Preprint NCU/CCS-1998-1120; NSC-CTS-981101

Global Comparison of Thirteen Bacterial Complete Genomes and Phylogeny

Guo-An Wu[†], Zi-Hao Wang[†], H.C. Lee[†‡1] and Bai-lin Hao[†‡2]

[†] *Center for Complex Systems, National Central University, Chungli, Taiwan*
and
[‡] *National Center for Theoretical Sciences, P.O. Box 2-131, Hsinchu 300, Taiwan*

(December 28, 2000)

Abstract

A global comparison of the thirteen complete bactrial genomes available from the GenBank to May of 1998 is made by studying genome-genome correlations of distributions of frequency of occurrence of oligonuceotides (6-mers and 7-mers, respectively) tagged by under-represented palindromes (4-mers and 5-mers respectively) that are mostly recognition sites of bacterial restriction enzymes. Whereas frequency distributions of untagged oligonuceotides have little species dependence and correlations of such distributions show no structure, frequency distributions of the tagged oligonuceotides are strongly species dependent, and correlations of such distributions are highly structured and seem to yield useful phylogenetic information. In particular, correlations between species distant on the rRNA phylogenetic tree are observed, several of which corborrate results of recent analyses of gene homology suggesting the frequent ocurrence of lateral gene transfer between species.

[1]Send correspondence to: hclee@halley.phy.ncu.edu.tw
[2]On leave from The Institute of Theoretical Physics, Academia Sinica, Beijing 100080, China

1 Introduction

The accumulation of completely sequenced and annotated genomes in the past few years has, among many other things, already impacted [1] on our understanding of the phylogeny of bacteria [4, 5] based on the comparative analysis of the 16S rRNA sequences [2, 3]. In particular, apparent widespread and frequent occurrence of lateral gene transfers even between organisms that are thought to be phylogenetically far apart on the rRNA tree [7, 8, 9, 10] has raised the specter that the phylogeny of the earliest organisms may be more a network than a tree [11].

Since even the same sequences evolve at different rates in different lineages, phylogenies based on homologies of specific genes are bound to have discrepancies even in the absence of lateral gene transfers. Inconsistencies in a gene-based phylogenetic tree are thus impossible to avoid, and the best one can hope for is a consensus tree that admits the fewest number of conflicts. There is therefore motivation to look for alternative microbial phylogenetic methods that may provide additional information to gene-based analysis. Indeed, the number of possible trees containing more than a few species being huge - for thirteen species there are 13.75 billion possible rooted trees and 316.1 billion unrooted ones - and the correct tree being unique, independent information are always useful for pinning down the correct tree. A method most radically different from the single gene approach is to view each complete genome as a single entity and to make nongenic whole-genome comparisons. This may be more suitable for bacterial genomes, which have relatively few introns and practically no intergenic regions, than it would be for eukaryotes. The basis for the validity of such an approach is the assumption that, for species that are mutually not too distant, if two species are phylogenetically closer to each other than either is to a third species, then the genomes of the two species would be, on the whole, more similar to each other than either is to that of the third. In this work we look at a certain type of nongenic whole-genome comparison of thirteen complete bacterial genomes and show that it is a potential tool for studying relationship between bacterial species.

We view a genome as a long, continuous, unpunctuated text composed of the four letters A, C, G and T. Since the genomes are of differing lengths and are very long, with a typically length of one million letters, it is not useful to compare genomes letter by letter. A coarser-grain comparison is needed. A primal characterization of a long text is the frequencies of occurrence (hereafter, simply frequency) of words. Here we shall call an m-letter word an m-mer, and for the purpose of the our whole-genome comparison we shall pretend that we do not know the significance of any m-mer. (The only m-mers with definite and universal meanings are the 3-mers of the genetic code, but even they may be properly interpreted only when read relative to a correct starting position within a gene-coding frame.) Once a viable nongenic method is identified, gene related information can always be incorporated into it as an improvement or a refinement. The literature on statistical studies of m-mer frequencies and their possible biological significance is substantial, here we cite several that pertain to this work [27, 28, 12, 13, 14].

Since the number of m-mers grows rapidly with m, such that the set of frequencies for all m-mers up to a sufficiently large m would *uniquely* define a genome, comparing genomes frequency by frequnecy would be too fine grained and almost as useless as comparing them letter by letter. A very coarse grained measure of the difference between two genomes is the deviation, for a given m, of the distributions normalized to account for different sequence lengths, of the m-mer frequencies of two genome sequences. For reasons concerned with statistics and explained later, we concentrate on deviations for $m = 6$ and 7. In practice we used a *correlation* (see below for definition) defined to be proportional to the above mentioned deviation as a measure of the similarity between two species. It turned out that such correlations computed from 6-mer and 7-mer frequencies were uninteresting from two aspects: species-species correlations varied weakly, and the correlations were in general much weaker than those between random distributions of m-mer frequencies.

The situation changes drastically for correlations between distributions of frequencies of 6-mers and 7-mers tagged by under-represented palindromes. Recall n by recalling that the most under-represented 4-mers in the complete bacterial genomes are invariably tetrapalindromes; some of these are common to several genomes; the most under-represented 5-mers either contain an under-represented 4-mer or are pentapalidromes themselves; there is a close connection of the under-represented palindromic 4-mers and 5-mers with bacterial restriction enzymes [12, 13, 14]. We extend previous work to cover the larger number of complete genomes now available and find results, where comparable, in substantial agreement with results previously obtained. The fact that the rarity of oligopalidromes is a shared trait of bacterial genomes make them potential evolutionary markers.

In order to exploit these potential markers we look at the *correlations* of the bacterial frequency distributions of the oligonucleotides tagged by rare palindromes. When considering distributions it is necessary to have a sufficient sample size. There are 4096, 16762 and 67038 different 6-mers, 7-mers and 8-mers, respectively. Taking into account that the length of the genomes are of the order of one million bases, so that the average frequencies of 6-mers and 7-mers are about 244 and 61, respectively, while an 8-mers would on average have only a frequency of 15, and that there are 48 different 6-mers tagged by a 4-mer, and 48 different 7-mers tagged by a 5-mer, we chose to look at the frequency distributions of these two kinds of tagged oligonucleotides and do not consider 8-mers in this work. It turns out that whereas the frequency distributions of untagged 6-mers and 7-mers show little variation from species to species and are essentially uncorrelated pairwise, the frequency distributions of the tagged oligonucleotides exhibit high species dependence and their pairwise correlations span a wide range from very strong to very weak.

2 Materials and Method

Nomenclature. A genome sequence is composed of the four letters A, C, G, and T, representing the

nucleotides adenine, cytosine, guanine, and thymine, respectively. A string of length m, referred to as m-mers hereafter, is a set of m contiguous letters in the sequence. A tagged m-mer is a string containing a specific (shorter) substring. For example, $AACTAG$, $GCTAGC$ and $CTAGGG$ are all 6-mers tagged by the 4-mer $CTAG$. Although three of the species studied belong to the *archaea* kingdom, unless clear distinction demands otherwise, for convenience we shall address all the thirteen species as bacteria.

Data. Data include the thirteen complete bacterial genomes in GenBank as of May of 1998: *Haemophilus influenzae (H. influenzae)* (1.83 Mb) [15], *Mycoplasma genitalium (M. genitalium)* (0.580 Mb) [16], *Synechocystis PCC6803 (Synechocystis)* (3.57 Mb) [17], *Methanococcus jannaschii (M. jannaschii)* (1.66 Mb) [18], *Mycoplasma pneumoniae (M. pneumoniae)* (0.816 Mb) [19], *Rhizobium sp. NGR234 (Rhizobium sp.)* (0.536 Mb) [20], *Helicobater pylori (H. pylori)* (1.67 Mb) [21], *Escherichia coli (E. coli)* (4.64 Mb) [22], *Methanobacterium thermoautotrophicum (M. ther)* (1.75 Mb) [23], *Bacillus subtilis (B. subtilis)* (4.21 Mb) [24], *Archaeoglobus fulgidus (A. fulgidus)* (2.18 Mb) [25], *Borrelia burgdorferi (B. burgdorferi)* (0.911 Mb) [26], and *Aquifex aeolicus (A. aeolicus)* (1.55 Mb) [7].

Monomer frequency, random sequences and normalization. The symbol x is used to generically denote any one of the four letters A, C, G, and T; w to denote A or T; s to denote C or G. The percentage monomer frequencies of x in a genome is denote by p_x. In all genomes the relations $p_A \cong p_T$ and $p_G \cong p_C$ hold to within two percent. The doublet $d = (p_A,\ p_C)$ of the genomes distinctively falls into four types: type I (*A. fulgidus*, *M. thermoautotrophicum*, *E. coli*, *Synechocystis*) with $d \cong (25\%,\ 25\%)$, type II (*M. jannaschii*, *M. genitalium*, *B. burgdorferi*) with $d \cong (35\%,\ 15\%)$, type $IIIa$ (*B. subtilis*, *M. pneumoniae*, *H. influenzae*, *H. pylori*, *A. aeolicus*) with $d \cong (30\%,\ 20\%)$ and type $IIIb$ (*Rhizobium sp.*) with $d = (20\%,\ 30\%)$. For each type (type $IIIa$ and $IIIb$ are represented by a single type III) a million-word long reference random sequence is generated. Let $\sigma = \{x_1 x_2 \cdots x_m\}$ denote an m-mer and $f(\sigma)$ denote the observed frequency of σ in a sequence of length $N >> m$. Since the length and monomer distribution of the genomes vary considerably, for comparison purposes we scale $f(\sigma)$ to a sequence length of one million and remove its dependence on the monomer distribution to first order by defining the normalized frequency

$$\bar{f}(\sigma) = f(\sigma) \frac{10^6\ 4^{-m}}{N\ p_A^{n_A} p_C^{n_C} p_G^{n_G} p_T^{n_T}} \tag{1}$$

where n_x is the frequency of x in the word σ with $n_A + n_C + n_G + n_T = m$.

Mean and deviation. The mean value of the frequencies of the set $\mathcal{S}$ of $N_{\mathcal{S}}$ m-mers is

$$\langle \bar{f} \rangle_{\mathcal{S}} = N_{\mathcal{S}}^{-1} \sum_{\sigma \in \mathcal{S}} \bar{f}(\sigma) \tag{2}$$

and the root-mean-square (rms) of the deviation from this mean is

$$\left(\Delta \bar{f}\right)_{\mathcal{S}} = \left(N_{\mathcal{S}}^{-1} \sum_{\sigma \in \mathcal{S}} \left(\bar{f}(\sigma) - \langle \bar{f} \rangle_{\mathcal{S}} \right)^2 \right)^{1/2} \tag{3}$$

When $\mathcal{S}$ is the set $\{\sigma\}_m$ of all m-mers, $N_{\mathcal{S}} = 4^m$. For a normalized random sequence $\langle \bar{f} \rangle_{random} = 10^6\ 4^{-m}$.

Relative abundance. An m-mer σ has substrings $\sigma_{(1,0)} = \{x_2 \cdots x_m\}$, $\sigma_{(0,1)} = \{x_1 x_2 \cdots x_{m-1}\}$ and $\sigma_{(1,1)} = \{x_2 \cdots x_{m-1}\}$, etc. Based on estimations such as

$$\bar{f}(\sigma) \approx \bar{f}(\sigma_{(0,1)}) \bar{f}(x_m) \approx \bar{f}(\sigma_{(1,0)}) \bar{f}_{x_1} \approx \bar{f}(\sigma_{(2,0)}) \bar{f}(\{x_1 x_2\}) \approx \bar{f}(\sigma_{(0,2)}) \bar{f}(\{x_{m-1} x_m\}) \tag{4}$$

we define a relative abundance:

$$\tau^{(2)}(\sigma) = \frac{\bar{f}(\sigma) \bar{f}(\sigma_{(1,1)})}{\bar{f}(\sigma_{(0,1)}) \bar{f}(\sigma_{(1,0)})} \tag{5}$$

Although more elaborate definitions of relative abundance may be devised and used [13, 14], the one given here is sufficient for our purpose. Strings that contain a substring with a zero observed frequency are not considered in our analysis. Note that f and $\bar{f}$ have the same relative abundance.

Moments of frequency distribution. Given a distribution $D(y)$, the rank-k moment of the distribution is

$$\langle y^k \rangle = \frac{\int dy y^k D(y)}{\int dy D(y)} \tag{6}$$

The rank-k moment of the distribution $\bar{f}(\sigma)$ is

$$\langle \bar{f}^k \rangle_{\mathcal{S}} = N_{\mathcal{S}}^{-1} \sum_{\sigma \in \mathcal{S}} \left(\bar{f}(\sigma) \right)^k \tag{7}$$

Gamma and inverse-gaussian distributions. The distribution of frequencies of m-mers are well approximated by two two-parameter distributions with exponential tails: the gamma distribution

$$\Gamma(y; \alpha; \beta) = \frac{1}{\Gamma(\alpha)\beta^\alpha} y^{\alpha-1} \exp\left(-\frac{y}{\beta}\right) \qquad 0 \le y < \infty, \tag{8}$$

that has mean $\alpha\beta$ and mean-square deviation $\alpha\beta^2$, and the inverse-gaussian distribution

$$D(y; \mu; \lambda) = \sqrt{\frac{\lambda}{2\pi}} y^{-3/2} \exp\left(-\frac{\lambda(y-\mu)^2}{2\mu^2 y}\right), \qquad 0 \le y < \infty, \tag{9}$$

that has mean μ and mean-square deviation μ^3/λ. If $\Gamma(y; \alpha; \beta)$ and $D(y; \mu; \lambda)$ have the same first and second moments, then $\alpha = \lambda/\mu,\ \beta = \mu^2/\lambda$, or $\lambda = \alpha^2\beta,\ \mu = \alpha\beta$.

χ^2 difference and correlation between two distributions. Given two frequency distributions of m-mers $\bar{f}^a(\sigma)$ and $\bar{f}^b(\sigma)$, the χ^2 difference is

$$\chi^2(\bar{f}; a, b) = N_{\mathcal{S}}^{-1} \sum_{\sigma \in \mathcal{S}} \left(\bar{f}^a(\sigma) - \bar{f}^b(\sigma) \right)^2 \tag{10}$$

In this paper we define the correlation between distributions a and b as

$$\rho(\bar{f}; a, b) = \frac{\langle \bar{f} \rangle^2}{\chi^2(\bar{f}; a, b)} \tag{11}$$

where $\langle \bar{f} \rangle$ is some mean distribution independent of a and b.

3 Results

Under-represented Tetrapalindromes. We say a tetranucleotide σ is under-represented if it has a relative abundance that is less than unity, and *significantly* under-represented if $\bar{f}(\sigma)$ is less than 975 (a fourth of the mean frequency) and $\tau(\sigma) \le 0.95$, or if $\bar{f}(\sigma)$ is less than 1300 and $\tau(\sigma) \le 0.60$. It has been noticed that palindromes of 4-6 nucleotides that are recognition sites of type II restriction enzymes are under-represented in *E. coli* and some other bacterial DNA [14]; in particular, that $CTAG$ is so has been known for some time [22, 27, 28, 13]. Of the 256 tetranucleotides, sixteen are palindromic and all are recognition sites. Not including *M. thermoautotrophicum*, the other twelve complete

genomes have on average only 2.8 ± 1.7 palindromic tetranucleotides that are not under-represented. In *M. thermoautotrophicum*, which is highly exceptional in this aspect, ten palindromic tetranucleotides are not under-represented. Table 1 lists the palindromic tetranucleotides in the complete genomes that are significantly under-represented. The list of under-represented tetranucleotides in [13], where a relative abundance more elaborately defined than $\tau^{(2)}$ is used, are also given. For those under-represented tetranucleotides that are identified in both lists, either the observed frequency or the relative abundance is exceptionally low, or both are at least moderately low. For those under-represented tetranucleotides that are identified in either one of the lists but not both, only one of the abundance quantities are low. The minor differences in the two lists, where they occur, are explained by the fact that we examine normalized frequency while [13] looks at observed frequency and that differently defined relative abundances are used in the two studies. The most exceptionally under-represented tetrapalindromes are $CTAG$, which appears in *M. jannaschii* only 73 times with $\tau^{(2)} = 0.027$ and $GTAC$ which appears in *H. pylori* only 81 times with $\tau^{(2)} = 0.10$. The six most prominent rare tetrapalindromes, with the genomes in which each is under-represented given in brackets, are: $CTAG$ (*A. fulgidus*, *M. jannaschii*, *M. thermoautotrophicum*, *A. aeolicus*, *B. subtilis*, *Rhizobium sp.*, *E. coli*, *H. influenzae*); $CGCG$ (*A. fulgidus*, *M. jannaschii*, *M. thermoautotrophicum*, *A. aeolicus*, *Synechocystis*, *B. burgdorferi*); $GCGC$ (*A. fulgidus*, *M. jannaschii*, *M. thermoautotrophicum*, *A. aeolicus*); $TCGA$ (*M. jannaschii*, *M. genitalium*, *H. pylori*, *A. aeolicus*, *B. burgdorferi*); $GATC$ (*A. fulgidus*, *M. jannaschii*, *A. aeolicus*); $GTAC$ (*M. jannaschii*, *H. pylori*, *B. burgdorferi*). We infer from Table 1 a strong relationship among *A. fulgidus*, *M. jannaschii*, *M. thermoautotrophicum* and *A. aeolicus*. The first three belong to the Euryarchaeota phylum in the archaea subkingdom, but the last is in the Aquificales phylum in the Eubacteria subkingdom. The two phyla share their most important phenotypes that are also believed to be characteristic of the earliest organisms: tolerance of high-temperature habitats and lack of ability to grow on sugar or protein.

Under-represented Pentanucleotides. We say a pentanucleotide σ is significantly under-represented if the normalized frequency (see Eq. (1)) $\bar{f}(\sigma)$ is less than a fourth of the mean frequency and the relative abundance (see Eq. (5)) $\tau(\sigma) \leq 1$ or if $\bar{f}(\sigma)$ is less than a third of the mean frequency and $\tau(\sigma) \leq 0.7$. We shall refer to 5-mer as a pentapalindrome if it becomes a tetrapalindrome when its middle letter is removed. An example is $GGwCC$. The first thing one notices about under-represented pentanucleotides in the thirteen complete genomes is the preponderance of those containing a tetrapalindrome as a substring. Table 2 summarized the result on significantly under-represented pentanucleotides. The ratio of all pentanucleotides to pentanucleotides with tetrapalindromes as substrings to pentapalindromes is $16 : 2 : 1$. For the under-represented pentanucleotides in Table 2 the ratio is $16 : 5.2 : 2.4$; with *H. pylori* excluded the ratio is $16 : 5.0 : 1.9$; for *H. pylori* alone the ratio is $16 : 5.9 : 5.9$. Thus across all species pentapalindromes are indeed under-represented. They are highly under-represented in *H. pylori*, which also has by far the largest number of under-represented pentapalindromes - 14 compared with an average 2.2 ± 1.6 over the other species. Moreover, it also has the

six pentanucleotides with the lowest relative abundance among all the under-represented pentanucleotides for all the thirteen species. It is not understood why *H. pylori* should be so peculiar in this respect.

The 41 entries of under-represented pentapalindromes in Table 2 are not equally spread among the 16 possible pentapalindromes: $GGwCC$ and $CGwCG$ each appears eight times (as four pairs), $GTsAC$ six times and $TCsGA$ and $CTwAG$ four times each. Table 2 also shows that $GGwCC$ is the only pair that appears more than once (three times) with especially low relative abundance. We thus conclude that given in order of under-representation, $GGwCC$, $CGwCG$, $GTsAC$, $TCsGA$ and $CTwAG$ are the most noticeable under-represented pentapalindromes. In contrast, the pentapalindromes $AAxTT$, $ATxAT$, $CCxGG$, $GCxGC$, $TAxTA$ and $TTxAA$ are *never* under-represented.

Frequency distributions are approximated by gamma and inverse-gaussian distributions. The 4096 6-mers form a large enough pool for statistical analysis. Fig. 1 are plots of the number of 6-mers versus frequency $f(\sigma)$ (normalized to a total length of 1 million but *not* corrected for different mononucleotide compositions) for the three random and thirteen bacterial sequences. It is immediately noticed that whereas the distribution of the type I random sequence (Fig. 1(d)) is normal (i.e. gaussian) and centered at the mean frequency of 244, the distributions from the bacterial sequences have the characteristics of the two-parameter gamma and inverse-gaussian distributions (see Eq.(8) and Eq.(9)), both of which rise from zero at the origin, reach an asymmetrical peak, then decay with an exponential tail.

That this is so may be partially understood by comparing the distribution for the type I random sequence, which is G+C neutral, with the distribution for the type II (Fig. 1(e)) or III (Fig. 1(f)) random sequence, which are significantly (G+C)-poor. The latter distribution is composed of seven subdistributions, respectively for the subset of 6-mers containing no G or C, one G or C, and so on up to six G's and/or C's. Thus the subdistribution sitting at the high-frequency end is of the subset of 6-mers with zero G+C content, and the subdistribution sitting nearest zero frequency is of the subset of 6-mers entirely composed of G's and/or C's. In each case, the relative size of the subdistribution is roughly determined by the coefficients of the binomial expansion. This dependence in monomer composition is reflected in the frequency distributions of bacterial genomes, as is seen when we compare the distributions of, say, *A. fulgidus* with type I random sequence, *M. jannaschii* with type II random sequence, and *Rhizobium sp.* with type III random sequence in Fig. 1.

Figs. 2(a-c) show the frequency distributions of normalized $\bar{f}$ (see Eq.(1)) of 6-mers in *A. fulgidus*, *M. jannaschii* and *Rhizobium sp.*. There is no longer a pronounced dependence on monomer composition. In particular the seven separate subdistributions in both the type II and III random sequences (Fig. 1(e, f)) are now consolidated to a single gamma distribution (Fig. 2(d)). It happens that the bacterial distributions are well approximated by gamma and inverse-gaussian distributions, presumably due to further biases in the frequencies of dimers, trimers, etc. The parameters of a distribution α and β in the case of gamma distribution, and μ and λ in the case of inverse-gaussian distribution, that best represent a bacterial distribution are determined by the mean

and deviation of the latter. The parameters thus determined for the thirteen bacterial distributions are given in Table 3. In Fig. 2(a-c), the gamma (green line) and inverse-gaussian (red line) distributions that best fit the distributions derived from *A. fulgidus*, *M. jannaschii* and *Rhizobium sp.* are shown. Later we shall note that the gamma distribution gives somewhat better approximation to data. For the moment, because all the bacterial distributions have a mean of about 244, and this fixes the parameter μ for the inverse-gaussian distribution so that differences in bacterial distributions are expressed in the variation only in the parameter λ, we shall discuss the bacterial distributions in terms of their inverse-gaussian representations. Table 3 shows that the bacterial distributions fall roughly into three groups respectively having the ranges of values of λ: $330 \leq \lambda \leq 450$; $530 \leq \lambda \leq 670$; $750 \leq \lambda \leq 920$. Recall that λ is inversely proportional to the mean-square deviation of the distribution. Hence the distribution from *H. pylori*, with the longest tail - a reflection of the fact that *H. pylori* has the largest number of 6-mers with far above average frequencies, has a λ value of 328, and that from *Rhizobium sp.*, with the shortest tail, has a λ value of 921. Later we will examine the correlations between bacterial distributions and for the purpose of establishing a norm for such correlations we generate three random inverse-gaussian distributions with $\mu = 244$ and $\lambda =$ 300, 600 and 900, which we call random distribution I, II and III, respectively. Two of these are shown in Fig. 2 ((e) and (f)).

Given that both the gamma and inverse-gaussian distributions are able to reproduce exactly the mean and deviation of a set of data, the better distribution is the one which gives better fits to the third and higher moments of the data. Table 4 compares the observed moments (Eqs.(6) and (7)) and those computed from gamma distribution and inverse-gaussian distribution fits to several representative bacterial 6-mer frequency distributions. In every case the gamma distribution gives the better values for the higher moments. In fact it is quite remarkable how well the gamma distributions agree with data, even up to the sixth moment. The reason for the superiority of the gamma over the inverse-gaussian distribution is not understood.

Much of the discussion on 6-mer frequency distributions applies as well to 7-mer frequency distributions. Parameters for the gamma and inverse-gaussian fits to all the bacterial 7-mer frequency distributions and for the three random distributions are given in Table 5.

Correlation of frequency distributions of 6-mers and 7-mers. The *correlation* in two distributions, defined in Eq.(11), gives a measure of the similarity between two genomes. For $\langle \bar{f} \rangle$ we use the mean frequency in a sequence normalized to a length of one million bases, namely $\langle \bar{f} \rangle = 244$ for 6-mers and $\langle \bar{f} \rangle = 61$ for 7-mers. Pair-wise correlations of 6- and 7-mer frequency distributions of the thirteen bacteria and the three random inverse-gaussian distributions are color-coded and displayed in the two triangular lattices shown in Fig. 3. The coordinates of the lattice, counting from top to bottom for the ordinate and right to left for the abscissa, are the bacterial and random distributions represented by numerals indicated in the figure. Each lattice site gives the color-coded value of the correlation between the pair of distributions represented by the coordinates of the lattice site. Black and purple indicate the strongest correlation and

yellow and white indicate the opposite. A rather surprising result is that correlations among the random distributions have values that are on average one order of magnitude greater than correlations among bacterial frequency distributions and those between a bacterial distribution and a random distribution. We have verified that this feature is not particular to any specific set of random distributions; changing random distributions (but not the parameters that characterize them) does not alter the global color patterns of the plots. The cause of this effect may be understood by remembering that the random distributions are true inverse-gaussian distributions, and as such do not have exceptionally high frequency components as the bacterial distributions do, and that the χ^2 deviation is heavily weighed by high-frequency components in two distributions that do not match. The absence of even moderate correlation between any pair of bacterial frequency distributions suggests such distributions occupy the space of frequencies with a very high degree of ergodicity. In any case, the 6-mer and 7-mer frequency distributions, being insufficiently genome specific for whole-genome comparisons, are clearly not useful targets for revealing phylogenetic information.

Correlation from distributions of tetrapalindrome-tagged 6-mers. It was seen that the frequencies of tetrapalindromes and pentapalindromes were very species specific. We therefore examine correlations in frequency distributions in subsets of 6-mers *tagged* by specific tetrapalindromes and, in the next section, 7-mers tagged by specific pentapalindromes. The set of 6-mers tagged by, say, $CTAG$ is composed of the forty-eight 6-mers that contain $CTAG$ as a substring. Figs. 4, and 5 show distributions of 6-mers tagged by $CTAG$ and $TATA$, respectively, two of the most prominent rare tetrapalindromes. The distributions have clear and strong species dependence. In species where the tag is under-represented, the frequencies are - as expected - bunched below the average 6-mer frequency. Otherwise the frequency distributions are essentially coarse-grained approximations of gamma and inverse-gaussian distributions. Not shown are frequency distributions of 6-mers tagged by tetranucleotides such as $AATT$ and by $TGCA$ that are never under-represented and by non-palindromic 4-mers such as $GGTT$. In these distributions species dependence is absent and all are essentially coarse-grained gamma or inverse-gaussian distributions. From what was seen in the last section, we expect that whereas correlation in the frequency distributions of 6-mers of two genome sequences both under-represented in the tag may or may not be strong (relative to correlation in random distributions), correlation will be weak if the tag is under-represented in no more than one of the two sequences.

Color-coded correlations computed from distributions of tagged 6-mers are given in the sixteen triangular lattices in Fig. 6, each lattice for a specific tetrapalindromic tag. The convention is the same as in Fig. 3 where darker color indicates stronger correlation. From these plots can be read off species that are strongly correlated and the tetrapalindrome through which this correlation is expressed. Below is a summary of the most strongly correlated pairs of species (black and purple in Fig. 6) and the tag.

Tag: $CTAG$; pairs: *A. fulgidus* with *M. jannaschii*, *M. thermoautotrophicum*, *B. subtilis*, *E. coli* and *A. aeolicus*; *M. jannaschii* with *M. thermoautotrophicum* and *E. coli*; *M. thermoautotrophicum* with *B. subtilis*, *Rhizobium sp.*, *E. coli* and

A. aeolicus; *B. subtilis* with *Rhizobium sp.*, ecoli, *H. influenzae* and *A. aeolicus*; *Rhizobium sp.* with *H. influenzae* and *A. aeolicus*; *E. coli*-*A. aeolicus*.

Tag: $TATA$; pairs: *M. genitalium* with *M. pneumoniae* and *Synechocystis*; *H. influenzae*-*Synechocystis*; *M. thermoautotrophicum* with *H. influenzae*, *Synechocystis* and *B. burgdorferi*.

Tag: $CGCG$; pairs: *Synechocystis* with *A. fulgidus*, *M. jannaschii* and *M. thermoautotrophicum*; *M. thermoautotrophicum* with *A. fulgidus* and *M. jannaschii*.

Tag: $TCGA$; pairs: *M. genitalium* with *M. jannaschii* and *H. pylori*; *H. influenzae*-*Synechocystis*.

Tag: $GTAC$; pair: *M. jannaschii*-*H. pylori*.

Tag: $ACGT$; pair: *M. jannaschii*-*B. burgdorferi*.

Distributions of frequencies of 6-mers tagged with the palindromes $AATT$, $AGCT$, $CATG$, $CCGG$, $GATC$, $GGCC$, $TGCA$ and $TTAA$ are uncorrelated.

Although most bacterial sequences involved in moderate or stronger correlations have the corresponding tetranucleotides under-represented, there are exceptions. *H. influenzae* and *Synechocystis* are exceptionally strongly or strongly correlated through $TATA$ and $TCGA$ but are barely or not at all under-represented in these two tetranucleotides; see Table 1. Generally, if two bacterial sequences are both under-represented in a tetrapalindrome, then the frequency distributions of their respective tetranucleotide-tagged 6-mers would be at least moderately correlated. Two noticeable exceptions are the pair *H. pylori* and *B. burgdorferi*, both very rare in $TCGA$, and the triplet *A. fulgidus*, *M. jannaschii* and *A. aeolicus*, all rare or moderately rare in $GATC$, whose corresponding frequency distributions of tagged 6-mers are not correlated. Thus under-representation is neither necessary nor sufficient condition for strong correlation.

Correlation from distributions of pentapalindrome-tagged 7-mers. The plots in Fig. 7 give the color-coded correlations of frequency distributions of tagged 7-mers between pairs of distributions from the thirteen bacterial genomes and the three random distributions. The tags include the 13 pentapalindromes appearing in Table 2 and, for control, the two palindromes $CCAGG$ and $GCAGC$ that do not appear there and the non-palindrome $GAACT$.

From Fig. 7 it is seen that correlations among the random frequency distributions of 7-mers are generally not as strong as their counterparts for 6-mers. This is understood as follows. The most under-represented 7-mers have tetrapalindromes as substrings, not pentapalindromes. For 7-mers we do not use tetrapalindrome as tags for the reason that correlations thus obtained would not be information independent from correlations obtained from tagged 6-mers. Compared with correlations of frequency distributions of tagged 6-mers, high correlation between distributions of pentapalindrome-tagged 7-mers is even less predictable on the bases on the under-representation of pentanucleotides. In the strong and moderately correlated pairs *H. influenzae* and *H. pylori* ($GGACC$; black), *H. influenzae* and *M. genitalium* ($TCGGA$; purple), *M. jannaschii* and *M. pneumoniae* ($CGACG$; red) and *M. thermoautotrophicum* and *E. coli* ($CTAAG$; red), the tag is

under-represented in both bacterial sequences, whereas in the pairs *A. fulgidus* and *A. aeolicus* ($GTGAC$; purple), *M. genitalium* and *H. influenzae* ($AGACT$; purple), *M. pneumoniae* and *Synechocystis* ($AGACT$; purple), *B. subtilis* and *E. coli* ($GAGTC$; purple), *E. coli* and *H. influenzae* ($TCGGA$; purple), *M. thermoautotrophicum* and *M. pneumoniae* ($CGACG$; red), *M. genitalium* and *M. pneumoniae* ($CGACG$; red), *B. subtilis* and *Rhizobium sp.* ($CTAAG$; red) and *Rhizobium sp.* and *E. coli* ($CTAAG$; red) the tag is at most under-represented in one of the pair and is often not under-represented in either. There are also cases where moderate correlation is absent between the two bacterial distributions when the tag is under-represented in both bacterial sequences. For example, $GGACC$ in under-represented in *E. coli*, *H. influenzae* and *H. pylori* but the $GGACC$-tagged 7-mer distribution of *E. coli* is not strongly correlated with those of the other two bacteria.

Combined correlation from distributions of tagged 6-mers and 7-mers. The data shown in Figs. 6 and 7 suggest that we take as a background for correlation the value 200, which is the typical correlation between random distributions. Fig. 8 shows the color-coded, summed, above background values (in units of 100) of the thirty-two correlations shown in Figs. 6 and 7. Note that correlations that are shown in white, yellow, light-blue and green in Figs. 6 and 7 do not contribute to Fig. 8. The contrast between the correlation plot in Fig. 8 and the plots in Fig. 3 is striking. In Fig. 3, species-species correlations show little variation and are up to an order of magnitude weaker than random-random correlations, while the exact opposite is true in Fig. 8. In the latter, since all correlations in which at least a random sequence is involved are in the three weakest colors - white, yellow and light-green - we may say any genome-genome correlation that is colored in, purple dark blue or black is statistically significant.

It is reassuring that among the pairs of bacteria that exhibit significant correlations are the two closest pairs on the rRNA phylogenetic tree [4]: *M. jannaschii-M. thermoautotrophicum* and *M. pneumoniae-M. genitalium*. As far as conforming to this phylogenetic tree is concerned, the three *Euryarchaeotas* are strongly correlated, and the four *Proteobacterials* have four moderate correlations: *H. influenzae-H. pylori*, *E. coli-H. influenzae* *E. coli-Rhizobium sp.* and *Rhizobium sp.-H. influenzae*. On the other hand, *B. subtilis* shows no correlation with its distant fellow *Firmicuteons*, *M. genitalium* and *M. pneumoniae* and *H. pylori* shows no correlation with its distant fellow *Proteobacterials*, *Rhizobium sp.* and *E. coli*.

Of correlations not expected from the rRNA phylogenetic tree, the very strong ones are *E. coli* with the three *Euryarchaeotas*, *M. jannaschii-H. pylori*, *B. subtilis-Rhizobium sp.* and *H. influenzae-Synechocystis*, and the moderate ones are *A. aeolicus* with *A. fulgidus*, *M. thermoautotrophicum* and *B. subtilis*, *B. subtilis-E. coli*, *M. jannaschii-B. burgdorferi* and *M. thermoautotrophicum-Synechocystis*.

As a group the three *Euryarchaeotas* are the most correlated with other genomes; of the total of 17 significant correlations they are involved in 11. In comparison, the four *Proteobacterials* are involved in 8, and the three *Firmicuteons* in 4, three of which involve *B. subtilis*. Interestingly, there is not a single significant correlation between the *Euryarchaeotas* and the *Firmicuteons*. As a group the two *Firmicuteons*, *M. pneumo-*

niae and *M. genitalium*, are the least correlated; they are strongly correlated with each other but with no other genome. Among all genomes *M. jannaschii* with five significant correlations is the most connected genome. Next come *A. fulgidus*, *M. thermoautotrophicum* and *E. coli* with four significant correlations each, and *B. subtilis* and *A. aeolicus* with three each. These six include all the four thermophiles in the thirteen genomes, the three *Euryarchaeotas* plus *A. aeolicus*.

Some aspects of the correlation seen in Fig. 8 between distant genomes (according to rRNA phylogeny) are corroborated in recent analyses on gene homology. Analysis of genes encoding the protein FtsY and the enzyme for the synthesis of tryptophan link *A. aeolicus* with *B. subtilis* and with Archaea, respectively [29]. The class I-type lysyl-tRNA synthetase, previously known to be contained only in some archaeons including *M. jannaschii* but not seen in bacteria and eukarya has now been identified in *B. burgdorferi* [30]. Comparison of genes encoding trisephosephate isomerase (TPI) show that eukaryote TPI genes are most closely related to the homologue from the α-proteobacterial genomes (of which *B. subtilis* is one) and most distantly related to archaeal homologues [6] which suggests that *B. subtilis* should have low correlation with the archaeon in Fig. 8, as indeed they do.

There are also a number of strong correlations between distant genomes seen in Fig. 8 that, as far as we are aware, have no other corroboration: *M. jannaschii-H. pylori*, *B. subtilis-E. coli*, *H. influenzae-Synechocystis* and the links between *E. coli* and the archaeons. It will be interesting to see if these "predicted" relations would be supported by future gene based analysis.

Discussion. The fact that comparitve studies based on 16S rRNA homology and on homology of genes encoding other proteins do not yield a consistent phylogenetic tree and the apparent widespread ability of one organism to uptake genes from phylogenetically distant organisms and express them [9] has made the issue of the tree of life a subject of intense recent debate, and has promoted a need to develop methods for studying microbial evolution that do not look only at gene homology. If evolutionary pressure mainly works on long contiguous sections of DNA and lineage is mainly corrupted by mutations that mostly affect single nucleotides, then whole-genome homology must carry phylogenetic information. The question is how to isolate and enhance such information other than keying on gene homology.

We have shown that, for the thirteen complete bacterial genomes published by May of 1998, frequencies distributions of 6-mers and 7-mers tagged by under-represented tetranucleotides and pentanucleotides, respectively, are highly genome specific and pairwise correlation of such distributions has very high signal-to-noise ratio and display strong dependence on the genome pairs. Although the method only looks at a very limited aspect of whole-genome correlation, the pattern of species relationship that emerges from this study exhibits a surprisingly high degree of consistency to relationship obtained from gene-based analysis. This suggests that the method has the potential for development into a phylogenetic tool for whole-genome analysis. The tags used in this study are under-represented short palindromic oligonucleotides that are mostly recognition sites for restriction enzymes. In the present study spatial information on the tagged

oligonucleotides were ignored and a refinement of the method is to incorporate this information into the correlation. Another obvious extension of the present work is to look at distributions of oligonucleotides tagged by over-represented strings. Since such tags are known to be species dependent, correlations of the distributions can be expected to yield phylogenetic information.

Acknowledgment. HCL thanks the Zoology Department of the University of British Columbia for hospitality, where part of this work was done. This work was partly supported by National Science Council (ROC) grants NSC 87-2119-M-007-004 to BLH and NSC 87-2112-M-008-002 to HCL.

References

[1] Pennisi, E. *Science* **280**, 672-674 (1998).

[2] Woese, C.R. and Fox, G.E. *Proc. Natl. Acad. Sci. USA* **74** 5088-5090 (1977).

[3] Woese, C.R. Kandler, O. and Wheelis, M.L. *Proc. Natl. Acad. Sci. USA* **87** 4576-4579 (1990).

[4] Olsen, G.J., Woese, C.R. and Overbeek, R. *J. Bacteriology* **176**, 1-6 (1994).

[5] Cavalier-Smith, T. *Microbial. Rev.* **57** 953-994 (1993).

[6] Keeling, P.J. and Doolittle, W.F. *Proc. Natl. Acad. Sci. USA* **94**, 1270-1275 (1997).

[7] Deckert, G. *et al. Nature* **392**, 353-358 (1998).

[8] Feng, D.F., Cho, G. and Doolittle, R.F. *Proc. Natl. Acad. Sci. USA* **94**, 13028-13033 (1997).

[9] Mazel, D., Dychinco, B., Webb. V.A. and Davies, J. Science **280**, 605-608 (1998).

[10] Doolittle, W.F. *Trends in Genetics* **18**, 307-311 (1998).

[11] Woese, C. *Proc. Natl. Acad. Sci. USA* **95**, 6854-6859 (1998).

[12] Burge, C., Campbell, A. M. & Karlin, S. *Proc. Natl. Acad. Sci. USA* **89**, 1358-1362 (1992).

[13] Karlin, S., Mrázek, J. & Campbell, A. M. *J. Bacteriol.* **179**, 3899-3913 (1997).

[14] Gelfand, M. S. & Koonin, E. V. *Nucleic Acids Research* **25**, 2430-2439 (1997).

[15] Fleischmann, R. D. *et al.* and assembly of *Haemophilus influenzae* Rd. *Science* **269**, 496-512 (1995).

[16] Fraser, C. M. *et al.* The minimal gene complement of *Mycoplasma genitalium. Science* **270**, 397-403 (1995).

[17] Kaneko, T. *et al.* PCC6803, *I* & II. *DNA Res.* **2**, 153-166 (1995); **3**, 109-136 (1996).

[18] Bult, C. J. *et al. Science* **273**, 1058-1073 (1996).

[19] Himmelreich, R. *et al. Nucl. Acids Res.* **24**, 4420-4449 (1996).

[20] Freiberg, C. *et al. Nature* **387**, 394-401 (1997).

[21] Tomb, J.-F. *et al. Nature* **388**, 539-547 (1997).

[22] Blattner, F. R. *et al. Science* **277**, 1453-1462 (1997).

[23] Smith, D. R. *et al. J. Bacteriol.* **179**, 7135-7155 (1997).

[24] Kunst, F. *et al. Nature* **390**, 249-256 (1997).

[25] Klenk, H.-P. *et al. Nature* **390**, 364-370 (1997).

[26] Fraser, C. M. *et al. Nature* **390**, 580-586 (1997).

[27] McClelland, M. *et al. Nucl. Acids Res.* **15**, 5985-6005 (1987).

[28] McClelland, M. & Bhagwat, A. S. *Nature* **355**, 595-596 (1992).

[29] Feldman, M., (*private communication*, 1998).

[30] Ibba, M. *et al. Science* **278**, 1119-1122 (1997); *Proc. Natl. Acad. Sci. USA* **94**, 14383-14388 (1997).

TABLES

Table 1: Under-represented palindromic tetranucleotides σ.

	This work			Ref. [13]	
Bacteria	σ	$\bar{f}(\sigma)$	$\tau^{(2)}$	σ	$\tau*^{b}$
Archaeoglobus fulgidus [a]	CTAG	240	0.18		
	CGCG	794	0.50		
	GATC	1136	0.27		
	GCGC	1244	0.60		
Methanococcus jannaschii	CTAG	73	0.027	CTAG	0.06
	GATC	204	0.053	GATC	0.11
	GTAC	270	0.22	GTAC	0.78
	CGCG	388	0.90	CGCG	0.70
	GCGC	458	0.30	GCGC	0.52
	TCGA	754	0.88	*unlisted*	
Methanobacterium therm.	CTAG	349	0.33	CTAG	0.29
	CGCG	675	0.85	*unlisted*	
Bacillus subtilis	CTAG	763	0.85	*unlisted*	
Mycoplasma genitalium	TCGA	717	0.92	*unlisted*	
	TATA	890	0.68	TATA	0.78
Mycoplasma pneumoniae	TATA	775	0.71	TATA	0.78
	ATAT	924	0.93	*unlisted*	
Rhizobium sp. NGR234	CTAG	820	0.58		
Escherichia coli	CTAG	191	0.26	CTAG	0.73
Haemophilus influenzae	CCGG	951	0.27	CCGG	0.37
	CTAG	1198	0.71	CTAG	0.63
	CATG	1273	0.38	CATG	0.43
	TATA	1431	0.69	*unlisted*	
	GGCC	1630	0.28	GGCC	0.50
Helicobater pylori [a]	GTAC	81	0.10		
	ACGT	187	0.089		
	TCGA	201	0.12		
Synechocystis PCC6803	CGCG	396	0.21	CGCG	0.37
	TATA	1256	0.74	*unlisted*	
	GCGC	1519	0.59	GCGC	0.63
Borrelia burgdorferi [a]	ACGT	785	0.80		
	CGCG	986	0.66		
	GTAC	1075	0.67		
Aquifex aeolicus [a]	CTAG	649	0.36		
	GCGC	696	0.32		
	TCGA	1110	0.44		
	GGCC	1169	0.37		
	GATC	1235	0.54		

[a] Sequences not studied in [13]; See [b] [13] for definition.

Table 2: Under-represented pentanucleotides. The columns under N_1 and N_2 give the total number of under-represented pentanucleotides and the number of those that contain a palindromic tetranucleotide; the fourth column gives the tetrapalindromes that appear in the under-represented pentanucleotides; the fifth column lists the pentapalindromes.

Bacteria	N_1	N_2	Contained tetrapalindromes	Pentapalindromes
A. fulgidus	9	7	CTAG GATC CGCG GTAC ACGT[a]	GGwCC[b]
M. jannaschii	56	26	CTAG GATC GCGC	CGTCG GTsAC AGwCT
M. therm.	21	10	CTAG CGCG	CTwAG
B. subtilis	8	6	ctag	-
M. genitalium	53	15	GGCC[a]TCGA GTAC[a]	CGTCG CGGCG TCsGA
M. pneumoniae	6	4	ATAT TATA	CGwCG
Rhizobium sp.	5	4	CTAG	-
E. coli	15	7	CTAG	CTwAG GGwCC
H. influenzae	10	5	CCGG CATG CTAG	GGwCC GAsTC
H. pylori	39	14	TCGA GTAC ACGT TATA[a]ATAT[a]	GTsAC ACwGT TCsGA GGwCC CGwCG ACsGT TCwGA
Synechocystis	15	8	CGCG TATA	-
B. burgdorferi	14	8	CGCG ACGT GTAC	CGwCG
A. aeolicus	20	17	GCGC TCGA CTAG GATC	CAwTG GTsAC GAwTC

[a]These tetrapalindromes do not appear in Table 1.
[b]Relative abundance of underlined pentapalindromes are less than 0.6.

Table 3: Parameters for gamma and inverse-gaussian distributions of frequencies of 6-mers of bacterial sequences and random inverse-gaussian distributions.

Bacteria	*A. ful.*	*M. jan.*	*M. the.*	*B. sub.*	*M. gen.*	*M. pneu.*	*R. sp.*	*E. coli*
α	2.521	1.397	2.188	3.266	1.943	3.249	3.788	3.098
β	96.60	176.2	111.3	75.01	119.6	74.93	64.19	78.64
μ	243	246	244	245	232	243	243	244
λ	614	344	533	800	451	790	920	754
Bacteria	*H. inf.*	*H. pyl.*	*Synecho.*	*B. bur.*	*A. aeol.*	*Inv.G. I*	Inv.G. *II*	Inv.G. *III*
α	2.701	1.312	2.302	2.156	1.781			
β	91.55	190.4	105.9	121.0	136.6			
μ	247	250	244	261	243	244	244	244
λ	668	328	562	562	433	300	600	900

Table 4: Observed moments and moments from gamma distribution and inverse-gaussian distribution fits to bacterial 6-mer frequency distributions.

Bacteria	Moments of distribution $\langle \bar{f}^k \rangle$					
	$k=3$			$k=4$		
	(a)	(b)	(c)	(a)	(b)	(c)
H. pylori	7.24×10^{7}	7.01×10^{7}	7.93×10^{7}	6.13×10^{10}	5.77×10^{10}	8.26×10^{10}
E. coli	3.13×10^{7}	3.15×10^{7}	3.30×10^{7}	1.49×10^{10}	1.51×10^{10}	1.76×10^{10}
A. fulgidus	3.76×10^{7}	3.69×10^{7}	3.91×10^{7}	2.11×10^{10}	1.98×10^{10}	2.41×10^{10}
Rhizobium sp.	2.75×10^{7}	2.80×10^{7}	2.90×10^{7}	1.16×10^{10}	1.22×10^{10}	1.38×10^{10}
	$k=5$			$k=6$		
	(a)	(b)	(c)	(a)	(b)	(c)
H. pylori	6.10×10^{13}	5.86×10^{13}	11.5×10^{13}	6.76×10^{16}	7.07×10^{16}	20.4×10^{16}
E. coli	8.23×10^{12}	8.43×10^{12}	11.7×10^{12}	5.21×10^{15}	5.37×10^{15}	9.30×10^{15}
A. fulgidus	1.41×10^{13}	1.25×10^{13}	1.87×10^{13}	1.10×10^{16}	0.917×10^{16}	1.78×10^{16}
Rhizobium sp.	5.46×10^{12}	6.12×10^{12}	7.93×10^{12}	2.84×10^{15}	3.46×10^{15}	5.42×10^{15}

(a) Observed moment (Eq.(7)); (b) computed from gamma distribution(Eqs.(8,6));
(c) computed from inverse-gaussian distribution(Eqs.(9,6)).

Table 5: Parameters for gamma and inverse-gaussian distributions of frequencies of 7-mers of bacterial sequences and random inverse-gaussian distributions.

Bacteria	*A. ful.*	*M. jan.*	*M. the.*	*B. sub.*	*M. gen.*	*M. pneu.*	*R. sp.*	*E. coli*
α	1.764	0.925	1.544	2.418	1.268	2.195	2.697	2.240
β	34.30	66.67	39.19	25.22	45.25	27.56	22.39	27.01
μ	60.5	61.7	60.5	61.0	57.4	60.5	60.4	60.5
λ	106.7	57.1	93.4	147.5	72.8	132.8	162.9	135.5
Bacteria	*H. inf.*	*H. pyl.*	*Synecho.*	*B. bur.*	*A. aeol.*	*Inv.G. I*	Inv.G. *II*	Inv.G. *III*
α	1.716	0.914	1.476	1.434	1.266			
β	35.90	67.89	41.03	46.22	47.69			
μ	61.6	62.1	60.6	66.3	60.4	61.0	61.0	61.0
λ	105.7	56.8	89.5	95.1	76.5	70	110	150

FIGURES

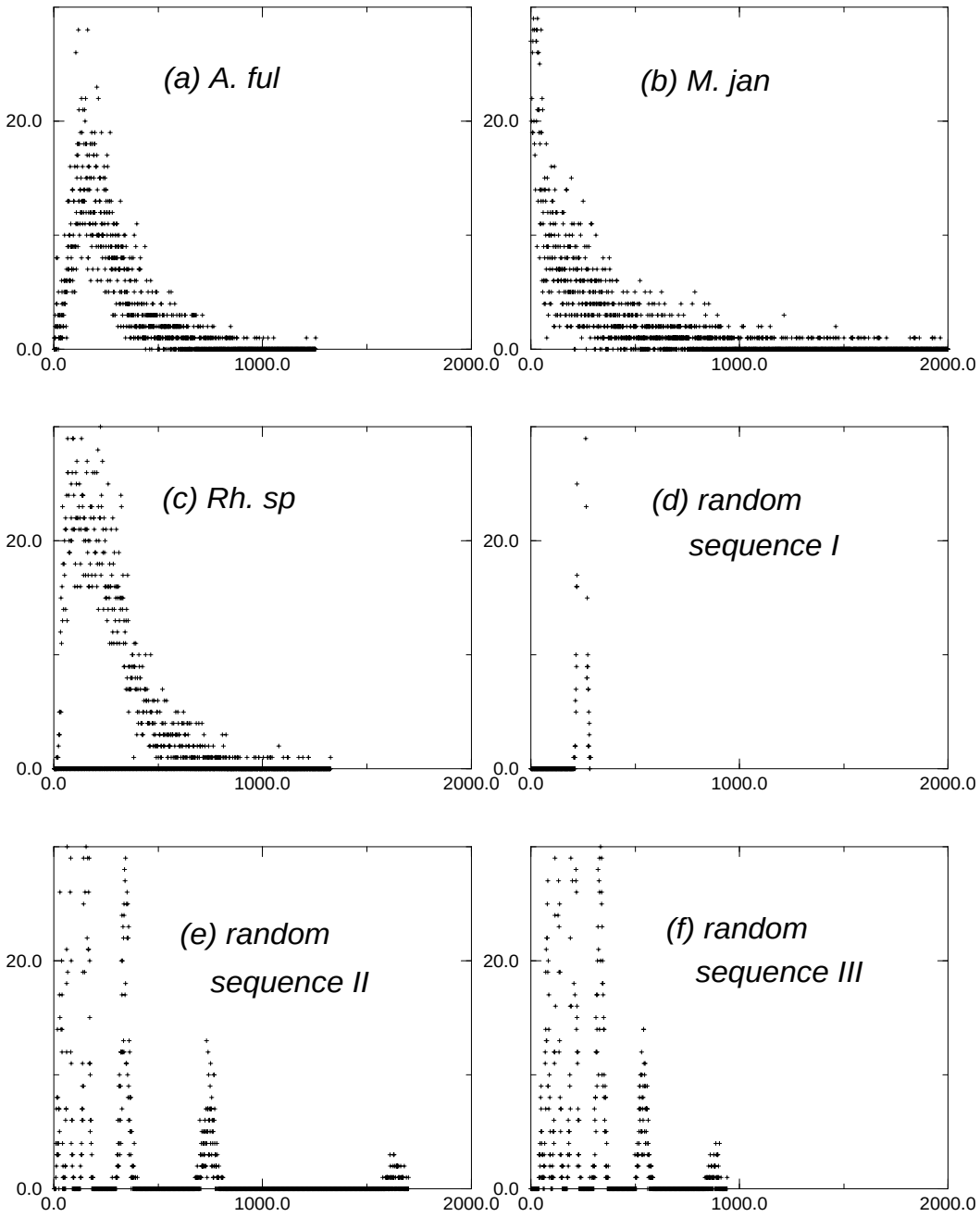

Figure 1: Distributions of frequencies $f(\sigma)$ of 6-mers σ in three bacterial and three random sequences, all normalized to a total length of one million bases. In each plot the abscissa is the frequency $f(\sigma)$ and the ordinate is the number of σ's having that frequency. $f(\sigma)$ retains dependence on mononucleotide composition of the sequence.

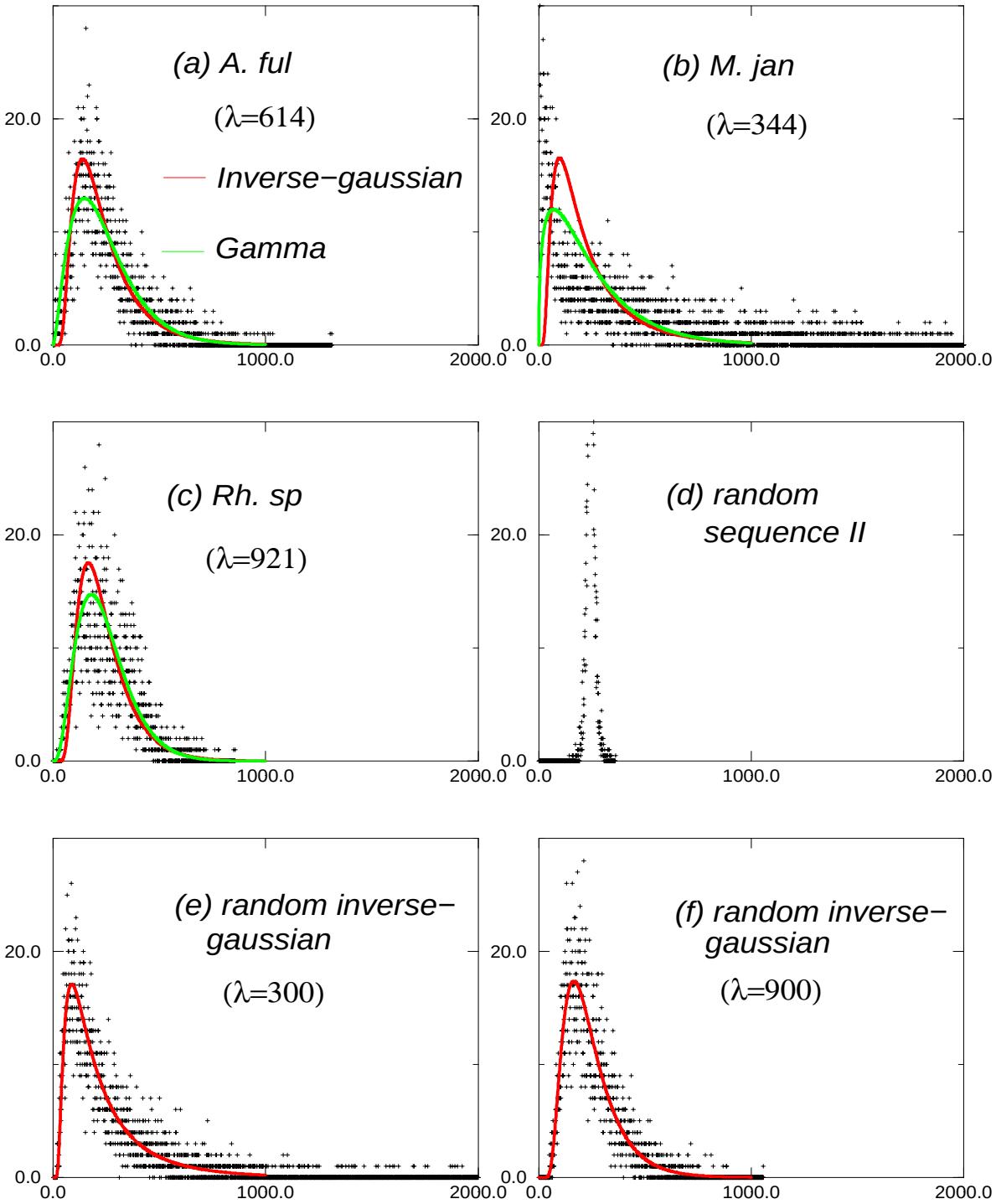

Figure 2: Distributions of normalized frequencies $\bar{f}(\sigma)$ of 6-mers σ in three bacterial sequences (a - c), a random sequence (d) and two random distributions constructed to fit inverse-gaussian distributions. Gamma (green line) and inverse-gaussian (red line) fits to the bacterial distributions are also indicated in (a - c). See Fig. 1 for further description of plots.

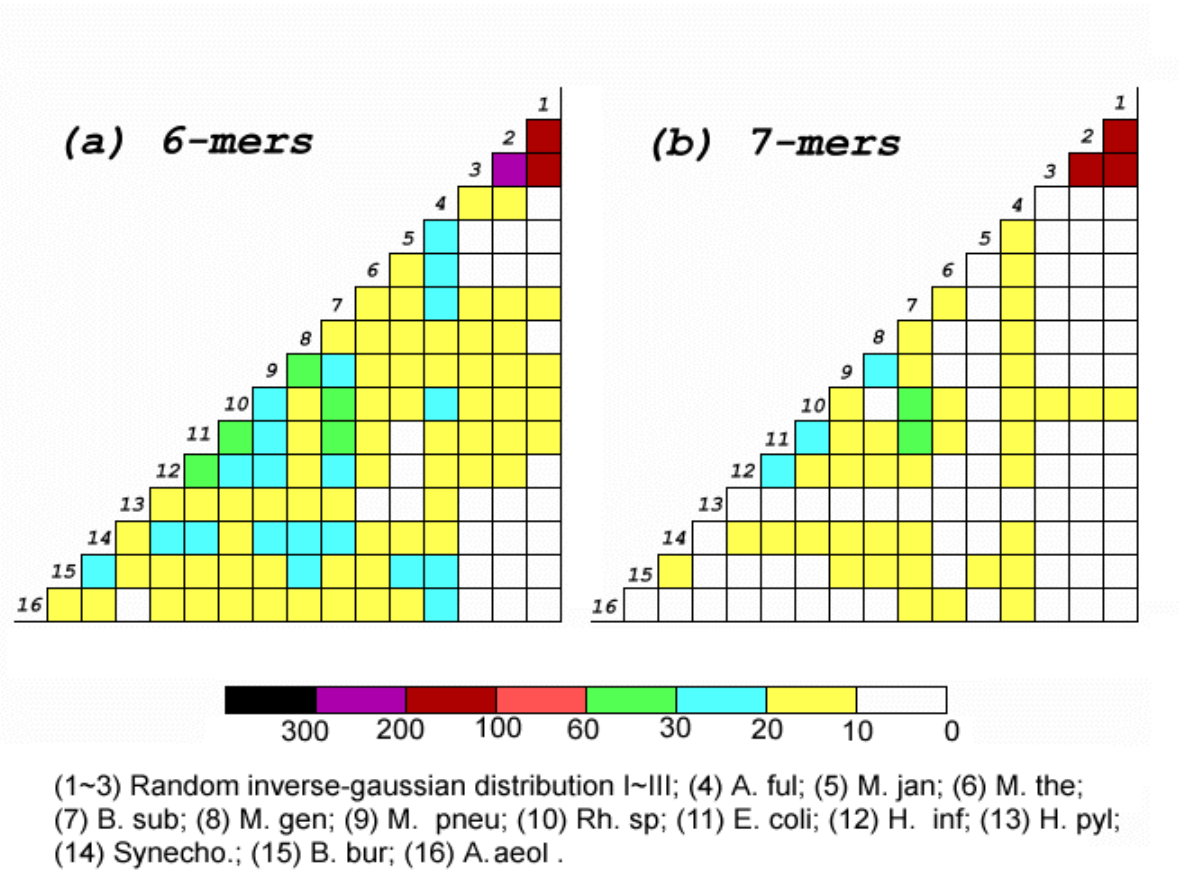

Figure 3: Color-coded correlations of normalized frequency distributions of thirteen bacterial sequences and three random inverse-gaussian distributions of (a) 6-mers and (b) 7-mers. The coordinates of the lattice are given by numerals representing distributions, top to bottom for the ordinate and right to left for the abscissa. The correlation of a pair of distributions is given by the color of the lattice site whose coordinates represent the pair. Darker colors correspond to stronger correlations.

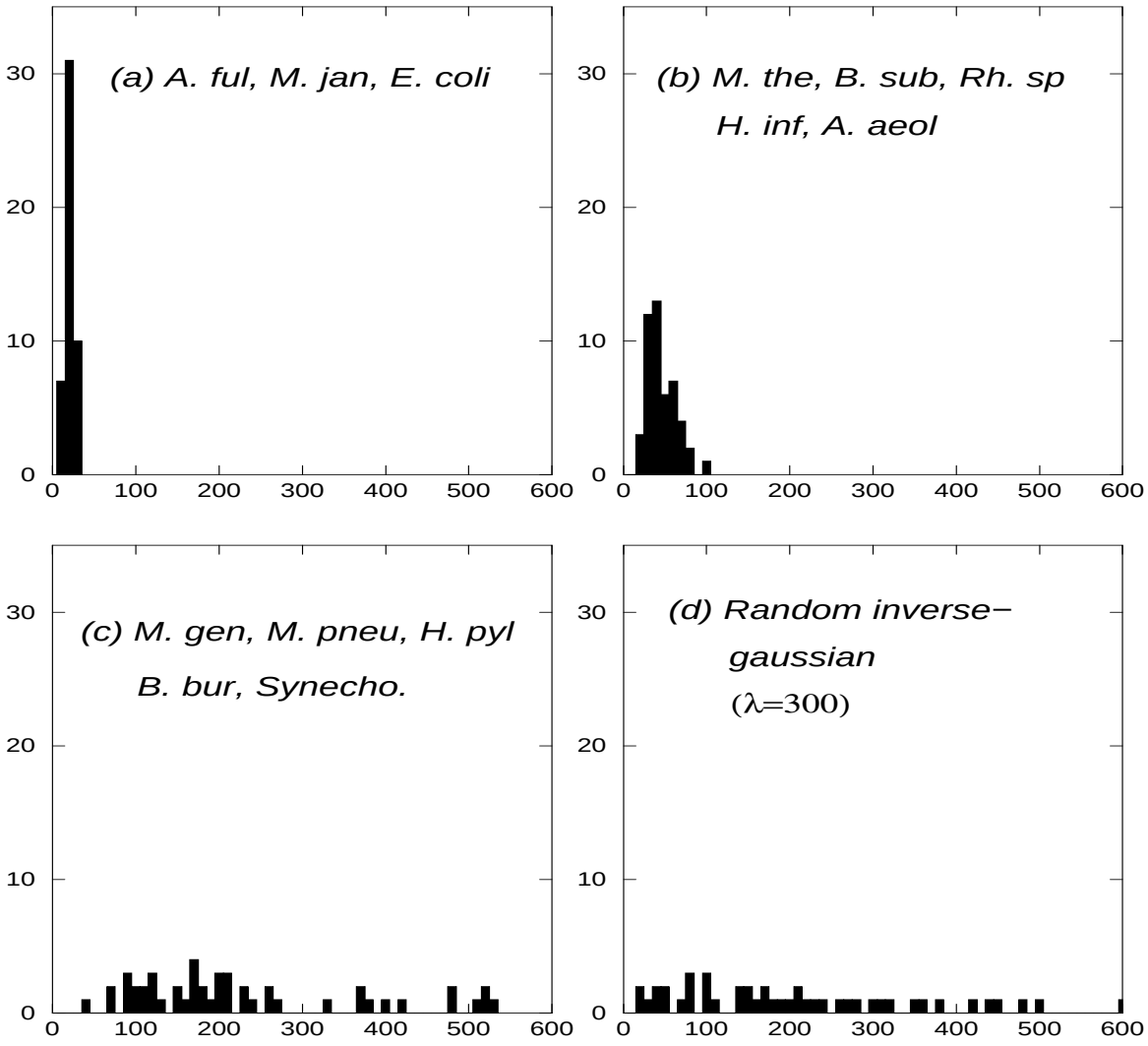

Figure 4: Distributions of normalized frequencies of the $CTAG$-tagged 6-mers. See Fig. 1 for further description of plots. (a), (b) and (c) are composites of similar genome distributions, (d) is from a random inverse-gaussian distribution.

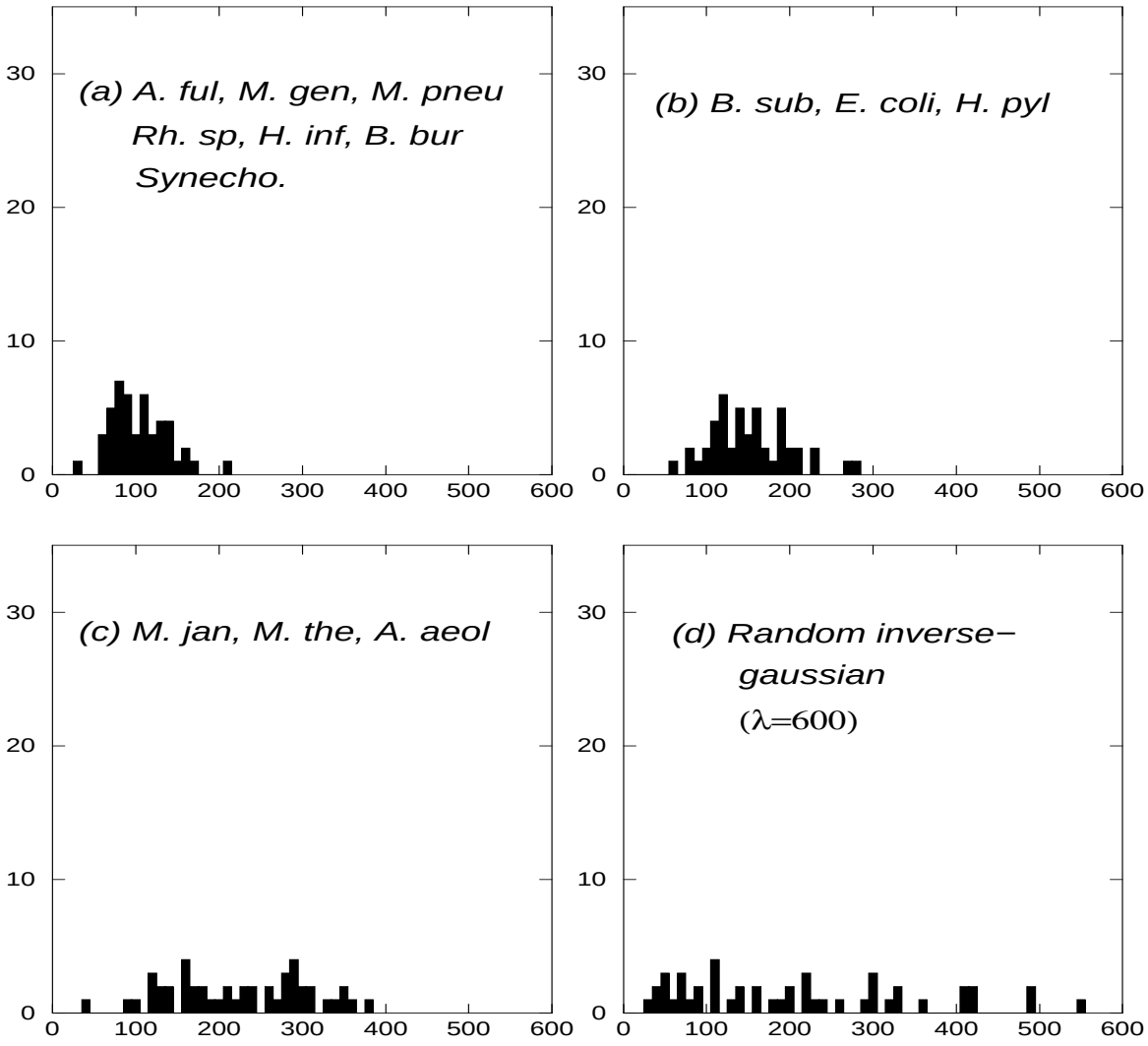

Figure 5: Distributions of normalized frequencies of $TATA$-tagged 6-mers. See Fig. 1 for further description of plots. (a), (b) and (c) are composites of similar genome distributions, (d) is from a random inverse-gaussian distribution.

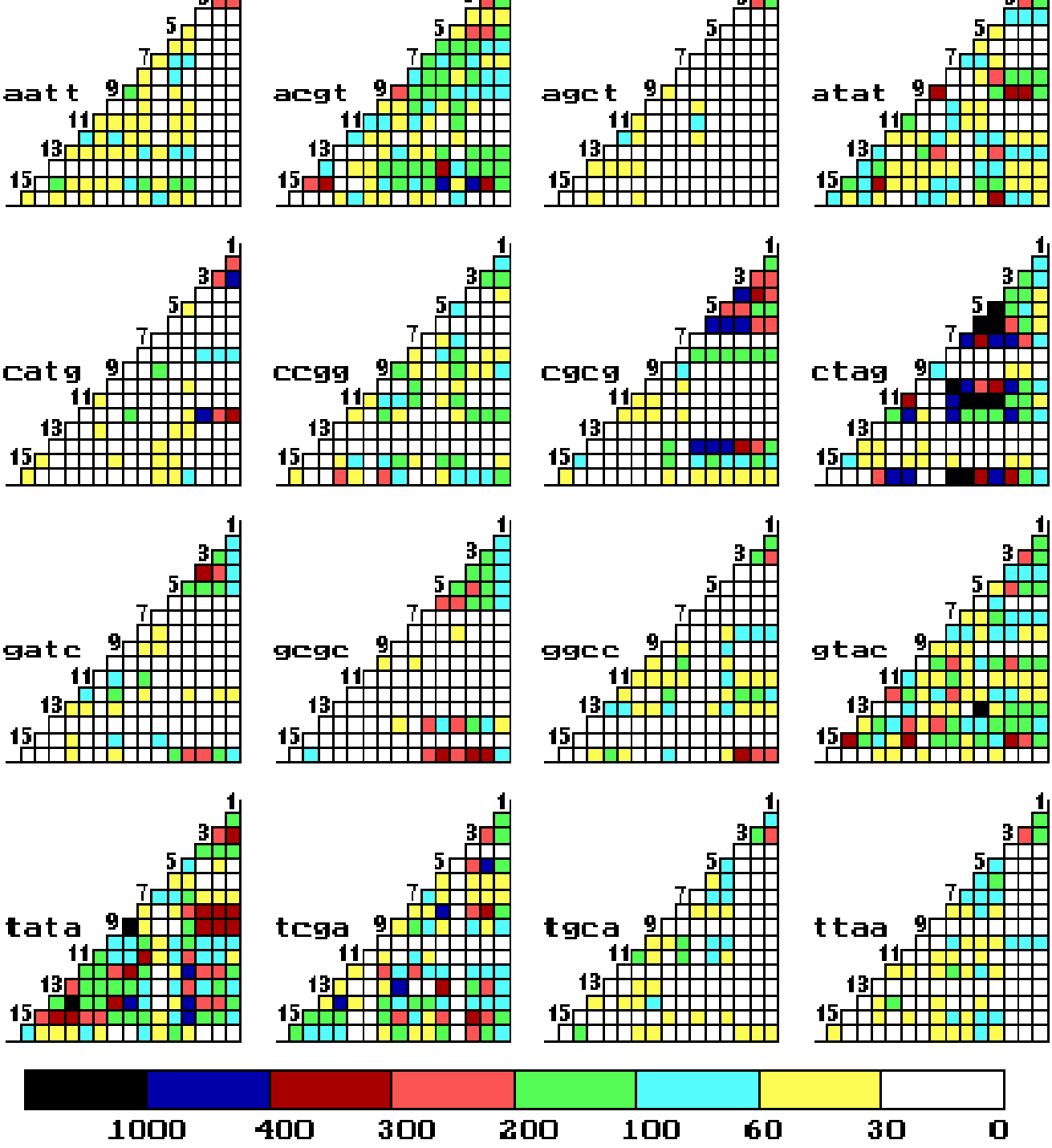

Figure 6: Color-coded correlations in frequency distributions of the 16 tetrapalindrome-tagged 6-mers in bacterial sequences and random inverse-gaussian distributions. Darker colors correspond to stronger correlations. Confer caption of Fig. 3 for a more detailed explanation of the lattices. Numeral representations of sequences are: (1) Random inverse-gaussian distribution I; (2) Random inverse-gaussian distribution II; (3) Random inverse-gaussian distribution III; (4) *A. fulgidus*; (5) *M. jannaschii*; (6) *M. thermoautotrophicum*; (7) *B. subtilis*; (8) *M. genitalium*; (9) *M. pneumoniae*; (10) *Rhizobium sp.*; (11) *E. coli*; (12) *H. influenzae*; (13) *H. pylori*; (14) *Synechocystis*; (15) *B. burgdorferi*; (16) *A. aeolicus*.

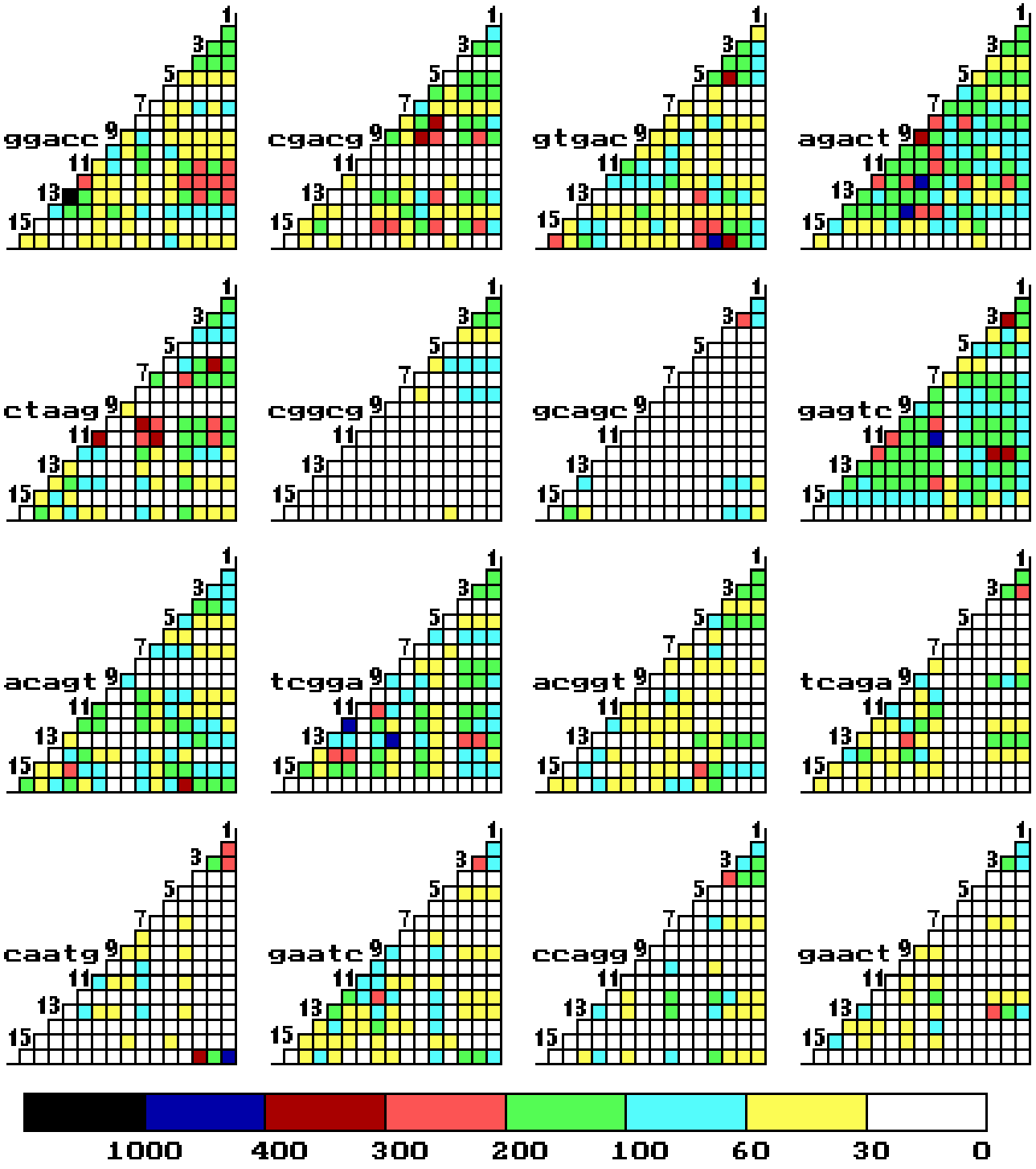

Figure 7: Color-coded correlations in frequency distributions of 16 pentanucleotide-tagged 7-mers in bacterial sequences and random inverse-gaussian distributions. Darker colors correspond to stronger correlations. Confer caption of Fig.3 for a more detailed explanation of the lattices. Numeral representations of sequences same as in Fig. 6.

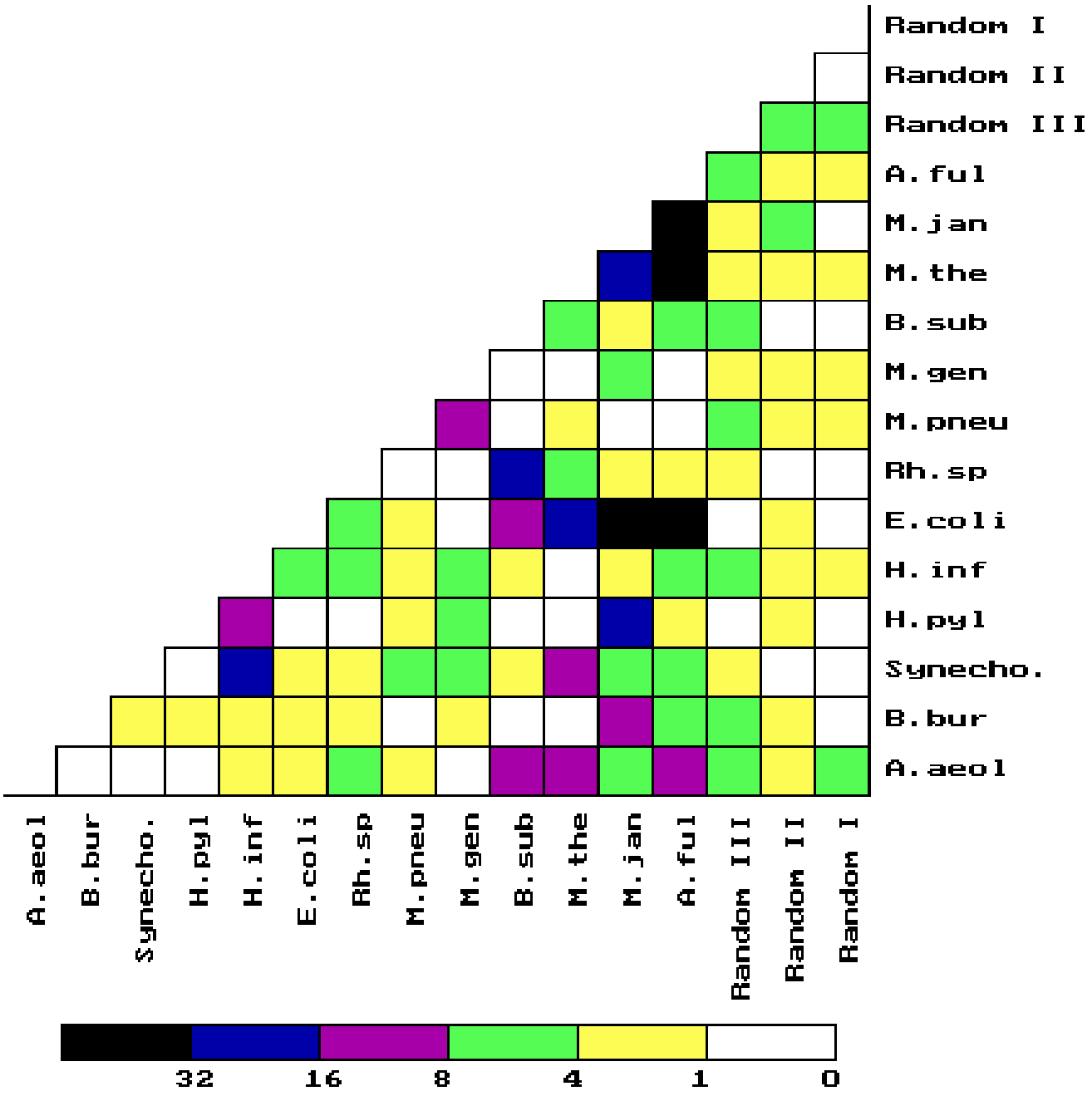

Figure 8: Color-code genome-genome correlations obtained summing the correlations in Fig. 6 and Fig. 7 above the background value 200. See text for detailed description. Darker colors correspond to stronger correlations. Confer caption of Fig.3 for a more detailed explanation of the lattice. Numeral representations of sequences same as in Fig. 6.