well as their last common ancestor. Even if evolution was not completely parsimonious in this case it does not negate our polarization that places PyrD 1B as the derived state. We argue that independent insertion events in this region are more probable than the homotetramer being the ancestral structure. At the very least, PyrD should be considered inconclusive for excluding the root since the sequence and structure arguments disagree. PyrD is another structural argument that the Archaea must be derived from the Gram-positives in line with previous arguments on proteasome evolution [62, 71, 208].

A maximum likelihood tree for PyrD 1B has good separation between the Firmicutes and Archaea (Figure 6-5). This implies this distribution is not the result of horizontal transfer, but rather each of these groups ancestrally had a derived form of the protein. The Crenarchaea and Euryarchaea each cluster separately too. The Archaeal ancestor probably had PyrD 1B, but it was lost in several Crenarchaea. It must be noted that PyrD 1B is present in the *Dehalococcoides* (a subgroup of Chloroflexi). Based on their position in the tree this could be a horizontal transfer from the Firmicutes. However, even if the *Dehalococcoides* invented PyrD 1B its presence across a single genus does not exclude the root from the Chloroflexi.

Ribosomal protein S12 and RpoC are probably not paralogs

Indel polarization requires special attention to the choice of paralogs. It has been argued that the indel in ribosomal protein S12 can be polarized using RpoC (DNA-directed RNA polymerase subunit beta') [34]. The authors claim this excludes the root from the Firmicutes and Archaea. This apparently derived insertion is present in all the

Chloroflexi, which is not discussed by the authors. Ribosomal protein S12 belongs to the “OB-fold” in SCOP. RpoC belongs to the fold “beta and beta-prime subunits of DNA dependent RNA-polymerase”. The overall structure of these proteins is different enough they are considered to be different folds. This alone is enough evidence that a sequence alignment between these proteins is probably meaningless. However the authors only claim the regions around the indel are homologous. They calculate an e-value of .002 that these 30 residues are paralogous in both proteins. This e-value is much worse than that of their other paralogous pairs (by up to several orders of magnitude). It is possible for there to be homology between proteins at a subdomain level as discussed in our recent review [224], but we see no evidence of that in this case. A pairwise alignment between ribosomal protein S12 (1J5E:L) and RpoC (2A69:D), both from *Thermus thermophilus*, was performed using FATCAT [225]. The regions in the sequence alignment do not align in the structure alignment at all. FATCAT concludes these structures are not similar (P-value of 9.96e-01). None of this is evidence these regions can be considered paralogous. It is possible these two regions do share a common ancestor, but since their structural context has changed it does not make sense to align their sequences. This raises the question of whether a sequence alignment can ever be considered meaningful without structural conservation. We conclude the indel in ribosomal protein S12 cannot be polarized using this out group. A structure search of the Molecular Modeling Database [226] revealed that no solved structures are homologous to ribosomal protein S12 in the region of interest despite their being many structures in the same family in SCOP. This indel probably cannot be polarized properly, and is not evidence against a root within the Chloroflexi.

Discussion

There are four conclusions of Lake *et al.* that would disprove the rooting of the tree of life proposed by Cavalier-Smith if they are correct. We show that all four of these arguments have flaws and that is evidence that Cavalier-Smith's rooting is probably correct. The fact that the HisA indel excludes all the Gram-negatives except the Eobacteria is certainly a novel piece of evidence that supports the rooting within the Eobacteria. There are only a limited number of paralogs sets that are ubiquitous enough to be useful for rooting the tree. These sets will probably be exhausted without truly contradicting the Eobacterial root. Indel analysis reliably excludes the root from the Archaea, Actinobacteria and all Gram-negatives except the Eobacteria (summarized in Figure 6-6). Our polarization of PyrD's quaternary structure excludes the root from the Archaea and the Firmicutes, and their last common ancestor. If we combine these new interpretations of the indel data with Cavalier-Smith's 13 polarizing arguments there are no contradictions. All of this data supports the notion that LUCA must be near the Chloroflexi.

One of the major unresolved questions about LUCA is whether it had a DNA or RNA genome [227]. We argue that if LUCA was Chloroflexi-like, then it might have had a u-DNA genome. Thymidylate synthase is an essential enzyme that catalyzes formation of dTMP. There are two unrelated enzymes that perform this function, ThyA and ThyX [228]. The other 4 DNA nucleotides are converted from their RNA counterparts by ribonucleotide reductase. This implies there was a stage in evolution where DNA used uracil instead of thymine [229].

Thymidylate synthase follows 2 distinct patterns of evolution in the fully sequenced Chloroflexi genomes. All the *Dehalococcoides* have both ThyA and ThyX. This must be a derived state since one of the enzymes must have arisen before the other. It is possible that LUCA contained both of these enzymes, but very few species retain both of them. In many cases horizontal transfer displaces one with the other instead of retaining both as can be seen by looking at the distribution of these enzymes on a species tree (data not shown). It is far more likely that at least one of these enzymes is the result of a later horizontal transfer to the *Dehalococcoides*. The rest of the Chloroflexi contain only ThyX. However, this ThyX contains a domain duplication. This duplicated version of the protein has not been characterized, but is present in a few other species. It is possible that LUCA had a duplicated ThyX and the rest of the species have lost a domain, but this is clearly less parsimonious than this form of ThyX being derived.

We postulate that LUCA was Chloroflexi-like with a u-DNA genome. One of the first major branching points in the modern tree of life would be the origin of thymidylate synthase. All u-DNA genomes would eventually be out competed by t-DNA genomes in similar niches. Any u-DNA genome that received thymidylate synthase from a horizontal transfer should be all right since they already had all the machinery necessary for dealing with u-DNA. They would out compete similar u-DNA species. The distribution of thymidylate synthase in the Chloroflexi can only be explained by multiple horizontal transfers. It is possible one of the thymidylate synthases in *Dehalococcoides* represents the ancestral form and there were two later displacements (or a displacement and duplication event). Distinguishing between these scenarios is very difficult as an ancient horizontal transfer so close to the root of the tree would mimic vertical descent.

Unfortunately all the sequence trees constructed have low bootstrap values for the critical edges, so we cannot conclude that each of the 3 different thymidylate synthases in the Chloroflexi are the result of horizontal transfer (data not shown). Therefore this might not be a falsifiable hypothesis, but it certainly is an interesting idea.

One could argue we have excluded the root from the ancestor of the Archaea and Firmicutes based on the presence of a derived PyrD, but have come up with an odd scenario to justify a rooting we like that has a derived thymidylate synthase. The major difference is that PyrD 1B appears to be present in almost every Archaeal and Firmicute genome, including the apparently deep branching Thaumarchaeota and Korarchaea, which makes horizontal transfers after their last ancestor unlikely. It is very unlikely that the history of thymidylate synthase in the Chloroflexi involved no horizontal transfers. There is also ample evidence that ThyX and ThyA frequently replace each other through horizontal transfers. It is reasonable for the Chloroflexi to have some derived traits even if they represent the most ancestral branch of the tree of life, as long as these traits appear to be the result of later transfers as in this case. Even with the correct rooting, it might not be possible to reconstruct LUCA because of later horizontal transfers, but this is a good example of how one can tell when a derived trait has been transferred to an ancient group.

There is growing evidence that the Archaea are derived from the Gram-positive Bacteria [34, 35, 62, 208, 219]. However there is still disagreement on whether that Gram-positive ancestor was a member of the Firmicutes or Actinobacteria (which is why the Archaea are placed ambiguously in Figure 6-6), as well as the source of selective pressure that was great enough to give rise to a novel superkingdom. Serious objections

have been raised to the possibility that Archaea are derived from Bacteria based on differences in DNA replication machinery [230], but our own analysis suggests this divide is not as vast as some have suggested (in preparation). But that is beyond the scope of this work.

One of the corollaries of a rooting near the Chloroflexi is that the first true cells had two membranes. This may seem counterintuitive, but is actually well explained by Cavalier-Smith's obcell theory [211]. The idea is that the first organisms were cells that had nucleozymes anchored by short hydrophobic peptides to the outside of the membrane. In other words life started on the outside of cells, not inside them. This gets around the difficulty of forming transmembrane pores, a major difficulty in the RNA world. Heredity would still be based on the division of membranes, just as it is now. If two obcells were to fuse, the result would be a double membrane protocell. Very little follow-up has been done to test the plausibility of the obcell theory. We hope the additional evidence for a Gram-negative root provided in this work motivates other to investigate this idea further. There are currently about a dozen sequenced Eobacterial genomes compared to the hundreds of proteobacterial genomes discussed above. More research into the genomics of these possibly early branching bacteria could bring many obscured details about LUCA into the light.

Methods

The multiple structural alignment in Figure 6-1A was performed using the MUSTANG webserver[231] (<http://www.cs.mu.oz.au/~arun/mustang/>) and the one in Figure 3 was performed using the CE-MC webserver [218]

(<http://pathway.rit.albany.edu/~cemc/>). We used different alignment algorithms for these 2 data sets because in each case one program gave a higher quality alignment than the other. Pairwise structural alignments were performed using FATCAT [225]. Molecular graphics images were produced using the UCSF Chimera package [232].

The phylogenetic tree in Figure 6-5 was constructed using Phym1 [201], packaged as part of Geneious (<http://www.geneious.com/>). The tree was built from the multiple alignment of PRK07259 in the NCBI Protein Clusters Database [222]. The tree was drawn and colored using FigTree (<http://tree.bio.ed.ac.uk/software/figtree/>).

The data used to generate Table 6-1 was taken by looking at the most populated clusters for each elongation factor in the NCBI Protein Clusters Database [222]. EF-Tu statistics were calculated using the sequences in PRK00049, PRK12735, and PRK12736. EF-1 sequences are from PRK12317. EF-G sequences are from PRK13351, PRK12740, PRK12739, and PRK00007. EF-2 sequences are from PRK07560. Only sequences that had both the RG(IV)T and PGH motifs were used to calculate insertion lengths, as these are then only cases that can be unambiguously compared without actually aligning the paralogs.

Competing interest

This work is a critical evaluation of the work done by Lake *et al.*, so there is an academic competing interest.

Authors' contributions

REV conceived the study and analyzed the data. PEB assisted in writing the manuscript.

Reviewers' comments

Reviewer's report 1

Greg Fournier, Department of Molecular and Cell Biology, University of Connecticut, Storrs, CT 06269-31258, USA (nominated by J. Peter Gogarten, Department of Molecular and Cell Biology, University of Connecticut, Storrs, CT 06269-31258, USA)

In this article, Valas and Bourne attempt to resolve two disparate indel-based rootings of the tree of life proposed by Cavalier-Smith [208] and Lake [27], respectively. In each case, the presence of indels is used as a polarizing character for pairs of paralogous genes that duplicated before the LUCA, allowing for the exclusion of the root from particular branches of the tree. Using different sets of genes containing indels, Cavalier-Smith argues for a root within the Actinobacteria, while Lake argues for a root within the Firmicutes. The authors determine that the work presented by Lake *et al.* [27] does not necessarily exclude the rooting reported by Cavalier-Smith, thus supporting a root within the Actinobacteria, near the Chloroflexi.

Author's response:

This is incorrect. Cavalier-Smith is not arguing that the root is anywhere near the Actinobacteria. He is arguing for a Gram-negative root and the Actinobacteria are Gram-positive. Lake et al. are also not arguing for a root within the Firmicutes either.

We realize this is a confusing subject since each group is referring to non-standard higher level taxa. We have tried to be more explicit about the traditional vs non-traditional names as well be clear as to the number of membranes each of these groups has. It is vital to understand that Cavalier-Smith's root is based on multiple types of polarized evidence including but certainly not limited to indels.

Various methods exist for rooting the tree of life, primarily either using polarizing characters, or reciprocal rooting of paralogs. While the authors mention paralog rooting as an alternative method, their claim that it is not “self consistent” is an overstatement, as the supermajority of paralogs used in these analyses support a rooting within the bacterial stem, with a few others showing weaker support for a root at an undertermined position within the bacterial domain. None support a rooting within either the archaea or eukarya. That being said, the authors' stated objective is to root the tree relying only on indel-based evidence, and it is only fair to evaluate their conclusions solely within that context.

Author's response:

We actually disagree with the statement that this work should be judged on indels alone. Our goal is to root the tree using any data possible. This work is just an evaluation of Lake et al.'s works on indels, but its important to remember the context of the argument rests on other data sources. We do not think we have overstated the inconsistencies created from paralog rooting. In the most comprehensive search for informative paralog rootings the true supermajority (137 out of 154) were inconclusive because they made both the Archaea and Bacteria polyphyletic due to horizontal transfer

[32]. Of the 17 remaining paralogs sets 9 supported the rooting between the stem Archaea and Bacteria, 7 supported rootings within Bacteria, and 1 supported a rooting within the Archaea. The authors advised to use caution when accepting the rooting between the superkingdoms, because it is consistent with the tree long branch attraction would cause. In reviewing their own work the authors say “Large-scale search of anciently duplicated genes did not bring any consensus”[233]. The reason paralogs rooting is not self consistent is because there are many reasons why a sequence based tree will not reflect the evolution of cells. We hope the moral of this work is that structure is an untapped resource in rooting the tree.

The authors correctly mention that the quality of alignment is the major limitation to indel-based rooting approaches, and that the addition of structural information greatly improves the reliability of the method. However, even with the additional confidence that observed indels are real (and not alignment artifacts), their utility as polarizing characters still varies greatly based on the length and context of the indel. For example, the support for excluding the root from the archaeal/eukaryal grouping [11] shown in Figure 1 consists of a large region of protein sequence (14 AA) within EF-2 corresponding to a discrete helical surface structure within the protein. Compared to the homologous regions of EF-G and EF-Tu, it is clear that this is a derived, polarizing character. Indel evidence used for excluding the root from within bacterial groups seems to be far weaker in both hypotheses being compared, as small indels are more likely to be the result of convergence.

Author's response:

We completely agree that larger indels make better phylogenetic markers and that our evidence excluding the root from Archaea and Eukaryotes is stronger than its placement within Bacteria. However, some still claim the root is within either of these groups [26, 234] or between them (see Eugene Koonin's review below). In our opinion reaching a consensus on a root within the Bacteria would be a big step forward.

While the authors clearly show that the evidence provided by Lake et al. is insufficient to exclude the root from near the Chloroflexi based on their improved analyses, there is a logical fallacy at work in their conclusions. “We show that all four of these arguments have flaws and that is evidence that Cavalier-Smith’s rooting is probably correct” is clearly a false dichotomy, as lack of evidence for the former does not correspond to any increase of evidence for the latter. The authors should have seriously considered (or at least discussed) the third possibility that there is simply not enough indel-based information for a reliable rooting of the tree of life (except for its exclusion from the archaeal/eukaryal branch).

Author's response:

We disagree. The very nature of exclusive rooting is to prove a branch of the tree has a derived trait. Therefore, it is impossible to ever truly prove a rooting using this method. If one cannot exclude a root using more and more data our confidence in that rooting should increase. Every argument that could (or has been claimed) to exclude the root from the Chloroflexi but does not can be taken as evidence that rooting is correct. To us this is the first real independent test of Cavalier-Smith’s hypothesis. We never

claim that indel-based data are enough to root the tree. In fact we claim the opposite, since our PyrD quaternary structure argument goes against what some of the indel data implies. Polarized indel arguments will be limited in nature since they require universal paralogs, but there might be many polarizable transition in quaternary structure. The position that there is not enough polarizing data to root the tree was certainly defensible before this work because there were numerous disagreements in the data. We have resolved all of these, so for now it seems there is enough polarizing data to reliably root the tree. The 4 polarized arguments presented here are not enough to root the tree reliably on their own. Our point is that independent lines of reasoning are beginning to converge on a single rooting. Its time to test (attack) Cavalier-Smith's rooting using every piece of reliable data out there until it breaks. Then we think it would be worth discussing the possibility that there is not enough polarizable data to root the tree of life.

An interesting and novel rooting approach is also presented, based on the quaternary structure of the PyrD enzyme. Mapping the phylogenetic distribution of each type of enzyme (monomer, homodimer, heterotetramer) using the NCBI Protein Clusters Database, the authors show that the most parsimonious path for the evolution of these types (i.e., one of increasing subunit complexity) effectively excludes the root from being within the archaea, the firmicutes, or their most recent common ancestor, and thus requires it to be within a bacterial non-firmicute group (which contain the monomeric type, PyrD2). However, a preliminary phylogenetic investigation shows that PyrD homologs which are present as a homodimer (PyrD1A) are clearly a derived group within

the heteratetramer (Pyr1B) set most closely related to the bacillus group within the firmicutes. Therefore, the presented parsimonious model of subunit evolution cannot be made to agree with any possible rooting of the tree. A more extensive phylogenetic analysis involving all PyrD homologs is clearly needed before this character can be used to exclude the root from any part of the tree.

Author's response:

It is hard for us rebut a tree that we have not seen but here goes. Our hypothesis on quaternary structure evolution is not contradicted by our own analysis of the PyrD tree, using HisA as outgroup (data not shown). We do not see evidence that PyrD 1A is clearly derived from PyrD 1B. Even if we did, we would still argue that result could be an artifact. All 3 PyrD families are going to be under very different selective pressures since they each have different protein-protein interaction sites. This is a case where it would be completely possible for PyrD 1B to evolve rapidly out of PyrD 1A but look exactly as you described in the sequence tree. It is possible that PyrD 1A is derived from PyrD 1B, but until one provides an outgroup that explains the origin of the heterotetramer we are not going to find a sequence argument very convincing here.

Reviewer's report 2

Purificación López-García, Unité d'Ecologie, Systématique et Evolution, UMR CNRS 8,079, Université Paris-Sud, bâtiment 360, 91405 Orsay Cedex, France

This work is a reanalysis of several conserved paralogous genes used by Lake et al. to place the root of the tree of life based on indel sharing that were in apparent disagreement with a rooting proposed by Cavalier-Smith between the Chloroflexi (Eobacteria) and the rest of bacterial + archaeal groups. A key factor to attempt such analyses is the quality of the alignment. Structural alignment data shows that the analysis those paralogous genes based on indels yields results that would be compatible with the rooting proposed by Cavalier-Smith.

This work is interesting as it shows some of the drawbacks that can be linked to this kind of indel analysis, particularly in what concerns ambiguous alignment and convergence. Yet, rooting the tree of life is a difficult task and there is possibly too little information left in ancient duplicated genes that allows answering that question with meaningful statistical support. From the four sets of genes studied here on a structural basis, two of them are discarded as unable to provide polarizing indel information: S12/RpoC because they may not be homologous, and Hsp70/MreB, which are inconclusive. The alignment of EF-G/EF-Tu proteins is ambiguous at the indel region used for polarization of character states, although quaternary structure-based alignment was not possible. Only two couples of paralogs yield what might be useful information according to Valas and Bourne, the HisA/HisF and PyrD/HemE.

Author's response:

We do not consider the PyrD/HemE indel argument to be robust, but we reach the same conclusion using our polarization of quaternary structure. We also think the insert

in GyrA robustly excludes the root from within the Actinobacteria, but there was no need to reanalyze that result here as with the other indels. We also accept the conclusion the EF-G/EF-Tu indel despite the problems with aligning these sequences.

Though of interest, one can wonder whether this information is enough to confidently exclude the root of the tree of life from everything outside the Chloroflexi. In addition of the alignment quality, there are other factors that can be of crucial importance here. One is convergence, as rightly pointed out by the authors, and the other is horizontal gene transfer (HGT). HGT is but very briefly mentioned here to exclude the possibility that PyrD has been transferred between archaea and bacteria recently. However, other protein genes, notably gyrase and Hsp70 genes are very likely cases of HGT from bacteria to archaea. Despite so, they are included in this kind of indel work. Careful phylogenetic analyses should be done for all the genes that are included in these attempts so that only vertically inherited genes are used. The possibility of important HGT levels between Firmicutes and/or Actinobacteria and archaea is not discarded. Finally, other problems such as hidden paralogies or even selective paralog loss cannot be excluded.

Author's response:

It is true that everyone of these factors can mislead an indel analysis. We are mainly trying to improve the alignment step for indel sets that appear to meet these criteria to a reasonable level. Horizontal transfer of Hsp70 is irrelevant since this indel is not polarizable using an outgroup regardless of its distribution within species. Gyrase has probably been horizontally transferred, but it is clear from looking at sequence alignments that the Actinobacterial insert has not been transferred in a way that could

confound the results. Horizontal transfers do not necessarily destroy an indel argument, we just need to be careful about whether they actually affect the conclusion or not. If a derived trait is horizontally transferred to ancient group the early branching members of that group will be unaffected. Unless there is a selective sweep or poor sampling we should be able to identify these cases. One also needs to keep in mind that horizontal transfers are defined by our assumption of what the correct tree is. This paper challenges the traditional rooting between the Archaea and Bacteria so many horizontal transfers presented in the literature may actually be better explained by vertical inheritance in this model. Loss is trickier since inferring where a loss as opposed to a gain occurs requires independent lines of polarization. We feel we are being very conservative in only really accepting 3 indel polarizations that seem robust to these sources of noise.

Another comment is that the fact that these results are compatible with a rooting between Chloroflexi/Eobacteria and anything else does not necessarily imply that the root is actually close to these organisms. First, more than 50% of the bacterial phylum-level groups correspond to candidate divisions without cultivated members. A similar trend occurs in archaea. Therefore, there is a considerable ignorance about the indel distribution in at least half of the bacterial diversity. Also, though the most parsimonious scenarios appear more likely to us, this is not proof that evolution proceeds that way. A cautionary vision should be held in this regard.

Author's response:

We agree that this work is certainly not the last word on the rooting of the tree of life. As discussed above we feel this work is a good test of independent lines of reasoning

to Cavalier-Smith's Eobacterial rooting. For now his hypothesis seems to be the one to beat. We are open to new data may change the picture, but the point of this paper is that the present data does not contradict itself in the manner it appears to in the current literature. There is an important difference between parsimony and polarization. To us parsimony can be used to analyze events where gain and loss have nearly equal probabilities, while polarizations imply that one direction would evolve more easily than the other. Consider the example of the proteasome discussed in detail in [208]. A parsimony argument would be that the 20s proteasome is the result of a duplication so a non duplicated structure must precede it. The polarization argument involves considering the structure and function of proteasomes as well as the fitness of the intermediates to argue that evolution towards the 20s proteasome is much more plausible than the reverse direction. There are probably many cases where evolution has not been parsimonious, and we do not think parsimony is a safe or productive assumption. However, there appears to be many polarizable transitions and hopefully there are many more waiting to be discovered. If the new data continues to support the Eobacterial rooting then our confidence in it will increase.

Reviewer's report 3

Eugene V. Koonin, National Center for Biotechnology Information, National Library of Medicine National Institutes of Health, Bethesda, Maryland 20894, USA

This manuscript addresses very important issues of the position of the primary divide among eukaryotes and, by implication, the relationships between archaea and

bacteria, and the nature of the LUCA. The discussion is presented in somewhat obsolete terms of the “root of the tree of life”. As there is no such thing as a single “tree of life”, speaking of a single root is somewhat misleading but the central question is nonetheless meaningful and crucial.

Author's response:

We do not believe that the issue of the nature of the tree of life has been settled. It is certainly true that recent work has shown that there is too little signal to resolve the tree of life using sequence alone [5, 57], and that many genes have histories distinct from the species in which they reside. However, it is limiting to assume that the only data useful for building the tree of life is sequence data, and that additional data will be unworthy in this pursuit. We are aware that tree representations have many shortcomings, but we still believe it is the single best metaphor to describe the major events in evolution. We are working with a novel data source to argue the tree of life is realer than studies in genomics have led us to believe (in preparation). The phrase "root of the tree of life" may not be as accurate as "the first polarizable transition between extant groups", but it certainly rolls off the tongue better.

Without going into the minute details of the analysis of indels in specific protein families, I will state my firm view of this issue. The nature of the primary divide in prokaryotes – and actually among all cellular life forms is clear, and it is between archaea and bacteria. This view is supported by the fundamental differences between archaeal and bacterial systems of DNA replication, core transcription, translation, and membrane biogenesis – essentially, all central cellular systems (not just the replication system as

noted in the present paper). I believe these differences are sufficient to close the “root debate” (regardless of the appropriateness or lack thereof of the very notion of a root in this context) and to base analyses and discussions aimed at the elucidation of the nature of LUCA on that foundation.

Author's response:

One cannot disagree with the fact there are vast differences between the Archaea and Bacteria, and we are well aware of the details of that argument. We believe that none of these differences are as great they appear at first glance, and we are working on a scenario to detail the transition between the Bacterial and Archaeal DNA replication system. It certainly makes sense the greatest splits in the tree would be the most ancient. However, we are proposing that the alternative hypothesis that a unique event in evolution occurred between the Bacteria and Archaea must be taken seriously. A rooting between the Archaea and Bacteria would imply the first Bacteria were Gram-positive. As Cavalier-Smith pointed out no one has adequately described how the transition from a Gram-positive to a Gram-negative bacterium could occur [208]. So the rooting you assume to be true has its own problems too. Cavalier-Smith has already proposed a detailed scenario that covers many of these transitions you mention [62] and Lake et al. have recently discussed the issue of membrane biogenesis [35]. That said there are aspects of both of these hypotheses on the origin of the Archaea that we feel are incomplete. At this point it seems reasonable to keep an open mind about the root, but this work argues against all evidence that has ever been used to support a Gram-positive Bacterial rooting. Until there is a scenario that describes all the major transitions starting from the root that has been robustly tested by multiple lines of evidence we

suggest this debate is not over. We think the data presented here offers a compelling reason to continue to look at the details of the Gram-negative rooting.

Indel analysis is a legitimate method of phylogenetic inference but is seriously hampered not only by horizontal gene transfer but, more importantly, by the possibility of homoplasy, that is, independent insertions in the same region of homologous proteins. The contradicting macro-phylogenetic inferences made by Lake, Cavalier-Smith, Gupta and others using this approach serve to illustrate the point. The use of structures to corroborate indels is, of course, a good idea in principle but changes very little substantially. Somewhat parenthetically, I find it strange that the alignments of this paper, aimed primarily at clarification of the relationship between archaea and bacteria, include only bacterial and in some cases eukaryotic proteins.

Author's response:

We disagree that this changes nothing substantial. Every single disagreement between these different groups is, in our view, caused by bad alignments. It appears to us the only substantial difference between Gupta and Cavalier-Smith's phylogenies were based off of Gupta's polarization of the Mreb/Hsp70 indel, which we have demonstrated is inconclusive. This work has significantly improved the quality of the alignments and resolved all contradictions within these data sets. We hope this begins to forge a consensus between them, and stimulates brain storming on how the systems you mention above could undergo such dramatic changes. In our opinion, one must deal with the differences between the macro-phylogenies one detail at a time instead of assuming they are a quixotic pursuit. If these macro-phylogenies are truly incompatible it should not be

possible to get them to converge on a single tree as we believe we have done here. Again, it is true that transfers and convergent evolution complicate indel analysis, but they do not invalidate it a priori. Multiple transitions are necessary to make a tree robust to these problems. We have been very stringent in accepting an indel as informative, so we are confident our conclusions are not a result of these factors. Our analysis of PyrD demonstrates that our conclusion is not the result of horizontal transfer, and the clustering of PyrD 1B shows it is not the result of convergent evolution. The distribution of this derived structure across a Gram-positive group and the Archaea must be considered seriously as evidence that the root might not be between the Archaea and Bacteria. We agree that it would be ideal to use more Archaeal structures in our alignments. However, we are limited by currently available crystal structures. At the present time it appears very few people besides us are really considering structure to be a useful tool in studying the major events in evolution. Therefore there is no directed effort aimed at widely sampling the same structure across the tree of life. We believe this paper shows that structure has a role to play in every aspect of studying the events that separate the major taxa. The landscape of the continent of genomics is being filled in rapidly, but the continent of protein structure, especially quaternary structure, lags far behind. We are optimistic that structure may still contain enough signal to resolve a single backbone to the tree of life where sequence has failed.

Acknowledgements

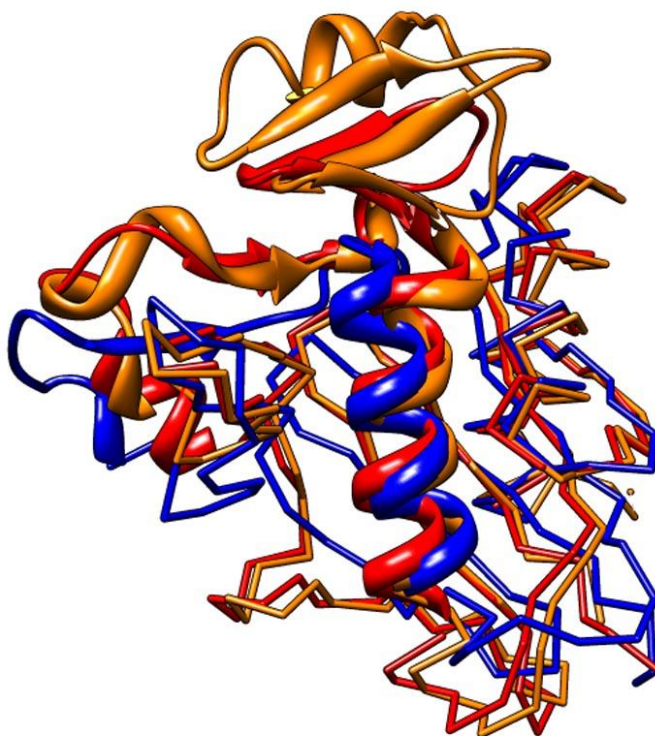
We would like to thank Russell Doolittle, William Loomis, and the entire Bourne Laboratory for useful discussions.

<i>Streptomyces_coelicolor_A3</i>	VASGGVSSLDLRAIAGLVPAGVEGAIVGKALY
<i>Synechocystis_sp._PCC_6803</i>	IASGGVSQVEDLLNLLSLESLGVNGVIIGKALY
<i>Bacillus_clausii_KSM-K16</i>	IASGGVSSIADLAQLRAA - - - SLAGAIVGKALY
<i>Thermus_thermophilus_HB27</i>	IVGGG IASPEDLHLLQAL - - - GVDGALIGKALY
<i>Thermus_thermophilus_HB8</i>	IVGGG IASPEDLHLLQAL - - - GVDGALVGKALY
<i>Deinococcus_geothermalis_DSM_11300</i>	IVGGGVATLDDVRLLEAE - - - GIQGAIVGRAIY
<i>Deinococcus_radiodurans_R1</i>	IVGGGVADTNDVRLLEAE - - - GIEGAIVGRAIY
<i>Herpetosiphon_aurantiacus_ATCC_23779</i>	IASGG IARVEHLARLAE - - - GISGAIIGTALY
<i>Chloroflexus_aurantiacus_J-10-fl</i>	IASGG IASIDHVRQLAQL - - - GVEGAIIIGKALY
<i>Roseiflexus_castenholzii_DSM_13941</i>	IASGG ISTLDHLRRLADV - - - GVEGAIVGRALY
<i>Roseiflexus_sp._RS-1</i>	IASGG ISTIDHLRRLAEV - - - GVEGAIVGRALY
<i>Dehalococcoides_ethenogenes_195</i>	IASGGVSSLSHLRLLKDI - - - GAEGAIVGKAIY
<i>Dehalococcoides_sp._BAV1</i>	IASGGVSSLSHLRLLKDI - - - GAEGAIVGKAIY
<i>Dehalococcoides_sp._CBDB1</i>	IASGGVSSLSHLRLLKDI - - - GAEGAIVGKAIY

Figure 6-2. HisA does not exclude the root from the Eobacteria.

A MUSCLE based alignment of all the HisA sequences in Eobacteria. Representatives from Actinobacteria (*Streptomyces coelicolor* A3) and other Gram-negatives (*Synechocystis* sp. PCC 6803) are included to show the indel. All the Eobacterial sequences share the relative deletion with the Firmicutes (*Bacillus clausii*).

A)



B)

```

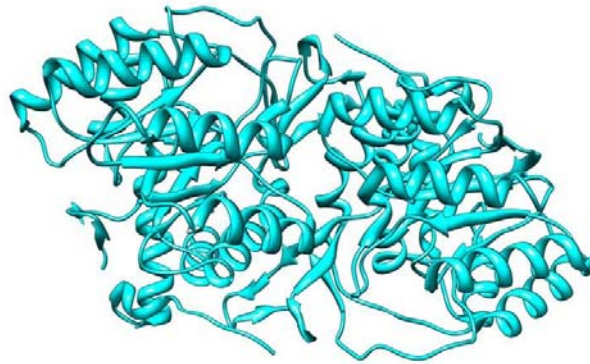
Hsp70+ AITN...PNTIISIKRHMGT.....DYKVEIEGKQYTPQEISAILQYLKSYA
Hsp70- AVTN...PQNTLFAIKRLIGRRFQDEEVQRDVSIMPFKIIAADNGDAWVEVKGQKMAPPQISAEVLKKMKKTA
Mreb   IGKTPATIKAIRPMRDGVI.....ADYTVALVMLRYFINKA

```

Figure 6-3. Structural Alignment of MreB/Hsp70

A) A multiple structural alignment of the MreB/Hsp70 C-terminal actin-like ATPase domain. The region around the indel is highlighted as a ribbon diagram. The backbone of the rest of the domain demonstrates high conservation between these structures. The blue chain is MreB from *Thermotoga Maritima* (1JCF:A 51-86 drawn as ribbon). The red chain is Hsp70 from the Gram-positive bacterium *Geobacillus Kaustophilus* (2V7Y:A 57-101 drawn as ribbon). The orange chain is Hsp70 from the Gram-negative bacterium *Escherichia coli* (1DKG:D 56-125 drawn as ribbon). B) The sequences corresponding to the highlighted portion of the structure alignment in A.

A)



B)

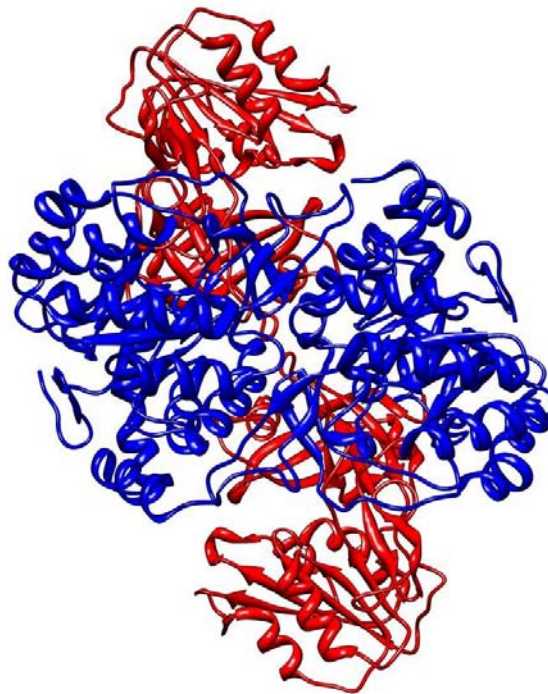


Figure 6-4. Quaternary Structure of PyrD

A) PyrD 1A from *Lactococcus lactis* is a homodimer (1JUB colored cyan). B) PyrD from *L. lactis* 1B is a heterotetramer. The homodimer interface at the center of PyrD 1B (1EP3 colored blue) is similar to the interface in PyrD 1A. PyrD 1B has an additional subunit PyrK (colored red). This implies that PyrD 1B is derived from PyrD 1A.

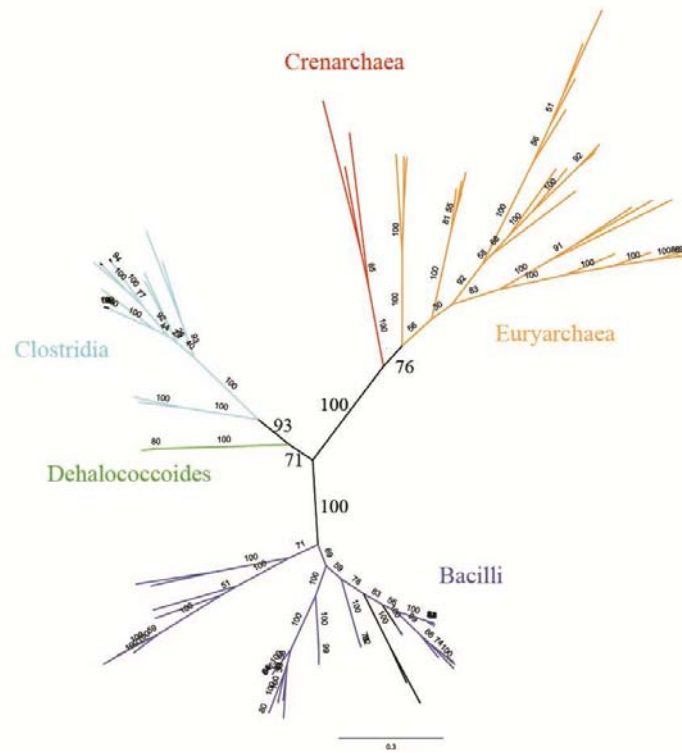


Figure 6-5. Maximum likelihood tree of PyrD 1B

Each of the major groups is separated by significant bootstrap values which indicates the distribution of PyrD 1B cannot be due to recent horizontal transfer.

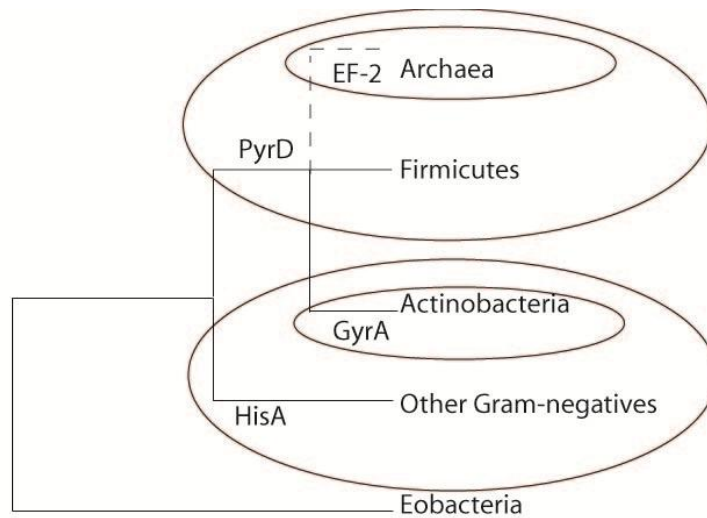


Figure 6-6. Summary of data.

Each circle corresponds to an argument presented above that excludes the root of the tree of life from a particular branch. The Archaea are placed with the Gram-positives, but drawn with a dashed line because we do not wish to argue which Gram-positive group was their ancestor at this time.

Tables

Table 6-1. Length between motifs in elongation factors.

A summary of the length of the region between the well conserved RG(IV)T/PGH motifs. This data implies the Archaeal ancestor of EF-2 had a 4 residue derived insertion regardless of which alignment is used.

	Total sequences	Do not have perfect match to motifs	% of sequences that have motifs	Length of region between motifs	Sequen ces with that length	% out of sequences with motifs
Ef-Tu	1013	12	98.82%	20	1001	100.00%
EF-1	53	3	94.34%	20	50	100.00%
EF-G	816	160	80.39%	20	367	55.95%
				24	13	1.98%
				25	22	3.35%
				26	2	0.30%
				27	211	32.16%
				28	12	1.83%
				29	11	1.68%
				31	9	1.37%
				33	4	0.61%
				35	3	0.46%
				40	1	0.15%
EF-2	52	9	82.69%	24	43	100.00%

Chapter 7: The origin of a derived superkingdom: how a gram-positive bacterium crossed the desert to become an archaeon

Abstract

Introduction

We have previously presented three arguments that support placing the root of the tree of life in the Bacteria. This data have been dismissed because those who support the canonical rooting between the prokaryotic superkingdoms cannot imagine how the vast divide between the prokaryotic superkingdoms could be crossed.

Results

We review the evidence the Archaea are derived, as well their biggest differences with the Bacteria. Novel data that demonstrate this is not an insurmountable gap. We also analyze whether the Archaea are holophyletic or paraphyletic which is essential to understanding their origin. We argue for a Firmicute ancestry of the Archaea over proposals for actinobacterial ancestry.

Conclusion

We present a novel hypothesis that a combination of antibiotic warfare and a viral endosymbiosis in the Bacilli led to dramatic changes in a bacterium that resulted in the birth of the Archaea and Eukaryotes.

Introduction

Archaea were first discovered because of a distinct sequence signature in their ribosomal RNA [15]. This remains one of the strongest signals found anywhere in the phylogenetic tree. It was truly a revolution in thought when the world realized there were two distinct types of prokaryotes. Besides placement on sequence trees, there are three major areas where the Archaea and Bacteria differ greatly. The structures of archaeal and bacterial ribosomes each have many unique proteins [235]. Archaeal membranes are composed of glycerol-ether lipids, while bacterial membranes are composed of glycerol-ester lipids [236]. The glycerols have different stereochemistries between the superkingdoms as well. The DNA replication machinery of these two superkingdoms is also very different; many key proteins have a superkingdom specific distribution [230].

These vast differences as well as the rRNA tree has convinced most that the root of the tree of life must be between the prokaryotic superkingdoms. The proposal that Archaea were a different kingdom was originally considered ridiculous because no one could imagine two distinct groups of prokaryotes [16]. In 30 years we went from the prevailing opinion that the Archaea were similar enough to Bacteria to be just prokaryotes, to the view they are so different they must each be primordial lineages.

Locating the root of the tree of life is a prerequisite for understanding the origin and evolution of life. There are many examples of conclusions that become radically different if one assumes a different rooting of the tree. For example the proposal that LUCA was acellular relies on a rooting between the Archaea and Bacteria [23]. Each of the estimates for divergence times of the prokaryotic taxa [237] would change drastically if the Archaea are not the same age as LUCA. There are many results that rely on a reference tree which would be interpreted differently under a different rooting.

However, there are several proposals that root the tree of life within the Bacteria and place the Archaea as a taxon derived from the Gram-positive bacteria [27, 208, 215]. These hypotheses are often dismissed for two reasons: 1) they do not agree on a single rooting 2) the immense gap between the Archaea and Bacteria in sequence trees and the systems mentioned above. We have addressed the differences between these alternative rooting in [38] and concluded that it is possible for them to converge on a single root in the Gram-negative bacteria. The point of this work is to address the objection to rooting the Archaea within the Gram-positives.

This work is a synthesis of many creative ideas that came before us; as a result, much of what we say here has been said in some form before by others. However, the arrangement of the pieces is novel and sheds light on the strengths and weaknesses of the various rootings of the tree of life. First we must discuss the ideas in their original form and consider what we see as the strengths and weaknesses of each. We take the stance that closing the debate prematurely deprives one of the ability to see the many strengths of each of these hypotheses and the large common ground between them. We

offer novel data that help refine some of these ideas and show the potential for testing them further.

Radhey Gupta has created a detailed tree of life using rarely fixed indels (insertion-deletions) in prokaryotic groups [215]. He concludes that the root of the tree of life is within the Gram-positive bacteria, and he places the Archaea as derived from the Firmicutes. The major driving force in his scenario is antibiotic warfare. He argues the differences between the Archaea and Bacteria coincide with the targets of many antibiotics produced by Gram-positive bacteria. He lists antibiotics and their general targets that differ between the prokaryotic superkingdoms. We provide a more detailed list generated from recent work in the field that supports this idea. The strength of his phylogeny rests on the fact that many of the branch orders are supported by several independent indels.

However, there are several points that concern us about Gupta's hypothesis. First of all we disagree with his polarization of Hsp70 which is used to justify the root of the tree of life [38]. But the focus of the present paper is the origin of the Archaea, so that debate is probably better left in our other work. The transition between the Gram-positives and Archaea must have been a drastic event that must be confronted in any hypothesis that roots the tree of life in the Bacteria. Antibiotic warfare is a powerful evolutionary force, but in Gupta's hypothesis there seems to be a special battle that resulted in the Archaea. But he does not explain why antibiotic warfare only gave rise to one other prokaryotic superkingdom. Should one expect there to be several different modified ribosomes in response to antibiotic pressure? We will invoke antibiotic warfare as a major driver in the origin of the Archaea, but we feel our scenario better sets the

stage for why this was a unique event. Antibiotic warfare on its own is not enough to account for the vast differences between the prokaryotic superkingdoms, but it certainly was important.

James Lake has also constructed a detailed tree of life using indels [27]. His group has focused more on indels that can be polarized using paralogous outgroups. The strength of this method is it provides evidence for derived and ancestral groups which we feel is essential for understanding evolutionary histories. The polarizations are largely independent. This allows one to refine the tree because a flawed polarization will only affect one part of the tree. Like Gupta, his group roots the Archaea within the Firmicutes and provides several independent reasons why this makes sense [34]. Lake has also proposed that the Eukaryotes had a crenarchaeal (Eocyte) origin based on a shared indel in EF-1 and similarities in their ribosomal structure [20, 238]. We find arguments like this appealing as it is a synthesis of both sequence and structural data. We discuss the strengths and weaknesses of that particular hypothesis at length below.

The weakness of that method in general is the difficulty in properly aligning the paralogs as we have argued in [38]. However, the polarizations are mostly independent, so changing a polarization does not invalidate the whole tree, it just refines it. We argue that the refined version of Lake's tree is completely consistent with Cavalier-Smith's [38]. There are very few universal paralogs, so this method certainly needs to be supplemented with other data sources.

Cavalier-Smith has discussed the relationship and origin of the superkingdoms at length [62, 208, 239]. The major difference between his hypothesis and these other two is the placement of the root in the Gram-negative bacteria. He also roots the Archaea within

or next to the Actinobacteria. He constructed this tree by polarizing multiple types of data including indels, membrane structure, and quaternary structure. Again, if any one of these polarizations is brought into question it does not weaken the rest of them. Cavalier-Smith has included discussions of the prokaryotic fossil record that appear to be lacking altogether from these other hypotheses. That analysis concludes there is no fossil evidence that argues Archaea are older than Eukaryotes, despite much evidence that the Bacteria are older than Eukaryotes. His synthesis of different data sources and attention to detail is second to none.

That said there are several aspects of Cavalier-Smith's tree that still do not sit well with us. His hypothesis relies on the assumption that the Archaea are holophyletic (the Eukaryotes are their sisters, not their descendents). He provides some justification for this, but we will discuss below why this is not a completely safe assumption to make at this time. His rooting of the neomura (his term for Archaea, Eukaryotes and their last common ancestor (LAECA)) is in the Actinobacteria. He cites traits shared between the Eukaryotes and Actinobacteria to support this hypothesis, but they are only relevant if the Archaea are holophyletic. We provide an alternative interpretation of this distribution, by invoking an actinobacterial endosymbiont near the root of the eukaryotes. He argues thermophily was the major force that lead to the neomuran revolution. We feel this argument falls short for the same reason as Gupta's, it just does not seem to be a unique enough selective pressure to create a novel superkingdom. Cavalier-Smith prefers the labels Archaeobacteria and Eubacteria because he feels the labels Archaea and Bacteria over emphasize the difference between these superkingdoms. We disagree, these

superkingdoms are fundamentally different. Despite that, we still believe the Archaea evolved from within the Bacteria.

None of these scenarios adequately addresses the origin of the DNA replication machinery shared between the Archaea and Eukaryotes. Therefore we invoke the ideas of Patrick Forterre, who has proposed that cells received the ability to replicate DNA from viruses. He proposes this occurred three times; each event resulting in the birth of a superkingdom [227, 240]. The amazing variation in DNA replication machinery found throughout the virosphere supports this idea. All extant cells use double stranded DNA, but viruses can have many other forms of genetic material (reviewed in [241]). The plasticity of replication in the virus certainly could lead to innovations of great importance in the cellular world.

There are two weaknesses of this view of our opinion. First, it is DNA centric so it necessarily neglects the many other important differences between the superkingdoms. Second, it is firmly placed within the framework of the classical rRNA tree. Forterre even assumes the Eukaryotes are a primordial lineage, as a consequence of taking the sequence tree too literally. We will demonstrate that this view is also highly informative if the Archaea are derived from the Bacteria. It has also been noted that other extra chromosomal elements could play key roles in the evolution of the different DNA replication systems [242], but that discussion was also firmly grounded in the canonical rooting.

It seems it would be quite an uphill battle to argue the Archaea are a derived superkingdom. One would have to provide compelling evidence the Archaea are derived, so we will review our data that supports that view. Any hypothesis that addresses how a

bacterium could become an archaeon would have to explain dramatic changes in membranes, DNA replication, and ribosomes. We will demonstrate that the ribosome can have great plasticity under certain circumstances. It has been previously argued that the Firmicutes have many of the enzymes needed to make Archaeal membranes [243]. We will invoke a viral endosymbiosis to explain the differences in DNA replication. For the reasons discussed below the hypothesis must work if the Archaea are paraphyletic or holophyletic. Finally, it must also address the rarity of the event that lead to this revolution. If a hypothesis could do all of these things, it would make a compelling argument for the origin of the Archaea. Without any one of these pieces this hypothesis would be incomplete.

Three robust reasons the Archaea are derived

Several large indels are shared between the Archaea and Gram-positive bacteria, as well as the fact that both groups only have one membrane [215]. If there is a direct relationship between the Gram-positives and the Archaea the root is either between them or one is derived from the other. Every piece of evidence that is polarizable implies the Archaea are derived from the Bacteria. Arguments that the Archaea and Bacteria are so different that they both evolved from LUCA sidestep directionality all together. The only recent work that explicitly roots the tree in the Archaea is [26]. Many of their arguments are based on assumptions about the nature of LUCA and assumptions of what a primitive state would look like. None of their arguments are true polarizations. To the best of our

knowledge there is no single polarized argument for an archaeal rooting that is on par with the three we shall discuss that place the Archaea as derived.

The first of these is the proteasome. Proteasomes are self compartmentalized atp-dependent proteases that are found in varying degrees of complexity across the tree of life. All Archaea contain a 20s proteasome which is composed of 28 subunits and is encoded by at least two genes that are clearly homologs. Therefore the 20s proteasome must be the result of a duplication. Cavalier-Smith has argued that the simpler bacterial homolog HslV (heat shock locus v) could be duplicated to generate a 20s proteasome [62, 208]. Loss of a subunit in the 20s proteasome would result in an open proteasome with no ATPase. Such a protein would lose the essential function of controlled degradation found in proteasomes, and does not make sense as an intermediate. It is more likely that the 20s proteasome is derived from a simpler structure. Cavalier-Smith excludes the root from the Archaea because all Archaea contain a clearly derived protein structure.

However there is a counter argument to that proposal that goes as follows; LUCA had HslV and LACA (last archaeal common ancestor) is the point in the tree where HslV evolves into the 20s proteasome (Figure 7-1A). This would still exclude the root from the crown Archaea, but it still allows for the possibility that the root is between the extinct stems of Archaea and Bacteria. Excluding the root from the Archaea will never be enough because one can always invoke stem lineages that show up before the derived trait. This would imply the 20s proteasome present in Actinobacteria is probably the result of a horizontal transfer from the Archaea. However, we have observed that the two proteasome genes are often in the same operon in Actinobacteria, but rarely together in Archaea. This weakly polarizes the direction of the horizontal transfer to the Archaea.

However, there is stronger evidence that narrows the root to within the Bacteria. Our own work argues that the Anbu proteasome (or peptidase according to [244]) is more likely than HslV to be the 20s proteasome's direct ancestor based on both sequence data and structure predictions [71]. This is much stronger than Cavalier-Smith's conclusion because HslV is widespread in the Gram-positives but Anbu appears to be missing in them altogether (Figure 7-1B). If the divide between Archaea and Bacteria is the earliest split in the tree, and our hypothesis on proteasome evolution is correct, then LUCA must have had Anbu. This would mean that all extant Gram-positives need to have lost Anbu while the Gram-negatives (that must be derived from Gram-positives in this scenario) somehow retained Anbu. One would have to invoke a selective sweep of the 20s proteasome in Archaea, and of HslV in the Gram-positives. It is plausible that the 20s proteasome outcompeted Anbu or HslV since they are almost never found in the same genome. However, Anbu and HslV are found together in many genomes, which is evidence neither totally displaces the other in terms of function. Our arguments about Anbu are based on structure predictions, but a crystal structure could potentially verify them. If we are correct it may be the smoking gun for a Gram-negative rooting, but even without that there is ample evidence to support Cavalier-Smith's position. Even if HslV is the direct ancestor of the 20s proteasome the root can still be excluded from all extant archaeal lineages.

The recent analysis of proteins found in Anbu's operon [244] presented evidence we are wrong in labeling Anbu a proteasome because it lacks an associated ATP-dependent protein required for unfolding substrates. HslV and the 20s proteasome clearly have associated ATPases dedicated to unfolding substrates. Therefore the transition to

both of them is easier from Anbu as no ATPase would have to be lost. The origin of HslV and the 20s proteasome would both involve the recruitment of distinct ATPases subunits. Therefore we think this new work strengthens our hypothesis that Anbu is ancestral to the 20s proteasome because no intermediate would ever lose the regulatory ATPase. If our hypothesis is correct, proteasomes would be polyphyletic if they are defined by the presence of the ATPase subunit as suggested in [244].

The indel in EF-2 shared between the Archaea and Eukaryotes has been polarized using EF-Tu as an outgroup [35]. Our alignment free analysis of this indel agrees with the authors' conclusions despite there being a sequence artifact in their original alignment [38]. This polarization robustly excludes the root from within Archaea, but does not narrow it to within the Bacteria.

In that analysis we also present a novel argument for polarizing the Archaea. The quaternary structure of PyrD 1B is a homotetramer across the Firmicutes and Archaea. We argue that the homotetramer is probably derived from the homodimer PyrD 1A based on a conserved interface. That is probably a derived structure relative to the monomeric and homodimeric versions found in the Gram-negatives and Actinobacteria. PyrD 1B is found across a Gram-positive group and the Archaea, so it would have to be present in their last common ancestor, which is LUCA under the canonical rooting. This could be explained by the presence of both PyrD 1A and 1B in LUCA. But that scenario would require PyrD 1A to be lost in every Archaea and some Firmicutes, and reversion to the monomeric form, PyrD, across the Gram-negatives and Actinobacteria. PyrD 1B is probably derived, so the Archaea, Firmicutes, and their last common ancestor are also derived.

The polarization of the indel in EF-2 excludes the root from the extant Archaea. Our novel polarizations of Anbu and PyrD argue the root is within the Bacteria. If these arguments only excluded the root from all extant Archaea one is left wondering why all the Archaea that are not clearly derived went extinct. The combination of all three of them together strongly supports the bacterial rooting of the tree. If the Archaea are derived, there must be some way of reconciling the major differences between them and the Bacteria.

Ribosomal revolutions are historical fact, not science fiction

The Archaea cluster separately on phylogenetic trees based on ribosomal RNA [15]. This split has remained robust in many trees derived since then. We will propose and discuss three scenarios that can explain this. The first scenario is that the ribosomal sequences are pretty good molecular clocks. The great splits seen in the tree reflects the fact that this is the most ancient divide in cellular life. This is obviously in accordance with the canonical rooting.

The second scenario does not contradict that rooting either. It goes as follows: The ribosome in LUCA was incomplete. It did not have all the proteins found in extant Archaea or Bacteria, only the core that is universal between them. The addition of proteins after the split of the superkingdoms would start a quantum evolution event. Some sites would be free to mutate from increased stability, while others would be under stricter pressure to keep a protein in the correct place. The rate of mutation at different

sites on the ribosome could vary wildly and exaggerate the true distance between the superkingdoms, even if they do represent a very ancient split.

The third scenario, which we champion here, is that the Bacterial ribosome evolved into an Archaeal one. Again this would be a quantum evolutionary event and sequences of both rRNA and ribosomal proteins would evolve rapidly. The point we are trying to make is that these three scenarios would result in exactly the same sequence tree. So we must look towards independent lines of reasoning to determine what scenario can best be described as the cause of the tree.

We can exclude the first scenario by comparing the structure of the ribosome in Archaea and Bacteria. In the 50s subunit there are six ribosomal proteins that are in the same position on the rRNA, but have non homologous structures in Archaea and Bacteria [245, 246]. These must have changed in at least one lineage since LUCA regardless of LUCA's nature. Therefore we should expect that the distance between the Archaea and Bacteria would be exaggerated due to compensatory mutations in the rRNA and ribosomal proteins.

It is certainly reasonable to object to the third scenario because it seems so implausible that a ribosome would ever change so much when it has stayed so well conserved within a superkingdom. However, there are two examples where we know the ribosome structure has changed greatly. Mitochondrial ribosomes have changed dramatically from their bacterial ancestors. They have lost about half their rRNA and replaced it with additional proteins [247]. The eukaryotic ribosome evolved from an archaeal one (or technically some sort of proto-archaeal ribosome if the archaea are holophyletic). There are eleven ribosomal proteins found only in the Eukaryotes, nine of

which are conserved across the superkingdom [235]. And there is good separation on rRNA trees between the Eukaryotes and Archaea. In the two cases where the ribosome structure has changed we know it changed from another fully functional ribosome. Why would it be out of the question for it to happen between the Archaea and Bacteria? There are five ribosomal proteins present across the Crenarchaea, but absent in the Euryarchaea [235]. These proteins were either lost or gained in one of these groups after they split. In either case there would be a transition between two complete ribosomes. In each of these cases we can clearly see that a ribosome can undergo dramatic changes in macromolecular structure when there is proper selective pressure (or relaxation of selective pressures).

The tree presented in [42] was constructed by concatenating 31 universal proteins. 23 of these are ribosomal proteins and many more are directly involved in translation. Many taxa on the tree cluster together with high bootstrap values (greater than 80%). However, there appears to be only three connections between high level taxa that are supported with that strength. The clustering of Crenarchaea and Euryarchaea is well supported, as is the clustering of Eukaryotes and Archaea. There is also a long, well supported, branch between the archaeal-eukaryal clade and the Bacteria. We doubt it is a coincidence that these splits correspond to the greatest changes in ribosomal structure on the tree. It appears the sequence tree in [42] and rRNA trees could be merely a reflection of the large changes in ribosomal structure that have occurred throughout the true tree of life. This protein set would be expected to work better as a clock within groups that have the same ribosomal proteins. Even if ones uses more sophisticated tree building techniques, such as those in [21], the major changes in the ribosome are still going to

problematic. The authors concatenated many translational proteins and the resulting tree supported the paraphyly of the Archaea. The Eukaryotes were placed near the archaeal species with the most similar ribosomal structures. However, a single gene tree of RNA polymerase alpha subunit (RPOA) supported holophyly. This implies some of their results are an artifact caused by structural changes in a ribosomal revolution.

The third scenario could certainly be weakened if it was found that all the ribosomal proteins were essential in Bacteria and there was absolutely no way they could be tinkered with. We examined what ribosomal proteins are essential in eleven different bacterial species using the Database of Essential Genes [248]. There are sixteen ribosomal proteins that would need to be lost in the transition from a bacterium to an archaeon, as they are found across the Bacteria but never in the Archaea. None of these ribosomal proteins was found to be essential in all species, which is the first sign it could be possible to lose and replace some of the ribosomal proteins. Four of the sixteen proteins are essential in all species except *Mycobacterium tuberculosis* (Table 7-1). In fact only four of these proteins are essential in *M. tuberculosis*, the least of any species in this data set.

To determine whether this portion of the ribosome is significantly flexible we calculated a p-value assuming a binomial distribution. The essentiality of each subunit can be considered a success or a failure. The p-value measures the odds of seeing at most n essential subunits in a set of sixteen random ribosomal proteins. The odds of a random ribosomal protein being essential were estimated as the proportion of ribosomal proteins found to be essential in that species. This was done to eliminate experimental biases between the species sets, as some of the knockout experiments are more thorough than

others. Several species had p-values under .05, but *M. tuberculosis* was by far the most significant with a p-value of .0031. This implies that *M. tuberculosis*'s ribosome is under different selective pressure than most bacteria, and that it is the most preadapted ribosome in this dataset to evolve into an archaeal ribosome.

It is highly counterintuitive that nearly universal proteins could be nonessential. The difference between essential and persistent genes was discussed in [249]. The authors point out that essentiality differs in the wild and laboratory settings. Many of the ribosomal proteins listed as nonessential are still highly deleterious to lose. But the point is they can be lost under the right circumstances. It might be our proteasome centric view of the world, but we think the presence of the 20s proteasome in *Mycobacterium* could partially explain this observation. It has been proposed the major cost of mutations and mistranslation comes from dealing with misfolded proteins [250]. The ribosomal proteins are among the most highly translated proteins in the cell, so there is lots of pressure to ensure they fold correctly. A highly advanced degradation system, like the 20s proteasome with a Pup targeting system [251], could greatly relax that selective pressure. If the initial tinkering is not lethal, one can easily imagine a scenario where compensatory mutations and structures could significantly change the ribosome rapidly if there is proper selective pressure. We will describe such a scenario below.

It has been observed that many Bacteria contain paralogs of ribosomal proteins where one form binds Zn and the other does not [252]. *M. tuberculosis* has duplicates of several ribosomal proteins, which could explain why some (but not all) of the ribosomal proteins are not essential in that genome. The authors note that thermophilic bacteria seem to prefer the Zn binding forms of the ribosomal proteins, and that there are seven

Zn binding ribosomal proteins conserved across the Archaea and Eukaryotes that are absent in the Bacteria. This is consistent with some ideas our group has presented that major historical changes in the availability of Zn in the ocean were a major constraint on protein structure evolution [142, 253]. Bacteria vary their ribosomes to optimize for both high and low Zn conditions. One can imagine this strategy being taken to an extreme where the tweaks are not just simple displacements, but larger rearrangements. Increased availability of Zn as the ocean became oxic could be a factor that made toying with the ribosome favorable for the early archaea depending on the timing of these events. That combined with the antibiotic pressures discussed below could lead to a ribosomal revolution, just as the presence of two ribosomes lead to a revolution at the root of the Eukaryotes.

There truly is a great divide in DNA replication machinery, but it can be bridged

The differences between the archaeal and bacterial replication machinery is vast [230]. Leipe *et al.* claim this difference is so great that it is unreasonable to argue that one prokaryotic superkingdom evolved from the other. They list four key functions of DNA replication that are performed by completely non-homologous proteins in Archaea and Bacteria: the main polymerase's polymerization domain, the phosphatase that powers the polymerase, the gap filling polymerase, and the DNA primase. We will argue that the differences between Archaea and Bacteria do not imply the root of the tree of life has to be between them.

We must keep in mind there is some flexibility in DNA replication machinery despite the division across the superkingdoms. For instance many proteobacteria use a PolB family polymerase as a repair protein [254], which is almost certainly the result of HGT. PolD appears to have been present in LACA, but was lost in the Crenarchaea [255]. These are just two examples of major changes in replication machinery that happened in DNA based genomes with fully functional replication systems. We are just arguing an even larger event occurred between the prokaryotic superkingdoms. This event entailed viral transfers and novel innovations, but there are several proteins whose origins can be better described by vertical inheritance from the Gram-positive bacteria which we will review first.

Koonin *et al.* have demonstrated that many bacterial proteins have a region that is homologous to the small subunit of the archaeo-eukaryotic primase [256]. This domain is present in DNA ligase D from *M. tuberculosis*, which can act as a DNA-dependent RNA polymerase [257]. The rest of the protein is homologous to the ATP-dependent DNA ligase found in Archaea and Eukaryotes. Therefore DNA ligase D is perfectly preadapted to replace the primase function of DnaG. The fission of the two halves of the protein would allow for the preservation of ligase activity while developing enhanced primase activity. A recent analysis of DNA ligases revealed many transfers between the Archaea, Bacteria, and Viruses [258]. This history is very complicated, so it is hard to say with certainty where the Archaeal enzymes originated. The large subunit of the primase may be a true innovation since it has no detectable bacterial homologs, but the small subunit of the primase and ATP-dependent DNA ligases both could have been inherited from the Gram-positive ancestors of the Archaea.

The main helicase in Bacteria is DnaB, while Archaea use MCM6. Of great interest to this discussion is a recent biochemical analysis of a protein in a prophage element in *Bacillus cereus* that has domains homologous to the MCM6-AAA domain as well as the small subunit of the archaeal primase [259]. The authors found that the protein was a functional helicase but had no primase activity. The narrow distribution of this prophage element implies its insertion was probably too recent to play a role in the origin of the Archaea. However, it demonstrates that there can be a selective advantage for a DNA based genome to take novel DNA handling machinery from a virus and use it in a different context. We will come back to this point later.

The bacteria use DnaA to define the origin of replication, while Archaea use Cdc6. These proteins have a homologous AAA+ ATPase domain, but have little similarity besides that. However, the bacterial protein RuvB has the same domain combination as Cdc6. RuvB, Cdc6, and DnaA were all put in the same superfamily in a recent classification of AAA+ domains [260]. RuvB is recruited to Holliday junctions by RuvA where it forms a hexamer around the DNA [261], just like Cdc6. It is plausible that Cdc6 evolved from RuvB.

Archaea use a protein called Hjc to resolve Holliday junctions instead of the bacterial RuvABC system. Hjc is related to the alternative bacterial system RecU [262]. The only group of Bacteria that use RecU are the Firmicutes, and they also have RuvABC. We argue below the Archaea are derived from within the Firmicutes. It is possible the redundancy of Holliday junction systems allowed RuvB to drift in function. The homology between RecU and Hjc could be explained by the presence of a Holliday junction resolvase in LUCA under the canonical rooting. However if the hypothetical

RNA-DNA hybrid LUCA proposed in [230] was dealing with Holliday junctions we argue it probably would also need topoisomerases at that point. But since the distribution of topoisomerases is different across the prokaryotic superkingdoms [263, 264] that would imply the ancestral topoisomerase was displaced in at least one lineage. This weakens the proposal in [230]. We feel it is more likely the archaeal topoisomerases evolved from the bacterial ones as Cavalier-Smith has proposed [62].

There are certainly large differences between the Archaeal and Bacterial DNA replication machineries. We have demonstrated the divide between replication systems has some flexibility, and this opens the door for a replication revolution. It is possible to come up with detailed scenarios for how each of the Archaea replication proteins originated. These results are summarized in

Table 7-2. We will elaborate on this scenario below. However, there are several Archaeal replication proteins that do not appear to have any homologs in the Bacteria: histones, PolD, and the large subunit of the archaeal-eukaryal primase. These are true innovations, but there really are not that many of them; certainly not enough to make the transitions seem unreasonable in light of the polarizations presented above.

The proposal for two independent inventions of DNA replication has recently been challenged [265]. The authors argue that ribonucleotide reduction is thermodynamically unfavorable, so convergent evolution is highly unlikely. They note that all ribonucleotide reductases have been shown to have a monophyletic origin. Finally they argue that the proteins that are universally conserved imply a high fidelity replication system in LUCA that could not have been RNA based. The hypothesis the

root must between the superkingdoms is diminished when one combines these arguments with the scenarios we have outlined here.

Are the Archaea Paraphyletic or Holophyletic? We're agnostic

So far we have presented several independent arguments that strongly polarize the Archaea as a taxa derived from within the Bacteria. We have demonstrated that although there are vast differences between the ribosomes and DNA replication machinery between the prokaryotic superkingdoms, none of them seem totally insurmountable. We will soon present a novel hypothesis to account for this. But first we must pinpoint the bacterial roots of the Archaea. However, can not properly reason about that without first discussing whether the Archaea are paraphyletic (the Eukaryotes branch within them) or holophyletic (the Eukaryotes are their sisters). As there is clearly a relationship between the Archaea and the Eukaryotes it is vital to differentiate between these two scenarios to understand their origins. We will review the current available data, and argue that for now precise agnosticism seems the best course. Therefore any hypothesis on the origin of the Archaea needs to accommodate both of these models. That said, we lean towards holophyly and our hypothesis does as well.

The Eukaryotes and Archaea are sister clades under the standard three domain model. However, James Lake proposed that the Eukaryotes had a crenarchaeal (eocyte) origin based on a shared indel in EF-1 and similarities in their ribosomal structures [20, 238]. This hypothesis never gained much support because there was little phylogenetic

evidence to corroborate it. However, recent work [21] has shown that there is sequence data that implies the Archaea are paraphyletic and the Eukaryotes have a crenarchaeal-like ancestor. However, an analysis done around the same time supported a deep branching archaeon as the host of the mitochondria [266], which would be inconsistent with the eocyte hypothesis. They demonstrate the Eukaryotes inherited both crenarchaeal and euryarchaeal specific proteins, so ancestry from either group alone is not enough to explain the eukaryotic protein repertoire. However, several deep branching archaeal genomes from Korarchaeota and Thaumarchaea are now available and change the context of some of these conclusions [267, 268]. Both of these groups appear to contain a mix of crenarchaeal and euryarchaeal genes so the observation in [266] could be explained by a member of one of these groups being ancestral to the eukaryotes.

Cavalier-Smith's hypothesis on the origin of the neomura relies on a sisterhood relationship between the Archaea and Eukaryotes [62]. As discussed below, he mainly roots the neomura using traits the Actinobacteria share with the Eukaryotes, but not the Archaea. This only makes sense if the Archaea and Eukaryotes are sisters, otherwise the traits should be present in at least some Archaea. He lists eight properties that are unique and ubiquitous in the Archaea [62]. All of these traits strongly imply the Archaea are monophyletic. However, most of them do not differentiate between whether the Archaea are holophyletic or paraphyletic.

For instance, the unique isoprenoid ether lipids in all archaeal membranes are best explained by their presence in LACA. Eukaryotes have lipids that are more similar to those of Bacteria. It would be more parsimonious for the Archaea and Eukaryotes to be sisters with a single change in lipid structure. Any other scenario requires a reversion in

the Eukaryotes back to the bacterial state. Even though this is not parsimonious, it is not out of the question because the mitochondrial ancestor would have all the necessary genes to make bacterial membranes [269]. We have to admit that does not seem unreasonable relative to the innovations we are discussing in this work. This certainly seems like a case where simple parsimony in terms of any one trait, even membrane structure, will be highly misleading.

The only one of these properties that appeared really informative in regard to this problem was the split gene for RPOA. RPOA is the only single gene tree that supported the three domain model in [21], so it is clear the Eukaryotes did not get this protein from the mitochondrial ancestor. Reassembling the split gene is highly improbable, so there is no reason to doubt the fused genes are monophyletic. This strongly contradicts the original eocyte hypothesis. However, novel genomic data has revealed that representatives from the deep branching phyla Korarchaeota and Thaumarchaeota have the non-split form of this gene [267, 268]. This opens the door for a more specific version of the Eocyte hypothesis where the Eukaryotes stem from either of these groups. Therefore, we have examined what additional data have to say about these taxa. The branch order between them has not yet been resolved, but it appears safe to assume they both branch before the split between the Crenarchaea and Euryarchaea. This branching is supported by several phylogenetic trees as well as the non-split RPOA. This assumption will be key to our reasoning in several points.

It seems impossible to come up with a scenario that deals with all the traits we will discuss below that is completely parsimonious for all traits at the same time. With that in mind we have tried to reason which traits can better explained by convergent

evolution than others. When we observe convergent evolution happening at an indel site we do not consider it informative. Independent loss in any form is much easier than independent invention. Loss seems to be the rule rather than exception in the Archaea. Both the Thaumarchaeota and Korarchaeota have traits that were thought to be specific to either Euryarchaea or Crenarchaea. For instance Euryarchaea use FtsZ for cell division while Crenarchaea use the *cdvABC* system. Intriguingly the Thaumarchaeotal genomes have orthologs of both of these systems [270]. This implies that the Crenarchaea and Euryarchaea each lost one of these systems. This is not the most parsimonious solution, but it is the only one that is consistent with the apparent branch order of these taxa. Many other traits have the same distribution pattern. It is clear that groups of Archaea can lose proteins of major functional importance. We will attempt to address these distributions in our hypothesis below.

Beyond the EF-1 indel that implies paraphyly, six highly conserved indels were found to be informative on the relationship between the Archaea and Eukaryotes in [266]. The authors only looked at derived insertions with well conserved sequences. The authors state that four indels argue for the holophyly of the Archaea. There is one indel that is shared between the Eukaryotes and Crenarchaea, as well as one shared between the Euryarchaea and Eukaryotes. This implies there was a reversion in at least one lineage or a horizontal transfer.

We have analyzed those six indels as well as EF-1 in the context of the new deep branching genomes. The results are summarized in Table 7-3. Only the indels that differ between the archaeal groups are useful for determining their branch order. Therefore we created alignments that only contained archaeal sequences to ensure these indels were not

artifacts created by including eukaryotic and bacterial sequences. We used structural alignments from representatives of the superkingdoms where possible to further ensure the larger indels were real as we did in [38].

First of all, the reported indel shared between the Euryarchaea and Eukaryotes in the DNA repair protein RadA appears to be an artifact to us. In our archaea only alignment the euryarchaeal and crenarchaeal sequences align well in this region (Figure 7-2). This is important because it was the only line of evidence in that work that implied a relationship between the Euryarchaea and Eukaryotes. This new alignment in conjunction with the split RPOA gene implies the Eukaryotes either descend from within the deep branching archaea or are their sisters.

We also argue that the two reported indels in the alignments of Beta-glucosidase/6-phospho-beta-glucosidase/beta-galactosidase (PBG) and ribosomal protein S12 are both uninformative based off the authors' own analyses (supplemental data from [266]). The indel in ribosomal S12 is conserved across all Archaea and Eukaryotes, so it implies nothing about their branch order. The indel in PBG is uninformative because the authors conclude the eukaryotic version of this gene is of probably bacterial origin (supplemental data from [266]). Therefore the state of the gene in the Archaea implies nothing about the branch order of these groups.

Two of the remaining four indels are only a single residue. The glycine insertion in SecY is present in the Thaumarchaeota and Eukaryotes, but is absent in Korarchaeota. That weakly implies a relationship between the Eukaryotes and Thaumarchaeota. However, the fact that the insertion is present in some of the deep branching taxa, but not in all Euryarchaea, implies there was at least one secondary loss of this insertion. This is

reasonable since the insertion is a single glycine residue, and will not have a dramatic effect on protein structure.

The single residue insertion in prolyl-tRNA aminoacyl synthetase initially implied the Archaea were holophyletic. However, the insert is missing in the thaumarchaeal genomes. However, when these genes are used to seed a BLAST [183] search they hit the Firmicutes much more highly than other Archaea. This implies there was probably a horizontal transfer to the Thaumarchaeota. If that is the case this insert could still support holophyly, but that cannot be concluded with absolute certainty.

This leaves us with two larger indels in EF-1 and glutamyl-tRNA amidotransferase subunit D (gatD). The seven AA insert in gatD is well conserved in the archaeal alignment. A structural alignment with a bacterial homolog reveals this indel is not an artifact caused by the sequence alignment (data not shown). The phylogenetic tree for this family presented in (supplemental data of [266]) places the Archaea and Eukaryotes as sisters with 100% bootstrap support. This is remarkable because the archaeal proteins have a different domain combination and quaternary structure than the Eukaryotes and Bacteria [271]. However, it seems that tree is too good to be true. We have attempted to verify the history of this indel, and found that the tree in [266] was missing a bacterial paralog. *E. coli* has members of two paralogous families of l-asparaginase [272], and it appears only one of them was present in the initial tree. The tree in Figure 7-3 shows that the Fungi and the rest of the eukaryotes received the same domain superfamily from two distinct sources. Their sequences are mixed in with some bacteria, which implies there were some recent horizontal transfers. This tree is not well resolved, but it is certainly does not support the notion the Eukaryotes inherited this

protein from their archaeal ancestor. That, as well as the differences in domain combination and quaternary structure, implies this indel is inconclusive with regards to holophyly versus paraphyly.

EF-1 also appears inconclusive to us. The insert shared between the Crenarchaea and Eukaryotes is present in Thaumarchaeota, but not Korarchaeota. Our alignment revealed there are actually four different forms of the indels at this site found in the Archaea (Figure 7-5). This implies there is some plasticity in this region in the Archaea. This is in contrast to the bacterial alignment which has no indels in this region. A structural alignment between a bacterial representative from *E. coli* and an archaeal one from *Sulfolobus solfataricus* reveals the conserved glycines in the sequence alignments are very close in their position in both form of this indel (Figure 7-4). It is possible there were two insertions near the root of the Archaea that preserved the position of that residue. This indel's history does not appear to be parsimonious, which weakens its usefulness as a marker. Therefore this indel appears to weakly support archaeal paraphyly, but we consider it inconclusive.

The ribosomal proteins are the other side to this story. In a previous study, five ribosomal proteins were found in at least one crenarchaeon, but not in any of the Euryarchaea (L38e, L13e, S25e, S26e and S30e) [235]. These, as well as four others that are not universal in the Archaea, are conserved across the Eukaryotes. We examined what ribosomal proteins are present in the thaumarchaeal and korarchaeal genomes (Table 7-4). It still appears that Lake is correct that the Crenarchaea have more similar ribosomal proteins to the Eukaryotes than any other group of Archaea.

The Korarchaeota are missing three ribosomal proteins found in some Crenarchaea and Eukaryotes. They have five ribosomal proteins that are present across the Eukaryotes that are absent in the Thaumarchaeota. There are two ways we can interpret this trend. If the Archaea are paraphyletic then this distribution is best explained by the invention of ribosomal proteins after LACA. LECA could branch between the Korarchaeota and Crenarchaea, before the RPOA gene split. The alternative interpretation is that the Archaea are holophyletic and the Archaeal ancestor had all the ribosomal proteins that are in any archaeon and at least one eukaryote. There would have to be several independent losses of each these ribosomal proteins. Again this is not parsimonious, but there is evidence it has occurred several times so we must consider it. Again it can be argued that if a protein is present in the Korarchaeota and Crenarchaea, but absent in the Euryarchaea, it must have been lost. The archaeal ribosomal proteins are more dispensable than their counterparts in the other superkingdoms [235], so they might not be a reliable marker for rooting the Eukaryotes in the Archaea.

For now it seems the only reasonable stance in light of all of this evidence is agnosticism. Only when the Thaumarchaeota and Korarchaeota are sampled better, and their positions in the Archaeal tree are determined robustly, will it be possible to state with confidence whether the Archaea are holophyletic or paraphyletic. We might always be left trying to weigh whether reversion of ribosomal proteins or indels is the more parsimonious scenario. However, several of these traits clearly exclude the root of the Eukaryotes from within the Crenarchaea and Euryarchaea. Therefore any hypotheses on the origin of the Eukaryotes that invoke specific taxa within those groups can be rejected with confidence (for a discussion of the many hypotheses on this subject see [273]).

However, it may be possible those scenarios could be reworked to fit Thaumarchaeota or Korarchaeota once they are sampled better.

Where are the bacterial roots of the Archaea?

Weakening the neomuran hypothesis

Now that we have discussed the true distance between the superkingdoms we can begin to address how it could be bridged. From our discussion above we feel we must be cautious about declaring the debate closed on the holophyly of the Archaea. Therefore we are more interested in traits shared between a group of Bacteria and all Archaea than those shared with the Eukaryotes. Cavalier-Smith has presented fourteen reasons why the root of the neomura is probably within or next to the Actinobacteria [62]. Two of these traits are shared between the Actinobacteria and neomura, but the other twelve are only shared between the Eukaryotes and Actinobacteria. Under this scenario these twelve traits would be lost in the ancestor of Archaea, which implies the Archaea are holophyletic. We will review these fourteen traits, and argue that placing the Archaeal ancestor in the Bacilli makes more sense. We use the term neomura to refer to the clade of Eukaryotes and Archaea, but when we refer to neomuran hypothesis we refer to rooting that clade in the Actinobacteria.

The first piece of evidence that places the neomuran root near the Actinobacteria is the proteasome. Phylogenetic trees of the 20s proteasome subunits have the Actinobacteria and Archaea well separated which implies the presence of the 20s

proteasome across these groups is not the result of recent horizontal transfers. Recently 20s proteasomes have also been found in sequenced genomes from the Verrucomicrobia [274] and metagenomic sequences from *Leptospirillum* [178]. This somewhat weakens the Actinobacterial argument for ancestry, as the Archaea could inherit a proteasome from these other groups. However, that does not weaken the polarization argument at all; it just excludes the root from these additional groups.

The second trait apparently shared between the Actinobacteria and all Neomura is the post translational addition of CCA to the 3' end of tRNAs. The gene performing that function in the Archaea is tRNA CCA-pyrophosphorylase (protein cluster PRK13300 [222]). One of the domains, "PAP/Archaeal CCA-adding enzyme", does not hit any bacteria in the Superfamily database [72]. Since the CCA addition is performed by nonhomologous enzymes this is not strong evidence for rooting the Neomura. There is also an analogous enzyme conserved across the Bacilli (protein cluster PRK13299). Therefore, it is not clear which gram-positive group is ancestral to the Archaea even if they inherited this function from their bacterial ancestors.

Now we must address the dozen traits shared between the Actinobacteria and Eukaryotes. Although there were initial reports of sterol synthesis in the Actinobacteria [275, 276], the latest work has found no evidence for a complete pathway [277]. The authors report that the few cases of the full pathway in bacteria (all outside the actinobacteria) are probably the result of horizontal transfer. However, they find several sterol synthesis enzymes are present in many Actinobacteria. They conclude these are probably the result of a transfer from the Eukaryotes, but this is not supported by their trees which have good separation between the Eukaryotes and Actinobacteria. Several

sterol enzymes appear to have been inherited vertically from the Actinobacteria to the Eukaryotes. This is certainly consistent with Cavalier-Smith's hypothesis. This is a good example of the dangers of closing the debate on the position of the root too soon. Their trees clearly support an alternative hypothesis, but that data is buried in the supplemental material without discussion of the opposing view.

Initial reports also claimed the presence of chitin in Actinobacteria[278]. However, there is no gene for chitin synthase in the actinobacterial genomes. Several of them have chitinase which breaks chitin down. Also, chitin is found in Metazoa and Fungi, but not in the Archaeplastida which implies this enzyme was not in LECA.

It is true that Actinobacteria have many serine/threonine signaling systems related to cyclin-dependent kinases [279]. This would be a key preadaptation to the cell cycle. However, it has recently been shown *Bacillus subtilis* also has an extensive network of such regulation [280]. Therefore this line of evidence is consistent with either Gram-positive group being ancestral to the neomura.

Phosphatidylinositol is an interesting case. Recent work on this subject confirms the presence of phosphatidylinositol synthase as well the eukaryotic form of cardiolipin synthase in many Actinobacteria [281]. These enzymes are paralogs. We could not create a quality tree for this superfamily because the alignment was low quality. However BLAST searches had a good separation between prokaryotic and eukaryotic sequences which implies this is not the result of a recent HGT. It is difficult to determine exactly what family each prokaryotic homolog belongs to, so it is hard to say with certainty what other groups of bacteria have phosphatidylinositol. It is certainly possible the Eukaryotes inherited this from the Actinobacteria.

Some Actinobacteria do have an α -amylase with similar primary structure to the form found in Metazoa, but a recent comprehensive study found several other bacteria that did as well [282]. The authors concluded this was probably the result of a horizontal transfer due to their position in the phylogenetic tree as well the extremely sparse distribution of this form in the Actinobacteria. Therefore, this is not evidence for actinobacterial ancestry of the neomura.

The fatty acid synthetase (FAS) complex of Actinobacteria is unique among bacteria. They have the same form as some of the fungi [283]. Some fungi have this complex split into two genes, but the Actinobacteria have it fused. Our phylogenetic trees are consistent with actinobacterial ancestry (Figure 7-6). However, the distribution of the fungal type complex in Eukaryotes does not conclusively prove that this enzyme had to be in LECA. The only group outside the Fungi with this complex are Stramenopiles. However, the animal type FAS is also present in some Alveolata, so there could be some functional displacements. The Actinobacteria probably played a role in the evolution of this enzyme in Eukaryotes, but not necessarily via the neomuran hypothesis.

The argument that the exospore structure of Actinobacteria appears like it could be a good precursor to eukaryotic spore structures seems sound [284], but we are unable to locate a list of proteins involved in exospore formation. Without specific proteins homologs we cannot begin to evaluate this with bioinformatics. However, this argument becomes irrelevant if one invokes a viral ancestor of the nucleus as in [285].

Cavalier-Smith has also suggested that the C-terminal HEH domain found in the Ku proteins of some actinobacteria is ancestral to the HEH domain found in the eukaryotic Ku70 protein. However, the sequence analysis in [286] conclusively

demonstrates eukaryotes did not inherit the HEH domain from the Actinobacteria. This domain is very compact and common. Therefore, it is not out of the question for it to be recruited twice to the C-terminal of similar structures. Therefore we do not take this as evidence that the Eukaryotes inherited Ku from the Actinobacteria.

Several traits initially listed as unique to the Actinobacteria are now found in enough other bacterial groups that they are not ambiguous markers. The Actinobacteria do have tyrosine kinases, but they have recently been put into a bacterial specific family, BY-kinase [287]. This family is present across the Actinobacteria, Firmicutes, and Proteobacteria, so it does not exclusively support an actinobacterial rooting to the neomura. Many groups of bacteria have HU (histone H1 homologs) according to the Superfamily database. This protein is relatively short, so we should not expect sequence to resolve its history. It is possible this protein was inherited from the Actinobacteria, but there are too many other possibilities to state that with any certainty. Calmodulin-like proteins are now found in many bacteria, so this trait is not specific enough to root the neomura near the Actinobacteria as Cavalier-Smith now admits [208]. The Superfamily database reveals “Trypsin-like serine protease” are present in many groups of bacteria, but are absent in the Archaea. This appears to be another trait that is too general to be useful for rooting the neomura. There are certainly characters in common between the Actinobacteria and Eukaryotes, but they do solely support the neomuran hypothesis.

Evidence that supports a Firmicute ancestry

Skophammer *et al.* compiled several reasons to argue the Archaea are derived from the Bacilli [34]. There is an insert in ribosomal protein S12 that is present in the Archaea and Bacilli (and maybe Chloroflexi). Skophammer *et al.* conclude this indel is derived, but we argue elsewhere this polarization is flawed [38]. The insertion appears well conserved between the Archaea and Bacilli regardless of whether it is ancestral or derived.

Skophammer *et al.* also note that there is a shared deletion between the Firmicutes and Archaea in PyrD. Our own work strengthens this connection by considering the quaternary structure of PyrD. The form that has the deletion also has an additional subunit, PyrK. The sequence and structure of the Firmicute PyrD 1B are both shared by the Archaea. Our phylogenetic analysis of this protein implies this is not the result of recent horizontal transfers [38].

Skophammer *et al.* note that many enzymes involved in the biosynthesis of the unique archaeal membranes have been found in Firmicutes in a previous study [243]. The isoprenoid lipid precursors of the archaeal membranes are made via the mevalonate pathway, which is five enzymes long. The KEGG database [288] reveals the entire mevalonate pathway is present in several bacilli as well as some Actinobacteria (KEGG module M00191). The unique stereochemistry of the archaeal membranes is determined by the enzyme geranylgeranylglycerol phosphatase. Homologs of this enzyme are present in the Bacilli (protein cluster PRK04169), but appear to be absent in the Actinobacteria. The authors of an analysis of archaeal membrane biosynthesis propose the Archaea became genetically isolated from Bacteria once their membrane chemistry changed [289]. They suggest the Archaea branch early from within the Bacteria, but their

hypothesis is also consistent with a later Gram-positive origin. Cavalier-Smith's own analysis [208] suggest the eukaryotic enzymes that make n-linked glycoproteins, which are necessary for the loss of peptidoglycan, evolved from the firmicute specific gene EspE. The Firmicutes the bacterial group most preadapted to gain archaeal membranes for several reasons.

Homologs to ribosomal proteins L30e and L7ae are found across the Firmicutes. This is novel evidence of the link between the Firmicutes and the Archaea. Pfam [290] shows this family in several other groups, but many Firmicutes contain two copies of this family. One of these paralogs has been characterized as a ribosomal protein, but neither is essential [291]. We constructed phylogenetic trees to see if they are consistent with vertical inheritance (Figure 7-7). There is good separation between the paralogs in the Firmicutes, which implies the duplication occurred early in the Firmicutes. All the archaeal and eukaryotic genomes contain at least two copies of this family. The phylogenetic tree of the archaeal and firmicute sequences places the firmicute paralogs between the archaeal paralogs. The firmicute sequences are paraphyletic although with very weak support. If these proteins are the result of independent duplications the archaeal sequences should cluster together, not on opposite ends of the tree. However, it is possible one of the archaeal sequences evolved rapidly after the duplication.

One of the paralogs in *Bacillus subtilis* was found to localize to a different portion of the ribosome than either of the archaeal paralogs [291]. The proteins would not only have to jump superkingdoms for a transfer to occur, they would also have to bind to a different region of the rRNA without interfering with ribosome assembly. We argue it would be less disruptive for a protein already present to gradually bind a different piece

of rRNA. The separation between the superkingdoms in the phylogenetic trees also argues against HGT. If this is the result of vertical inheritance only two possibilities explain it. Either the Firmicutes are ancestral to the Archaea or the root lies between the Archaea and Firmicutes. Our polarization of PyrD 1B's quaternary structure eliminates the latter rooting as a possibility. This tree appears to support a Firmicute ancestry for the Archaea, although it may just be the result of rapid evolution of structures in different contexts.

As discussed above almost all the firmicute genomes have a unique Holliday junction resolvase RecU that is found very sparsely in other bacterial groups. It is homologous to the Archaeal Holliday junction resolvase Hjc [262]. Therefore the Firmicutes have more similar DNA repair to the Archaea than any other bacterial group.

Hsp90 is missing in all archaeal genomes, so its presence across the Eukaryotes and Bacteria implies it was inherited from the mitochondrial ancestor. However, a detailed analysis of this family did not reveal a relationship between eukaryotic and proteobacterial sequences [292]. Instead, the eukaryotic sequences branches within the Gram-positive bacteria. The authors argue this supports the classical neomuran hypothesis, but the Eukaryotes are sisters to the Firmicutes rather than Actinobacteria in that tree (albeit with moderate support). This would slightly favor Firmicutes ancestry over the Actinobacteria. In either case it supports the view that the Archaeal ancestor lost Hsp90.

Peroxisomes: the red herring?

There are certainly several traits present in either the Firmicutes or Actinobacteria that argue they are ancestral to either the Eukaryotes or Archaea. The only one that strongly argues the Actinobacteria are ancestral to the Neomura is the proteasome. Several more make compelling arguments that the Actinobacteria are ancestral to the Eukaryotes, but certainly not the full dozen listed in [62]. In Cavalier-Smith's most recent version of the neomuran hypothesis he admits the Firmicutes contributed a significant amount of genes to the neomuran ancestor [293]. He proposed the neomura originated as sisters of the Actinobacteria, and both of these taxa are descendents of the Firmicutes. That proposal is dependent on his argument the Actinobacteria are derived from the Firmicutes, which is one of the less developed ideas in [208]. We believe he is wrong in his assertion our analysis of the indel in ribosomal S12 [38] does not support firmicute ancestry of the Archaea. It is only shared (and well conserved) between the Bacilli and the Archaea regardless of the polarization of that indel. Cavalier-Smith is also not aware of the arguments about L7AE paralogs and RecU we present here for the first time. So we are left with a stronger list of reasons for firmicute ancestry and a weaker list for actinobacterial ancestry. However, there are still some key eukaryotic proteins that appear to be descended from the Actinobacteria. We will try to reconcile this apparent anomaly.

The peroxisome is an organelle with a single membrane found across the Eukaryotes that has various oxidative functions including the synthesis of some lipids [294]. They have been observed to divide independently of the rest of the cell which

initially led some, including Cavalier-Smith, to question whether they had an endosymbiotic origin [295, 296]. Two recent studies both concluded that the peroxisome was likely derived from the endoplasmic reticulum [297, 298], which led those initial proponents of peroxisomal endosymbiosis to abandon that idea.

However, [297] found that many peroxisomal proteins likely originated in Cyanobacteria, α -proteobacteria, or Actinobacteria. The authors explain that the proteobacterial genes were probably transferred from the mitochondria which is consistent with observations that mitochondrial genes are often retargeted to other organelles [299]. However, recent work argues for an endosymbiotic origin of the peroxisome from an actinobacterium [300]. They demonstrate that at least two proteins imported into the peroxisome are of actinobacterial origin, and that the peroxisomal proteome has higher average BLAST scores to Actinobacteria than any other group of prokaryotes. They argue that the retargeting of mitochondrial proteins after their genes migrate to the host's genome is easier than *de novo* targeting of peroxisomal proteins. They propose this masks the true history of the peroxisome.

The literature proposes two scenarios to explain the origins of the peroxisome: either the peroxisome was an endosymbiont or the Actinobacteria were not endosymbionts. Clearly there is a 3rd possibility; there was an actinobacterial endosymbiont, but the peroxisome is not a descendent of that membrane. That is to say, genes of an endosymbiotic origin are targeted into the peroxisome, but historically they are foreigners there. How could this be? A primitive peroxisome derived from the endomembrane system would be beneficial because it would separate dangerous oxidative chemistry from the rest of the cell. Proteins would be targeted to the organelle

with relative ease since that system would be developed in the mitochondrial endosymbiosis. Genes would be copied from the actinobacterial endosymbiont to the host genome (but not necessarily lost in the actinobacterium), and then imported into the peroxisome. This would be more advantageous because some of these reactions would do better in that specialized environment instead of their original host. There would be probably less cost involved in maintaining an organelle that already existed versus an entire endosymbiont. Once enough genes were present in the host, the actinobacterial endosymbiont would essentially be a parasite, and its complete loss would be beneficial.

That would be in contrast to organelles such as plastids and mitochondria that have both retained genomes and membranes long after they have become organelles. Some have questioned why they retain any genes at all [301]. The authors note that most genes retained in plastids and mitochondria are membrane spanning proteins involved in the core of the photosynthetic and respiratory systems. They agree with an earlier proposal that these proteins must be kept in the organelle to be able to quickly respond to and balance redox gradients [302]. In other words, plastids and mitochondria have retained membranes and genes because their functions are centered around membrane based chemistry. The stripped down endosymbionts perform these functions better than a novel organelle initially could, so they are left with a few essential genes and membranes they inherited from the endosymbiosis. Those genes come with a high cost because the organelles need to import the machinery to translate them as well as the machinery to replicate the genes that encode them. Therefore it is not a stretch to imagine that other endosymbionts whose functions are not as membrane-centric could be replaced by organelles that are not of endosymbiotic origin. Unfortunately these two organelles

have greatly shaped our expectations that endosymbionts will leave both membranes and genomes behind. That is an over simplistic expectation.

We argue that an actinobacterial endosymbiosis accounts for the traits shared between the Eukaryotes and the Actinobacteria, as well the phylogenetic trees that place Actinobacteria as sisters of the peroxisomal proteins. The fact that numerous mitochondrial proteins are imported into the peroxisome is evidence this endosymbiosis - occurred after the mitochondrial endosymbiosis. This would reconcile the apparently conflicting signals in terms of which Gram-positive group is ancestral to Archaea and Eukaryotes. We find this scenario more reasonable than invoking an extinct lineage of Gram-positives that has all the traits listed in Table 7-5 and Table 7-6. However, if a genome is sequenced that contains the actinobacterial specific traits as well as firmicute specific listed traits here we would very quickly have no need to invoke endosymbiosis. It is also possible to reconcile the canonical rooting with the traits shared by the Actinobacteria by invoking this endosymbiotic hypothesis.

Viruses are the missing link between the prokaryotic superkingdoms

Now that we have described the true distance between the Archaea and Bacteria, the time has come to cross that desert. As we have asserted above, this is a unique event in evolution, so we must properly set the stage. The selective pressures of extreme environments and antibiotic warfare are ancient. They cannot cause a revolution on their

own; a great relaxation in selective pressure is necessary. We will argue that a viral endosymbiosis could relax selective pressure enough to start such a revolution.

Koonin has observed that the PolB family of polymerases are the most common DNA polymerase in viruses[303]. Koonin *et al.* also observed that archaeo-eukaryotic DNA primase was a hallmark viral protein [241]. This hints at some connection in DNA replication between the Archaea, Eukaryotes, and Viruses. We examined the distribution of all protein families in Pfam [290] that originated at the root of the Archaea and Eukaryotes to see if this connection could be extended. We defined Pfams that were present in least 90% of the archaeal genomes (out of 46) and 90% of the eukaryotic genomes (out of 35) and in less than 50% of the bacterial genomes (out of 939) as originating at the root of the Archaea and Eukaryotes. The 90% cutoff is strict enough to imply the protein was present in LAECA, while the 50% cutoff is loose enough to accommodate recent horizontal transfers. Most of these Pfams are well below the 50% cutoff in the Bacteria.

There are 74 Pfam domains that originated in LAECA by this definition. 24 of these are found in at least one viral genome. Each of those Pfams is in an average of 36.38 viral genomes (14.36 if one excludes PolB). As a loose measure of the significance of that result we took 10000 random samples of 74 Pfams that are found in at least one cellular genome to see often one finds 24 or more Pfams in at least one viral genome. None of the random sets had that many viral pfams, which implies this set is significantly enriched in viral proteins. However, we must keep in mind our sampling of the viral world is still highly biased (discussed in [303]) and that viral genomes evolve rapidly. They are sampled so poorly that no viral genomes had the MCM domain in Pfam, even

though it is found in a prophage region in some bacilli as discussed above. Also 18 of the remaining Pfams that originate in LAECA are ribosomal proteins, which are probably less advantageous for viruses to encode than DNA replication machinery (although we did find several ribosomal proteins in viruses in this set).

We can also verify if this result is significant by looking at the set of proteins that would be present in LBCA (last bacterial common ancestor), but not LEACA under the same definition: Pfams present in at least 90% of bacterial genomes and less than 50% of archaeal and 50% of eukaryal genomes. There are 106 such Pfams, and 15 of them are found in at least one viral genome (p-value of .2457). Each of those is in an average of about 8.33 viral genomes. It must be noted this is an underestimate for LBCA's content since there are so many parasitic bacteria with genomes sequences available. However, in general Viruses share more Pfams with LAECA than LBCA.

Koonin proposes that PolB's distribution reflects the fact the Archaea arose from an acellular ancestor and then retained the more ancient polymerase [303]. We find this view very hard to reconcile with the three independent arguments for the Archaea being derived from the Bacteria provided above. Forterre has argued that DNA originated from a viral endosymbiosis in each of the superkingdoms[240], but our data argues against that scenario in the origin of the Bacteria. We propose the alternative hypothesis that a viral endosymbiosis occurred in the Bacteria, and gave rise to the Archaea. This virus would supply the missing link in terms of DNA replication machinery between the prokaryotic superkingdoms. We think this would have to be an endosymbiosis and not just a horizontal transfer because of the distribution and interdependencies of these systems in cellular life.

Roughly speaking it seems safe to say that there are three components that define the propensity of a genome to get permanently damaged. The first is the environment. Many different extreme environments are damaging to DNA including radiation, high temperature, and desiccation [304]. The second component is the size of the genome. The larger the piece of DNA is, the more likely damage is to occur, and the more of it must be mediated. Thirdly, one must also consider the state of the active repair system. If one has poor active repair even rare damages will eventually accumulate. Therefore we argue that systems that are extreme in any one of these components are going to have to expect to deal with DNA damage during replication as the norm.

Archaea, in general, fit the description of extremophile better than any other major taxa. It has been proposed that the unifying trait of all Archaea is adaptation to chronic energy stress [305]. The author argues that Archaea outcompete Bacteria in niches that are under chronic stress. The Archaea have become successful by dealing with environments that other superkingdoms cannot handle. The author noted that Archaea do better in environments that are consistently extreme, and are outcompeted by Bacteria in environments that fluctuate more.

A corollary of chronic energy stress is chronic DNA damage. Many of the extremophilic environments the Archaea have made home severely damage DNA. On the other hand, Bacteria may face a stressful situation and require DNA repair. Therefore it is disadvantageous for them play it safe and have their repair systems on all the time. The Archaea need to constantly repair their DNA, so it would make sense if the line is blurred between their replication and repair systems. An example of this prepare for the worst

strategy is the unique ability of PolB to read ahead and stall replication if a uracil is encountered in the Archaea [306].

In terms of large genomes Eukaryotes win hands down (see figure 1 in [73]). A polymerase is more likely to encounter damage somewhere in the replication of these large genomes than in a smaller genomed prokaryote in a similar environment. This is supported by evidence that the Eukaryotes use a separate repair system during replication for the large non-transcribed regions of their genome [307, 308].

What other situation besides chronic DNA stress and large genome size would put similar pressure on the DNA replication machinery? We argue, somewhat counter intuitively, that a total lack of active DNA repair systems would create a similar situation. Again it is optimal for the replicative system to expect to encounter damage. Viruses fit that description perfectly as they are unable to actively maintain their genomes without their host.

If the repair systems were turned on more and more of the time, the main replicative system would become free to drift. Under this scenario the ancestors of the Archaea could mix and match bacterial repair and replication proteins with several molecular innovations, and some transfers from the viral endosymbiont that would create a system that is more robust to chronic stress. The canonical rooting implies that the components of the replication machinery that are homologous but not orthologous were independently recruited from proteins that initially processed RNA. Under either scenario the same amount of molecular innovation is required. The question becomes whether it is easier to innovate function in an RNA based organism or a DNA based organism that is under extreme relaxations in pressure. We argue that difference cannot be quantified, as

both scenarios predict exactly what we observe: some proteins are orthologs, some are homologs, and some are unrelated. Therefore the way to tell the difference between these scenarios is independent lines of evidence. The polarizations presented above imply the bacterial repair machinery was recruited to become the replication machinery of the Archaea.

It is also tempting to speculate that many of the features shared between Viruses and the eukaryotic nucleus described in the viral eukaryogenesis hypothesis [285, 309, 310] could be extended to this hypothesis. Bell notes many similarities between nuclei and viral replication factories. One can imagine the ancestry of these traits going back to LAECA with some being lost in the Archaea, and others not developing until the root of the Eukaryotes. This would only work be consistent with our hypothesis if the Archaea are holophyletic, but for now it is certainly worth considering.

The greatest battle ever fought

So far we have demonstrated that there is robust evidence the Archaea are a derived superkingdom. We have shown the bacterial ribosome could have enough plasticity to evolve into an archaeal one. We have presented evidence that there is some link between DNA replication in the Archaea, Eukaryotes, and Viruses that could be the result of an endosymbiosis. Now we will try to combine these into the larger story of why a bacterium would evolve into an archaeon.

As we discussed above, we feel the greatest weakness of Gupta's invocation of antibiotics is it is not a special enough pressure to cause a revolution on the scale necessary to create the differences between the prokaryotic superkingdoms. The observations of the vast differences in DNA replication machinery and the evidence of a viral endosymbiosis in a bacillus before LEACA will set the stage in our hypothesis.

The Gram-positives in the traditional antibiotic battle are capable of evolving resistance to each other. This leads to what is commonly referred to as a Red Queen game [311]. Neither group ever really gets ahead in the long term war as each defensive innovation is usually matched by an offensive one. But that does not mean there are never winners in battles on shorter time scales. Winning a battle is not a good thing in the long run. The winners will increase in population size and consume more of an environment's resources. The corollary is that they become a better target for less dominant species to kill. If a species evolves a more resistant ribosome it just puts more pressure on the rest of the community to hit their other targets.

One can imagine a firmicute deeply entrenched in such warfare endowed with the gift of a complete and novel replication system from a virus. This is supported by the distribution of viral Pfams discussed above. The virosphere contains so much diversity that even rare combinations of genes would eventually end up in the same capsid at the same time as long as they have some advantage to any virus. It would be an incredibly rare event for the virus to be just right for the bacterium to take up the entire replication system. And thus the stage is partly set for why the revolution happened but once.

The core of the DNA replication system does not appear to be as common of an antibiotic target as the ribosomes or RNA polymerase. A search of DrugBank revealed

no antibiotics that target PolC [312]. However, there are several that target gyrase. We propose the difference is that inhibition of PolC just stops a population from growing but the damage induced by a loss of a functional gyrase invokes an SOS response, and leads to cell death. There are probably natural antibiotics that target PolC, but they would not be as effective as the numerous ones that target the ribosome and RNA polymerase. Therefore the introduction of PolB into the bacillus genome would not be the enough to start the revolution. This is supported by the fact that many proteobacteria use PolB as a repair enzyme, the result of a HGT that did not start a revolution.

As discussed above there are no Bacteria that have archaeal histones. This strongly implies they are only compatible with the archaeal-eukaryal replication machinery. Therefore we argue the viral endosymbiosis was a relaxation in selective pressure that in combination with pressure from antibiotics targeting gyrase led to the innovation of histones. This is not a trivial difference with Cavalier-Smith's hypothesis that the numerous differences between the DNA-handling machinery of the Bacteria and Archaea are the result of histones dramatically changing the way in which this machinery could interact with the DNA [62]. He argues this was an adaptation to thermophily.

However, Forterre has presented several arguments against Cavalier-Smith's scenario. He argues that the bacterial histone-like proteins that have replaced the archaeal ones in *Thermoplasma acidophilum* work just fine with the archaeal replication machinery [240]. He also notes that many hyperthermophilic bacteria do not use histones. That can be supported further by considering the fact that hyperthermophilic bacteria exchange many genes with Archaea [58]. Therefore the standard bacterial replication machinery could probably not tolerate the invention of histones even under

selective pressure from an extreme environment. Euryarchaea appear to have gained DNA gyrase via several independent horizontal transfers from the Bacteria [263]. The fact that several Euryarchaea retain both histones and gyrase is evidence against Cavalier-Smith's idea that gyrase became totally redundant with the advent of histones. That view is weakened further by the fact that gyrase was found to be essential in several of those genomes [313].

Since pressure from thermophily alone could not force the innovation of histones, we invoke the viral endosymbiont. In other bacteria an alternative system to gyrase would not be much of an advantage, as getting rid of gyrase would just put more pressure on targets like the ribosome and peptidoglycan synthesis. However, as discussed above the Bacilli have several unique ribosomal proteins. That means they could already have some adaptations and preadaptation to antibiotic warfare that makes them a difficult target to hit. As discussed above they have EpsE [314], which could preadapt them for functioning without peptidoglycan. Once gyrase was no longer a useful target they could quickly lose peptidoglycan in their cell walls. The loss of these two major targets would be a huge advantage and increase pressure on the ribosomes as a target.

At this point any change to the ribosome would be highly beneficial. One can imagine a Red Queen game where the neomura have a distinct advantage over the Gram-positives but need constant innovations in their ribosomes to maintain that advantage. The observation that many archaeal-eukaryal ribosomal proteins bind Zn would be consistent with pressure to ensure proper assembly despite the antibiotics. This is supported by the fact that bacterial hyperthermophiles, whose environment interferes with ribosomal assembly, have more Zn binding sites than most other bacteria [252].

This means the initial neomura would have an advantage in antibiotic warfare as well as the ability to replicate DNA even in the presence of damaging pressures. Their genomes could be much larger than extant prokaryotes. A large robust genome would allow the neomuran to be oligotrophic and handle extreme environments. This would put them in direct competition with many bacteria in diverse environments. Their larger genome size would allow for more gene duplications, which could lead to structural innovations like the ribosomal proteins found in neomura but not the Bacteria.

The strongest support of this view comes from the antibiotic target site most studied in the ribosome: the 23s RNA between ribosomal proteins L22 and L4. L22 and L4 are conserved across the Archaea, Bacteria, and Eukaryotes. They bind the same positions on the ribosome in all three superkingdoms. There are numerous crystal structures, from both prokaryotic superkingdoms, of these sites bound with antibiotics. [315, 316]. Those studies demonstrated that nine different antibiotics that bind this site well in Bacteria bind with much less affinity in Archaea. A2058 (*E. coli* numbering) is one of the sites on the 23s RNA directly involved in binding these drugs. A2058 is conserved across 99.4% of sequenced bacterial 23s rRNA[317]. That site is almost universally guanine in Archaea and Eukaryotes. The mutation A2058G makes many bacteria macrolide-resistant [318], while the reverse mutation can make Archaea macrolide-sensitive [319]. These differences in antibiotic affinities are well conserved across the divide between the Bacteria and neomura, and appear to be the result of intense selective pressure from antibiotics.

Even though Bacteria are able to gain resistance through a similar mutation, it is probably not fixed because there is a slight decrease in fitness that can be reduced with

other mutations [317]. If there was constant pressure on that site other mutations and changes in structure could relax those costs and fix that position. That would be completely consistent with the scenario outlined here. If the divide between the Archaea and Bacteria is primordial, it is much harder to explain this difference. Ribosomal proteins L22 and L4 must have been present in LUCA. If the ancestor of the Archaea was an extremophile they should not have been in competition with enough Bacteria to need the resistance inferred by this mutation.

It would be tempting to speculate that this mutation is an adaptation to thermophily or some other extreme environment to answer this nagging issue of antibiotic pressure at the root of the Archaea. However, this can be tested by examining that position in bacterial hyperthermophiles. In both the hyperthermophiles *Aquifex aeolicus* and *Thermotoga maritime* this position is 100% conserved as adenine, as it is in their thermophilic relatives (Figure 7-8). The thermophile *Thermus thermophilus* has two copies of the 23s RNA that both have usually have adenine at that position unless they are under selective pressure from antibiotics [320]. The only explanation that appears to hold water is some extreme antibiotic pressure at the root of the Archaea.

The mark of antibiotic pressure can also be seen in the proteins that would be lost at the origin of the Archaea. We searched Pfam and DrugBank for antibiotic targets that are conserved across the Bacteria but were clearly not in LACA. Eight of these are listed in

Table 7-8. Several of these appear to have been horizontally transferred to the Archaea, such as DNA gyrase. That is consistent with this scenario because once the Archaea were no longer under strong antibiotic pressure these systems would be free to become essential again. It would be interesting to look at each of these eight predicted

losses, and see what preadaptations and environmental conditions can make them non-essential.

Why would this war end and who would the winners be? We will invoke the two novel niches that are central to the neomuran hypothesis; phagotrophy and hyperthermophily [62]. The large genomed oligotrophic neomuran would be able to form many symbioses with prokaryotes because of its diverse metabolism. Such an environment would favor the preadaptations to phagotrophy discussed in [321]. This could lead to several endosymbiotic events in a short span of time. These would force the nucleus to become a better separator to deal with the selective pressures proposed in several hypotheses: invasion of introns [322], differing metabolisms [323], and ribosome chimerism [324]. The successful phagotroph would eat prokaryotes, so at first it would be to the advantage of the prey to try to kill the neomura. However, that is not the optimal strategy for dealing with phagotrophs. It is much better to persist inside them and eat them from the inside out, as can be seen by the numerous bacterial taxa that have independently evolved the ability to infect Eukaryotes. Once it is possible to infect the phagotrophs, killing them with antibiotics becomes counterproductive. And thus a truce (or new war) would be declared on one front of the great antibiotics war.

The early eukaryotes would outcompete and eat many of the initial neomura, but would be at a disadvantage in extreme environments as they began to rely on their cytoskeletons and larger cell size more and more. It would be easier for the neomura to drift into more extreme environments because of their DNA replication machinery. The proto-archaea would begin to emerge as the neomura began moving into previously unoccupied niches of extremophily. The conversion of their membranes would probably

be the committed step in the process. Once they began settling into environments that are constantly extreme they would be under pressure to streamline their genomes.

This is consistent with a recent study on gene content evolution in the Archaea that concluded that most archaeal genomes have been streamlined from larger ancestral genomes [325]. They conclude that the archaeal ancestor could have had 2000 gene families, and the extant archaeal groups are mostly created through differential loss. The authors note this repeated loss is consistent with the energy shock state of the archaea described in [305], as specialization and loss are highly favorable in consistently extreme environments. The trend of euryarchaeal and crenarchaeal specific traits both being present in the deep branching archaea is also consistent with the idea the Archaea became specialized from a more generalized genomic ancestor. The redundancy in archaeal systems such as two replicative polymerases and two cell division systems could be remnants of the antibiotic war. That redundancy would become unnecessary once the Archaea committed to extremophily. It was noted in [235] that ribosomal protein loss is much more common in the Archaea than in the Bacteria. Our hypothesis implies that the distribution of ribosomal proteins in the Archaea is the result of independent losses once they were no longer under antibiotic pressures. Some of these novel proteins developed other roles to deal with extremophily so they have been retained. The ancestral Archaeal ribosome could very well have contained all of the proteins found in any archaeal genome, which would certainly weaken that aspect of the Eocyte hypothesis.

And what about the neomura themselves? They would be stuck in the middle. The Eukaryotes would be eating them, and they would still be in competition with the Bacteria. Their only viable strategy would be constant innovation, as they would not

really have a novel niche. However, the wave caused by viral endosymbiosis would not go on forever. There would be diminishing returns in terms of the resistance provided by the new innovations. Eventually the innovations would become a disadvantage as Bacteria can then release compounds that only target the new systems. For instance Aphidicolin inhibits DNA replication in Archaea and Eukaryotes but not Bacteria by targeting their unique polymerase [326, 327]. So the initial advantage the neomura have in terms of antibiotic resistance is not a stable niche. They were outcompeted from three sides, and thus we are left with a hole in the middle of the branches of the tree of life that often gets mistaken as the root.

Discussion

It is fair to ask how different the Archaea and Bacteria would have to be for us to consider the rooting debate closed. If the genetic material was different between the superkingdoms it would be strong evidence of life being polyphyletic. If the genetic codes were somewhat different (even a few codons), that would certainly be evidence that both groups were primordial. If membrane proteins like SecY were not universally conserved, we would take that as evidence LUCA was acellular. The differences between the prokaryotic superkingdoms seem small if we consider the fact that the last prokaryotic common ancestor had a membrane and a ribosome that used the same genetic code as all extant life. They have more in common than can be described by any tree.

None of the differences between the Archaea and Bacteria are great enough to imply a transition between the superkingdoms is impossible. The three independent polarizations provide compelling evidence the transition occurred. A viral endosymbiosis in a firmicute host could be the relaxation in selective pressure that acted in combination with pressure from antibiotics to cause a revolution in terms of membranes, ribosomes, and DNA replication machinery. This is supported by the association between proteins found in viral genomes and those that appear at the root of the Archaea and Eukaryotes. Gupta's hypothesis that antibiotics led to the differences between the superkingdoms is well supported by the data generated in the past decade.

The Archaea would certainly need to have several innovations in terms of protein structure. None of these are deal breakers. They are present in extant cells, and are not found in viruses. So they would have been an innovation in a cell at some point. There is no reason to assume all structural innovations happened near the root of the tree of life. Work from our group probing the relationship between ancient ocean chemistry and protein structure evolution is an example of one source of later innovations [142, 253]. The modern ocean has several orders of magnitude more Zn than the ocean of LUCA's time [328]. Many Zn extant binding sites evolved after that transition. As noted above, several of the ribosomal proteins unique to the neomura have Zn binding sites. One of the innovations needed, PolD, is predicted to have two Zn fingers [255]. Increasing levels of Zn would not be the only factor, but it is another example of how the revolutionary planetary changes shape evolution as discussed in [329]. This observation makes sense if one places the origin of the Archaea after the great oxidation event, and considers the fossil record as a supplement to phylogenetic data. If we look at the details we may find

the rhyme and reason to the other novel structures at the root of the Archaea as well.

There are also many structural innovations at the root of the Eukaryotes [141]. The fact that the Archaea have many unique protein structures does not imply they are primordial.

As we have hinted above, one of the strengths of this hypothesis is it does not rely on the Archaea being holophyletic. The scenario we have described implies holophyly, but if something conclusively proved the Archaea were ancestral to the Eukaryotes, it could be adapted. There is no explanation in the neomuran hypothesis for the traits shared between the Actinobacteria and Eukaryotes besides vertical descent. The link Cavalier-Smith has justified could be the result of an endosymbiosis that did not leave its mark with an extant organelle. If the Archaea are paraphyletic it just means the Eukaryotes did not originate for the reasons we have hinted at here, but rather more along the lines of traditional endosymbiotic hypotheses. It does not change the way we have to think about the origin and rooting of the Archaea, which is the central focus of this paper.

The hypothesis we have proposed can be refined with experiment. It seems if one really wants to understand the likelihood of intermediates between the Archaea and Bacteria we need to understand why hybrid systems are unheard of. For instance, what other proteins need to be placed into a bacterium to allow them to use histones? How would an Archaea with a bacterial ribosome function? Trying to recreate the intermediates we believe went extinct would certainly give insight into their plausibility. It definitely would give better insight the functional nuances of the proteins with homologous function across the prokaryotic superkingdoms that appear to be highly resistant to horizontal transfer. That would be highly informative regardless of the location of the root of tree of life.

We have drawn our data from diverse sources that are not usually the primary tools for studying evolution. Viruses have been getting more attention as players in shaping the tree of life recently [227] and better sampling will clarify the plausibility of the endosymbiosis we have proposed. However, essentiality and protein structure are non-traditional tools in this field. If essentiality experiments were performed across the ribosomes and DNA replication machinery of the Bacilli under different conditions it could give us hints of what selective pressures would need to be relaxed for the major transition to begin. Further study of natural antibiotics will also continue to increase the resolution of the hypothesis. We argue this line of experiment would be useful in its own right, since many of the Firmicutes are pathogens that effect human health.

Of course in depth sampling of Thaumarchaeota and Korarchaeota is going to be invaluable to this endeavor. If the ribosomal proteins that are currently missing in these groups are found in new genomes it would imply independent losses and make holophyly seem a little more appealing. The redundancy left in these genomes could just be the first surprise. Deeper sampling may reveal redundancy in the forms of the some of the archaeal-bacterial hybrid systems we discussed above. Finding a deep branching archaeon that uses a bacterial system is not necessary for validating this hypothesis, but it would be sufficient!

There are aspects of this transition we did not cover in this work. However one that we have neglected that is not outside our scope of expertise is the proteasome. If one roots the Archaea in the Bacilli it does not explain the presence of the proteasome across the entire superkingdom. We think Cavalier-Smith is correct in pointing it out as a link between the Archaea and Actinobacteria, but in light of the other evidence raised here we

do not find that argument convincing on its own. It is not clear what direction the proteasome was inherited. Even if the proteasome was horizontally transferred it does not weaken the polarization of the Archaea; it would still be a derived structure that was present at the root of the Archaea, so they must still be derived.

There are many instances where data is only presented under the canonical rooting, when in fact it is better explained by an alternative scenario. This quickly leads to circular logic; a hypothesis gets buried because no data supports it, data gets buried (in supplemental data or *ad hoc* invocations of HGT) because it does not fit with the canonical rooting. As an example look at how much data from Eugene Koonin's group we have cited in this work to support our hypothesis even though he has made it clear he thinks this rooting is unsupportable (see his reviews of [38, 293]). The view that there is no reasonable data to support a rooting within the Bacteria is simply not true. But for some reason it's the prevailing opinion.

For instance, one of the biggest problems with the canonical rooting is the origin of cells. The term "RNA world" is sometimes invoked as a miracle that could explain anything that happened in evolution before cells looked the way they do now. But one thing RNA definitely cannot do is make transmembrane pores. This problem is addressed well by the obcell hypothesis of Blobel and Cavalier-Smith [211, 330]. They propose proto-cells had very little going on inside them initially. Rather, they were collections of ribozymes tethered to the outside of a cell. The details of their proposal get around the problems of transmembrane RNA structures, but also imply the first true cell had a double membrane (like the Gram-negative bacteria). Our point here is not about which hypothesis is correct; but that both of them are understood better in terms of the strengths

and weaknesses of the other; throwing either out is essentially operating without a null hypothesis. Their differences were recently reviewed in [331].

Our view is that the debate should not be closed, but we acknowledge the difficulty in making meaningful contributions to that debate due to the complexity of the problem. Clearly DNA or protein sequence data alone does not suffice to provide a satisfactory answer. Data from different scales of biology – structure, function, biochemical processes, cell morphology etc. as well as the fossil record and earth's environment at different time points have to be applied. Fortunately, in our view, the increasing availability of these data and the tools to manipulate that data promise to keep the debate alive and opinion will continue to see-saw as it has done for the past 33 years since the pioneering work of Woese.

This novel combination of hypotheses on the origin of the Archaea has the strengths of each piece and strengthened their weaknesses greatly. We think Cavalier-Smith has the best method for rooting the tree. His attention to detail and multiple sources of data allows one to refine his ideas as we have done here. Lake (and Gupta) has the right root for the Archaea, and despite our critiques indel polarization is a useful methodology. Gupta has the right idea about antibiotics being a major force in this story and of course his work on indels laid the ground work for our own work as well as Lake's. Forterre is right about viruses being major players in this event. Of course many others have shaped our thoughts on this subject, but we have clearly taken the most from the work of these four. In doing so we have demonstrated the value of using opposing ideas as null hypotheses to each other.

Have we provided a scenario that explains every detail of how the Archaea evolved out of the Gram-positive bacteria? We certainly have not. What we have presented is a variety of data that shows it is a plausible and defensible stance. This would certainly be one of the greatest events in the history of life, so we are not arguing it was easy or simple. But if we close the debate we close our eyes to the large body of evidence that supports the polarization of this transition.

We have provided a novel take on the origin of the Archaea that makes it clear very little is settled on this subject. We have provided a scenario that covers most of the transition between the Bacteria and Archaea. The ideas we propose here can be refined with experiment and more observation. They currently are supported by a very diverse range of data. The study of these hypotheses will give us insight into several tangentially related topics that are worth pursuing. The hypothesis we present and support here reconciles many opposing viewpoints and strongly argues the Archea are derived from the Bacilli.

Figures

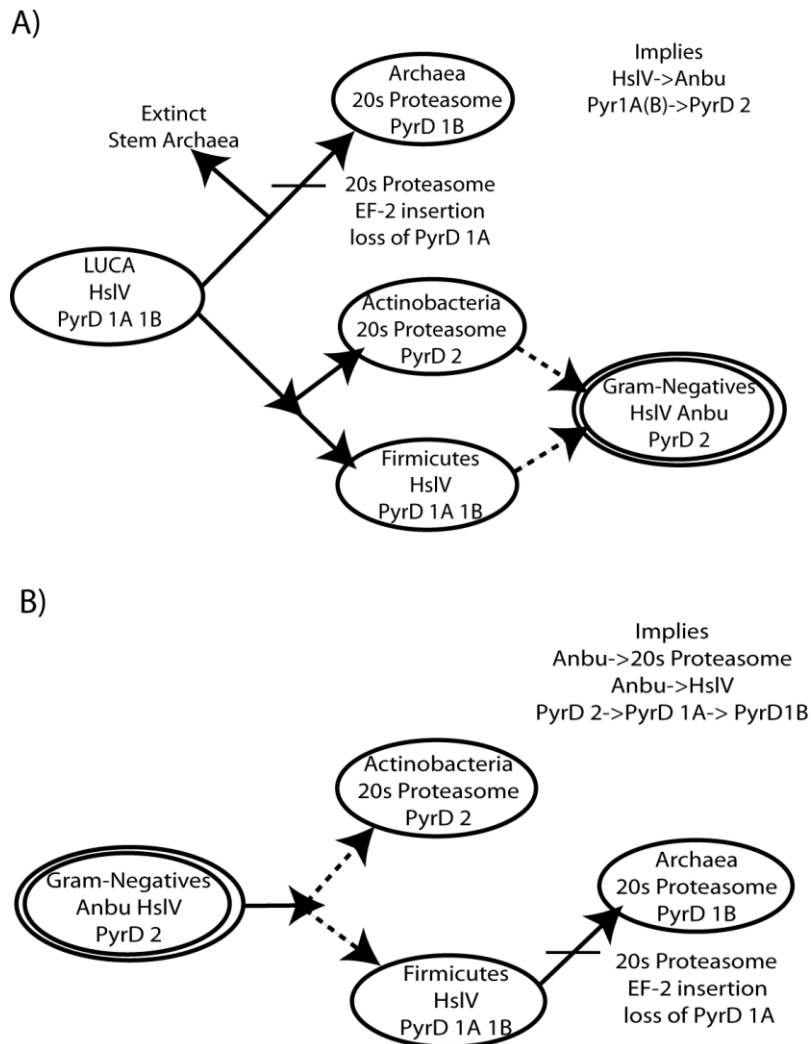


Figure 7-1. Two scenarios for interpreting the three polarizations.

A) Under the canonical rooting proteasome evolution would require several selective sweeps and large scale losses. The monomer PyrD B would have evolved from one of the more complex quaternary structures, and the derived insert in EF-2 would occur after LACA. B) Under the Gram-negative rooting, Anbu could be ancestral to both HsIV and the 20s proteasome. PyrD could evolve via stepwise increases in structural complexity, and there is no need to invoke extinct stem Archaea to explain the EF_G insert. We believe these transitions argue for a Gram-negative rooting.

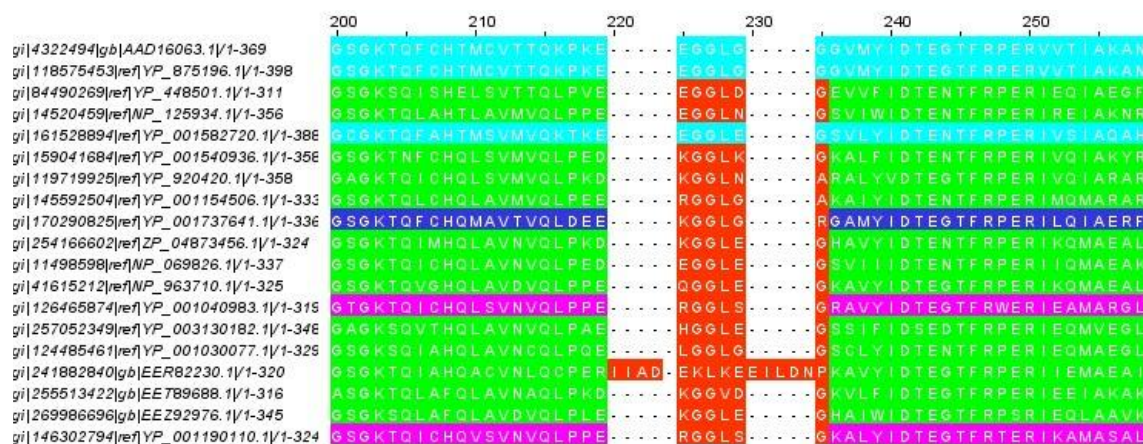


Figure 7-2. Alignment of RadA sequences from representative Archaea.

Euryarchaeal sequences are highlighted in green, Crenarchaea are magenta, Thaumarchaeota are cyan, and Korarchaeota are blue. The region of the indel is highlighted in red. There is no informative indel in this gene as was initially reported.

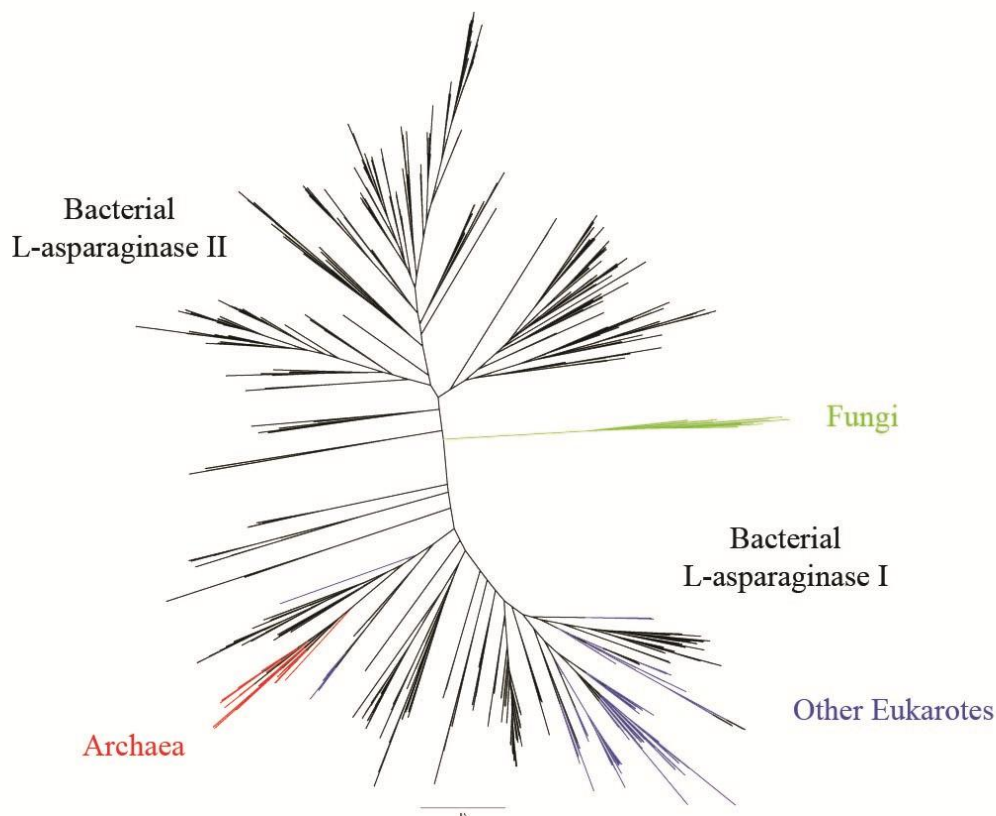


Figure 7-3. Maximum likelihood tree of GatD argues for multiple horizontal transfers.

This tree is not well resolved, but it does not support archaeal ancestry for the eukaryotic proteins.

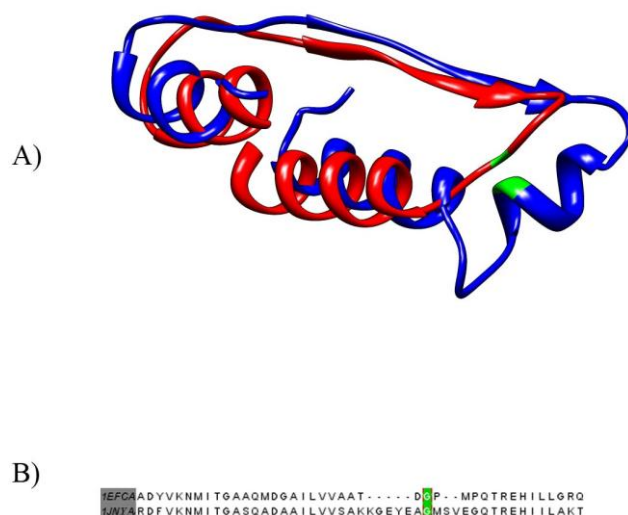


Figure 7-4. Structural alignment of EF-1 and EF-Tu.

The Structural alignment of EF-1 (1JNYA) and EF-Tu (1EFC) in A, and the corresponding sequence alignment in B, show the potential for two independent indels in this region which confounds analysis.

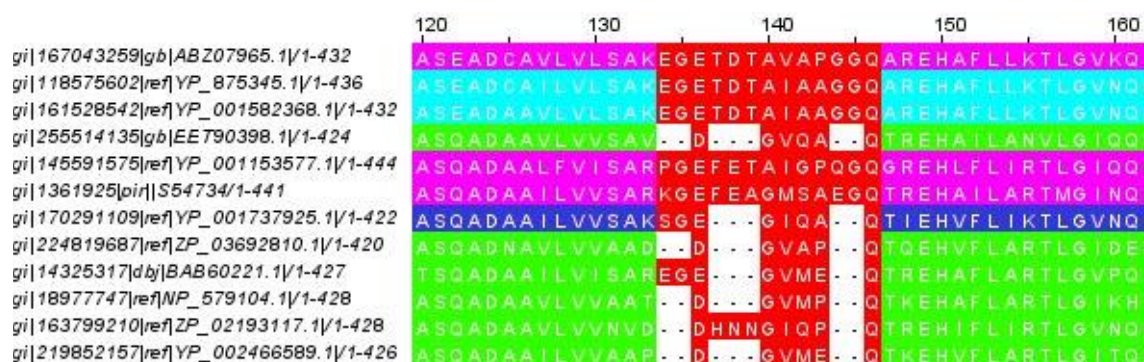


Figure 7-5. Sequence alignment of EF-1 from representative Archaea.

Euryarchaeal sequences are highlighted in green, Crenarchaea are magenta, Thaumarchaea are cyan, and Korarchaeota are blue. The region of the indel is highlighted in red. This alignment implies several reversions. Therefore this indel is not robust enough to determine whether the Archaea are holophyletic or paraphyletic.

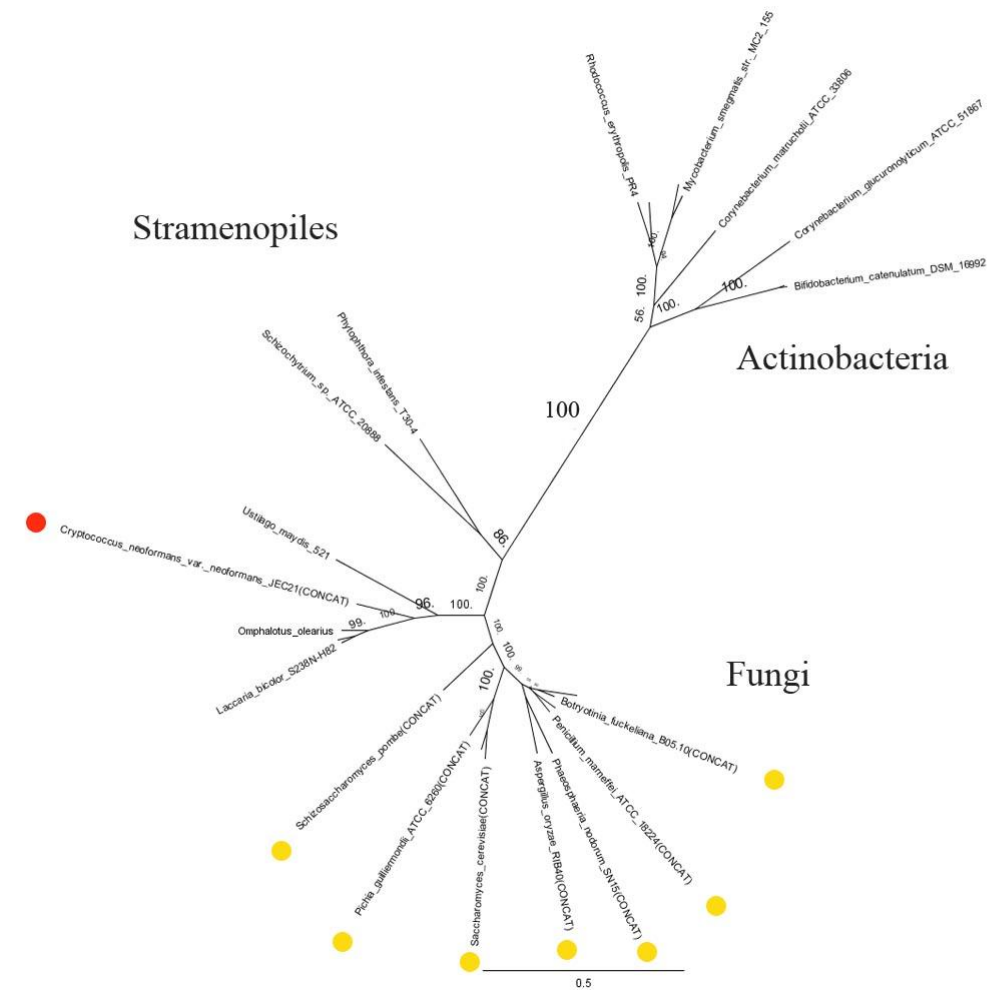


Figure 7-6. Maximum likelihood tree of fungal type Fatty Acid Synthase complex.

This tree implies the Eukaryotes did not get FAS from a recent transfer, but it is also not clear whether or not it was in LECA. Circles indicate the split form of the gene. There gene is split in two different places indicated by the yellow and red circles.

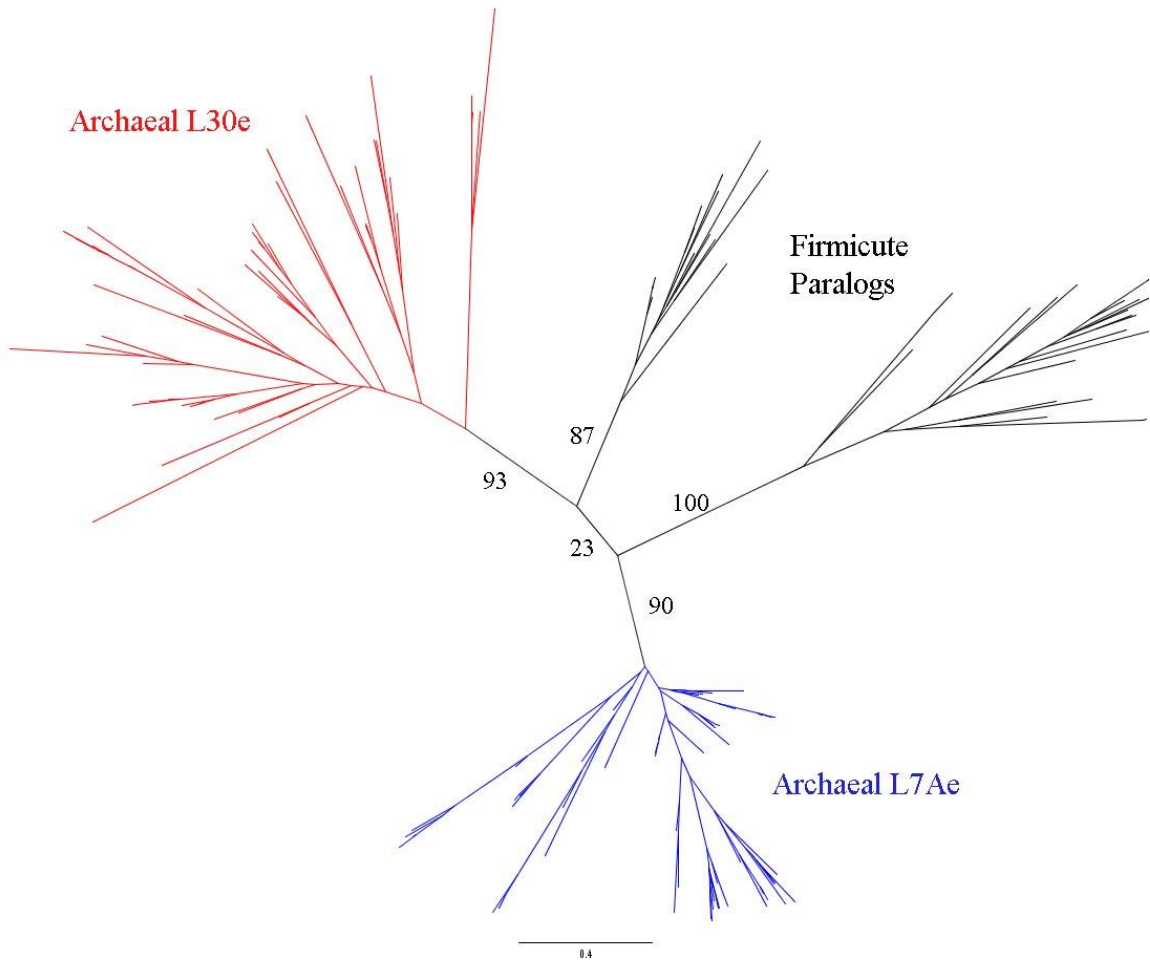


Figure 7-7. Alignment of L7Ae paralogs in Archaea and Firmicutes.

This tree is consistent with a firmicute origin for two archaeal ribosomal proteins.

<i>Petrotoga mobilis</i>	CCCGCAGCUAGACGGAAAGACCCCGUGGA
<i>Thermosipho africanus</i>	CCCGCAGCUAGACGGAAAGACCCCGUGGA
<i>Marinitoga camini</i>	CCCGCAGUUAGACGGAAAGACCCCGUGGA
<i>Fervidobacterium islandicum</i>	CCCGCAGCUAGACGGAAAGACCCCGUGGA
<i>Hydrogenivirga</i> _sp._12B-5-R1-1	CCCGCGGCAGGACGGAAAGACCCCGUGAA
<i>Thermotoga maritima</i> _MSB8	CCCGCGGCUAGACGGAAAGACCCCGUGGA
<i>Hydrogenobaculum</i> _sp._Y04AAS1	CCCACGGCAGGACGGAAAGACCCCGUGAA
<i>Aquifex aeolicus</i> _VF5	CCCGCGGCAGGACGGAAAGACCCCGUGAA

Figure 7-8. 23s rRNA A2058 (*E. coli* numbering) is well conserved across bacterial hyperthermophiles.

This implies the conserved guanine in that position in the Archaea is not an adaptation to thermophily.

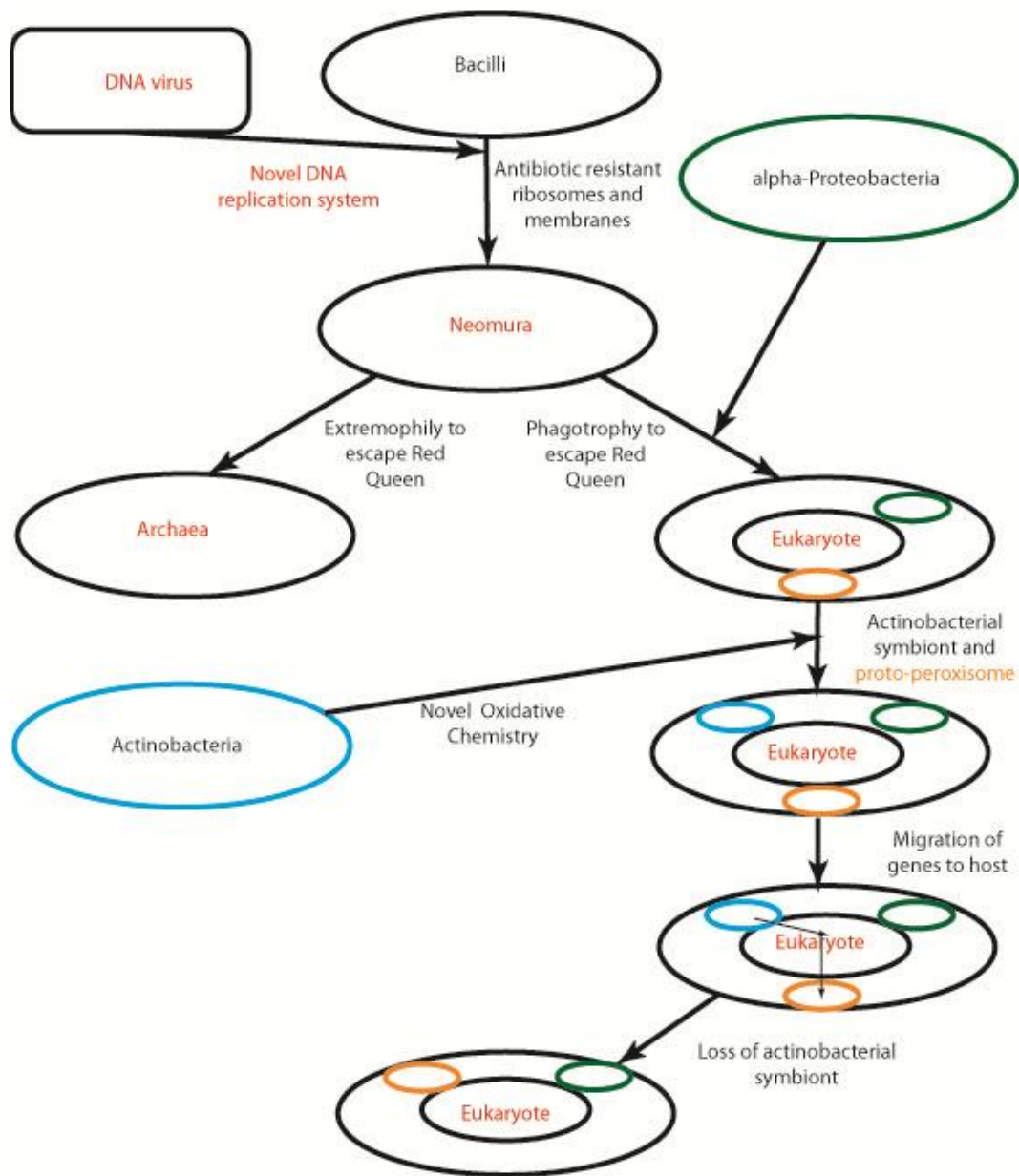


Figure 7-9. Summary of our hypothesis.

A viral endosymbiosis bridges the gap in DNA machinery between the superkingdoms. That triggered an antibiotic war that resulted in the birth of the Eukaryotes and Archaea. The antibiotic war ended when Archaea became extremophiles and the Eukaryotes became phagotrophs. Traits shared between the Eukaryotes and Actinobacteria are the result of an endosymbiosis, the peroxisome is not the direct descendent of an actinobacterium.

Tables

Table 7-1. Essentiality of ribosomal proteins.

The essentiality of proteins that would need to be lost in the transition from an extant bacterial ribosome to an archaeal one varies from species to species. *M. tuberculosis* appears preadapted for the losses that would be necessary in the transition to an archaeon.

Species	Essential ribosomal proteins	P (ribosome is essential)	Essential but absent in Archaea	P- value
<i>Pseudomonas aeruginosa</i> UCBPP-PA14	32	0.5818	8	0.3373
<i>Escherichia coli</i> MG1655	45	0.8182	10	0.055
<i>Bacillus subtilis</i> 168	51	0.9273	14	0.3263
<i>Mycoplasma pulmonis</i> UAB CTIP	47	0.8545	10	0.0204
<i>Francisella novicida</i> U112	49	0.8909	13	0.2505
<i>Helicobacter pylori</i> 26695	29	0.5273	10	0.8493
<i>Mycoplasma genitalium</i> G37	51	0.9273	14	0.3263
<i>Acinetobacter baylyi</i> ADP1	48	0.8727	11	0.0435
<i>Mycobacterium tuberculosis</i> H37Rv	34	0.6182	4	0.0031
<i>Staphylococcus aureus</i> N315	51	0.9273	14	0.3263
<i>Haemophilus influenzae</i> Rd KW20	39	0.7091	9	0.1546

Table 7-2. Summary of differences in DNA replication machinery of Archaea and Bacteria.

The list of protein functions was compiled from box 1 in [242] and table 1 in [332]. Italics indicate a probable horizontal transfer to a superkingdom. There are very few proteins in Archaea that are true innovations. Many of their unique replication proteins could be recruited from bacterial or viral systems. A * indicates the Superfamily database was used to predict domain assignments of PDB entries not yet classified in SCOP.

Function	Superkingdom	Protein	PDB ID (if applicable)	SCOP Superfamily Combination	Proposed Origin in Archaea
Origin recognition	Bacteria	DnaA	1l8q	_gap_52540,48295	
	Archaea	Cdc6/ORC (single or multiple homologues)	1fnn	52540, 46785	evolved from RuvB
Replicative helicase	Bacteria	DnaC	3ec2	52540 *	
	Archaea	MCM	3f9v 1b79 (n-terminal domain)		viral transfer
Helicase loader ssDNA-binding protein	Bacteria	DnaB		gap, 48024	
	Archaea, Bacteria	SSB (one subunit)	1o7i	50249	archaeal SSB evolved from bacterial SSB
	Archaea	RPA (one or three subunits)	2b28	50249, 50249, 50249	evolved from SSB
Primase	Bacteria	DnaG	1dd9	56731	
	Archaea	PriA (small)	1zt2:a	56747, gap	viral transfer or vertically inherited
Replicative Polymerase	Archaea	PriB(large)	1zt2:b	140914	innovation
	Archaea, Bacteria	PolB (one or multiple homologues)	1q8i	gap, 56672	viral transfer

Table 7-2 continued

DNA sliding clamp	Archaea	PolD (small)	none	gap, 56300	innovation
	Archaea	PolD (large)	none	gap	innovation
	Bacteria	PolC (DnaE)	2hqa	89550, large gap*	evolved from β -clamp
	Archaea	PCNA (three subunits)	3a2f	55979, 55979	
	Bacteria	β -clamp (dimer)	1ok7	55979	
Clamp loader		γ -Complex (three subunits)			
	Bacteria	RFC (two subunits)	1jr3:a	52540, 48019	
Removal of primers	Archaea		1iqp	52540, 48019	evolved from γ -Complex
	Archaea	Fen1	1rxw	88723, 47807	viral origin
	Bacteria	RNase H	1jl1	53098	archaeal RNAase HII evolved from bacterial RNAase HI
	Archaea	RNase HII	1eke 2kfz (mssing c-terminal domain)	53098	
Preinitiation complex	Bacteria	PolA DNA		53098, 56672	
	Archaea, Bacteria	ligase(ATP-DEP) DNA	1xn9 (human)	117018, 56091, 50249	viral origin or vertical inheritance
	Bacteria	ligase(NAD-DEP)	1dgs	56091, 50249, 47781	
	Archaea	gins	2eho(human)	158573	true innovation

Table 7-3. Analysis of potentially informative gene structures in Korarchaeota and Thaumarchaeota.

Each indel was analyzed by creating an alignment of archaeal sequences from BLAST searches. We consider these results to be inconclusive until Thaumarchaeota and Korarchaeota are sampled better.

Sequence Property	Initially reported to support	In Korarchaeota	In Thaumarchaea	Now Supports	Notes
Split RpoA	Holophyly	X	X	Holophyly or paraphyly (Thaumarchaeota or Korarchaeota)	
7 AA Insert in EF-1	Paraphyly (Crenarchaea)	2 aa insert shared with thermoplasma	X	Paraphyly, but weakly	Must be reversion in Euryarchaea
6 AA insert in RadA	Paraphyly (Euryarchaea)	?	?	?	There is probably an artifact in the original alignment.
1 AA insert in SecY	Paraphyly (Crenarchaea)	absent	X	Paraphyly (Crenarchaea or Thaumarchaeota)	Single Glycine, but anchors are really nicely conserved
1 AA insert in proAS	Holophyly	X	absent	Probably holophyly, maybe paraphyly (Thaumarchaeota)	BLAST reveals Thaumarchaeota may have HGT from Firmicutes.
7 AA insert in GatD	Holophyly	X	X	Inconclusive	Completely conserved, but its not clear the Eukaryotes inherited

Table 7-3 continued

					this from Archaea
2 AA insert in PBG	Holophyly	Gene is absent	Gene is absent	Inconclusive (bacterial origin)	This protein is present in bacteria, so the Thaumarchaeota and Korarchaeota probably lost it
2 AA insert in ribosomal S12	Holophyly	X	X	Inconclusive	It is conserved across the Eukaryotes and Archaea

Table 7-4. Informative ribosomal proteins in Thaumarchaeota and Korarchaeota.

This table was constructed from [235]. The values listed were taken from searches of the Pfam website.

Ribosomal proteins L20A and L30E were not well defined in Pfam so a BLAST searches were performed instead. These results on their own support the eocyte hypothesis, but it seems quite plausible that there were independent losses of ribosomal subunits in the Archaea based on additional data.

Ribosomal protein	Pfam	Eukaryotes	Bacteria	Crenarchaea	Euryarchaea	Korarchaeota	Thaumarchaeota
L38a	PF01781	182	0	7	0	0	0
L13e	PF01294	202	0	19	0	1	0
S25e	PF03297	195	0	20	0	1	0
S26e	PF01283	189	0	23	1	1	3
S30e	PF04758	179	0	22	0	1	3
L20a		No hits	No hits	Hits	Hits	No hits	No hits
L35ae	PF01247	172	0	5	10	0	0
L14e	PF01929	196	0	16	8	1	0
L34e	PF01199	208	0	21	21	1	0
L30e		No hits	No hits	hits	hits	Hits	No hits

Table 7-5. Summary of data used to support actinobacterial ancestry of the Archaea.

Many of these traits argue for an actinobacterial role in eukaryogenesis but not the origin of the Archaea.

This list of informative characters is taken from [62].

Trait	Supports actinobacterial ancestry of Neomura	Supports actinobacterial Ancestry of Eukaryotes	Exclusive to Actinobacteria among Bacteria
20s proteasome	Yes	Yes	In a few other bacterial genomes
CCA added post- transcriptionally	No	No	Also in Firmicutes
Sterols	No	Yes	In a few other bacterial genomes (HGT), but only ones that could be vertical are actinobacterial
Chitin	No	No	Not in Actinobacteria
Serine/threonine signaling system	No	Yes	In Firmicutes too
Tyrosine Kinase	No	Yes	In many bacteria

H1 linker histone	No	Yes	In many bacteria
Calmodulin	No	Yes	In many bacteria
Phosphatidylinositol	No	Yes	Apparently unique to Actinobacteria, but related enzymes in other groups
Serine proteases	No	Yes	In many bacteria
Primary structure of α -amylase	No	No	In a few other bacterial genomes
Fatty Acid synthetase complex	No	Maybe	Exclusive to actinobacteria
Exospore formation	No	Yes	Exclusive to actinobacteria
HEH domain in Ku protein	No	No	Exclusive to actinobacteria

Table 7-6. Summary of data that supports Bacilli ancestry for the Archaea.

The Bacilli are more similar to the Archaea in terms of DNA repair, ribosome structure, and lipid metabolism than any other group of Bacteria.

Trait	Supports Bacilli as ancestor of Neomura	Supports Bacilli as ancestor of the Archaea	Exclusive to Bacilli among Bacteria
Ribosomal S12 insert	Yes	Yes	Might be also be in Chloroflexi
Geranylgeranylglyceryl Phosphatase	No	Yes	Also in Bacteroides
Mevanolate pathway	Yes	Yes	In a few other bacterial genomes
Ancestors of n-linked glycoprotein	Yes	Yes	Closest hits to eukaryote genes are all from Bacilli
PyrD 1B	No	Yes	In Firmicutes and some Thermotogae
Two L7AE paralogs	Yes	Yes	Exclusive to Firmicutes
RecU	Yes	Yes	Exclusive to Firmicutes

Table 7-7. Pfams that originated near LAECA and their distribution in the viral world.

The Pfams that originated in LAECA are more common in viruses than those that originated in LBCA.

PFAM ID	Description	Archaea	Bacteria	Eukaryotes	Viruses
PF00136	DNA polymerase family B	45	195	31	384
	DNA polymerase family B, exonuclease				
PF03104	domain	45	210	31	173
PF01068	ATP dependent DNA ligase domain	45	372	31	50
PF01096	Transcription factor S-II (TFIIS)	45	1	31	44
PF00867	XPG I-region	43	0	31	37
PF04566	RNA polymerase Rpb2, domain 4	45	0	32	31
PF04567	RNA polymerase Rpb2, domain 5	45	0	32	31
	Eukaryotic and archaeal DNA primase small				
PF01896	subunit	44	159	31	30
PF04675	DNA ligase N terminus	44	167	31	20
	ATP dependent DNA ligase C terminal				
PF04679	region	44	248	31	20
PF00752	XPG N-terminal domain	43	0	31	16
	Proliferating cell nuclear antigen, N-terminal				
PF00705	domain	45	0	33	10
PF01191	RNA polymerase Rpb5, C-terminal domain	44	0	32	5
	Transcription factor TFIID (or TATA-				
PF00352	binding protein, TBP)	45	0	33	4
PF01194	RNA polymerases N / 8 kDa subunit	45	0	30	3
PF01981	Peptidyl-tRNA hydrolase PTH2	45	26	30	3
PF00382	Transcription factor TFIIB repeat	45	0	32	3
	RNA polymerase Rpb7-like, N-terminal				
PF03876	domain	42	0	31	2
PF08542	Replication factor C	44	0	31	2
PF02933	Cell division protein 48 (CDC48), domain 2	45	18	31	1
PF01599	Ribosomal protein S27a	41	0	32	1
	Cell division protein 48 (CDC48), N-				
PF02359	terminal domain	45	44	32	1
PF01873	Domain found in IF2B/IF5	45	0	33	1
PF01253	Translation initiation factor SUI1	45	287	33	1

Table 7-8. Drug targets found across Bacteria that were probably not in LACA.

We argue these proteins were lost in the archaeal ancestor in response to a unique antibiotic warfare scenario. Targets in italics appear to have transferred to the Archaea after LACA.

Target	Example Drug
<i>Alanine Racemase</i>	Cycloserine
<i>Beta-lactamase</i>	Cefoxitin
Hsp90	Geldanamycin
<i>Cytochrome c</i>	Minocycline
<i>DNA gyrase</i>	Trovafloxacin
<i>3-oxoacyl-[acyl-carrier-protein] synthase 1</i>	Cerulenin
<i>Penicillin-binding proteins</i>	Cefoperazone
Peptidoglycan synthetase ftsI	Ertapenem

Table 7-9. Examples of drugs targets sites with resistance in Archaea.

These drugs bind targets sites present in both Bacteria and Archaea (or Eukaryotes), but with very different affinities. We argue this is a molecular fossil of the unique antibiotic war that resulted in the origin of the Archaea.

Target Site	Example drugs
23s RNA between L22 and L4	Macrolides[316] and PTF inhibitors [315]
RNA Polymerase	GE23077[333]
EF-Tu	Kirromycin [334]

Chapter 8: The evolution of amino acids: better, faster, cheaper

Introduction

The genetic code is LUCA's greatest legacy to us. The code is conserved in all living cells with only minor exceptions. But as I've already stressed LUCA was not the first cell, and the genetic code and amino acid (AA) alphabet in LUCA were certainly not the first either. The complex machinery needed to make the current code must have evolved from a simpler predecessor. There are numerous theories on the origin of the genetic code (reviewed in [335]). Each of these provides a unique ordering of when codons and AAs were entered the code.

It does not make sense to code for an AA that does not exist, so presence of AAs is one constraint that can be placed on the code's history. Numerous prebiotic soup experiments have produced amino acids, and they have also been found in several meteorite samples (reviewed in [336]). The conditions used in each of these experiments vary greatly. Despite this, there is strong overlap in terms of which AAs are produced even though their ranks in terms of abundance vary. There are several AAs that have never been abiotically synthesized under any of these conditions, so the code must have gotten started without them. Many AAs are precursors to others in the extant pathways. Trifinov has constructed a consensus order of the appearance of the AAs based on numerous lines of reasoning including prebiotic availability, biosynthetic pathways, and codon preference in the extant code [337].

We will probably never have a precise order for the appearance of AAs or know for sure whether some have been displaced and are no longer in the code. But it is clear there is some consensus in the data give or take a rank. If the code and alphabet developed in stages, is it possible to find molecular fossils created during the intermediate periods and hence shed light on evolution before LUCA?

Another interesting question on the origin of the code is why are these 20 AAs used by the code [338]? Optimization of the code itself has recently been examined by trying to maximize redundancy and diversity while minimizing cost[339]. Are there other alphabets that could be just as useful and need very different codes?

Old AAs conserved in extant proteins

Since some AAs are older than others it could be possible to find proteins that originated before the AA alphabet was finalized. A comparison between the ribosome (which evolved before LUCA) and spliceosome (which evolved before LECA) showed a statistically significant enrichment for older amino acids in 100% conserved sites [147]. Any site that is 100% conserved can safely be assumed to be under very strong pressure since LUCA. Their result makes sense since parts of the ribosome should predate the finalized code.

I pursued this idea further by counting the portion of prebiotic conserved sites in every alignment in the protein clusters database [222]. GroEl stands out as an ancient

protein. 23 of its 26 100% conserved sites are prebiotic amino acids. Its role as a chaperone could have been vital to primitive alphabets would not fold as well as the modern one. However, no other protein had such a clear signal. It is more difficult to do this with single proteins as opposed to large complexes because there are so few 100% conserved sites. It may be possible to extend this result to larger evolutionary units such as metabolic pathways.

The ribosome probably coevolved with the code, so it would make sense that the older parts (proteins) of the ribosome use older AAs than the younger parts. But how can we date the parts of the ribosome? I propose that any ribosomal protein that does not directly contact the rRNA must be younger than the interface protein between it and the rRNA. This idea is supported by recent work that has described evolution of the ribosome in terms of adding layers to an onion [340]. Therefore, in general the proteins that directly touch the rRNA should be younger than those do not. We should expect them to have more conserved prebiotic amino acids. It turns out this is the case (Table 8-1), but the p-value was only .07.

The problem is probably that I bin both proteins and AAs into too coarse a view of old and young, but a better age system for either of them may not help. A finer grained view would mean less 100% conserved sites, which would make it very difficult to get statistically significant results. When I rank the AAs according to Trifinov's ordering there is no clear distinction between how the younger and older parts of the ribosome use most AAs with the exceptions of proline and lysine (Figure 8-1).

Since there are no proteins composed of only prebiotic AAs, I searched for catalytic sites that only contain prebiotic AAs. There are many catalytic sites that can be

made from only prebiotic AAs, which led me to ask whether some entire pathways could be made using older amino acids. If pathways existed that only used prebiotic AAs, it would be evidence that pathway originated before the code was frozen.

I created a mapping of the catalytic site atlas (CSA)[341] onto KEGG [342]. The CSA curates PDB structures' active residues. I started by mapping every PDB structure onto the KEGG metabolic pathways, and then mapped the site data onto those pathways. The module for lysine biosynthesis is shown below (Figure 8-2). Not only did this pathway use many biosynthesized AAs, it also uses lysine in five different catalytic sites. I was unable to find any pathways that only used prebiotic AAs in their catalytic sites. I expected that catalytic sites used to make an AA would not use that AA or one that should have evolved later. However, almost every pathway used to synthesize a specific AA uses that AA somewhere in a catalytic site. This is interesting because it guarantees that at least some of these enzymes cannot be the initial machine that perform that specific reaction. That makes sense since AA metabolism should have been initially performed by ribozymes that were later replaced by enzymes. The question that I cannot answer is whether an early enzyme was ever replaced with a younger one. I presume this would be due to convergent evolution of function, as tinkering with the existing catalytic site could be highly deleterious. However, a novel enzyme with the same functionality would not be deleterious and could eventually outcompete the older one.

It seems there are hints of the evolution of the AA alphabet in some extant proteins, but it is very hard to say anything with certainty beyond the fact the abiotic amino acids predate biosynthetic ones. The last AAs to enter the code were superior to their

predecessor in many proteins, so there is no record of that initial alphabets used to create many of the early folds.

Why this Alphabet?

So far I have explored ideas on the order of the appearance of 20 AAs, but we can also ask why there are 20 AAs, and why they are these 20? The evolution of the genetic code is a balancing act of diversity and redundancy. Amino acids with different biochemistry are the most useful for expansion of the alphabet, but come with the highest cost in terms of mistranslation. If all codons encoded the same amino acid there would be no translational errors, but proteins would be extremely limited in their functionality. If all the codons coded for a different amino acids the functionality of proteins would be higher but there would a high cost to mistranslation.

An initial exploration of the biochemical diversity of possible alphabets was recently preformed to test the hypothesis the AA alphabet we under pressure to maximize biochemical diversity [343]. The authors generated a set of plausible AAs from prebiotic sources (Murchison meteorite) or AAs that are synthesized in extant organisms but are not part of the code. They computed values of hydrophobicity, volume, and charge for each of the 92 AAs in their test set as measures of biochemical diversity. Variance was used to measure diversity for each variable. They then permuted random alphabets out of the 92 AAs to calculate the odds of getting an alphabet that is as diverse as the standard alphabet for each of these categories.

In general their data supports the idea that the alphabet was under some selective pressure to maximize biochemical diversity. For instance they found that random

alphabets of 20 AAs drawn from the abiotic AAs always have less variance in volume than the standard alphabet. Their results are often improved by invoking a weighting scheme in the randomization step. They weight prebiotic amino acids by their abundance in samples from the Murchison meteorite, assuming it's easier to incorporate an abundant AA into a code than a rare one. For instance only 38% of random alphabets drawn from the abiotic set have less variance in hydrophobicity than the standard code. This results improves 99% when the abiotic AAs are weighted by abundance. The authors admit that this is not ideal as different conditions on prebiotic earth could greatly shift these abundances. Although abundance probably shaped the alphabet I wanted refrain from using it due to this uncertainty.

Their results do not adequately explain why the eight prebiotic amino acids (from the meteorite sample) present in the standard alphabet were selected for. For instance about only 33% of random eight letter alphabets had less diversity in terms of hydrophobicity than those eight amino acids that made it into the code. Over half the alphabets had more diversity in terms of volume when abundance weighting was used. This suggests that there must be some important selective factor besides biochemical diversity.

I think one additional constraint on the alphabet beside diversity is redundancy. The genetic code is highly optimized to minimize errors. Some of this is due to redundancy between the amino acids. Leucine and isoleucine are highly interchangeable but have slightly different probabilities of transitioning to other amino acids. A code that includes both leucine and isoleucine gains very little in diversity, but has a much larger neutral space. Redundancy is necessary for adding new amino acids into the alphabet. It

allows for a stage where a single tRNA is ambiguous for both AAs to be near neutral. This means that the alphabet should maximize the overall distance between all amino acids while minimizing the distances between nearest neighbors to maximize diversity as well as redundancy.

Efficiency is also important. Every amino acid has a metabolic cost. It has been shown that proteins are optimized to use minimal energy by having cheaper amino acids in more highly expressed proteins [47] (some of their published values turn out to be wrong thanks to a bug I spotted). Likewise, any alphabet is also going to be optimized to work as cheaply as possible. Since many of the amino acids in the test set are not biosynthesized, we cannot use the energy of their pathway as estimate of their cost as in [47]. Instead I make the simplifying assumption that the mass of each amino acid is proportional to the energy used to make it. This assumption has been justified in [344] because all the matter incorporated into the amino acid required energy at some point in the pathway. Mass is constant across species, while biochemical cost can vary based on growth conditions. Early abundance is closely related to energy because the cost of incorporating a highly abundant amino acid into the code is zero. Even if we can estimate the initial abundance of each amino acid there would still be no way of measuring abundance at the birth of the genetic code because those amino acids could have been used by life before that stage. One must look at the energy efficiency of the standard alphabet compared to random alphabets in addition to diversity and redundancy.

The obvious need to consider all three of these factors can be seen if we consider the most extreme alphabets possible. An alphabet that is only glycine contains no biochemical diversity, but it is also impossible for there to be a translation error. In this case

it would also be very efficient. An alphabet of 63 amino acids would have a tremendous amount of diversity, but any translational error or mutation would result in a change in protein sequence. If the 63 amino acids were diverse as possible there would be very few changes that could be tolerated, and many of the AAs would be energetically costly. Clearly the standard alphabet has balanced these factors. If two alphabets have the same diversity and redundancy the cheaper of them would win out. The optimal alphabet is one that maximizes biochemical diversity without loss of redundancy using the cheapest amino acids possible.

There is also a downside to measuring one aspect of diversity at a time. It is possible to create an alphabet where there is extreme variance for charge, but all the AAs have the same mass. That alphabet would not be diverse at all, but it would get ranked as such in a single variable ranking. To get around this problem I compared all three measures at biochemical diversity at once. These three variables can be considered as three dimensions in space. Figures 3-5 are 2-d plots of the three dimensions of chemical variance. One loses a lot of information if the variance is measured one dimension at a time. The distance between AAs was defined using both Euclidean and Manhattan distances which did not seem to greatly effect the results (data not shown).

In each case I took a fixed set of AAs and compared them to 10000 random ones of the same size as done in the initial study. Diversity was measured in terms of average distance. Redundancy was measured as the distance to the nearest neighbor (the one that would be least costly to replace it with). The minimization of this distance ensures that each AA has at least one neighbor in the genetic code table that would be a near neutral

mistranslation error. Energy was defined in terms of mass. In each case the standard alphabet is optimized more than would be expected by chance (

Table 8-2). The eight prebiotic amino acids in the standard alphabet are highly optimized in terms of diversity, redundancy, and efficiency. Only about 6% of random alphabets use less energy (mass), so this has clearly been optimized with respect to the initial alphabet. This result was not seen directly in the initial study, which explains why their results do not explain the initial alphabet. However, the lighter AAs are also among the most abundant, so their weighting scheme indirectly captures it.

The standard alphabet is more optimized than 97.2% of random alphabets with regards to diversity, redundancy, and efficiency. 75% of those alphabets are more diverse, which shows the other two factors I have introduced are important. For the reasons discussed above the benefits of diversity must balance its cost.

I also built alphabets where I kept those eight prebiotic AAs fixed to explore how biosynthetic AAs were added on top of the prebiotic ones. The standard alphabet is more optimized than 92% of all alphabets that have those eight AAs. Again the result demonstrated the standard alphabet is highly optimized in terms of these three factors. All three factors would be important throughout the expansion of the alphabet.

Discussion

The major limitation of this work is my definition of cost. Right now I cannot differentiate between isomers. The relative free energies of each of the standard AAs and their isomers has been calculated experimentally [345]. In each case the isomers used in

the standard alphabet is close to the lowest free energy. If we combine this observation with my data it implies the alphabet is heavily optimized in terms of energetic cost. Unfortunately this data was lost (lost hard drive) so I was unable to incorporate it into my simulations (Zhang, personal communication). However, I think it is safe to conclude that the alphabet evolved under the motto “better, faster, and cheaper”.

None of the measures I used here are dependent on environment, so it seems reasonable to speculate about general properties of amino acid codes. Glycine and alanine are present in relatively high abundance in many of the abiotic samples [336], so it is not a stretch of the imagination to say that alien life would probably use those amino acids. They are just too cheap and useful not to. It seems more reasonable to me that aliens would use proteins than they would DNA. Proteins are written in such a wonderfully rich alphabet that they can do almost anything. We’ve only seen a very tiny portion of sequence space sampled by life in four billions years on this planet. Any genetic material that holds the information for those blueprints would do. The code could be very different, but it would be surprising if it encoded much more than 20 amino acids. There just aren’t that many amino acids that would be useful enough additions after a certain amount of chemical biodiversity in the alphabet is reached. At that point it is probably safer to add additional diversity through posttranslational modifications in the special cases where it is needed. No matter how alien of life we might find we should still expect what ever they use for function to be diverse, redundant, and cheap.

Figures

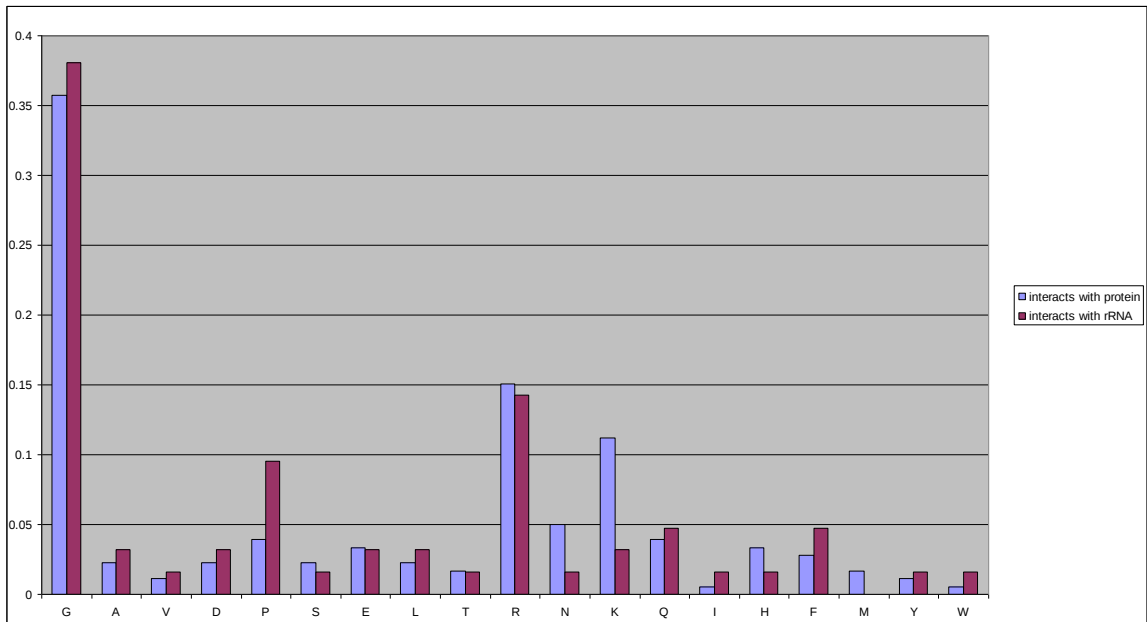


Figure 8-1. Normalized comparison of AA usage between old and young ribosomal proteins.

Only sites that 100% conserved were used. The x-axis is ordered based on Trifonov consensus ordering for the appearance of AAs.

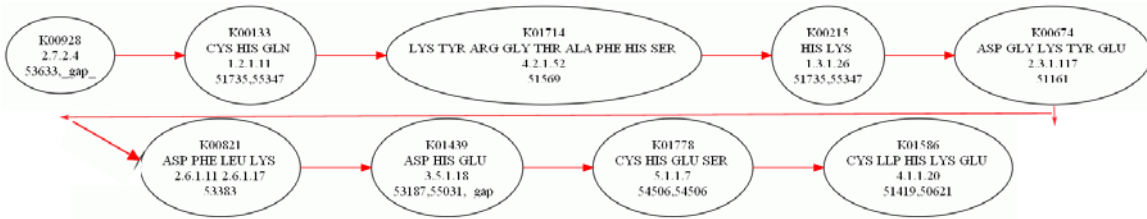


Figure 8-2. Catalytic sites of lysine biosynthesis module.

Each node corresponds to an enzyme in the pathway. The labels are from top to bottom KEGG's cluster identifier, the AAs used in the catalytic site, the EC number of that reaction, and the SCOP combination for that structure.

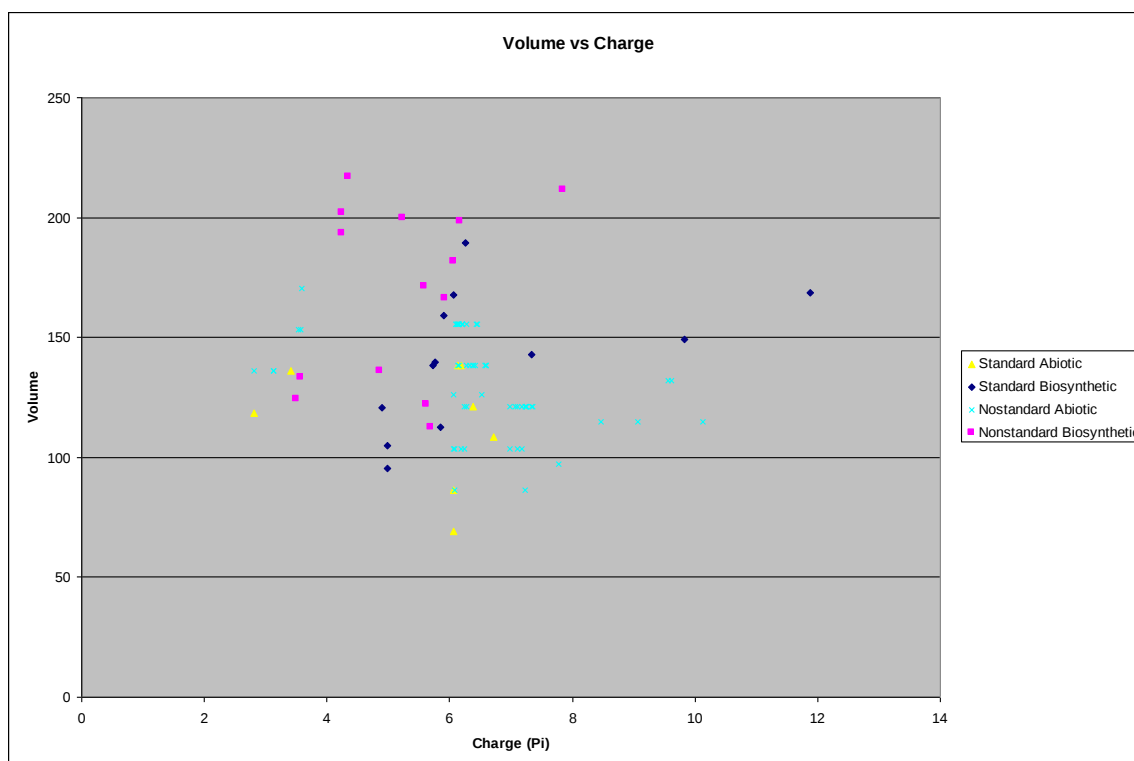


Figure 8-3. Plot of Van der Waals volume vs charge of standard and possible amino acids.

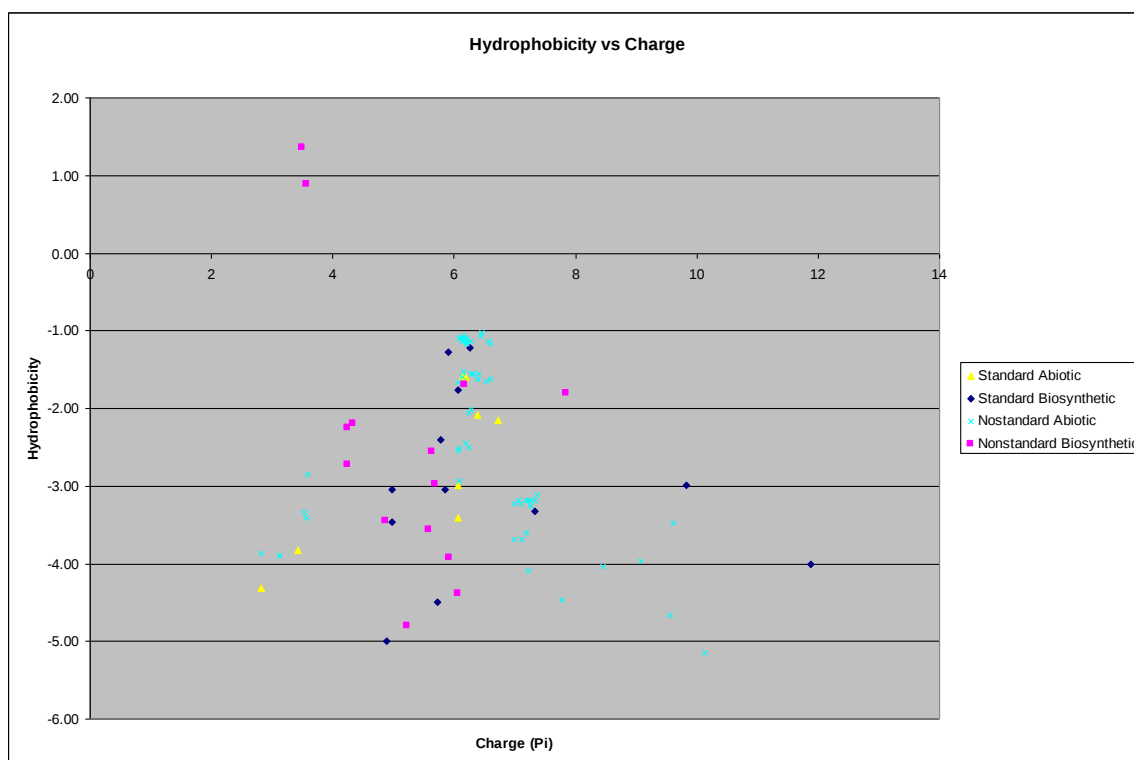


Figure 8-4. Plot of hydrophobicity vs charge of standard and possible amino acids.

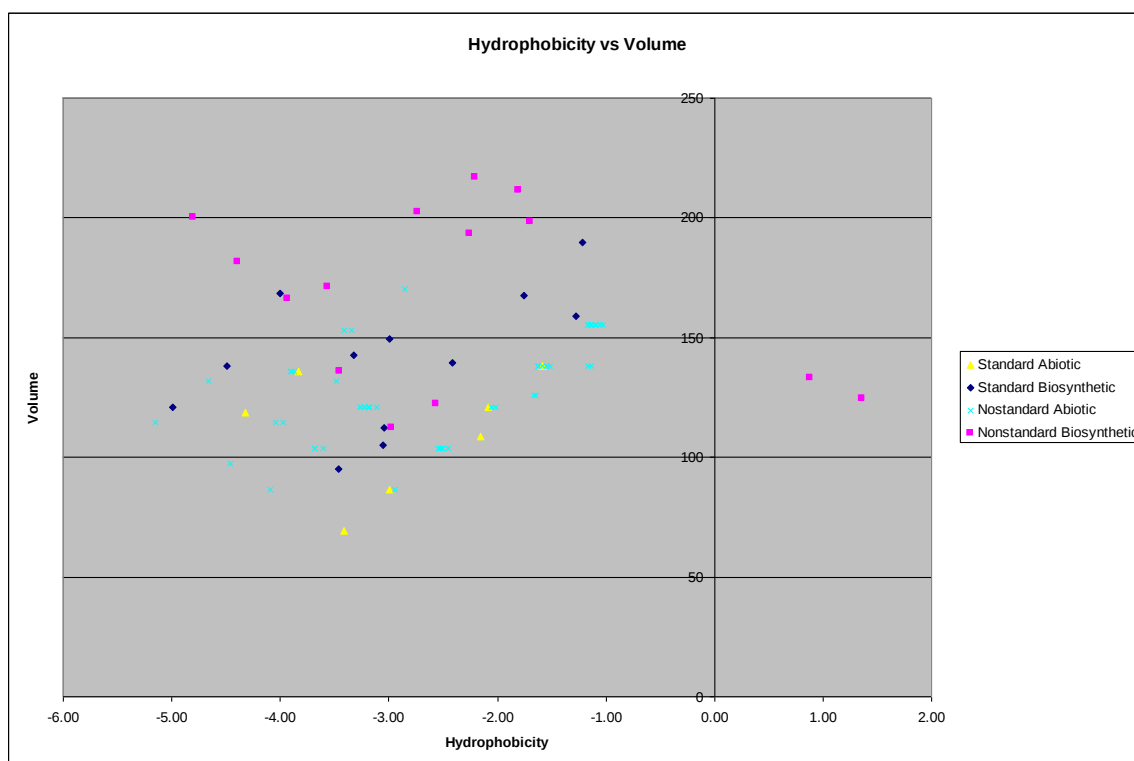


Figure 8-5. Plot of hydrophobicity vs volume of standard and possible amino acids.

Tables

Table 8-1. Comparison of prebiotic amino acid use in old and young ribosomal proteins.

The proteins that directly contact the rRNA use prebiotic amino acids a little more frequently than those that do not.

	Prebiotic sites	total 100% conserved sites	% prebiotic
interacts with rRNA	42	63	66.7%
interacts with protein	99	179	55.3%

Table 8-2. Comparison of standard AA alphabet to 10000 random alphabets.

The standard alphabet is highly optimized in terms of diversity, redundancy, and efficiency.

Fixed set	Random sets drawn from	Alphabets with more diversity	Alphabets with less redundancy	Alphabets that use less energy	Alphabets that are better at all three	Expect alphabets better at all three
8 used prebiotic AAs	66 possible prebiotic AAs	3803	2618	616	5	6.133
20 standard AAs	92 possible AAs	2597	7900	4288	283	819.74
20 standard AAs	Keep 8 prebiotic fixed and select rest from 86 possible AAs	1940	8882	8684	755	1496.34

References

1. Darwin, E., *Zoonomia; or, The laws of organic life*. 1794, Dublin: P. Byrne, and W. Jones. 2 v.
2. Darwin, C., *On the origin of species by means of natural selection*. 1859, London: J. Murray. ix, [1], 502 p.
3. Ragan, M.A., *Trees and networks before and after Darwin*. Biol Direct, 2009. **4**: p. 43; discussion 43.
4. Baptiste, E., et al., *Prokaryotic evolution and the tree of life are two different things*. Biol Direct, 2009. **4**(1): p. 34.
5. Doolittle, W.F. and E. Baptiste, *Pattern pluralism and the Tree of Life hypothesis*. Proc Natl Acad Sci U S A, 2007. **104**(7): p. 2043-9.
6. Raff, R.A., *Written in stone: fossils, genes and evo-devo*. Nat Rev Genet, 2007. **8**(12): p. 911-20.
7. Rosing, M.T., *¹³C-Depleted carbon microparticles in >3700-Ma sea-floor sedimentary rocks from west greenland*. Science, 1999. **283**(5402): p. 674-6.
8. Furnes, H., et al., *Early life recorded in archean pillow lavas*. Science, 2004. **304**(5670): p. 578-81.
9. Allwood, A.C., et al., *Stromatolite reef from the Early Archaean era of Australia*. Nature, 2006. **441**(7094): p. 714-8.
10. Lowe, D.R., *Abiological origin of described stromatolites older than 3.2 Ga*. Geology, 1994. **22**: p. 387-90.
11. Buick, R., *When did oxygenic photosynthesis evolve?* Philos Trans R Soc Lond B Biol Sci, 2008. **363**(1504): p. 2731-43.
12. Blank, C.E., *Phylogenomic dating--a method of constraining the age of microbial taxa that lack a conventional fossil record*. Astrobiology, 2009. **9**(2): p. 173-91.
13. Stanier, R.Y. and C.B. Van Niel, *The concept of a bacterium*. Arch Mikrobiol, 1962. **42**: p. 17-35.
14. Zuckerkandl, E. and L. Pauling, *Molecules as documents of evolutionary history*. J Theor Biol, 1965. **8**(2): p. 357-66.
15. Woese, C.R. and G.E. Fox, *Phylogenetic structure of the prokaryotic domain: the primary kingdoms*. Proc Natl Acad Sci U S A, 1977. **74**(11): p. 5088-90.

16. Woese, C.R., *The Archaeal Concept and the World it Lives in: A Retrospective*. Photosynth Res, 2004. **80**(1-3): p. 361-72.
17. Fuerst, J.A., *Intracellular compartmentation in planctomycetes*. Annu Rev Microbiol, 2005. **59**: p. 299-328.
18. Howe, C.J., *Cellular evolution: what's in a mitochondrion?* Curr Biol, 2008. **18**(10): p. R429-31.
19. Embley, T.M. and W. Martin, *Eukaryotic evolution, changes and challenges*. Nature, 2006. **440**(7084): p. 623-30.
20. Lake, J.A., et al., *Eocytes: a new ribosome structure indicates a kingdom with a close relationship to eukaryotes*. Proc Natl Acad Sci U S A, 1984. **81**(12): p. 3786-90.
21. Cox, C.J., et al., *The archaeobacterial origin of eukaryotes*. Proc Natl Acad Sci U S A, 2008. **105**(51): p. 20356-61.
22. Mushegian, A., *Gene content of LUCA, the last universal common ancestor*. Front Biosci, 2008. **13**: p. 4657-66.
23. Koonin, E.V. and W. Martin, *On the origin of genomes and cells within inorganic compartments*. Trends Genet, 2005. **21**(12): p. 647 - 654.
24. Mulkidjanian, A.Y., M.Y. Galperin, and E.V. Koonin, *Co-evolution of primordial membranes and membrane proteins*. Trends Biochem Sci, 2009. **34**(4): p. 206-15.
25. Forterre, P., *The two ages of the RNA world, and the transition to the DNA world: a story of viruses and cells*. Biochimie, 2005. **87**(9-10): p. 793-803.
26. Wong, J.T., et al., *Polyphasic evidence delineating the root of life and roots of biological domains*. Gene, 2007. **403**(1-2): p. 39-52.
27. Lake, J.A., et al., *Evidence for a new root of the tree of life*. Syst Biol, 2008. **57**(6): p. 835-43.
28. Cavalier-Smith, T., *Rooting the tree of life by transition analyses*. Biol Direct, 2006. **1**: p. 19.
29. Kurland, C.G., L.J. Collins, and D. Penny, *Genomics and the irreducible nature of eukaryote cells*. Science, 2006. **312**(5776): p. 1011-4.
30. Iwabe, N., et al., *Evolutionary relationship of archaeobacteria, eubacteria, and eukaryotes inferred from phylogenetic trees of duplicated genes*. Proc Natl Acad Sci U S A, 1989. **86**(23): p. 9355-9.

31. Gogarten, J.P., et al., *Evolution of the vacuolar H⁺-ATPase: implications for the origin of eukaryotes*. Proc Natl Acad Sci U S A, 1989. **86**(17): p. 6661-5.
32. Zhaxybayeva, O., P. Lapierre, and J.P. Gogarten, *Ancient gene duplications and the root(s) of the tree of life*. Protoplasma, 2005. **227**(1): p. 53-64.
33. Saier, M., *Structure and Evolution of Prokaryotic Cell Envelopes*, in *Microbe magazine*. 2008.
34. Skophammer, R.G., et al., *Evidence for a gram-positive, eubacterial root of the tree of life*. Mol Biol Evol, 2007. **24**(8): p. 1761-8.
35. Skophammer, R.G., et al., *Evidence that the root of the tree of life is not within the Archaea*. Mol Biol Evol, 2006. **23**(9): p. 1648-51.
36. Servin, J.A., et al., *Evidence excluding the root of the tree of life from the actinobacteria*. Mol Biol Evol, 2008. **25**(1): p. 1-4.
37. Lake, J.A., et al., *Rooting the tree of life using nonubiquitous genes*. Mol Biol Evol, 2007. **24**(1): p. 130-6.
38. Valas, R.E. and P.E. Bourne, *Structural analysis of polarizing indels: an emerging consensus on the root of the tree of life*. Biol Direct, 2009. **4**: p. 30.
39. Levy, E.D., et al., *Assembly reflects evolution of protein complexes*. Nature, 2008. **453**(7199): p. 1262-5.
40. Gribaldo, S. and C. Brochier, *Phylogeny of prokaryotes: does it exist and why should we care?* Res Microbiol, 2009.
41. Virchow, R.L.K., *Die cellularpathologie in ihrer begründung auf physiologische und pathologische gewebelehre*. 1858, Berlin: A. Hirschwald. xvi, 440 p.
42. Ciccarelli, F.D., et al., *Toward automatic reconstruction of a highly resolved tree of life*. Science, 2006. **311**(5765): p. 1283-7.
43. Dagan, T. and W. Martin, *The tree of one percent*. Genome Biol, 2006. **7**(10): p. 118.
44. Dagan, T., Y. Artzy-Randrup, and W. Martin, *Modular networks and cumulative impact of lateral transfer in prokaryote genome evolution*. Proc Natl Acad Sci U S A, 2008. **105**(29): p. 10039-44.
45. Ishihama, Y., et al., *Protein abundance profiling of the Escherichia coli cytosol*. BMC Genomics, 2008. **9**: p. 102.

46. Malmstrom, J., et al., *Proteome-wide cellular protein concentrations of the human pathogen Leptospira interrogans*. Nature, 2009. **460**(7256): p. 762-5.
47. Akashi, H. and T. Gojobori, *Metabolic efficiency and amino acid composition in the proteomes of Escherichia coli and Bacillus subtilis*. Proc Natl Acad Sci U S A, 2002. **99**(6): p. 3695-700.
48. Baptiste, E., et al., *Alternative methods for concatenation of core genes indicate a lack of resolution in deep nodes of the prokaryotic phylogeny*. Mol Biol Evol, 2008. **25**(1): p. 83-91.
49. Galtier, N. and V. Daubin, *Dealing with incongruence in phylogenomic analyses*. Philos Trans R Soc Lond B Biol Sci, 2008. **363**(1512): p. 4023-9.
50. Doolittle, R.F. and J. Handy, *Evolutionary anomalies among the aminoacyl-tRNA synthetases*. Curr Opin Genet Dev, 1998. **8**(6): p. 630-6.
51. Zaneveld, J.R., D.R. Nemergut, and R. Knight, *Are all horizontal gene transfers created equal? Prospects for mechanism-based studies of HGT patterns*. Microbiology, 2008. **154**(Pt 1): p. 1-15.
52. Omelchenko, M.V., et al., *Non-homologous isofunctional enzymes: a systematic analysis of alternative solutions in enzyme evolution*. Biol Direct, 2010. **5**: p. 31.
53. Wolf, Y.I., et al., *Inaugural Article: The universal distribution of evolutionary rates of genes and distinct characteristics of eukaryotic genes of different apparent ages*. Proc Natl Acad Sci U S A, 2009. **106**(18): p. 7273-80.
54. Uchiyama, I., *Multiple genome alignment for identifying the core structure among moderately related microbial genomes*. BMC Genomics, 2008. **9**: p. 515.
55. Seshadri, R., et al., *Comparison of the genome of the oral pathogen Treponema denticola with other spirochete genomes*. Proc Natl Acad Sci U S A, 2004. **101**(15): p. 5646-51.
56. von Mering, C., et al., *STRING: a database of predicted functional associations between proteins*. Nucleic Acids Res, 2003. **31**(1): p. 258-61.
57. Puigbo, P., Y.I. Wolf, and E.V. Koonin, *Search for a 'Tree of Life' in the thicket of the phylogenetic forest*. J Biol, 2009. **8**(6): p. 59.
58. Beiko, R.G., T.J. Harlow, and M.A. Ragan, *Highways of gene sharing in prokaryotes*. Proc Natl Acad Sci U S A, 2005. **102**(40): p. 14332-7.
59. Koonin, E.V. and Y.I. Wolf, *The fundamental units, processes and patterns of evolution, and the Tree of Life conundrum*. Biol Direct, 2009. **4**(1): p. 33.

60. Rokas, A. and S.B. Carroll, *Bushes in the tree of life*. PLoS Biol, 2006. **4**(11): p. e352.
61. Koonin, E.V., *The Biological Big Bang model for the major transitions in evolution*. Biol Direct, 2007. **2**: p. 21.
62. Cavalier-Smith, T., *The neomuran origin of archaeobacteria, the negibacterial root of the universal tree and bacterial megaclassification*. Int J Syst Evol Microbiol, 2002. **52**(Pt 1): p. 7-76.
63. Bloch, K.E., *Speculations on the evolution of sterol structure and function*. CRC Crit Rev Biochem, 1979. **7**(1): p. 1-5.
64. Bekker, A., et al., *Dating the rise of atmospheric oxygen*. Nature, 2004. **427**(6970): p. 117-20.
65. Gupta, R.S. and E. Griffiths, *Critical issues in bacterial phylogeny*. Theor Popul Biol, 2002. **61**(4): p. 423-34.
66. Simpson, G.G., *Tempo and mode in evolution*. Columbia biological series; no. 15. 1944, New York: Columbia University Press. xviii, 237 p.
67. Ota, T., *Population size and rate of evolution*. J Mol Evol, 1972. **1**(3): p. 305-14.
68. Cordero, O.X. and P. Hogeweg, *Large changes in regulome size herald the main prokaryotic lineages*. Trends Genet, 2007. **23**(10): p. 488-93.
69. Drummond, D.A., et al., *Why highly expressed proteins evolve slowly*. Proc Natl Acad Sci U S A, 2005. **102**(40): p. 14338-43.
70. Kuo, C.H. and H. Ochman, *Inferring clocks when lacking rocks: the variable rates of molecular evolution in bacteria*. Biol Direct, 2009. **4**(1): p. 35.
71. Valas, R.E. and P.E. Bourne, *Rethinking proteasome evolution: two novel bacterial proteasomes*. J Mol Evol, 2008. **66**(5): p. 494-504.
72. Wilson, D., et al., *The SUPERFAMILY database in 2007: families and functions*. Nucleic Acids Res, 2007. **35**(Database issue): p. D308-13.
73. Isambert, H. and R.R. Stein, *On the need for widespread horizontal gene transfers under genome size constraint*. Biol Direct, 2009. **4**: p. 28.
74. Kurtak, K., Yin, Yang, *Darwin and the Essence of the Universe*, in *What Matters*. 2009.
75. Retchless, A.C. and J.G. Lawrence, *Temporal fragmentation of speciation in bacteria*. Science, 2007. **317**(5841): p. 1093-6.

76. Theobald, D.L., *A formal test of the theory of universal common ancestry*. Nature, 2010. **465**(7295): p. 219-22.
77. Bernstein, F.C., et al., *The Protein Data Bank: a computer-based archival file for macromolecular structures*. J Mol Biol, 1977. **112**(3): p. 535-42.
78. Berman, H.M., et al., *The Protein Data Bank*. Nucleic Acids Res, 2000. **28**(1): p. 235-42.
79. Qian, J., et al., *PartsList: a web-based system for dynamically ranking protein folds based on disparate attributes, including whole-genome expression and interaction information*. Nucleic Acids Res, 2001. **29**(8): p. 1750-64.
80. Marti-Renom, M., et al., *Structure Comparison and Alignment*. 2nd Edition ed, ed. J. Gu and P. Bourne. 2009, New York: Wiley-Blackwell.
81. Godzik, A., *The structural alignment between two proteins: is there a unique answer?* Protein Sci, 1996. **5**(7): p. 1325-38.
82. Kolodny, R., P. Koehl, and M. Levitt, *Comprehensive evaluation of protein structure alignment methods: scoring by geometric measures*. J Mol Biol, 2005. **346**(4): p. 1173-88.
83. Holm, L., et al., *Using Dali for structural comparison of proteins*. Curr Protoc Bioinformatics, 2006. **Chapter 5**: p. Unit 5 5.
84. Holm, L. and C. Sander, *Dali/FSSP classification of three-dimensional protein folds*. Nucleic Acids Res, 1997. **25**(1): p. 231-4.
85. Cuff, A.L., et al., *The CATH classification revisited--architectures reviewed and new ways to characterize structural divergence in superfamilies*. Nucleic Acids Res, 2008.
86. Andreeva, A., et al., *Data growth and its impact on the SCOP database: new developments*. Nucleic Acids Res, 2008. **36**(Database issue): p. D419-25.
87. Hadley, C. and D.T. Jones, *A systematic comparison of protein structure classifications: SCOP, CATH and FSSP*. Structure, 1999. **7**(9): p. 1099-112.
88. Jefferson, E.R., T.P. Walsh, and G.J. Barton, *A comparison of SCOP and CATH with respect to domain-domain interactions*. Proteins, 2008. **70**(1): p. 54-62.
89. Holland, T.A., et al., *Partitioning protein structures into domains: why is it so difficult?* J Mol Biol, 2006. **361**(3): p. 562-90.
90. Veretnik, S., et al., *Toward consistent assignment of structural domains in proteins*. J Mol Biol, 2004. **339**(3): p. 647-78.

91. Goldstein, R.A., *The structure of protein evolution and the evolution of protein structure*. Curr Opin Struct Biol, 2008. **18**(2): p. 170-7.
92. Taylor, W.R., *Evolutionary transitions in protein fold space*. Curr Opin Struct Biol, 2007. **17**(3): p. 354-61.
93. Grishin, N.V., *Fold change in evolution of protein structures*. J Struct Biol, 2001. **134**(2-3): p. 167-85.
94. Alexander, P.A., et al., *The design and characterization of two proteins with 88% sequence identity but different structure and function*. Proc Natl Acad Sci U S A, 2007. **104**(29): p. 11963-8.
95. Andreeva, A. and A.G. Murzin, *Evolution of protein fold in the presence of functional constraints*. Curr Opin Struct Biol, 2006. **16**(3): p. 399-408.
96. Shindyalov, I.N. and P.E. Bourne, *An alternative view of protein fold space*. Proteins, 2000. **38**(3): p. 247-60.
97. Yeats, C., et al., *Gene3D: comprehensive structural and functional annotation of genomes*. Nucleic Acids Res, 2008. **36**(Database issue): p. D414-8.
98. Chothia, C., et al., *Evolution of the protein repertoire*. Science, 2003. **300**(5626): p. 1701-3.
99. Copley, R.R., et al., *Protein domain analysis in the era of complete genomes*. FEBS Lett, 2002. **513**(1): p. 129-34.
100. Doolittle, R.F., *Evolutionary aspects of whole-genome biology*. Curr Opin Struct Biol, 2005. **15**(3): p. 248-53.
101. Koonin, E.V., Y.I. Wolf, and G.P. Karev, *The structure of the protein universe and genome evolution*. Nature, 2002. **420**(6912): p. 218-23.
102. Orengo, C.A. and J.M. Thornton, *Protein families and their evolution-a structural perspective*. Annu Rev Biochem, 2005. **74**: p. 867-900.
103. Gerstein, M., *A structural census of genomes: comparing bacterial, eukaryotic, and archaeal genomes in terms of protein structure*. J Mol Biol, 1997. **274**(4): p. 562-76.
104. Wolf, Y.I., et al., *Distribution of protein folds in the three superkingdoms of life*. Genome Res, 1999. **9**(1): p. 17-26.
105. Apic, G., J. Gough, and S.A. Teichmann, *Domain combinations in archaeal, eubacterial and eukaryotic proteomes*. J Mol Biol, 2001. **310**(2): p. 311-25.

106. Dokholyan, N.V., B. Shakhnovich, and E.I. Shakhnovich, *Expanding protein universe and its origin from the biological Big Bang*. Proc Natl Acad Sci U S A, 2002. **99**(22): p. 14132-6.
107. Karev, G.P., et al., *Birth and death of protein domains: a simple model of evolution explains power law behavior*. BMC Evol Biol, 2002. **2**(1): p. 18.
108. Ranea, J.A., et al., *Evolution of protein superfamilies and bacterial genome size*. J Mol Biol, 2004. **336**(4): p. 871-87.
109. Vogel, C. and C. Chothia, *Protein family expansions and biological complexity*. PLoS Comput Biol, 2006. **2**(5): p. e48.
110. Doolittle, R.F., *The multiplicity of domains in proteins*. Annu Rev Biochem, 1995. **64**: p. 287-314.
111. Moore, A.D., et al., *Arrangements in the modular evolution of proteins*. Trends Biochem Sci, 2008. **33**(9): p. 444-51.
112. Vogel, C., et al., *Structure, function and evolution of multidomain proteins*. Curr Opin Struct Biol, 2004. **14**(2): p. 208-16.
113. Bashton, M. and C. Chothia, *The generation of new protein functions by the combination of domains*. Structure, 2007. **15**(1): p. 85-99.
114. Basu, M.K., et al., *Evolution of protein domain promiscuity in eukaryotes*. Genome Res, 2008. **18**(3): p. 449-61.
115. Vogel, C., et al., *Supra-domains: evolutionary units larger than single protein domains*. J Mol Biol, 2004. **336**(3): p. 809-23.
116. Snel, B., P. Bork, and M.A. Huynen, *Genomes in flux: the evolution of archaeal and proteobacterial gene content*. Genome Res, 2002. **12**(1): p. 17-25.
117. Kummerfeld, S.K. and S.A. Teichmann, *Relative rates of gene fusion and fission in multi-domain proteins*. Trends Genet, 2005. **21**(1): p. 25-30.
118. Fong, J.H., et al., *Modeling the evolution of protein domain architectures using maximum parsimony*. J Mol Biol, 2007. **366**(1): p. 307-15.
119. Ekman, D., A.K. Bjorklund, and A. Elofsson, *Quantification of the elevated rate of domain rearrangements in metazoa*. J Mol Biol, 2007. **372**(5): p. 1337-48.
120. Itoh, M., et al., *Evolutionary history and functional implications of protein domains and their combinations in eukaryotes*. Genome Biol, 2007. **8**(6): p. R121.

121. Yang, S. and P.E. Bourne, *The Evolutionary History of Protein Domains Viewed by Species Phylogeny*. PLoS Comput Biol, To be published.
122. Friedberg, I. and A. Godzik, *Fragnostic: walking through protein structure space*. Nucleic Acids Res, 2005. **33**(Web Server issue): p. W249-51.
123. Berezovsky, I.N., A.Y. Grosberg, and E.N. Trifonov, *Closed loops of nearly standard size: common basic element of protein structure*. FEBS Lett, 2000. **466**(2-3): p. 283-6.
124. Koczyk, G. and I.N. Berezovsky, *Domain Hierarchy and closed Loops (DHcL): a server for exploring hierarchy of protein domain structure*. Nucleic Acids Res, 2008. **36**(Web Server issue): p. W239-45.
125. Sobolevsky, Y. and E.N. Trifonov, *Protein modules conserved since LUCA*. J Mol Evol, 2006. **63**(5): p. 622-34.
126. Frenkel, Z.M. and E.N. Trifonov, *From protein sequence space to elementary protein modules*. Gene, 2008. **408**(1-2): p. 64-71.
127. Xie, L. and P.E. Bourne, *Detecting evolutionary relationships across existing fold space, using sequence order-independent profile-profile alignments*. Proc Natl Acad Sci U S A, 2008. **105**(14): p. 5441-6.
128. Bashton, M., I. Nobeli, and J.M. Thornton, *Cognate ligand domain mapping for enzymes*. J Mol Biol, 2006. **364**(4): p. 836-52.
129. Mintz, S., et al., *Generation and analysis of a protein-protein interface data set with similar chemical and spatial patterns of interactions*. Proteins, 2005. **61**(1): p. 6-20.
130. Shulman-Peleg, A., et al., *MultiBind and MAPPIS: webserver for multiple alignment of protein 3D-binding sites and their interactions*. Nucleic Acids Res, 2008. **36**(Web Server issue): p. W260-4.
131. Levy, E.D., et al., *3D complex: a structural classification of protein complexes*. PLoS Comput Biol, 2006. **2**(11): p. e155.
132. Ranea, J.A., et al., *Protein superfamily evolution and the last universal common ancestor (LUCA)*. J Mol Evol, 2006. **63**(4): p. 513-25.
133. Ma, B.G., et al., *Characters of very ancient proteins*. Biochem Biophys Res Commun, 2008. **366**(3): p. 607-11.
134. Di Giulio, M., *The origin of genes could be polyphyletic*. Gene, 2008. **426**(1-2): p. 39-46.

135. Yang, S., R.F. Doolittle, and P.E. Bourne, *Phylogeny determined by protein domain content*. Proc Natl Acad Sci U S A, 2005. **102**(2): p. 373-8.
136. Deeds, E.J., H. Hennessey, and E.I. Shakhnovich, *Prokaryotic phylogenies inferred from protein structural domains*. Genome Res, 2005. **15**(3): p. 393-402.
137. Caetano-Anolles, G. and D. Caetano-Anolles, *An evolutionarily structured universe of protein architecture*. Genome Res, 2003. **13**(7): p. 1563-71.
138. Caetano-Anolles, G. and D. Caetano-Anolles, *Universal sharing patterns in proteomes and evolution of protein fold architecture and life*. J Mol Evol, 2005. **60**(4): p. 484-98.
139. Snel, B., P. Bork, and M.A. Huynen, *Genome phylogeny based on gene content*. Nat Genet, 1999. **21**(1): p. 108-10.
140. Sadreyev, R.I., B.H. Kim, and N.V. Grishin, *Discrete-continuous duality of protein structure space*. Curr Opin Struct Biol, 2009. **19**(3): p. 321-8.
141. Aravind, L., L.M. Iyer, and E.V. Koonin, *Comparative genomics and structural biology of the molecular innovations of eukaryotes*. Curr Opin Struct Biol, 2006. **16**(3): p. 409-19.
142. Dupont, C.L., et al., *Modern proteomes contain putative imprints of ancient shifts in trace metal geochemistry*. Proc Natl Acad Sci U S A, 2006. **103**(47): p. 17822-7.
143. Tu, B.P. and J.S. Weissman, *Oxidative protein folding in eukaryotes: mechanisms and consequences*. J Cell Biol, 2004. **164**(3): p. 341-6.
144. Karev, G.P., et al., *Gene family evolution: an in-depth theoretical and simulation analysis of non-linear birth-death-innovation models*. BMC Evol Biol, 2004. **4**: p. 32.
145. Murzin, A.G., et al., *SCOP: a structural classification of proteins database for the investigation of sequences and structures*. J Mol Biol, 1995. **247**(4): p. 536-40.
146. Wang, M. and G. Caetano-Anolles, *Global phylogeny determined by the combination of protein domains in proteomes*. Mol Biol Evol, 2006. **23**(12): p. 2444-54.
147. Fournier, G.P. and J.P. Gogarten, *Signature of a primitive genetic code in ancient protein lineages*. J Mol Evol, 2007. **65**(4): p. 425-36.
148. Cosentino Lagomarsino, M., et al., *Universal features in the genome-level evolution of protein domains*. Genome Biol, 2009. **10**(1): p. R12.

149. Deeds, E.J., B. Shakhnovich, and E.I. Shakhnovich, *Proteomic traces of speciation*. J Mol Biol, 2004. **336**(3): p. 695-706.
150. Retief, J.D., *Phylogenetic analysis using PHYLIP*. Methods Mol Biol, 2000. **132**: p. 243-58.
151. Wilgenbusch, J.C. and D. Swofford, *Inferring evolutionary trees with PAUP**. Curr Protoc Bioinformatics, 2003. **Chapter 6**: p. Unit 6 4.
152. Groll, M., et al., *Molecular machines for protein degradation*. Chembiochem, 2005. **6**(2): p. 222-56.
153. Goldberg, A.L., *Protein degradation and protection against misfolded or damaged proteins*. Nature, 2003. **426**(6968): p. 895-9.
154. Glickman, M.H. and A. Ciechanover, *The ubiquitin-proteasome proteolytic pathway: destruction for the sake of construction*. Physiol Rev, 2002. **82**(2): p. 373-428.
155. Rabinovich, E., et al., *Proteasome Channel Opening as a Rate-Limiting Step in the Ubiquitin- Proteasome System*. Israel Journal of Chemistry, 2006. **46**(2): p. 219-224.
156. Groll, M., et al., *Structure of 20S proteasome from yeast at 2.4 Å resolution*. Nature, 1997. **386**(6624): p. 463-71.
157. Ruepp, A., et al., *Proteasome function is dispensable under normal but not under heat shock conditions in Thermoplasma acidophilum*. FEBS Lett, 1998. **425**(1): p. 87-90.
158. Coux, O., K. Tanaka, and A.L. Goldberg, *Structure and functions of the 20S and 26S proteasomes*. Annu Rev Biochem, 1996. **65**: p. 801-47.
159. Smith, D.M., N. Benaroudj, and A. Goldberg, *Proteasomes and their associated ATPases: a destructive combination*. J Struct Biol, 2006. **156**(1): p. 72-83.
160. Iyer, L.M., et al., *Evolutionary history and higher order classification of AAA+ ATPases*. J Struct Biol, 2004. **146**(1-2): p. 11-31.
161. Darwin, K.H., et al., *Characterization of a Mycobacterium tuberculosis proteasomal ATPase homologue*. Mol Microbiol, 2005. **55**(2): p. 561-71.
162. Zwickl, P., et al., *An archaeobacterial ATPase, homologous to ATPases in the eukaryotic 26 S proteasome, activates protein breakdown by 20 S proteasomes*. J Biol Chem, 1999. **274**(37): p. 26008-14.

163. Bochtler, M., et al., *The proteasome*. Annu Rev Biophys Biomol Struct, 1999. **28**: p. 295-317.
164. Hill, C.P., E.I. Masters, and F.G. Whitby, *The 11S regulators of 20S proteasome activity*. Curr Top Microbiol Immunol, 2002. **268**: p. 73-89.
165. Ustrell, V., et al., *PA200, a nuclear proteasome activator involved in DNA repair*. Embo J, 2002. **21**(13): p. 3516-25.
166. Tanahashi, N., et al., *Hybrid proteasomes. Induction by interferon-gamma and contribution to ATP-dependent proteolysis*. J Biol Chem, 2000. **275**(19): p. 14336-45.
167. Hershko, A., *The ubiquitin system for protein degradation and some of its roles in the control of the cell division cycle*. Cell Death Differ, 2005. **12**(9): p. 1191-7.
168. Orłowski, M. and S. Wilk, *Ubiquitin-independent proteolytic functions of the proteasome*. Arch Biochem Biophys, 2003. **415**(1): p. 1-5.
169. Iyer, L.M., A.M. Burroughs, and L. Aravind, *The prokaryotic antecedents of the ubiquitin-signaling system and the early evolution of ubiquitin-like beta-grasp domains*. Genome Biol, 2006. **7**(7): p. R60.
170. Burton, R.E., T.A. Baker, and R.T. Sauer, *Nucleotide-dependent substrate recognition by the AAA+ HslUV protease*. Nat Struct Mol Biol, 2005. **12**(3): p. 245-51.
171. Gille, C., et al., *A comprehensive view on proteasomal sequences: implications for the evolution of the proteasome*. J Mol Biol, 2003. **326**(5): p. 1437-48.
172. Gough, J., et al., *Assignment of homology to genome sequences using a library of hidden Markov models that represent all proteins of known structure*. J Mol Biol, 2001. **313**(4): p. 903-19.
173. Lowe, J., et al., *Crystal structure of the 20S proteasome from the archaeon T. acidophilum at 3.4 Å resolution*. Science, 1995. **268**(5210): p. 533-9.
174. Seemuller, E., A. Lupas, and W. Baumeister, *Autocatalytic processing of the 20S proteasome*. Nature, 1996. **382**(6590): p. 468-71.
175. Seemuller, E., et al., *Proteasome from Thermoplasma acidophilum: a threonine protease*. Science, 1995. **268**(5210): p. 579-82.
176. Honda, D., A. Yokota, and J. Sugiyama, *Detection of seven major evolutionary lineages in cyanobacteria based on the 16S rRNA gene sequence analysis with new sequences of five marine Synechococcus strains*. J Mol Evol, 1999. **48**(6): p. 723-39.

177. Woese, C.R., *Bacterial evolution*. Microbiol Rev, 1987. **51**(2): p. 221-71.
178. De Mot, R., *Actinomycete-like proteasomes in a Gram-negative bacterium*. Trends Microbiol, 2007. **15**(8): p. 335-8.
179. Gupta, R.S. and E. Lorenzini, *Phylogeny and molecular signatures (conserved proteins and indels) that are specific for the Bacteroidetes and Chlorobi species*. BMC Evol Biol, 2007. **7**: p. 71.
180. Kelley, L.A., R.M. MacCallum, and M.J. Sternberg, *Enhanced genome annotation using structural profiles in the program 3D-PSSM*. J Mol Biol, 2000. **299**(2): p. 499-520.
181. Shindyalov, I.N. and P.E. Bourne, *Protein structure alignment by incremental combinatorial extension (CE) of the optimal path*. Protein Eng, 1998. **11**(9): p. 739-47.
182. Bochtler, M., et al., *Crystal structure of heat shock locus V (HslV) from Escherichia coli*. Proc Natl Acad Sci U S A, 1997. **94**(12): p. 6070-4.
183. Altschul, S.F., et al., *Basic local alignment search tool*. J Mol Biol, 1990. **215**(3): p. 403-10.
184. Reith, F., et al., *Biomineralization of gold: biofilms on bacterioform gold*. Science, 2006. **313**(5784): p. 233-6.
185. Finneran, K.T., C.V. Johnsen, and D.R. Lovley, *Rhodoferrax ferrireducens* sp. nov., a psychrotolerant, facultatively anaerobic bacterium that oxidizes acetate with the reduction of Fe(III). Int J Syst Evol Microbiol, 2003. **53**(Pt 3): p. 669-73.
186. Goris, J., et al., *Classification of the biphenyl- and polychlorinated biphenyl-degrading strain LB400T and relatives as Burkholderia xenovorans* sp. nov. Int J Syst Evol Microbiol, 2004. **54**(Pt 5): p. 1677-81.
187. Singh, A.K., et al., *The heat shock response in the cyanobacterium Synechocystis sp. Strain PCC 6803 and regulation of gene expression by HrcA and SigB*. Arch Microbiol, 2006. **186**(4): p. 273-86.
188. Huang, L., et al., *Global gene expression profiles of the cyanobacterium Synechocystis sp. strain PCC 6803 in response to irradiation with UV-B and white light*. J Bacteriol, 2002. **184**(24): p. 6845-58.
189. Kanasaki, Y., et al., *Salt stress and hyperosmotic stress regulate the expression of different sets of genes in Synechocystis sp. PCC 6803*. Biochem Biophys Res Commun, 2002. **290**(1): p. 339-48.

190. Li, H., et al., *Differential gene expression in response to hydrogen peroxide and the putative PerR regulon of Synechocystis sp. strain PCC 6803*. J Bacteriol, 2004. **186**(11): p. 3331-45.
191. Zeller, T., et al., *Transcriptome and physiological responses to hydrogen peroxide of the facultatively phototrophic bacterium Rhodobacter sphaeroides*. J Bacteriol, 2005. **187**(21): p. 7232-42.
192. Dominguez-Cuevas, P., et al., *Transcriptional tradeoff between metabolic and stress-response programs in Pseudomonas putida KT2440 cells exposed to toluene*. J Biol Chem, 2006. **281**(17): p. 11981-91.
193. Alm, E.J., et al., *The MicrobesOnline Web site for comparative genomics*. Genome Res, 2005. **15**(7): p. 1015-22.
194. Pfister, P., et al., *Molecular analysis of Methanobacterium phage psiM2*. Mol Microbiol, 1998. **30**(2): p. 233-44.
195. Seitz, A., et al., *Enzymatic cross-linking of purple membranes catalyzed by bacterial transglutaminase*. Biomacromolecules, 2001. **2**(1): p. 233-8.
196. Helmann, J.D., et al., *Global transcriptional response of Bacillus subtilis to heat shock*. J Bacteriol, 2001. **183**(24): p. 7318-28.
197. Zhang, W., et al., *Oxidative stress and heat-shock responses in Desulfovibrio vulgaris by genome-wide transcriptomic analysis*. Antonie Van Leeuwenhoek, 2006. **90**(1): p. 41-55.
198. Akhtar, M.W., et al., *Oligomeric Hsp33 with enhanced chaperone activity: gel filtration, cross-linking, and small angle x-ray scattering (SAXS) analysis*. J Biol Chem, 2004. **279**(53): p. 55760-9.
199. Quail, M.A. and J.R. Guest, *Purification, characterization and mode of action of PdhR, the transcriptional repressor of the pdhR-aceEF-lpd operon of Escherichia coli*. Mol Microbiol, 1995. **15**(3): p. 519-29.
200. Edgar, R.C., *MUSCLE: multiple sequence alignment with high accuracy and high throughput*. Nucleic Acids Res, 2004. **32**(5): p. 1792-7.
201. Guindon, S. and O. Gascuel, *A simple, fast, and accurate algorithm to estimate large phylogenies by maximum likelihood*. Syst Biol, 2003. **52**(5): p. 696-704.
202. Drummond, A., et al., *Geneious v2.5*. 2006.
203. Moreland, J.L., et al., *The Molecular Biology Toolkit (MBT): a modular platform for developing molecular visualization applications*. BMC Bioinformatics, 2005. **6**: p. 21.

204. Martin, W., et al., *Hydrothermal vents and the origin of life*. Nat Rev Microbiol, 2008. **6**(11): p. 805-14.
205. Philippe, H. and P. Forterre, *The rooting of the universal tree of life is not reliable*. J Mol Evol, 1999. **49**(4): p. 509-23.
206. Price, P.B., *Microbial life in glacial ice and implications for a cold origin of life*. FEMS Microbiol Ecol, 2007. **59**(2): p. 217-31.
207. Trevors, J.T. and G.H. Pollack, *Hypothesis: the origin of life in a hydrogel environment*. Prog Biophys Mol Biol, 2005. **89**(1): p. 1-8.
208. Cavalier-Smith, T., *Rooting the tree of life by transition analyses*. Biol Direct, 2006. **1**(1): p. 19.
209. Lake, J.A., et al., *Genome beginnings: rooting the tree of life*. Philos Trans R Soc Lond B Biol Sci, 2009. **364**(1527): p. 2177-85.
210. Liu, S.V., *A Fundamentally New Perspective on the Origin and Evolution of Life*. Pioneer, 2008. **3**: p. 7-17.
211. Cavalier-Smith, T., *Obcells as proto-organisms: membrane heredity, lithophosphorylation, and the origins of the genetic code, the first cells, and photosynthesis*. J Mol Evol, 2001. **53**(4-5): p. 555-95.
212. Martin, W. and M.J. Russell, *On the origins of cells: a hypothesis for the evolutionary transitions from abiotic geochemistry to chemoautotrophic prokaryotes, and from prokaryotes to nucleated cells*. Philos Trans R Soc Lond B Biol Sci, 2003. **358**(1429): p. 59-83; discussion 83-5.
213. Di Giulio, M., *The evidence that the tree of life is not rooted within the Archaea is unreliable: a reply to Skophammer et al*. Gene, 2007. **394**(1-2): p. 105-6.
214. Liolios, K., et al., *The Genomes On Line Database (GOLD) in 2007: status of genomic and metagenomic projects and their associated metadata*. Nucleic Acids Res, 2008. **36**(Database issue): p. D475-9.
215. Gupta, R.S., *Protein phylogenies and signature sequences: A reappraisal of evolutionary relationships among archaeobacteria, eubacteria, and eukaryotes*. Microbiol Mol Biol Rev, 1998. **62**(4): p. 1435-91.
216. Philippe, H., K. Budin, and D. Moreira, *Horizontal transfers confuse the prokaryotic phylogeny based on the HSP70 protein family*. Mol Microbiol, 1999. **31**(3): p. 1007-10.
217. Chang, Y.W., et al., *Crystal structures of the 70-kDa heat shock proteins in domain disjoining conformation*. J Biol Chem, 2008. **283**(22): p. 15502-11.

218. Guda, C., et al., *CE-MC: a multiple protein structure alignment server*. Nucleic Acids Res, 2004. **32**(Web Server issue): p. W100-3.
219. Gupta, R.S., *What are archaeobacteria: life's third domain or monoderm prokaryotes related to gram-positive bacteria? A new proposal for the classification of prokaryotic organisms*. Mol Microbiol, 1998. **29**(3): p. 695-707.
220. Soding, J., M. Remmert, and A. Biegert, *HHrep: de novo protein repeat detection and the origin of TIM barrels*. Nucleic Acids Res, 2006. **34**(Web Server issue): p. W137-42.
221. Norager, S., et al., *E. coli dihydroorotate dehydrogenase reveals structural and functional distinctions between different classes of dihydroorotate dehydrogenases*. Structure, 2002. **10**(9): p. 1211-23.
222. Klimke, W., et al., *The National Center for Biotechnology Information's Protein Clusters Database*. Nucleic Acids Res, 2009. **37**(Database issue): p. D216-23.
223. Rowland, P., et al., *Structure of dihydroorotate dehydrogenase B: electron transfer between two flavin groups bridged by an iron-sulphur cluster*. Structure, 2000. **8**(12): p. 1227-38.
224. Valas, R.E., S. Yang, and P.E. Bourne, *Nothing about protein structure classification makes sense except in the light of evolution*. Curr Opin Struct Biol, 2009.
225. Veeramalai, M., Y. Ye, and A. Godzik, *TOPS++FATCAT: fast flexible structural alignment using constraints derived from TOPS+ Strings Model*. BMC Bioinformatics, 2008. **9**: p. 358.
226. Wang, Y., et al., *MMDB: annotating protein sequences with Entrez's 3D-structure database*. Nucleic Acids Res, 2007. **35**(Database issue): p. D298-300.
227. Forterre, P., *The origin of viruses and their possible roles in major evolutionary transitions*. Virus Res, 2006. **117**(1): p. 5-16.
228. Myllykallio, H., et al., *An alternative flavin-dependent mechanism for thymidylate synthesis*. Science, 2002. **297**(5578): p. 105-7.
229. Forterre, P., *The origin of DNA genomes and DNA replication proteins*. Curr Opin Microbiol, 2002. **5**(5): p. 525-32.
230. Leipe, D.D., L. Aravind, and E.V. Koonin, *Did DNA replication evolve twice independently?* Nucleic Acids Res, 1999. **27**(17): p. 3389-401.
231. Konagurthu, A.S., et al., *MUSTANG: a multiple structural alignment algorithm*. Proteins, 2006. **64**(3): p. 559-74.

232. Pettersen, E.F., et al., *UCSF Chimera--a visualization system for exploratory research and analysis*. J Comput Chem, 2004. **25**(13): p. 1605-12.
233. Zhaxybayeva, O. and J. Gogarten, *Horizontal gene transfer, gene histories, and the root of the tree of life*, in *Planetary systems and the origin of life*, R.E. Pudritz, P.G. Higgs, and J.R. Stone, Editors. 2007, Cambridge University Press: Cambridge; New York.
234. Glansdorff, N., Y. Xu, and B. Labedan, *The last universal common ancestor: emergence, constitution and genetic legacy of an elusive forerunner*. Biol Direct, 2008. **3**: p. 29.
235. Lecompte, O., et al., *Comparative analysis of ribosomal proteins in complete genomes: an example of reductive evolution at the domain scale*. Nucleic Acids Res, 2002. **30**(24): p. 5382-90.
236. De Rosa, M., A. Gambacorta, and A. Gliozzi, *Structure, biosynthesis, and physicochemical properties of archaeobacterial lipids*. Microbiol Rev, 1986. **50**(1): p. 70-80.
237. Battistuzzi, F.U. and S.B. Hedges, *A major clade of prokaryotes with ancient adaptations to life on land*. Mol Biol Evol, 2009. **26**(2): p. 335-43.
238. Rivera, M.C. and J.A. Lake, *Evidence that eukaryotes and eocyte prokaryotes are immediate relatives*. Science, 1992. **257**(5066): p. 74-6.
239. Cavalier-Smith, T., *The phagotrophic origin of eukaryotes and phylogenetic classification of Protozoa*. Int J Syst Evol Microbiol, 2002. **52**(Pt 2): p. 297-354.
240. Forterre, P., *Three RNA cells for ribosomal lineages and three DNA viruses to replicate their genomes: a hypothesis for the origin of cellular domain*. Proc Natl Acad Sci U S A, 2006. **103**(10): p. 3669-74.
241. Koonin, E.V., T.G. Senkevich, and V.V. Dolja, *The ancient Virus World and evolution of cells*. Biol Direct, 2006. **1**: p. 29.
242. McGeoch, A.T. and S.D. Bell, *Extra-chromosomal elements and the evolution of cellular DNA replication machineries*. Nat Rev Mol Cell Biol, 2008. **9**(7): p. 569-74.
243. Boucher, Y., M. Kamekura, and W.F. Doolittle, *Origins and evolution of isoprenoid lipid biosynthesis in archaea*. Mol Microbiol, 2004. **52**(2): p. 515-27.
244. Iyer, L.M., et al., *Amidoligases with ATP-grasp, glutamine synthetase-like and acetyltransferase-like domains: synthesis of novel metabolites and peptide modifications of proteins*. Mol Biosyst, 2009. **5**(12): p. 1636-60.

245. Klein, D.J., P.B. Moore, and T.A. Steitz, *The roles of ribosomal proteins in the structure assembly, and evolution of the large ribosomal subunit*. J Mol Biol, 2004. **340**(1): p. 141-77.
246. Harms, J., et al., *High resolution structure of the large ribosomal subunit from a mesophilic eubacterium*. Cell, 2001. **107**(5): p. 679-88.
247. O'Brien, T.W., *Properties of human mitochondrial ribosomes*. IUBMB Life, 2003. **55**(9): p. 505-13.
248. Zhang, R., H.Y. Ou, and C.T. Zhang, *DEG: a database of essential genes*. Nucleic Acids Res, 2004. **32**(Database issue): p. D271-2.
249. Fang, G., E. Rocha, and A. Danchin, *How essential are nonessential genes?* Mol Biol Evol, 2005. **22**(11): p. 2147-56.
250. Drummond, D.A. and C.O. Wilke, *Mistranslation-induced protein misfolding as a dominant constraint on coding-sequence evolution*. Cell, 2008. **134**(2): p. 341-52.
251. Sutter, M., et al., *A distinct structural region of the prokaryotic ubiquitin-like protein (Pup) is recognized by the N-terminal domain of the proteasomal ATPase Mpa*. FEBS Lett, 2009. **583**(19): p. 3151-7.
252. Makarova, K.S., V.A. Ponomarev, and E.V. Koonin, *Two C or not two C: recurrent disruption of Zn-ribbons, gene duplication, lineage-specific gene loss, and horizontal gene transfer in evolution of bacterial ribosomal proteins*. Genome Biol, 2001. **2**(9): p. RESEARCH 0033.
253. Dupont, C.L., et al., *History of biological metal utilization inferred through phylogenomic analysis of protein structures*. Proc Natl Acad Sci U S A, 2010. **107**(23): p. 10567-72.
254. Filee, J., et al., *Evolution of DNA polymerase families: evidences for multiple gene exchange between cellular and viral proteins*. J Mol Evol, 2002. **54**(6): p. 763-73.
255. Tahirov, T.H., et al., *Evolution of DNA polymerases: an inactivated polymerase-exonuclease module in Pol epsilon and a chimeric origin of eukaryotic polymerases from two classes of archaeal ancestors*. Biol Direct, 2009. **4**: p. 11.
256. Koonin, E.V., et al., *Bacterial homologs of the small subunit of eukaryotic DNA primase*. J Mol Microbiol Biotechnol, 2000. **2**(4): p. 509-12.
257. Della, M., et al., *Mycobacterial Ku and ligase proteins constitute a two-component NHEJ repair machine*. Science, 2004. **306**(5696): p. 683-5.

258. Yutin, N. and E.V. Koonin, *Evolution of DNA ligases of Nucleo-Cytoplasmic Large DNA viruses of eukaryotes: a case of hidden complexity*. Biol Direct, 2009. **4**: p. 51.
259. Samuels, M., et al., *A biochemically active MCM-like helicase in Bacillus cereus*. Nucleic Acids Res, 2009. **37**(13): p. 4441-52.
260. Ammelburg, M., T. Frickey, and A.N. Lupas, *Classification of AAA+ proteins*. J Struct Biol, 2006. **156**(1): p. 2-11.
261. Chen, Y.J., X. Yu, and E.H. Egelman, *The hexameric ring structure of the Escherichia coli RuvB branch migration protein*. J Mol Biol, 2002. **319**(3): p. 587-91.
262. Komori, K., et al., *A Holliday junction resolvase from Pyrococcus furiosus: functional similarity to Escherichia coli RuvC provides evidence for conserved mechanism of homologous recombination in Bacteria, Eukarya, and Archaea*. Proc Natl Acad Sci U S A, 1999. **96**(16): p. 8873-8.
263. Forterre, P., et al., *Origin and evolution of DNA topoisomerases*. Biochimie, 2007. **89**(4): p. 427-46.
264. Forterre, P. and D. Gadelle, *Phylogenomics of DNA topoisomerases: their origin and putative roles in the emergence of modern organisms*. Nucleic Acids Res, 2009. **37**(3): p. 679-92.
265. Becerra, A., et al., *The Very Early Stages of Biological Evolution and the Nature of the Last Common Ancestor of the Three Major Cell Domains*. Annual Review of Ecology, Evolution, and Systematics, 2007. **38**(1): p. 361-379.
266. Yutin, N., et al., *The deep archaeal roots of eukaryotes*. Mol Biol Evol, 2008. **25**(8): p. 1619-30.
267. Brochier-Armanet, C., et al., *Mesophilic Crenarchaeota: proposal for a third archaeal phylum, the Thaumarchaeota*. Nat Rev Microbiol, 2008. **6**(3): p. 245-52.
268. Elkins, J.G., et al., *A korarchaeal genome reveals insights into the evolution of the Archaea*. Proc Natl Acad Sci U S A, 2008. **105**(23): p. 8102-7.
269. Martin, W. and M. Muller, *The hydrogen hypothesis for the first eukaryote*. Nature, 1998. **392**(6671): p. 37-41.
270. Ettema, T.J. and R. Bernander, *Cell division and the ESCRT complex: A surprise from the archaea*. Commun Integr Biol, 2009. **2**(2): p. 86-8.
271. Schmitt, E., et al., *Structural basis for tRNA-dependent amidotransferase function*. Structure, 2005. **13**(10): p. 1421-33.

272. Schwartz, J.H., J.Y. Reeves, and J.D. Broome, *Two L-asparaginases from E. coli and their action against tumors*. Proc Natl Acad Sci U S A, 1966. **56**(5): p. 1516-9.
273. Poole, A.M. and D. Penny, *Evaluating hypotheses for the origin of eukaryotes*. Bioessays, 2007. **29**(1): p. 74-84.
274. Hou, S., et al., *Complete genome sequence of the extremely acidophilic methanotroph isolate V4, Methylacidiphilum infernorum, a representative of the bacterial phylum Verrucomicrobia*. Biol Direct, 2008. **3**: p. 26.
275. Nes, W.R. and W.D. Nes, *Lipids in evolution*. 1980, New York: Plenum Press. xi, 244 p.
276. Lamb, D.C., et al., *A sterol biosynthetic pathway in Mycobacterium*. FEBS Lett, 1998. **437**(1-2): p. 142-4.
277. Desmond, E. and S. Gribaldo, *Phylogenomics of Sterol Synthesis: Insights into the Origin, Evolution, and Diversity of a Key Eukaryotic Feature*. Genome Biol Evol, 2009. **2009**(0): p. 364-381.
278. Smucker, R.A., *Biochemistry of the Streptomyces spore sheath*, in *Biological, biochemical, and biomedical aspects of actinomycetes*, L. Ortiz-Ortiz, L.F. Bojalil, and V. Yakoleff, Editors. 1984, Academic Press: Orlando. p. xv, 643 p.
279. Wehenkel, A., et al., *Mycobacterial Ser/Thr protein kinases and phosphatases: physiological roles and therapeutic potential*. Biochim Biophys Acta, 2008. **1784**(1): p. 193-202.
280. Macek, B., et al., *The serine/threonine/tyrosine phosphoproteome of the model bacterium Bacillus subtilis*. Mol Cell Proteomics, 2007. **6**(4): p. 697-707.
281. Sandoval-Calderon, M., et al., *A eukaryote-like cardiolipin synthase is present in Streptomyces coelicolor and in most actinobacteria*. J Biol Chem, 2009. **284**(26): p. 17383-90.
282. Da Lage, J.L., G. Feller, and S. Janecek, *Horizontal gene transfer from Eukarya to bacteria and domain shuffling: the alpha-amylase model*. Cell Mol Life Sci, 2004. **61**(1): p. 97-109.
283. Leibundgut, M., et al., *The multienzyme architecture of eukaryotic fatty acid synthases*. Curr Opin Struct Biol, 2008. **18**(6): p. 714-25.
284. Cavalier-Smith, T., *The origin of eukaryotic and archaeobacterial cells*. Ann N Y Acad Sci, 1987. **503**: p. 17-54.

285. Bell, P.J., *Sex and the eukaryotic cell cycle is consistent with a viral ancestry for the eukaryotic nucleus*. J Theor Biol, 2006. **243**(1): p. 54-63.
286. Aravind, L. and E.V. Koonin, *Prokaryotic homologs of the eukaryotic DNA-end-binding protein Ku, novel domains in the Ku protein and prediction of a prokaryotic double-strand break repair system*. Genome Res, 2001. **11**(8): p. 1365-74.
287. Grangeasse, C., et al., *Tyrosine phosphorylation: an emerging regulatory device of bacterial physiology*. Trends Biochem Sci, 2007. **32**(2): p. 86-94.
288. Kanehisa, M., *The KEGG database*. Novartis Found Symp, 2002. **247**: p. 91-101; discussion 101-3, 119-28, 244-52.
289. Payandeh, J. and E.F. Pai, *Enzyme-driven speciation: crystallizing Archaea via lipid capture*. J Mol Evol, 2007. **64**(3): p. 364-74.
290. Finn, R.D., et al., *The Pfam protein families database*. Nucleic Acids Res, 2008. **36**(Database issue): p. D281-8.
291. Sojka, L., et al., *YbxF, a protein associated with exponential-phase ribosomes in Bacillus subtilis*. J Bacteriol, 2007. **189**(13): p. 4809-14.
292. Stechmann, A. and T. Cavalier-Smith, *Evolutionary origins of Hsp90 chaperones and a deep paralogy in their bacterial ancestors*. J Eukaryot Microbiol, 2004. **51**(3): p. 364-73.
293. Cavalier-Smith, T., *Origin of the cell nucleus, mitosis and sex: roles of intracellular coevolution*. Biol Direct, 2010. **5**(1): p. 7.
294. van den Bosch, H., et al., *Biochemistry of peroxisomes*. Annu Rev Biochem, 1992. **61**: p. 157-97.
295. De Duve, C., *Blueprint for a cell: the nature and origin of life*. 1991, Burlington, N.C.: N. Patterson. 275 p.
296. Cavalier-Smith, T. *Symbiotic origin of peroxisomes*. in *Endocytobiology IV*. 1990. Paris: Institut national de la recherche agronomique.
297. Gabaldon, T., et al., *Origin and evolution of the peroxisomal proteome*. Biol Direct, 2006. **1**: p. 8.
298. Schluter, A., et al., *The evolutionary origin of peroxisomes: an ER-peroxisome connection*. Mol Biol Evol, 2006. **23**(4): p. 838-45.
299. Gabaldon, T. and M.A. Huynen, *Reconstruction of the proto-mitochondrial metabolism*. Science, 2003. **301**(5633): p. 609.

300. Duhita, N., et al., *The origin of peroxisomes: The possibility of an actinobacterial symbiosis*. Gene, 2009. **450**(1-2): p. 18-24.
301. Race, H.L., R.G. Herrmann, and W. Martin, *Why have organelles retained genomes?* Trends Genet, 1999. **15**(9): p. 364-70.
302. Allen, J.F., *Control of gene expression by redox potential and the requirement for chloroplast and mitochondrial genomes*. J Theor Biol, 1993. **165**(4): p. 609-31.
303. Koonin, E.V., *Temporal order of evolution of DNA replication systems inferred by comparison of cellular and viral DNA polymerases*. Biol Direct, 2006. **1**: p. 39.
304. Rothschild, L.J. and R.L. Mancinelli, *Life in extreme environments*. Nature, 2001. **409**(6823): p. 1092-101.
305. Valentine, D.L., *Adaptations to energy stress dictate the ecology and evolution of the Archaea*. Nat Rev Microbiol, 2007. **5**(4): p. 316-23.
306. Fogg, M.J., L.H. Pearl, and B.A. Connolly, *Structural basis for uracil recognition by archaeal family B DNA polymerases*. Nat Struct Biol, 2002. **9**(12): p. 922-7.
307. Heddle, J.A. and J.H. Bielas, *Unifying concept of DNA repair: the polymerase scanning hypothesis*. Environ Mol Mutagen, 2005. **45**(2-3): p. 143-9.
308. Bielas, J.H., *Non-transcribed strand repair revealed in quiescent cells*. Mutagenesis, 2006. **21**(1): p. 49-53.
309. Bell, P.J., *The viral eukaryogenesis hypothesis: a key role for viruses in the emergence of eukaryotes from a prokaryotic world environment*. Ann N Y Acad Sci, 2009. **1178**: p. 91-105.
310. Bell, P.J., *Viral eukaryogenesis: was the ancestor of the nucleus a complex DNA virus?* J Mol Evol, 2001. **53**(3): p. 251-6.
311. Vermeij, G.J., *Evolution and escalation: an ecological history of life*. 1987, Princeton, N.J.: Princeton University Press. xv, 527 p.
312. Wishart, D.S., et al., *DrugBank: a knowledgebase for drugs, drug actions and drug targets*. Nucleic Acids Res, 2008. **36**(Database issue): p. D901-6.
313. Sioud, M., et al., *Coumarin and quinolone action in archaebacteria: evidence for the presence of a DNA gyrase-like enzyme*. J Bacteriol, 1988. **170**(2): p. 946-53.
314. Jolly, L., et al., *Lactobacillus helveticus glycosyltransferases: from genes to carbohydrate synthesis*. Glycobiology, 2002. **12**(5): p. 319-27.

315. Hansen, J.L., P.B. Moore, and T.A. Steitz, *Structures of five antibiotics bound at the peptidyl transferase center of the large ribosomal subunit*. J Mol Biol, 2003. **330**(5): p. 1061-75.
316. Hansen, J.L., et al., *The structures of four macrolide antibiotics bound to the large ribosomal subunit*. Mol Cell, 2002. **10**(1): p. 117-28.
317. Pfister, P., et al., *23S rRNA base pair 2057-2611 determines ketolide susceptibility and fitness cost of the macrolide resistance mutation 2058A-->G*. Proc Natl Acad Sci U S A, 2005. **102**(14): p. 5180-5.
318. Vester, B. and S. Douthwaite, *Macrolide resistance conferred by base substitutions in 23S rRNA*. Antimicrob Agents Chemother, 2001. **45**(1): p. 1-12.
319. Tu, D., et al., *Structures of MLSBK antibiotics bound to mutated large ribosomal subunits provide a structural explanation for resistance*. Cell, 2005. **121**(2): p. 257-70.
320. Gregory, S.T., J.H. Cate, and A.E. Dahlberg, *Spontaneous erythromycin resistance mutation in a 23S rRNA gene, rrlA, of the extreme thermophile Thermus thermophilus IB-21*. J Bacteriol, 2001. **183**(14): p. 4382-5.
321. Cavalier-Smith, T., *Predation and eukaryote cell origins: a coevolutionary perspective*. Int J Biochem Cell Biol, 2009. **41**(2): p. 307-22.
322. Martin, W. and E.V. Koonin, *Introns and the origin of nucleus-cytosol compartmentalization*. Nature, 2006. **440**(7080): p. 41-5.
323. Lopez-Garcia, P. and D. Moreira, *Selective forces for the origin of the eukaryotic nucleus*. Bioessays, 2006. **28**(5): p. 525-33.
324. Jekely, G., *Origin of the nucleus and Ran-dependent transport to safeguard ribosome biogenesis in a chimeric cell*. Biol Direct, 2008. **3**: p. 31.
325. Csuros, M. and I. Miklos, *Streamlining and large ancestral genomes in Archaea inferred with a phylogenetic birth-and-death model*. Mol Biol Evol, 2009. **26**(9): p. 2087-95.
326. Krokan, H., E. Wist, and R.H. Krokan, *Aphidicolin inhibits DNA synthesis by DNA polymerase alpha and isolated nuclei by a similar mechanism*. Nucleic Acids Res, 1981. **9**(18): p. 4709-19.
327. Forterre, P., C. Elie, and M. Kohiyama, *Aphidicolin inhibits growth and DNA synthesis in halophilic archaeobacteria*. J Bacteriol, 1984. **159**(2): p. 800-2.
328. Anbar, A.D. and A.H. Knoll, *Proterozoic ocean chemistry and evolution: a bioinorganic bridge?* Science, 2002. **297**(5584): p. 1137-42.

329. Cavalier-Smith, T., *Cell evolution and Earth history: stasis and revolution*. Philos Trans R Soc Lond B Biol Sci, 2006. **361**(1470): p. 969-1006.
330. Blobel, G., *Intracellular protein topogenesis*. Proc Natl Acad Sci U S A, 1980. **77**(3): p. 1496-500.
331. Griffiths, G., *Cell evolution and the problem of membrane topology*. Nat Rev Mol Cell Biol, 2007. **8**(12): p. 1018-24.
332. Kelman, Z. and M.F. White, *Archaeal DNA replication and repair*. Curr Opin Microbiol, 2005. **8**(6): p. 669-76.
333. Sarubbi, E., et al., *Mode of action of the microbial metabolite GE23077, a novel potent and selective inhibitor of bacterial RNA polymerase*. Eur J Biochem, 2004. **271**(15): p. 3146-54.
334. Masullo, M., et al., *Valine 114 replacements in archaeal elongation factor 1 alpha enhanced its ability to interact with aminoacyl-tRNA and kirromycin*. Biochemistry, 2002. **41**(49): p. 14482-8.
335. Koonin, E.V. and A.S. Novozhilov, *Origin and evolution of the genetic code: the universal enigma*. IUBMB Life, 2009. **61**(2): p. 99-111.
336. Higgs, P.G. and R.E. Pudritz, *A thermodynamic basis for prebiotic amino acid synthesis and the nature of the first genetic code*. Astrobiology, 2009. **9**(5): p. 483-90.
337. Trifonov, E.N., *Consensus temporal order of amino acids and evolution of the triplet code*. Gene, 2000. **261**(1): p. 139-51.
338. Lu, Y. and S. Freeland, *On the evolution of the standard amino-acid alphabet*. Genome Biol, 2006. **7**(1): p. 102.
339. Tlusty, T., *Casting polymer nets to optimize noisy molecular codes*. Proc Natl Acad Sci U S A, 2008. **105**(24): p. 8238-43.
340. Hsiao, C., et al., *Peeling the onion: ribosomes are ancient molecular fossils*. Mol Biol Evol, 2009. **26**(11): p. 2415-25.
341. Porter, C.T., G.J. Bartlett, and J.M. Thornton, *The Catalytic Site Atlas: a resource of catalytic sites and residues identified in enzymes using structural data*. Nucleic Acids Res, 2004. **32**(Database issue): p. D129-33.
342. Kanehisa, M., et al., *KEGG for linking genomes to life and the environment*. Nucleic Acids Res, 2008. **36**(Database issue): p. D480-4.

- 343. Lu, Y. and S.J. Freeland, *A quantitative investigation of the chemical space surrounding amino acid alphabet formation*. J Theor Biol, 2008. **250**(2): p. 349-61.
- 344. Seligmann, H., *Cost-minimization of amino acid usage*. J Mol Evol, 2003. **56**(2): p. 151-61.
- 345. Zhang, H.Y., *Exploring the evolution of standard amino-acid alphabet: when genomics meets thermodynamics*. Biochem Biophys Res Commun, 2007. **359**(3): p. 403-5.