A Markovian Analysis of Bacterial Genome Sequence Constraints

The arrangement of nucleotides within a bacterial chromosome is influenced by numerous factors. The degeneracy of the third codon within each reading frame allows some flexibility of nucleotide selection; however, the third nucleotide in the triplet of each codon is at least partly determined by the preceding two. This is most evident in organisms with a strong G+C bias, as the degenerate codon must contribute disproportionately to maintaining that bias. Therefore, a correlation exists between the first two nucleotides and the third in all open reading frames. If the arrangement of nucleotides in a bacterial chromosome is represented as a Markov process, we would expect that the correlation would be completely captured by a second-order Markov model and an increase in the order of the model (e.g. third-, fourth-…order) would not capture a significant additional uncertainty in the process. In this manuscript, we present the results of a comprehensive study of the Markov property that exists in the DNA sequences of 906 bacterial chromosomes. All of the 906 bacterial chromosomes studied exhibit a statistically significant Markov property that extends beyond second-order, and therefore cannot be fully explained by codon usage. An unrooted tree containing all 906 bacterial chromosomes based on their transition probability matrices of third-order shares ~25% similarity to a tree based on sequence homologies of 16S rRNA sequences. This congruence to the 16S rRNA tree is greater than for trees based on lower-order models (e.g. second-order), and higher-order models result in diminishing improvements in congruence. A nucleotide correlation most likely exists within every bacterial chromosome that extends past three nucleotides. This correlation places significant limits on the number of nucleotide sequences that can represent probable bacterial chromosomes. Transition matrix usage is largely conserved by taxa, indicating that this property is likely inherited, however some important exceptions exist that may indicate the convergent evolution of some bacteria.

1   Roy D Welch

2   Department of Biology, Syracuse University, United States

3   rowelch@syr.edu

4

5   Aaron D Skewes

6   Department of Biology, Syracuse University, United States

7   Department of Mathematics, Syracuse University, Syracuse, NY, United States

8   askewes@syr.edu8

9 **Introduction**

10 For more than twenty years, the nucleotide composition of bacterial genomes has been the focus of

11 many studies attempting to identify patterns in nucleic acid sequences. One of the first analyses of

12 nucleotide sequences by Muto and Osawa noted that nucleotide biases exist and are likely influenced

13 by selection ([47]). Later work By Karlin and Burge proposed that a bacterial signature could be defined

14 by certain statistical properties of complete sequences ([34]). They discovered that correlations exist

15 between neighboring nucleotides (dinucleotides) in bacteria, and that dinucleotide frequencies can be

16 used as a genomic signature which may result from: 1) the chemistry of dinucleotide stacking; 2) DNA

17 conformational tendencies; 3) species-specific properties of DNA replication and repair mechanisms;

18 4) the selection of restriction endonucleases ([35]); and 5) codon usage, as it effects translational

19 efficiency ([27], [28], [60]). These and other pioneering studies were narrow in scope because, at that

20 time, available data was limited to single gene sequences, partial chromosomes, and the complete

21 genomes of a small number of model organisms, such as *Escherichia coli K-12* ([10])*, Haemophilus*

22 *influenzae* ([21]) and *Bacillus subtilis* ([44]). Nevertheless, these analyses were instrumental in laying the

23 foundation for statistical genomics. In this early period, researchers were forced to focus on very

24 specific phenomena or draw broad conclusions from data sets that were insignificant when compared

25 to the size of the global metagenome. The situation is beginning to change. Genome sequences are

26 now available for more than 2,000 bacterial species, which may represent as much as ~0.002% of all

27 bacteria ([15], [58]). With this expanded data set we can begin to address new types of questions. For

28 example, we can begin to identify sequence features that may constrain nearly all bacterial genomes,

29 and thereby describe a set of heuristics that may eventually help define the statistical boundaries of

30 what constitutes a bacterium.

31 One established method used to model genome sequences is the finite state Markov chain model

32 (see Methods) ([3], [6], [9], [12], [24]). Markov models are defined by a transition matrix, which stores the

33   conditional probabilities of the kth symbol following the previous k-1 symbols in a word of length k; they

34   are akin to word frequency counts. Applying this type of analysis to a complete genome sequence

35   provides information about dynamic and stationary statistics that cannot be captured from a single

36   gene or set of genes. One of the first applications of Markov models to the analysis of genetic

37   sequences was their use as a method to identify sequence bias. Pioneering work by researchers

38   including Phillips ([50](#)) Rocha ([54](#)), and Burge and Karlin ([13](#)) established that Markov analysis of DNA

39   sequences can be useful in identifying over- and under-represented sequences. Work by Elhai ([19](#))

40   compared several different statistical methods of finding bias in the relative abundance of

41   oligonucleotides in DNA sequences. All these methods were based on comparing observed

42   oligonucleotide frequencies to their expectation under several models, and all concluded that Markov

43   model based methods underperformed some more complex methods, when the purpose of the

44   method was to determine abundance.

45   Determining relative abundance is not the only reason for examining DNA sequences, however, and

46   when looking for other patterns an empirically derived Markov model does contain valuable

47   information. For example, lateral gene transfer events produce a localized nucleotide bias that can be

48   detected with variations of Markov models, although they must be recent events, as the bias tends to

49   disappear in a short period of evolutionary time ([45](#), [52](#)). Many studies have examined this

50   phenomenon and have concluded that the lateral transfer of genetic material is a very important factor

51   in bacterial evolution ([14](#), [17](#), [31](#), [41](#), [65](#)). This conclusion may seem obvious now but, at the time, it

52   challenged many assumptions about vertical descent, the meaning of phylogenetics, and how

53   phylogenies are constructed ([46](#)). This and similar Markov model based methods have also revealed

54   niche and habitat influences in the genomic composition of bacteria at the G+C content level ([22](#)), the

55   amino-acid level ([62](#)), and the whole genome level ([49](#)). One of the earliest gene prediction methods

56   ([11](#)) used non-homogenous Markov models to estimate the probability that a particular location along

57 the genome of an organism contains genetic information. Clearly, Markov models have a purpose in

58 evaluating patterns in DNA sequences, although they may not always be the best choice.

59 Many studies have explored the use of Markov models to infer phylogenies, as an alternative to

60 methods based on multiple sequence alignment. Höhl et.al ([29](#)) compared several alignment-free

61 methods for inference of phylogeny based on bacterial amino acid sequences. They made two

62 important conclusions: 1) methods based on k-mer frequencies are generally inferior to approaches

63 based on maximum-likelihood distance estimates of multiply aligned sequences and; 2) there is an

64 optimal word length (k) which produces a stable inferred tree, beyond which there is only a negligible

65 improvement in stability. A similar conclusion was reached by Jun et. al. ([33](#)) and Dai et. al, ([16](#)) using

66 proteome sequences of prokaryotes. Again, it must be noted that these studies were looking for

67 optimal ways to identify a particular set of data; their conclusions do not mean that Markov methods

68 are inherently inferior. There is a significant amount of information contained in the transition matrix of

69 a bacterial genome beyond what these studies were looking for, and the existence of an optimal word

70 length may indicate that a lower-order Markov model can capture the majority of the information

71 contained in higher-order models. The application of finite state Markov chain models to identify

72 patterns that exist in bacterial genomes can help in understanding molecular change, in developing

73 molecular criteria for classification, and in exploring the boundaries of what may (or may not)

74 constitute a viable genome sequence.

75 Sequenced bacterial genomes span a size range of approximately two orders of magnitude, from

76 *Carsonella ruddii* (~0.15MB) ([48](#)) to *Sorangium cellulosum* (~13MB) ([59](#)), and a range of %G+C

77 content from a low of ~17% in *Carsonella ruddii* to ~75% in *Anaeromyxobacter dehalogenans* ([56](#)). If

78 we consider the set of all possible bacterial chromosomes to include every closed circular DNA

79 sequence that fits within these ranges, the number of distinct chromosomal sequences would be

80 overwhelming. Determining the subset of probable bacterial chromosomes from the set of possible

81   bacterial chromosomes is a problem whose complexity is analogous to protein structure prediction. To

82   begin addressing this problem, we can apply heuristics based on biological phenomena considered to

83   be ubiquitous. For example, we might propose that a sequence must contain codons, open reading

84   frames, regulatory sequences, and a certain set of "essential" genes in order for it to be included in the

85   probable subset. Applying these kinds of heuristics renders the subset of probable chromosomes

86   much smaller that the set of possible chromosomes, but it would still be an overwhelmingly large

87   number. Also, the boundaries of the subset would not be hard, since consensus on parameters such

88   as the number of open reading frames and the list of essential genes would be impossible.

89   An independent and complementary approach to developing heuristics to limit the subset of probable

90   bacterial chromosomes would be to base them on sequence patterns identified either as ubiquitous or

91   extremely rare. This type of heuristic would not rely on a biological interpretation of sequence data, but

92   rather on definable sequence patterns that are highly likely or unlikely to occur in the population based

93   on their appearance within a representative sub-population. A few heuristics have already been

94   proposed. For example, Lawrence and Ochman summarized four salient features of prokaryotic

95   genomes ([45]): 1) base composition varies widely among bacterial species; 2) base composition is

96   related to phylogeny; 3) base composition is relatively homogeneous over the entire bacterial

97   chromosome; and 4) within each species, the first, second, and third positions of codons, as well as

98   the genes for structural RNAs, have characteristic base compositions. Once defined, these features

99   can be explored and parameterized into models capturing certain properties..

100 Despite limited available data, early studies made some very important observations. Karlin et al.

101 identified correlations between neighboring nucleotides ([34]) (i.e., the probability of appearance of the

102 n^{th} nucleotide depends on the $(n-1)^{th}$ nucleotide), and concluded that dinucleotide frequencies carry a

103 phylogenic signal. Goldman and others discovered that tri- and tetranucleotide correlations exist in

104 bacterial sequences ([26], [36], [38]) . Tetranucleotide frequencies have also been found to carry a

105 phylogenetic signal, and to reflect high-order information beyond third codon biases that are not

106 present in the analysis of single genes ([51](#)). The study by Pride et. al ([51](#)) looked at tetranucleotide

107 usage conservation in 27 microbial genomes, and compared a tree based on tetranucleotide usage

108 departures to that of 16S rRNA trees. They concluded that tetraucleotide usage patterns are

109 conserved by taxa, and that usage departure is a measure of how far tetranucleodide frequencies

110 diverge from the expectations under a null-model, which in their case was designed to remove any

111 sequence bias. This approach has been useful in identifying under- and over-represented

112 olglionucleotides ([3](#), [37](#), [57](#)).

113 **Materials and Methods**

114 The complete DNA sequences and 16S ribosomal DNA sequences were collected for 906 closed

115 bacteria from GenBank ([8](#)) (for a complete list see S1). For organisms having multiple chromosomes,

116 the major chromosome was selected as representative of the genomic sequences of the respective

117 organism. Our analysis indicates that the DNA sequence of the major chromosome in bacteria has

118 similar statistical properties in regards to nucleotide probabilities as a sequence constructed by

119 appending all the chromosomes for that organism, excluding plasmids (data not shown). All software

120 developed for this work was written in C++, except where otherwise noted.

121 **Constructing the 16S rRNA tree**

122 The ribosomal DNA sequences for each of the 906 bacteria were obtained from GenBank, and the

123 DNA sequence corresponding to the 16S ribosome was written to a single FASTA file. In organisms

124 having multiple copies of 16S rRNA, the first copy relative to the 5' direction was chosen as

125 representative of the organism ([2](#)). The 16S rRNA sequences were aligned using MUSCLE ([18](#)).

126 Aligned 16S rRNA sequences were bootstrapped with 100 replicates and transformed into distances

127 using the F84 ([40](#)) model available in the Phylip package (Felsenstein, J. 1993). We chose to use the

128 F84 distance method because, unlike other methods (e.g. Jukes and Cantor's ([32](#)) and K80 ([39](#))), it

129　allows for both unequal base frequencies and unequal transition/transversion probabilities. The base

130　frequencies and transition/transversion probabilities are estimated from the data, and the distances

131　can be interpreted as a maximum likelihood estimate of the divergence time; this provides an accurate

132　representation of bacterial sequence dynamics. Each replicated distance matrix was clustered using

133　the Neighbor-joining method ([55]). Neighbor-joining was used because of its speed and accuracy when

134　given a correct distance matrix. A majority-rule consensus tree was calculated using Phylip (Phylip

135　formatted tree available as S2). Tree visualizations shown in this paper were produced using

136　Dendroscope ([30]) and ladderized right.

137　Constructing the transition tree

138　The frequency of each genomic subsequence and its reverse complement (3' $\longrightarrow$ 5') of length n

139　appearing in each bacterial genome was explicitly counted. The transition probabilities were estimated

140　for the kth-order transition matrix (k = n-1), for $0 \leq k \leq 5$, from the subsequence frequencies. The

141　Euclidean distance was computed between each transition matrix describing each of the 906 bacterial

142　sequences for a given order of Markov chain model. The Euclidean distances were clustered using the

143　Neighbor-joining method available in the Phylip package (Phylip formatted tree available as S3). Tree

144　visualizations were produced using Dendroscope ([30]) and ladderized right.

145　Determination of tree similarity

146　A direct method of assessing tree similarity comes from set theory and is referred to as the symmetric

147　difference ([53]). The symmetric difference of a tree structure is the total number of partitions that differ

148　between the two trees. We used the percent symmetric difference, which is the symmetric difference

149　(D_s) divided by the maximum symmetric difference (D_{max}), with $D_{max} \approx 2n-6$ for n-number of taxa. The

150　significance of D_s for a given number of taxa can be estimated empirically, and is shown to be

151　asymptotic, with a convergence rate dependent on n ([61]). For n = 30, any $D_s < (D_{max} - 2)$ is significant,

152   with p < 0.01. The symmetric difference method as implemented in the Phylip package was used for

153   the data presented in figures 1 and 2.

154   **Markov models of bacterial chromosomes**

155   A chromosome sequence can be modeled as a finite state space Markov chain, with each of the four

156   nucleotides (A,T,G,C) represented by a single state with transition probabilities P_A, P_T, P_G and P_C

157   respectively. This representation is memoryless, in that the appearance of any nucleotide at any

158   position is completely independent of any other. This is also referred to as a 0th order Markov model,

159   and in this context can only capture biases in the relative frequency of appearance of the nucleotides

160   (e.g., G+C bias and A/T fraction bias). The transition matrix, θ, for the 0th order Markov model

161   describing the finite state space Markov chain is:

$$\theta = \left[P_A, P_T, P_G, P_C \right].$$

162

163   In higher-order Markov models, transition probabilities are conditional on the previous k bases (for

164   k>0). For example, we can consider a 1st order Markov model with transition probabilities $P_{A|A}$, $P_{T|A}$,

165    $P_{G|A}$, $P_{C|A}$ ... $P_{G|C}$, $P_{C|C}$, where $P_{i|j}$ is the probability of the ith nucleotide following the jth nucleotide. We can

166   easily generalize this to describe the transition matrix for a *k*-order Markov model representing a

167   genomic sequence, with $\theta = [i,j] 4^k \times 4^1.$ $\theta = [i,j]_{4^k \times 4^1}$

168   **Testing for Markovity**

169   We adapted an existing framework that is rooted in information theory to estimate the significance of a

170   Markov property of order k. The framework tests for contingency tables and calculates the value of the

171   X^2 statistic ([4], [43]), which tests the null hypothesis that the sequence is a realization of a stationary

172   Markov chain of order k-1, against the alternative hypothesis that it is it a realization of a k-order

173   stationary Markov chain. The application of this method to DNA sequences is discussed by Avery et. al

174 ([7](#)). For very long sequences, such as chromosomal DNA, the resulting X^2 statistic for lower-order

175 Markov models is almost guaranteed to be significant (p-value near zero). However, what can be

176 assessed is the change in value of the statistic as the order and degrees of freedom increases. In

177 other words, for increasing order k, we can observe the change in uncertainty between k and k-1 order

178 models in terms of the X^2 statistic, and look for asymptotic behavior in its value. A constant X^2 value

179 becomes less significant as the degrees of freedom increase. When the statistic's value is asymptotic

180 in the presence of increasing degrees of freedom, the static value will inevitably become insignificant.

181 For these data in this paper, the value of the X^2 statistic is significant for all sequences with k≤3 (see

182 S4 for a table of the statistics for each chromosome considered in this work) and nearly asymptotic at

183 k=3 for the majority. Any violations to the second observation are because of sequences with relatively

184 short lengths, high G+C bias, or some combination of the two.

185 **Results and Discussion**

186 The goal of this work is not to devise a new or improved method of phylogenetic inference, or to imply

187 that Markov models are superior to other methods. Rather, our goal is to address the following three

188 questions 1) is there a universal Markov property present in whole bacterial DNA sequences; 2) to

189 what extent (order) does this property hold true; and 3) is the existence of the Markov property

190 biologically relevant.

191 Using the complete nucleotide sequences of 906 bacteria, including its complement, and excluding

192 plasmids and minor chromosomes ([8](#)) (see S1 for a complete list or organisms), we estimated the 0th -

193 5th order transition matrices, describing the respective order Markov chain model for each. The 5th

194 order model intersects at least two codons and, given the length of bacterial genomes, it is still short

195 enough to allow sufficient oligonucleotide frequencies to avoid sparse transition matrices. We then

196 calculated the Euclidean distance between each pair of transition matrices for each order model (one

197 distance matrix for each order Markov chain model for all chromosomes) and produced a cladogram

198   from the distances based on the Neighbor Joining method ([55]). We refer to this kind of tree as a

199   "transition tree".

200   Branching patterns of trees based on alignments of 16S ribosomal RNAs are an accepted method to

201   represent phylogeny ([23], [66]). To see if this is also a characteristic of the transition tree, we performed

202   a comparison between each transition tree and a 16S rRNA tree constructed in similar fashion (see

203   Methods for a detailed description). Briefly, 16S rRNA sequences for each of 906 bacteria were

204   collected from GenBank and aligned against one another using MUSCLE ([18]). Alignments were

205   bootstrapped with replacement ([20]), transformed into a distance matrix, clustered using the Neighbor

206   Joining method, and a cladogram was produced for visual comparison.

207   **Comparisons between the 16S rRNA and transition tree topologies**

208   Using the symmetric difference method ([53]) of comparing tree topologies, we calculated the percent

209   symmetric difference of each transition tree (1st - 5th order) relative to the16S rRNA tree (fig 1, right)

210   and to the 0th order transition tree (fig 1, left). Previous research on the distribution of D_s from

211   simulation data has shown it to be asymptotic in nature, with convergence dependent on the number

212   of taxa. These findings are summarized in Steel and Penny ([61]), and suggests that for trees with more

213   than a moderate number of taxa, any $D_s < D_{max}$ is significant, (e.g., for n = 30, any similarity of more

214   than a few partitions is very unlikely). Therefore, no similarity in topology is predicted between

215   randomly placed nodes in trees with a large number of taxa. As shown in figure 1 (left panel), the

216   congruence, $(1- (D_s/D_{max}))\times100$, between the 0th order tree, which is a function of G+C content alone,

217   and the 16S rRNA tree is low ($D_s/D_{max}\times100=96.7\%$). However, as summarized by Steel and Penny,

218   even this small difference from D_{max} is significant, and this suggests that there is some influence of

219   G+C content reflected in the 16S rRNA tree. A similar conclusion can be reached by examining figure

220   1 (right panel). The percent symmetric difference between the 0th order transition tree and the

221   higher-order transition trees is large (90.7% - 93.5%), but even this small degree of congruence is

222 considered significant, and it reflects the influence of the 0th order model on the higher-order models.

223 Interestingly, the effects of G+C content are rather stable beyond the 2nd order model, in that the

224 percent symmetric difference between the 0th order model and higher order models (beyond 2nd order)

225 does not change by a large amount (93.0% - 93.5%). These observations lead us to conclude that

226 G+C content bias has a real but relatively small influence on both the 16S rRNA tree and the transition

227 tree.

228 The symmetric difference between the 16S rRNA tree and each of the transition trees decreases most

229 between the 0th and 3rd orders (96.7% - 75.5%), with little additional decrease between the 3rd and 5th

230 order (75.2% - 74.8%). These data lead us to conclude the following: 1) the 3rd order transition tree

231 shares ~25% similarity to the 16S rRNA tree; 2) this congruence is greater than for trees based on

232 lower-order models; 3) this congruence is similar to trees based on higher-order models. These

233 conclusions are further supported by data presented in figure 2. We calculated the percent symmetric

234 difference between subsequent orders of transition trees and observed that from 3rd until 5th order,

235 each order transition tree shares approximately 65% of its partitions with its previous and subsequent

236 order trees. This leads us to conclude that large decreases in symmetric difference between

237 subsequent orders of transition trees stop after the 3rd order. Of course, if we continued to increase the

238 order of the Markov models indefinitely, the subsequent tree topologies produced by k-1 and k order

239 models would eventually converge. This is due to the increasingly sparse transition matrices. For a

240 given sequence, the transition matrix would approach the null set, with only two elements populated

241 (that corresponding to (1…n-1) and to (2…n) for a sequence of length n) with a frequency count of

242 1.The resulting distance matrix, based on the sparse transition matrices, will reach steady-state. For

243 any particular sequence, the complexity of the model necessary to achieve this convergence depends

244 on many factors, including sequence length and G+C content bias. Convergence is inevitable,

245 however, because it is inherent in the model. In other words, in the most extreme case, we can always

246 find a model of order equal to the sequence length -1.

247 **Data bias**

248 Bias must be considered because it exists in the collection of sequenced bacteria. Some Genera (e.g.

249 *Escherichia, Streptococcus*, and *Bacillus*) are overrepresented, while others are underrepresented.

250 We must therefore consider the possibility that the 16S rRNA tree and the transition tree show a

251 greater degree of congruence in more closely related species, so that the overrepresented genera

252 would inflate the overall congruence in topology between the transition trees and the 16SrRNA tree. To

253 determine if this effect exists, four overrepresented genera, *Escherichia* (29 species), *Streptococcus*

254 (43 species), *Bacillus* (24 species) and *Burkholderia* (23 species), totaling 119 species (~13% of the

255 data collection) were chosen, and 16S rRNA and 3rd order transition trees were constructed. This

256 subset of species was selected to represent an exaggerated sequencing bias so that, if the observed

257 congruence between 16S rRNA trees and transition trees is partly due to this bias, it should be

258 amplified in this subset. Instead, the symmetric difference between these trees was calculated as

259 76.7%, which very close to the 75.5% measured using the entire 906 bacteria. We therefore conclude

260 that sequence bias has no significant impact on these results.

261 **Topology of 16S rRNA tree versus 3rd order transition tree**

262 The symmetric difference between the 16S rRNA tree and the 3rd order transition tree is presented in

263 figure 3 as the 16S rRNA tree, with branches in red representing disagreement between it and the 3rd

264 order transition tree. The 16S rRNA tree and 3rd order transition tree from which figure 3 is derived are

265 provided in supplementary materials (S2 and S3, respectively). In figures 4 – 6, the taxa of interest are

266 shown in red, with the 16S rRNA tree on the left and the transition tree on the right. Comparisons are

267 made relative to the transition tree, with all organisms of a particular genera of interest accounted for

268 in both trees (as either a group member or outlier in the transition tree).

269 There is good agreement between the 16S rRNA tree and the 3rd order transition tree in several

270 places; figure 4 presents a large collection of Enterobacteriaceae as an example. This grouping

271 includes *Salmonella*, *Escherichia* and *Shigella*, and the transition tree shows consistent grouping of

272 each genus as compared to the 16S rRNA tree. The 16S rRNA sequences of *Shigella* and *Escherichia*

273 are very homologous, and this results in some species from each genus being shuffled within the 16S

274 rRNA tree as opposed to the transition tree, which is more sensitive to sequence bias. This shuffling is

275 not observed in the transition tree.

276 Figure 5 illustrates differences in how the genus *Streptococcus* clusters in the 16S rRNA tree versus

277 the transition tree. In the 16S rRNA tree, all of the Streptococci form one cluster, whereas in the

278 transition tree there are two separate clusters. The two clusters do not divide based on hemolytic

279 properties, serogroup or habitat, however each group has a distinct G+C content ($p<0.05$ with

280 Students t-test) with group one $\mu=40.43\%$, $\sigma^2=1.05\%$, $n=21$ and group two $\mu= 37.92\%$, $\sigma^2= 1.43\%$,

281 $n=22$, where μ is the mean G+C content, σ^2 is the variance about the mean, and n is the number of

282 samples. There is a distinct difference in nucleic acid content between the two groups of Streptococci

283 that does not appear to follow the typical physiological traits used to define these organisms. In this

284 case, the transition tree is detecting clear molecular differences between otherwise similar organisms.

285 Figure 6 highlights a group of bacteria that cluster tightly in the transition tree (with outliers in boldface

286 type), but are separated into distinct groups in the 16S rRNA tree. This group includes members of the

287 *Polynucleobacter, Psychrobacter, Marinomonas, Shewanella* and *Vibrio* genera, with a G+C content

288 range of approximately 40% - 49%. Most of these organisms are associated with a cold-water aquatic

289 habitats. It is known that thermophiles exhibit preferences in the first codon for G+C, due to the higher

290 melting temperatures ([42](), [63]()); the converse of this is a similarly reasonable explanation.

291 Thermophobic bacteria may prefer A+T at the first codon due to lower separation energies required

292 during replication. Although members of Yersinia and six species of Lactobacillus may initially appear

293 to contradict this observation, this may not be the case. *Yersinia pseudotuberculosis* is a soil- and

294 waterborne human pathogen, and the closest known ancestor of *Yersinia pestis* ([1]()), and many species

295   of Lactobacillus can be found in marine sediment. There is further evidence in support of our aquatic

296   hypothesis within the other genera. Two species of *Shewanella* are located outside the cluster, *S.*

297   *amazonensis* and *S. loihica.* Both of these organisms are psychrophobic, whereas the *Shewanella*

298   species within the cluster grow well at low to moderate temperatures. Also, *Vibrio* is a genus of

299   proteobacteria that are a common cause of food-borne illness resulting from infected seafood. *V.*

300   *fischeri*, which is the only Vibrio outlier, is unique among *Vibrio* species because it is apathogenic and

301   found predominantly in symbiosis with various marine animals. We hypothesize that these species

302   represent outliers on the transition tree because they occupy different habitats from their 16S nearest

303   neighbors.

304   Habitat has been shown to influence genomic composition ([22](#)). Perhaps the difference between the

305   16S rRNA tree and 3rd order transition tree illustrated in Figure 4 is an example of that influence.

306 **Conclusions**

307 These data and analyses lead us to the following three observations: 1) in nearly all bacterial

308 chromosomes there is a significant long-range nucleotide correlation that extends beyond the

309 2nd order; 2) similarity trees constructed on matrices derived from these correlations have a

310 statistically significant overlap with 16S RNA trees and, when divergent, may reveal functional

311 differences between species; 3) the apparent ubiquity of these correlations may place practical

312 limitations on what will or will not evolve to become a bacterium.

313 These observations cannot easily be explained by our understanding of biology. Overall G+C

314 bias is a 0th order property, so that its influence is completely defined by the independent

315 probabilities of each of the four nucleotides. A codon is three nucleotides long, so codon bias

316 within open reading frames is a1^{st} order (binucleotide) or 2nd order (trinucleotide) correlation. Any

317 correlations that extend beyond 2nd order reflect a mechanism or mechanisms that drive the

318 nucleic acid order beyond the length of a codon.

319 We have also shown that the transition matrices for a large number of chromosomes exhibit a

320 phylogenetic correlation. From the matrices we can build transition trees that are statistically

321 similar to 16S rRNA trees, and we propose that some of the differences between transition and

322 16S trees may be due to influences from ecological niche and/or habitat. Proximity would

323 present organisms that occupy similar habitats, such as cold water, with the opportunity to share

324 genetic material that increases their likelihood for survival, such as anti-freeze genes ([25]).

325 Although transfer of small bits of genetic material would not account for similarity of transition

326 matrices between whole chromosomes of distantly related organisms, DNA sharing has been

327 previously observed on a much larger scale. Specialized bacteria that occupy the same habitat

328 or ecological niche also may experience convergent evolution ([5], [64]). Horizontal gene transfer

329 (HGT) is known to play a major role in how bacteria acquire new genetic material. It seems

330   logical that organisms within the same habitat might acquire similar genomic characteristics via

331   HGT.

332   The Markov property we have described appears to be ubiquitous. We were able to identify the

333   property in all of the 906 chromosomes we studied, and it has been estimated that there are

334   ~10^8 bacterial species on the Earth ([15], [58]). Using the statistical "rule of three", we can be 95%

335   confident that the rate of this phenomenon is no less frequent than 301 in 302 bacterial

336   chromosomes. We therefore conclude that the majority of all of the bacterial species will have

337   this Markov property in their chromosomes, and this likely represents a statistical heuristic that

338   limits the sequence space of probable bacterial chromosomes.

339   What, if any effect this conclusion has on our ability to select the probable from the much larger

340   number of possible bacterial chromosomes is impossible to quantify, but we may be able to

341   illustrate some of its impact and provide context through example. If any one of the set of

342   possible bacterial chromosomes can be represented as a random closed circle of nucleotides,

343   and if we assume one biology-based heuristic –that it can be any integral length between the

344   smallest (0.15Mb) and the longest (15Mb) sequenced chromosome, then there are

345   $$\sum_{n=\alpha}^{\beta} 4^n = \frac{4}{3}\left(4^\beta - 4^{\alpha-1}\right) \approx$$ $10^{9,000,000}$ possible bacterial chromosomes (α=0.15M and β=15M) each

346   with an equal probability of occurrence. If we now consider that most bacterial chromosomes

347   have a compositional bias (e.g. G+C content), some of the possible combinations become

348   more or less probable. If we then consider higher-order compositional biases, say an A/T

349   fractional bias in addition to a G+C content bias, then we can be even more specific about

350   probable and improbable chromosomes.

351   With the existence of a high-order Markov process, the number of variables (states) increases

352   exponentially with each increase in model order. This allows a more precise determination of the

353   probability of a particular sequence (i.e. greater resolution of transition probabilities), and

354   thereby the identification of more sequences that are unlikely to be bacterial chromosomes. Let

355   X_L^K define a sequence of K letters over an alphabet of L characters, then the probability of

356   sequence X_L^K is:

357   $$P(x_L^K) = \prod_{j=1}^{K} P(X_j = x_j \mid X_L^{j-L} = x_L^{j-L})$$, where X_j represents the nucleotide at position j with x_j as

358   its realization. For a DNA sequence (and assuming a 3rd-order Markov Model), L=K=4. In the

359   trivial case, where each character (nucleotide) is equally likely to occur, it can be easily shown

360   that $P(x_L^K) = \dfrac{1}{L^K}$ and the expected frequency $f(x_L^K) = \dfrac{N - K - 1}{L^K} \approx \dfrac{N}{L^K}$ for K << N. For any

361   sequence that is the result of a 3rd-order Markov process and modeled as such, we get $L^K = 4^4$

362   times more states than with a 0-order model. In other words, we get 256 times greater

363   resolution of transition probabilities than if we just consider limitations of G+C bias and

364   chromosome length.

365   We know that many of the biological constraints placed on an organism limit the number of

366   possible combinations that can result in a viable genomic sequence, but these constraints seem

367   difficult to quantify. Now that we have a significant sample size of sequenced bacterial

368   chromosomes, we can identify some of the constraints through statistical methods, and perhaps

369   also uncover new biological phenomena.

370   **Acknowledgments**

373 review of this manuscript. We also want to express our appreciation to Barry Goldman,

374 Monsanto Corp. for his insights regarding the possible data bias and into the lifestyle of many of

375 the 906 bacteria included in this study and his time spent reading the drafts. We would also like

376 to thank Laura Welch for her editorial comments.

References

378 1. **Achtman, M., K. Zurth, G. Morelli, G. Torrea, A. Guiyoule, and E. Carniel.** 1999.
379   Yersinia pestis, the cause of plague, is a recently emerged clone of Yersinia
380   pseudotuberculosis. Proceedings of the National Academy of Sciences of the United
381   States of America **96:**14043-14048.
382 2. **Acinas, S. G., L. A. Marcelino, V. Klepac-Ceraj, and M. F. Polz.** 2004. Divergence and
383   Redundancy of 16S rRNA Sequences in Genomes with Multiple rrn Operons. J.
384   Bacteriol. **186:**2629-2635.
385 3. **Almagor, H.** 1983. A Markov analysis of DNA sequences. Journal of Theoretical Biology
386   **104:**633-645.
387 4. **Anderson, T. W., and L. A. Goodman.** 1957. Statistical-Inference About
388   Markov-Chains. Ann Math Stat **28:**89-110.
389 5. **Audic, S., C. Robert, B. Campagna, H. Parinello, J. M. Claverie, D. Raoult, and M.
390   Drancourt.** 2007. Genome analysis of Minibacterium massiliensis highlights the
391   convergent evolution of water-living bacteria. Plos Genet **3:**1454-1463.
392 6. **Avery, P. J.** 1987. The Analysis of Intron Data and Their Use in the Detection of Short
393   Signals. Journal of Molecular Evolution **26:**335-340.
394 7. **Avery, P. J., and D. A. Henderson.** 1999. Fitting Markov chain models to discrete state
395   series such as DNA sequences. J Roy Stat Soc C-App **48:**53-61.
396 8. **Benson, D. A., I. Karsch-Mizrachi, D. J. Lipman, J. Ostell, and D. L. Wheeler.** 2004.
397   GenBank: update. Nucl. Acids Res. **32:**D23-26.
398 9. **Blaisdell, B. E.** 1985. Markov-Chain Analysis Finds a Significant Influence of
399   Neighboring Bases on the Occurrence of a Base in Eukaryotic Nuclear-DNA Sequences
400   Both Protein-Coding and Noncoding. Journal of Molecular Evolution **21:**278-288.
401 10. **Blattner, F. R., G. Plunkett, III, C. A. Bloch, N. T. Perna, V. Burland, M. Riley, J.
402   Collado-Vides, J. D. Glasner, C. K. Rode, G. F. Mayhew, J. Gregor, N. W. Davis, H.
403   A. Kirkpatrick, M. A. Goeden, D. J. Rose, B. Mau, and Y. Shao.** 1997. The Complete
404   Genome Sequence of Escherichia coli K-12. Science **277:**1453-1462.
405 11. **Borodovsky, M., and J. McIninch.** 1993. GENMARK: Parallel gene recognition for both
406   DNA strands. Comput Chem **17:**123-133.
407 12. **Brendel, V., J. S. Beckmann, and E. N. Trifonov.** 1986. Linguistics of
408   Nucleotide-Sequences - Morphology and Comparison of Vocabularies. Journal of
409   Biomolecular Structure & Dynamics **4:**11-21.
410 13. **Burge, C., A. M. Campbell, and S. Karlin.** 1992. Over- and under-representation of
411   short oligonucleotides in DNA sequences. Proc Natl Acad Sci U S A **89:**1358-1362.
412 14. **Campbell, A. M.** 2000. Lateral gene transfer in prokaryotes. Theoretical Population
413   Biology **57:**71-77.
414 15. **Curtis, T. P., W. T. Sloan, and J. W. Scannell.** 2002. Estimating prokaryotic diversity
415   and its limits. Proceedings of the National Academy of Sciences of the United States of
416   America **99:**10494-10499.

417 16. **Dai, Q., and T. M. Wang.** 2008. Comparison study on k-word statistical measures for
418   protein: From sequence to 'sequence space'. Bmc Bioinformatics **9**.
419 17. **Doolittle, W. F.** 1999. Phylogenetic classification and the universal tree. Science
420   **284:**2124-2128.
421 18. **Edgar, R. C.** 2004. MUSCLE: multiple sequence alignment with high accuracy and high
422   throughput. Nucleic Acids Res **32:**1792-1797.
423 19. **Elhai, J.** 2001. Determination of bias in the relative abundance of oligonucleotides in
424   DNA sequences. J Comput Biol **8:**151-175.
425 20. **Felsenstein, J.** 1985. Confidence-Limits on Phylogenies - an Approach Using the
426   Bootstrap. Evolution **39:**783-791.
427 21. **Fleischmann, R., M. Adams, O. White, R. Clayton, E. Kirkness, A. Kerlavage, C.**
428   **Bult, J. Tomb, B. Dougherty, J. Merrick, and e. al.** 1995. Whole-genome random
429   sequencing and assembly of Haemophilus influenzae Rd. Science **269:**496-512.
430 22. **Foerstner, K. U., C. von Mering, S. D. Hooper, and P. Bork.** 2005. Environments
431   shape the nucleotide composition of genomes. EMBO Rep **6:**1208-1213.
432 23. **Fox, G. E., E. Stackebrandt, R. B. Hespell, J. Gibson, J. Maniloff, T. A. Dyer, R. S.**
433   **Wolfe, W. E. Balch, R. S. Tanner, L. J. Magrum, L. B. Zablen, R. Blakemore, R.**
434   **Gupta, L. Bonen, B. J. Lewis, D. A. Stahl, K. R. Luehrsen, K. N. Chen, and C. R.**
435   **Woese.** 1980. The phylogeny of prokaryotes. Science **209:**457-463.
436 24. **Gelfand, M. S., C. G. Kozhukhin, and P. A. Pevzner.** 1992. Extendable words in
437   nucleotide sequences. Comput Appl Biosci **8:**129-135.
438 25. **Gilbert, J. A., P. J. Hill, C. E. R. Dodd, and J. Laybourn-Parry.** 2004. Demonstration of
439   antifreeze protein activity in Antarctic lake bacteria. Microbiology **150:**171-180.
440 26. **Goldman, N.** 1993. Nucleotide, dinucleotide and trinucleotide frequencies explain
441   patterns observed in chaos game representations of DNA sequences. Nucleic Acids Res
442   **21:**2487-2491.
443 27. **Gouy, M., and C. Gautier.** 1982. Codon usage in bacteria: correlation with gene
444   expressivity. Nucl. Acids Res. **10:**7055-7074.
445 28. **Grantham, R., C. Gautier, M. Gouy, M. Jacobzone, and R. Mercier.** 1981. Codon
446   Catalog Usage Is a Genome Strategy Modulated for Gene Expressivity. Nucleic Acids
447   Res **9:**R43-R74.
448 29. **Hohl, M., and M. A. Ragan.** 2007. Is multiple-sequence alignment required for accurate
449   inference of phylogeny? Systematic Biol **56:**206-221.
450 30. **Huson, D. H., D. C. Richter, C. Rausch, T. Dezulian, M. Franz, and R. Rupp.** 2007.
451   Dendroscope: An interactive viewer for large phylogenetic trees. BMC Bioinformatics **8:**-.
452 31. **Jain, R., M. C. Rivera, and J. A. Lake.** 1999. Horizontal gene transfer among genomes:
453   The complexity hypothesis. Proceedings of the National Academy of Sciences of the
454   United States of America **96:**3801-3806.
455 32. **Jukes, T., and C. Cantor.** 1969. Evolution of Protein Molecules. New York: Academic
456   Press**:**21-132.
457 33. **Jun, S.-R., G. E. Sims, G. A. Wu, and S.-H. Kim.** 2010. Whole-proteome phylogeny of
458   prokaryotes by feature frequency profiles: An alignment-free method with optimal feature
459   resolution. Proceedings of the National Academy of Sciences **107:**133-138.
460 34. **Kariin, S., and C. Burge.** 1995. Dinucleotide relative abundance extremes: a genomic
461   signature. Trends in Genetics **11:**283-290.
462 35. **Karlin, S., A. M. Campbell, and J. Mrazek.** 1998. Comparative DNA analysis across
463   diverse genomes. Annu Rev Genet **32:**185-225.
464 36. **Karlin, S., A. M. Campbell, and J. Mrazek.** 1998. Comparative DNA analysis across
465   diverse genomes. Annu Rev Genet **32:**185-225.
466 37. **Karlin, S., J. Mrazek, and A. Campbell.** 1997. Compositional biases of bacterial
467   genomes and evolutionary implications. J. Bacteriol. **179:**3899-3913.

468 38. **Karlin, S., J. Mrazek, and A. M. Campbell.** 1997. Compositional biases of bacterial
469   genomes and evolutionary implications. J Bacteriol **179:**3899-3913.
470 39. **Kimura, M.** 1980. A Simple Method for Estimating Evolutionary Rates of Base
471   Substitutions through Comparative Studies of Nucleotide-Sequences. J Mol Evol
472   **16:**111-120.
473 40. **Kishino, H., and M. Hasegawa.** 1989. Evaluation of the maximum likelihood estimate of
474   the evolutionary tree topologies from DNA sequence data, and the branching order in
475   hominoidea. Journal of Molecular Evolution **29:**170-179.
476 41. **Koonin, E. V., K. S. Makarova, and L. Aravind.** 2001. Horizontal gene transfer in
477   prokaryotes: Quantification and classification. Annual Review of Microbiology
478   **55:**709-742.
479 42. **Kreil, D. P., and C. A. Ouzounis.** 2001. Identification of thermophilic species by the
480   amino acid compositions deduced from their genomes. Nucleic Acids Research
481   **29:**1608-1615.
482 43. **Kullback, S., M. Kupperman, and H. H. Ku.** 1962. Tests for Contingency Tables and
483   Markov Chains. Technometrics **4:**573-&.
484 44. **Kunst, F., N. Ogasawara, I. Moszer, A. M. Albertini, G. Alloni, V. Azevedo, M. G.**
485   **Bertero, P. Bessieres, A. Bolotin, S. Borchert, R. Borriss, L. Boursier, A. Brans, M.**
486   **Braun, S. C. Brignell, S. Bron, S. Brouillet, C. V. Bruschi, B. Caldwell, V. Capuano,**
487   **N. M. Carter, S. K. Choi, J. J. Codani, I. F. Connerton, N. J. Cummings, R. A. Daniel,**
488   **F. Denizot, K. M. Devine, A. Dusterhoft, S. D. Ehrlich, P. T. Emmerson, K. D. Entian,**
489   **J. Errington, C. Fabret, E. Ferrari, D. Foulger, C. Fritz, M. Fujita, Y. Fujita, S. Fuma,**
490   **A. Galizzi, N. Galleron, S. Y. Ghim, P. Glaser, A. Goffeau, E. J. Golightly, G. Grandi,**
491   **G. Guiseppi, B. J. Guy, K. Haga, J. Haiech, C. R. Harwood, A. Henaut, H. Hilbert, S.**
492   **Holsappel, S. Hosono, M. F. Hullo, M. Itaya, L. Jones, B. Joris, D. Karamata, Y.**
493   **Kasahara, M. Klaerr-Blanchard, C. Klein, Y. Kobayashi, P. Koetter, G. Koningstein,**
494   **S. Krogh, M. Kumano, K. Kurita, A. Lapidus, S. Lardinois, J. Lauber, V. Lazarevic,**
495   **S. M. Lee, A. Levine, H. Liu, S. Masuda, C. Mauel, C. Medigue, N. Medina, R. P.**
496   **Mellado, M. Mizuno, D. Moestl, S. Nakai, M. Noback, D. Noone, M. O'Reilly, K.**
497   **Ogawa, A. Ogiwara, B. Oudega, S. H. Park, V. Parro, T. M. Pohl, D. Portetelle, S.**
498   **Porwollik, A. M. Prescott, E. Presecan, P. Pujic, B. Purnelle, et al.** 1997. The
499   complete genome sequence of the Gram-positive bacterium Bacillus subtilis. Nature
500   **390:**249-256.
501 45. **Lawrence, J. G., and H. Ochman.** 1997. Amelioration of Bacterial Genomes: Rates of
502   Change and Exchange. Journal of Molecular Evolution **44:**383-397.
503 46. **Ludwig, W., and H.-P. Klenk.** 2005. Overview: A Phylogenetic Backbone and Taxonomic
504   Framework for Procaryotic Systematics, p. 49-66. *In* D. J. Brenner, N. R. Krieg, J. T.
505   Staley, and G. M. Garrity (ed.), Bergey's Manual® of Systematic Bacteriology. Springer
506   US.
507 47. **Muto, A., and S. Osawa.** 1987. The guanine and cytosine content of genomic DNA and
508   bacterial evolution. Proceedings of the National Academy of Sciences of the United
509   States of America **84:**166-169.
510 48. **Nakabachi, A., A. Yamashita, H. Toh, H. Ishikawa, H. E. Dunbar, N. A. Moran, and M.**
511   **Hattori.** 2006. The 160-Kilobase Genome of the Bacterial Endosymbiont Carsonella.
512   Science **314:**267.
513 49. **Perry, S. C., and R. G. Beiko.** 2010. Distinguishing Microbial Genome Fragments
514   Based on Their Composition: Evolutionary and Comparative Genomic Perspectives.
515   Genome Biol Evol **2:**117-131.
516 50. **Phillips, G. J., J. Arnold, and R. Ivarie.** 1987. Mono-through hexanucleotide
517   composition of the Escherichia coli genome: a Markov chain analysis. Nucleic Acids
518   Research **15:**2611-2626.

519 51.   **Pride, D. T., R. J. Meinersmann, T. M. Wassenaar, and M. J. Blaser.** 2003.
520   Evolutionary Implications of Microbial Genome Tetranucleotide Frequency Biases.
521   Genome Res **13:**145-158.
522 52.   **Reva, O., and B. Tummler.** 2004. Global features of sequences of bacterial
523   chromosomes, plasmids and phages revealed by analysis of oligonucleotide usage
524   patterns. BMC Bioinformatics **5:**90.
525 53.   **Robinson, D. F., and L. R. Foulds.** 1981. Comparison of Phylogenetic Trees.
526   Mathematical Biosciences **53:**131-147.
527 54.   **Rocha, E. P. C., A. Viari, and A. Danchin.** 1998. Oligonucleotide bias in Bacillus
528   subtilis: General trends and taxonomic comparisons. Nucleic Acids Research
529   **26:**2971-2980.
530 55.   **Saitou, N., and M. Nei.** 1987. The neighbor-joining method: a new method for
531   reconstructing phylogenetic trees. Mol Biol Evol **4:**406-425.
532 56.   **Sanford, R. A., J. R. Cole, and J. M. Tiedje.** 2002. Characterization and Description of
533   Anaeromyxobacter dehalogenans gen. nov., sp. nov., an Aryl-Halorespiring Facultative
534   Anaerobic Myxobacterium. Appl. Environ. Microbiol. **68:**893-900.
535 57.   **SCHBATH, S., B. PRUM, and E. DE TURCKHEIM.** 1995. Exceptional Motifs in Different
536   Markov Chain Models for a Statistical Analysis of DNA Sequences. Journal of
537   Computational Biology **2:**417-437.
538 58.   **Schloss, P. D., and J. Handelsman.** 2004. Status of the Microbial Census. Microbiol.
539   Mol. Biol. Rev. **68:**686-691.
540 59.   **Schneiker, S., O. Perlova, O. Kaiser, K. Gerth, A. Alici, M. O. Altmeyer, D. Bartels, T.**
541   **Bekel, S. Beyer, E. Bode, H. B. Bode, C. J. Bolten, J. V. Choudhuri, S. Doss, Y. A.**
542   **Elnakady, B. Frank, L. Gaigalat, A. Goesmann, C. Groeger, F. Gross, L. Jelsbak, L.**
543   **Jelsbak, J. Kalinowski, C. Kegler, T. Knauber, S. Konietzny, M. Kopp, L. Krause, D.**
544   **Krug, B. Linke, T. Mahmud, R. Martinez-Arias, A. C. McHardy, M. Merai, F. Meyer, S.**
545   **Mormann, J. Munoz-Dorado, J. Perez, S. Pradella, S. Rachid, G. Raddatz, F.**
546   **Rosenau, C. Ruckert, F. Sasse, M. Scharfe, S. C. Schuster, G. Suen, A.**
547   **Treuner-Lange, G. J. Velicer, F.-J. Vorholter, K. J. Weissman, R. D. Welch, S. C.**
548   **Wenzel, D. E. Whitworth, S. Wilhelm, C. Wittmann, H. Blocker, A. Puhler, and R.**
549   **Muller.** 2007. Complete genome sequence of the myxobacterium Sorangium cellulosum.
550   Nat Biotech **25:**1281-1289.
551 60.   **Sharp, P. M., M. Stenico, J. F. Peden, and A. T. Lloyd.** 1993. Codon Usage -
552   Mutational Bias, Translational Selection, or Both. Biochemical Society Transactions
553   **21:**835-841.
554 61.   **Steel, M. A., and D. Penny.** 1993. Distributions of Tree Comparison Metrics--Some New
555   Results. Syst Biol **42:**126-141.
556 62.   **Suen, G., B. S. Goldman, and R. D. Welch.** 2007. Predicting Prokaryotic Ecological
557   Niches Using Genome Sequence Analysis. Plos One **2:**-.
558 63.   **Tekaia, F., E. Yeramian, and B. Dujon.** 2002. Amino acid composition of genomes,
559   lifestyles of organisms, and evolutionary trends: a global picture with correspondence
560   analysis. Gene **297:**51-60.
561 64.   **Welch, R. D., G. Suen, and B. S. Goldman.** 2007. Predicting Prokaryotic Ecological
562   Niches Using Genome Sequence Analysis. Plos One **2**.
563 65.   **Woese, C.** 1998. The universal ancestor. Proceedings of the National Academy of
564   Sciences of the United States of America **95:**6854-6859.
565 66.   **Woese, C. R., and G. E. Fox.** 1977. Phylogenetic structure of the prokaryotic domain:
566   the primary kingdoms. Proceedings of the National Academy of Sciences of the United
567   States of America **74:**5088-5090.

568

569

Figure 1

Percent symmetric difference of each order transition tree relative to the16S rRNA tree (left) and the zero-order transition tree (right).

The greatest change in symmetric difference between the 16S rRNA tree and that based on transition matrices occurs between 0th order and 3rd order, with only a very small change thereafter. Similarly, the greatest the symmetric difference between the 0th order transition tree and higher-order trees becomes relatively asymptotic after the 3rd order.
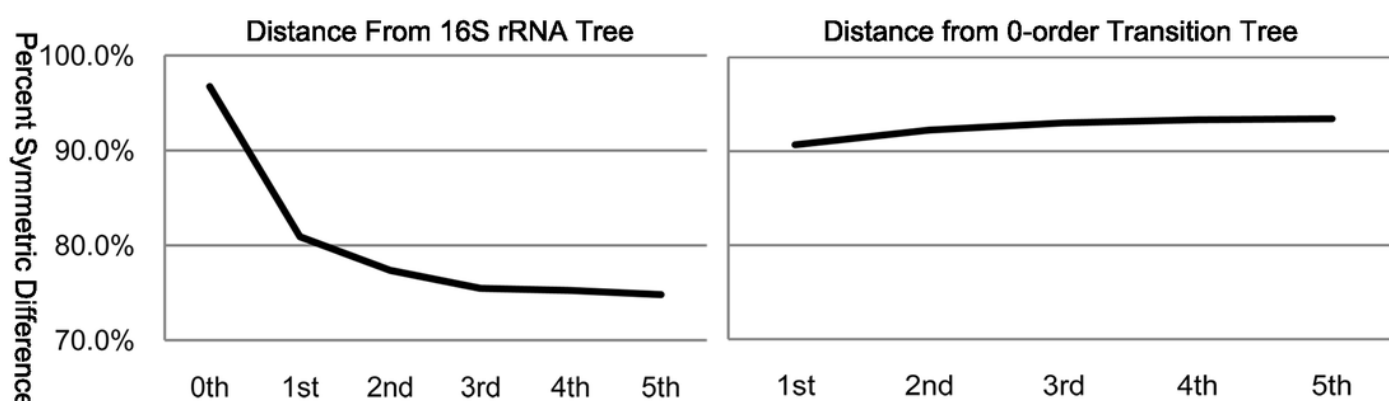
Figure 2

Percent symmetric difference between subsequent orders of transition trees.

T he symmetric difference between the subsequent order transition trees becomes relatively asymptotic after the 3rd | 4th order.
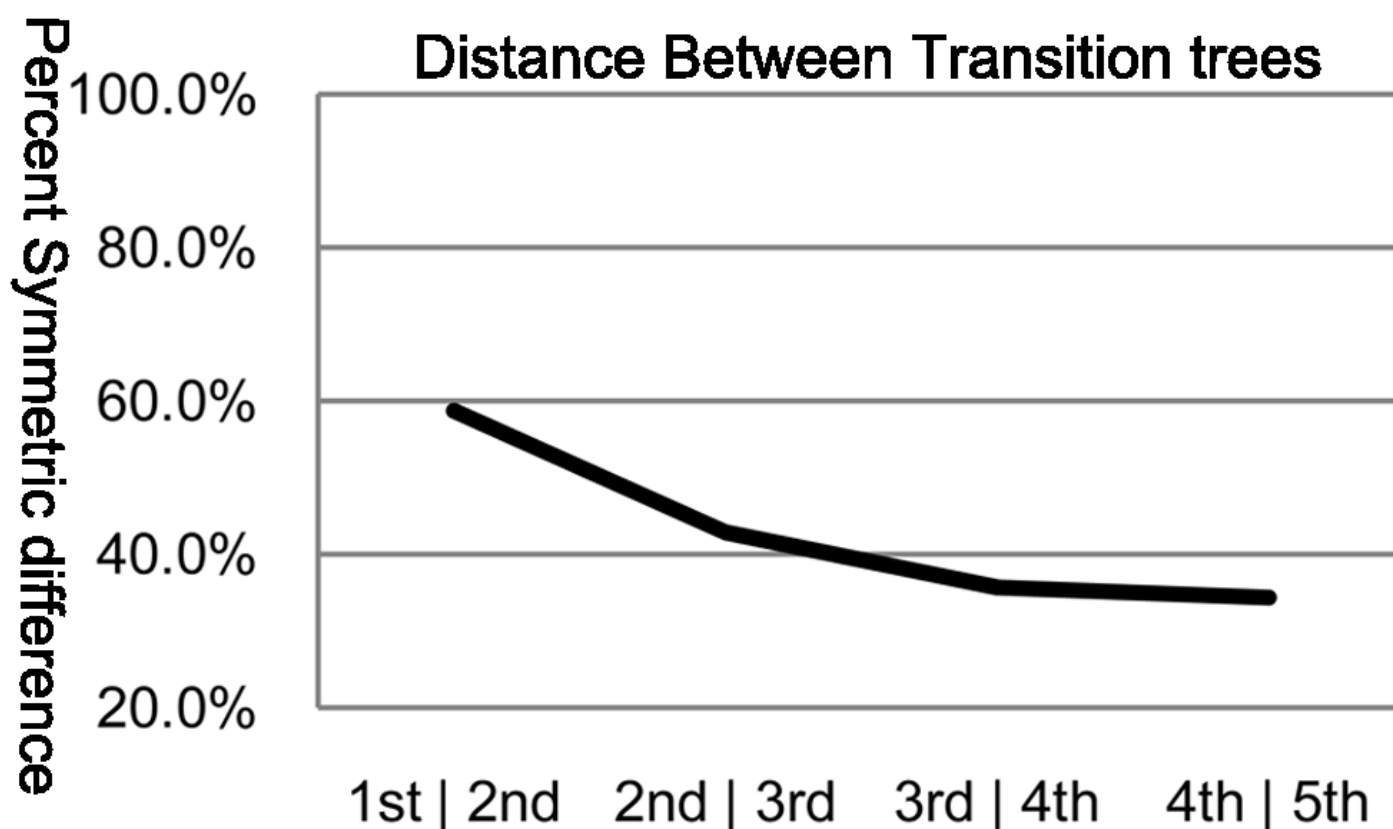
Figure 3

The symmetric difference between the 16S rRNA tree and the third-order transition tree.

Branches marked in red represent disagreement in topology between the trees.
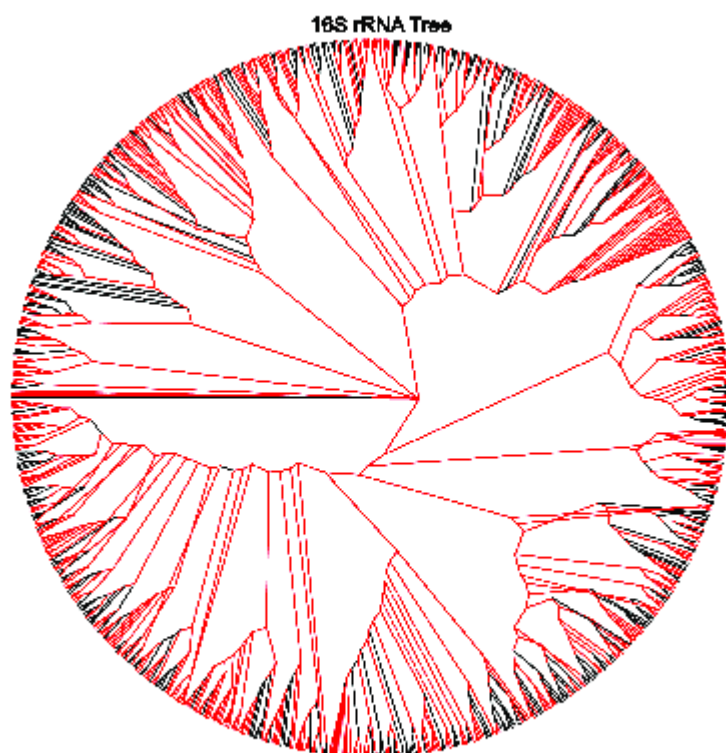
16S rRNA Tree

Figure 4

A collection of Enterobacteriaceae consisting of *Salmonella*, *Escherichia* and Shigella as example of taxa which cluster similarly in the 16SrRNA and third-order transition trees.

The genus of interest appear in red in the radial cladogram. A list of the organisms is given, with species that are not included in the transition tree, but are included in the 16S rRNA tree in boldface type. *A.macleodii__Deep_ecotype, H.baltica_ATCC_49814, I.loihiensis_L2TR, K.koreensis_DSM_16069,* **L.acidophilus_NCFM***, L.brevis_ATCC_367, L.casei_ATCC_334,* **L.delbrueckii_bulgaricus, L.delbrueckii_bulgaricus_ATCC_BAA-365, L.fermentum_IFO_3956, L.gasseri_ATCC_33323, L.helveticus_DPC_4571, L.johnsonii_FI9785, L.johnsonii_NCC_533***, L.plantarum, L.plantarum_JDM1,* **L.reuteri_DSM_20016, L.reuteri_F275_Kitasato***, L.rhamnosus_GG, L.rhamnosus_Lc_705,* **L.sakei_23K, L.salivarius_UCC118***, Marinomonas_MWYL1, M._mobilis_JLW8, P.profundum_SS9, P.necessarius_asymbioticus_QLW_P1DMWA_1, P.necessarius_STIR1, P.atlantica_T6c, P.haloplanktis_TAC125, P.arcticum_273-4, P.cryohalolentis_K5, Psychrobacter_PRwf-1, S.degradans_2-40,* **S.amazonensis_SB2B***, Shewanella_ANA-3, S.baltica_OS155, S.baltica_OS185, S.baltica_OS195, S.baltica_OS223, S.denitrificans_OS217, S.frigidimarina_NCIMB_400, S.halifaxensis_HAW_EB4,* **S.loihica_PV-4***, Shewanella_MR-4, Shewanella_MR-7, S.oneidensis, S.pealeana_ATCC_700345, S.piezotolerans_WP3, S.putrefaciens_CN-32, S.sediminis_HAW-EB3, Shewanella_W3-18-1, S.woodyi_ATCC_51908, T.crunogena_XCL-2,* **T.denitrificans_ATCC_33889***, V.cholerae, V.cholerae_M66_2, V.cholerae_MJ_1236, V.cholerae_O395, Vibrio_Ex25,* **V.fischeri_ES114***, V.harveyi_ATCC_BAA-1116, V.parahaemolyticus, V.splendidus_LGP32, V.vulnificus_CMCP6, V.vulnificus_YJ016, Y.enterocolitica_8081, Y.pestis_Angola, Y.pestis_Antiqua, Y.pestis_biovar_Microtus_91001, Y.pestis_CO92, Y.pestis_Nepal516, Y.pestis_Pestoides_F, Y.pseudotuberculosis_IP_31758, Y.pseudotuberculosis_IP32953, Y.pseudotuberculosis_PB1, Y.pseudotuberculosis_YPIII*

16S rRNA Tree

3rd-order Transition tree
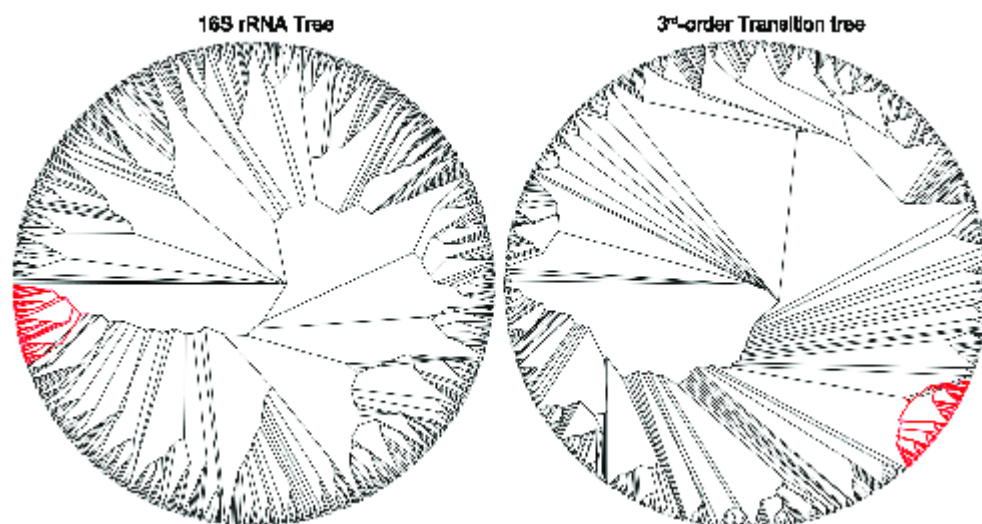
Figure 5

Genus Streptococcus appear in two distinct clusters in the third-order transition tree, but are assigned one cluster in the 16SrRNA tree.

The genus of interest appears in red in the radial cladogram. A list of the organisms is given. **Group 1** : *S.equi_4047, S.equi_zooepidemicus, S.equi_zooepidemicus_MGCS10565, S.gordonii_Challis_substr_CH1, S.sanguinis_SK36, S.pneumoniae_70585, S.pneumoniae_JJA, S.pneumoniae_D39, S.pneumoniae_R6, S.pneumoniae_P1031, S.pneumoniae_G54, S.pneumoniae_Taiwan19F_14, S.pneumoniae_ATCC_700669, S.pneumoniae_CGSP14, S.pneumoniae_Hungary19A_6, S.pneumoniae_TIGR4, S.suis_05ZYH33, S.suis_98HAH33, S.suis_SC84, S.suis_P1_7, S.suis_BM407* **Group 2:** *S. agalactiae_2603, S.agalactiae_NEM316, S.agalactiae_A909, S.dysgalactiae_equisimilis_GGS_124, S.pyogenes_M1_GAS, S.pyogenes_MGAS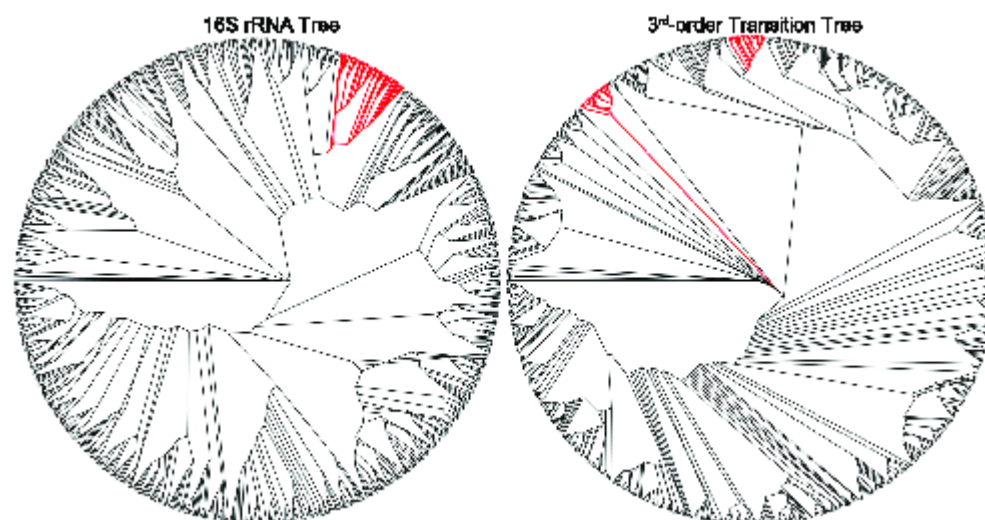9429, S.pyogenes_MGAS10270, S.pyogenes_NZ131, S.pyogenes_MGAS10750, S.pyogenes_MGAS10394, S.pyogenes_MGAS8232, S.pyogenes_MGAS315, S.pyogenes_MGAS5005, S.pyogenes_MGAS6180, S.pyogenes_MGAS2096, S.pyogenes_Manfredo, S.pyogenes_SSI-1, S.thermophilus_CNRZ1066, S.thermophilus_LMG_18311, S.thermophilus_LMD-9, S.uberis_0140J, S.mutans*

Figure 6

A group of mostly Aquatic Bacteria that cluster together in the third-order transition tree, but are dispersed in the 16S rRNA tree.

The genus of interest appear in red in the radial cladogram. A list of the organisms is given with those that appear outside the cluster in the transition tree in boldface type.

Shewanella_sediminis_HAW-EB3, Shewanella_woodyi_ATCC_51908,

Alteromonas_macleodii__Deep_ecotype_, Saccharophagus_degradans_2-40,

Pseudoalteromonas_haloplanktis_TAC125, Methylotenera_mobilis_JLW8,

Psychrobacter_arcticum_273-4, Psychrobacter_cryohalolentis_K5, Psychrobacter_PRwf-1,

Pseudoalteromonas_atlantica_T6c, Shewanella_ANA-3, Shewanella_MR-4, Shewanella_MR-7,

Shewanella_baltica_OS155, Shewanella_baltica_OS185, Shewanella_baltica_OS195,

Shewanella_baltica_OS223, Shewanella_oneidensis, Shewanella_putrefaciens_CN-32,

Shewanella_W3-18-1, Shewanella_denitrificans_OS217, Shewanella_halifaxensis_HAW_EB4,

Shewanella_pealeana_ATCC_700345, Shewanella_piezotolerans_WP3,

Shewanella_frigidimarina_NCIMB_400, Photobacterium_profundum_SS9, Vibrio_cholerae,

Vibrio_cholerae_M66_2, Vibrio_cholerae_O395, Vibrio_cholerae_MJ_1236,

Vibrio_vulnificus_CMCP6, Vibrio_vulnificus_YJ016, Vibrio_Ex25,

Vibrio_harveyi_ATCC_BAA-1116, Vibrio_parahaemolyticus, Vibrio_splendidus_LGP32,

Marinomonas_MWYL1, Hirschia_baltica_ATCC_49814,

Polynucleobacter_necessarius_asymbioticus_QLW_P1DMWA_1,

Polynucleobacter_necessarius_STIR1, Idiomarina_loihiensis_L2TR, Yersinia_enterocolitica_8081,

Yersinia_pestis_Angola, Yersinia_pestis_Nepal516, Yersinia_pestis_Antiqua,

Yersinia_pestis_biovar_Microtus_91001, Yersinia_pestis_CO92,

Yersinia_pseudotuberculosis_IP32953, Yersinia_pseudotuberculosis_PB1_,

Yersinia_pseudotuberculosis_IP_31758, Yersinia_pseudotuberculosis_YPIII,

Yersinia_pestis_Pestoides_F, Lactobacillus_brevis_ATCC_367, Lactobacillus_plantarum,

Lactobacillus_plantarum_JDM1, Lactobacillus_casei_ATCC_334, Lactobacillus_rhamnosus_GG,

Lactobacillus_rhamnosus_Lc_705, Kangiella_koreensis_DSM_16069,

Thiomicrospira_crunogena_XCL-2 **Vibrio_fischeri_ES114, Lactobacillus_sakei_23K,**

Lactobacillus_reuteri_DSM_20016, Shewanella_amazonensis_SB2B, Shewanella_loihica_PV-4,

Lactobacillus_delbrueckii_bulgaricus, Thiomicrospira_denitrificans_ATCC_33889,

Lactobacillus_acidophilus_NCFM