RNA Biology 11:3, 176–185; March 2014; © 2014 Landes Bioscience

Molecular phylogenetics before sequences

Oligonucleotide catalogs as *k*-mer spectra

Mark A Ragan*, Guillaume Bernard, and Cheong Xin Chan

Institute for Molecular Bioscience, and ARC Centre of Excellence in Bioinformatics; The University of Queensland; Brisbane, QLD, Australia

Keywords: molecular phylogenetics, 16S ribosomal RNAs, oligomers, *k*-mers

From 1971 to 1985, Carl Woese and colleagues generated oligonucleotide catalogs of 16S/18S rRNAs from more than 400 organisms. Using these incomplete and imperfect data, Carl and his colleagues developed unprecedented insights into the structure, function, and evolution of the large RNA components of the translational apparatus. They recognized a third domain of life, revealed the phylogenetic backbone of bacteria (and its limitations), delineated taxa, and explored the tempo and mode of microbial evolution. For these discoveries to have stood the test of time, oligonucleotide catalogs must carry significant phylogenetic signal; they thus bear re-examination in view of the current interest in alignment-free phylogenetics based on *k*-mers. Here we consider the aims, successes, and limitations of this early phase of molecular phylogenetics. We computationally generate oligonucleotide sets (*e-catalogs*) from 16S/18S rRNA sequences, calculate pairwise distances between them based on D_2 statistics, compute distance trees, and compare their performance against alignment-based and *k*-mer trees. Although the catalogs themselves were superseded by full-length sequences, this stage in the development of computational molecular biology remains instructive for us today.

Introduction

"A basic goal of biology is to account for the evolution of the cell. Emergence of the translation apparatus is the single most important event in this evolution, for capacity to translate is what defines genotype and phenotype."[1]

From our vantage point, informed as we are by petabytes of sequence and structural data from all manner of organisms, it is easy to forget how little was known of the molecular basis of life when Carl Woese began his career in research. Carl was awarded his PhD in 1953, the year Watson and Crick published the double-helical structure of DNA. At about the same time Keller et al.[2] localized protein synthesis to a "microsomal" fraction of the cell, within which RNA-rich particles, later termed *ribosomes*, were soon discovered.[3] In 1959–1961 two large RNA molecules were revealed as components of the ribosome—one sedimenting at 16S, the other at 23S.[4]

*Correspondence to: Mark A Ragan; Email: m.ragan@uq.edu.au
Submitted: 12/10/2013; Accepted: 12/12/2013; Published Online: 01/14/2014
http://dx.doi.org/10.4161/rna.27505

In early publications, Carl described how DNA, RNA, and microsomal fractions behaved during the germination of bacterial spores.[5,6] In late 1960 he initiated research on the genetic code, and over the next few years made fundamental contributions to our understanding of its origin, universality, and specificity. He was among the first to consider translation in an explicitly evolutionary perspective[7-9] and emphasized the role of RNA, for example in refocusing the basis of genetic code specificity away from steric interactions among amino acids: "in an important sense, the codon 'chooses' its amino acid, not the reverse."[10]

Through these early years, the structure of RNAs remained unclear; indeed, not until the early 1960s was it established that RNAs were linear polymers, i.e., can be referred to as having a *sequence*. In the early 1950s, Fred Sanger and collaborators had developed a stepwise experimental strategy to reveal the structure of insulin as a sequence of amino acids. Each chain was enzymatically cleaved into oligopeptides; these were separated and laboriously characterized, and from the fragmentary sequences large portions of the original protein sequence were reassembled.[11] By about 1960 it was becoming clear that protein sequences were non-random and contained regions with different degrees of conservation.

A similar strategy was soon applied to elucidate the structure of some viral RNAs. RNAs were digested with pancreatic ribonuclease and the products separated using chromatography and electrophoresis, yielding mono-, di-, and tri-nucleotides consistent with an unbranched linear sequence of ribonucleotides linked by phosphodiester bonds.[12] Ten of these products, with lengths from one to four nucleotides, could be readily identified based on their electrophoretic mobility alone, while others were identified via a combination of strategies.[13] Differences in the relative abundance of dinucleotides were interpreted as demonstrating differences in "the sequential arrangement of nucleotides" that were imagined to underlie biological differences among viruses.[14]

The Sanger protocol was later modified to introduce an initial digestion with ribonuclease T_1, thereby generating only a single product from G.[15] With the separation technology then available, about 40 oligomers in the length range from one to five nucleotides could be resolved; this was not sufficient to distinguish *Escherichia coli* 16S from 23S rRNA, although in due course methodological improvements provided access to higher oligomers.[16,17] Sequences of tRNA[18] and 5S rRNA[19] yielded to other protocols that generated overlapping fragments; but in 1964, when Carl took up an appointment to the faculty of the University of Illinois, no RNA molecule had been fully sequenced.

The evolutionary approach to conserved structure and function

Ideas of interrelationships among structure, function, and evolution run deep in the history of biology. Karl von Baer, the French transcendental morphologists and others variously glimpsed parts of this nexus, albeit from pre-Darwinian perspectives and with a mechanical interpretation of function. The appearance of protein sequences in the early 1960s brought renewed interest in relationships among ancestry, conserved and variable regions of sequence and structure, and molecular function.[19-27] In a 1969 letter to Francis Crick,[28] Carl referred to this history embedded in molecules as the cell's "internal fossil record."

Carl understood that a comparative approach would likewise reveal which regions of RNAs were conserved, hence, functionally important. Already in 1961 he had compared nucleotide compositions of 16S and 23S rRNA fractions in different bacteria[29] and in the 1969 letter to Crick he wrote of his "important and nearly irreversible decision" to "determine primary structures for a number of genes in a very diverse group of organisms, on the hope that by deducing rather ancient ancestor sequences for these genes, one will eventually be in the position of being able to see features of the cell's evolution. The obvious choice of molecules here lies in the components of the translation apparatus. What more ancient lineages are there?"[28]

Carl directed some effort to 5S rRNA[30] and 23S rRNA[31] but his main focus was on 16S rRNA. Beginning in 1971, Carl and his coworkers at Illinois, and in due course collaborators in Halifax and Munich, generated oligonucleotide catalogs for 16S rRNAs from about 400 organisms.[32,33] Their comparative approach quickly bore fruit, with the observation that the sets of oligonucleotides from *Escherichia coli* and *Bacillus megaterium* 16S rRNAs were much more similar than expected by chance:

"It is important to explain the existence of sequence homology between these two 16S rRNA species. If it reflects the fact that certain portions of their common ancestral primary structure are locked into the present sequences due to stringent constraints imposed by structural and/or functional considerations, then the conservation becomes highly significant. However, were the frequency of occurrence of mutations in rRNA cistrons to be sufficiently low for some reason, then the bulk of the observed conservation could merely reflect the fact that mutations had not occurred in those regions in either organism, and conservation would be of trivial significance."[1]

The second, alternative hypothesis is amenable to experiment, and their comparison with a third 16S rRNA, represented by a partial catalog from *Alcaligenes faecalis*, and with the "unrelated" 14S rRNA of *Rhodopseudomonas spheroides* and the 18S rRNA of yeast, may have constituted the first validation in computational molecular biology. Although not a proof, additional sequences could be brought into the comparison until the argument for homology becomes undeniable.

Pechman and Woese[1] also concluded that "(i)n a molecule as large as the 16S rRNA, all residues are clearly not equivalent in their importance to molecular function." Some residues are neutral and would be replaced quickly on an evolutionary timescale, whereas others are functionally constrained such their replacement would have to be compensated by a "more or less simultaneous" change of other residues. The more deeply such a "replacement unit" is entangled into molecular function, the longer its mutational "half-life," and the more informative it might be on basal features in the tree. 16S/18S rRNA was a "compound, non-linear chronometer"[34] whose broad-range applicability arises not from its size per se, but rather because each of its more-or-less independent structural domains embeds covariance sets that inform on different scales of evolutionary time, much as the hands of a clock separately indicate hours, minutes, and seconds.[35]

For Carl, the "ultimate goal in comparative studies of rRNA sequence is to construct a chronometric model of the molecule that permits its potential as an evolutionary measuring device to be fully exploited." [36] He formalized this deeply structural (i.e., not purely statistical or cladistic) concept as covariance sets of nucleotides. In due course, sets of co-varying positions would be mapped onto folded structure; but in the meantime, the path to covariance sets lay through oligonucleotide catalogs and signature analysis.

Oligonucleotide catalogs

In this context, a catalog is the list of oligomers identified following enzymatic digestion of an RNA or protein. Complete digestion of an RNA with T_1 ribonuclease[15] yields non-overlapping oligonucleotides that end in G. Although at first only short oligonucleotides could be resolved and identified, by the mid-1970s the upper limit on length had been pushed well into the teens, and in one case to 24.[37] Incompletely characterized oligomers, those with modified bases, and termini, were often included in these catalogs; short oligomers (for 16S rRNAs, 5-mers and below) contributed no additional resolving power, and were often not reported.

An RNA dinucleotide catalog was presented by Reddi[14] and catalogs with larger oligonucleotides were published by Rushizky and Knight,[38] Sanger,[15] and others. More than 30 16S/18S rRNAs had been oligo-cataloged by 1975,[39] more than 170 by 1980,[40] and more than 400 by 1985.[36] Most of these data were transferred to punch cards[17] and organized as a database with search, comparison, and tree-inference tools.[41]

Comparing catalogs and computing trees

Sydney Fox and Paul Homeyer[42] compared partial amino acid compositions in seed globulins of six plants,[43] and in 24 protein types mostly from animals (Table I of ref. 44). The idea of combinatorially based diversity can be discerned in their publication, but Fox and Homeyer did not discuss sequences per se. Importantly, however, they interpreted these composition data as showing that "protein synthesis has not, in the main, yet become sufficiently diverse through molecular evolution to yield substantially unrelated proteins." In modern terminology, protein structure as reflected in 1-mers did not seem to be evolving so fast that historical signal would be lost. This had not been shown before, and set the scene for the subsequent development of molecular phylogenetics.

As we mention above in the context of primary-structural determination, peptides from protease digests could be separated in two dimensions by paper chromatography and electrophoresis, and compared by eye for similarities and differences.[22,45,46]

František Šorm and colleagues[47,48] were arguably the first to use patterns and frequencies of di-, tri-, and tetra-peptides not only to explore regularities in protein structure, but also to compare "proteins which have the same function but differ in their origin (different animal species), and proteins of similar function and a common origin."[48] As summarized by Williams et al.,[49] Šorm thought that his work demonstrated that "even where complete sequences are not known, the number of peptides common to two proteins can be used to show similarity of their primary structures."

New techniques were needed to compare sets of oligomers. Two sequences might be compared by eye (e.g., ref. 50), but this is neither scalable nor statistically rigorous. Citing a standard statistical text[51] Carl selected for this purpose the binary association coefficient (S_{AB}): twice the sum of nucleotides in oligonucleotides common to a pair of catalogs, divided by the total number of nucleotides in the two catalogs.[52] Short oligomers were omitted, and no background correction was made (see ref. 53). Carl was nonetheless distrustful about comparing catalogs in this (or any other automated) way: the oligonucleotide data were biased (ribonuclease T_1 does not cleave randomly, and electrophoresis at low pH separates some oligonucleotides more cleanly than it does others), and families of similar, probably homologous, oligonucleotides were mostly ignored. But more fundamentally for Carl, S_{AB} values could not capture molecular structure.

Later, when full sequences had become available, Carl plotted pairwise S_{AB} values between catalogs against percent similarity of aligned 16S rRNA sequences, revealing an imprecise relationship for S_{AB} less than about 0.40, i.e., most of them.[35] Carl criticized his earlier catalog approach as (1) not having resolved branching orders among major bacterial divisions and subdivisions, and (2) failing to resolve branching order of rapidly evolving lineages such as the planctomycetes. Catalogs and pairwise S_{AB} values could not offer the resolving power that was available from the rRNA chronometer as read via sequences; nor should we "consider the second hand when timing the seasons."[35]

Given a matrix of pairwise S_{AB} values, a dendrogram could be computed by average linkage clustering. The first rRNA oligonucleotide trees appeared in 1976[53] and 1977.[37,54] Fox et al.[37] asserted that although this approach is phyletic, "it is clear from the molecular nature of the data" that the topology "would closely resemble, if not be identical to, that of a phylogenetic tree based upon such ancestral catalogs." These trees might be a guide to relatedness and relative antiquity (e.g., ref. 40), but Carl did not delineate taxa solely on the basis of trees.[36]

Signatures

More important than trees was the "internal fossil record" revealed through signatures. Carl defined a signature as a "set of oligonucleotides that is characteristic of (unique to) a group of organisms," but immediately relaxed this to allow oligonucleotides to "occur in half or more of the members of the group, but are either not found in other organisms or occur only sporadically therein."[55] Slightly different formulations were offered later.[35,36] Modulo this relaxation, signatures were synapomorphies (Carl Woese, personal communication to MAR, 30 August 1988).

Carl immersed himself in the details. As related by George Fox, during the heyday of the oligonucleotide work "Carl had established routines that allowed him to be with the fingerprints 8 hours a day, 5 days a week. He went to great lengths to avoid interruptions and non-research related activities."[17] Carl's knowledge of patterns of oligonucleotide occurrence and co-variation, and his ability to map details immediately onto folded structure, convinced one of us (MAR) that he had an exquisitely detailed mental map of 16S rRNA structure and evolution, as Emanuel Margoliash surely had for cytochrome *c*. In any case, to Carl a signature was a deeply structural and chronometric construct,[36] not to be entrusted to generic (or even purpose-built) software.

Carl's group managed and compared signatures, and computed trees, with the aid of mainframe computing. Tom Macke wrote a program "sig" that could map the distribution of oligonucleotides, including degenerate ones, across a set of catalogs.[36,56,57] Similar programs are mentioned by Sobieski et al.[41] In those years, hardware and operating systems were far less standardized than today, and it was not straightforward to exchange programs, much less to offer remote access.

All these factors conspired to make signature analysis à la Woese somewhat opaque to outsiders, including the numerical taxonomy and cladistics communities. Zablen et al.[39] clearly articulate the value of shared derived characters; Fox et al.[52] describe an approach seemingly inspired by parsimony; and Carl mentions parsimony analyses elsewhere in passing, e.g., reference 35. Once 16S rRNA structures became available,[58,59] Carl mapped these signatures onto folded structure. Taxa could at last be recognized by three criteria (page 236 of ref. 35): coherence by S_{AB}, shared sequence signature, and higher-order molecular structure.

Oligonucleotides and *k*-mers

Sequences or regions thereof can be arranged relative to each other to reveal similarities and differences; the term "alignment" was introduced for such operations in 1960,[60] although the concept has deeper roots in genetics, computer science, and other fields. Peptides and proteins were aligned first, then tRNAs in 1966[61] and 5S rRNA in 1971.[50] These early alignments were based on visual inspection, but as the comparison problem began to be described more precisely for analysis using electronic computers, three not unrelated classes of approaches emerged.[62] In today's terminology these are the sliding-window, dot-matrix, and *k*-mer spectrum approaches.

Dot-matrix methods were prefigured by Walter Fitch[63] and others prior to their formal description by Gibbs and MacIntyre.[64] Adrian Gibbs and colleagues considered the dot-matrix to subsume the sliding-window approach,[62] and to be "similar in principle"[64] to a method explained by Saul Needleman to Fitch[65] in 1965 and later introduced as the first algorithm for full-length sequence alignment.[66] Applied to molecular sequences, all these approaches find regions of local identity (or similarity). Like oligonucleotides matched between two catalogs, these local regions are not of predefined length; rather, their frequency spectrum (number at each increment of length) is determined by the degree and pattern of pairwise sequence similarity, and by data quality.

Alternatively, sequence analysis can be approached using a fixed word length. In the BLAST algorithm[67] for example, the query

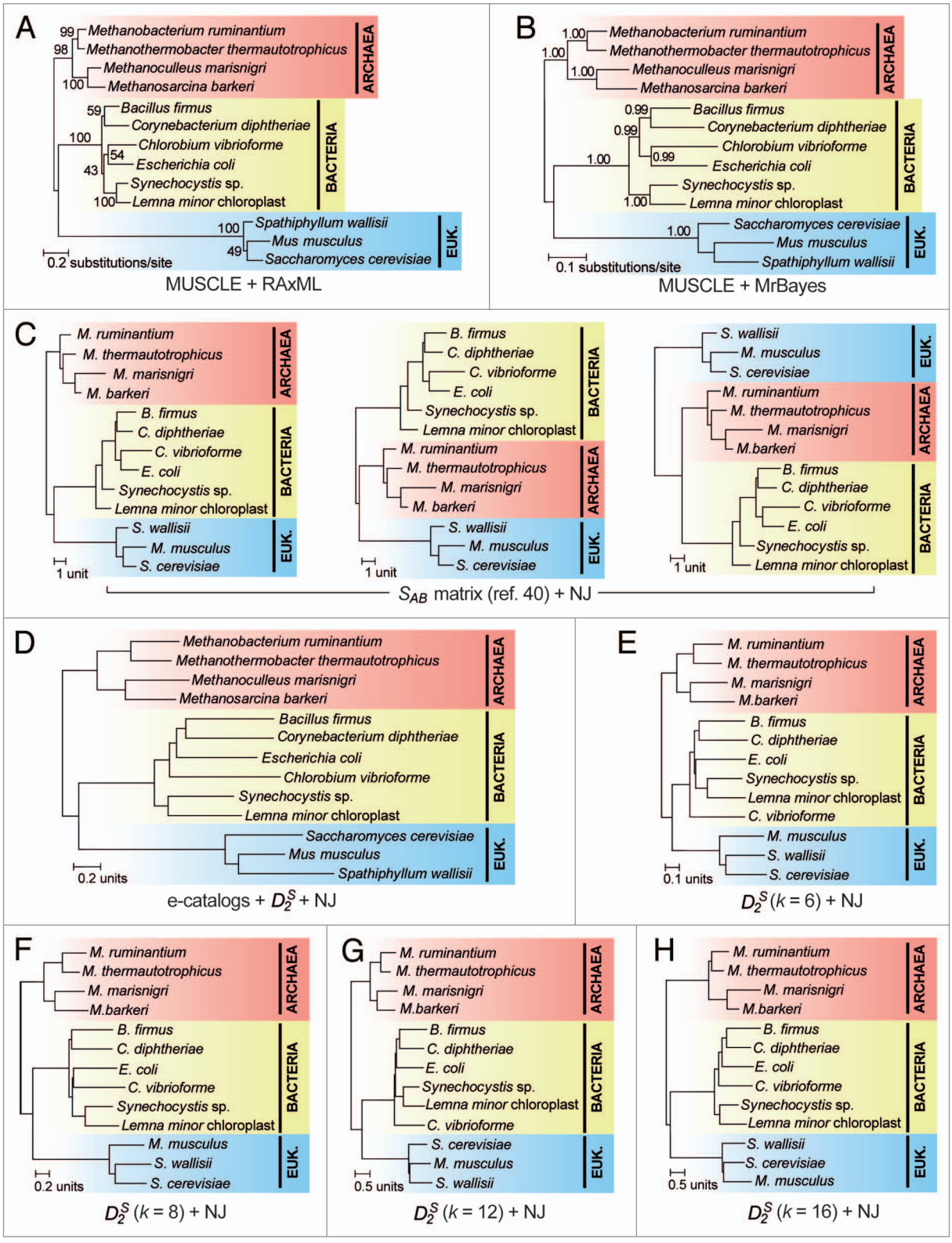

Figure 1. Trees for 16S/18S rRNAs in the three-kingdom data set[74] inferred via multiple sequence alignment of full-length rRNAs using MUSCLE and (**A**) RAxML or (**B**) MRBAYES; (**C**) computed via neighbor-joining from the similarity matrix in reference 74; (**D**) calculated via D_2^S and neighbor-joining from our e-catalogs; and calculated via D_2^S and neighbor-joining from *k*-mer spectra at (**E**) $k = 6$, (**F**) $k = 8$, (**G**) $k = 12$, or (**H**) $k = 16$. To facilitate comparison, all trees were rooted similarly (arbitrarily on archaea), except for (**C**) in which trees were rooted independently on archaea (left), bacteria (middle), and eukaryotes (right).

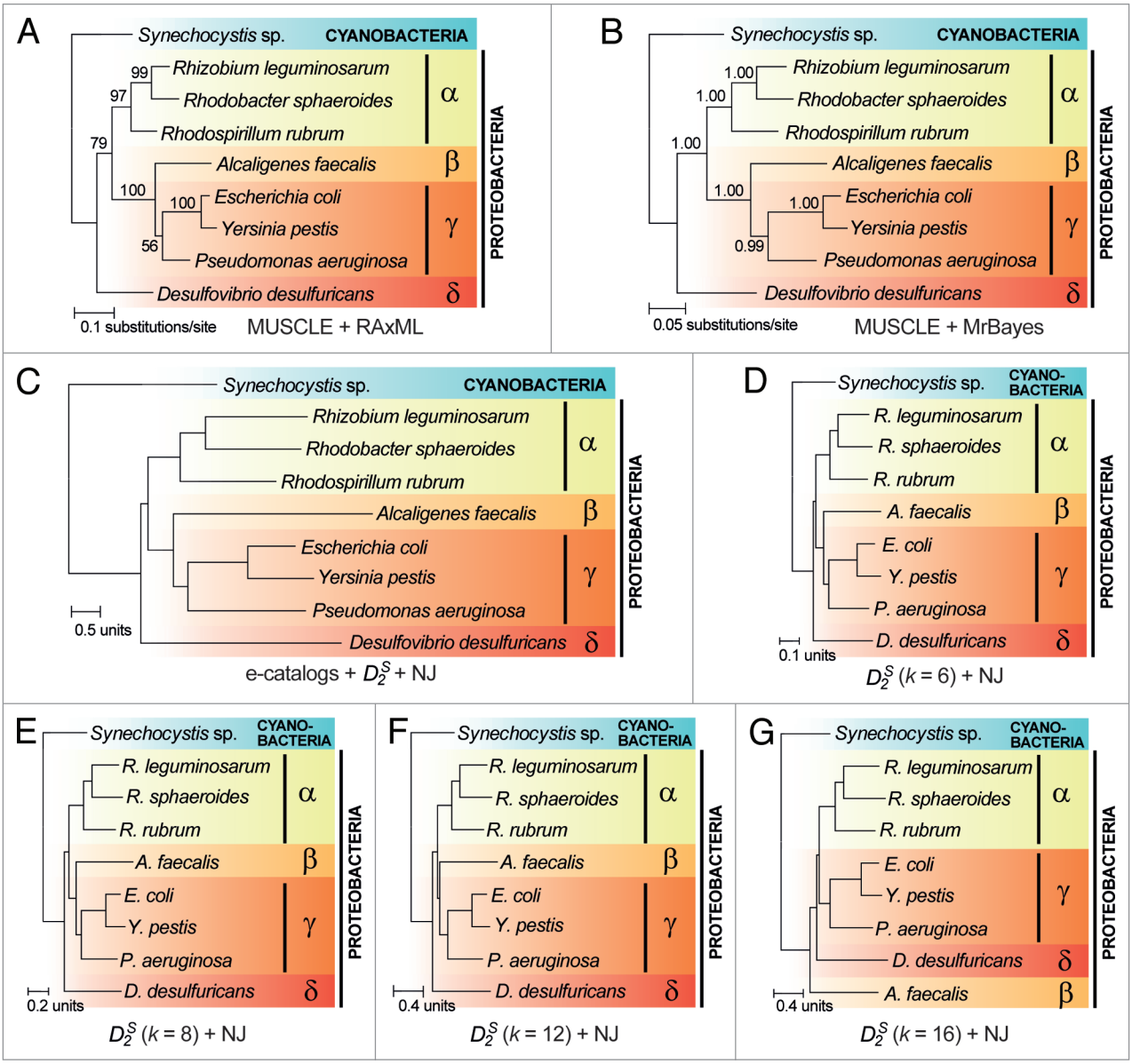

Figure 2. Trees for 16S rRNA in the proteobacterial data set[40] inferred via multiple sequence alignment of full-length rRNAs using MUSCLE and (**A**) RAxML or (**B**) MRBAYES; (**C**) calculated via D_2^S and neighbor-joining from our e-catalogs; and calculated via D_2^S and neighbor-joining from *k*-mer spectra at (**D**) k = 6, (**E**) k = 8, (**F**) k = 12, or (**G**) k = 16. To facilitate comparison, all trees were rooted similarly on the 16S rRNA of the cyanobacterium *Synechocystis*.

sequence is hashed into regions of predetermined length. Similar operations are encountered in diverse areas of mathematics, computer science, and information theory e.g., for sequence compression, indexing, or retrieval. Reflecting these diverse origins and applications, short perfectly matched strings of predetermined length are variously termed *k*-mers, words, or *n*-grams. A common thread is that these strings provide a fast approach to detecting a signal of similarity. *K*-mers find utility in many areas of genomics including genome size estimation, assembly, clustering, and studies on sequence periodicity and lateral genetic transfer.[68,69]

In molecular phylogenetics, *k*-mers have long been used to capture phylogenetic signal. Gibbs et al.[62] used dipeptide frequencies (k = 2) to compute phylogenetic trees based on sequences of cytochromes *c*, hemoglobins, and other proteins. Blaisdell[70] did likewise for a broader set of proteins, with k = 2 and k = 3. More recently, tree inference has used values of k in the range 3–5 for proteins,[71,72] and longer k has been proposed for nucleotides.[73]

Below we look back on Carl's oligonucleotide catalogs as a source of data for phylogenetic inference. With the benefit of complete 16S/18S rRNA sequences, we ask about the accuracy and coverage of T_1 oligonucleotide catalogs, and compare Carl's clustering diagrams with trees based on multiple alignment of complete sequences and inference methods. Because most original T_1 catalogs are no longer accessible in an electronic format, we computationally reconstruct e-catalogs from full-length rRNA sequences of the 13 organisms examined by Woese and Fox,[74] compare them

Table 1. All 16S ribosomal rRNA sequences used in this study, their GenBank accession numbers, and their inclusion in our re-analysis of rRNAs from (A) three kingdoms[74] and (B) proteobacteria (Fig. 4 of ref. 40). For proteobacteria in our analysis B, we identify class (α, β, γ, or δ-proteobacteria).

Source organism	GenBank accession	Analysis
Mus musculus	X00686.1	A
Saccharomyces cerevisiae	V01335.1	A
Spathiphyllum wallisii	AF207023.1	A
Methanobacterium ruminantium	NR_074117.1	A
Methanoculleus marisnigri	NR_074174.1	A
Methanosarcina barkeri	NR_074253.1	A
Methanothermobacter thermautotrophicus	NR_074260.1	A
Bacillus firmus	JQ282815	A
Chlorobium vibrioforme	M62791	A
Corynebacterium diphtheriae	NR_103937.1	A
Lemna minor chloroplast	NC_010109.1*	A
Synechocystis sp.	NR_074311.1	A, B
Rhodobacter sphaeroides (α)	NR_029215.1	B
Rhodospirillum rubrum (α)	NR_074249.1	B
Rhizobium leguminosarum (α)	D14513.1	B
Alcaligenes faecalis (β)	AF155147.1	B
Desulfovibrio desulfuricans (δ)	NR_036778.1	B
Escherichia coli (γ)	NR_102804.1	A, B
Yersinia pestis (γ)	NR_074199.1	B
Pseudomonas aeruginosa (γ)	NR_074828.1	B

*, positions 106162–107648.

with selected empirical catalogs, calculate D_2 statistics,[75-78] and compute a neighbor-joining (NJ) tree.[79] We then do the same for a more complete set of bacteria.[40] Thereafter, we extract *k*-mers (at different values of *k*) from the full-length sequences, and again calculate D_2 statistics, and compute NJ trees. This allows us to explore similarities and differences between oligonucleotide catalogs and modern *k*-mer spectra in phylogenetics.

Results

Trees from aligned full-length sequences

As a reference topology, we inferred a tree based on full-length 16S/18S rRNA sequences of the 13 organisms in Woese and Fox[74] or very close relatives. Multiple sequence alignment (i.e., not leveraging the folded structure of rRNA) followed by fast maximum-likelihood (**Fig. 1A**) or Bayesian inference (**Fig. 1B**) yielded trees differing from each other in two respects: the position of the cyanobacterial/chloroplast subtree within the bacteria, and branching order within eukaryotes. These disagreements correspond to very short internal edges and poor bootstrap support in the likelihood tree (**Fig. 1A**). We followed the same approach to infer trees from aligned full-length 16S rRNA sequences from eight proteobacteria (Fig. 4 of ref. 40), with a *Synechocystis* rRNA as outgroup (**Fig. 2A and B**).

Trees from published S_{AB} matrices

Woese and Fox[74] present a matrix of pairwise association coefficients (S_{AB}) between oligonucleotide catalogs (length ≥ 6), but do not depict the tree these data imply. We converted these S_{AB} values to distances, and computed the NJ tree (**Fig. 1C**). Rooted at any point outside the three clusters of sequences, this tree clearly reveals three main lines of descent. Woese and Fox[74] do not treat branching structure within each kingdom, but the topology we reconstruct within the bacterial lineage is congruent with the cluster diagram published at about the same time by Balch et al.[54] Later, with data from additional bacteria, *Chlorobium* assumed a more-basal position.[40] However, *Synechocystis* and the *Lemna* chloroplast appear paraphyletic, as do *Methanobrevibacter* and *Methanothermobacter* among Archaea.

Computational generation of e-catalogs

We had hoped to generate trees from the original oligonucleotide catalog data underlying Woese and Fox[74] but were able to access only six of the 13 catalogs, and part of a seventh (George Fox has more recently recovered others for us). So instead, starting with full-length 16/18S rRNA sequences from the same or very closely related organisms (**Table 1**), we computationally generated sets of oligonucleotides, mimicking digestion with ribonuclease T_1. Fragments at the 5′ and 3′ termini were included, and oligomers of length < 6 were removed. We refer to these sets as e-catalogs.

Comparison of empirical and e-catalogs

To determine the extent to which our e-catalogs recapitulate Carl's empirical T_1 catalogs (and can thus stand in for the latter in tree inference), we compared e-catalogs and original T_1 catalogs for *Escherichia coli* and *Methanobacterium ruminantium* M-1 (later renamed *Methanobrevibacter ruminantium* M1). For the purpose of this comparison, we ignored base modifications (e.g., treated A* as identical to A) and copy number, and resolved ambiguities in the empirical data in favor of a match. **Table 2** demonstrates that our e-catalogs recapitulate the empirical oligonucleotides very well, although not perfectly. Mismatches likely arise due to weak, diffuse (e.g., Figure 1 of ref. 52), or incompletely resolved spots on paper electrophoresis (e.g., Figure 1 of ref. 16), although sequencing errors, covalent modifications, and/or strain differences cannot be ruled out. It is clear from **Table 3** that the landmark recognition of three kingdoms,[74] and molecular-systematic studies on numerous groups of bacteria and archaea, were based on data representing fewer than 40% of the positions in the 16S rRNA. This is less worrisome than might be thought; many of these oligonucleotides map to one side of a helical region, such that much of the "missing" information is in fact represented as the reverse complement (see Figure 2 of ref. 17).

Trees from e-catalogs

From the e-catalogs we calculated pairwise distances via the D_2^S statistic (Materials and Methods), and computed an NJ tree (**Fig. 1D**). This tree shows the three-kingdom structure. Topology within the archaeal (methanogen) subtree agrees with that in Fox et al.[37,40] and with our *k*-mer trees (**Fig. 1E–H**, for which see below); for simplicity we call this the 2M+2M topology within Archaea. Among bacteria, the *Synechocystis*-chloroplast and

Bacillus-Corynebacterium pairs seen in the alignment-based trees are apparent here too, but *Escherichia* and *Chlorobium* rRNAs no longer form a monophyletic group, instead appearing as adjacent branches. Pairwise S_{AB} values for these bacterial catalogs are in the range 0.19–0.34. Recalling that Carl[35] called attention to the imprecise relationship between S_{AB} and full-length sequence similarity especially at S_{AB} < 0.40, we selected a different bacterial data set (from Fig. 4 of ref. 40) with pairwise S_{AB} values in the range 0.31–0.78, and again calculated D_2^S values and distances. The topology of this tree (**Fig. 2C**) agrees with the alignment-based references (**Fig. 2A and B**) and differs from that implied by Fox et al.[40] only in the relative branching positions of the most-basal branches; that is, again the differences correspond with the smallest S_{AB} values, and short internal edges.[40]

Table 2. Numbers of unique oligonucleotides in empirical 16S rRNA catalogs, and of *k*-mers in e-catalogs

Oligomer length or *k*	*Escherichia coli*			*Methanobacterium ruminantium*		
	empirical	e-catalog	match	empirical	e-catalog	match
6[a]	21[b]	21	21	22	22	20
7	17	16	16	15	16	13
8	10	11	10	14	15	13
9	13	12	12	10	9	8
³10	11	13	10	11	12	10
Total	72	73	69	72	74	64

Escherichia coli empirical catalog from Uchida et al.[16] as corrected by Magrum et al.,[88] and *Methanobacterium ruminantium* M-1 (later renamed *Methanobrevibacter ruminantium* M1) empirical catalog from Fox et al.[37] For the calculation of matching, modifications of bases are ignored and ambiguities are resolved favorably. Includes the 5′ termimus. [b]Uchida et al.[16] report one 6-mer sequence twice, once as unmodified and once as modified; for the purposes of this table we count them once.

K-mer trees from full-length sequences

We extracted k-mers from full-length sequences at selected values of *k* between 6 and 16, calculated pairwise D_2^S values and distances, and used these to compute NJ trees for the three-kingdom[74] and bacterial data sets.[40] The three-kingdom structure, and branching order within Archaea, do not depend on choice of *k* within this range; branching order within bacteria, and within eukaryotes, does (**Fig. 1E–H**). The expected cyanobacterium–chloroplast and *Bacillus–Corynebacterium* pairs are apparent across all *k* = 6, 8, 12, or 16, while the other two bacterial sequences, *Escherichia coli* and *Chlorobium vibrioforme*, show no consistent position. This is perhaps unsurprising, as even today basal branching in the bacterial tree can scarcely be resolved.[80] As above, we therefore examined a less-divergent bacterial data set (from Fig. 4 of ref. 40). At *k* = 6, 8, or 12 (**Fig. 2D–F**) our D_2^S -based NJ trees agree with the alignment-based reference (**Fig. 2A and B**). Even at *k* = 16 (**Fig. 2G**), much of the expected internal structure is preserved.

Discussion

From about 1971 through the mid-1980s, Carl Woese and colleagues generated T_1 oligonucleotide catalogs for more than 400 organisms, mostly bacteria and archaea, with the aim of understanding the nexus among structure, function, and evolution for the RNA components of the translational apparatus. Using tools that in retrospect seem basic—nuclease digestion, radiolabelling, paper electrophoresis, binary association coefficients, clustering algorithms, and simple statistical models of expected similarity—Carl and his colleagues revolutionized the way we view the living world. Recognition of the three kingdoms of life, a phylogenetic backbone of the microbial world, and natural groupings of various size, taxonomic depth, and biological specialization all arose from Carl's interpretation of the molecular fossil record internal to 16S/18S rRNA, via the deeply structural idea of the molecular chronometer that intertwines structure, sequence, and evolution for sufficiently large rRNA molecules.[35,36]

For this fundamental biology to have emerged and withstand the test of time, T_1 oligonucleotide catalogs—incomplete sets of unordered, short, and somewhat noisy sequences—must carry phylogenetic signal. To be sure, their power of resolution wears thin at greater depths (low S_{AB} values), but this is true as well for complete sequences using modern methods.[80]

Empirical oligonucleotide catalogs sample surprisingly little of the full-length sequence (**Table 3**), although rather more of its information content (see above). Carl, who was using these catalogs (along with other approaches) to reconstruct full-length sequences, was well aware of this, but argued that oligonucleotides of length ≤ 4, which accounted for much of the sequence not represented in the catalogs, were in any case uninformative about homology;[81] length 5 was "marginal." The same argument had earlier been made for short oligopeptides in tryptic digests (e.g., ref. 22). By contrast, *k*-mers represent the entire sequence, base-paired, and uninformative regions along with informative ones.

Three kingdoms are apparent in all the trees we compute from e-catalogs or *k*-mers, as is the 2M+2M arrangement within archaea (methanogens). By contrast, within bacteria the branching order is somewhat unstable, particularly for the more-basal branches. Interestingly, the same features are poorly resolved in a modern curated resource, with structure-guided multiple alignment of full-length sequences, and RAxML inference of trees.[80] As for the eukaryotic subtree, the inability of 18S rRNA sequence analysis to resolve the branching order of the green plant, fungal, and animal lineages is well known.[82]

It has not been our aim here to illustrate the full spectrum of so-called alignment-free approaches and methods, nor to compute *k*-mer trees for other genes, proteins, concatenated gene sets, or full genomes. We hope that these analyses will stimulate reflection and deeper analysis where warranted, on how and why catalog-based methods could underpin the revolutionary era in microbiology associated with Carl Woese. Thanks to next-generation and community sequencing technologies, microbiology again faces large, imperfect, and not entirely familiar data; new analytical, comparative, and computational approaches are in play, while non-evolutionary directions beckon. Carl understood

Table 3. Nucleotide coverage of full-length 16S rRNA sequence by oligonucleotides in empirical catalogs, and *k*-mers in e-catalogs, of *Escherichia coli* and *Methanobacterium ruminantium* M-1 (*Methanobrevibacter ruminantium* M1)

16S rRNA source	Number (empirical)	Coverage (%)	Number (e-catalog)	Coverage (%)
E. coli	584/1542	37.9	590/1542	38.3
M. ruminantium	572/1436	39.8	601/1436	41.9

For catalogs, see **Table S1**. Multiple (non-unique) instances are counted (note that Fox et al.[37] do not report multiple occurrences, which in any case were rare for oligonucleotides ≥ 6). Full-length sequences are NR_102904.1 and NR_074117.1, respectively.

that only an evolutionary framework could link genotype with phenotype, and molecular structure with function. With the support of colleagues and great personal determination, Carl built that framework. His life and achievements are, and will long remain, an inspiration.

Materials and Methods

Data

All 16S/18S rRNA sequences used in this study are listed in **Table 1**. We obtained full-length rRNA sequences from organisms and strains that are identical, or as closely related as possible, to those examined by Woese and Fox[74] (**Table S1**). For closer examination of the bacterial lineage, we selected from the organisms in Figure 4 of reference 40 in a way that gives representation of the four major proteobacterial groupings: α (3), β (1), δ (1), and γ (3), with the 16S rRNA of the cyanobacterium *Synechocystis* sp. as outgroup (**Table S2**).

Generation of e-catalogs

To mimic T_1 RNase digestion, we computationally cleaved each full-length sequence (**Table 1**) immediately 3′ of each guanine (G) residue, yielding a set of strings that end in G. Terminal fragments were included in each set, while strings of length < 6 were removed. For ease of handling, we ordered each list first by increasing size, then alphabetically. Our e-catalogs of the Woese and Fox[74] three-kingdom organism set are presented in **Table S1**, and those of the bacterial set[40] in **Table S2**.

Phylogenetic analysis

For each of the two sets of full-length rRNA sequences, we performed multiple sequence alignment using MUSCLE[83] at default settings, then inferred trees using MRBAYES[84] and RAxML.[85] MRBAYES parameter settings were: MCMC ngen = 5 000 000 generations, nchain = 4, burnin = 2 500 000 generations. RAxML (-m GTRGAMMA) was run with 100 bootstraps. For the *k*-mer-based approach, for each sequence set we applied D_2^S statistics[77] independently at k = 6, 8, 12, and 16, yielding a score for each possible pair of sequences within each set. These scores were transformed via logarithmic representation of the geometric mean, to generate a distance. The pairwise distance d_{ab} between sequences a and b is defined as

$$d_{ab} = \left| \ln \frac{D_{ab}}{\sqrt{D_{aa} \times D_{bb}}} \right|$$

where D_{ab} is the pairwise score, and D_{aa} and D_{bb} are the respective self-matching scores. The resulting distance matrix generated for each k was used to reconstruct a phylogenetic tree using neighbor in PHYLIP v3.69 (evolution.genetics.washington.edu/phylip). Similarly, using the in silico oligomer catalog for each full-length sequence as input for D_2^S we calculated pairwise scores and distances for all pairs within a sequence set. The resulting distance matrix was used to compute a tree using neighbor in PHYLIP v3.69.

Disclosure of Potential Conflicts of Interest

No potential conflicts of interest were disclosed.

Acknowledgments

Biographical information on Carl Woese was sourced in part from references 17, 86, and 87. David Lipman drew our attention to reference 60. We thank Linda Bonen, Russell Doolittle, George Fox, Nigel Goldenfeld, Robin Gutell, Norm Pace, David Sankoff, Jan Sapp, and Sahotra Sarkar for information and advice on this manuscript.

Supplemental Materials

Supplemental materials may be found here:
www.landesbioscience.com/journals/rnabiology/article/27505

References

1. Pechman KJ, Woese CR. Characterization of the primary structural homology between the 16s ribosomal RNAs of *Escherichia coli* and *Bacillus megaterium* by oligomer cataloging. J Mol Evol 1972; 1:230-40; PMID:4219625; http://dx.doi.org/10.1007/BF01660242
2. Keller EB, Zamecnik PC, Loftfield RB. The role of microsomes in the incorporation of amino acids into proteins. J Histochem Cytochem 1954; 2:378-86; PMID:13192327; http://dx.doi.org/10.1177/2.5.378
3. Roberts RB. Introduction. In: Roberts RB, ed. Microsomal particles and protein synthesis. New York: Pergamon Press, 1958: vii-viii.
4. Bąkowska-Żywicka K, Tyczewska A. The structure of the ribosome – short history. Biotechnologia 2009; 1:14-23
5. Woese CR, Forro JR. Correlations between ribonucleic acid and deoxyribonucleic acid metabolism during spore germination. J Bacteriol 1960; 80:811-7; PMID:13786172
6. Woese CR, Langridge R, Morowitz HJ. Microsome distribution during germination of bacterial spores. J Bacteriol 1960; 79:777-82; PMID:13845563
7. Woese CR. Universality in the genetic code. Science 1964; 144:1030-1; PMID:14137944; http://dx.doi.org/10.1126/science.144.3621.1030
8. Woese CR. Order in the genetic code. Proc Natl Acad Sci U S A 1965; 54:71-5; PMID:5216368; http://dx.doi.org/10.1073/pnas.54.1.71
9. Woese CR. On the evolution of the genetic code. Proc Natl Acad Sci U S A 1965; 54:1546-52; PMID:5218910; http://dx.doi.org/10.1073/pnas.54.6.1546
10. Woese CR, Dugre DH, Saxinger WC, Dugre SA. The molecular basis for the genetic code. Proc Natl Acad Sci U S A 1966; 55:966-74; PMID:5219702; http://dx.doi.org/10.1073/pnas.55.4.966
11. Stretton AOW. The first sequence. Fred Sanger and insulin. Genetics 2002; 162:527-32; PMID:12399368

12. Markham R, Smith JD. The structure of ribonucleic acids. II. The smaller products of ribonuclease digestion. Biochem J 1952; 52:558-65; PMID:13018278
13. Sanger F. Sequences, sequences, and sequences. Annu Rev Biochem 1988; 57:1-28; PMID:2460023; http://dx.doi.org/10.1146/annurev.bi.57.070188.000245
14. Reddi KK. Structural differences in the nucleic acids of some tobacco mosaic virus strains. II. Di- and trinucleotides in ribonuclease digests. Biochim Biophys Acta 1959; 32:386-92; PMID:14436793; http://dx.doi.org/10.1016/0006-3002(59)90611-0
15. Sanger F, Brownlee GG, Barrell BG. A two-dimensional fractionation procedure for radioactive nucleotides. J Mol Biol 1965; 13:373-98; PMID:5325727; http://dx.doi.org/10.1016/S0022-2836(65)80104-8
16. Uchida T, Bonen L, Schaup HW, Lewis BJ, Zablen L, Woese C. The use of ribonuclease U_2 in RNA sequence determination. Some corrections in the catalog of oligomers produced by ribonuclease T_1 digestion of *Escherichia coli* 16S ribosomal RNA. J Mol Evol 1974; 3:63-77; PMID:4597768; http://dx.doi.org/10.1007/BF01795977
17. Sapp J, Fox GE. The singular quest for a universal tree of life. Microbiol Mol Biol Rev 2013; 77:541-50; PMID:24296570; http://dx.doi.org/10.1128/MMBR.00038-13
18. Holley RW, Apgar J, Everett GA, Madison JT, Marquisee M, Merrill SH, Penswick JR, Zamir A. Structure of a ribonucleic acid. Science 1965; 147:1462-5; PMID:14263761; http://dx.doi.org/10.1126/science.147.3664.1462
19. Brownlee GG, Sanger F, Barrell BG. Nucleotide sequence of 5*S*-ribosomal RNA from *Escherichia coli*. Nature 1967; 215:735-6; PMID:4862513; http://dx.doi.org/10.1038/215735a0
20. Ingram VM. Gene evolution and the haemoglobins. Nature 1961; 189:704-8; PMID:13717731; http://dx.doi.org/10.1038/189704a0
21. Hill RL, Buettner-Janusch J, Buettner-Janusch V. Evolution of haemoglobin in primates. Proc Natl Acad Sci U S A 1963; 50:885-93; PMID:14082353; http://dx.doi.org/10.1073/pnas.50.5.885
22. Zuckerkandl E, Jones RT, Pauling L. A comparison of animal hemoglobins by truptic peptide pattern analysis. Proc Natl Acad Sci U S A 1960; 46:1349-60; PMID:16590757; http://dx.doi.org/10.1073/pnas.46.10.1349
23. Zuckerkandl E, Pauling L. Evolutionary divergence and convergence in proteins. In: Bryson V, Vogel HJ, eds. Evolving Genes and Proteins. New York and London: Academic Press, 1965: 97-166.
24. Fitch WM, Margoliash E. Construction of phylogenetic trees. Science 1967; 155:279-84; PMID:5334057; http://dx.doi.org/10.1126/science.155.3760.279
25. Fitch WM, Margoliash E. The usefulness of amino acid and nucleotide sequences in evolutionary studies. Evol Biol 1970; 4:67-109
26. Dickerson RE. The structures of cytochrome c and the rates of molecular evolution. J Mol Evol 1971; 1:26-45; PMID:4377446; http://dx.doi.org/10.1007/BF01659392
27. Zuckerkandl E. On the molecular evolutionary clock. J Mol Evol 1987; 26:34-46; PMID:3125336; http://dx.doi.org/10.1007/BF02111280
28. Woese CR. (Department of Microbiology, University of Illinois). Letter to: Francis H.C. Crick (Medical Research Council, Cambridge). 1969 Jun 24. 2 leaves. Located at: Wellcome Library, London
29. Woese CR. Composition of various ribonucleic acid fractions from micro-organisms of different deoxyribonucleic acid composition. Nature 1961; 189:920-1; PMID:13786175; http://dx.doi.org/10.1038/189920a0
30. Sogin SJ, Sogin ML, Woese CR. Phylogenetic measurement in procaryotes by primary structural characterization. J Mol Evol 1972; 1:173-84; PMID:24173440; http://dx.doi.org/10.1007/BF01659163
31. Woese CR. Primary structure homology within the 23*S* ribosomal RNA. Nature 1968; 220:923; PMID:4881007; http://dx.doi.org/10.1038/220923a0
32. Woese CR, Stackebrandt E, Ludwig W. What are mycoplasmas: the relationship of tempo and mode in bacterial evolution. J Mol Evol 1985; 21:305-16; PMID:6085735; http://dx.doi.org/10.1007/BF02115648
33. McGill TJ, Jurka J, Sobieski JM, Pickett MH, Woese CR, Fox GE. Characteristic archaebacterial 16S rRNA oligonucleotides. Syst Appl Microbiol 1986; 7:194-7; PMID:11542064; http://dx.doi.org/10.1016/S0723-2020(86)80005-4
34. Woese CR. (Department of Microbiology, University of Illinois). Lecture at Nobel Symposium 70 (Karskoga, Sweden) from notes of Mark A. Ragan (then Atlantic Research Laboratory, National Research Council Canada). 1988 Aug 29.
35. Woese CR. Bacterial evolution. Microbiol Rev 1987; 51:221-71; PMID:2439888
36. Woese CR, Stackebrandt E, Macke TJ, Fox GE. A phylogenetic definition of the major eubacterial taxa. Syst Appl Microbiol 1985; 6:143-51; PMID:11542017; http://dx.doi.org/10.1016/S0723-2020(85)80047-3
37. Fox GE, Magrum LJ, Balch WE, Wolfe RS, Woese CR. Classification of methanogenic bacteria by 16S ribosomal RNA characterization. Proc Natl Acad Sci U S A 1977; 74:4537-41; PMID:16592452; http://dx.doi.org/10.1073/pnas.74.10.4537
38. Rushizky GW, Knight CA. An oligonucleotide mapping procedure and its use in the study of tobacco mosaic virus nucleic acid. Virology 1960; 11:236-49; PMID:14440244; http://dx.doi.org/10.1016/0042-6822(60)90064-7
39. Zablen LB, Kissil MS, Woese CR, Buetow DE. Phylogenetic origin of the chloroplast and prokaryotic nature of its ribosomal RNA. Proc Natl Acad Sci U S A 1975; 72:2418-22; PMID:806081; http://dx.doi.org/10.1073/pnas.72.6.2418
40. Fox GE, Stackebrandt E, Hespell RB, Gibson J, Maniloff J, Dyer TA, Wolfe RS, Balch WE, Tanner RS, Magrum LJ, et al. The phylogeny of prokaryotes. Science 1980; 209:457-63; PMID:6771870; http://dx.doi.org/10.1126/science.6771870
41. Sobieski JM, Chen KN, Filiatreau JC, Pickett MH, Fox GE. 16S rRNA oligonucleotide catalog data base. Nucleic Acids Res 1984; 12:141-8; PMID:6694898; http://dx.doi.org/10.1093/nar/12.1Part1.141
42. Fox SW, Homeyer PG. A statistical evaluation of the kinship of protein molecules. Am Nat 1955; 89:163-8; http://dx.doi.org/10.1086/281876
43. Smith EL, Greene RD. Further studies on the amino acid composition of seed globulins. J Biol Chem 1947; 167:833-42; PMID:20287917
44. Haurowitz F. Chemistry and biology of proteins. New York: Academic Press, 1950.
45. Ingram VM. A specific chemical difference between the globins of normal human and sickle-cell anaemia haemoglobin. Nature 1956; 178:792-4; PMID:13369537; http://dx.doi.org/10.1038/178792a0
46. Ingram VM. Abnormal human haemoglobins. I. The comparison of normal human and sickle-cell haemoglobins by fingerprinting. Biochim Biophys Acta 1958; 28:539-45; PMID:13560404; http://dx.doi.org/10.1016/0006-3002(58)90516-X
47. Šorm F. On proteins. XXV. The chemical structure of proteins. Introduction. [in Russian]. Collect Czech Chem Commun 1954; 19:1003-5
48. Šorm F, Keil B. Holeyšovský, Knesslová V, Kostka V, Mäsiar P, Meloun B, Mikeš O, Tomášek V, Vaněček J. On proteins. XXXIX. Structural resemblance in certain proteins. Collect Czech Chem Commun 1957; 22:1310-29
49. Williams J, Clegg JB, Mutch MO. Coincidence and protein structure. J Mol Biol 1961; 3:532-40; PMID:14007178; http://dx.doi.org/10.1016/S0022-2836(61)80019-3
50. DuBuy B, Weissman SM. Nucleotide sequence of *Pseudomonas fluorescens* 5 S ribonucleic acid. J Biol Chem 1971; 246:747-61; PMID:5542687
51. Anderberg MR. Cluster analysis for applications. New York: Academic Press, 1973.
52. Fox GE, Pechman KR, Woese CR. Comparative cataloging of 16*S* ribosomal ribonucleic acid: molecular approach to procaryotic systematics. Int J Syst Bacteriol 1977; 27:44-57; http://dx.doi.org/10.1099/00207713-27-1-44
53. Bonen L, Doolittle WF. Partial sequences of 16S rRNA and the phylogeny of blue-green algae and chloroplasts. Nature 1976; 261:669-73; PMID:819841; http://dx.doi.org/10.1038/261669a0
54. Balch WE, Magrum LJ, Fox GE, Wolfe RS, Woese CR. An ancient divergence among the bacteria. J Mol Evol 1977; 9:305-11; PMID:408502; http://dx.doi.org/10.1007/BF01796092
55. Woese CR, Maniloff J, Zablen LB. Phylogenetic analysis of the mycoplasmas. Proc Natl Acad Sci U S A 1980; 77:494-8; PMID:6928642; http://dx.doi.org/10.1073/pnas.77.1.494
56. Woese CR, Stackebrandt E, Weisburg WG, Paster BJ, Madigan MT, Fowler VJ, Hahn CM, Blanz P, Gupta R, Nealson KH, et al. The phylogeny of purple bacteria: the alpha subdivision. Syst Appl Microbiol 1984; 5:315-26; PMID:11541974; http://dx.doi.org/10.1016/S0723-2020(84)80034-X
57. Woese CR, Weisburg WG, Paster BJ, Hahn CM, Tanner RS, Keieg NR, Koops H-P, Harms H, Stackebrandt E. The phylogeny of purple bacteria: the beta subdivision. Syst Appl Microbiol 1984; 5:327-36; http://dx.doi.org/10.1016/S0723-2020(84)80035-1
58. Ehresmann C, Stiegler P, Mackie GA, Zimmermann RA, Ebel JP, Fellner P. Primary sequence of the 16S ribosomal RNA of Escherichia coli. Nucleic Acids Res 1975; 2:265-78; PMID:1091918; http://dx.doi.org/10.1093/nar/2.2.265
59. Brosius J, Palmer ML, Kennedy PJ, Noller HF. Complete nucleotide sequence of a 16S ribosomal RNA gene from *Escherichia coli*. Proc Natl Acad Sci U S A 1978; 75:4801-5; PMID:368799; http://dx.doi.org/10.1073/pnas.75.10.4801
60. Lanni F. Analysis of sequence patterns in ribonuclease, I. Sequence vectors and vector maps. Proc Natl Acad Sci U S A 1960; 46:1563-76; PMID:16590782; http://dx.doi.org/10.1073/pnas.46.12.1563
61. Zachau HG, Dütting D, Feldmann H. The structures of two serine transfer ribonucleic acids. Hoppe Seylers Z Physiol Chem 1966; 347:212-35; PMID:5991670; http://dx.doi.org/10.1515/bchm2.1966.347.1.212
62. Gibbs AJ, Dale MB, Kinns MR, MacKenzie HG. The transition matrix method for comparing sequences; its use in describing and classifying proteins by their amino acid sequences. Syst Zool 1971; 20:417-25; http://dx.doi.org/10.2307/2412117
63. Fitch WM. An improved method of testing for evolutionary homology. J Mol Biol 1966; 16:9-16; PMID:5917736; http://dx.doi.org/10.1016/S0022-2836(66)80258-9
64. Gibbs AJ, McIntyre GA. The diagram, a method for comparing sequences. Its use with amino acid and nucleotide sequences. Eur J Biochem 1970; 16:1-11; PMID:5456129; http://dx.doi.org/10.1111/j.1432-1033.1970.tb01046.x
65. Fitch WM. Locating gaps in amino acid sequences to optimize the homology between two proteins. Biochem Genet 1969; 3:99-108; PMID:5364927; http://dx.doi.org/10.1007/BF00520346

66. Needleman SB, Wunsch CD. A general method applicable to the search for similarities in the amino acid sequence of two proteins. J Mol Biol 1970; 48:443-53; PMID:5420325; http://dx.doi.org/10.1016/0022-2836(70)90057-4
67. Altschul SF, Gish W, Miller W, Myers EW, Lipman DJ. Basic local alignment search tool. J Mol Biol 1990; 215:403-10; PMID:2231712
68. Vinga S, Almeida J. Alignment-free sequence comparison-a review. Bioinformatics 2003; 19:513-23; PMID:12611807; http://dx.doi.org/10.1093/bioinformatics/btg005
69. Chan CX, Ragan MA. Next-generation phylogenomics. Biol Direct 2013; 8:3; PMID:23339707; http://dx.doi.org/10.1186/1745-6150-8-3
70. Blaisdell BE. A measure of the similarity of sets of sequences not requiring sequence alignment. Proc Natl Acad Sci U S A 1986; 83:5155-9; PMID:3460087; http://dx.doi.org/10.1073/pnas.83.14.5155
71. Höhl M, Ragan MA. Is multiple-sequence alignment required for accurate inference of phylogeny? Syst Biol 2007; 56:206-21; PMID:17454975; http://dx.doi.org/10.1080/10635150701294741
72. Ragan MA, Chan CX. Biological intuition in alignment-free methods: response to Posada. J Mol Evol 2013; 77:1-2; PMID:23877343; http://dx.doi.org/10.1007/s00239-013-9573-0
73. Forêt S, Kantorovitz MR, Burden CJ. Asymptotic behaviour and optimal word size for exact and approximate word matches between random sequences. BMC Bioinformatics 2006; 7(Suppl 5):S21; PMID:17254306; http://dx.doi.org/10.1186/1471-2105-7-S5-S21
74. Woese CR, Fox GE. Phylogenetic structure of the prokaryotic domain: the primary kingdoms. Proc Natl Acad Sci U S A 1977; 74:5088-90; PMID:270744; http://dx.doi.org/10.1073/pnas.74.11.5088
75. Chor B, Horn D, Goldman N, Levy Y, Massingham T. Genomic DNA *k*-mer spectra: models and modalities. Genome Biol 2009; 10:R108; PMID:19814784; http://dx.doi.org/10.1186/gb-2009-10-10-r108
76. Forêt S, Wilson SR, Burden CJ. Empirical distribution of *k*-word matches in biological sequences. Pattern Recognit 2009; 42:539-48; http://dx.doi.org/10.1016/j.patcog.2008.06.026
77. Reinert G, Chew D, Sun F, Waterman MS. Alignment-free sequence comparison (I): statistics and power. J Comput Biol 2009; 16:1615-34; PMID:20001252; http://dx.doi.org/10.1089/cmb.2009.0198
78. Wan L, Reinert G, Sun F, Waterman MS. Alignment-free sequence comparison (II): theoretical power of comparison statistics. J Comput Biol 2010; 17:1467-90; PMID:20973742; http://dx.doi.org/10.1089/cmb.2010.0056
79. Saitou N, Nei M. The neighbor-joining method: a new method for reconstructing phylogenetic trees. Mol Biol Evol 1987; 4:406-25; PMID:3447015
80. Yarza P, Richter M, Peplies J, Euzeby J, Amann R, Schleifer K-H, Ludwig W, Glöckner FO, Rosselló-Móra R. The All-Species Living Tree project: a 16S rRNA-based phylogenetic tree of all sequenced type strains. Syst Appl Microbiol 2008; 31:241-50; PMID:18692976; http://dx.doi.org/10.1016/j.syapm.2008.07.001
81. Woese CR, Fox GE, Zablen L, Uchida T, Bonen L, Pechman K, Lewis BJ, Stahl D. Conservation of primary structure in 16S ribosomal RNA. Nature 1975; 254:83-6; PMID:1089909; http://dx.doi.org/10.1038/254083a0
82. Baldauf SL, Roger AJ, Wenk-Siefert I, Doolittle WF. A kingdom-level phylogeny of eukaryotes based on combined protein data. Science 2000; 290:972-7; PMID:11062127; http://dx.doi.org/10.1126/science.290.5493.972
83. Edgar RC. MUSCLE: multiple sequence alignment with high accuracy and high throughput. Nucleic Acids Res 2004; 32:1792-7; PMID:15034147; http://dx.doi.org/10.1093/nar/gkh340
84. Ronquist F, Teslenko M, van der Mark P, Ayres DL, Darling A, Höhna S, Larget B, Liu L, Suchard MA, Huelsenbeck JP. MrBayes 3.2: efficient Bayesian phylogenetic inference and model choice across a large model space. Syst Biol 2012; 61:539-42; PMID:22357727; http://dx.doi.org/10.1093/sysbio/sys029
85. Stamatakis A. RAxML-VI-HPC: maximum likelihood-based phylogenetic analyses with thousands of taxa and mixed models. Bioinformatics 2006; 22:2688-90; PMID:16928733; http://dx.doi.org/10.1093/bioinformatics/btl446
86. Pace NR, Sapp J, Goldenfeld N. Phylogeny and beyond: Scientific, historical, and conceptual significance of the first tree of life. Proc Natl Acad Sci U S A 2012; 109:1011-8; PMID:22308526; http://dx.doi.org/10.1073/pnas.1109716109
87. Nair P. Woese and Fox: Life, rearranged. Proc Natl Acad Sci U S A 2012; 109:1019-21; PMID:22308527; http://dx.doi.org/10.1073/pnas.1120749109
88. Magrum L, Zablen L, Stahl D, Woese C. Corrections in the catalogue of oligonucleotides produced by digestion of Escherichia coli 16S rRNA with T1 RNase. Nature 1975; 257:423-6; PMID:809718; http://dx.doi.org/10.1038/257423a0