
In silico comparison of bacterial strains using mutual information

D SWATI

Department of Physics, MMV, Banaras Hindu University, Varanasi 221 005, India

(Email, swati@bhu.ac.in)

Fast-sequencing throughput methods have increased the number of completely sequenced bacterial genomes to about 400 by December 2006, with the number increasing rapidly. These include several strains. *In silico* methods of comparative genomics are of use in categorizing and phylogenetically sorting these bacteria. Various word-based tools have been used for quantifying the similarities and differences between entire genomes. The simple di-nucleotide frequency comparison, codon specificity and k-mer repeat detection are among some of the well-known methods.

In this paper, we show that the Mutual Information function, which is a measure of correlations and a concept from Information Theory, is very effective in determining the similarities and differences among genome sequences of various strains of bacteria such as the plant pathogen *Xylella fastidiosa*, marine Cyanobacteria *Prochlorococcus marinus* or animal and human pathogens such as species of *Ehrlichia* and *Legionella*. The short-range three-base periodicity, small sequence repeats and long-range correlations taken together constitute a genome signature that can be used as a technique for identifying new bacterial strains with the help of strains already catalogued in the database.

There have been several applications of using the Mutual Information function as a measure of correlations in genomics but this is the first whole genome analysis done to detect strain similarities and differences.

[Swati D 2007 *In silico* comparison of bacterial strains using mutual information; *J. Biosci.* **32** 1169–1184]

1. Introduction

The DNA molecule is made up of four nucleotides or bases – adenine, thymine, guanine and cytosine. Its primary sequence can be visualized as a language written with an alphabet of four letters, the symbols being A, T, G, C. The DNA sequence can also be treated as a quaternary symbol string. As a language has words with meaning or message content, DNA has coding sequences. Earlier all non-coding sequences were thought to be ‘gibberish’ or ‘junk’ DNA. Sustained research has established that non-coding DNA has biological functions no less important than those of the coding sequences. DNA, like any language has its dialects,

accents and pauses. Dialects correspond to different genomes. Accents in the spoken language help locate people geographically. The structure of repeats – type and copy number – can be compared to the accent of an individual and taken as a genome signature. The *Alu* repeats for the human genome are an example of a well-known accent. Just as the information content of a language can be analysed, the information content of a DNA sequence can be analysed and lead to a better understanding about the sequence itself.

Since DNA encrypts the genetic code it cannot be a random sequence of four symbols. Computational biologists are interested in studying correlations between the occurrences

Keywords. Bacterial strains; correlations; DNA alphabet; Mutual Information; repeat structure

Abbreviations used: C – C, cytosine – cytosine; GC, guanine and cytosine; SSR, short sequence repeats; VNTR, variable number of tandem repeats

Supplementary Material Folder pertaining to this article is available on the *Journal of Biosciences* Website at <http://www.ias.ac.in/jbiosci/sept2007/pp1169-1184-suppl.pdf>

of nucleotides (bases) at different sites or positions occupied by the bases along the sequence. For computing correlations between nucleotide – nucleotide pairs we have to find whether there are dependencies of occurrence of nucleotides along the sequence. That is, given a nucleotide, say α at site ‘ i ’ of the DNA sequence we want to find the probability of occurrence of nucleotide β at site ‘ j ’ of the same sequence, where ‘ j ’ is a site ‘ k ’ bases downstream. From this conditional probability, we can then find the correlation function between these two nucleotides or bases.

There are several ways in the literature of mathematically defining what is meant by ‘correlations’. As a measure for estimating correlations, we have chosen the Mutual Information function as defined by Shannon’s theorem (Shannon 1948) from Information Theory, since it is uniformly applicable to any system of study and has an easily understood definition. The objective (system independent) nature of Mutual Information as a measure of correlations is discussed in detail in Li (1990).

In order to compute the Mutual Information function, we have to reduce the DNA sequence to a quaternary string of numeric symbols say 0, 1, 2, 3 for bases A, T, G, C, respectively. Alternatively we may use a property of the DNA molecule to make the four symbols collapse to two, thereby converting it to a binary string of symbols.

One possible choice is to take the strength of hydrogen bonding between A - T pair which is weak (W), and that between the G - C pair which is strong (S). This makes the DNA sequence a string of binary symbols W and S .

If we have a DNA sequence of length N composed of bases α_i , where α is one of the nucleotides A, T, G or C at position ‘ i ’ on the sequence,

$$\alpha_1, \alpha_2, \alpha_3, \dots, \alpha_N.$$

Then the sequence of symbols α_i has to be reduced to string of numbers before one can study the correlations hidden in the sequence.

For this purpose, we use the projection operator method of Voss (1992) and use $U_\alpha(i) = 1$, if α_i the symbol at position i is α ; otherwise we take it as zero.

Let us consider the following small sequence for illustration.

ATTGCACAG

$$\left. \begin{aligned} \text{Where } U_A(1) = 1 \text{ and } U_C(1) = 0; U_G(1) = 0; U_T(1) = 0 \\ U_T(2) = 1 \text{ and } U_A(2) = 0; U_G(2) = 0; U_T(2) = 0 \\ U_G(9) = 1 \text{ and } U_A(9) = 0; U_C(9) = 0; U_T(9) = 0 \text{ etc.} \end{aligned} \right\}$$

If we consider site ‘ j ’, which is ‘ k ’ bases downstream from site ‘ i ’ that is

$$j = i + k.$$

We can define a correlation function

$$X_{\alpha\beta}(k) = \langle U_\alpha(i)U_\beta(i+k) \rangle. \tag{1}$$

Where α and β belong to the set {A, T, G, C} and the symbol $\langle \dots \rangle$ denotes taking an average over the entire sequence. When α and β are different, equation (1) defines a cross-correlation function and when α and β are identical, it defines on auto-correlation function.

However this is not the only way correlation functions are defined in literature. Sometimes, the expression is divided by ‘ k ’; in some references the definition that includes the covariance and variance of a distribution (of symbols in this case) is used.

1.1 Mutual Information function

We have used the Mutual Information function $M(k)$ as defined by Shannon (1948) and applied to biosequences by Li (1990, 1997) which, by definition is a weighted sum over all sixteen nucleotide – nucleotide correlation functions. It indicates in a single parameter the correlations, if any, between the symbols in a string.

If $\{x_i\}$ $i = 1, 2, \dots, N$ is to symbolize the sequence string with N being the total number of bases or nucleotides and

$$p_\alpha = \frac{N_\alpha}{N}; p_\beta = \frac{N_\beta}{N}$$

are the independent frequencies of occurrence of base $\alpha \in$ (A, T, G, C) and base $\beta \in$ (A, T, G, C) at position i and j of the sequence and N_α and N_β are the number of bases of type α and β , respectively, then the Mutual Information function $M(k)$ is defined as:

$$M(k) = \sum_{\alpha}^n \sum_{\beta}^n p_{\alpha\beta}(k) \ln_2(p_{\alpha\beta}(k) / p_\alpha p_\beta). \tag{2}$$

Where $p_{\alpha\beta}(k)$ is the conditional probability of a base $\alpha \in$ (A, T, G, C) existing at site i with a base $\beta \in$ (A, T, G, C) at a site j where $j = i + k$, and n is the number of symbols in the DNA alphabet or symbol string. The conventional units of $M(k)$ are bits if the logarithm is taken to the base 2. If there was no mutual dependence of occurrence of nucleotides along the DNA sequence, then

$$p_{\alpha\beta}(k) = p_\alpha p_\beta.$$

And in that case the Mutual Information function $M(k)$ as defined above is zero.

However, if

$$p_{\alpha\beta}(k) > p_\alpha p_\beta$$

then the occurrence of nucleotides (bases) α and β are said to be positively correlated and the value of $M(k)$ as defined above is positive.

If the sequence was totally random then $M(k)$ would have a value that would be inversely proportional to N , the total number of nucleotides in the DNA sequence string (Li 1997). N is taken to connote genome size or genome length and is measured in units of nt (nucleotides) or kb (kilobases) or Mb (megabases). Therefore, the minimum value of $M(k)$ for a given DNA sequence for a prokaryote of genome size 4 Mb is typically 10^{-6} . This is also called the 'volume exclusion factor' (Chechetkin and Turynin 1996).

We have $n = 2$ for binary and $n = 4$ for quaternary strings and these are said to be the alphabets denoting the number of symbols needed to define the symbol string in either case.

There are two other possible choices for binary alphabets to represent the DNA symbol string based on the chemical structure of the nucleotides A, T, G and C. These are:

Purine (A/G) and pyrimidine (C/T)

Amino (A/C) and keto (T/G).

The species independence of the three-base periodicity which arises in coding sequences due to codon usage bias is shown clearly for all DNA sequences when we make a plot of $M(k)$ versus the base separation k . We have analysed the entire genome analysis with the nucleotide – nucleotide correlation function for all possible sixteen combinations of correlation functions. $M(k)$ is computed by varying base separation from 1 to 10^5 or 0.1 Mb. For a typical bacterium, the size of the genome or genome length may be taken to be 2 Mb.

This means that we are probing short-range correlations say from (1–100) bases and also for base separation of the order of (100–3000) in length. The second upper limit corresponds roughly to the maximum possible length of coding sequences which can also be estimated from $M(k)$ versus k plots. Finally, there is the really long range estimation of correlations for (10^4 – 10^5) base separation, which is termed 'asymptotic' in this paper.

We find that after we take the average of the data to smoothen out the three base periodicity, the repeat structure inherent in the DNA is revealed. In this work we look at the effect of strain difference on Mutual Information and then use it to detect similarities or differences in controversial cases such as that of certain strains of *Bacillus thuringiensis* and *Bacillus cereus* mimicking the *Bacillus anthracis* pathogen as claimed in Helgason *et al* (2000).

Earlier studies involving Mutual Information function include those by Berryman and Abbott (2004), Grosse *et al* (2000) and Holste *et al* (2003), to cite a few. They have applied it study to coding and non-coding regions separately or to analyse repeat structures but to our knowledge this is the first study to use $M(k)$ as a strain- discriminating function.

Another simple method often used for analysing periodicities of the DNA sequence is the method of Fourier

Transforms and spectral analysis (Silverman and Linsker 1986; Tiwari *et al* 1997; Lobzin and Chechetkin 2000; Swati 2007). For circular sequences there is a simple relation between the autocorrelation functions as defined in equation (1) and the Fourier transform of the power spectra (Wiener-Khinchine Theorem 1949). But this does not apply to linear sequences of prokaryotic and eukaryotic genomes (Lobzin and Chechetkin 2000). For such cases, direct evaluation of correlation functions is required and it is shown in this paper that determining the Mutual Information function gives information not only on the over- all content of the genome in a single parameter but each term of equation (1) may be used to observe the individual nucleotide - nucleotide correlation functions as well.

2. Materials and methods

The study has been done on several bacterial strains. The bacteria chosen were mostly different types of pathogens of differing genome lengths and compositions i.e. guanine and cytosine (GC) content. The genomic data were downloaded from Genbank which is available at the site maintained by NCBI: <http://www.ncbi.nlm.nih.gov/>. The compositional details, GC%, which is the percentage of GC taken together, relative to the total number of nucleotides in a given DNA sequence, has been taken from the Comparative Genomics site maintained by the University of Lausanne at: <http://www2.unil.ch/comparativegenomics/>.

All genomic data are given in table 1 of the Supplementary Material folder.

A program in C++ was written by Rithun Mukherjee¹ to compute the Mutual Information function for all possible base separation beginning with $k=1$ and ending at $(N/2)$ for circular sequences and $(N-k)$ for linear sequences (chromosome and plasmids) where N is total number of bases or genome length.

The word size could be changed. For one word i.e. for value of k equalling 1, the 16 nucleotide – nucleotide correlation functions could be examined separately and their properly scaled sum generated to give the Mutual Information function $M(k)$. The maximum alignment and repeat finding software MUMmer by Kurtz *et al* (2004) was used to corroborate the results obtained in this study. This is an open source software available at: <http://mummer.sourceforge.net/>. This software uses the Suffix Tree method to find maximal pairwise matches between megabases long genomic sequences, where regions of maximal match, insertions or deletions are shown visually through plots, known as MUMmerplots.

¹Formerly at Center for Computational Biology and Bioinformatics, SIT, Jawaharlal Nehru University, New Delhi 110 057.

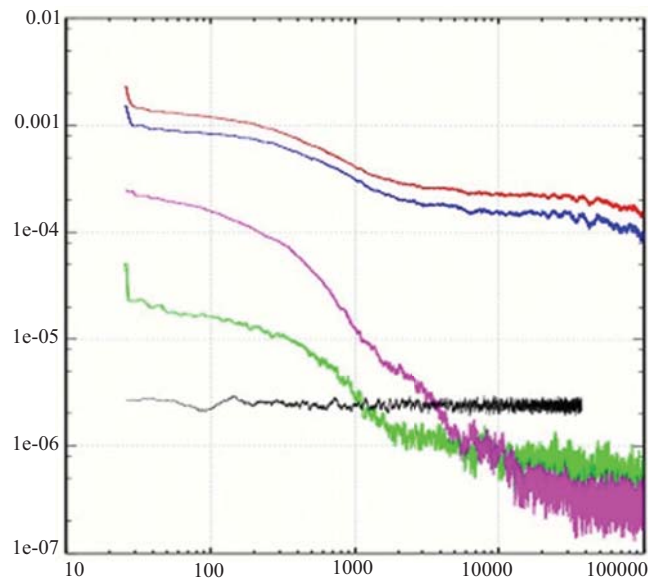


Figure 1. Mutual information $M(k)$ vs base separation k : *T. tengcongensis*: Quaternary – red; quaternary (randomized sequence) – black; binary M/K – green; binary R/Y – blue; binary S/W – pink.

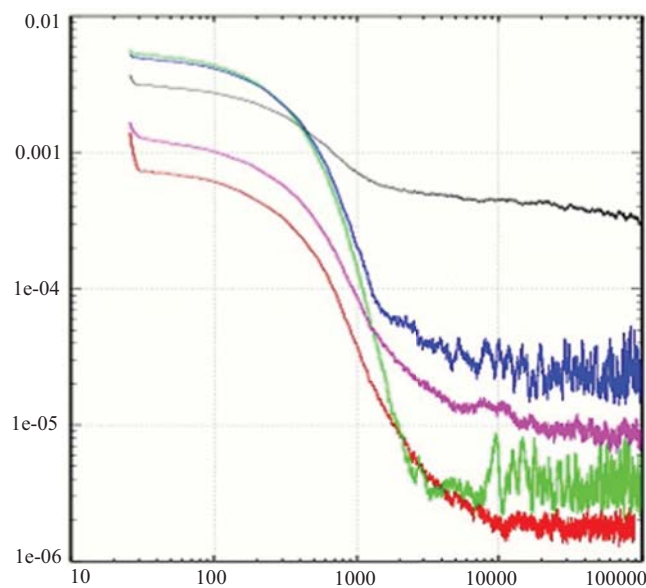


Figure 2. Mutual information $M(k)$ vs base separation k : for different bacteria: *E. coli* – red; *D. radiodurans* I – green; *D. radiodurans* II – blue; *C. diphtheriae* – pink; *C. tetani* – black.

3. Results

The results for computing Mutual Information function $M(k)$ for a given base separation k are plotted for different bacterial species and strains, the details of which are given in the legends below the figures. The $M(k)$ for a few randomized sequences are shown as the minimum value expected for Mutual Information function which may be used as a basal reference level. These values agree well

with the predicted value termed ‘volume exclusion factor’ (Chechetkin and Turygin 1996).

The nature of figures can be categorized in four main classes:

- (i) The Mutual Information function $M(k)$, first plotted without any smoothing as in figures 3 for *Xylella fastidiosa* strains and in figure 1 of Supplementary Material Folder. Then a smoothed plot of each strain is superimposed on the earlier plot.

- (ii) Figures 5, 9 and 14 are examples of $M(k)$ smoothed by doing a sliding window averaging method over the entire data, taking the window size to be 3 bases to smoothen out the three-base periodicity due to codon usage bias shown in all the Mutual Information function plots. This smoothening results in the short sequence repeats (SSR) structure seen in the $M(k)$ plots.
- (iii) The three-base periodicity can be removed by choosing a window size to be a multiple of three, for example any of the numbers 45, 51, 99 or 102 may be taken to smoothen out the SSRs and reveal the structure for large-base separation. We have chosen the window size to be 51. Figures 1, 2 and 11 are of this type.
- (iv) The remaining figures are a superimposition of two types of plots. First a plot of Mutual Information data is smoothed by window size 3 and then the same $M(k)$ data smoothed by window size 51 is superimposed on that. It may be noted that with values of $M(k)$ getting smaller for k -values larger than 50,000 the variations in the asymptotic value of $M(k)$, that is $M(k)$ for large values of k are visible only after the random fluctuations are smoothed out. The window size of 51 is an optimum size for this purpose.
- (v) Figures 6–20 are generated using the software MUMmer (Kurtz *et al* 2004). The reference sequence is plotted along the x-axis and the query sequence is plotted along the y-axis of each plot. If the match is maximal, one gets a perfect diagonal.

4. Discussion

4.1 General features of $M(k)$

The effect of computing Mutual Information for a randomized sequence of a DNA molecule is shown in figure 1 for *T. tengcongensis* along with the relative information content of different binary string choices of S/W , R/Y , M/K . The highest information content - revealed by values of $M(k)$ - is for the quaternary string, which is logically expected. It is worth noting that *T. tengcongensis* is an AT-rich genome and the R/Y binary choice has the highest information content among possible binary alphabet choices. The $M(k)$ for the randomized sequence of this DNA is shown to be flat as expected and corresponds to the minimum possible value coming from 'volume exclusion effect' (Chechetkin and Turygin 1996). Figure 2 is a comparison of four phylogenetically dissimilar bacteria: *E. coli*, *D. radiodurans* chromosomes I and II, *C. tetani* and *C. diphtheriae*. This is included to indicate how $M(k)$ patterns can vary widely with genomes of different species.

Figure 3 shows a comparison of two different strains of *Xylella fastidiosa* which have very similar information content. The very large base separation plots show some minor variation. The three-base periodicity is clearly shown and a smoothed plot over window size 51 is also included. The values of $M(k)$ for asymptotic values of k show some variation as noted above. The repeat structure of the two strains may be different (Colleta-Filho *et al* 2001) and this may be responsible for the observed variation.

Figure 4 shows a remarkable difference between the two strains of *P. marinus*_MIT-9313, which is a high light-adaptative type and is able to grow at very low irradiances (greater ocean depth) and *P. marinus*_MIT-9312, which is a low light-adaptative type, and can only grow at higher irradiances (close to the ocean surface as discussed by Rocap *et al* 2002). Both are marine species that flourish at different oceanic depths and use different mechanisms for photosynthesis. When five different strains of the marine Cyanobacteria *Prochlorococcus* spp were compared, four seem to group together and *P. marinus*_MIT-9313, the high light adaptive type seems quite different. It is larger in length and has a much higher GC%. (see table 1 and figure 2 in Supplementary Material Folder). From that figure it is apparent that the other three strains, *P. marinus*_NATL2A, *P. marinus*_marinus and *P. marinus*_pastoris give Mutual Information plots which are very similar to *P. marinus*_MIT-9312. It may tentatively be concluded that all three are low light-adaptative strains.

In figure 5 we see the Mutual Information plots of six *Mycoplasma* species. Each is distinguishable and shows a different structure of SSR exhibiting a genomic difference at the primary sequence level. This may be contrasted with figure 10 of different strains of *M. hyopneumoniae*, which show an almost identical nature.

However, we can deduce that different species appear to show different dependence of $M(k)$ on base separation k and the difference in behaviour of $M(k)$ points to a difference in the structure of the DNA sequence at the primary level. Bacterial strains on the other hand should show overall similar behaviour of $M(k)$. We shall see that there are notable exceptions to both assumptions. One exception to $M(k)$ of strains showing similar patterns is the case of the *P. marinus* MIT 9312 and *P. marinus*_MIT- 9313 strains as delineated above.

4.2 Clostridia: Clostridium perfringens strains

Figure 6 shows the plots for four strains of *C. perfringens*, *C. novyi*_NT and *C. tetani*_E88 with *C. acetobutylicum* as a counterpoint. The plots for smoothening over window size 3 are remarkably similar in SSR structure and look different only for very large base separation. The very recently sequenced genome of *C. novyi*_NT seems to have slightly

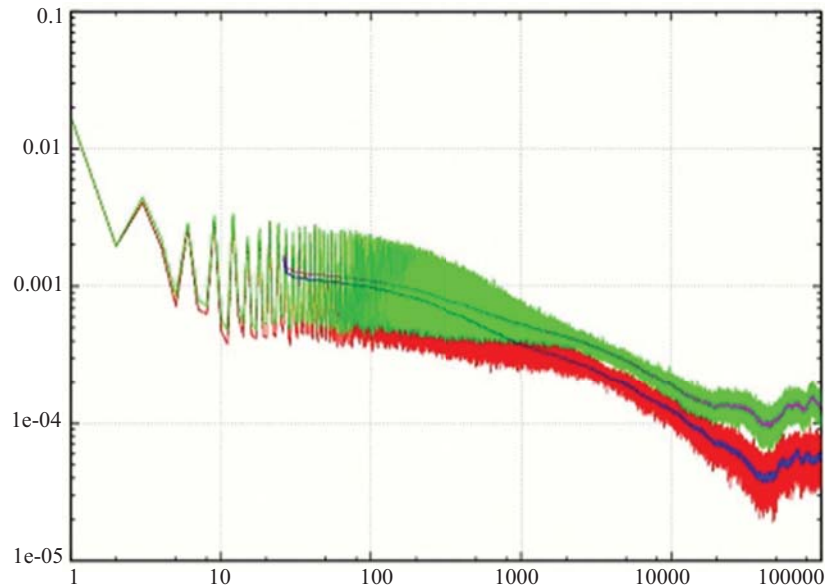


Figure 3. Mutual information $M(k)$ vs base separation k : *X. fastidiosa* strains 9a5c c – red; *Temecula* 1 – green; smoothed plot (window size 51) of mutual information: 9a5c – blue; *Temecula* 1 – pink.

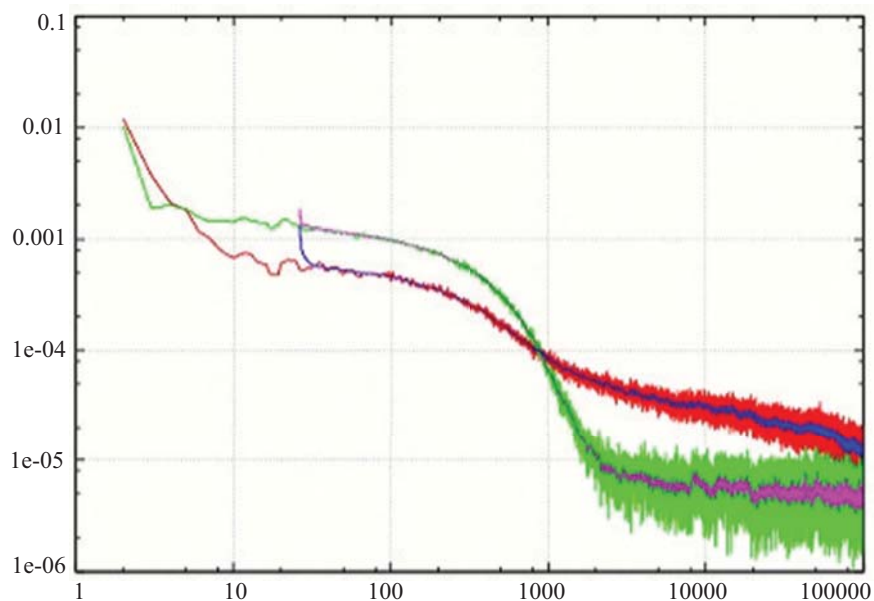


Figure 4. Mutual information $M(k)$ vs base separation k : *P. marinus* strains: MIT-9313 – red; MIT-9312 – green; smoothed plot (window size 51) of mutual information: MIT-9313 – blue; MIT-9312 – pink.

higher mutual information content for large base separations and indicates the possibility of large copies of repeats. The unusually large value of $M(k)$ for all Clostridia can possibly be related to the strand asymmetry seen with respect to the location of their coding sequences (Shimizu *et al* 2002). More

than 80% of the coding sequences lie on the leading strand. The analysis of the reasons leading to relatively large values of asymptotic $M(k)$ is being done by the author at present.

The results of MUMmer software applied to genome sequences of *C. tetani* E88 and *C. perfringens*_13, and two

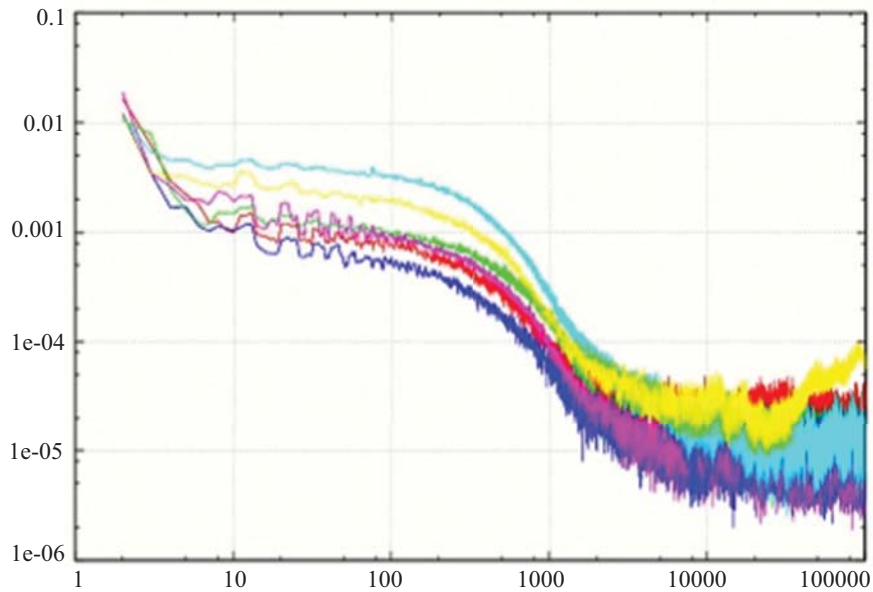


Figure 5. Mutual information $M(k)$ vs base separation k : *Mycoplasma*: *M. genitalium* – red; *M. gallisepticum* – green; *M. pneumoniae*_M129 – blue; *M. hyopneumoniae*_232 – pink; *M. capricolum_capricolum* – light blue; *M. synoviae* – yellow.

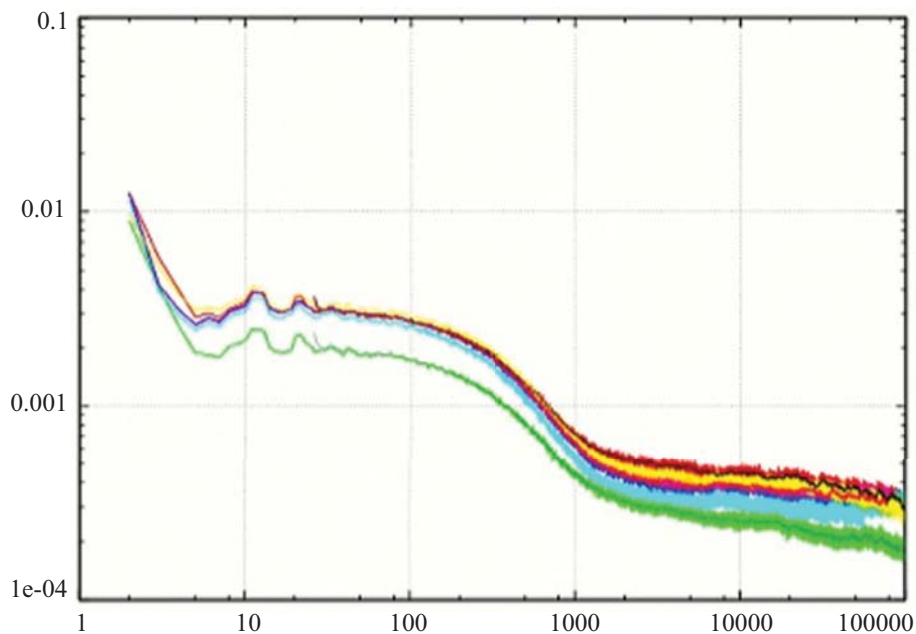


Figure 6. Mutual information $M(k)$ vs base separation k : *Clostridia*: *C. tetani* – red; *C. acetobutylicum* – green; *C. perfringens* strains: 13 – blue; ATCC-13124 – light blue; SM101 – pink; novyi-NT – orange; smoothed plot of mutual information: *C. tetani* – black; *C. perfringens*_13 – rust; *C. perfringens*_novyi-NT – yellow; *C. acetobutylicum* – grey.

strains of *C. perfringens*_13 and _SM101 are shown in figures 10 and 11 of the Supplementary Material Folder. There is little match seen between *C. tetani*_E88 and *C. perfringens*_13, but except for the deletion at around 1.2

Mb, there is a good match in the sequence of the two strains of *C. perfringens*_13 and _SM101. Their plots of $M(k)$ are also somewhat different at large values of base separation k . To search for sequence similarity between the plasmid

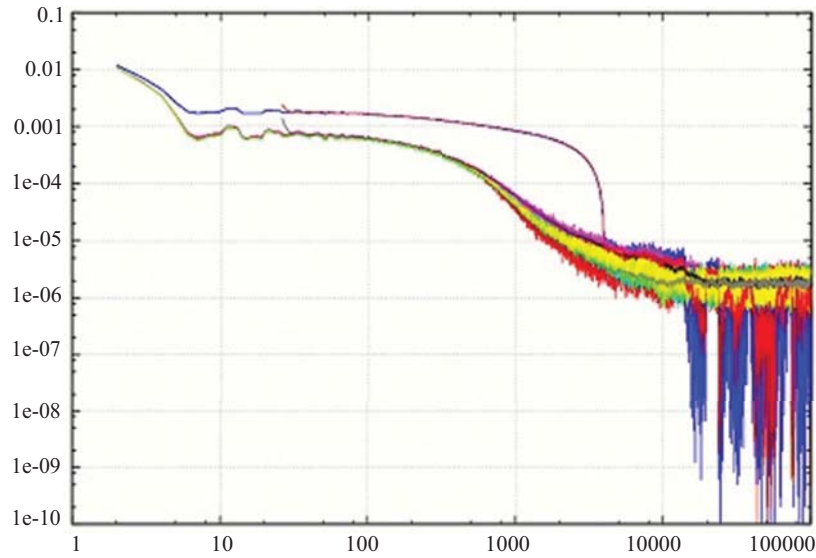


Figure 7. Mutual information $M(k)$ vs base separation k : *E. coli* strains: K12 – red; CFT073 – green; O157: H7-EDL933 – blue; O157: H7 –Sakai-pink; UTI89 – light blue; 536 – yellow; smoothed plots of mutual information O157: H7-EDL933 – rust; O157: H7 – Sakai-black; CFT073 – grey.

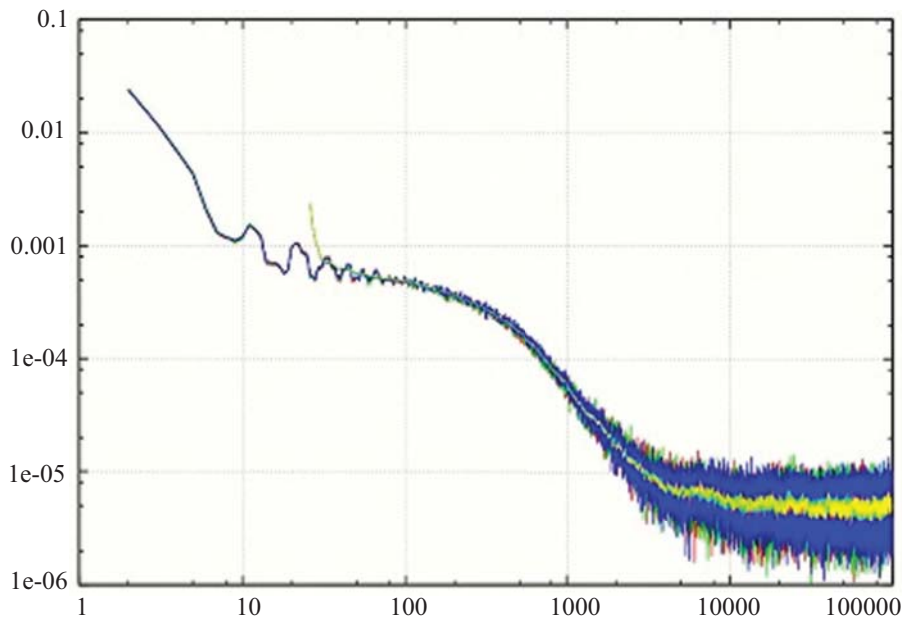


Figure 8. Mutual information $M(k)$ vs base separation k : *Helicobacter pylori* strains: 26695 – red; HPAG1 – green; J99 – blue; smoothed plot of mutual information: 26695 – pink; HPAG1 – light blue; J99 –yellow.

DNA, the C-C Correlation function for the plasmids of *C. perfringens*_SM101 and plasmids of *C. perfringens* 13 and *C. tetani* were plotted. *C. perfringens*_SM101 has also encapsulated a prophage phi SM101, the C-C correlation function for which was also plotted. It appears that there is no sequence similarity between the various plasmid DNA. It can safely be concluded that there will be a difference in the functional ability of each.

4.3 Escherichia coli strains

Figure 7 is plot for six strains of *Escherichia coli*: K12, UTI189, O157: H7 Sakai, O157:H7-EDL933, CFT073 and the recently sequenced 536 (Brzuszkiewicz *et al* 2006). Five of the strains show identical SSR structure but there is a striking departure in the $M(k)$ plot of *E. coli*_O157: H7-EDL933. The $M(k)$ plots, smoothed by window

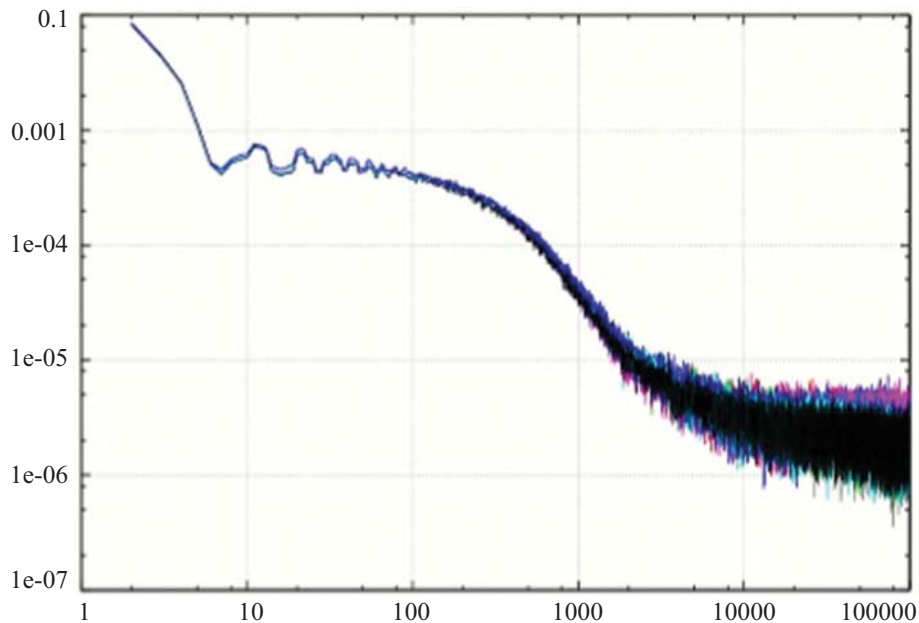


Figure 9. Mutual information $M(k)$ vs base separation k : *Yersinia*: *Y. pestis* strains: KIM – red; CO92 – green; Antiqua – black; Nepal516 – pink; biovar_Microtus – light blue; *Y. pseudotuberculosis* – blue.

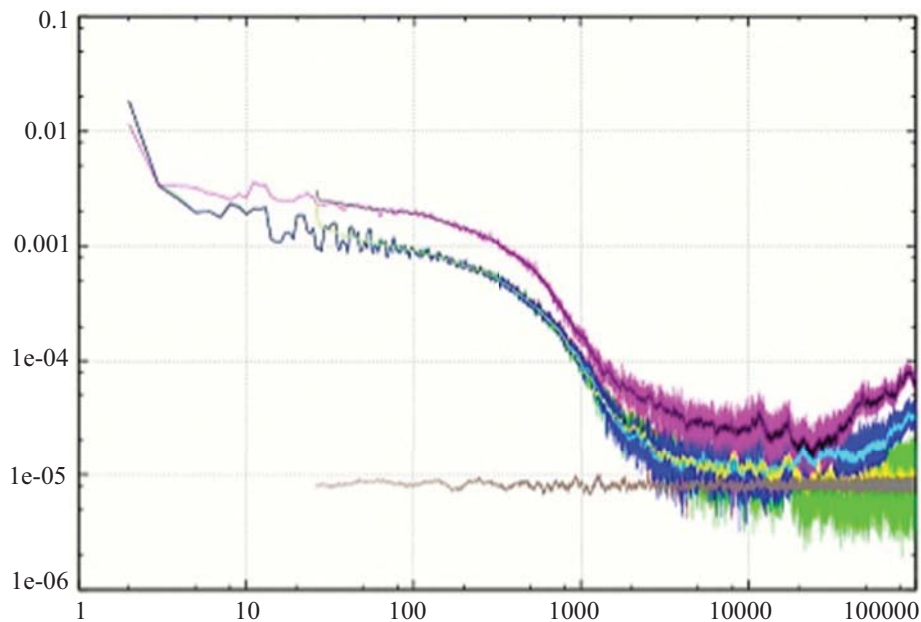


Figure 10. Mutual information $M(k)$ vs base separation k : *M. hyopneumoniae* strains: 7448 – red; 232 – green; J – pink; *M. synoviae_53* – blue; smoothed plot of mutual information: *M. synoviae_53* – light blue; *M. hyopneumoniae* strains: J – black; 232 – yellow; *M. synoviae_53* (randomized sequence) – grey.

size 51 show similar results in the variation of long-range behaviour. The two O157:H7 strains are compared with the laboratory strain K12 by Hayashi *et al* (2000) and by the author in a comparative study using the versatile

software MUMmer (Kurtz *et al* 2004), the results of which are shown in the Supplementary Materials Folder in figures 12–15. The MUMmer plots obtained show that the pathogenic strain O157:H7-EDL933 has an insertion

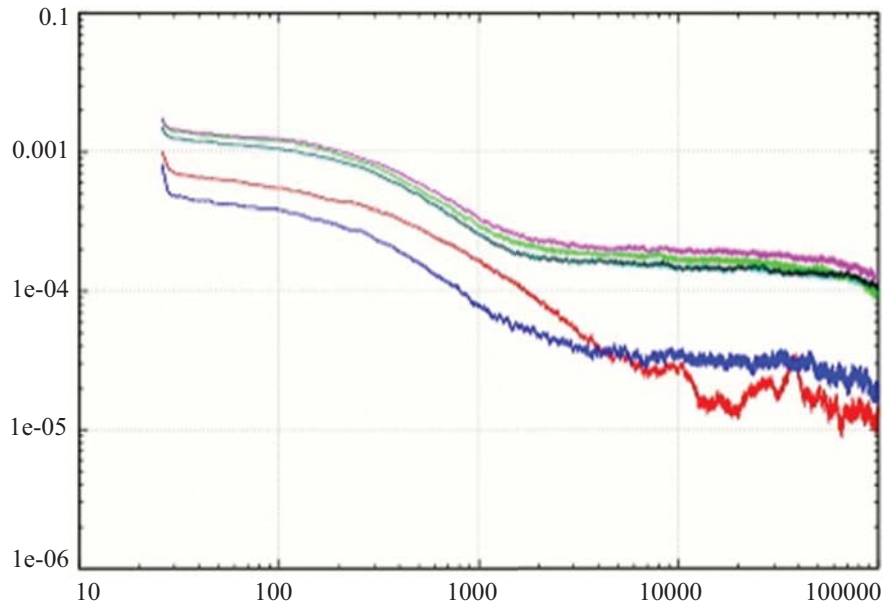


Figure 11. Mutual information $M(k)$ vs base separation k : *Ehrlichia* and *Anaplasmataceae*: *A. phagocytophilum* – red; *E. chaffeensis*_Arkansas – green; *E. canis*_Jake – pink; *E. ruminantium*_Gardel – light blue; *E. ruminantium*_Welgevonden – black; *N. sennetsu*_Miyayama – blue.

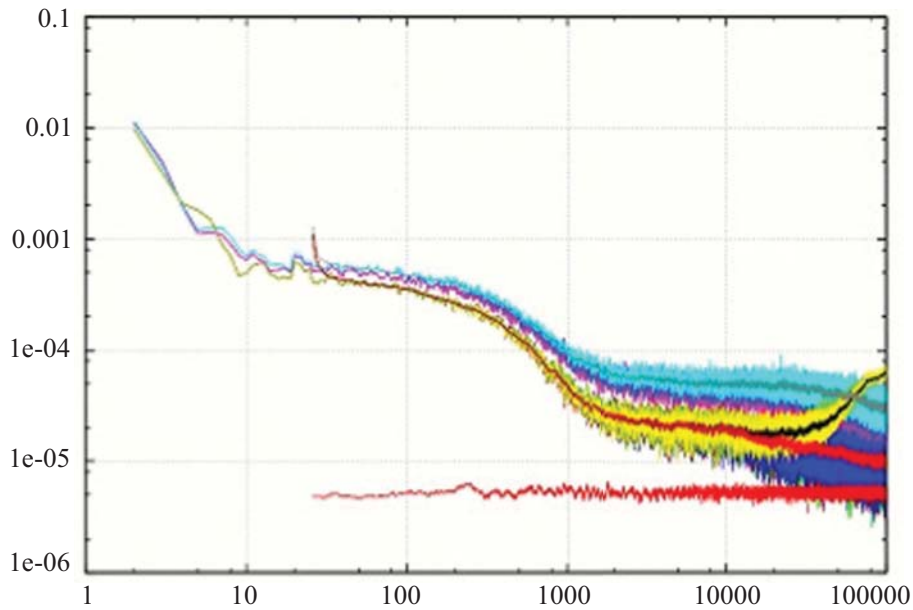


Figure 12. Mutual information $M(k)$ vs base separation k : *Chlamydiae*: *C. pneumoniae* strains: CWL029 – red; AR39 – green; TW183 – blue; J138 – yellow; *C. muridarum* – light blue; *C. trachomatis* – pink; smoothed plot of mutual information: *C. muridarum* – grey; *C. pneumoniae* strains: CWL029 – orange; J138 – black; J138 (randomized sequence) – rust.

sequence which is different from that of all the other strains sequenced and studied till date, including O157:H7-Sakai. The other strains have large regions of maximal matches with each other, one example of which is

shown in the figure 13 of the Supplementary Materials Folder. This is similar to the almost identical variation of the $M(k)$ plots of all strains other than O157:H7-EDL933 in figure 7.

4.4 *Helicobacter pylori* strains

Figure 8 shows highly similar plots for three strains of *H. pylori*. Even the large-range behaviour is strikingly similar. This may be due to the more than 90% sequence similarity as reported in Salama *et al* (2000). However, there are two regions of deletion in *H. pylori* 26695 and two distinct regions of insertion in *H. pylori* J138, which may account for the difference in their $M(k)$ plots. Further analysis is required before firm conclusions can be drawn.

4.5 *Yersinia*: *Yersinia pestis* strains

Yersinia pseudotuberculosis is shown as a counterpoint to the *Y. pestis* strains in figure 9. *Y. pestis* diverged from *Y. pseudotuberculosis* as recently as 20,000 years ago (Huang *et al* 2006) and there is a high level of homology at the DNA level. Their sequences are more than 90% alike and their 16 S rRNA are identical (Brubaker 1991), yet they differ markedly in their pathogenicity (Radnedge *et al* 2002; Hinchcliffe *et al* 2003). This behaviour is quite like that obtained for the *B. cereus*, *B. anthracis* and *B. thuringiensis*. These strains too possess plasmids with toxic genes, one that is common to all and others that are species- and strain-specific. The common plasmid called pLCR or pYV is found to be necessary for virulence. If it is lost, then the strain becomes avirulent. This holds even for the highly virulent strain *Y. pestis* which causes bubonic plague (Huang *et al* 2006).

4.6 *Mycoplasmataceae*: *Mycoplasma hyopneumoniae* strains

Figure 10 is a plot for the Mutual Information function for three strains of the porcine pathogen *M. hyopneumoniae*_232, _7448 and J along with the avian pathogen *Mycoplasma synoviae*_53. There is some variation among the strains for long-range correlations as evinced by the $M(k)$ plot in the figure. *Mycoplasma synoviae* also clearly shows a different SSR structure indicating that it belongs to a different species (Minion *et al* 2004; Vasconcelos 2005). *M. hyopneumoniae*_J has a larger asymptotic value of $M(k)$ than the other two *M. hyopneumoniae* strains. This is a noteworthy fact and in-depth analysis will confirm or negate as to whether the large asymptotic value of $M(k)$ is due to the presence of a large number of copies of repeats (Berryman and Abbott 2004; Johansson *et al* 2004).

4.7 *Ehrlichiae*: *Ehrlichia ruminantium*

The study of four different strains of *Ehrlichia*, compared and contrasted to *Anaplasma phagocytophilum* and *Neorickettsia sennetsu* Miyayama yield very interesting

results as shown in figures 11 and 4 of the Supplementary Material Folder. The $M(k)$ for *Ehrlichia canis*_Jake and *Ehrlichia chaffeensis* and three strains of *Ehrlichia ruminantium* fall close together, giving remarkably similar patterns but *Anaplasma phagocytophilum*, *Anaplasma marginale* and *Neorickettsia sennetsu*_Miyayama $M(k)$ are strikingly different. The similarity and differences between these closely related species is given in Dunning-Hotopp *et al* (2006). The asymptotic value of $M(k)$ of the *Ehrlichiae* is ten times larger than the *Anaplasma* species. Eight percent of the genomes of the *E. ruminantium* strains is occupied by large copies of variable number of tandem repeats (VNTR) as shown by Collins *et al* (2005). A recent study by Frutos *et al* (2006) has shown that *E. ruminantium* strains have a much higher number of repeats than the *Anaplasmas* and that these repeat structures in the intergenic regions of *E. ruminantium* may be taken as a marker of their subspecies. The $M(k)$ of *E. ruminantium*_Gardel and *E. ruminantium*_Welgevonden are identical, so much so that apart from being cultured at different places from different sources they seem to be the same. There are also two accession numbers assigned to *E. ruminantium*_Welgevonden, NC_006832 and NC_005925, as can be checked from the NCBI database. They too yield identical patterns.

4.8 *Chlamydiaceae*: *Chlamydia pneumoniae* strains

Figure 12 shows the Mutual Information function plots for four *C. pneumoniae* strains along with *C. trachomatis* and *C. muridarum*. The plot of $M(k)$ smoothed over window size 3 shows the SSR structure and the difference between *C. muridarum*, *C. trachomatis* and *C. pneumoniae* strains is visible. The superimposed plot of $M(k)$ smoothed over window size 51 shows that the asymptotic value of $M(k)$ is larger for *C. pneumoniae*_J138 strain than the others shown in figure 12. The reason for this variation is still to be studied. The Mutual Information function for the randomized sequence of *C. pneumoniae*_J138 is also plotted to give the minimum possible value for $M(k)$.

4.9 *Francisella tularensis* and *Legionella pneumophila* strains

Figure 13 shows the Mutual Information function plots for five strains of *F. tularensis*. Of these *F. tularensis*_tularensis strains are highly virulent and cause the disease tularaemia, whereas *F. tularensis*_holarctica strains are mildly virulent and *F. tularensis*_novidicida is avirulent. Figures 16 and 17 of the Supplementary Materials Folder also show a dissimilarity at primary sequence level of *F. tularensis*_Schu4 and *F. tularensis*_holarctica and similarity of the two strains of *F. tularensis*_holarctica. Of the two strains of

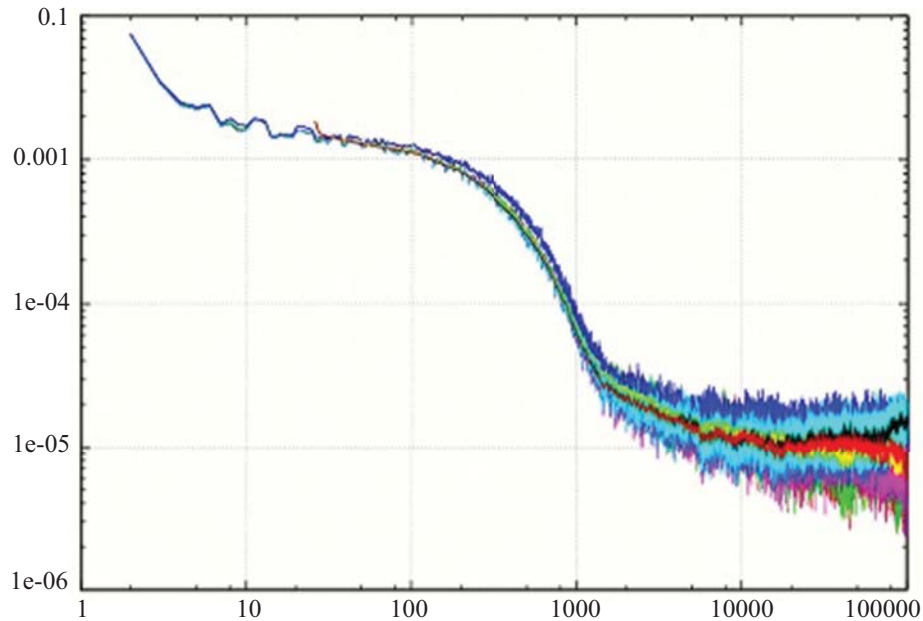


Figure 13. Mutual information $M(k)$ vs base separation k : *F. tularensis* strains: Schu4 – red; *tularensis* – green; *holartica* – pink; *holartica*-OSU18 – light blue; *novicida* – blue; smoothed plot of mutual information: Schu4 – yellow; *holartica* – rust; *holartica*-OSU18 – black.

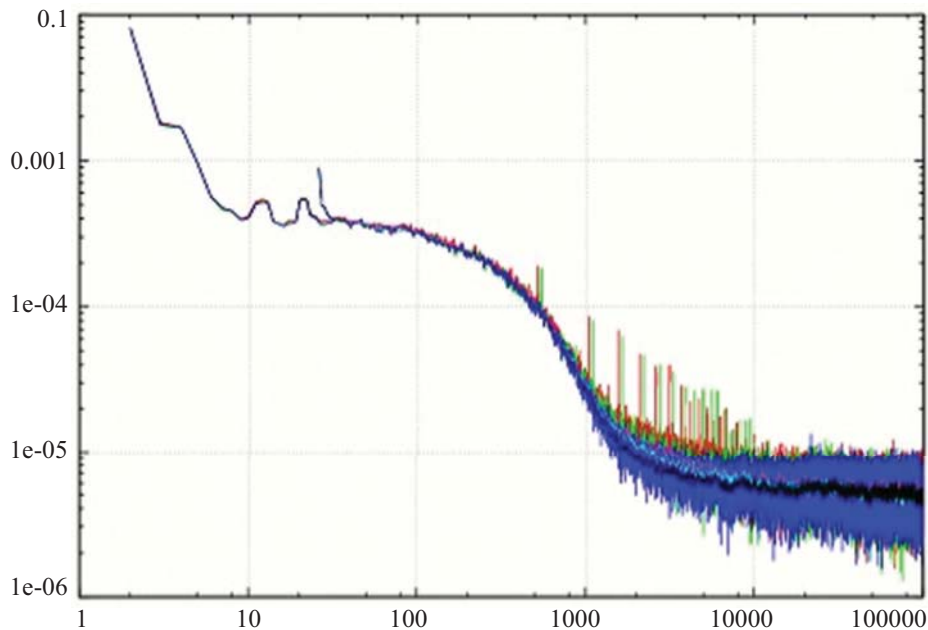


Figure 14. Mutual information $M(k)$ vs base separation k : *L. pneumophila* strains: Lens – red; Paris – green; Philadelphia-1 – blue; smoothed plot of mutual information: Lens – pink; Paris – light blue; Philadelphia-1 – black.

F. tularensis_holarctica, one has a circular chromosome like most other bacteria and the other, *F. tularensis_holarctica* OSU18 has a linear chromosome. The superimposed plot in figure 13 shows a slight increase in asymptotic value of $M(k)$ for *F. tularensis_holarctica* OSU18 vis-à-vis the asymptotic value of $M(k)$ for *F. tularensis_holarctica*. There is a hint here that the behaviour of $M(k)$ might be different for very

large values of k . This bears further investigation since the Mutual Information function for linear chromosomes of eukaryotes shows an increasing trend which may be due to the effect of telomeres. (D Swati, unpublished results).

Figure 14 shows the remarkable repeat structure in the $M(k)$ plots of the strains Paris and Lens of *Legionella*

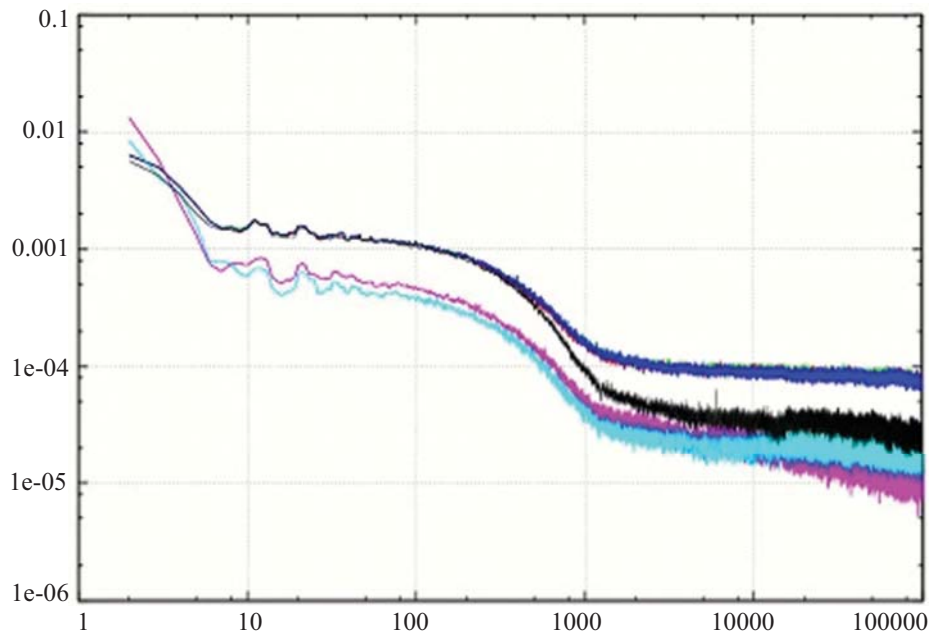


Figure 15. Mutual information $M(k)$ vs base separation k : *Bacillus*: *B. anthracis*_Ames_ancestor – red; *B. cereus*_ATCC-10987 – green; *B. thuringiensis*_konkukian-9727 – blue; *B. subtilis*_subtilis – pink; *B. halodurans*_C125 – light blue; *O. iheyensis*_HTE831 – black.

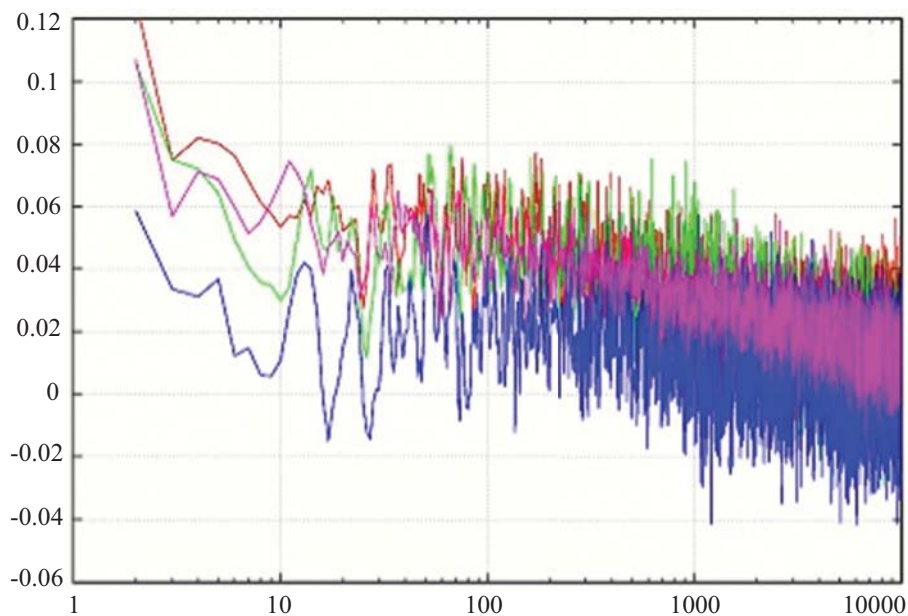


Figure 16. C-C correlation function vs base separation k : *B. anthracis*_Ames_ancestor-pXO1 – red; *B. anthracis*_Ames_ancestor-pXO2 – green; *B. cereus*_ATCC-10987-pBc – pink; *B. thuringiensis*_konkukian-9727-pBT – blue.

pneumophila. The plot shows an identical periodicity of 500 nucleotides or bases for these two strains but the *L. pneumophila*_Philadelphia-1 strain from a different continent does not have this signature SSR [verified through

appropriate window-averaging as well as independent repeat finding software (Kurtz *et al* 2004)]. The environmental effects on development of different strains may be surmised from this (Cazalet *et al* 2004).

4.10 Bacilli: *B. anthracis* and *B. cereus* strains

The comparison of three different strains of *B. anthracis* and *B. cereus* shows that they have remarkable identity (Ivanova *et al* 2003). It appears as if only one strain was being plotted. This is shown in figure 3 of the Supplementary Material Folder. Figure 15 shows the results of a comparison between five bacilli: *Bacillus subtilis subtilis*, *B. halodurans*_C-125, *B. cereus*_ATCC-10987, *B. thuringiensis serovar konkukian* 97-27, *B. anthracis*_Ames ancestor and *Oceanobacillus iheyensis*_HTE831. In this plot *B. subtilis subtilis* and *B. halodurans* and *O. iheyensis*_HTE831 all show different behaviour of Mutual Information function $M(k)$ with respect to base separation k for large k . This is indicative of a genome signature. There are a host of articles that have studied the closely related *B. anthracis*, *B. cereus* and *Bacillus thuringiensis*. Only few SNPs separate the genomes (Read *et al* 2002). A plethora of wet laboratory work involving PCR (Radnedge *et al* 2003), microarray (Zwick *et al* 2004), 16S and 23S rRNA analysis (Cherif *et al* 2003; Hoffmaster *et al* 2004) and the work of Daffonchio *et al* (2006) draw conflicting results about whether this near-identity in sequence implies that these particular strains of *B. cereus* and *B. thuringiensis* have become as pathogenic as *B. anthracis* and may cause anthrax.

After the plasmids of each *Bacillus* was sequenced and studied it was surmised that the opportunistic pathogen *B. cereus* or the insect pathogen *B. thuringiensis konkukian* could mimic a strain of *B. anthracis* only if it had acquired the plasmid pXO1 or pXO2 of *B. anthracis* by a horizontal gene transfer event. In figure 16 we see that the behaviour of $M(k)$ of the plasmid pBT of *B. thuringiensis* is different from the behaviour of $M(k)$ of the plasmid - pBc of *B. cereus* and the plasmids pXO1 and pXO2 of *B. anthracis*_Ames. To ensure that the roles of plasmids are different, the nucleotide–nucleotide correlation functions were studied. The results for the correlation function for C–C are displayed in figure 16. The difference in the C–C correlation functions for the various plasmids is clearly seen. Therefore the toxicity genes that reside on the plasmids pXO1 and pXO2 and are partly responsible for causing the disease anthrax are probably not generated by any subsequence of the plasmids pBT of *B. thuringiensis* and plasmid pBc of *B. cereus*. A MUMmer plot of plasmids pXO1 and pXO2 show some sequence identity as shown in figure 20 of the Supplementary Materials folder. The MUMmer plot of pBT and pXO1 does not show any match. We can tentatively conclude that near-identity of sequences of chromosomal DNA does not lead to similar pathogenic behaviour. The structure and function of the plasmid and chromosomal DNA taken as a whole should be considered before confirming the pathogenicity or otherwise of a particular bacteria.

5. Conclusion

To summarize, we can draw the following conclusions:

- $M(k)$ can be used to select coding regions. The scale of an average coding region can be easily read from the $M(k)$ versus k plot.
- The behaviour of $M(k)$ seems to depend on the GC%. More data should be generated to do a proper statistical analysis to draw this conclusion definitively.
- The plot of $M(k)$, after the three-periodicity is smoothed out, can be used to detect SSR if any.
- The similarity and difference between $M(k)$ of different strains are also revealed as shown by our limited but varied case studies. Mutual Information function $M(k)$ can be taken as a genome signature and used to identify new and controversial bacterial strains.
- The dramatic results of sequence similarity of *B. anthracis*_Ames with *B. cereus* ATCC 10987 and *B. thuringiensis serovar konkukian* is because the sequences are more than 90% similar. So much so, that 16S rRNA and 23S rRNA analysis cannot reveal the differences (Helgason *et al* 2000). However, the role of plasmids should be studied in detail before drawing conclusions about the presence or absence of pathogens causing anthrax or any other disease.
- The highly noteworthy difference shown in the $M(k)$ plot of *E. coli* O157:H7-EDL933 from others points to a probable horizontal gene transfer (HGT) event, but confirmation can only be done by doing wet-lab experiments or examining the sequence in segments and doing in-depth analysis.
- The use of the Mutual Information function is seen to be effective in identifying new strains *vis-a-vis* existing completely sequenced strains.
- The comparable method of spectral analysis using Fourier Transforms does not yield the varied results that the method of Mutual Information does. Even in this study, where we have not used it to study frame-shifts or dinucleotide–dinucleotide correlations and have thus limited its utility, we have got a measure of average coding sequence length, SSR structure and large-range correlations by the use of just one computational method. While specific features such as pathogenicity cannot be decided by sequence analysis alone without resorting to proteomics, Mutual Information function can serve as a genome signature and hence be used to identify strain similarities and differences.

Acknowledgements

The author would like to thank Professpr A Bhattacharya, Professor R Ramaswamy and Dr A Lynn of the Centre for Computational Biology and Bioinformatics, JNU, Delhi for

providing research facilities and encouragement for part of this work. The author is grateful to her senior colleagues, Professor Sushila Singh and Professor Uma Jaiswal and Dr D P Giri of MMV, BHU for help in developing the Computational Biology laboratory at MMV. Dr D Choudhury of CBT, JNU is accorded special thanks for stimulating discussions, so essential for progress in research.

References

- Benson G and Waterman M S 1994 A Method for Fast Database Search for all k-nucleotides Repeats; *Nucleic Acids Res.* **22** 4828–4836
- Berryman M J and Abbott R 2004 Mutual Information for Examining Correlations in DNA; *Fluctuation Noise Lett. (World Scientific)* **4** L237–L246
- Blattner F R, Plunkett G 3rd, Bloch C A, Perna N T, Burland V, Riley M, Collado-Vides J, Glasner J A D *et al* 1997 The Complete Genome Sequence of *Escherichia coli* K12; *Science* **277** 1453–1462
- Broekhuijsen M, Larsson P, Johansson A, Bystrom M, Eriksson U, Larsson E, Prior R G, Sjostedt A *et al* 2003 Genomewide DNA Microarray Analysis of *Francisella tularensis*; *J. Clin. Microbiol.* **41** 2924–2931
- Brubaker R R 1991 Factors promoting acute and chronic diseases caused by *Yersinia*; *Clin. Microbiol. Rev.* **4** 309–324
- Bruggemann H, Baumer S, Fricke W F, Weizer A, Leisegang H, Decker I, Herzberg C, Martinez-Arias R *et al* 2003 The Genome Sequence of *Clostridium tetani*, the causative agent of tetanus disease; *Proc. Natl. Acad. Sci. USA* **100** 1316–1334
- Brzuszkiewicz E, Bruggemann H, Liesegang H, Emmet M, Olschlager T, Nagy G, Alvermann K, Wagner C *et al* 2006 How to become a Uropathogen : Comparative Genomic Analysis of Extra-intestinal Pathogenic *Escherichia coli* Strains; *Proc. Natl. Acad. Sci. USA* **103** 12879–12884
- Cazalet C, Rusnoik C, Bruggemann H, Zidane N, Maguier A, Ma L, Tichit M, Jarraud S *et al* 2004 Evidence in the *Legionella pneumophila* genome for exploiting of host cell functions and gigh genome plasticity; *Nat. Genet.* **36** 1165–1173
- Chechetkin V R and Turygin V V 1996 Study of Correlation in DNA Sequences; *J. Theor. Biol.* **178** 205–217
- Chen S L, Huang C-S, Xu J, Reigstad C S, Magrini V, Sabo A, Blasier D, Bieri T *et al* 2006 Identification of Genome Subject to Positive Selection in Uropathogenic Strains of *Escherichia coli*: A comparative Genomics approach; *Proc Natl. Acad. Sci. USA* **103** 5977–5982
- Cherif A, Borin S, Rizzi A, Ouzari H, Boubadous A and Daffonchio D 2003 *Bacillus anthracis* diverges from related clades of the *Bacillus Cereus* group in 16S-23S ribosomal DNA intergenic transcribed spacers containing tRNA genes; *Appl. Environ. Microbiol.* **69** 33–40
- Colleta-Filho H D, Takita M A, de Souza A A, Aguilar-Vildoso C F and Machado M A 2001 Differentiating Strains of *Xyella fastidiosa* by a Variable Number of Tandem Repeat Analysis; *Appl. Environ. Microbiol.* **67** 4091–4105
- Collins N E, Liebenberg J, de Villiers E P, Brayton K A, Louw E, Pretorius A, Faber F E, van Heerde H *et al* 2005 The genome of the heart water agent *Ehrlichia ruminantium* contains multiple tandem repeats of actively variable copy number; *Proc. Natl. Acad. Sci. USA* **102** 834–843
- Daffonchio D, Raddadi N, Merabishvili M, Cherif A, Carmagnola L, Brusetti L, Rizzi A, Chanishvili N *et al* 2006 Strategy for Identification of *Bacillus cereus* and *Bacillus thuringiensis* and strains closely related to *Bacillus anthracis*; *Appl. Environ. Microbiol.* **72** 1295–1301
- Dunning-Hotopp J C, Lin M, Madupu R, Crabtree J, Angiuoli S V, Eisen J A, Seshadri R, Ren Q *et al* 2006 Comparative Genomics of Emerging Human Ehrlichiosis Agents; *PLOS Genet.* **2** 2008–2019
- Frutos R, Viari A, Ferraz C, Morgat A, Eychenie S, Kandassamy Y, Chantal I, Bensaïd A *et al* 2006 Comparative Genomics Analysis of Three Strains of *Ehrlichia ruminantium* Reveals an Active Process of Genome Size Plasticity; *J. Bacteriol.* **188** 2533–2542
- Grosse I, Herzel H, Buldyrev S V and Stanley H 2000 Species Independence of Mutual Information in Coding and Noncoding DNA; *Phys. Rev. E* **61** 5624–5629
- Han C S, Xie G, Challacombe J F, Altherr M R, Bhotika S S, Brown N, Bruce D, Campbell C S *et al* 2006 Pathogenomic Sequence Analysis of *Bacillus cereus* and *Bacillus thuringiensis* Isolates Closely related to *Bacillus anthracis*; *J. Bacteriol.* **188** 3382–3390
- Helgason E, Okstad O A, Caugent D A, Johansen H A, Fout A, Mock M, Hegna I, and Kolsto A-B 2000 *Bacillus anthracis* *Bacillus cereus* and *Bacillus thuringiensis* – One species on the basis of Genetic Evidence; *Appl. Environ. Microbiol.* **66** 2627–2630
- Herzel H, Trifonov E N, Weiss O and Grobe I 1998 Interpreting Correlations in Biosequences; *Physica A* **249** 449–459
- Hinchliffe S J, Isherwood K E, Stabler R A, Prentice M B, Rakin A, Nichols R A, Petra C F O, Hinds J *et al* 2003 Applications of DNA Microarrays to Study the Evolutionary Genomics of *Yersinia pestis* and *Yersinia pseudotuberculosis*; *Genome Res.* **13** 2018–2029
- Hoffmaster A R, Ravel J, Rasko D A, Chapman G D, Chute M D, Marston C K, De B K., Sachhi C T *et al* 2004 Identification of anthrax toxin genes in a *Bacillus cereus* associated with an illness resembling inhalation anthrax; *Proc. Natl. Acad. Sci. USA* **101** 8449–8454
- Holste D, Grosse I, Beirer S, Schieg P and Herzel H 2003 Repeats and correlations in human DNA sequences; *Phys. Rev. E* **67** 061913
- Huang X Z, Nikolich M P, and Lindler L E 2006 Current Trends in Plague Research: From Genomics to Virulence; *Clin. Med. Res.* **4** 189–199
- Ivanova N, Sorokin A, Anderson I, Galleron N, Candelon B, Kapatral V, Bhattacharya A, Reznik G *et al* 2003 Genome Sequence of *Bacillus Cereus* and Comparison to *Bacillus anthracis*; *Nature (London)* **423** 87–91
- Jensen G B, Hansen B M, Eilenberg J and Mahillon J 2003 The Hidden Lifestyles of *Bacillus Cereus* and Relatives; *Environ. Microbiol.* **5** 631–640
- Johansson A, Farlow J, Larsson P, Dukeich M, Chambers E, Bystrom M, Fox J, Chu M *et al* 2004 Worldwide Genetic Relationship among *Francisella tularensis* Isolates Determined by Multiple-Locus Variable Number Tandem Repeat Analysis; *J. Bacteriol.* **186** 5808–5818

- Keim P, Kalif A, Shlupp J, Hill K, Travis S E, Richmond K, Adair D M, Hugh-Jones M *et al* 1997 Molecular Evolution and diversity in *B anthracis* as detected by amplified fragment length polymorphism (AFLP); *J. Bacteriol.* **179** 818–824
- Kurtz S, Phillipy A, Delcher A L, Smoot M, Shumway M, Antonescu C and Salzberg S 2004 Versatile and open software for comparing large genomes; *Genome Biol.* **5** R12
- Li W 1990 Mutual Information Function versus Correlation Functions; *J. Stat. Phys. (Springer Verlag)* **60** 823–843
- Li W 1997 The Study of Correlation Structures of DNA Sequences: a Critical Review; *Comput. Chem.* **21** 257–271
- Lobozin V V and Chechetkin V R 2000 Order and correlations in genomic DNA sequences: The spectral approach; *Phys.-Uspekhi* **43** 55–78
- Minion F C, Lefkowitz E J, Madsen M L, Cleary D J, Swartzell S M and Mahairas G G 2004 Genome Sequence of Strain 230 The Agent of Swine Mycoplasmosis; *J. Bacteriol.* **189** 7123–7133
- Radnedge L, Agron P G, Worsham P L and Andersen G L 2002 Genome plasticity in *Yersinia pestis*; *Microbiology* **148** 1687–1698
- Radnedge L, Agron P G, Hill K K, Jackson P J, Ticknor L O, Keim P and Anderson G L 2003 Genome differences that distinguish *Bacillus anthracis* from *Bacillus cereus* and *Bacillus thuringiensis*; *Applied and Environmental Microbiology* **69** 2755–2764
- Rasko D A, Ravel J, Okstad O A, Helgason E, Cer R Z, Jiang L, Shores K A, Fouts D F *et al* 2004 The Genome sequence of *Bacillus cereus* ATCC 10987 reveals metabolic adaptations and a large plasmid related to *Bacillus anthracis* pX01; *Nucleic Acids Res.* **32** 977–988
- Read T D, Brunham R C, Shen C, Gill S R, Heidelberg J F, White O, Hickey E K, Petersen J *et al* 2000 Genome Sequences of *Chlamydia trachomatis* MoPn and *Chlamydia pneumoniae* AR39; *Nucleic Acids Res.* **28** 1397–1406
- Read T D, Salzberg S L, Pop M, Shumway M, Umayam L, Jiang L, Holtzapple E, Busch J D *et al* 2002 Comparative Genome Sequencing for Discovery of Novel Polymorphisms in *Bacillus anthracis*; *Science* **296** 2028–2033
- Read T D, Peterson S N, Tourasse N, Baillie L W, Paulsen I T, Nelson K E, Tettelin H, Fouts D E *et al* 2003 The Genome sequence of *Bacillus anthracis* Ames and comparison to closely related bacteria; *Nature (London)* **423** 871–888
- Rocap G, Larimer F W, Lamerdin J, Malfatti S, Chain P, Ahlgren A, Arellano A, Coleman M *et al* 2002 Resolution of *Prochlorococcus* and *Synechococcus* Ecotypes by using 16S – 23S Ribosomal DNA Internal Transcribed Spacer Sequences; *Appl. Environ. Microbiol.* **68** 1180–1191
- Rosselo-Mora R and Amann R 2001 The Species Concept for Prokaryotes; *FEMS Microbiol. Rev.* **25** 39–67
- Salama N, Guillermin K, McDaniel T K, Sherlock G, Tompkins L and Falkow S 2000 A Whole Genome Microarray Reveals Genetic Diversity among *Helicobacter pylori* strains; *Proc. Natl. Acad. Sci. USA* **97** 14668–14673
- Shannon C E 1948 A mathematical theory of communication; *Bell System Technical J.* **27** 379–423
- Shirai M, Hirakawa H, Kimoto M, Tabuchi M, Kishi F, Ouchi K, Shida T, Ishii K *et al* 2000 Comparison of Whole Genome Sequence of *Chlamydomonas pneumoniae* J138 from Japan and CWL029 from USA; *Nucleic Acids Res.* **28** 2311–2314
- Shimizu T, Ohtani K, Hirekawa H, Ohshima K, Yamashita A, Shiba T, Ogasawara N, Hattori M *et al* 2002 Complete Genome Sequence of *Clostridium Perfringens*, an Anaerobic Flesh-eater; *Proc. Natl. Acad. Sci. USA* **99** 996–1001
- Silverman B D and Linsker R 1986 A measure of DNA periodicity; *J. Theor. Biol.* **118** 295–300
- Swati D 2007 Use of Mutual Information Function and Power Spectra for Analyzing Some Prokaryotic Genomes; *Am. J. Math. Manag. Sci.* (in press)
- Tiwari S, Ramchandran S, Bhattacharya A, Bhattacharya S and Ramaswamy R 1997 Prediction of Probable Genes by Fourier Analysis of Genomic Sequences; *Comput. Appl. Biol. Sci.* **13** 263–270
- Tomb J-F, White O, Kerlavage A R, Clayton R, Sutton G G, Fleischmann R D, Ketchum K A, Klenk H P *et al* 1997 The complete genome sequence of the gastric pathogen *Helicobacter pylori*; *Nature (London)* **386** 515–516
- Vasconcelos A T, Ferreira H B, Bizirro C V, Banatto S L, Carvalho M O, Pinto P M, Almeida D F, Almeida L G P *et al* 2005 Swine and Poultry Pathogens: the Complete Genome Sequences of Two Strains of *Mycoplasma hyopneumoniae* and a Strain of *Mycoplasma synoviae*; *J. Bacteriol.* **187** 5568–5577
- Voss R F 1992 Evolution of Long-range Fractal Correlations and 1/f Noise in DNA Base Sequences; *Phys. Rev. Lett.* **68** 3805–3808
- Zwick M E, McAffee F, Cutler D J, Read T D, Ravel J, Bowman G R, Galloway D R and Mateczum A 2004 Microarray-based re-sequencing of multiple *Bacillus anthracis* isolates; *Genome Biol.* **6** R10

MS received 14 September 2006; accepted 14 April 2007

ePublication: 20 August 2007

Corresponding editor: SOMDATTA SINHNA