

Tensor Decomposition Reveals Concurrent Evolutionary Convergences and Divergences and Correlations with Structural Motifs in Ribosomal RNA

Chaitanya Muralidhara¹, Andrew M. Gross², Robin R. Gutell^{1,3}, Orly Alter^{4*}

1 Institute for Cellular and Molecular Biology, University of Texas at Austin, Austin, Texas, United States of America, **2** Bioinformatics and Systems Biology Program, University of California at San Diego, San Diego, California, United States of America, **3** Section of Integrative Biology and Center for Computational Biology and Bioinformatics, University of Texas at Austin, Austin, Texas, United States of America, **4** Scientific Computing and Imaging (SCI) Institute and Departments of Bioengineering and Human Genetics, University of Utah, Salt Lake City, Utah, United States of America

Abstract

Evolutionary relationships among organisms are commonly described by using a hierarchy derived from comparisons of ribosomal RNA (rRNA) sequences. We propose that even on the level of a single rRNA molecule, an organism's evolution is composed of multiple pathways due to concurrent forces that act independently upon different rRNA degrees of freedom. Relationships among organisms are then compositions of coexisting pathway-dependent similarities and dissimilarities, which cannot be described by a single hierarchy. We computationally test this hypothesis in comparative analyses of 16S and 23S rRNA sequence alignments by using a tensor decomposition, i.e., a framework for modeling composite data. Each alignment is encoded in a cuboid, i.e., a third-order tensor, where nucleotides, positions and organisms, each represent a degree of freedom. A tensor mode-1 higher-order singular value decomposition (HOSVD) is formulated such that it separates each cuboid into combinations of patterns of nucleotide frequency variation across organisms and positions, i.e., "eigenpositions" and corresponding nucleotide-specific segments of "eigenorganisms," respectively, independent of a-priori knowledge of the taxonomic groups or rRNA structures. We find, in support of our hypothesis that, first, the significant eigenpositions reveal multiple similarities and dissimilarities among the taxonomic groups. Second, the corresponding eigenorganisms identify insertions or deletions of nucleotides exclusively conserved within the corresponding groups, that map out entire substructures and are enriched in adenosines, unpaired in the rRNA secondary structure, that participate in tertiary structure interactions. This demonstrates that structural motifs involved in rRNA folding and function are evolutionary degrees of freedom. Third, two previously unknown coexisting subgenetic relationships between Microsporidia and Archaea are revealed in both the 16S and 23S rRNA alignments, a convergence and a divergence, conferred by insertions and deletions of these motifs, which cannot be described by a single hierarchy. This shows that mode-1 HOSVD modeling of rRNA alignments might be used to computationally predict evolutionary mechanisms.

Citation: Muralidhara C, Gross AM, Gutell RR, Alter O (2011) Tensor Decomposition Reveals Concurrent Evolutionary Convergences and Divergences and Correlations with Structural Motifs in Ribosomal RNA. PLoS ONE 6(4): e18768. doi:10.1371/journal.pone.0018768

Editor: Purificación López-García, Université Paris Sud, France

Received: January 7, 2011; **Accepted:** March 17, 2011; **Published:** April 29, 2011

Copyright: © 2011 Muralidhara et al. This is an open-access article distributed under the terms of the Creative Commons Attribution License, which permits unrestricted use, distribution, and reproduction in any medium, provided the original author and source are credited.

Funding: This research was supported by National Institute of General Medical Sciences R01 Grant GM-067317 (to RRG) and National Human Genome Research Institute R01 Grant HG-004302 and National Science Foundation CAREER Award DMS-0847173 (to OA). The funders had no role in study design, data collection and analysis, decision to publish, or preparation of the manuscript.

Competing Interests: The authors have declared that no competing interests exist.

* E-mail: orly@sci.utah.edu

Introduction

The ribosomal RNA (rRNA) is an essential component of the ribosome, the cellular organelle that associates the cell's genotype with its phenotype by catalyzing protein synthesis in all known organisms, and therefore also underlies cellular evolution. RNAs are thought to be among the most primordial macromolecules. This is because an RNA template, similar to a DNA template, can be used to synthesize DNA and RNA, while RNA, similar to proteins, can form three-dimensional structures and catalyze reactions. It was suggested, therefore, that rRNA sequences and structures, that are similar or dissimilar among groups of organisms, are indicative of the relative evolutionary pathways of these organisms [1–3]. Advances in sequencing technologies have resulted in an abundance of rRNA sequences from organisms spanning all taxonomic groups. Today, the small subunit

ribosomal RNA (16S rRNA) is the gene with the largest number of determined sequences.

Evolutionary relationships among organisms are commonly described by using a hierarchy derived from comparisons of rRNA sequences. Secondary and tertiary structure models of rRNAs are also being derived from comparisons of rRNA sequences by using sequence conservation and covariation analyses, and assuming that sequence positions with similar patterns of variation across multiple organisms are base-paired in the rRNA structure [4–7]. The determination of the high-resolution crystal structures of the ribosome [8–10] substantiate these structure models, with approximately 97% of the proposed base pairs present in the crystal structures [11].

We propose that even on the level of a single rRNA molecule, an organism's evolution is composed of multiple pathways [12] due to concurrent evolutionary forces that act independently upon

different rRNA degrees of freedom [13] of multiple types, i.e., different nucleotides in different positions of the molecule, that might correspond to different structural and functional components of the molecule, and in different groups of organisms. The relationships among organisms are then compositions of coexisting evolutionary pathway-dependent rRNA similarities and dissimilarities, which cannot be described by a single hierarchy.

We computationally test this hypothesis in comparative analyses of 16S and 23S rRNA sequence alignments [14] by using a tensor decomposition, i.e., a mathematical framework for modeling composite data with multiple types of degrees of freedom. Each alignment is encoded in a cuboid, i.e., a third-order tensor, where the nucleotides, positions and organisms, each represent a possible degree of freedom. We note that an encoding that transforms a sequence alignment into a second-order tensor, i.e., a matrix, effectively introduces into the analysis assumptions regarding the relations among the nucleotides in the rRNA structure. Similarly, in the analysis of the tensor unfolded into a matrix, some of the degrees of freedom are lost and much of the information in the alignment might also be lost. A tensor mode-1 higher-order singular value decomposition (HOSVD) [15–18] is, therefore, formulated such that it separates each cuboid into combinations of patterns of nucleotide frequency variation across organisms and positions, i.e., “eigenpositions” and corresponding nucleotide-specific segments of “eigenorganisms,” respectively, independent of a-priori knowledge of the taxonomic groups [19] or rRNA structures.

We find, in support of our hypothesis that, first, the significant eigenpositions reveal multiple similarities and dissimilarities among the taxonomic groups, some known and some previously unknown. Second, the corresponding eigenorganisms identify insertions or deletions of nucleotides exclusively conserved within the corresponding groups, that map out entire substructures, some known and some previously unknown, and are enriched in adenosines, unpaired in the rRNA secondary structure, that participate in tertiary structure interactions [20–26]. This demonstrates that structural motifs involved in rRNA folding and function are evolutionary degrees of freedom. Third, two previously unknown coexisting subgenetic relationships between Microsporidia [27–29] and Archaea [30,31] are revealed in both the 16S and 23S rRNAs, a convergence and a divergence, conferred by insertions and deletions of these motifs, which cannot be described by a single hierarchy.

These analyses show that the mode-1 HOSVD provides a mathematical framework for the modeling of rRNA sequence alignments, independent of a-priori knowledge of the taxonomic groups and their relationships, or the rRNA structures, where the mathematical variables represent biological reality [32]. The significant eigenpositions and corresponding nucleotide-specific segments of eigenorganisms represent multiple subgenetic evolutionary relationships of convergence and divergence and correlations with structural motifs, some known and some previously unknown, that are consistent with current biological understanding of the 16S and 23S rRNAs. Mode-1 HOSVD modeling of rRNA sequence alignments, therefore, might be used to computationally predict evolutionary mechanisms, i.e., evolutionary pathways and the underlying structural changes that these pathways are correlated, possibly even coordinated with.

Methods

To computationally test whether the mathematical nature of evolutionary relationships is that of a hierarchy or a composition, we analyze alignments of 339 16S and 75 23S rRNA sequences from the Comparative RNA Website (CRW) [14], representing all 16S and 23S sequences for which a secondary structure model is

available. The organisms corresponding to these sequences are from different National Center for Biotechnology Information (NCBI) Taxonomy Browser groups [19]: The 339 16S organisms include 21 Archaea, 175 Bacteria and 143 Eukarya (Datasets S1, S2, S3); the 75 23S organisms include six Archaea, 57 Bacteria and 12 Eukarya (Datasets S4, S5, S6). The alignments tabulate six sequence elements or “nucleotides,” i.e., A, C, G and U nucleotides, unknown (“N”) and gap (“–”), across 339 organisms and 3249 sequence positions for the 16S, or 75 organisms and 6636 sequence positions for the 23S, with A, C, G or U nucleotides, but not an unknown or a gap, in at least 1% of the 339 or 75 organisms, respectively. A six-bit binary encoding [33],

$$\begin{aligned} A &= (1,0,0,0,0,0) \\ C &= (0,1,0,0,0,0) \\ G &= (0,0,1,0,0,0) \\ U &= (0,0,0,1,0,0) \\ N &= (0,0,0,0,1,0) \\ - &= (0,0,0,0,0,1), \end{aligned} \tag{1}$$

transforms each alignment matrix into a cuboid, i.e., a third-order tensor, of six “slices,” one slice for each nucleotide, tabulating the frequency of this nucleotide across the organisms and positions, where the nucleotides, positions and organisms, each represent a possible degree of freedom (Figure 1 and Mathematica Notebooks S1, S2, S3, S4). Note that an encoding that transforms a sequence alignment into a second-order tensor, i.e., a matrix, effectively introduces into the analysis assumptions regarding the relations among the nucleotides in the RNA structure. Similarly, in the analysis of the tensor unfolded into a matrix, some of the degrees of freedom are lost and much of the information in the alignment might also be lost.

To comparatively analyze the 16S and 23S rRNA alignments, therefore, we use a tensor mode-1 higher-order singular value decomposition (HOSVD) [15,16]. We formulate the mode-1 HOSVD such that it transforms each K -organisms $\times L=6$ -nucleotides $\times M$ -positions tensor \mathcal{D} into the reduced and diagonalized K -“eigenpositions” $\times K$ -“eigenorganisms” matrix Σ , by using the K -eigenorganisms $\times L=6$ -nucleotides $\times M$ -positions transformation tensor \mathcal{U} and the K -organisms $\times K$ -eigenpositions transformation matrix V^T ,

$$\mathcal{D} = \mathcal{U} \Sigma V^T. \tag{2}$$

The mode-1 HOSVD is computed from the singular value decomposition (SVD) [17,18] of each data tensor unfolded along the K -organisms axis such that its nucleotide-specific slices D_i are appended along the organisms axis,

$$\begin{pmatrix} D_A \\ D_C \\ D_G \\ D_U \\ D_N \\ D_- \end{pmatrix} = \begin{pmatrix} U_A \\ U_C \\ U_G \\ U_U \\ U_N \\ U_- \end{pmatrix} \Sigma V^T. \tag{3}$$

The transformation tensor \mathcal{U} is obtained by stacking the nucleotide-specific slices U_i along the organisms axis.

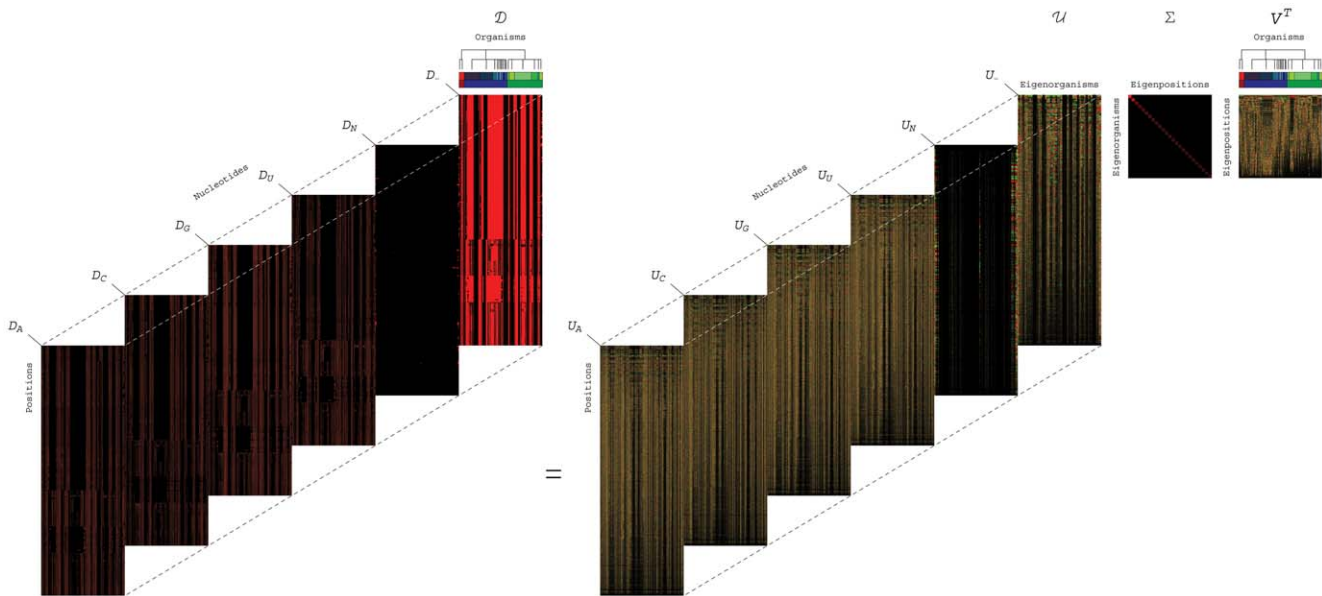


Figure 1. Mode-1 HOSVD of the 16S rRNA sequence alignment. Organisms, positions and sequence elements, each represent a degree of freedom in the alignment encoded in a cuboid (Equation 1). Mode-1 HOSVD (Equation 2) separates the alignment into combinations of “eigenpositions” and nucleotide-specific segments of “eigenorganisms,” i.e., patterns of nucleotide frequency variation across the organisms and positions, with increase (red), no change (black) and decrease in the nucleotide frequency (green) relative to the average frequency across the organisms and positions. It was shown that SVD provides a framework for modeling DNA microarray data [32]: The mathematical variables, significant patterns uncovered in the data, correlate with activities of cellular elements, such as regulators or transcription factors. The mathematical operations simulate experimental observation of the correlations and possibly causal coordination of these activities. Recent experimental results [16] demonstrate that SVD modeling of DNA microarray data can be used to correctly predict previously unknown cellular mechanisms [15,36]. We now show that mode-1 HOSVD, which is computed by using SVD (Equation 3), provides a framework for modeling rRNA sequence alignments: The mathematical variables, significant patterns of nucleotide frequency variation, represent multiple subgenomic evolutionary relationships of convergence and divergence among the organisms, some known and some previously unknown, and correlations with structural motifs. Our mode-1 HOSVD analyses of 16S and 23S rRNA alignments support the hypothesis that even on the level of a single rRNA molecule, an organism’s evolution is composed of multiple pathways due to concurrent forces that act independently upon different rRNA degrees of freedom. These analyses demonstrate that entire rRNA substructures and unpaired adenosines, i.e., rRNA structural motifs which are involved in rRNA folding and function, are evolutionary degrees of freedom. These analyses also show that mode-1 HOSVD modeling of rRNA alignments might be used to computationally predict evolutionary mechanisms, i.e., evolutionary pathways and the underlying structural changes that these pathways are correlated, possibly even coordinated with.
doi:10.1371/journal.pone.0018768.g001

The significance of each eigenposition and the corresponding eigenorganism is defined in terms of the fraction of the overall information that these orthogonal patterns of nucleotide frequency variation across the K -organisms and $L=6$ -nucleotides $\times M$ -positions, respectively, capture in the data tensor and is proportional to the corresponding singular value that is listed in Σ . These singular values are ordered in decreasing order, such that the patterns are ordered in decreasing order of their relative significance. We find that the seven and five most significant eigenpositions and corresponding eigenorganisms uncovered in the 16S and 23S alignments, respectively, capture $\sim 88\%$ and 87% of the nucleotide frequency information in these alignments (Figures S1 and S2 in Appendix S1). In both alignments, the most significant eigenposition is approximately invariant across the organisms and correlates with the average frequency of all nucleotides across the positions [34], with the correlation >0.995 . The correlation of each nucleotide-specific segment of the most significant eigenorganism with the average frequency of this nucleotide across the positions is >0.999 .

We interpret the remaining eigenpositions and the nucleotide-specific segments of the corresponding eigenorganisms as patterns of nucleotide frequency variation relative to these averages. We find that the patterns uncovered in the 16S and 23S are qualitatively similar. Note also that these patterns are robust to variations in the selection of organisms for the 16S and 23S

alignments, and therefore also to variations in the rRNA positions that each alignment spans.

Results

We find, in support of our hypothesis, that, first, the significant eigenpositions uncovered in both the 16S and 23S alignments, starting with the second most significant eigenposition in each alignment (Figure 2, and Figure S3 in Appendix S1) reveal multiple similarities and dissimilarities among the taxonomic groups, some known and some previously unknown. To biologically interpret the mathematical patterns of nucleotide frequency variation across the organisms, i.e., the eigenpositions, we correlate and anticorrelate each eigenposition with increased relative nucleotide frequency across a taxonomic group according to the NCBI Taxonomy Browser annotations [19] of the two groups of k organisms each, with largest and smallest levels of nucleotide frequency in this eigenposition among all K organisms, respectively. The P -value of a given association is calculated assuming hypergeometric probability distribution of the J annotations among the K organisms, and of the subset of $j \subseteq J$ annotations among the subset of k organisms, as described [35], $P(j; k, K, J) = \binom{K}{k}^{-1} \sum_{i=j}^k \binom{J}{i} \binom{K-J}{k-i}$ (Tables 1 and 2).

In both alignments, the second most significant eigenposition captures the dissimilar between the Eukarya excluding the

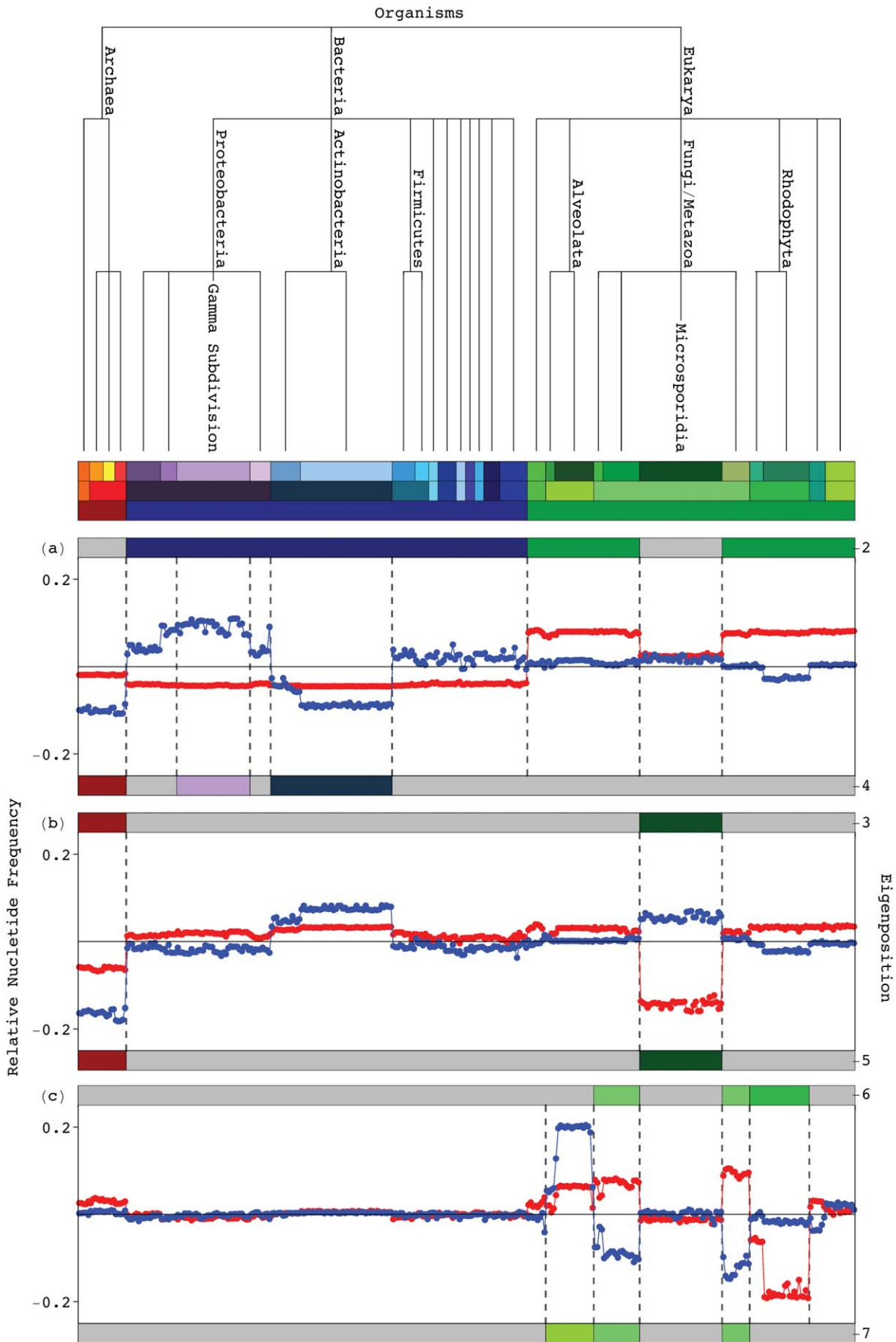


Figure 2. Significant 16S eigenpositions. Line-joined graphs of the second through seventh 16S eigenpositions, i.e., patterns of nucleotide frequency across the organisms, and their correlation with the taxonomic groups in the 16S alignment, classified according to the top six hierarchical levels of the NCBI Taxonomy Browser [19] (Figure S1 in Appendix S1). (a) The second most significant eigenposition (red) differentiates the Eukarya excluding the Microsporidia from the Bacteria, as indicated by the color bar (Table 1). The fourth (blue) distinguishes between the Gamma Proteobacteria and the Actinobacteria and Archaea. (b) The third (red) and fifth (blue) eigenpositions describe similarities and dissimilarities among the Archaea and Microsporidia, respectively. (c) The sixth (red) and seventh (blue) eigenpositions differentiate the Fungi/Metazoa excluding the Microsporidia from the Rhodophyta and the Alveolata, respectively.
doi:10.1371/journal.pone.0018768.g002

Microsporidia and the Bacteria. These patterns of relative nucleotide frequency across the organisms correlate with increased frequency across the Eukarya excluding the Microsporidia and decreased frequency across the Bacteria, with both P -values $< 10^{-25}$ and $< 10^{-2}$ in the 16S and 23S alignments, respectively. The fourth 16S eigenposition correlates with increased nucleotide frequency across the Gamma Proteobacteria and decreased frequency across the Actinobacteria and Archaea, with both P -values $< 10^{-23}$. Note that the Gamma Proteobacteria and the Actinobacteria are the two largest bacterial groups in this alignment. The fourth 23S eigenposition captures the dissimilar between the Proteobacteria and the Firmicutes, the two largest bacterial groups in this alignment. In both alignments, the third and fifth eigenpositions capture the similar and dissimilar between the Archaea and Microsporidia, respectively. In the 16S alignment, the sixth and seventh eigenpositions identify dissimilarities among the Fungi/Metazoa excluding the Microsporidia and the Rhodophyta and separately the Alveolata, respectively.

Second, we find that, consistent with the eigenpositions, the corresponding 16S and 23S eigenorganisms identify positions of nucleotides that are approximately conserved within the respective taxonomic groups, but not among them. These positions are significantly enriched in conserved sequence gaps which map out entire substructures inserted or deleted in the 16S and 23S rRNAs of one taxonomic group relative to another as well as adenosines, unpaired in the rRNA secondary structure, most of which participate in tertiary structure interactions and map to the same substructures. To biologically interpret the mathematical patterns of nucleotide-specific frequency variation across the positions, i.e., the eigenorganisms, we consider the m positions with largest increase or decrease in the relative nucleotide frequency in each nucleotide-specific segment of each eigenorganism (Tables 3 and 4). These positions exhibit the frequency variations across the

organisms that are most correlated or anticorrelated, respectively, with the corresponding eigenposition. We calculate the P -value of the enrichment of these positions in sequence and structure motifs conserved across the corresponding taxonomic groups by assuming hypergeometric probability distribution of the N conserved motifs among the M positions, and of the subset of $n \subseteq N$ motifs among the subset of m positions, as described [35], $P(n; m, M, N) = \binom{M}{m}^{-1} \sum_{i=n}^m \binom{N}{i} \binom{M-N}{m-i}$.

The positions identified by the eigenorganisms include entire substructures inserted or deleted in the structure of one taxonomic group relative to another. Consider for example the 124 positions with largest nucleotide frequency increase in the gap segment of the second most significant 16S eigenorganism, i.e., the positions for which the frequency of gaps across the organisms is most correlated with the second eigenposition. These positions are enriched in sequence gaps conserved in the Eukarya excluding the Microsporidia (Figure S4a in Appendix S1). The 100 positions with largest frequency decrease are enriched in gaps conserved in the Bacteria (Figure S4b in Appendix S1). Both P -values $< 10^{-93}$. Mapped onto the secondary structure models of the bacterium *E. coli* and the eukaryote *S. cerevisiae* [14], these positions map out known as well as previously unrecognized insertions and deletions of not only isolated nucleotides but entire substructures in the Eukarya with respect to the Bacteria [31] (Figure 3). Similarly, the positions identified by the gap segment of the second 23S eigenorganism map out entire substructures inserted and deleted in 23S rRNAs of the Eukarya relative to the Bacteria (Figures S5 and S6 in Appendix S1).

In addition, the eigenorganisms identify adenosines that are unpaired in the rRNA secondary structure and are conserved exclusively in the respective taxonomic groups. The majority of these adenosines participate in tertiary structure interactions, and

Table 1. Enrichment of the significant 16S eigenpositions in taxonomic groups.

Eigenposition	Correlated				Anticorrelated			
	Taxonomic Group	j	J	P -value	Taxonomic Group	j	J	P -value
2	Eukarya – Microsporidia	75	107	5.7×10^{-50}	Bacteria	75	175	1.5×10^{-26}
3					Archaea + Microsporidia	57	57	3.4×10^{-49}
4	Gamma Proteobacteria	32	32	2.0×10^{-24}	Actinobacteria + Archaea	72	74	2.5×10^{-67}
5	Microsporidia	29	36	1.8×10^{-15}	Archaea	21	21	1.5×10^{-15}
6	Fungi/Metazoa – Microsporidia	32	32	2.0×10^{-24}	Rhodophyta	26	26	1.8×10^{-19}
7	Alveolata	21	21	1.5×10^{-15}	Fungi/Metazoa – Microsporidia	32	32	2.0×10^{-24}

Probabilistic significance of the enrichment of the $k=75$ organisms, with largest relative nucleotide frequency increase or decrease in each of the significant eigenpositions, in the respective taxonomic groups. The P -value of each enrichment is calculated as described [35], assuming hypergeometric probability distribution of the J annotations among the $K=339$ organisms, and of the subset of $j \subseteq J$ annotations among the subset of $k=75$ organisms.

doi:10.1371/journal.pone.0018768.t001

Table 2. Enrichment of the significant 23S eigenpositions in taxonomic groups.

Eigenposition	Correlated				Anticorrelated			
	Taxonomic Group	<i>j</i>	<i>J</i>	<i>P</i> -value	Taxonomic Group	<i>j</i>	<i>J</i>	<i>P</i> -value
2	Eukarya – Microsporidia	8	8	3.8×10^{-7}	Bacteria	15	57	9.7×10^{-3}
3					Archaea + Microsporidia	10	10	3.6×10^{-9}
4	Proteobacteria	15	23	2.2×10^{-10}	Firmicutes	12	13	2.2×10^{-10}
5	Microsporidia	4	4	1.1×10^{-3}	Archaea	6	6	2.5×10^{-5}

Probabilistic significance of the enrichment of the $k = 15$ organisms, with largest relative nucleotide frequency increase or decrease in each of the significant eigenpositions, in the respective taxonomic groups. The *P*-value of each enrichment is calculated as described [35], assuming hypergeometric probability distribution of the *J* annotations among the $K = 75$ organisms, and of the subset of $j \subseteq J$ annotations among the subset of $k = 15$ organisms.
doi:10.1371/journal.pone.0018768.t002

some also map to the same substructures. Previous comparative studies observed that the A nucleotides are usually unpaired in the secondary structure models of rRNA, while most of the C, G and

U nucleotides are base-paired [6]. It was also noted that these unpaired adenosines are especially abundant in tertiary structure motifs, such as tetraloops, i.e., the four-base hairpin loops that cap

Table 3. Enrichment of the significant 16S eigenorganisms sequence and structure motifs exclusively conserved in taxonomic groups.

Eigenorganism	Nucleotide Segment	Structure Motif	Correlated				Anticorrelated			
			Taxonomic Group	<i>n</i>	<i>N</i>	<i>P</i> -value	Taxonomic Group	<i>n</i>	<i>N</i>	<i>P</i> -value
2	A	Unpaired A	Eukarya – Microsporidia	48	66	2.3×10^{-63}	Bacteria	50	50	1.2×10^{-82}
	Gap	Gap		*124	211	4.0×10^{-167}		57	58	9.8×10^{-94}
		Unpaired A	Bacteria	*13	50	2.1×10^{-8}				
3	C	Helix					Archaea + Microsporidia	76	1148	4.0×10^{-17}
	G							68	1148	1.5×10^{-11}
	U							65	1148	8.5×10^{-10}
	Gap	Gap						6	6	7.3×10^{-10}
4	A	Unpaired A	Gamma Proteobacteria	11	11	1.4×10^{-17}				
	Gap	Helix	Actinobacteria + Archaea	34	153	8.6×10^{-22}				
5	A	Unpaired A					Archaea	14	14	2.7×10^{-22}
	G	Helix						84	933	1.8×10^{-31}
	C		Microsporidia	58	947	9.6×10^{-10}		85	933	1.3×10^{-32}
	U			55	947	3.6×10^{-8}		57	933	1.8×10^{-9}
	Gap	Unpaired A	Archaea	7	14	6.1×10^{-8}				
6	A		Fungi/Metazoa – Microsporidia	9	16	1.7×10^{-10}	Rhodophyta	25	27	2.2×10^{-37}
7			Alveolata	25	31	4.3×10^{-34}	Fungi/Metazoa – Microsporidia	10	16	3.3×10^{-12}

Probabilistic significance of the enrichment of the $m = 100$ positions with largest frequency increase or decrease in each of the nucleotide-specific segments of the significant eigenorganisms (except for the gap segment of the second eigenorganism, where the largest nucleotide frequency increase is shared by * $m = 124$ positions) in sequence and structure motifs exclusively conserved in the corresponding taxonomic groups. The *P*-value of each enrichment is calculated assuming, for each nucleotide, hypergeometric probability distribution of the *N* conserved motifs among the $M = 3249$ positions, and of the subset of $n \subseteq N$ motifs among the subset of *m* positions. Exclusive sequence gap conservation is defined as conservation of gaps within at least 80% of the organisms of the corresponding taxonomic group but in less than 20% of the remaining organisms. Exclusive unpaired A nucleotide conservation is defined as conservation of an adenosine within at least 80% of the organisms of the group but in less than 20% of the remaining organisms, together with greater frequency of unpaired nucleotides within the group rather than among the remaining organisms. Helix conservation is defined as conservation of base-paired nucleotides within at least 60% of the organisms of the group.
doi:10.1371/journal.pone.0018768.t003

Table 4. Enrichment of the significant 23S eigenorganisms in sequence and structure motifs exclusively conserved in taxonomic groups.

Eigenorganism	Nucleotide Segment	Structure Motif	Correlated				Anticorrelated			
			Taxonomic Group	<i>n</i>	<i>N</i>	<i>P</i> -value	Taxonomic Group	<i>n</i>	<i>N</i>	<i>P</i> -value
2	A	Unpaired A	Eukarya – Microsporidia	59	59	1.7×10^{-94}	Bacteria	41	41	6.1×10^{-65}
	Gap	Gap		136	145	2.3×10^{-220}		*14	14	9.8×10^{-27}
		Unpaired A	Bacteria	15	41	2.9×10^{-13}	Eukarya – Microsporidia	*8	59	1.1×10^{-6}
3	A			28	41	4.8×10^{-34}	Archaea + Microsporidia	11	11	1.4×10^{-17}
	Gap	Gap		12	14	3.5×10^{-17}		*41	45	2.2×10^{-74}
		Unpaired A					Bacteria	*8	41	1.3×10^{-7}
4	A		Proteobacteria	8	8	5.9×10^{-13}	Firmicutes	5	5	2.4×10^{-8}
5			Microsporidia	16	31	5.1×10^{-17}	Archaea	39	49	6.6×10^{-52}
	Gap	Gap		191	387	3.3×10^{-245}		*15	59	1.1×10^{-10}
		Unpaired A					Microsporidia	*9	31	1.9×10^{-7}

Probabilistic significance of the enrichment of the $m=200$ positions with largest frequency increase or decrease in each of the nucleotide-specific segments of the significant eigenorganisms (except for the gap segments of the second, third and fifth eigenorganisms, where the largest nucleotide frequency decrease is shared by $m=91, 100$ and 199 positions, respectively) in sequence and structure motifs exclusively conserved in the corresponding taxonomic groups. The *P*-value of each enrichment is calculated assuming, for each nucleotide, hypergeometric probability distribution of the *N* conserved motifs among the $M=6636$ positions, and of the subset of $n \leq N$ motifs among the subset of *m* positions. Exclusive sequence gap conservation is defined as conservation of gaps within at least 80% of the organisms of the corresponding taxonomic group but in less than 20% of the remaining organisms. Exclusive unpaired A nucleotide conservation is defined as conservation of an adenosine within at least 80% of the organisms of the group but in less than 20% of the remaining organisms, together with greater frequency of unpaired nucleotides within the group rather than among the remaining organisms.

doi:10.1371/journal.pone.0018768.t004

many rRNA double helices [7]. Experimental observations of intra- and intermolecular interactions involving these loops and other motifs rich in unpaired adenosines [20] suggested a role for these unpaired nucleotides in a universal mode of RNA helical packing [21,22] as well as in the accuracy and specificity of the translational function of the rRNA protein synthesis [23–26].

We find the positions with largest nucleotide frequency increase in the A segment of the second 16S eigenorganism to be enriched in unpaired adenosines, which are exclusively conserved in the Eukarya excluding the Microsporidia (Figures S7 and S8 in Appendix S1), whereas the positions with largest decrease include all 50 unpaired adenosines exclusively conserved in the Bacteria (Figure 4). Both *P*-values $< 10^{-62}$. Note that 13 of these 50 unpaired adenosines exclusively conserved in the Bacteria, map to rRNA substructures that are deleted in Eukarya, with the corresponding *P*-value $< 10^{-7}$. Of these 50 unpaired adenosines, the crystal structure of the bacterium *T. thermophilus* [8] reveals that 28 are involved in tertiary base-base or base-backbone interactions. Similarly, the positions identified by the A nucleotide segment of the second 23S eigenorganism are enriched in unpaired adenosines exclusively conserved in the Eukarya excluding the Microsporidia and in the Bacteria, most of which map to the substructures inserted or deleted in the Eukarya with respect to the Bacteria, respectively (Figures S5 and S9 in Appendix S1).

We find a similar enrichment of unpaired A nucleotides exclusively conserved in the taxonomic groups identified by the fourth through seventh 16S eigenpositions and by the third through fifth 23S eigenpositions. The 100 positions with largest frequency increase or decrease in the A nucleotide segment of the fourth, fifth, sixth or seventh 16S eigenorganism, i.e., the positions for which the A nucleotide frequency across the organisms is most correlated or anticorrelated, respectively, with the fourth, fifth, sixth or seventh

eigenposition, include all or most of the unpaired A nucleotides exclusively conserved in either the Gamma Proteobacteria, Archaea, Rhodophyta, Alveolata or Fungi/Metazoa excluding the Microsporidia, respectively, with all *P*-values $< 10^{-9}$. The 200 positions with largest frequency increase or decrease in the A nucleotide segment of the third, fourth or fifth 23S eigenorganism include all or most of the unpaired A nucleotides exclusively conserved in either the Proteobacteria, Firmicutes, Archaea or Microsporidia, respectively, with all *P*-values $< 10^{-8}$.

These results demonstrate that an organism's evolutionary pathway is correlated and possibly also causally coordinated with insertions or deletions of entire rRNA substructures and unpaired adenosines, i.e., these structural motifs which are involved in rRNA folding and function are evolutionary degrees of freedom.

Third, we find that, two previously unknown coexisting subgenetic relationships between Microsporidia [27–29] and Archaea [30,31] are revealed in both the 16S and 23S rRNAs, a convergence and a divergence, that are conferred by insertions and deletions of these structural motifs. The Microsporidia are eukaryotic single cell intracellular parasites of small genomes, ribosomes and rRNAs, which lack mitochondria and most other membrane-bound cellular organelles. Early comparative studies of single rRNA molecules suggested that the Microsporidia are most dissimilar to all other eukaryotes [27], while more recent comparative studies of multiple genes revealed that Microsporidia are most similar to the Fungi [28,29]. The Archaea are single cell prokaryotes of extremely small genomes. Archaeal rRNAs are more similar to bacterial rather than eukaryotic rRNAs. Archaeal ribosomal proteins, however, are more similar to eukaryotic rather than bacterial ribosomal proteins [30,31].

In both 16S and 23S alignments, the third most significant eigenposition captures the similarities among the two taxonomic

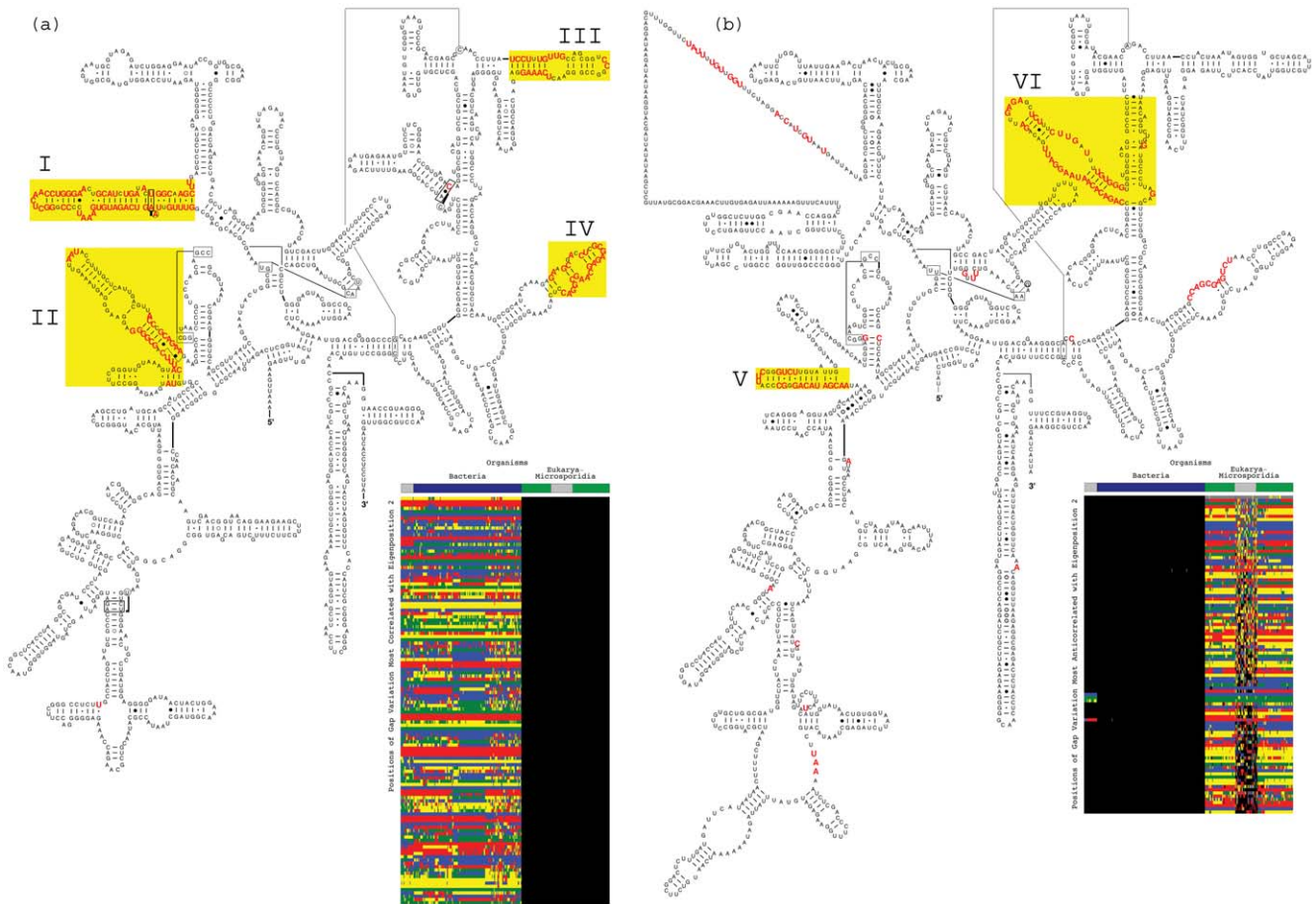


Figure 3. Sequence gaps exclusive to Eukarya or Bacteria 16S rRNAs. The second most significant 16S eigenorganism identifies gaps exclusively conserved in either the Eukarya excluding the Microsporidia or the Bacteria (Table 3) that map out known as well as previously unrecognized, entire substructures deleted or inserted, respectively, in the Eukarya relative to the Bacteria. (a) The 124 positions with largest increase in relative nucleotide frequency in the gap segment of the second eigenorganism, i.e., the 124 positions of gap variation across the organisms most correlated with the second eigenposition, map out the exclusively conserved known substructures [31] I and II and the previously unrecognized substructures III and IV in the secondary structure model of the bacterium *E. coli* [14]. These 124 positions are also displayed in the inset raster, ordered by their significance, with the most significant position at the top. The nucleotides are color-coded A (red), C (green), G (blue), U (yellow), unknown (gray) and gap (black). The color bars highlight the taxonomic groups that are differentiated by the second 16S eigenposition and eigenorganism, i.e., the Eukarya excluding the Microsporidia and the Bacteria. (b) Of the 100 positions of gap variation across the organisms most anticorrelated with the second eigenposition, 99 map out the substructures V and VI in the secondary structure model of the eukaryote *S. cerevisiae*. The 100th position is an unknown nucleotide at the 3'-end of the molecule, which is not displayed. These 100 positions are also displayed in the inset raster.

doi:10.1371/journal.pone.0018768.g003

groups and correlates with decreased nucleotide frequency across both the Archaea and Microsporidia relative to all other organisms with the P -values $< 10^{-23}$ and 10^{-9} , respectively. The 100 positions with largest nucleotide frequency decrease in the gap segment of the third 16S eigenorganism identify all six gaps exclusively conserved in both the Archaea and Microsporidia with the corresponding P -value $< 10^{-9}$. Mapped onto the secondary structure model of the bacterium *E. coli*, these 100 positions identify deletions of not only isolated nucleotides but entire substructures in the Archaea and Microsporidia with respect to the Bacteria (Figure 5a), indicating a convergent loss in both the Archaea and Microsporidia with respect to the Bacteria as well as the Eukarya. We observe that these same positions (Figure 5b) also identify nucleotides that are deleted in the metazoan mitochondrial 16S rRNA sequences (Figure 5c, and Dataset S7), suggesting that these similarities among the Archaea and Microsporidia may be explained by evolutionary forces that act to reduce the genome sizes of the Archaea and Microsporidia. Similarly, the 100 positions with largest nucleotide frequency

decrease in the gap segment of the third 23S eigenorganism identify 41 of the 45 gaps and all 11 unpaired adenosines that are exclusively conserved in both the Archaea and Microsporidia with the corresponding P -values $< 10^{-16}$ (Figures S10 and S11 in Appendix S1). Note that in the 16S alignment, the third eigenorganism also identifies positions of helices, i.e., base-paired nucleotides, that are exclusively conserved in both the Archaea and Microsporidia (Figure S12 in Appendix S1).

The fifth 16S and 23S eigenpositions both capture the dissimilarities between Archaea and Microsporidia and correlate with increased and decreased frequency across the Microsporidia and the Archaea, with the P -values $< 10^{-14}$ and 10^{-2} , respectively. In the gap segment of the 16S fifth eigenorganism, the 100 positions with largest nucleotide frequency increase include seven of the 14 unpaired A nucleotides exclusively conserved in the Archaea, implying that these seven unpaired adenosines are exclusively missing in the Microsporidia (Figure S13 in Appendix S1). The 100 positions with largest nucleotide

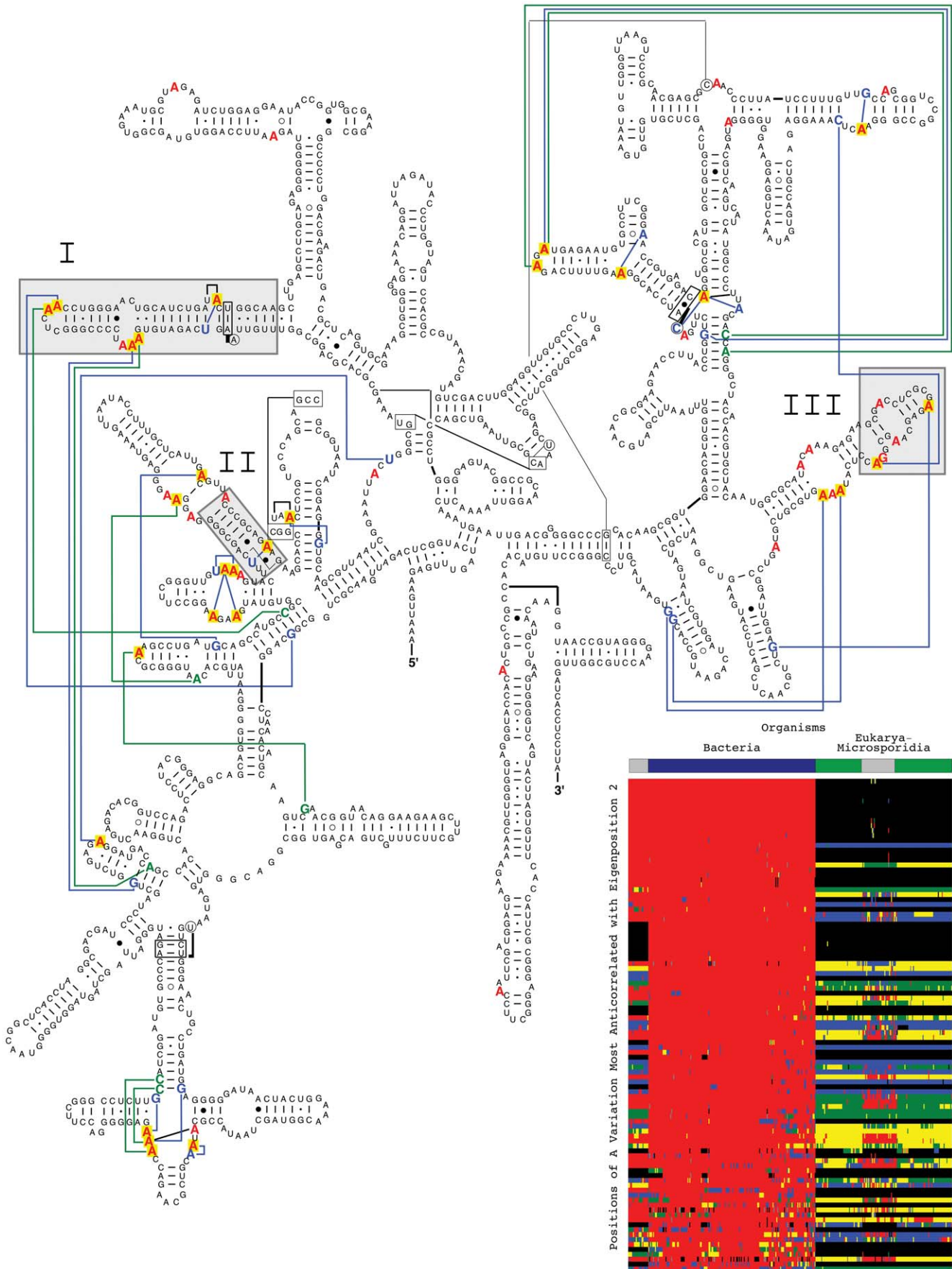


Figure 4. Unpaired adenosines exclusive to Bacteria 16S rRNAs. The 100 positions identified in the A nucleotide segment of the second 16S eigenorganism with the largest decrease in relative nucleotide frequency include all 50 positions (red) in the alignment with unpaired A nucleotides exclusively conserved in the Bacteria. Of these 50 positions, 28 (yellow) map to known tertiary interactions in the crystal structure of the bacterium *T. thermophilus* [8], plotted on the secondary structure model of the bacterium *E. coli* [14]. These include 22 base-base interactions (blue) and eight base-backbone interactions (green). Of the 50 positions of unpaired A nucleotides exclusively conserved in the Bacteria, 13 correspond to gaps exclusively conserved in the Eukarya excluding the Microsporidia. These 13 positions map to the entire 16S rRNA substructures that are deleted in the Eukarya with respect to the Bacteria (gray), identified by the gap segment of the second eigenorganism (Figure 3). These 100 positions identified in the A nucleotide segment of the second 16S eigenorganism are displayed in the inset raster, ordered by their significance, with the most significant position at the top. The nucleotides are color-coded A (red), C (green), G (blue), U (yellow), unknown (gray) and gap (black). The color bars highlight the taxonomic groups that are differentiated by the second eigenposition and eigenorganism, i.e., the Eukarya excluding the Microsporidia and the Bacteria.

doi:10.1371/journal.pone.0018768.g004

frequency decrease in the A nucleotide segment of this eigenorganism include all 14 unpaired A nucleotides exclusively conserved in the Archaea (Figure S14 in Appendix S1). These same positions in the mitochondrial 16S rRNA do not follow a trend similar to either the Archaea or the Microsporidia. Similarly, the gap segment of the 23S fifth eigenorganism identifies 191 of the 387 and 15 of the 59 sequence gaps exclusive to the Microsporidia and the Archaea, respectively, with both P -values $< 10^{-9}$, mapping out entire substructures deleted and inserted in the Microsporidia relative to the Archaea. The A nucleotide segment of this eigenorganism identifies 16 of the 31 and 39 of the 49 unpaired adenosines exclusively conserved in the Microsporidia and Archaea, respectively, with both P -values $< 10^{-16}$ (Figures S15–S17 in Appendix S1).

Together, the third and fifth eigenpositions and eigenorganisms, in both the 16S and 23S rRNA alignments, reveal two previously unknown coexisting subgenic relationships of similarity and dissimilarity, i.e., convergence and divergence, between the Archaea and Microsporidia. These two relationships might

correspond to two independent evolutionary pathways. The similarity among Microsporidia and Archaea in terms of their 16S and 23S rRNAs may be explained by evolutionary forces that act to reduce the genome sizes of these organisms. A single hierarchy cannot describe both these relationships of coexisting pathway-dependent similarity and dissimilarity.

These results support our hypothesis that even on the level of a single rRNA molecule, an organism's evolution is composed of multiple pathways due to concurrent evolutionary forces that act independently upon different rRNA degrees of freedom.

Discussion

It was shown that the SVD provides a mathematical framework for the modeling of DNA microarray data, where the mathematical variables and operations represent biological reality [32]: The variables, significant patterns uncovered in the data, correlate with activities of cellular elements, such as regulators or transcription factors. The operations, such as classification, rotation or

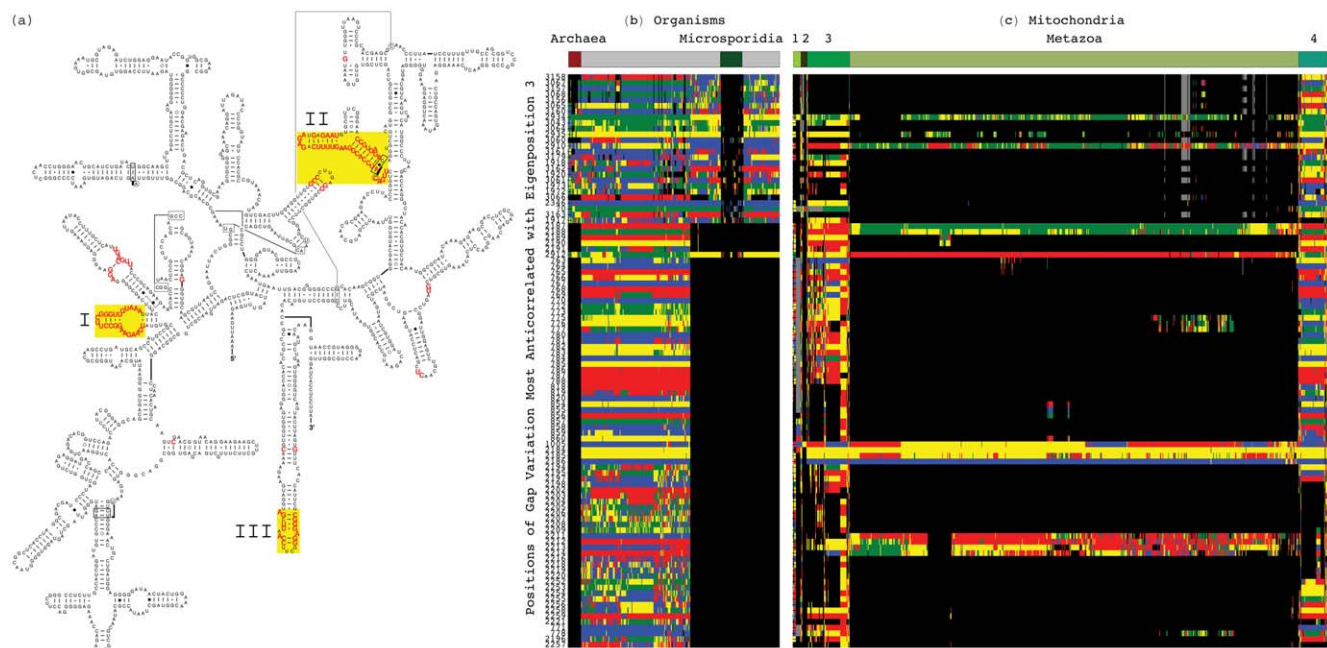


Figure 5. Sequence gaps exclusive to both Archaea and Microsporidia 16S rRNAs. The 100 positions identified in the gap segment of the third 16S eigenorganism with the largest decrease in relative nucleotide frequency map out entire substructures in the Bacteria 16S rRNAs that are convergently lost in the Archaea and the Microsporidia. (a) The 100 gaps conserved in both the Archaea and Microsporidia map to the entire substructures I–III in the secondary structure model of the bacterium *E. coli* [14]. (b) Raster display of the 100 positions of conserved gaps in both the Archaea and Microsporidia across the alignment. (c) Raster display of the same 100 positions across an alignment of 858 mitochondrial 16S rRNA sequences show gaps conserved in most Metazoa. The other groups of Eukarya represented in the mitochondrial alignment are Alveolata (1), Euglenozoa (2), Fungi (3) and Rhodophyta and Viridiplantae (4). The nucleotides are color-coded A (red), C (green), G (blue), U (yellow), unknown (gray) and gap (black). The color bars highlight the taxonomic groups.

doi:10.1371/journal.pone.0018768.g005

reconstruction in subspaces of these patterns, were shown to simulate experimental observation of the correlations and possibly even the causal coordination of these activities. Recent experimental results [16] verify a computationally predicted mechanism of regulation [15,36], demonstrating that SVD modeling of DNA microarray data can be used to correctly predict previously unknown cellular mechanisms.

We now show that the mode-1 HOSVD, which is computed by using the SVD, provides a mathematical framework for the modeling of rRNA sequence alignments, independent of a-priori knowledge of the taxonomic groups and their relationships, or the rRNA structures, where the mathematical variables, significant eigenpositions and corresponding nucleotide-specific segments of eigenorganisms, represent multiple subgenetic evolutionary relationships of convergence and divergence, some known and some previously unknown, and correlations with structural motifs, that are consistent with current biological understanding of the 16S and 23S rRNAs.

Our mode-1 HOSVD analyses of 16S and 23S rRNA sequence alignments support the hypothesis that even on the level of a single rRNA molecule, an organism's evolution is composed of multiple pathways due to concurrent evolutionary forces that act independently upon different rRNA degrees of freedom. These analyses demonstrate that entire rRNA substructures and unpaired adenosines, i.e., rRNA structural motifs which are involved in rRNA folding and function, are evolutionary degrees of freedom. These analyses also show that mode-1 HOSVD modeling of rRNA sequence alignments might be used to computationally predict evolutionary mechanisms, i.e., evolutionary pathways and the underlying structural changes that these pathways are correlated, possibly even coordinated with.

Supporting Information

Appendix S1 A PDF format file, readable by Adobe Acrobat Reader.
(PDF)

Mathematica Notebook S1 Mode-1 higher-order singular value decomposition (HOSVD) of the 16S rRNA alignment. A Mathematica 8 code file, executable by Mathematica 8 and readable by Mathematica Player, freely available at <http://www.wolfram.com/products/player/>.
(SITX)

Mathematica Notebook S2 Mode-1 higher-order singular value decomposition (HOSVD) of the 16S rRNA alignment. A PDF format file, readable by Adobe Acrobat Reader.
(PDF)

Mathematica Notebook S3 Mode-1 HOSVD of the 23S rRNA alignment. A Mathematica 8 code file, executable by Mathematica 8 and readable by Mathematica Player, freely available at <http://www.wolfram.com/products/player/>.
(NB)

Mathematica Notebook S4 Mode-1 HOSVD of the 23S rRNA alignment. A PDF format file, readable by Adobe Acrobat Reader.
(PDF)

Dataset S1 Taxonomy annotations of the organisms in the 16S rRNA alignment. A tab-delimited text format file, readable by both Mathematica and Microsoft Excel, reproducing the National Center for Biotechnology Information (NCBI) Taxonomy Browser [19] annotations of the 339 organisms in the 16S alignment.
(TXT)

Dataset S2 16S rRNA alignment. Tab-delimited text format files, readable by both Mathematica and Microsoft Excel, reproducing the alignment of 16S rRNA sequences from the Comparative RNA Website (CRW) [14], tabulating six sequence elements, i.e., A, C, G and U nucleotides, unknown ("N") and gap ("—"), across the 339 organisms and the 3249 sequence positions.
(TXT)

Dataset S3 Base-pairing of the positions of the 16S rRNA alignment. Tab-delimited text format files, readable by both Mathematica and Microsoft Excel, reproducing the base-pairing of the positions of the 16S rRNA alignment in the secondary structure models of the 16S sequences from the CRW [14], tabulating base-paired ("Y") and unpaired ("N") nucleotides as well as gaps ("—"), across the 339 organisms and the 3249 sequence positions.
(TXT)

Dataset S4 Taxonomy annotations of the organisms in the 23S rRNA alignment. A tab-delimited text format file, readable by both Mathematica and Microsoft Excel, reproducing the NCBI Taxonomy Browser [19] annotations of the 75 organisms in the 23S alignment.
(TXT)

Dataset S5 23S rRNA alignment. Tab-delimited text format files, readable by both Mathematica and Microsoft Excel, reproducing the alignment of 23S rRNA sequences from the CRW [14], tabulating six sequence elements, i.e., A, C, G and U nucleotides, unknown ("N") and gap ("—"), across the 75 organisms and the 6636 sequence positions.
(TXT)

Dataset S6 Base-pairing of the positions of the 16S rRNA alignment. Tab-delimited text format files, readable by both Mathematica and Microsoft Excel, reproducing the base-pairing of the positions of the 23S rRNA alignment in the secondary structure models of the 23S sequences from the CRW [14], tabulating base-paired ("Y") and unpaired ("N") nucleotides as well as gaps ("—"), across the 75 organisms and the 6636 sequence positions.
(TXT)

Dataset S7 Mitochondrial 16S rRNA alignment with taxonomy annotations of the organisms. Tab-delimited text format files, readable by both Mathematica and Microsoft Excel, reproducing the alignment of 858 mitochondrial 16S rRNA sequences from the CRW [14], tabulating six sequence elements across the 858 organisms and the 3249 sequence positions, as well as reproducing the NCBI Taxonomy Browser [19] annotations of the 858 organisms.
(TXT)

Acknowledgments

We thank GH Golub for introducing us to matrix and tensor computations, and the American Institute of Mathematics in Palo Alto and Stanford University for hosting the 2004 Workshop on Tensor Decompositions and the 2006 Workshop on Algorithms for Modern Massive Data Sets, respectively, where some of this work was done. We also thank JJ Cannone for technical assistance.

Author Contributions

Conceived and designed the experiments: OA. Performed the experiments: CM AMG OA. Analyzed the data: CM AMG RRG OA. Contributed reagents/materials/analysis tools: AMG RRG OA. Wrote the paper: CM AMG RRG OA.

References

1. Woese CR (1967) The genetic code: The molecular basis for genetic expression. New York: Harper & Row. 200 p.
2. Crick FH (1968) The origin of the genetic code. *J Mol Biol* 38: 367–379.
3. Orgel LE (1968) Evolution of the genetic apparatus. *J Mol Biol* 38: 381–393.
4. Gutell RR, Power A, Hertz GZ, Putz EJ, Stormo GD (1992) Identifying constraints on the higher-order structure of RNA: continued development and application of comparative sequence analysis methods. *Nucleic Acids Res* 20: 5785–5795.
5. Eddy SR, Durbin R (1994) RNA sequence analysis using covariance models. *Nucleic Acids Res* 22: 2079–2088.
6. Gutell RR, Cannone JJ, Shang Z, Du Y, Serra MJ (2000) A story: unpaired adenosine bases in ribosomal RNAs. *J Mol Biol* 304: 335–354.
7. Woese CR, Winker S, Gutell RR (1990) Architecture of ribosomal RNA: Constraints on the sequence of tetra-loops. *Proc Natl Acad Sci USA* 87: 8467–8471.
8. Schluenzen F, Tocilj A, Zarivach R, Harms J, Gluehmann M, et al. (2000) Structure of functionally activated small ribosomal subunit at 3.3 angstroms resolution. *Cell* 102: 615–623.
9. Ban N, Nissen P, Hansen J, Moore PB, Steitz TA (2000) The complete atomic structure of the large ribosomal subunit at 2.4 Å resolution. *Science* 289: 905–920.
10. Wimberly BT, Brodersen DE, Clemons WM Jr., Morgan-Warren RJ, Carter AP, et al. (2000) Structure of the 30S ribosomal subunit. *Nature* 407: 327–339.
11. Gutell RR, Lee JC, Cannone JJ (2002) The accuracy of ribosomal RNA comparative structure models. *Curr Opin Struct Biol* 12: 301–310.
12. Bock WJ (1959) Preadaptation and multiple evolutionary pathways. *Evolution* 13: 194–211.
13. Ridley CP, Lee HY, Khosla C (2008) Evolution of polyketide synthases in bacteria. *Proc Natl Acad Sci USA* 105: 4595–4600.
14. Cannone JJ, Subramanian S, Schnare MN, Collett JR, D'Souza LM, Du Y, et al. (2002) The comparative RNA web (CRW) site: an online database of comparative sequence and structure information for ribosomal, intron, and other RNAs. *BMC Bioinformatics* 3: 15.
15. Omberg L, Golub GH, Alter O (2007) Tensor higher-order singular value decomposition for integrative analysis of DNA microarray data from different studies. *Proc Natl Acad Sci USA* 104: 18371–18376.
16. Omberg L, Meyerson JR, Kobayashi K, Drury LS, Diffey JF, et al. (2009) Global effects of DNA replication and DNA replication origin activity on eukaryotic gene expression. *Mol Syst Biol* 5: 312.
17. Golub GH, Van Loan CF (1996) *Matrix Computations*. Baltimore: Johns Hopkins University Press, third edition. 694 p.
18. Alter O, Brown PO, Botstein D (2000) Singular value decomposition for genome-wide expression data processing and modeling. *Proc Natl Acad Sci USA* 97: 10101–10106.
19. Sayers EW, Barrett T, Benson DA, Bolton E, Bryant SH, et al. (2010) Database resources of the National Center for Biotechnology Information. *Nucleic Acids Res* 38: D5–D16.
20. Cate JH, Gooding AR, Podell E, Zhou K, Golden BL, et al. (1996) RNA tertiary structure mediation by adenosine platforms. *Science* 273: 1696–1699.
21. Doherty EA, Batey RT, Masquida B, Doudna JA (2001) A universal mode of helix packing in RNA. *Nat Struct Biol* 8: 339–343.
22. Nissen P, Ippolito JA, Ban N, Moore PB, Steitz TA (2001) RNA tertiary interactions in the large ribosomal subunit: The A-minor motif. *Proc Natl Acad Sci USA* 98: 4899–4903.
23. Ogle JM, Brodersen DE, Clemons WM Jr, Tarry MJ, Carter AP, et al. (2001) Recognition of cognate transfer RNA by the 30S ribosomal subunit. *Science* 292: 897–902.
24. Lentzen G, Klinck R, Matassova N, Aboul-ela F, Murchie AI (2003) Structural basis for contrasting activities of ribosome binding thiazole antibiotics. *Chem Biol* 10: 769–778.
25. Isaksson J, Acharya S, Barman J, Cheruku P, Chattopadhyaya J (2004) Single-stranded adenine-rich DNA and RNA retain structural characteristics of their respective double-stranded conformations and show directional differences in stacking pattern. *Biochemistry* 43: 15996–16010.
26. Lancaster L, Noller HF (2004) Involvement of 16S rRNA nucleotides G1338 and A1339 in discrimination of initiator tRNA. *Mol Cell* 20: 623–632.
27. Vossbrinck CR, Maddox JV, Friedman S, Debrunner-Vossbrinck BA, Woese CR (1987) Ribosomal RNA sequence suggests microsporidia are extremely ancient eukaryotes. *Nature* 326: 411–414.
28. Hirt RP, Logsdon JM Jr., Healy B, Dorey MW, Doolittle WF, et al. (1999) Microsporidia are related to Fungi: evidence from the largest subunit of RNA polymerase II and other proteins. *Proc Natl Acad Sci USA* 96: 580–585.
29. Baldauf SL, Roger AJ, Wenk-Siefert I, Doolittle WF (2000) A kingdom-level phylogeny of eukaryotes based on combined protein data. *Science* 290: 972–977.
30. Woese CR, Kandler O, Wheelis ML (1990) Towards a natural system of organisms: proposal for the domains Archaea, Bacteria, and Eucarya. *Proc Natl Acad Sci USA* 87: 4576–4569.
31. Winker S, Woese CR (1991) A definition of the domains Archaea, Bacteria and Eucarya in terms of small subunit ribosomal RNA characteristics. *Syst Appl Microbiol* 14: 305–310.
32. Alter O (2006) Discovery of principles of nature from mathematical modeling of DNA microarray data. *Proc Natl Acad Sci USA* 103: 16063–16064.
33. Sagara JI, Shimizu S, Kawabata T, Nakamura S, Ikeguchi M, et al. (1998) The use of sequence comparison to detect 'identities' in tRNA genes. *Nucleic Acids Res* 26: 1974–1979.
34. Cadima J, Jolliffe I (2009) On relationships between uncentered and column-centered principal component analysis. *Pak J Statist* 25: 473–503.
35. Tavazoie S, Hughes JD, Campbell MJ, Cho RJ, Church GM (1999) Systematic determination of genetic network architecture. *Nat Genet* 22: 281–285.
36. Alter O, Golub GH (2004) Integrative analysis of genome-scale data by using pseudoinverse projection predicts novel correlation between DNA replication and RNA transcription. *Proc Natl Acad Sci USA* 101: 16577–16582.