Phylogenetics without multiple sequence alignment

Mark Ragan

Institute for Molecular Bioscience

and

School of Information Technology & Electrical Engineering

The University of Queensland, Brisbane, Australia

IPAM Workshop on Multiple Sequence Alignment
UCLA, 13 January 2015

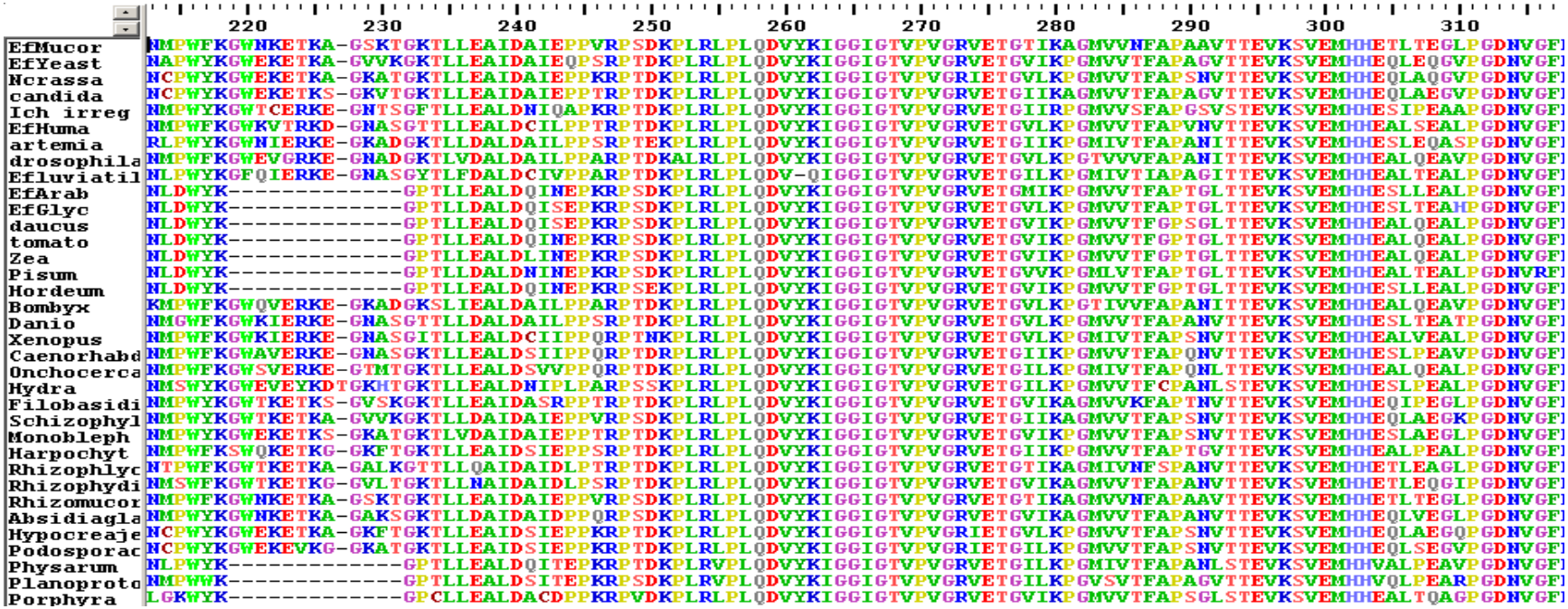
The University Of Queensland AUSTRALIA | IMB

Given a set of molecular sequences, MSA* gives us access to…

Patterns within columns

Local adjacency relationships within rows (across columns)

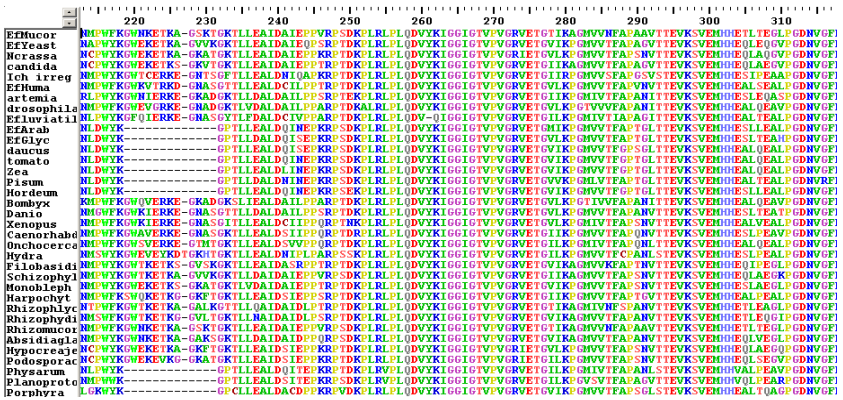
Global architecture

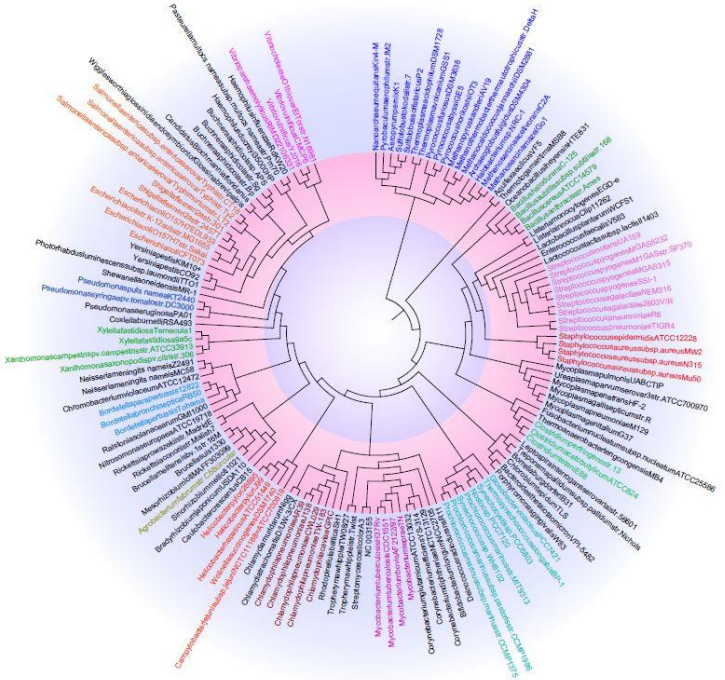


* *MSA = multiple sequence alignment*

For application in phylogenetic inference, we interpret the MSA as a

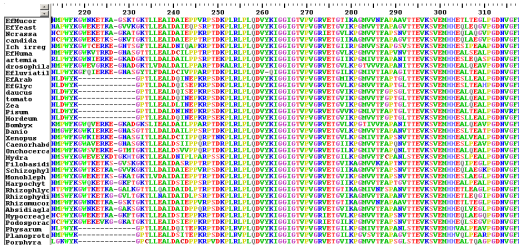
position-by-position (*i.e.* column-by-column) **hypothesis of homology**

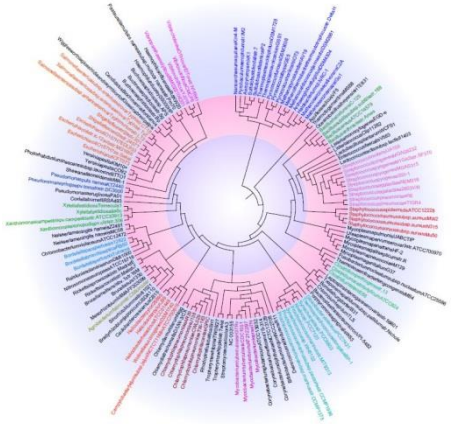




MSA by Mark Ragan; tree by Cheong Xin Chan

Homology signal





Phylogenetic inference is based on **homology signal**

Homology signal inheres in the sequences, not in their MSA

MSA can make it easier to see*, but doesn't create it

** and easier for existing computer programs to work with*

Homology signal (continued)

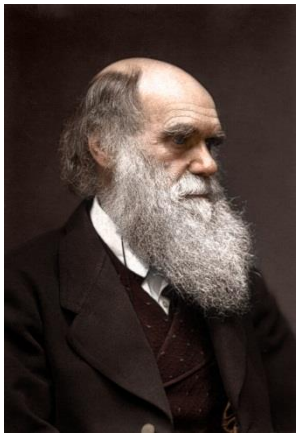
We shouldn't assume that MSA captures it all, or uses it optimally

MSA gives us access to

- Patterns within columns
- Local adjacency relationships
- Global architecture

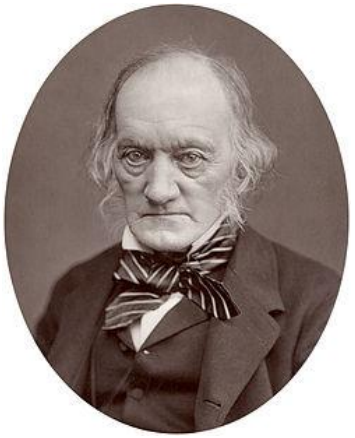
} Let's consider these to be *components* of the homology signal

Here we'll focus on the first two of these components



Pattern and adjacency

The column component needs to capture "sameness" of a character across sequences



For application in phylogenetics, "sameness" has to mean *homology* (or *orthology*). It's difficult to build a statistical case that a particular single character in one sequence is homologous with a particular one in a second sequence. MSA uses adjacency (and sometimes global) information to build this support. Alternatively we might compare sets of adjacent characters (strings), which are less likely to occur by chance.

The adjacency component doesn't just provide statistical support for the column component

Because conserved function arises in part from chemical properties of adjacent residues (*e.g.* in making that part of the molecule an active site or α-helix), we expect homology signal to have an adjacency component in its own right.

MSA: potential (and real) problems

Genomes are dynamic, data can be dirty, and MSA is hard

Within some but not all members of a gene set…

- Homologous regions may be inserted / deleted
- Homologous regions may be rearranged / duplicated
- Regions may have different evolutionary histories (LGT)
- Transcriptional variation → similar issues for protein sets

Sequences may be mis-assembled (or not assembled in the first place) and/or truncated

MSA is computationally difficult and/or heuristic

Can we extract enough/most/all of the homology signal without MSA ?

REVIEW

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

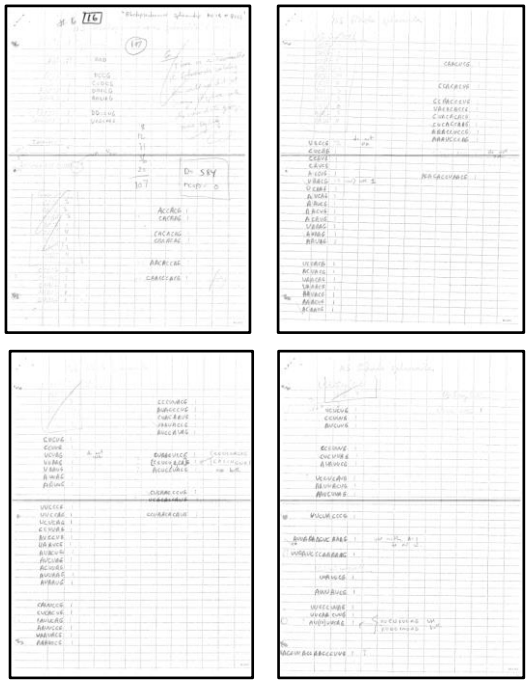
Molecular phylogenetics before sequences

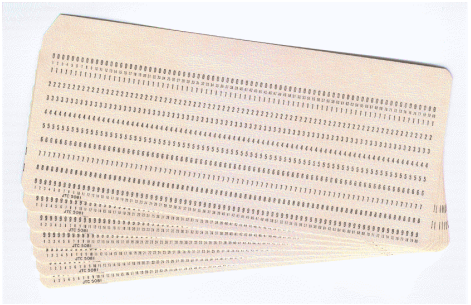
Oligonucleotide catalogs as *k*-mer spectra

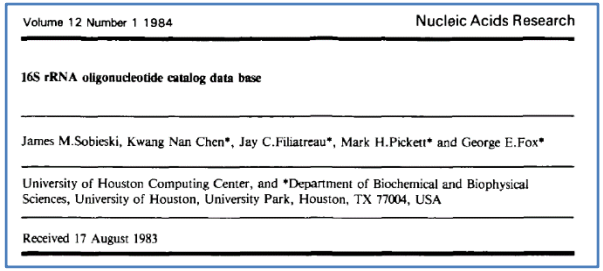
Mark A Ragan*, Guillaume Bernard, and Cheong Xin Chan

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

Carl Woese – photo by Ken Luehrsen







Volume 12 Number 1 1984 Nucleic Acids Research

16S rRNA oligonucleotide catalog data base

James M.Sobieski, Kwang Nan Chen*, Jay C.Filiatreau*, Mark H.Pickett* and George E.Fox*

University of Houston Computing Center, and *Department of Biochemical and Biophysical Sciences, University of Houston, University Park, Houston, TX 77004, USA

Received 17 August 1983

Courtesy of George Fox 2013

Oligonucleotide catalogs

Table 1. Oligonucleotide catalogs for 16S rRNA of 10 methanogens

Oligonu-cleotide sequence	Present in organism number
5-mers	
CCCCG	1-10;1,5,8
CCCAG	6
CCACG	10
ACCCG	10
CCAAG	9
CACAG	9
CAACG	1-10;8-9
ACACG	7-9
ACCAG	7
AACCG	1-10;10
ACAAG	1-6;1,5
AAACG	7-9
AAAAG	1,6,9-10
CUCCG	4,7
CCCUG	9
UCCAG	6-8,10
CUCAG	1-10
CCAUG	1-10
UCACG	1-2,4-5
UACCG	1-6,8
ACCUG	4-5;5
ACUCG	6
AUCCG	9
UAACG	4-9
CAAUG	1-6;4
ACUAG	2-3,8-9
ACAUG	10
AUACG	7
AAUCG	10
UAAAG	2
AUAAG	3-10;3,6-9;7
AAAUG	4
UUCCG	1-6,8;4
CUUCG	5-6,8
UCCUG	1-6;4
CCUUG	1
CUCUG	6,8
UCUAG	7
UUCAG	5,7-9;9
CUAUG	5
UACUG	7-10;8-10
UAUCG	7-8
ACUUG	1-6,10
AUCUG	3-5,7-8
AUUCG	2-3,10
UUAAG	1-10;1-2,4,6,8,10
UAAUG	1-2,5,10;2
AUAUG	3-4,9
AAUUG	1-10;1-2,4-6,9
AUUAG	1-10;1-7,9;7
UUUCG	4,7,9
UUCUG	3
UCUUG	8-9
CUUUG	1-3,5,10
UUUAG	2,7
UUAUG	4,9,9
UUUUG	2,9
6-mers	
CCCCAG	4,6
CCCAAG	6,10
CAACCG	8-9
ACCACG	7-9
ACACCG	6-10
AAACCG	8-10
CCCUCG	5,8,10
CCUCAG	5
CUCCAG	5
UCCCAG	2,7-10
CCACUG	4-5,9
ACCUCG	9
CCUAAG	1-3,5,10
CUCAAG	4-6
CCAUAG	4
CAUACG	1
ACACUG	4-5,7-9
AACCUG	1-6,10;1
AAUCCG	7,9-10
CUAAAG	7-9
UAAACG	1-6,8-10
ACUAAG	9
ACAAUG	1-10
AUAACG	10
AAUACG	1-6,10
AACAUG	10;10
AAACUG	1-10;8-9
AAAUCG	1-3,7
AAUAAG	1-2,4-6
CCCUUG	6,10
CCUCUG	7-9
UCCCUG	1-10
CCUUAG	4,7-8
CUCUAG	1-3
CUUCAG	9
UCCUAG	1-2
UUCCAG	1-6
CCUAUG	3
CUACUG	1-3,6
UCACUG	3,7-9
CUAUCG	7-10
UCAUCG	7,9
CAUCUG	7
ACUCUG	7-8
ACCUUG	4-6
AUCCUG	1-10
UCUAAG	7-8
UUACAG	8
UAUCAG	9
UAUACG	7
UAAUCG	1-10
AUACUG	3,7-8,10
ACAUUG	1
AACUUG	3
AAUCUG	5-9
UAAAUG	4
AUUAAG	1-8
AAUAUG	9
CCUUUG	1-2,5
CUUUCG	10
UCUCUG	1-2,4-6
UUCCUG	5
UCUUAG	5
CUAUUG	1-4,6
UUACUG	10
UAUUCG	3
AUUCUG	2,8-10
ACUUUG	2
UAUAUG	8
CUUUUG	1-5;1
UCUUUG	1,4
UUUUUG	7
7-mers	
ACCCACG	1-9
ACCACCG	7
AACCCCG	7
CCAACAG	7-8
CAACACG	1-2,5-6
CAAACCG	8-9
CCCUACG	1-10
CCCACUG	10
UCCACCG	4-6
CCACCUG	10
CCCUAAG	7-8
UCACACG	3
CUACACG	4,7-10
UAACCCG	5-6
AUACCCG	1-10
AACCUCG	8
CCUAAAG	1-6
UAACACG	1-10
AUAACCG	7
AAUCCAG	8-10
AACAUCG	10
AAAUCCG	7-9
UAAAAAG	1,3-6
CCCUUAG	1,3-6
CAUCCUG	7-10
UACUCCG	7
AUCUCCG	8
ACCUUCG	9
UCCUAAG	7
UUACCAG	1-2,4-6,10
CUAACUG	3-4
UAACUCG	1-4,7-8,10
AUUCCAG	7
AUCAUCG	6
AAUCUCG	3
AACCUUG	6
UCUAAAG	10
CUUAAAG	7-9
CAAUAUG	10
AUACUAG	1
AAUCUAG	1-2,4,7-8
AAAUCUG	10
UAAAAUG	10
CUCCUUG	1-3,5-10
UCCCUUG	9
UUCUCCG	7
CUCUUAG	2
UACUUCG	8
UACUCUG	10
UCAUAUG	10
UAAUCUG	4
AAUUUAG	3
UUCUUCG	10
UCUCUUG	7-8
CUUUAUG	10
UUUAUCG	1
UAUUUCG	1
AUUAUUG	10
UUCUUUG	4-6
UAUUUUG	3
UUUUUUG	1-3
8-mers	
CCACAACG	1-3,5-6,9-10
ACCCCAAG	1,5
AAACCCCG	9
UCCACCAG	9
CCCACAUG	7-8
CUCAACCG	8
ACCCUCAG	7
ACCACCUG	1,3-6,8,10
UAACACCG	1-6,10
AUCCCAAG	2-3
AAAUCCCG	1
CCCUCAUG	1,3-4
UACUCCCG	4
AU(CCUC)CG	5
CCUAUCAG	10
CCUAACUG	5
CUUAACCG	4,7,9
UAAUCCCG	9
CUACAAUG	1-10
UACUACAG	10
UAAUACCG	7-9
AUUACCAG	3
AUAACCUG	6-8,10

Table 1 continues on following page.

Table 1. (continued)

Oligonu-cleotide sequence	Present in organism number
ACAAUCUG	9
AAAUCCUG	1-2,6-9
AUAAACUG	3-6
AUAAAUAG	2
(CU,CCUU)CG	4
AUCCUUCG	4
UCUAACUG	1
CUUAACUG	2-3,5-6
UAAUCCUG	1-3,6
UCUAAAUG	1
UUAAAUCG	10
CAUAUAUG	10
AAAUCUUG	10
AAAUUCUG	2-3
AUAAAUUG	1
CUUUUCAG	6
UUCUCAUG	2
UUUAAUCG	9
UAUCAUUG	9
UUUAAAUG	2-3
UUUAAUUG	1-8,10
UUUUUUCG	2-3
UUUUAUUG	1
9-mers	
CCCACCAAG	4-5
CACACACCG	1-10
(CCA,CAA)CAG	8
CCCAACAAG	7-9
AACCCCAAG	6
AAACCCAAG	4
CCUCACCAG	8
CCUACCAAG	6
CCUACAACG	10
AUAACCCCG	6,8,10
AAACCUCCG	1-6
CACACUAAG	1-6
AUAAACCCG	6
UACUCCCAG	1-3,5-6
UAAUCCCCG	7
AAUCCCCUG	1,3-6
CUUACCAAG	1-3
(UC)ACACAUG	3
(UC)ACAAUCG	2-3
UCAUAACCG	4
CUAAUACCG	3
ACCCUUAAG	7
AUAAUCCCG	9
AUAACCCUG	1-5
AUAAUACCG	4-5
AUAUACAAG	9
UCUUACCAG	10
UCACUAUCG	6
UAAUCCCUG	10
UAAUCCUCG	8
AAUUUCCCG	10
AAUCCUCUG	2
UCAUAAUCG	1,5
CUAAUACUG	1
CAUCAUAUG	10
AUAAUUCCG	10
AAUUAUCCG	7-9
UUUAAAACG	7
UAAACUAUG	7
AUAAUACUG	2
CUAUUACUG	9
UUAAAUUCG	1
UUUAAUAAG	2
UUAUAUUCG	2
UAUUUCUAG	9
UUUAUUAAG	1
CUUUUAUUG	6
UUUUUAUUG	2,4
UUUUUUUCG	1
10-mers	
AAUAACCCCG	7
ACCACCUAUG	9
AAUCUCACCG	8
AAAUCUCACG	4
UAACUCAAAG	8
AAACUUAAAG	1-10
ACCUUACCUG	10
UUACCAUCAG	3
UACCUACUAG	10
AAUCACUUCG	5
AACCCUUAUG	6
UAAAUAACUG	9
UUCUUCACCG	6
ACUCUACUUG	9
CUUAACUAUG	1
AUACUAUUAG	2,4-5
UUCCCUAUUG	4
UCUUCUUAAG	4
AUUUUUUUCG	1
UUUUCUUUUG	5
11-mers	
ACAACUCACCG	10
AAAUCCCACAG	6
CAUCUCACCAG	7,9
UAACUCACCCG	9
AAAUCUCACCG	7,9
AAACACCUUCG	6
AAAUCCCAUAG	5
UCCCUCCCCUG	10
CAUAUCCUCCG	10
AAAUCCUAUAG	3
UUUCAACAUAG	7,9
A(UA,UCA,CUA)UG	6
UUUCAAUAUAG	10
CUUUUCUUAAG	1,3
CUUUUCAUUAG	2
UUCUUUAAUCG	7
UUUUUUUCCUG	1
UUUUUUUUAAG	2
12-mers	
CCACCCAAAAAG	1-2,4,6
UCAAACCACCCG	8-10
UCAAACCAUCCG	7
ACAUCUCACCAG	1-6
CCACUCUUAACG	4-6
CCAUUCUUAACG	1-3
CUCAACUAUUAG	10
CCACUAUUAUUG	7
CAAUUAUUCCUG	2
CCACUUUUAUUG	8
CCAUUUUUAUUG	5
(CUA,CUUUUA)UUG	3
13-mers	
UAAACUACACCUG	10
(CAA,CCA)CAUUCUG	6
UAAUACUCCAUAG	9
UUUCAAAAUAACG	8
AUAAUUUUUCCUG	3
(UUU,CUU,CU)AAAUG	5
14-mers	
AAAACUUUACCAUG	9
AAAACUUUACAAUG	7-8,10
AUUUUU(CCU,CU)UUG	2
15-mers	
UCUAAAACACACCUG	8
AUAACCUACCCUUAG	1-3
AUAACCUAACCUUAG	4
AAUAAUACCCUAUAG	8
AAUAAUACUCCAUAG	7
AUAAUCUACCCUUAG	5
16-mers	
UAAUCCCCUAAACCAG	6
AAAUCCUAUAAUCCCG	4
AAUCUCCUAAACAUAG	5
CAAUCUCUUAAACCUG	7
UAAUCUCCUAAACCUG	4
AAAUCCUAUAAUCCUG	5
17-mers	
CAAUCUUUUAAACCUAG	3
UAAU(CCU,CU)AAACUUAG	1-2
AUAAU(CCU,CU)AAACCUG	9
18-mer	
AACAAUCUCCUAAACCUG	8
24-mer	
(AAACA,UAAUCUCA)—CCCAUCCUUAG	10
termini	
5′ end	
pAG	4,6
pAAUCCG	5
pAAUCUG	1,3
pAUUCUG	2,7-10
3′ end	
$AUCACCUCCU_{OH}$	1-10

First column is oligonucleotide sequence; second column shows organisms in which that sequence is found. Organisms are designated by number (see Fig. 1) as follows: 1, *M. arbophilicum;* 2, *M. ruminantium* strain *PS;* 3, *M. ruminantium* strain *M-1;* 4, *M. formicicum;* 5, *M.* sp. strain *M.o.H.;* 6, *M. thermoautotrophicum;* 7, Cariaco isolate JR-1; 8, Black Sea isolate JR-1; 9, *Methanospirillum hungatii;* 10, *Methanosarcina barkeri.* Multiple occurrences of a sequence in a given organism are denoted by repeating the organism's number in column 2: e.g., *1–4,6–8;3,7,;3* signifies a double occurrence in organism 7 and a triple occurrence in organism 3.

$$S_{AB} = 2\,N_{AB} / (N_A + N_B)$$

where N = number of residues in oligomers of at least length L, and N_{AB} = total number of residues in coincident oligomers between catalogs A and B (Fox *et al. IJSB* 1977)

Table 2 Relatedness in pairwise comparisons

Similarity coefficient, S_{ij}

	E. coli	*B. subtilis*	*R. spheroides*	6301	6714	6701	6308	*Porphyridium*	*Euglena*
E. coli		33	37	31	31	28	28	27	27
B. subtilis	86		35	37	34	32	32	34	26
R. spheroides	88	87		32	33	30	29	30	23
6301	85	88	86		49	44	41	36	35
6714	85	87	86	92		58	52	47	32
6701	84	86	85	90	94		83	43	33
6308	84	86	84	89	92	98		42	32
Porphyridium	83	87	85	88	91	90	89		36
Euglena	83	82	80	87	86	86	86	87	

% Homology, C_{ij}

There are several ways in which the data of Table 1 can be used to derive single values (similarity coefficients S_{ij}) which measure the relatedness of any pair of RNAs i and j. That which seems most reasonable to us defines S_{ij} as

$$\left[200 \sum_{N=5}^{\infty} k_{N_{ij}} N\right] / \sum_{N=5}^{\infty} (t_{N_i} + t_{N_j}) N$$

where N is oligonucleotide length (in residues), $k_{N_{ij}}$ is the number of N-mers common to the two RNAs, and t_{N_i} and t_{N_j} are the total number of N-mers obtained from RNAs i and j, respectively. S_{ij} is thus equivalent to 200 times the number of individual nucleotide residues present in sequences (pentamers and larger) which are common to two rRNAs divided by the number of individual residues present in all sequences (pentamers and larger) found in both rRNAs. It will range in value from 100 for identical molecules down to 10–15 for totally unrelated species. (Chance sequence coincidences prevent S_{ij} from reaching zero in such cases.) Table 2 shows (above the diagonal) values of S_{ij} computed for the 36 possible pairwise comparisons involving the blue-green and chloroplast catalogues and, in addition, representative bacterial 16S rRNA catalogues obtained by Woese and his collaborators for *Escherichia coli*[13,14], *Bacillus subtilis*[19], and *Rhodopseudomonas spheroides*[20] (a photosynthetic bacterium). For certain purposes it is useful to convert values of S_{ij} to values (C_{ij}) for the probability of any single nucleotide residue having remained unchanged during the evolutionary divergence of RNAs i and j, since the latter has more obvious physical significance (and can be equated with "% homology"). Such conversion was made with

Bonen & Doolittle, *Nature* 1976

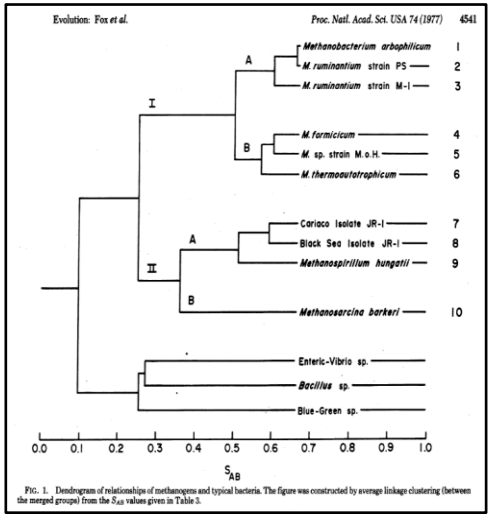


FIG. 1. Dendrogram of relationships of methanogens and typical bacteria. The figure was constructed by average linkage clustering (between the merged groups) from the S_{AB} values given in Table 3.

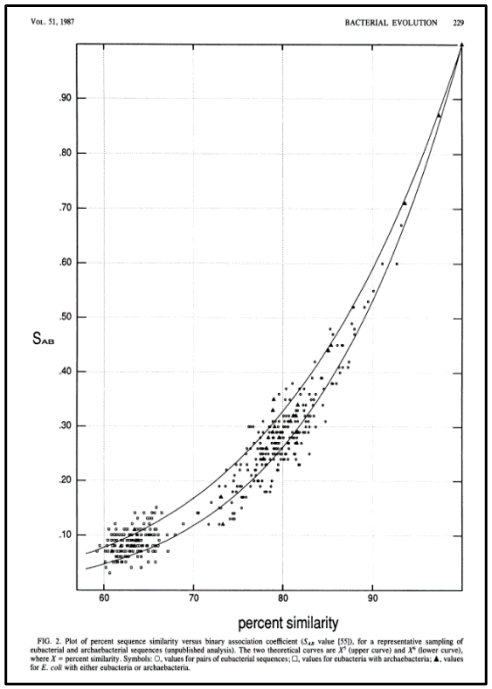


FIG. 2. Plot of percent sequence similarity versus binary association coefficient (S_{AB} value [55]), for a representative sampling of eubacterial and archaebacterial sequences (unpublished analysis). The two theoretical curves are X^2 (upper curve) and X^4 (lower curve), where X = percent similarity. Symbols: ○, values for pairs of eubacterial sequences; □, values for eubacteria with archaebacteria; ▲, values for *E. coli* with either eubacteria or archaebacteria.

Fox *et al., PNAS* 1977 (top)
Woese, *Microbiol. Rev*. 1987 (bottom)

The three kingdoms (domains) of life

Proc. Natl. Acad. Sci. USA
Vol. 74, No. 11, pp. 5088–5090, November 1977
Evolution

Phylogenetic structure of the prokaryotic domain: The primary kingdoms

(archaebacteria/eubacteria/urkaryote/16S ribosomal RNA/molecular phylogeny)

CARL R. WOESE AND GEORGE E. FOX*

Department of Genetics and Development, University of Illinois, Urbana, Illinois 61801

Communicated by T. M. Sonneborn, August 18, 1977

ABSTRACT **A phylogenetic analysis based upon ribosomal RNA sequence characterization reveals that living systems represent one of three aboriginal lines of descent: (*i*) the eubacteria, comprising all typical bacteria; (*ii*) the archaebacteria, containing methanogenic bacteria; and (*iii*) the urkaryotes, now represented in the cytoplasmic component of eukaryotic cells.**

The biologist has customarily structured his world in terms of certain basic dichotomies. Classically, what was not plant was animal. The discovery that bacteria, which initially had been considered plants, resembled both plants and animals less than plants and animals resembled one another led to a reformulation of the issue in terms of a yet more basic dichotomy, that of eukaryote versus prokaryote. The striking differences between eukaryotic and prokaryotic cells have now been documented in endless molecular detail. As a result, it is generally taken for granted that all extant life must be of these two basic types.

Thus, it appears that the biologist has solved the problem of the primary phylogenetic groupings. However, this is not the case. Dividing the living world into *Prokaryotae* and *Eukaryotae* has served, if anything, to obscure the problem of what extant groupings represent the various primeval branches from the common line of descent. The reason is that eukaryote/prokaryote is not primarily a phylogenetic distinction, although to construct phylogenetic classifications between domains: Prokaryotic kingdoms are not comparable to eukaryotic ones. This should be recognized by an appropriate terminology. The highest phylogenetic unit in the prokaryotic domain we think should be called an "urkingdom"—or perhaps "primary kingdom." This would recognize the qualitative distinction between prokaryotic and eukaryotic kingdoms and emphasize that the former have primary evolutionary status.

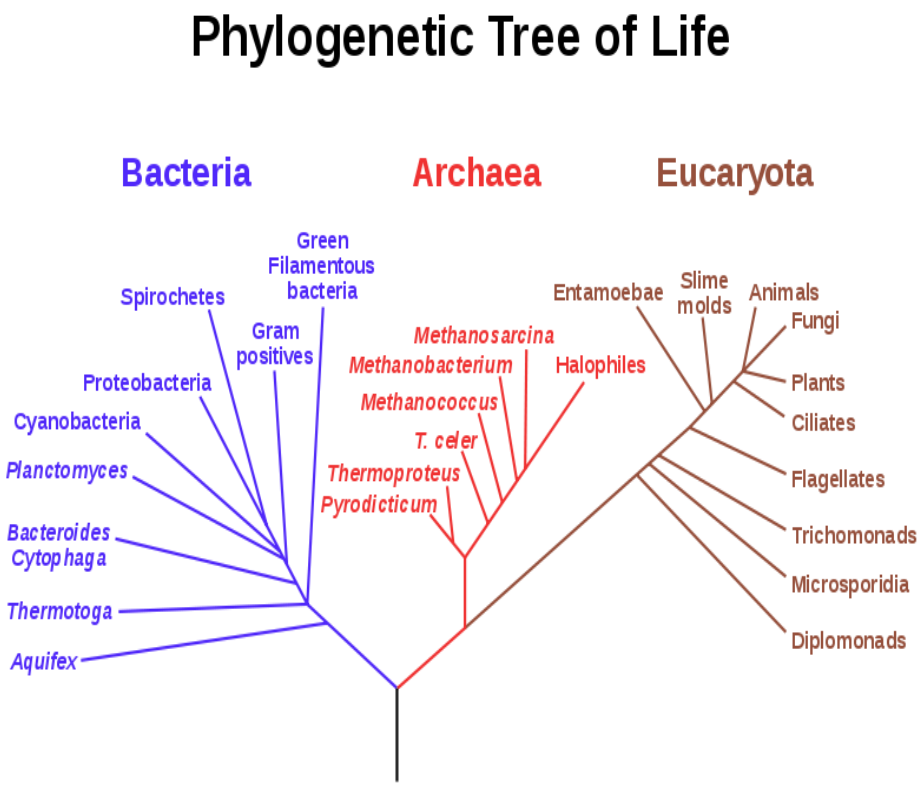
The passage from one domain to a higher one then becomes a central problem. Initially one would like to know whether this is a frequent or a rare (unique) evolutionary event. It is traditionally assumed—without evidence—that the eukaryotic domain has arisen but once; all extant eukaryotes stem from a common ancestor, itself eukaryotic (2). A similar prejudice holds for the prokaryotic domain (2). [We elsewhere argue (6) that a hypothetical domain of lower complexity, that of "progenotes," may have preceded and given rise to the prokaryotes.] The present communication is a discussion of recent findings that relate to the urkingdom structure of the prokaryotic domain and the question of its unique as opposed to multiple origin.

Phylogenetic relationships cannot be reliably established in terms of noncomparable properties (7). A comparative approach that can measure degree of difference in comparable

Table 1. Association coefficients (S_{AB}) between representative members of the three primary kingdoms

	1	2	3	4	5	6	7	8	9	10	11	12	13
1. *Saccharomyces cerevisiae*, 18S	—	0.29	0.33	0.05	0.06	0.08	0.09	0.11	0.08	0.11	0.11	0.08	0.08
2. *Lemna minor*, 18S	0.29	—	0.36	0.10	0.05	0.06	0.10	0.09	0.11	0.10	0.10	0.13	0.07
3. L cell, 18S	0.33	0.36	—	0.06	0.06	0.07	0.07	0.09	0.06	0.10	0.10	0.09	0.07
4. *Escherichia coli*	0.05	0.10	0.06	—	0.24	0.25	0.28	0.26	0.21	0.11	0.12	0.07	0.12
5. *Chlorobium vibrioforme*	0.06	0.05	0.06	0.24	—	0.22	0.22	0.20	0.19	0.06	0.07	0.06	0.09
6. *Bacillus firmus*	0.08	0.06	0.07	0.25	0.22	—	0.34	0.26	0.20	0.11	0.13	0.06	0.12
7. *Corynebacterium diphtheriae*	0.09	0.10	0.07	0.28	0.22	0.34	—	0.23	0.21	0.12	0.12	0.09	0.10
8. *Aphanocapsa* 6714	0.11	0.09	0.09	0.26	0.20	0.26	0.23	—	0.31	0.11	0.11	0.10	0.10
9. Chloroplast (*Lemna*)	0.08	0.11	0.06	0.21	0.19	0.20	0.21	0.31	—	0.14	0.12	0.10	0.12
10. *Methanobacterium thermoautotrophicum*	0.11	0.10	0.10	0.11	0.06	0.11	0.12	0.11	0.14	—	0.51	0.25	0.30
11. *M. ruminantium* strain M-1	0.11	0.10	0.10	0.12	0.07	0.13	0.12	0.11	0.12	0.51	—	0.25	0.24
12. *Methanobacterium* sp., Cariaco isolate JR-1	0.08	0.13	0.09	0.07	0.06	0.06	0.09	0.10	0.10	0.25	0.25	—	0.32
13. *Methanosarcina barkeri*	0.08	0.07	0.07	0.12	0.09	0.12	0.10	0.10	0.12	0.30	0.24	0.32	—

The 16S (18S) ribosomal RNA from the organisms (organelles) listed were digested with T1 RNase and the resulting digests were subjected to two-dimensional electrophoretic separation to produce an oligonucleotide fingerprint. The individual oligonucleotides on each fingerprint were then sequenced by established procedures (13, 14) to produce an oligonucleotide catalog characteristic of the given organism (3, 4, 13–17, 22, 23; unpublished data). Comparisons of all possible pairs of such catalogs defines a set of association coefficients (S_{AB}) given by: $S_{AB} = 2N_{AB}/(N_A + N_B)$, in which N_A, N_B, and N_{AB} are the total numbers of nucleotides in sequences of hexamers or larger in the catalog for organism A, in that for organism B, and in the interreaction of the two catalogs, respectively (13, 23).



From Wikimedia Commons
after Carl Woese and colleagues (~1972 ff.)

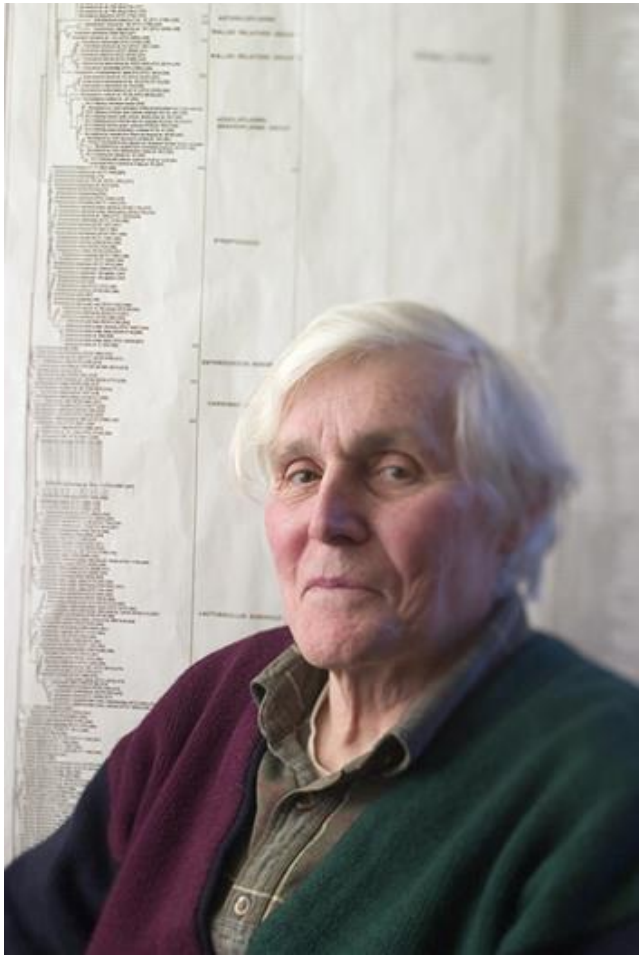
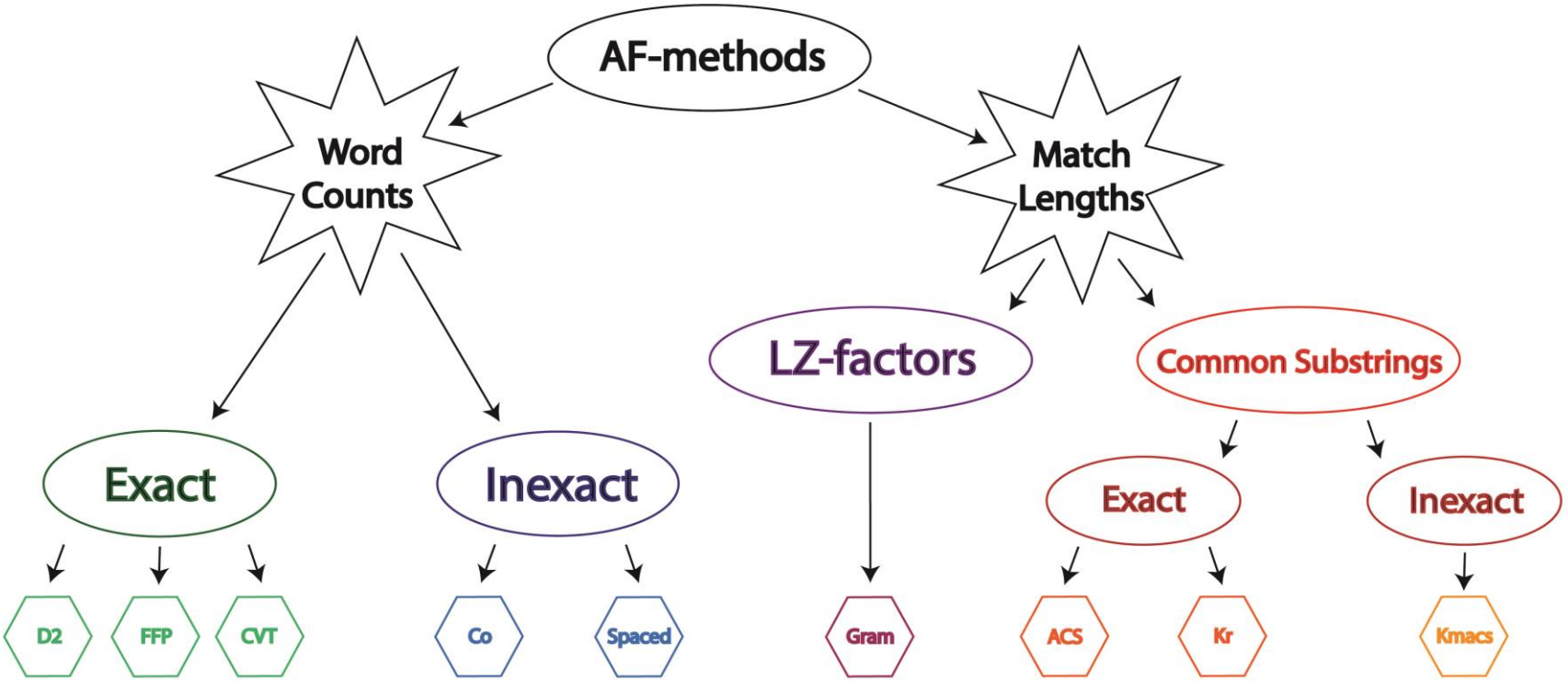


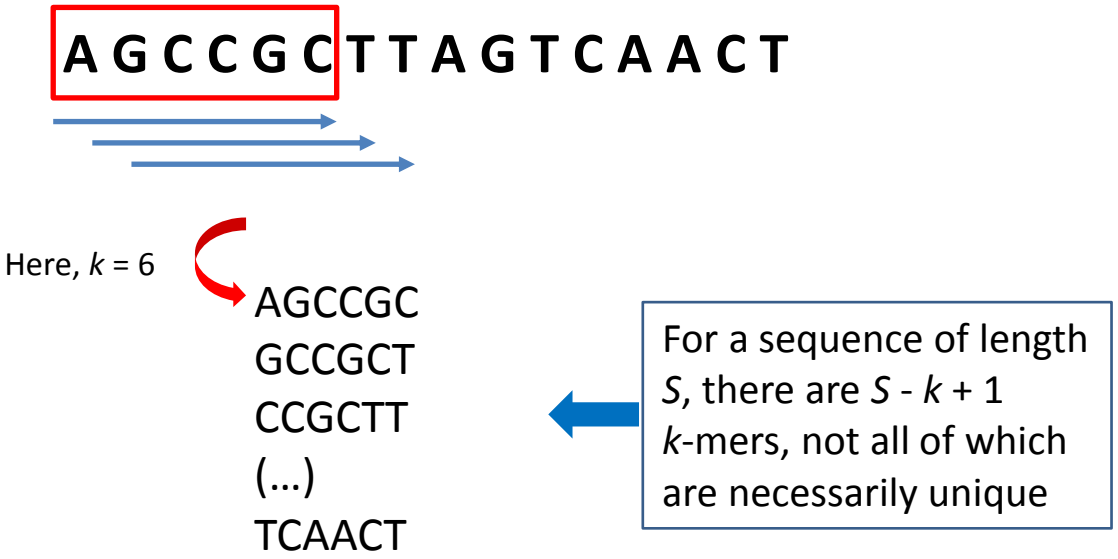
Image courtesy of Institute for Genomic Biology
University of Illinois

Alignment-free methods



Guillaume Bernard, after Haubold, *Briefings in Bioinformatics* (2013)

k-mers / *k*-tuples / *k*-words / *n*-mers / *n*-grams



There's also a parallel world of ***patterns***

Höhl, Rigoutsos & Ragan, *Evol Bioinf* 2:357-373 (2006)

D_2 statistics: a brief overview

The D_2 statistic is the count of exact word matches of length k between two sequences

For alphabet A, there are A^k possible words w of length k. Given sequences X and Y,

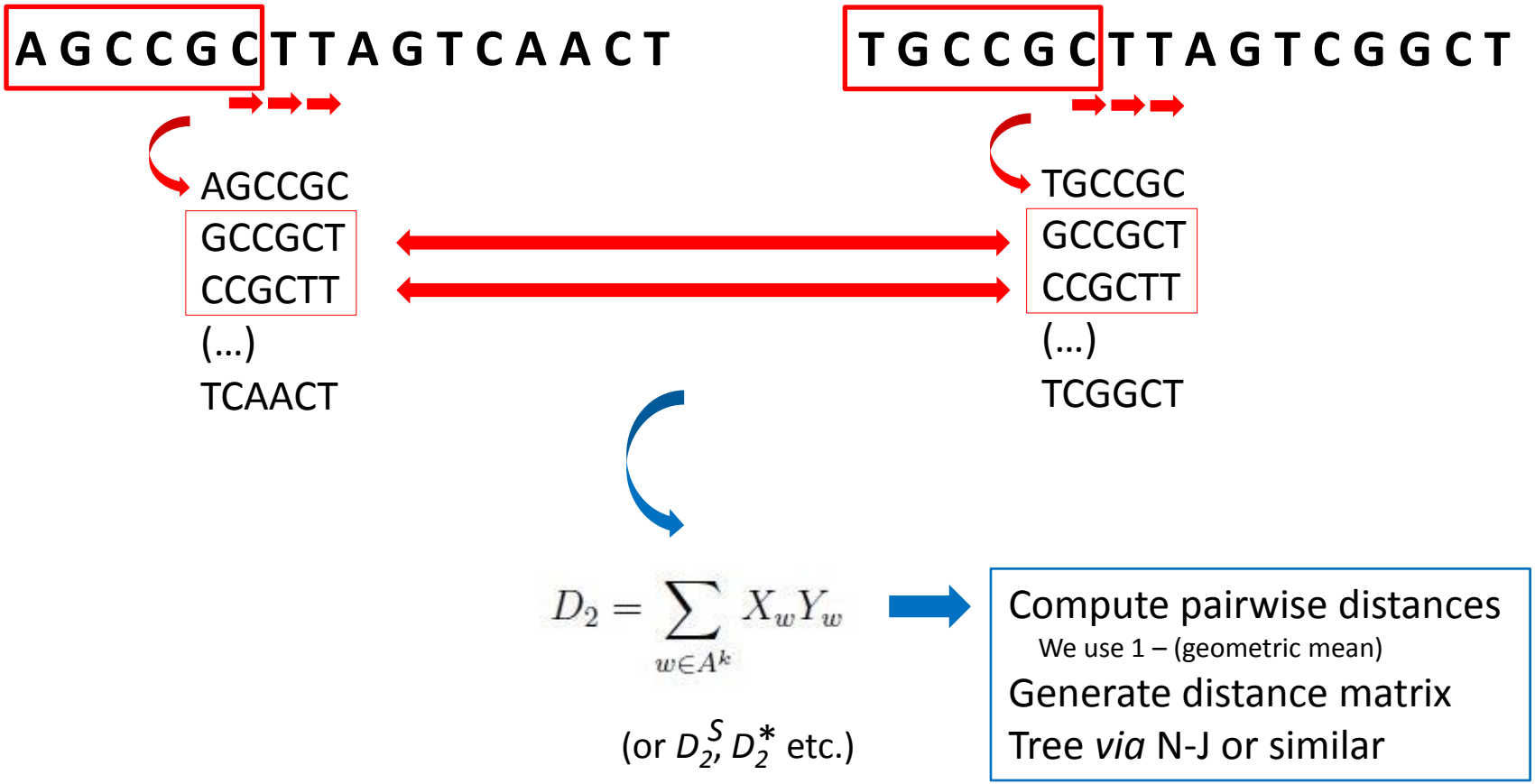
$$D_2 = \sum_{w \in A^k} X_w Y_w$$

Because D_2 is sensitive to sequence length, the statistic is often normalised by the probability of occurrence of specific words (D_2^S), or by assuming a Poisson distribution of word occurrence (D_2^*) for long words

Although defined for exact word matches, D_2 can be easily extended to n mismatches (neighbourhood of order n): D_2^n

Chor *et al*., *Genome Biol* 10:R108 (2009); Reinert *et al*., *J Comput Biol* 16:1615-1634 (2009); Reinert *et al*., *J Comput Biol* 17:1349-1372 (2010); Burden *et al*., *J Comput Biol* 21:41-63 (2014)

D_2-based distance



Other AF methods based on word counts

Feature frequency profile *Sims & Kim, PNAS 2011*	Compares *k*-mer frequency profiles (Jensen-Shannon divergence) & computes a pairwise distance
Composition vector *Wang & Hao, JME 2004*	FFP using word frequencies normalised by probability of chance occurrence
Word context **Co-phylog*:*** *Yi & Jin, NAR 2013*	Pairwise distances based on proportions of *k*-mers that differ in a certain position; more-realistic branch lengths
Spaced word frequencies *Leimeister, Bioinformatics 2014*	Considers word mismatches as well as matches; less statistical dependency between neighbouring matches

AF methods based on match length

In general, similar sequences share longer exact words

Grammar-based distance ***d*-gram:** Russell, *BMC Bioinf* 2010	The concatenate of two sequences is more compressible (*e.g.* by Lempel-Ziv) if the sequences are similar
Average common substring Ulitsky*, J Comp Biol* 2006	Mean of longest matches between sequences, starting from each position; unlike L-Z, word overlap is allowed
Shortest unique substring Haubold*, J Comp Biol* 2009	Longest common substring + 1, corrected for random matches: "AF version of Jukes-Cantor distance"
Underlying subwords Comin, *Algorith Mol Biol* 2012	Like ACS, but discards common subwords that are covered by longer (more-significant) ones
***k*-Mismatch ACS (kmacs)** Leimester, *Bioinformatics* 2014	ACS with *k* (in our notation, *n*) mismatches

Shortest unique substring (shustring) algorithm

Unique substrings remain unique upon extension, so use only the shortest ones

The length of the shortest unique substring is inversely related to information content of the sequence

- *C. elegans* autosomes L= 11 (one example of 10)
- human autosomes L= 11, but Y chromosome L= 12
- mouse autosomes L= 11, but Y chromosome L=12

Given a random sequence model, the probability of finding even one shustring of L= 11 in human is $<10^{-100}$

In human and mouse (and presumably other) genomes, shustrings are preferentially located within 1 kb of protein-coding genes

Haubold *et al*., *J Comp Biol* 2009

Under simplifying assumptions, there's a relationship between
d (≈ mutational distance between two sequences) and average shustring length

Haubold *et al*., *J Comp Biol* 2009

The probability that a shustring of length *X* is longer than a threshold *t* is given by

$$Pr\{X > t\} = (1 - m/l)^t \approx e^{-tm/l}$$

where *m*= number of mutations and *l*= length of sequence

If all nucleotides are equally frequent, the correction for random matches is

$$Pr\{X \leq t\} = (1 - e^{-tm/l})(1 - 4^{-t})^l$$

Correction for multiple substitutions yields an AF version of classical (J-C) distance

$$d_{kr} = -\frac{3}{4} ln \left(1 - \frac{4}{3l} m\right)$$

Can we compute accurate trees using AF-based distances ?

How do we best ask this question ?

Simulated data

- Generate replicate data on a known tree, varying data size, substitution model, tree shape, branch lengths etc.
- Extract *k*-mers & compute a tree; sweep over relevant parameters
- Compare topologies (R-F)
- Measure performance (precision, recall, sensitivity…)

Advantages/disadvantages

- We can study effects of different factors & scenarios individually
- Sequence models may be too simplistic

Empirical data

- Identify empirical datasets for which someone has ventured a phylogenetic tree
- Extract *k*-mers & compute a tree; sweep over *k*
- Compare topologies (R-F)
- Count congruent/incongruent edges & try to interpret

Advantages/disadvantages

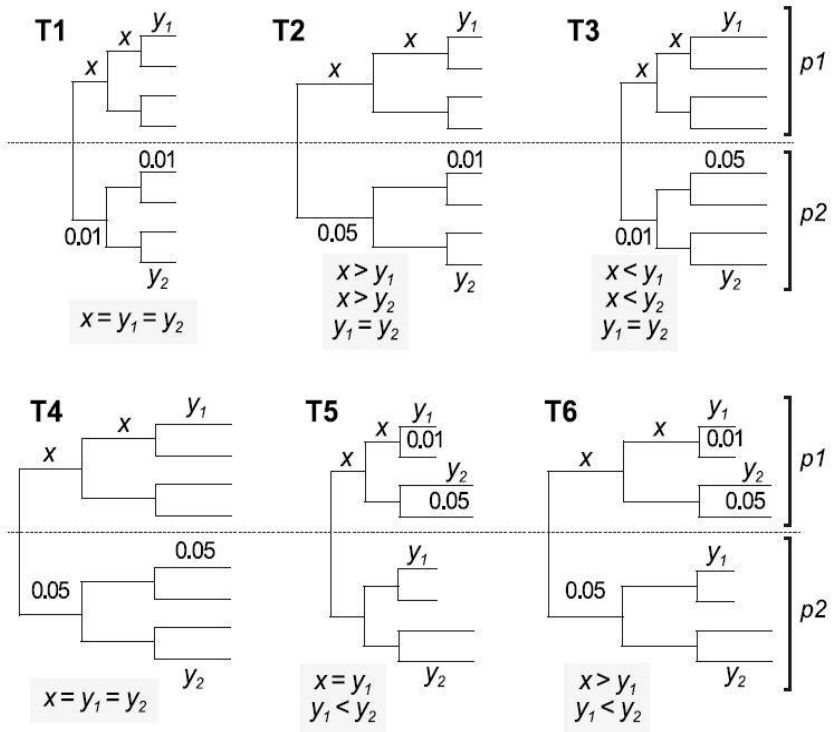
- Sequences are (by definition) real
- We can't study effects of different factors & scenarios individually
- The true tree remains unknown

First we simulated sequence data on a tree

Simulation software ranges from simplistic to maddeningly complex

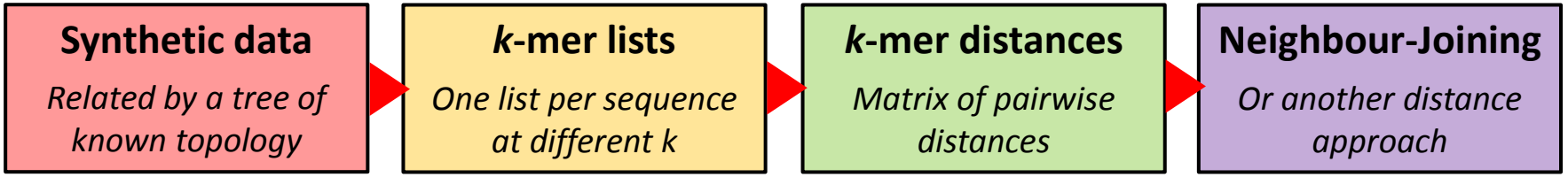
Using *evolver* (PAML) we simulated DNA and protein sequence sets on trees of different size (8 / 32 / 128 taxa), symmetry, and absolute and relative branch lengths

We also simulated DNA sequences on trees generated under a coalescent model (not shown)



Chan *et al*., *Scientific Reports* 4:6504 (2014)

We extracted *k*-mers at different *k*, computed distances under different variants of the D_2 statistic, and generated a N-J tree

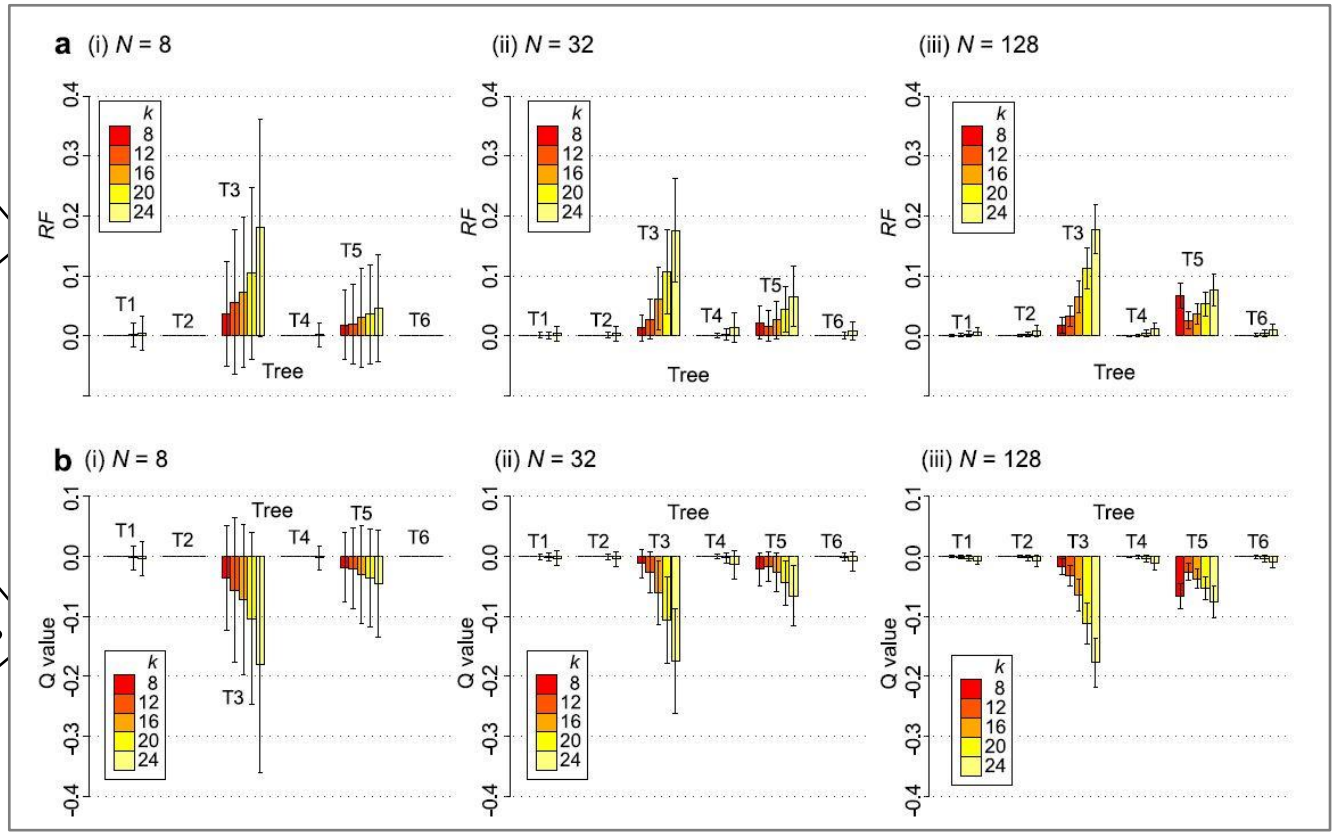


No method for confidence estimation is currently available, but one can imagine using a variant of the nonparametric bootstrap, or by jackknifing

Then we compared the D_2 + NJ tree with the known true topology, and with the topologies inferred using MSA + MrBayes

RF > 0 means D2n1 tree topology is incongruent

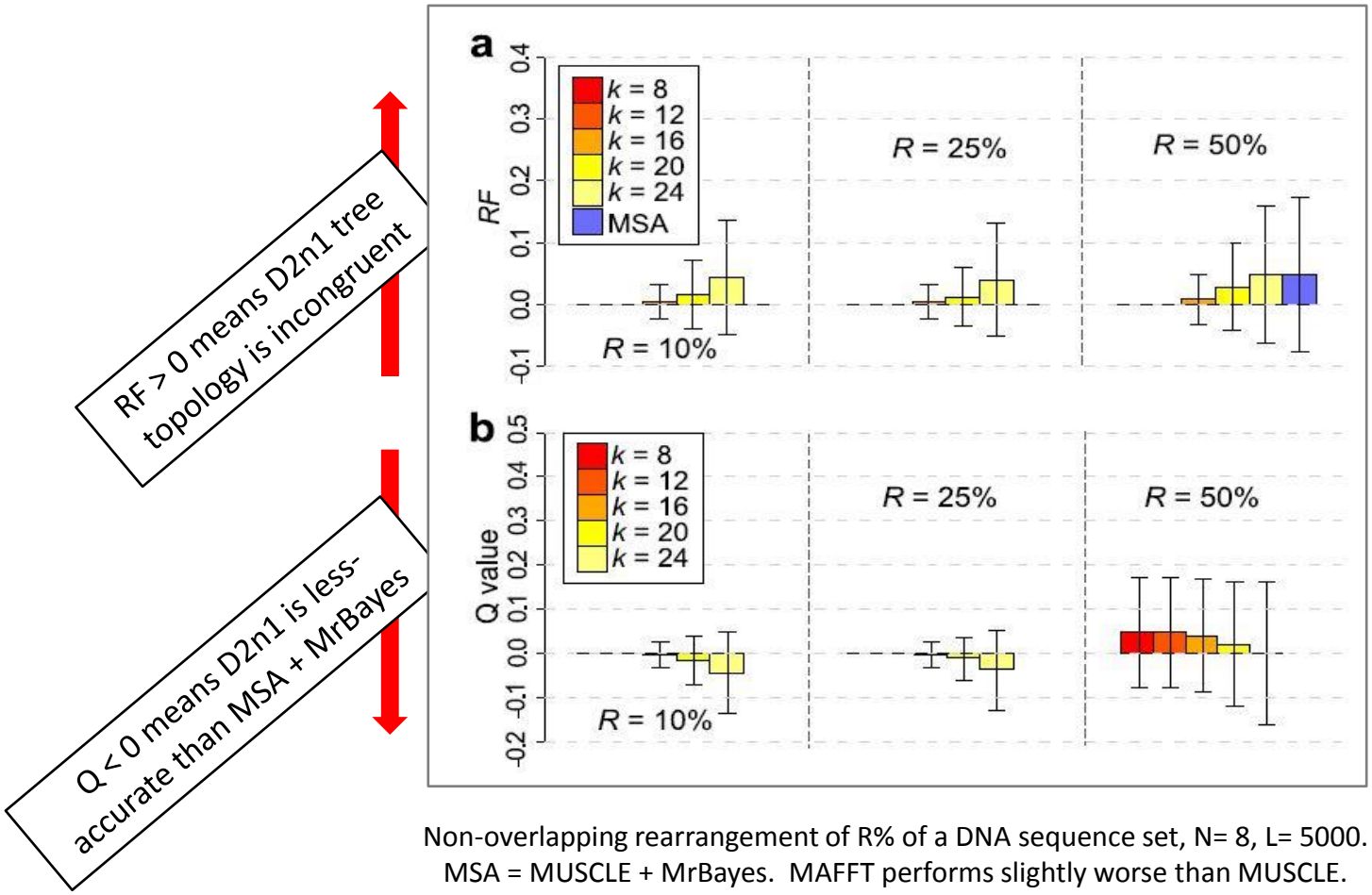
Q < 0 means D2n1 is less-accurate than MSA + MrBayes



DNA alphabet, L = 1500 nt, 100 replicates

Chan *et al.*, *Scientific Reports* 2014

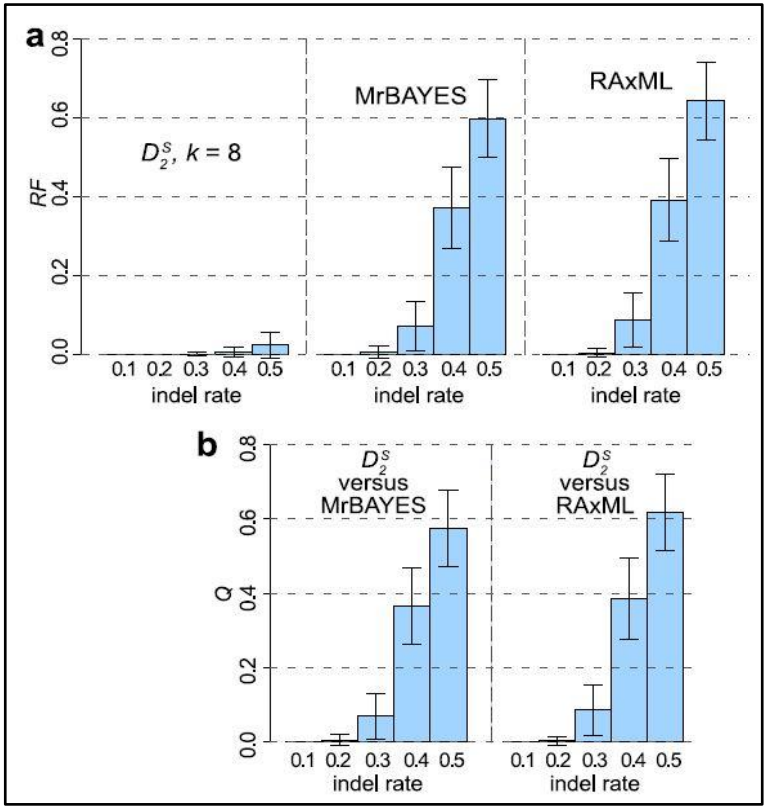
D_2 + NJ performs well with rearranged sequences



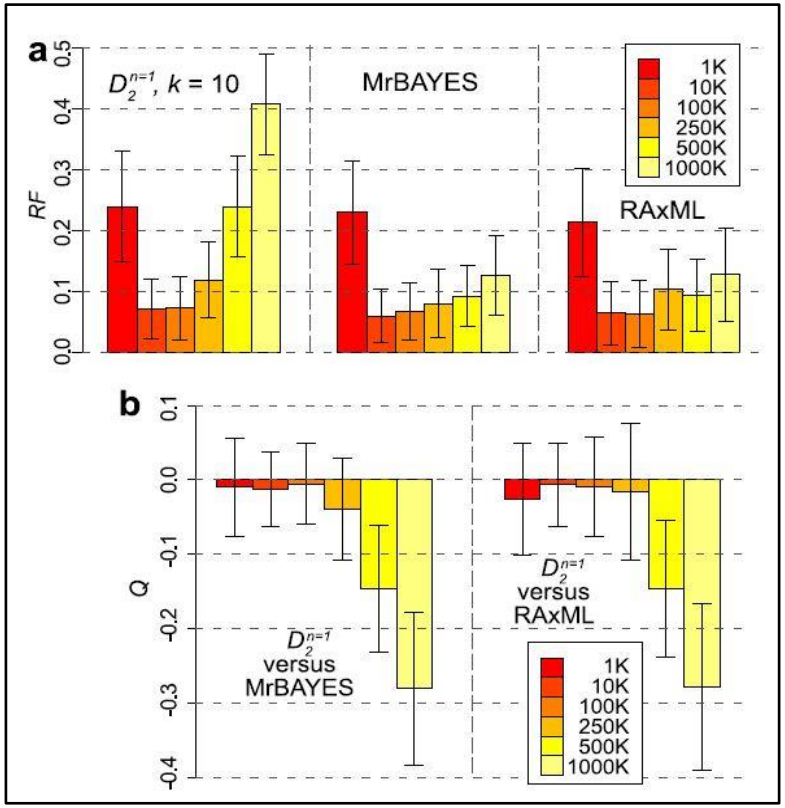
Non-overlapping rearrangement of R% of a DNA sequence set, N= 8, L= 5000.
MSA = MUSCLE + MrBayes. MAFFT performs slightly worse than MUSCLE.

Chan *et al*., *Scientific Reports* 2014

D_2^S + NJ is more-robust to indels than leading MSA methods



With data simulated under a coalescent model, D_2^{n1} + NJ results are similar to MSA except at high/low sequence divergence



Numbers in box are N_e = effective population size
Smaller N_e implies shorter branch lengths on the tree

Chan *et al.*, *Scientific Reports* 2014

Summary: trees computed from *k*-mer distances

Aspect	
Sequence length	**D_2**
Recent sequence divergence	MSA
Ancient sequence divergence	**D_2**
Among-site rate heterogeneity	**D_2** or MSA
Compositional bias	**D_2** or MSA
Genetic rearrangement	**D_2**
Incomplete sequence data	MSA
Insertions/deletions	**D_2**
Computational scalability	**D_2**
Memory consumption	MSA

Accuracy of D_2 methods increases with L

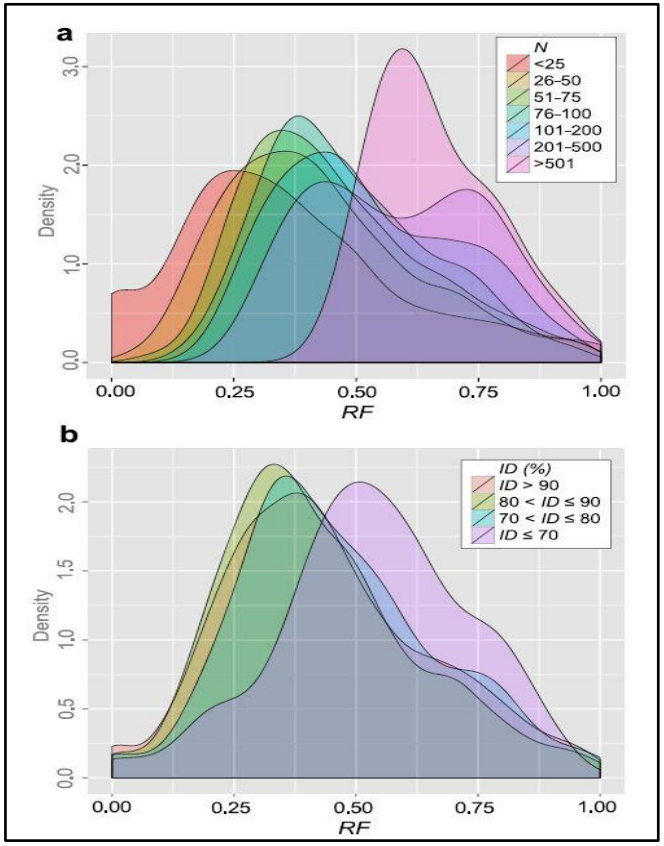
D_2 methods are more **robust** to ancient sequence divergence, to rearrangement and to indel frequency

D_2 methods are more **sensitive** to recent sequence divergence and to the presence of incomplete (truncated) data

Optimal *k* is negatively correlated with alphabet size, and is not greatly affected by N or L in a biologically relevant range

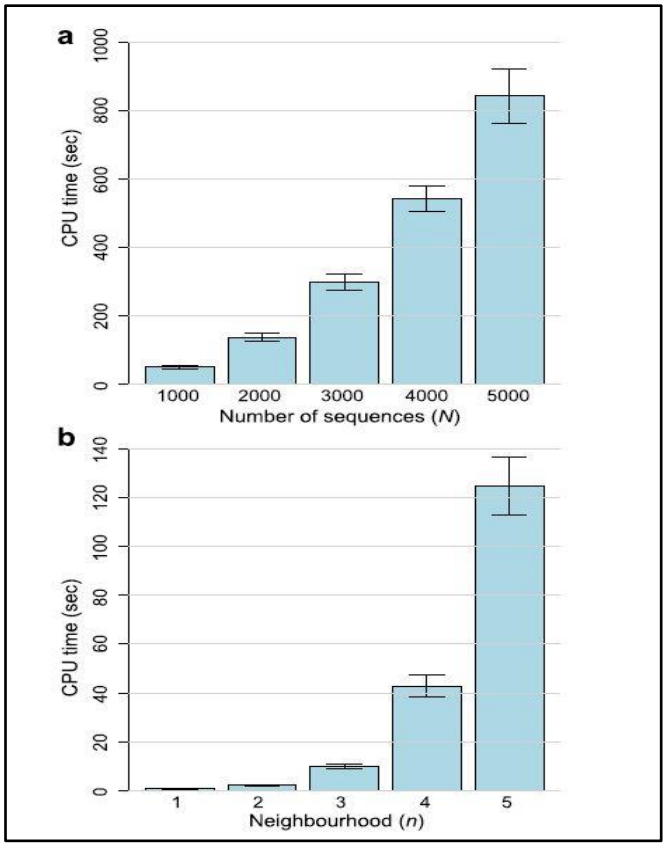
D_2 methods are more **scalable** to large data than are MSA-based approaches, but usually require more **memory**

D2 + NJ performs acceptably with empirical data, particularly if N is small and sequences are similar



RF probability densities (DNA data, D_2^S, k=8), 4156 trees from 2471 studies in TreeBASE (mean 59.4, median 41). We observed the identical tree in 106/4156 analyses.

The D_2 workflow scales almost linearly with sequence number if we keep to perfectly matched strings



D_2 walltime, 16S rRNA data (GreenGenes). Memory usage 378 MB (N=1000) to 2445 MB (N=5000).

Chan *et al*., *Scientific Reports* 2014

Nine AF methods are insensitive to frequency of inverted translocations

Synthetic genomes (ALF*)

30 genomes, ~2.5 Mbp each

Gene pool size 2500

Gene length 240-2000 nt

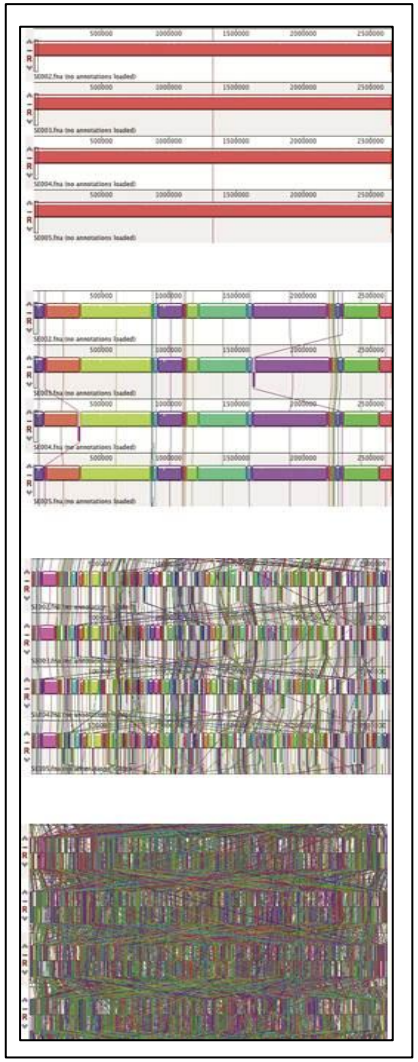
Speciation rate 0.5

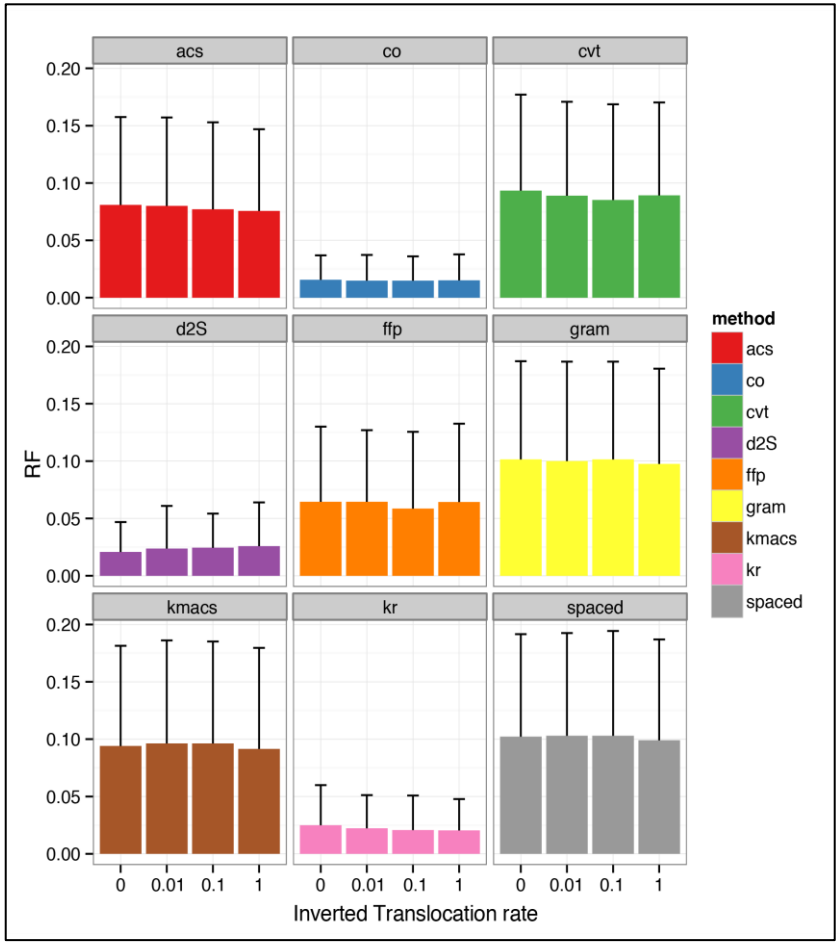
Extinction rate 0.1

Inverted translocations at rates {0, 0.01, 0.1, 1}

50 replicates

*Daniel *et al.*, *Mol Biol Evol* (2012)



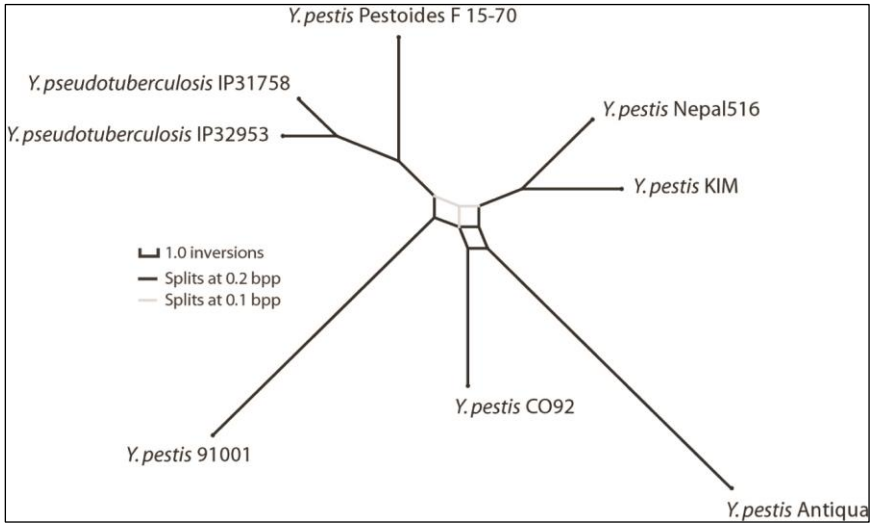


Guillaume Bernard

Eight *Yersinia* genomes: AF *versus* inversion phylogeny

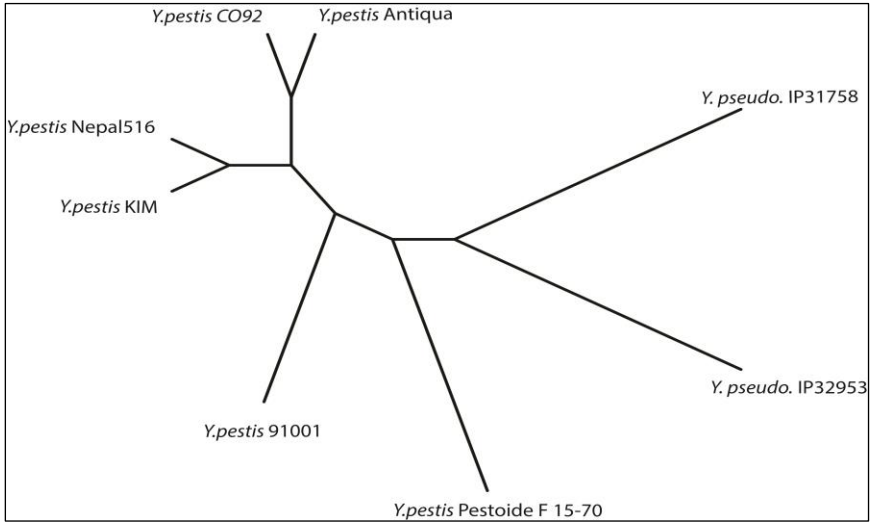
Consensus phylogenetic network based on inversions. Mauve (78 locally collinear blocks) then BADGER (Larget, *MBE* 2005). Requires extensive parameter estimation, with each run 500K MCMC generations.

Darling, Miklós & Ragan, *PLoS Genetics* (2008)

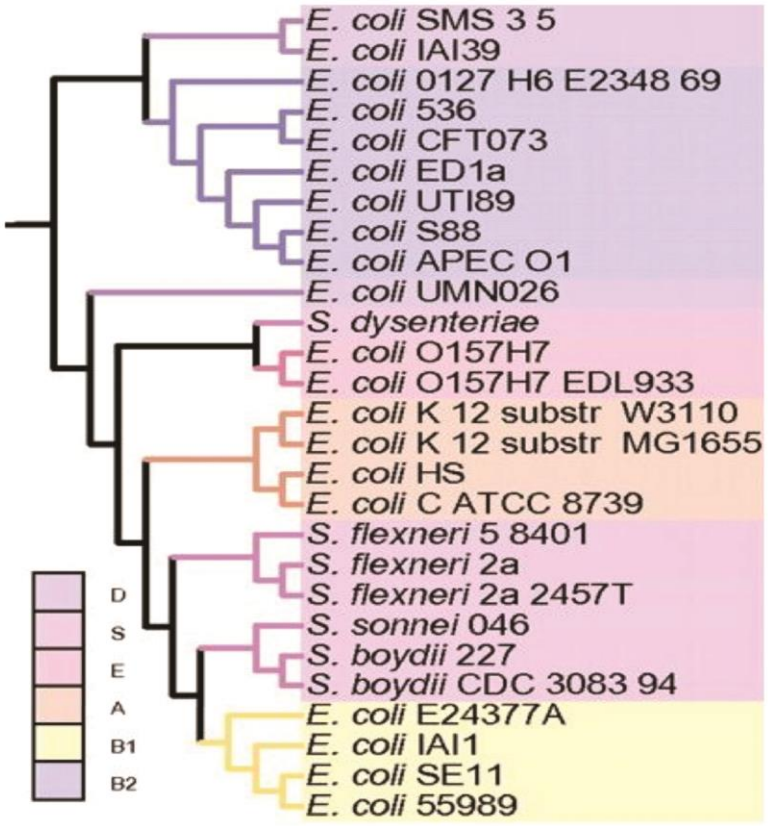


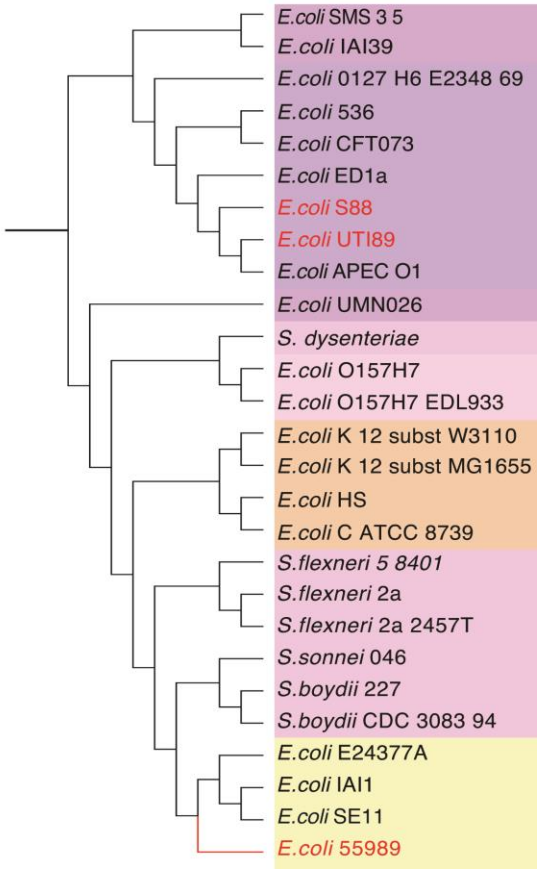
Kr (Haubold, *BMC Bioinformatics* 2005) yields a congruent phylogeny; no parameter optimisation, runtime 1 minute on laptop.

Bernard, Chan & Ragan, unpublished



27 *Escherichia coli* + *Shigella* genomes





ProgressiveMauve alignment (17 hours), extract 5282 single-copy gene sets $N \geq 4$, GBlocks, MrBayes (5M MCMC generations, 10 models) followed by MRP

Skippington & Ragan, *BMC Genomics* (2011)

Co-phylog (Yi & Jin, *NAR* 2013) with k=8, < 2 minutes on laptop

Bernard, Chan & Ragan, unpublished

Conclusions & outlook

AF methods hold considerable potential in phylogenetics & phylogenomics

But MSA-based approaches have a six-decade head start

With synthetic data, AF methods perform better than MSA-based approaches under some evolutionarily relevant scenarios, but worse under others

With empirical data, the jury is still out

(Some) AF methods could likely be subsumed under a rigorous model, although probably at the cost of speed & scalability

i.e. what makes them attractive in the first place

Efficient data structures & precomputation have much to offer

Other application areas include LGT analysis, and trees directly from NGS data

Song *et al*., *J Comp Biol* 2013; Yi & Jin, *NAR* 2013

Rob Beiko, Guillaume Bernard, Cheong Xin Chan, Xin-Yi Chua, Yingnan Cong, Aaron Darling, Leanne Haggerty, Michael Höhl & Elizabeth Skippington

Australian Research Council
James S. McDonnell Foundation
National Computational Infrastructure
- **National Supercomputing Facility**
- **Specialised Facility in Bioinformatics**

QFAB Bioinformatics

The University Of Queensland
AUSTRALIA

IMB Institute for Molecular Bioscience