

OMICS Journals are welcoming Submissions

OMICS International welcomes submissions that are original and technically so as to serve both the developing world and developed countries in the best possible way.

OMICS Journals are poised in excellence by publishing high quality research. OMICS International follows an Editorial Manager® System peer review process and boasts of a strong and active editorial board. Editors and reviewers are experts in their field and provide anonymous, unbiased and detailed reviews of all submissions.

The journal gives the options of multiple language translations for all the articles and all archived articles are available in HTML, XML, PDF and audio formats. Also, all the published articles are archived in repositories and indexing services like DOAJ, CAS, Google Scholar, Scientific Commons, Index Copernicus, EBSCO, HINARI and GALE.

For more details please visit our website:

<http://omicsonline.org/Submitmanuscript.php>

Alexander Bolshoy, Ph.D

Genomics based on gene lengths

Current research.

Molecular Evolution – Spring 2014

Topics:

COGs as Input Data

Phylogenomics

Robust Classification of Prokaryotic Genomes

Revealing Factors Affecting Gene Length

Input to clustering and ranking

COGS

Simplification: Genome as a Bag of Genes

- For our purposes: Genome = Text (over the alphabet of 4 letters); Gene = a substring of a Genome.
- There are two kinds of genes: Orphans and Family Members (having homologs).
- Prokaryotic Gene Family = COG.
- "COG" stands for Cluster of Orthologous Groups of proteins.
- The proteins that comprise each COG are assumed to have evolved from an ancestral protein, and are therefore either orthologs or paralogs. Orthologs are proteins from different species that evolved by vertical descent (speciation), and typically retain the same function as the original. Paralogs are proteins from within a given species that are derived from gene duplication.

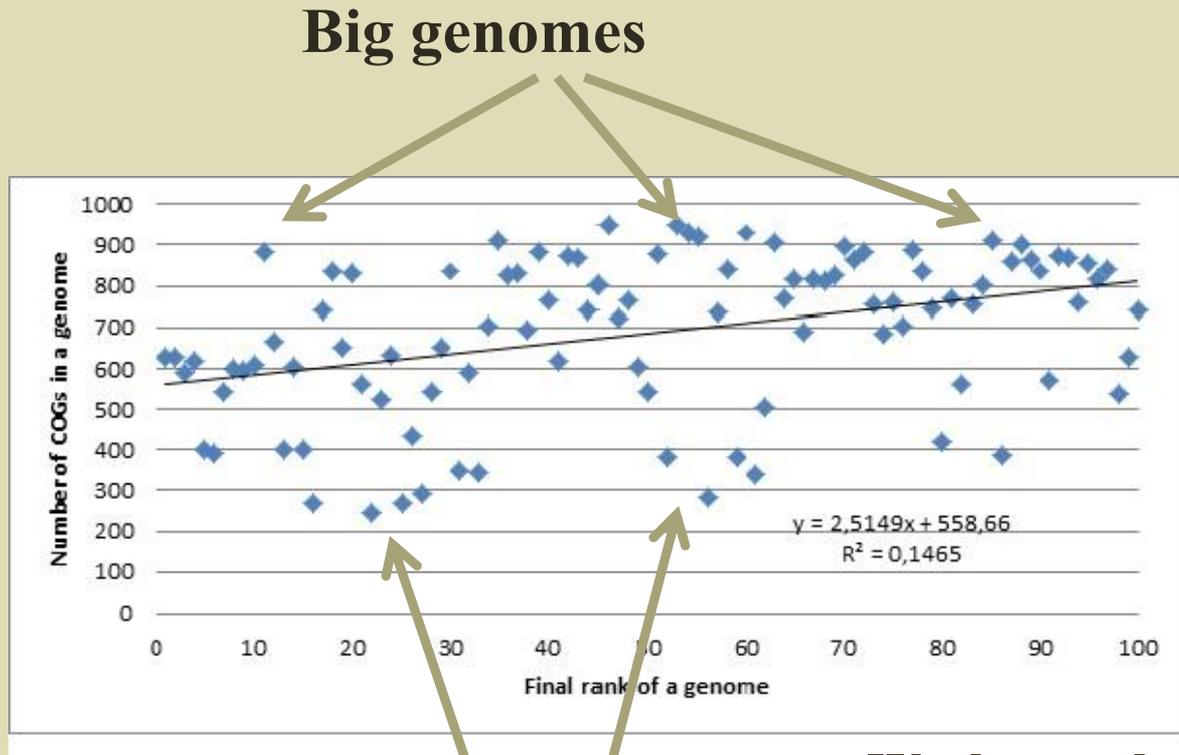
Filtering

- The current complete set consists of more than 1500 genomes and more than 5000 COGs.
- To test/debug our methods we take a small subset of randomly chosen 100 genomes and call it R1 subset.
- Naturally, there are COGs with a very small amount of proteins from R1 in them.
- We can apply filtering, 35% filtering means that we removed those COGs that have less than 35% of genomes from R1.
- After this filtering there are 1409 relevant COGs (next Figure, red bars).
- After taking only MEDIAN paralogs we get an input matrix 100 x 1409

R1

- A number of genes vary from genome to genome.
- Consequently, genomes of R1 are presented by different number of COGs: from small Mycoplasmas and Ureaplasma - the smallest and simplest self-replicating organisms with genome sizes from about 540 kb and less than 300 COGs inside to long genomes with more than 900 COGs
- $300 / 1409 = 0.21$
- $900 / 1409 = 0.64$

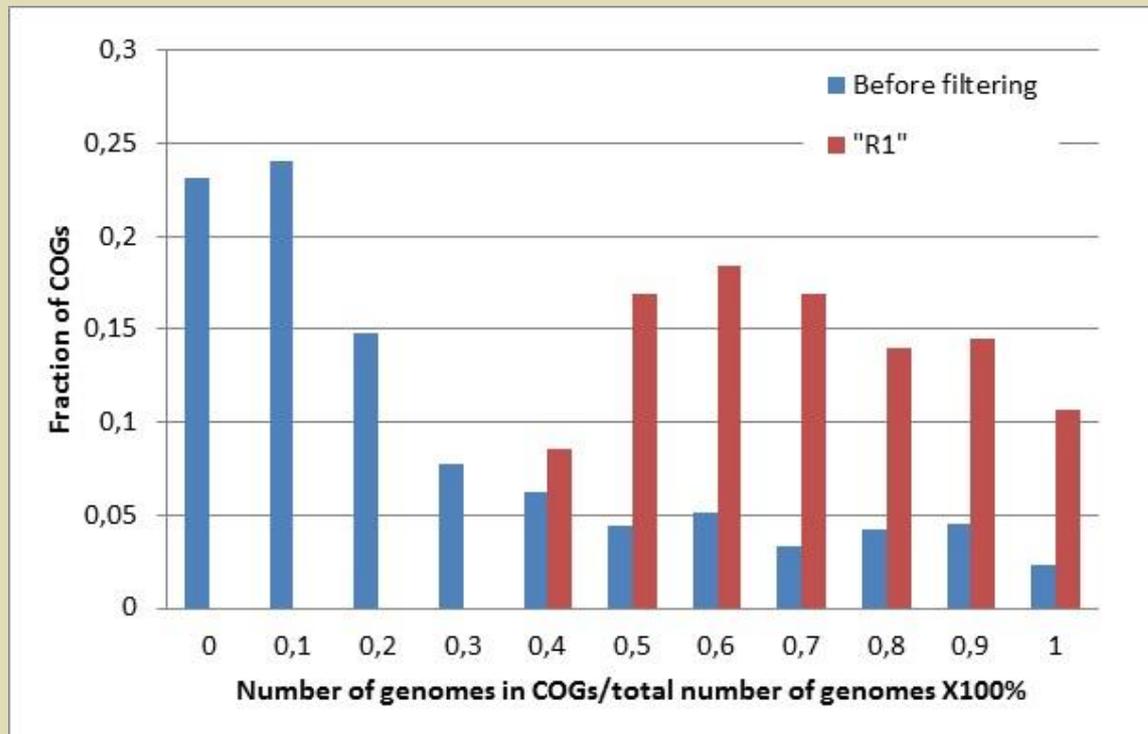
Sparse input data



- The available data is sparse: prokaryotic genomes do not contain every GF

Weak correlation exists
R1 genomes are ordered
accordingly to gene lengths

Histogram of number of genomes contained in each COG



Phylogenomics

As it was reported at
NCBI 2009

Research Objectives

- To build up a prokaryotic species tree using information bottleneck method on the whole-genome proteomes. Such a tree should include more than 500 genomes.
- To infer the produced tree of genomes using bootstrapping methods.
- To check robustness of topology of the produced Unicellular Genome Tree in general, and its correspondence with the “Archaea Tree” Hypothesis, in particular.
- To find out factors of lengthening or shortening of prokaryotic protein coding genes.
- ...

Genome Distances and Genome Clustering

- Clustering problems deal with partitioning of data items into groups of elements similar to each other. A similarity is used to find out group membership by means of a distance-like function that measures the distortion between the data points and reflects certain background information of the data's structure. The determination of such a function is an essential task in cluster analysis.
- The major difficulty arising in the distance function's selection problem links to a choice of the relevant data features involved in the function determination.
- The Information Bottleneck method of Tishby et al. suggested an approach based on the theory of information.

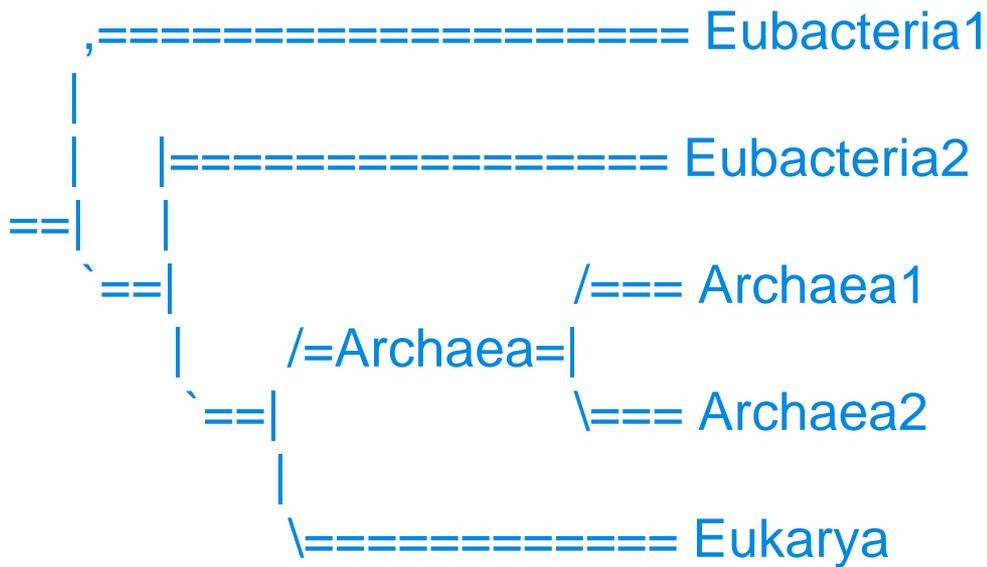
The information bottleneck method

- The information bottleneck method seeks for a compact (clustered) representation of X , which keeps as much information as possible on the applicable variable Y .
- Typically, X represents a variable which is intended to compress, and Y - a variable which we would like to predict.
- In our case, X is a set of genomes to be clustered, and Y is a property (for example, a length) of the genes presented in the COG.

W1

- The COG collection of 2003 consisted of 138,458 proteins, which form 4873 COGs and comprise 75% of the 185,505 (predicted) proteins encoded in 66 genomes of unicellular organisms.
- This set was used for construction of genome trees.

Results



The rooting of the tree produced by the **Sequential Clustering Information Bottleneck Algorithm**

Genome	Taxonomy	10Clusters	10Bool
Aae	B	1	1
Tma	B	1	5
Cje	P	1	1
Hpy	P	1	1
jHp	P	1	1
Afu	A	2	2
Hbs	A	2	2
Mac	A	2	2
Mja	A	2	2
Mka	A	2	2
Mth	A	2	2
Ape	A	3	2
Pab	A	3	2
Pho	A	3	2
Pya	A	3	2
Sso	A	3	2
Tac	A	3	2
Tvo	A	3	2
Pae	Pm	4	3
Atu	L	4	3
Bme	L	4	3
Ccr	L	4	3
Mlo	L	4	3
Sme	L	4	3
Rso	P	4	3
Bbu	Bsp	5	1
Cpn	Bch	5	7
Ctr	Bch	5	7
Tpa	Bsp	5	1
Buc	PE	5	1
Xfa	Px	5	3
Rco	Pr	5	7
Rpr	Pr	5	7
Bha	D	6	4
Bsu	D	6	4
Cac	D	6	5
Lin	D	6	4
Lla	D	6	4
Sau	D	6	4
Spn	D	6	4
Spy	D	6	4

Genome	Taxonomy	10Clusters	10Bool
Dra	Bde	7	5
Fnu	Bfu	7	5
Nos	Bcy	7	5
Syn	Bcy	7	5
Cgl	C	7	6
Mle	C	7	6
MtC	C	7	6
Mtu	C	7	6
Eco	PE	8	8
Ecs	PE	8	8
EcZ	PE	8	8
Hin	PP	8	8
Pmu	PP	8	8
Sty	PE	8	8
Vch	PV	8	8
Ype	PE	8	8
NmA	PN	8	8
Nme	PN	8	8
Ecu	K	9	9
Sce	K	9	9
Spo	K	9	9
Mge	D	10	10
Mpn	D	10	10
Mpu	D	10	10
Uur	D	10	10

Taxonomy		
Archaea	A	
Bacteria	B	<i>Fusobacteria</i> <i>Cyanobacteria</i> <i>Spirochaetes</i> <i>Chlamydiae</i> <i>Deinococcus</i>
Actinobacteria	C	
Gramplus	D	
Proteobacteria	P	<i>Enterobacteriales</i> <i>Pseudomonadales</i> <i>Pasteurellales</i> <i>Neisseriales</i> <i>Rickettsiales</i> <i>Vibrionales</i>
Alpha	L	
Eukaryotes	K	

Genome	Taxonomy	10Clusters	10Bool
Aae	B	1	1
Tma	B	1	5
Cje	P	1	1
Hpy	P	1	1
jHp	P	1	1
Afu	A	2	2
Hbs	A	2	2
Mac	A	2	2
Mja	A	2	2
Mka	A	2	2
Mth	A	2	2
Ape	A	3	2
Pab	A	3	2
Pho	A	3	2
Pya	A	3	2
Sso	A	3	2
Tac	A	3	2
Tvo	A	3	2
Pae	Pm	4	3
Atu	L	4	3
Bme	L	4	3
Ccr	L	4	3
Mlo	L	4	3
Snc	I	4	3

Genome	Taxonomy	10Clusters	10Bool
Dra	Bde	7	5
Fnu	Bfu	7	5
Nos	Bcy	7	5
Syn	Bcy	7	5
Cgl	C	7	6
Mle	C	7	6
MtC	C	7	6
Mtu	C	7	6
Eco	PE	8	8
Ecs	PE	8	8
EcZ	PE	8	8
Hin	PP	8	8
Pmu	PP	8	8
Sty	PE	8	8
Vch	PV	8	8
Ype	PE	8	8
NmA	PN	8	8
Nme	PN	8	8
Ecu	K	9	9
Sce	K	9	9
Spo	K	9	9
Mge	D	10	10
Mpn	D	10	10
Mpu	D	10	10
Sul	D	10	10

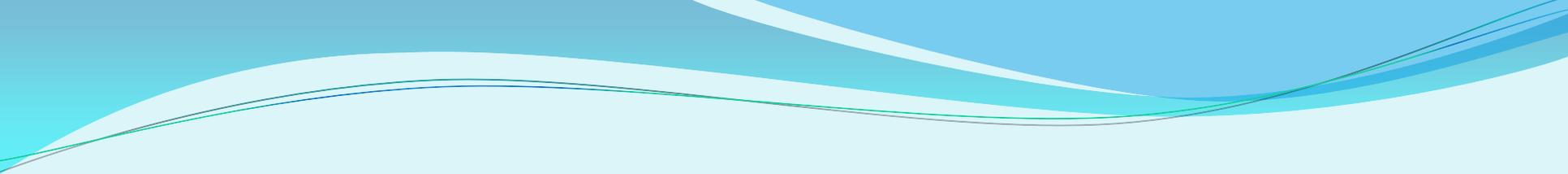
Sso	A	3	2
Tac	A	3	2
Tvo	A	3	2
Pae	Pm	4	3
Atu	L	4	3
Bme	L	4	3
Ccr	L	4	3
Mlo	L	4	3
Sme	L	4	3
Rso	P	4	3
Ebu	Bsp	5	1
Cpn	Bch	5	7
Ctr	Bch	5	7
Tpa	Bsp	5	1
Buc	PE	5	1
Xfa	Px	5	3
Rco	Pr	5	7
Rpr	Pr	5	7
Bha	D	6	4
Bsu	D	6	4
Cac	D	6	5
Lin	D	6	4
Lla	D	6	4
Sau	D	6	4
Spn	D	6	4
Spy	D	6	4

Nme	PN	8	8
Ecu	K	9	9
Sce	K	9	9
Spo	K	9	9
Mge	D	10	10
Mpn	D	10	10
Mpu	D	10	10
Uur	D	10	10

Taxonomy		
Archaea	A	
Bacteria	B	<i>Fusobacteria</i> <i>Cyanobacteria</i> <i>Spirochaetes</i> <i>Chlamydiae</i> <i>Deinococcus</i>
Actinobacteria	C	
Gramplus	D	
Proteobacteria	P	<i>Enterobacteriales</i> <i>Pseudomonadales</i> <i>Pasteurellales</i> <i>Neisseriales</i> <i>Rickettsiales</i> <i>Vibrionales</i>
Alpha	L	
Eukaryotes	K	

Table 1. Clustering to 10 groups based on gene lengths.

Cluster	Number of genomes	Taxonomy	Distribution of taxa
Cluster 1	5 <i>Proteobacteria</i>	Campylobacterales	3 of 3
		Aquificales	1 of 1
		Thermotogales	1 of 1
Cluster 2	6 <i>Archaea</i>	Euryarchaeota:	6 of 10
Cluster 3	7 <i>Archaea</i>	Euryarchaeota:	4 of 10
		Crenarchaeota	3 of 3
Cluster 4	7 <i>Proteobacteria</i>	Rhizobiales	4 of 4
		Burkholderiales	1 of 1
		Pseudomonadales	1 of 1
		Caulobacterales	1 of 1
Cluster 5	8 "Mix"	Spirochaetales	2 of 2
		Chlamydiales	2 of 2
		Rickettsiales	2 of 2
		Xanthomonadales	1 of 1
		Enterobacteriales <i>(Buchnera)</i>	1 of 6
Cluster 6	8 <i>Firmicutes</i> without <i>Mycoplasmatales</i>	Clostridiales, Bacillales, Lactobacillales	8 of 8
Cluster 7	8 "Mix"	<i>Cyanobacteria, Deinococcus-Thermus,</i> <i>Fusobacteria</i>	4 of 4
		<i>Actinobacteria</i>	4 of 4
Cluster 8	10 <i>Proteobacteria</i>	Enterobacteriales	5 of 6
		Pasteurellales	2 of 2
		Neisseriales	2 of 2
		Vibrionales	1 of 1
Cluster 9	3 Eukarya	<i>Ascomycota, Microsporidia</i>	3 of 3
Cluster 10	4 <i>Firmicutes</i> - <i>Mycoplasmatales</i>	Mycoplasmatales	4 of 4



Robust Classification of Prokaryotic Genomes

Computational Biology and Chemistry

Outline

- Background
- Goals
- Methods
- Results
- Summary
- Acknowledgement

Information Bottleneck

- The fundamental Information Bottleneck (IB) approach has been proposed by Tishby et al. having a declared purpose to avoid the arbitrary choice of a distance measure.
- In the framework of this general methodology given the empirical joint distribution of two random variables $P(X, Y)$, a compact representation of X is being constructed persevering as much information as possible about the relevant variable Y .
- In the previous work Top-Down variant was used, while in this study Bottom-Up (agglomerative) algorithm is used.

IB and problem definition

- A great success of the IB-approach motivates its application in many other problems.
- The key issue is here a representation of the considered objects by means of conditional probability distributions.
- In this study, the objects are prokaryotic genomes, the approach is the Bag-of-Tokens method, and the genomes are presented by lengths of protein coding genes.
- The length values were obtained using the database of Clusters of Orthologous Groups of proteins (COGs).
- Here may be asked three questions about:
 - **Relevance, suitability, robustness.**

Relevance

- The method is closely related to the group of methods based on the presence and absence of genes; in addition, it uses the information related to the lengths of genes.
- Several approaches pertaining to the method presented in the current study have been proposed.
- These approaches may be called “determination of genome phylogeny based on gene content”.
- By the way, we have already presented classifications produced for the same small group of genomes and based either on gene content or on genomic data related to lengths of homologous proteins were compared.
- It was showed that, as expected, the dendrogram based on usage larger evolutionary information presents a more taxonomically convincing genomic tree.
- The gene content certainly carries very strong phylogenetic component so even presence of a horizontally transferred genes as a part of considered gene repertoire does not corrupt produced classifications.

Suitability

- Our claim of suitability of gene lengths of a COG to be a truthful variable Y in the IB method is based on the observation that the most common ways of changing protein length during prokaryotic evolution are consequences of insertions and deletions (indels).
- So, taxonomically and evolutionary close organisms have very similar gene presence/absence profiles and lengths of orthologous genes in such organisms are very similar as well.

Primary goals

- In introducing a novel method of genome classification presented as a genome tree construction, we have had several primary goals.
 - **First**, to propose a fast method that in principle allows construction of genome tree from as large as possible subset of all available prokaryotic genomes. Naturally, we must show that the method is fast and reliable.
 - **Second**, to show the **robustness** of the proposed algorithm. The intention is to show that for a chosen genome dataset, tree structures obtained using different subsets of gene families are sufficiently similar.
 - **Third**, to demonstrate that a produced dendrogram, which presents classification of a selected small group of genomes, looks a lot like a phylogenetic tree.
 - **Fourth**, to fix the parameters of the method, basing on results of a few empirical case studies.

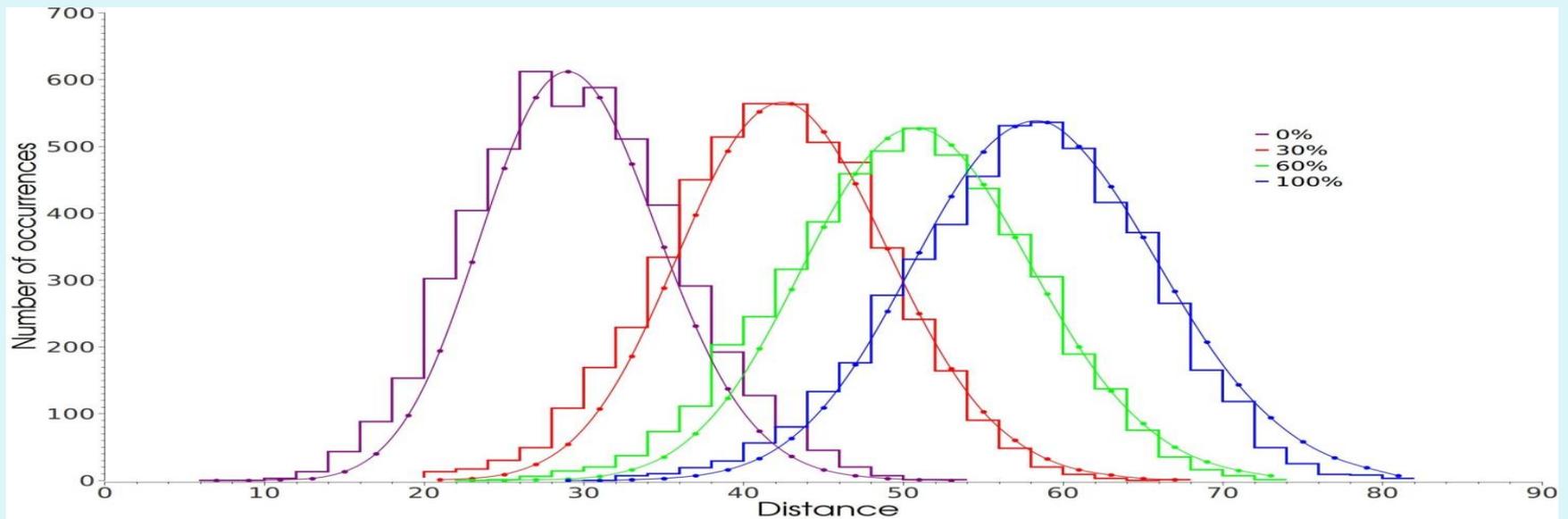
Tree robustness estimation

- Assessing the robustness of the phylogenetic trees remains contentious. The methods commonly used to assess support for cladograms include nonparametric bootstrapping and jackknifing.
- In case of our data (a matrix [genome, COG]), bootstrapping method means that while dimensions of a matrix and a number of non-empty elements in it are unchanged a certain randomly chosen fraction of columns (COGs) are reshuffled.
- In case of our data, jackknifing method means that a certain randomly chosen fraction of columns (COGs) are deleted from original input data. It means that instead of usage the full set of COGs only a subset of gene families is used.

Randomizing (bootstrapping)

- FOR ($Z = 0, 73, 346, 577$) DO {
 - do 100 times {
 - Randomly select 577 out of 888 COGs. It gives a sparse supermatrix $M [60, 577]$.
 - Randomly reshuffle Z chosen columns of the abovementioned matrix M .
 - Generate a tree for the obtained supermatrix M .
 - } end do
 - For each pair of constructed trees calculate a distance between these trees using partition metric.
 - Draw a histogram of distance distribution.
- } END FOR

Bootstrapping: randomizing and robustness are anticorrelated events



Randomizing

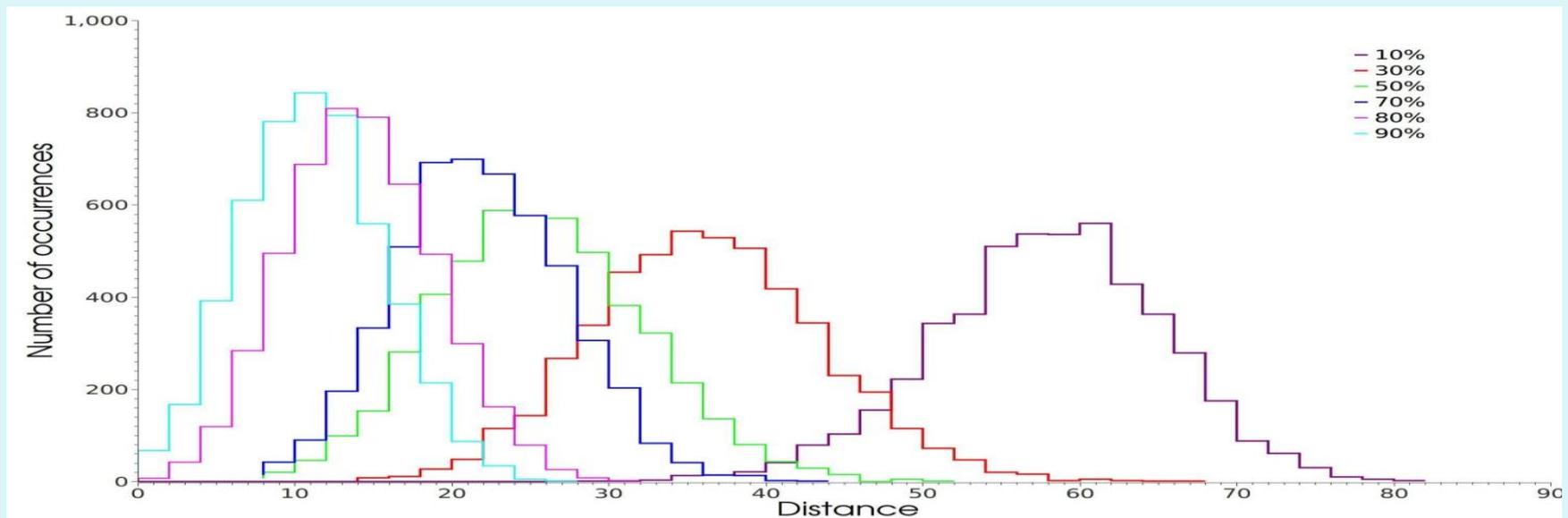
- First of all, the curves are Poisson-like $P(x = k)$
 $= e^{-\lambda} \frac{\lambda^k}{k!}$ as expected [15].
- The furthest to the left curve shows the distance distribution between the trees based on the original 65% datasets corresponding to the Poisson-like curve with the parameter $\lambda+6=30$, and the right one - the distance distribution for fully randomized data - to the Poisson-like curve with the parameter $\lambda+30=59$.
- Actually, this Figure justifies our expectations that randomization increases distances between corresponding trees, and the distance distributions are nearly Poisson.

Jackknifing

- FOR ($B=10\%$; $B < 100\%$; $B+10\%$) DO {
 - $Y = 888 * B$;
 - do 100 times {
 - Select random Y columns out of 888, which gives a sparse supermatrix $M [60, Y]$.
 - Generate a tree for the obtained supermatrix M .
 - } end do
 - For each pair of constructed trees calculate a distance between these trees using partition metric.
 - Draw a histogram of distance distribution.
- }

Jackknifing:

gene-families' subset size and tree robustness are correlated events – more is better



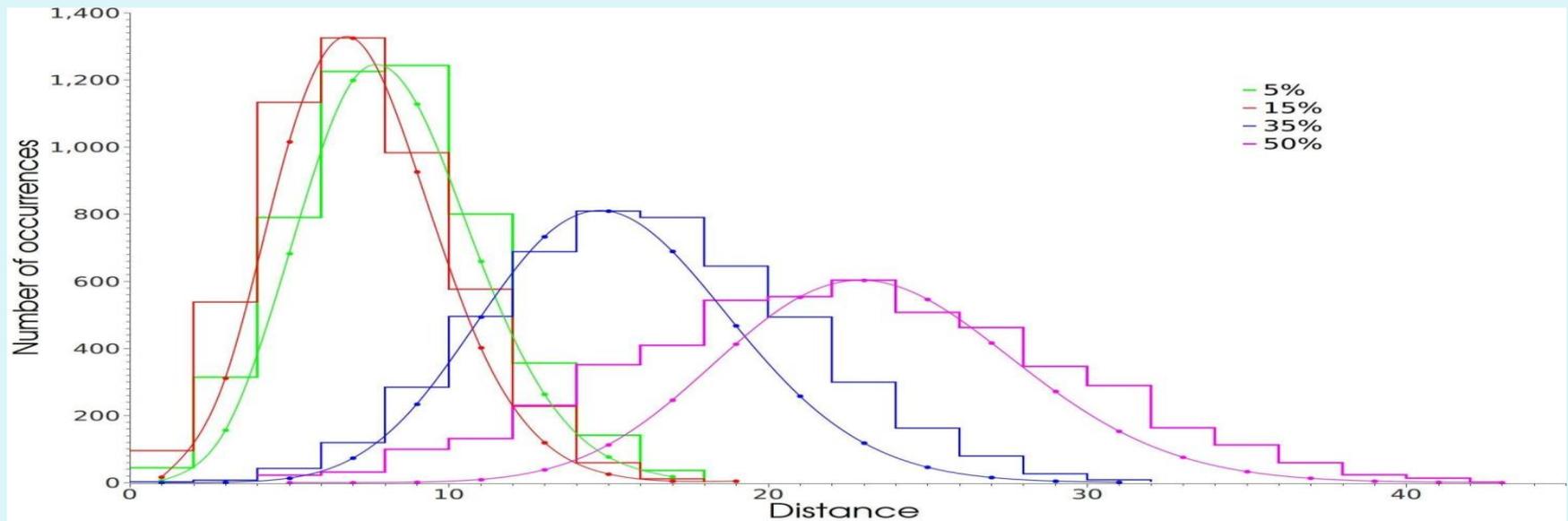
Analysis of the histograms may be concluded in three words "more is better".

Filtering

Here, we investigate whether it is worthwhile to consider as many gene families as possible or, it is better to filter out COGs with a small number of proteins in it. Higher values of a separation-out parameter A lead to stronger selection, which means that only those COGs that have broader representation of genomes remain for further consideration.

- FOR ($A = 5, 15, 35, 50$) DO {
 - Preprocess the complete COGs dataset with the separation-out parameter A . Obtain X COGs containing more than $A\%$ of the maximal COG size. $Y = X * 80\%$
 - do 100 times {
 - Randomly select Y COGs out of X . It gives a sparse supermatrix $M [60, Y]$.
 - Generate a tree for the obtained supermatrix M .
 - } end do
 - For each pair of constructed trees calculate a distance between these trees using partition metric.
 - Draw a histogram of distance distribution.
- }

Filtering: relationship between a gene-families' separation – preprocessing parameter and tree robustness



The best distribution (the far left curve) is related to the value 15%, while both 5% and 35% show distributions with worse parameters.

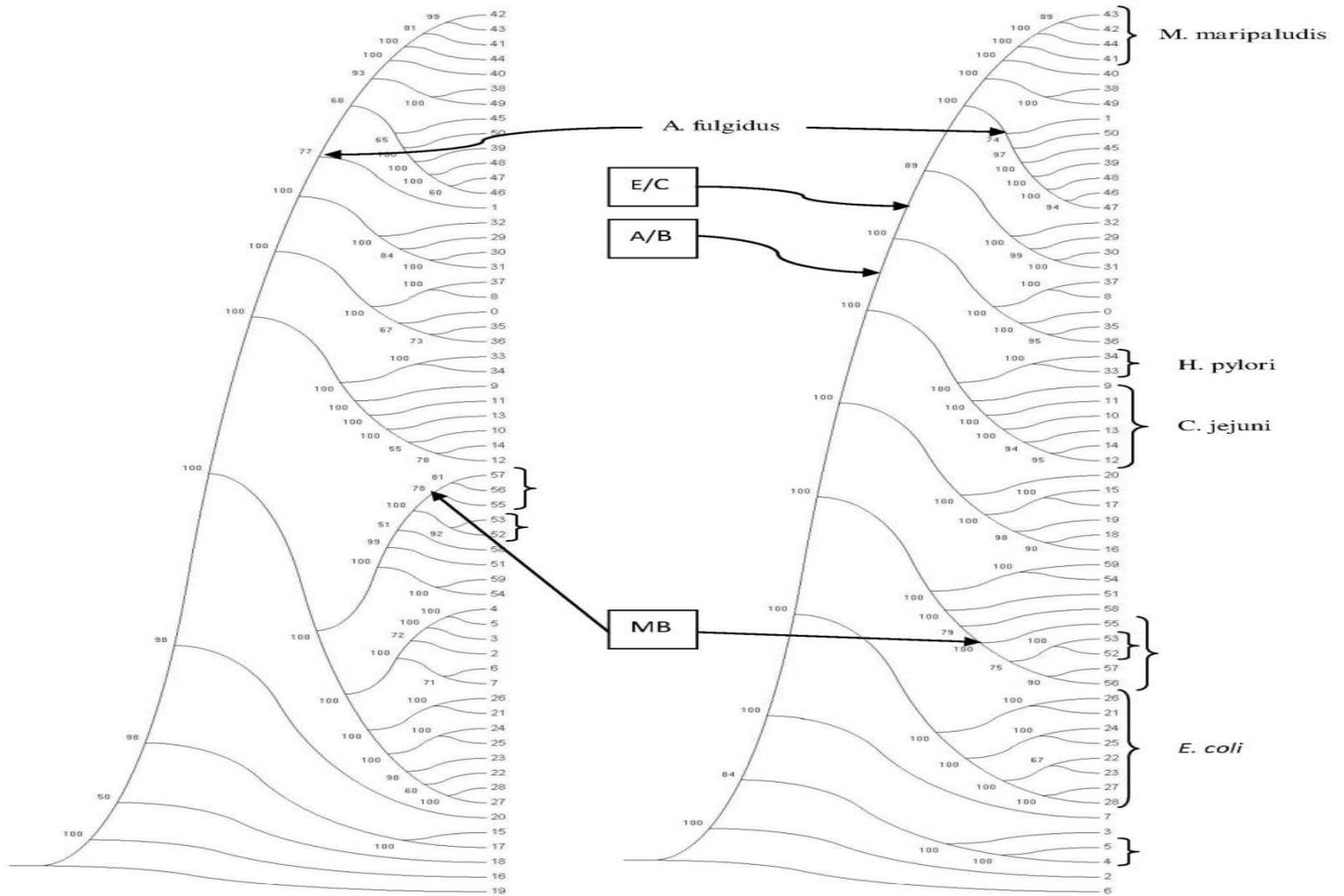
Filtering

The best distribution (the far left curve at Fig. 3) is related to the value 15%, while both 5% and 35% show distributions with worse parameters. It means that a value parameter of 5%, which leads to the utilization of very small COGs, is too small to be optimal, while the value of 15% appears to be close to optimal value. Therefore, we arrived empirically to the choice of the values of $A = 15\%$ and $B = 80\%$ as optimal parameters for further investigations.

- FOR ($A = 5, 15, 35, 50$) DO {
 - Preprocess the complete COGs dataset with the separation-out parameter A . Obtain X COGs containing more than $A\%$ of the maximal COG size. $Y = X * 80\%$
 - do 100 times {
 - Randomly select Y COGs out of X . It gives a sparse supermatrix M [$60, Y$].
 - Generate a tree for the obtained supermatrix M .
 - } end do
 - For each pair of constructed trees calculate a distance between these trees using partition metric.
 - Draw a histogram of distance distribution.
- }

Consensus trees

- Consensus trees are obtained by using a CONSENSE software from the Phylip package.
- A subset size B was chosen to be equal to 80%.
- The bootstrapping parameter C is equal to zero – no randomization.
- Two consensus trees for $A=15\%$ and $A=35\%$ are constructed.
- Taking $A=15\%$ get $X = 2088$ and $Y = 1670$. Taking $A=35\%$ get $X = 888$ and $Y = 710$.
- The general procedure presented in Schema 1 for this section is transformed into:
 - FOR ($A = 15, 35$) DO
 - {
 - Select only those X COGs which contain more than $A\%$ of the maximal COG size.
 - do 100 times {
 - Select randomly Y COGs out of X . It gives a sparse supermatrix M [$60, Y$].
 - Generate a tree for the obtained supermatrix M .
 - }
 - Construct a consensus tree for obtained 100 trees.
 - }
- The output is a collection of 100 near optimal and very similar trees.
- By far, the most common methods to build consensus trees.
- Their aim is simply to build a single tree displaying the frequencies of splits (or of clades) seen in the collection. These methods summarize all common properties to these trees, presenting the branch-labeled consensus tree.



Consensus trees

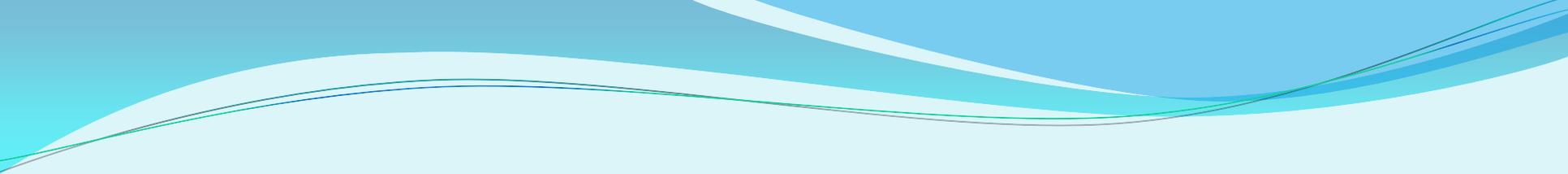
- Comparing consensus trees based on an 80%-subset of COGs, applying these two jackknifing rates (Fig. 4a for 35%-jackknifing, and Fig. 4b for 15%-jackknifing), we can see that while the topologies are quite similar, the branch labels are larger in Fig. 4b. It can probably be interpreted as follows: There is some essential information that is contained only in small COGs, e.g., information specific for some subgroup of organisms.
- Let us verify phylogenetic reasonableness of the tree.
- The representatives of both prokaryotic Kingdoms: Eubacteria and Archaea – are clustered separately. In other words, Archaeal organisms (genomes 0, 1, 8, 29-32, 35-50) form a monophyletic group (see A/B marked arrow in Fig. 4b).
- Euryarchaeota and Crenarchaeota form monophyletic groups (see E/C marked arrow).
- The representatives of different strains of the same bacteria are placed together: 4 strains of *M. maripaludis* (genomes 41-44), 5 strains of *C. jejuni* (genomes 10-14), 2 strains of *M. bovis* (genomes 52-53), 8 strains of *E. coli* (genomes 21-28), 2 strains of *H. pylori* (genomes 33-34), and 2 strains of *B. pseudomallei* (genomes 4 and 5).
- Unfortunately, 3 strains of *M. tuberculosis* (genomes 55-57) do not form a monophyletic group; however, all *Mycobacterium* do form a monophyletic group.
- Interestingly, we can observe in Fig. 4a that *M. tuberculosis* (genomes 55-57) form a monophyletic group - see MB marked arrow.

summarizing notes

- In this study, we have conducted extensive experiments to validate the performance of bootstrapping and jackknifing to estimate how robust the trees produced by proposed methodology are.
- To summarize, we are confident in our proposal to perform classification of prokaryotes, which results in dendrograms strongly resembling prokaryotic phylogenetic trees, using the fast and reliable method described in this manuscript with the parameter values equal to 15% of the maximal COG size for the preprocessing parameter and equal to 80% for the jackknifing parameter.

Collaboration

- This study was performed in collaboration with
- Dr. Zeev Volkovich
 - And
- Dr. Katerina Korenblat
- from ORT Braude Academic College.



Revealing Factors Affecting Gene Length

**Rating of Prokaryotic Genomes
According to their Gene Lengths**

Acknowledgement to my co-authors

- **Bilal Salih^{1, 2 +}, Irit Cohen^{1, 3 +}, Tatiana Tatarinova^{4*}**
- ¹ Department of Evolutionary and Environmental Biology and Institute of Evolution, the University of Haifa, Israel
- ² Department of Computer Science, the University of Haifa, Israel
- ³ The Tauber Bioinformatics Research Center at the University of Haifa
- ⁴ Children's Hospital Los Angeles, University of Southern California, Los Angeles, California, USA

Objectives

- To better understand the interaction between the environment and bacteria, whether in a human host or other ecosystem, one must understand the laws governing bacterial evolution and adaptation.
- Understanding mechanisms of adaptation of bacterial species to their environment will greatly help this process.
- For example, it is essential to understand how a change in pH or external temperature affects the bacterial genome and especially its coding sequences.

Objectives

- Unfortunately, the evolution of bacterial coding sequences remains unclear. Orthologous proteins may drastically differ in both codon usage and length across species.
- When a gene length changes, a protein may acquire a new or lose an existing function, hence, changing the entire ecosystem. However, it has been hard to predict the effect of a changing environment on gene length.

Combinatorial optimization

- The approach of using seriation as a combinatorial optimization problem is new in the data mining field; and, to the best of our knowledge, is completely novel in application of data mining techniques to evolutionary molecular biology.
- A basic problem in data analysis, called seriation or sometimes sequencing, is to arrange all objects in a set in a linear order given available data and some loss or merit function in order to reveal structural information. Together with cluster analysis and variable selection, seriation is an important problem in the field of combinatorial data analysis.

Revealing Factors Affecting Gene Length

Review of applicable methods of combinatorial optimization (Bioinformatics and Biology Insights 6: 317–327)

Alexander Bolshoy¹ and Tatiana Tatarinova²

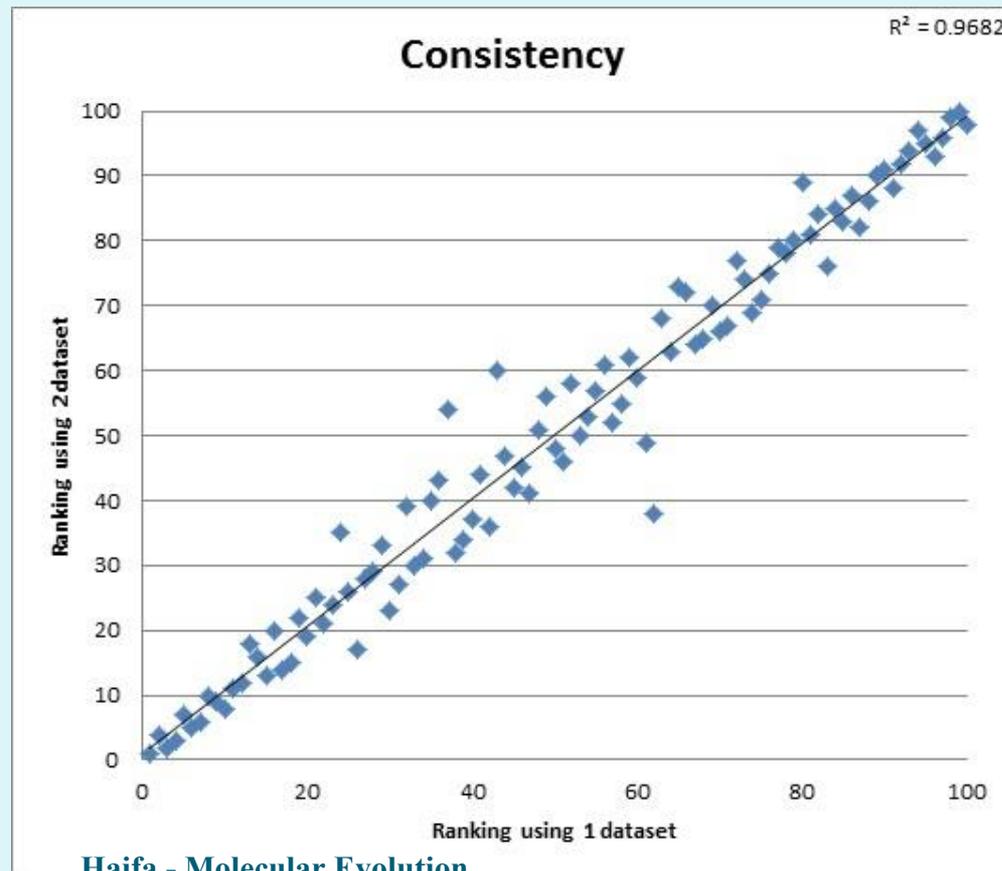
1 University of Haifa, Israel

2 University of Glamorgan, Wales, UK

Comparison of rankings produced by two different incomplete subsets of COGs

- We performed a random selection of 1050 COGs twice (overlap was 777 COGs).
- For the two subsets of COGs the resulting rankings are significantly correlated (Figure 3), Kendall tau correlation coefficient is 0.908 (2-sided p-value ≈ 0).
- Lowest and highest ranks agree the most, while genomes from the middle portion of the ordering show the most deviation.

Comparison of rankings produced by two different incomplete subsets of COGs

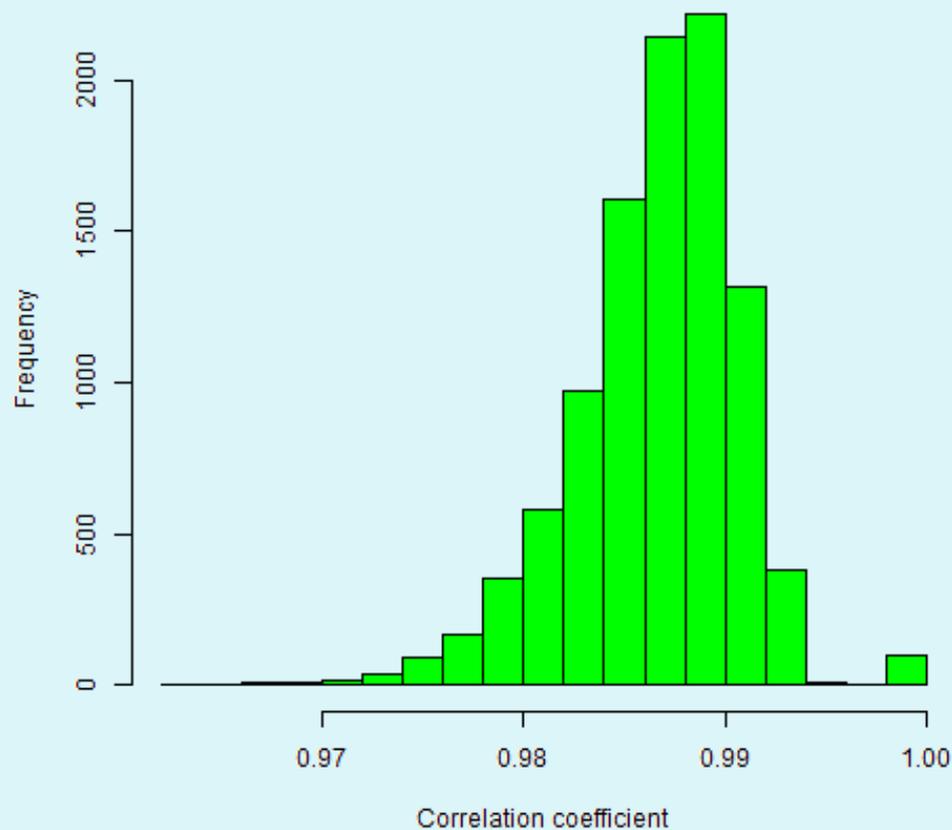


Results

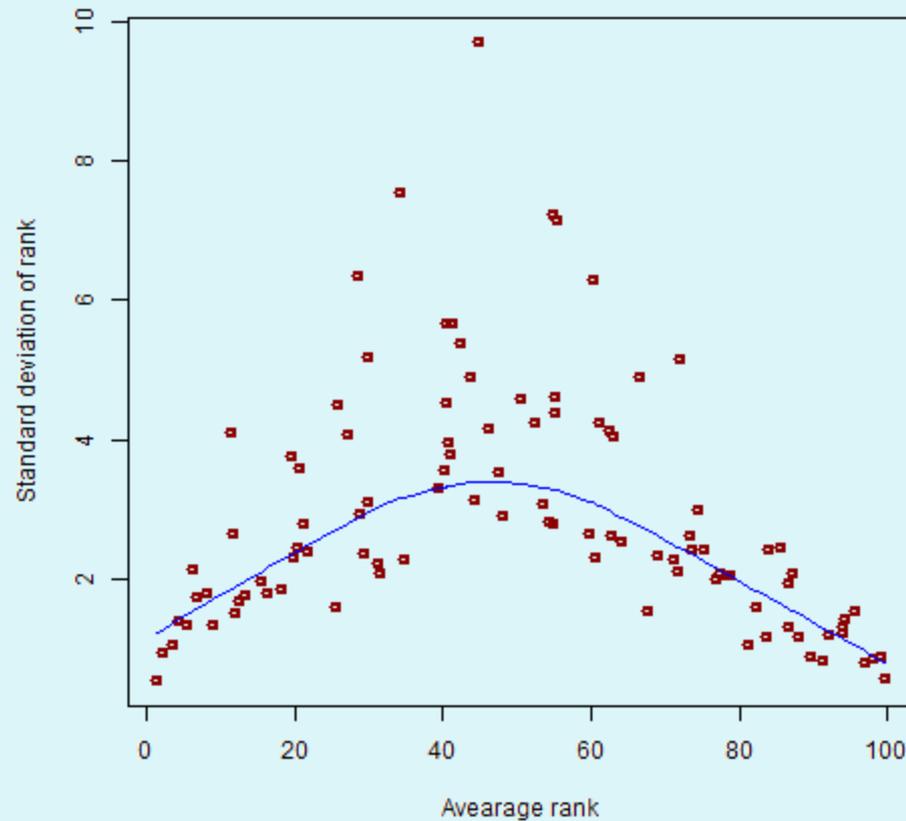
- First, we randomly selected 100 archaeal and bacterial genomes and selected only those COGs that were present in at least 40 genomes.
- We got a list of 1409 COGs.
- From this list, we randomly chose 1000 COGs and applied the optimization procedure.
- We repeated this selection and optimization process 100 times and compared rankings between the runs.
- The code was implemented using MPI R package.
- Due to the computational complexity of the problem we used parallel competition on HPC Wales cluster.
- Mean correlation coefficient between rankings is 0.98.

Correlations between rankings

Histogram of correlation coefficients between rankings



Extremes are more conserved



The resulting ranking of the genomes

- Taxonomic groups appear to be tightly clustered within the ordering. For example, majority of the Archaeal genomes are placed on the top of the ranking table
- Our calculations performed on other genome subsets (unpublished data) persuade us to discuss at this stage only the most stable groups: the top (ranked 1-16) and the bottom (ranked 85-100).
- Among the top sixteen 13 hyperthermophiles are the clear majority. There are both Archaea and Eubacteria in this group.
- In addition to hyperthermophiles two campylobacters and one helicobacter accomplish this group. There are no other campylobacters or helicobacters in R₁. The two species of Archaea that are not hyperthermophiles are placed at ranks 20 and 50. The opposite end of the spectrum is occupied by Actinobacteria

Ranking of Prokaryotic Genomes Based on Maximization of Sortedness of Gene Lengths

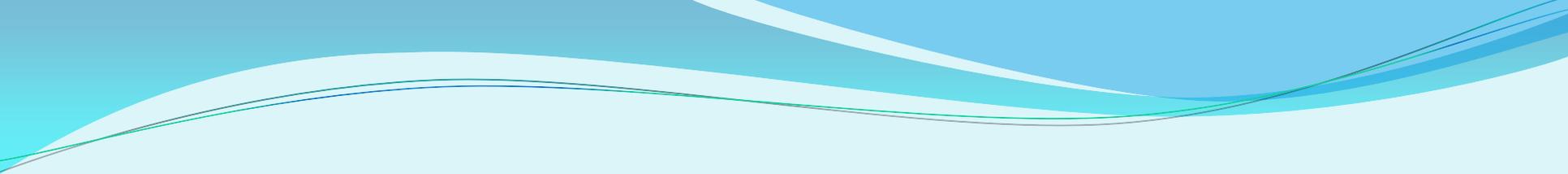
Journal of Data Mining in Genomics - 2014

Salih B. ⁺, Cohen I. ⁺, Tatarinova T. and Bolshoy A.

University of Haifa, Israel

University of South California





**Lengths of orthologous
prokaryotic proteins are
affected by evolutionary factors**

Results

- The Average ranking (A), the Simple Additive Ranking (SAR), and the Bubble sort ranking (B sort) methods were applied both to the non-filtered input (matrix of size 100x5664, Figure 1) and to a smaller matrix (filtered version: excluding columns that contain more than 65% null values given a matrix of size 100x1455, Figure 2). Simulated Annealing procedure (SAP) was applied only to the smaller matrix because of computational complexity. Each procedure produced a certain order of rows in the matrix, a ranking vector X . Calculating of the Kendall tau correlation coefficients between the ranking vector X and each column of the matrix yielded distribution of correlations between the global ranking and individual COGs. These distributions are shown in Figures 1 and 2.

Figure 1

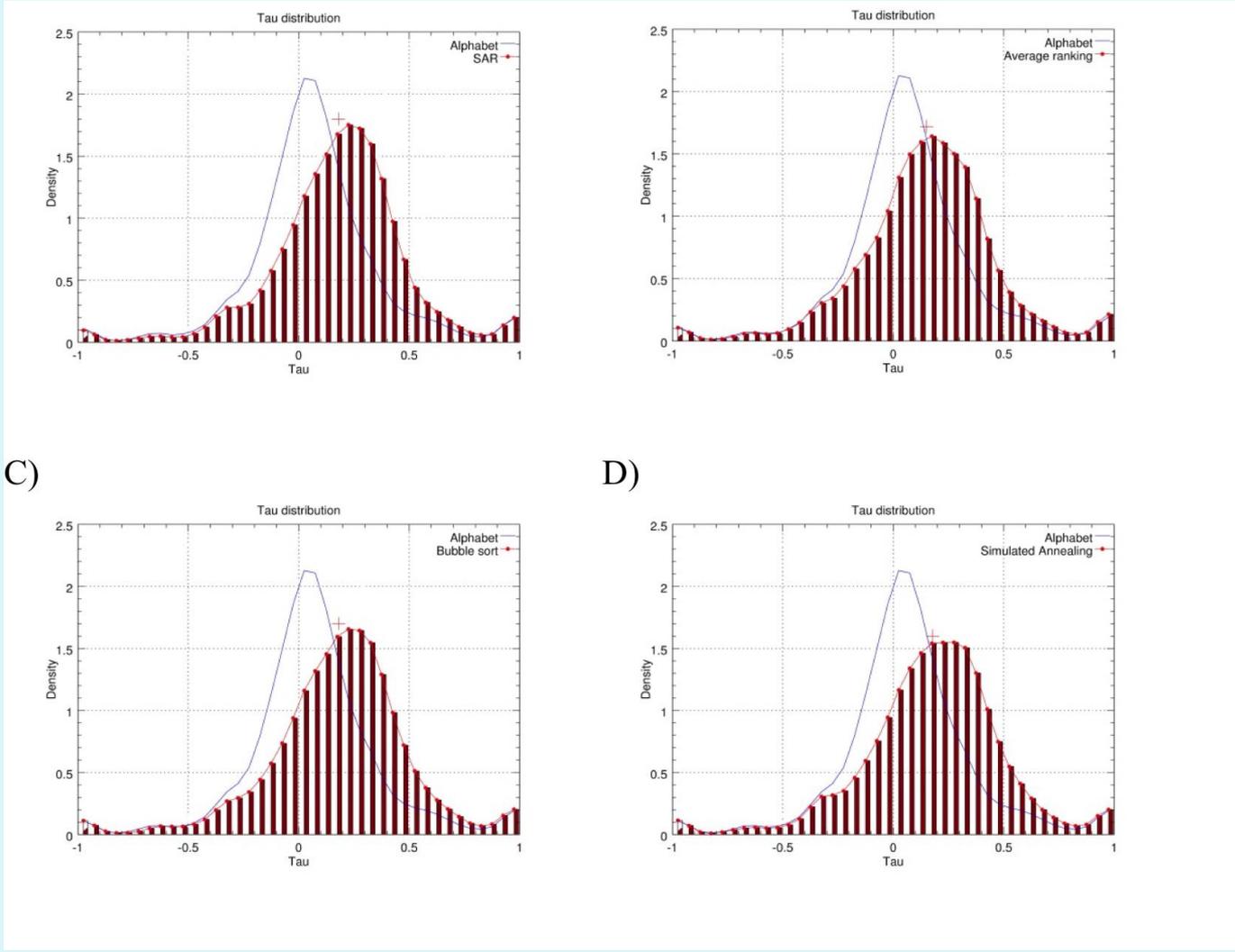
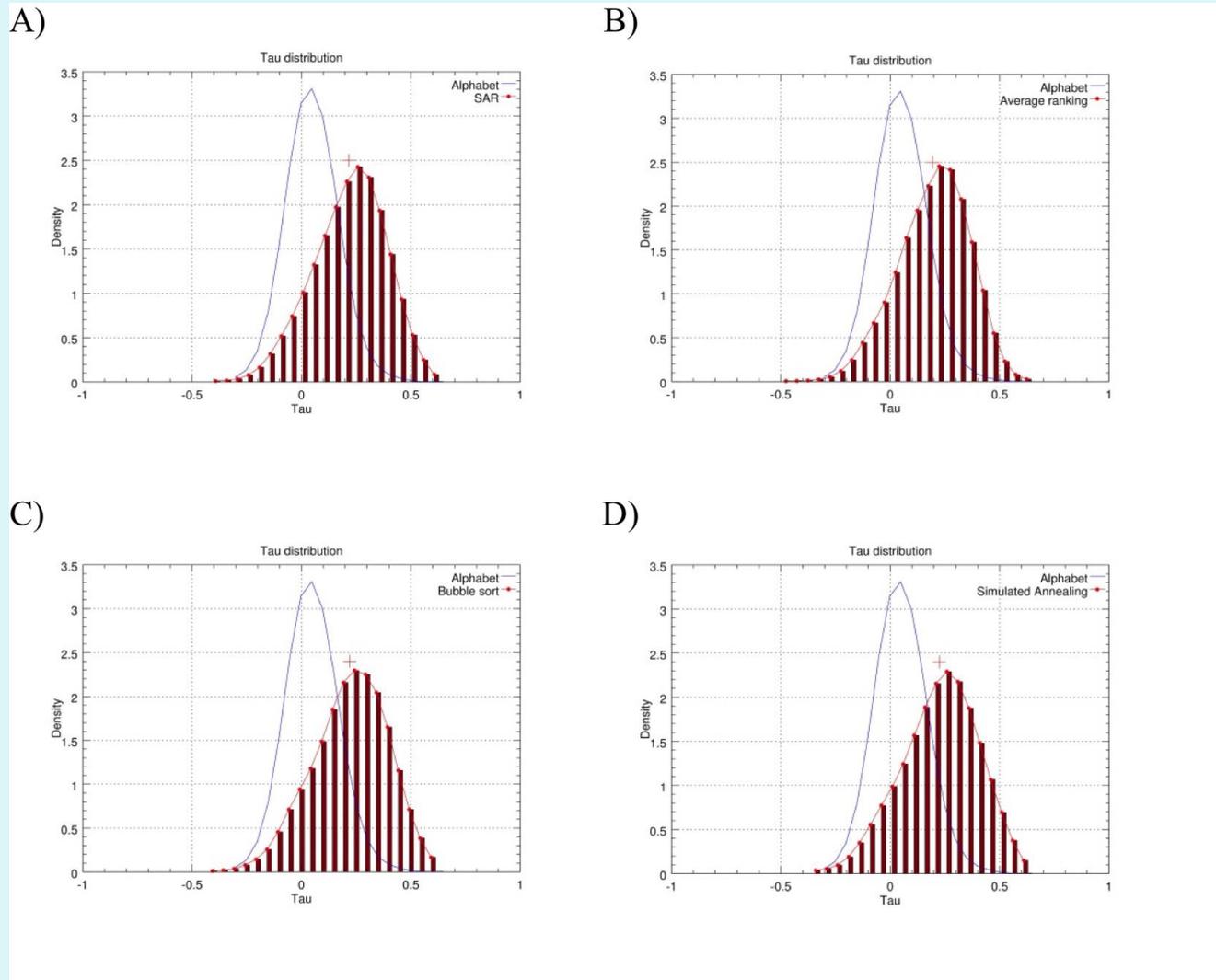


Figure 2



Pairwise comparisons of orderings

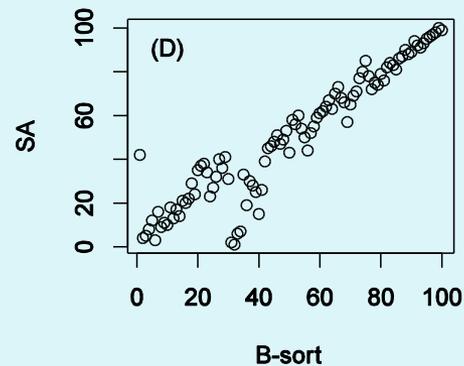
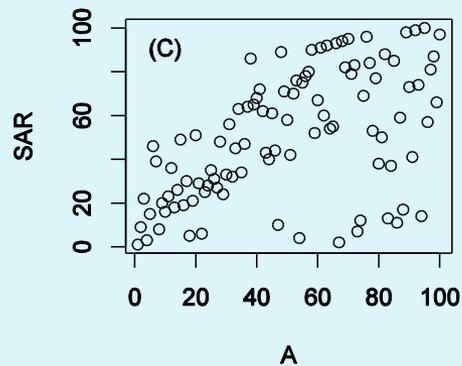
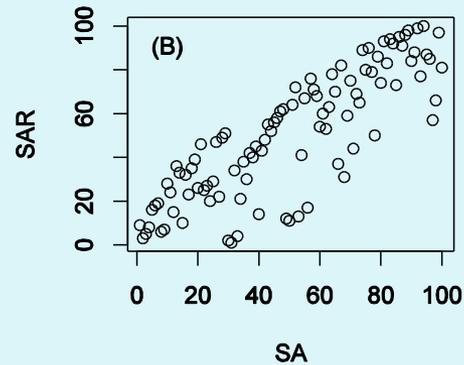
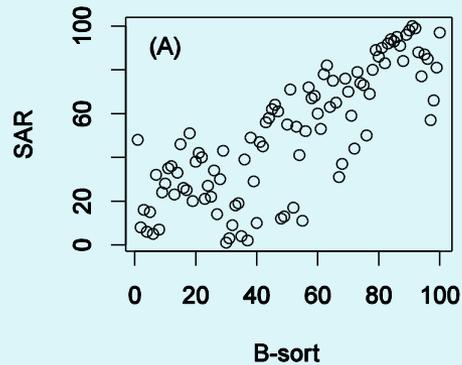


Table 4

Table 4 Goodness of fit of rankings measured by Kemeny measure

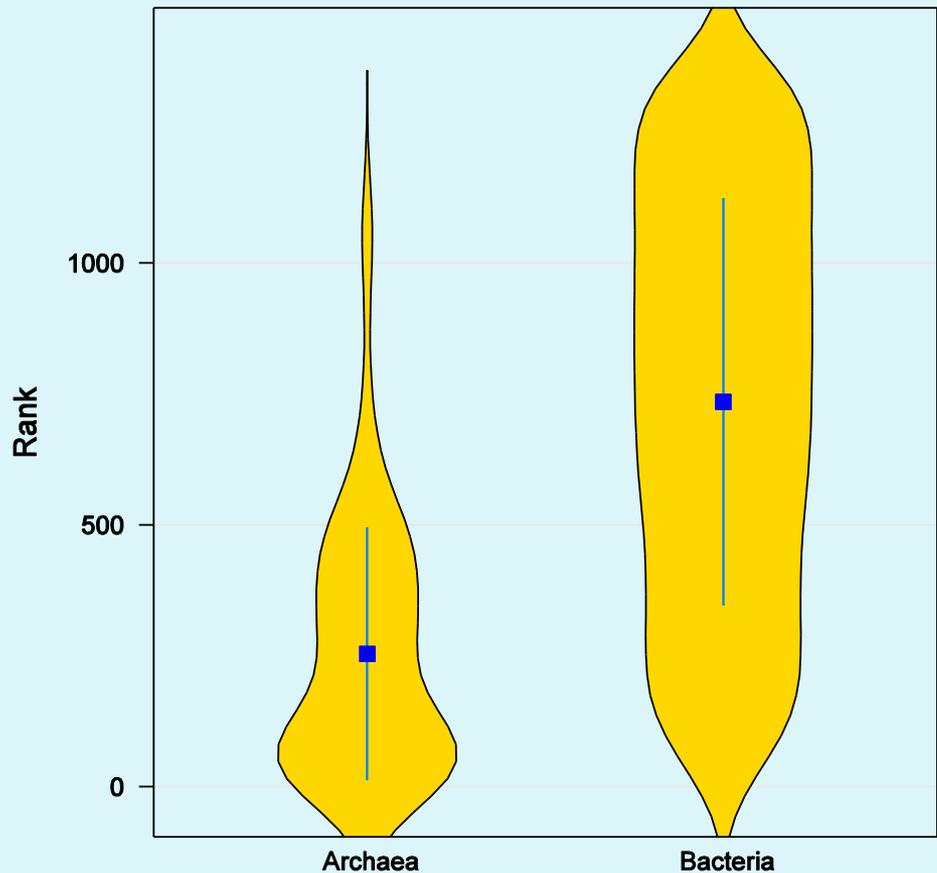
Ranking Method	Tau sortedness	
	Complete Matrix (100x5664)	Filtered Matrix (100x1455)
Average Ranking	0.15114	0.19459
Simple Additive Ranking	0.17990	0.21736
LOPI (Bubble Sorting)	0.18048	0.22057
Simulated Annealing (applied to the filtered matrix)	0.17841	0.22381

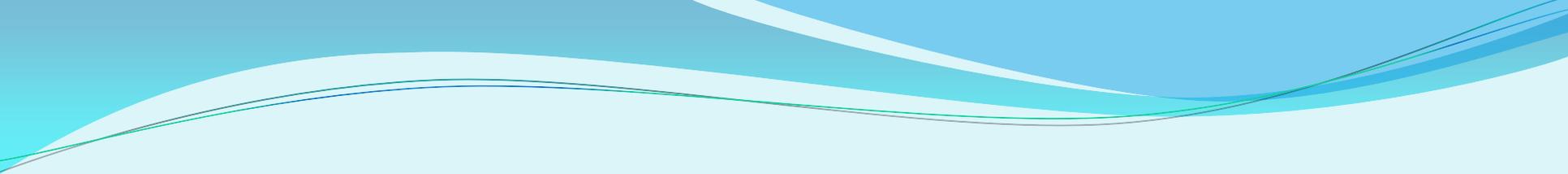
Table 5

Table 5 Pairwise Kendall coefficients of correlation between different rankings

	SAR-rank	A-rank	B-rank	SAR-rank (filtered)	A-rank (filtered)	B-rank (filtered)	SA-rank (filtered)
SAR-rank	1						
A-rank	0.47556	1					
B-rank	0.74949	0.47152	1				
SAR-rank (filtered)	0.87475	0.52081	0.80121	1			
A-rank (filtered)	0.5899	0.77253	0.56566	0.64242	1		
B-rank (filtered)	0.7503	0.48848	0.95313	0.81333	0.58424	1	
SA-rank (filtered)	0.73293	0.53333	0.81212	0.80646	0.63313	0.84444	1

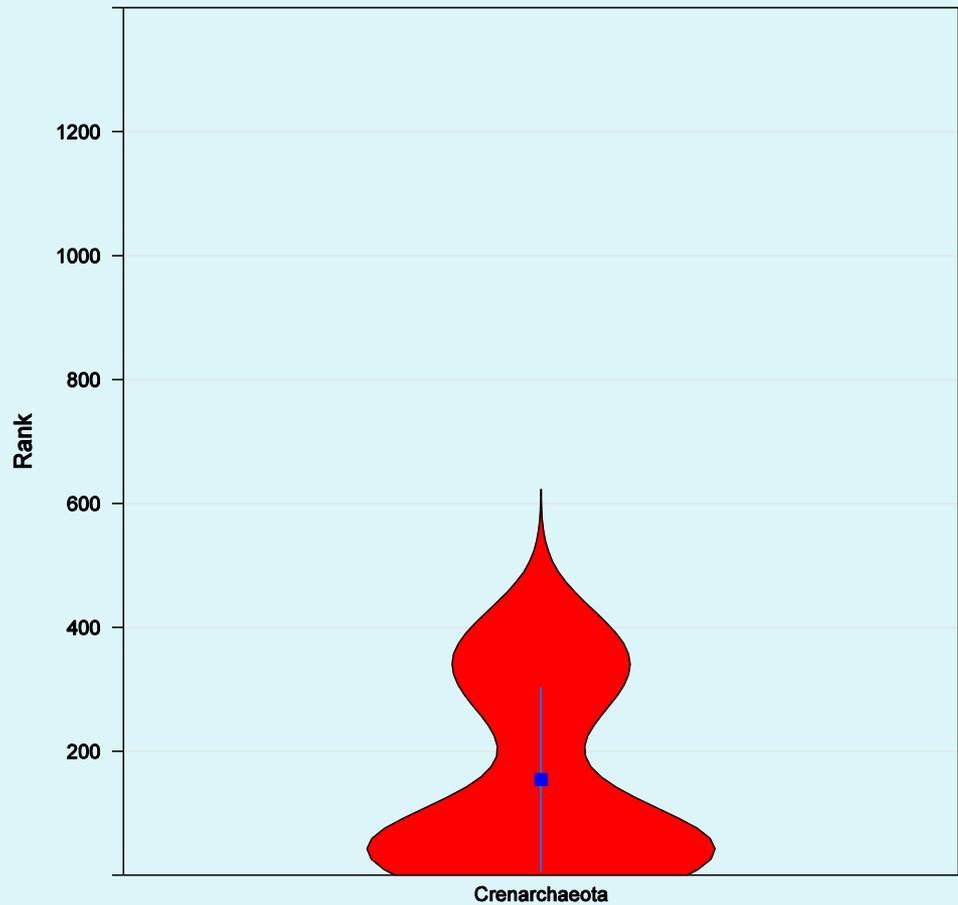
Violin plots of Bubble sort ranks of Archaea and Bacteria. Average rank of 1276 *Bacterial* genomes is 735 and average rank of 114 *Archaeal* genomes is 254.



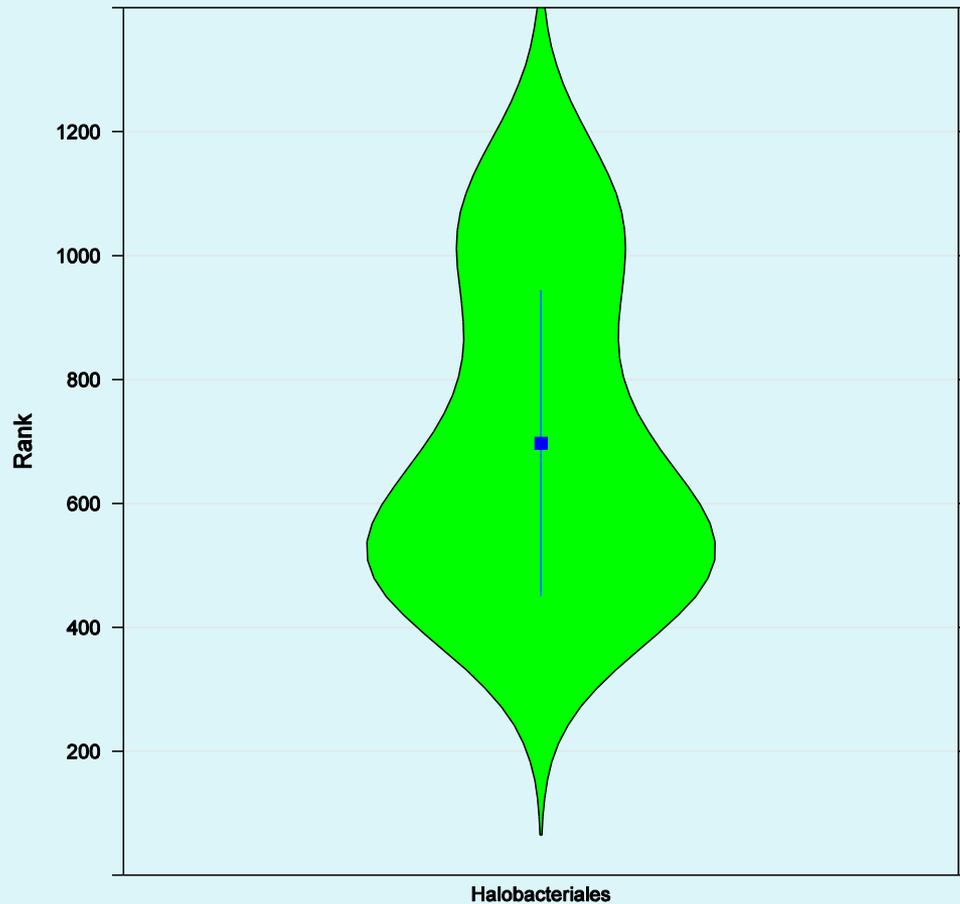


Violin plots of length distributions for six groups of prokaryotic genomes. Ranks calculated by applying Bubble sort to the filtered set of 1390 prokaryotic genomes.

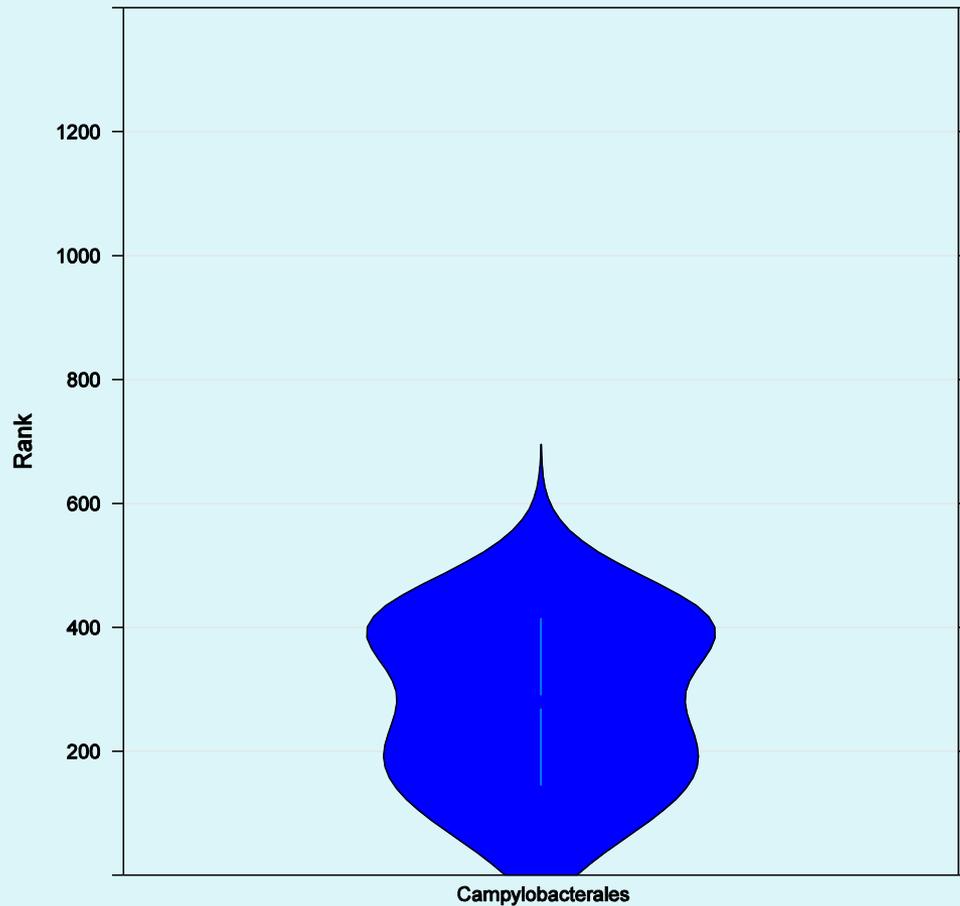
Crenarchaeota



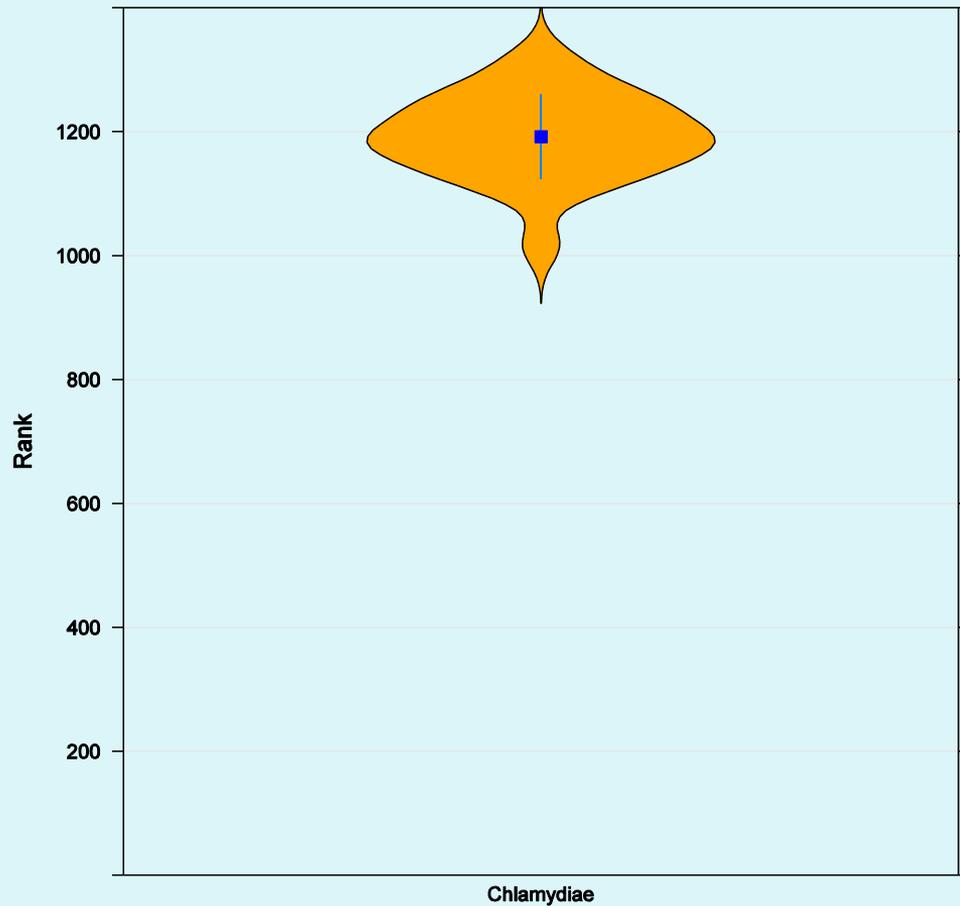
Halobacteriales



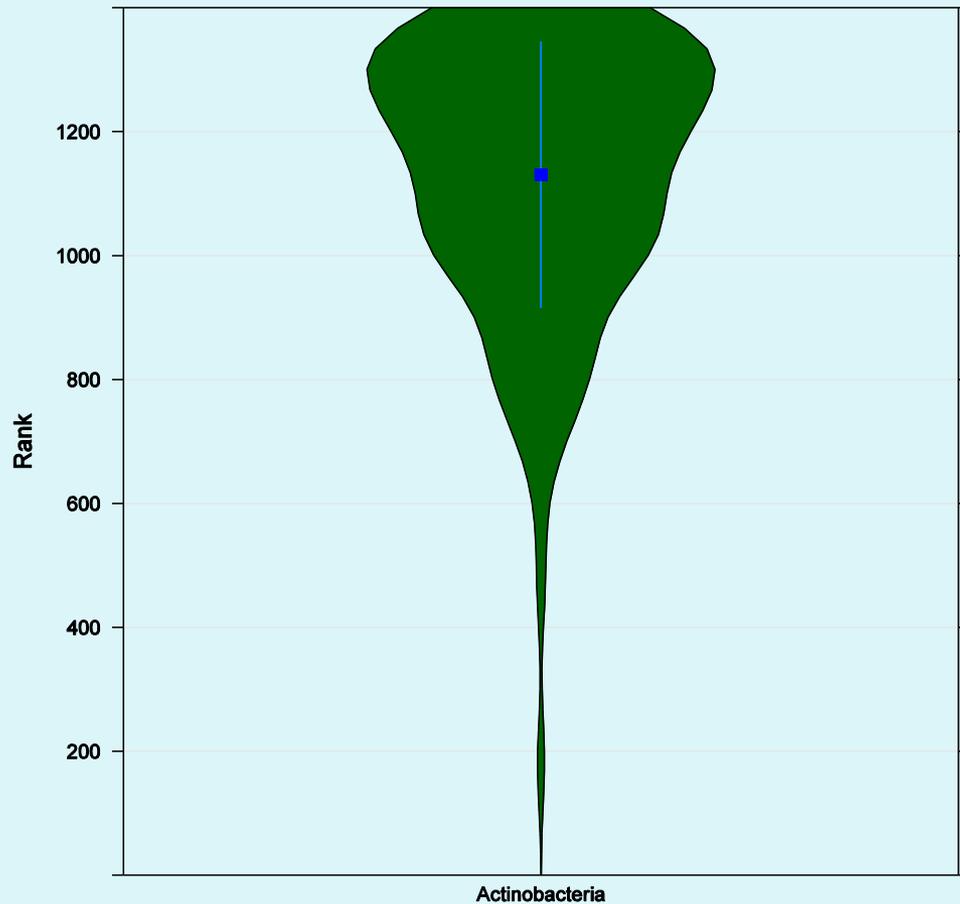
Campylobacterales



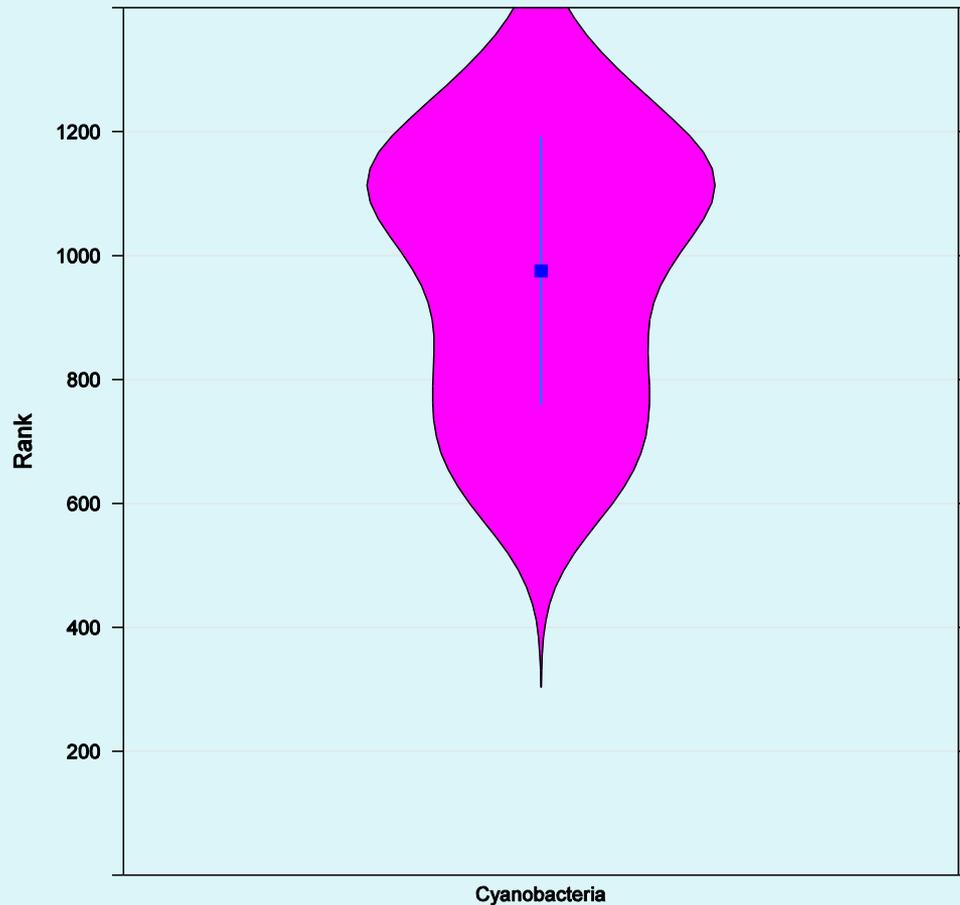
Chlamydiae



Actinobacteria



Cyanobacteria



Conclusions

- We have presented four methods of genome ranking and compared their performance using a dataset of 100 genomes randomly selected from the entire NCBI collection of *Eubacterial* and *Archaeal* genomes.
- We have demonstrated that all four methods produce consistent results and that Bubble Sort and Simulated Annealing have the best ranking.
- Given computational advantages of Bubble Sort, it appears to be the most appropriate method for the task of genome ordering, since Bubble Sort provides both good accuracy and speed.

Conclusions

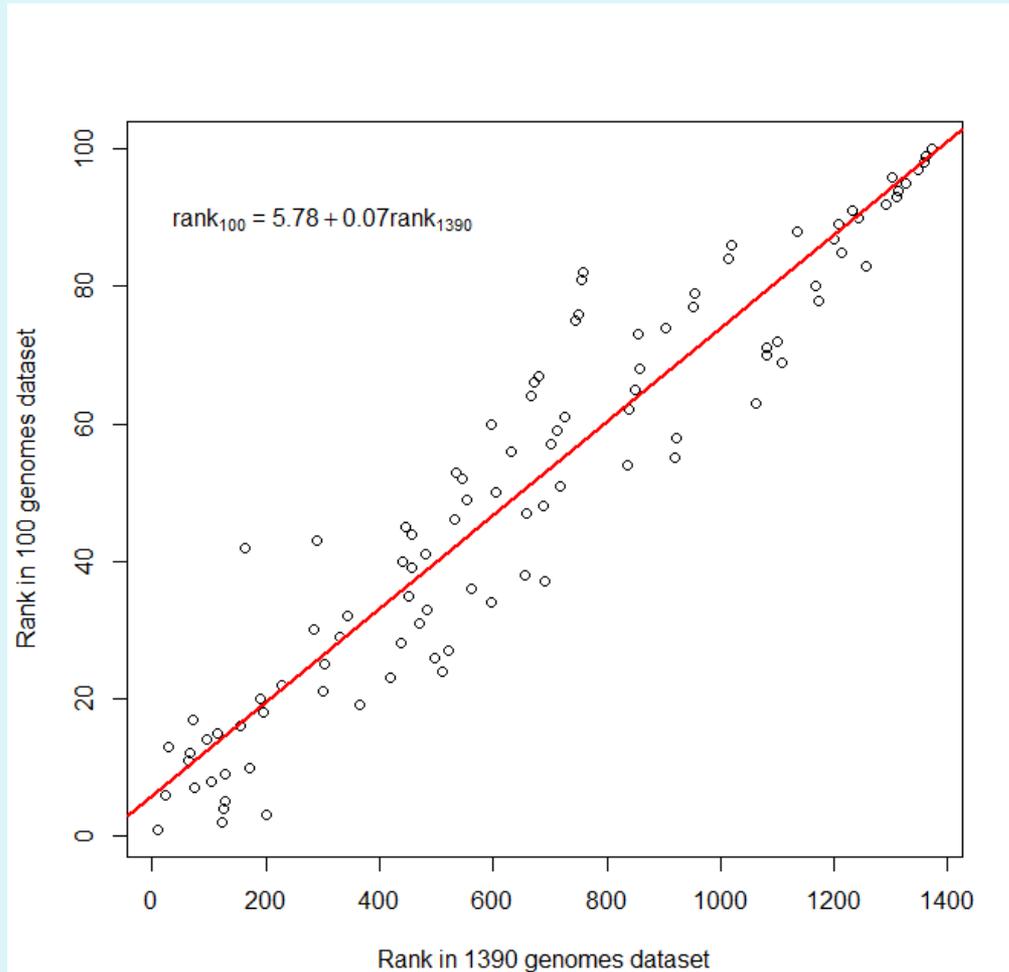
- When Bubble Sort was applied to the set of 1390 prokaryotic genomes, it revealed several interesting trends.
- First, the resulting ranking placed *Archaea* before *Bacteria*, implying that *Archaeal* species have genes that are shorter than *Bacterial* ones.
- Within each kingdom, different phyla have preferences for short or long genes.
- We demonstrated that thermophiles tend to have shorter genes than the soil-dwellers.
- Second, there is a significant correlation between GC₃ content and gene length.

Conclusions

- Genome ordering procedure is stable: inclusion of additional genomes does not affect relative ranking of genomes.
- We demonstrated that correlation coefficient between the ranks of the 100 genomes in the 100- genomes dataset and in the larger (1390) dataset is 0.95.
- Hyperthermophilic species are ranked on top both in 100 and 1390 genomes lists; soil dwelling species are consistently in the bottom of the list.

Consistency of Bubble Sort ranks in 1390 and 100 genomes datasets.

Pearson's correlation coefficient between two ranks is 0.95; Kendall tau correlation coefficient is 0.82.



Thank you

ALEXANDER BOLSHOY

PROFESSOR

DEPT. OF EVOLUTIONARY AND ENVIRONMENTAL
BIOLOGY

UNIVERSITY OF HAIFA

ISRAEL

OMICS International Open Access Membership

Open Access Membership with OMICS International enables academic and research institutions, funders and corporations to actively encourage open access in scholarly communication and the dissemination of research published by their authors.

For more details and benefits, click on the link below:

<http://omicsonline.org/membership.php>

