International Journal of Bifurcation and Chaos, Vol. 12, No. 11 (2002) 2625–2630

"SPATIAL-TEMPORAL" PATTERNS IN PROKARYOTE GENOMES

BAILIN HAO
Beijing Genomics Institute/Human Genome Center,
Institute of Theoretical Physics, Academia Sinica,
P. O. Box 2735, Beijing 100080, China

Received October 29, 2001

We have developed a simple scheme to visualize the string composition of long DNA sequences and applied it to all available prokaryote complete genomes. Each species has a specific "portrait" and most clearly seen patterns in these portraits are determined by short avoided and under-represented strings. The "spatial" patterns in the space of short strings might reflect some "temporal" events in the evolutionary history.

Keywords: Fractal; genome; evolution.

1. Introduction

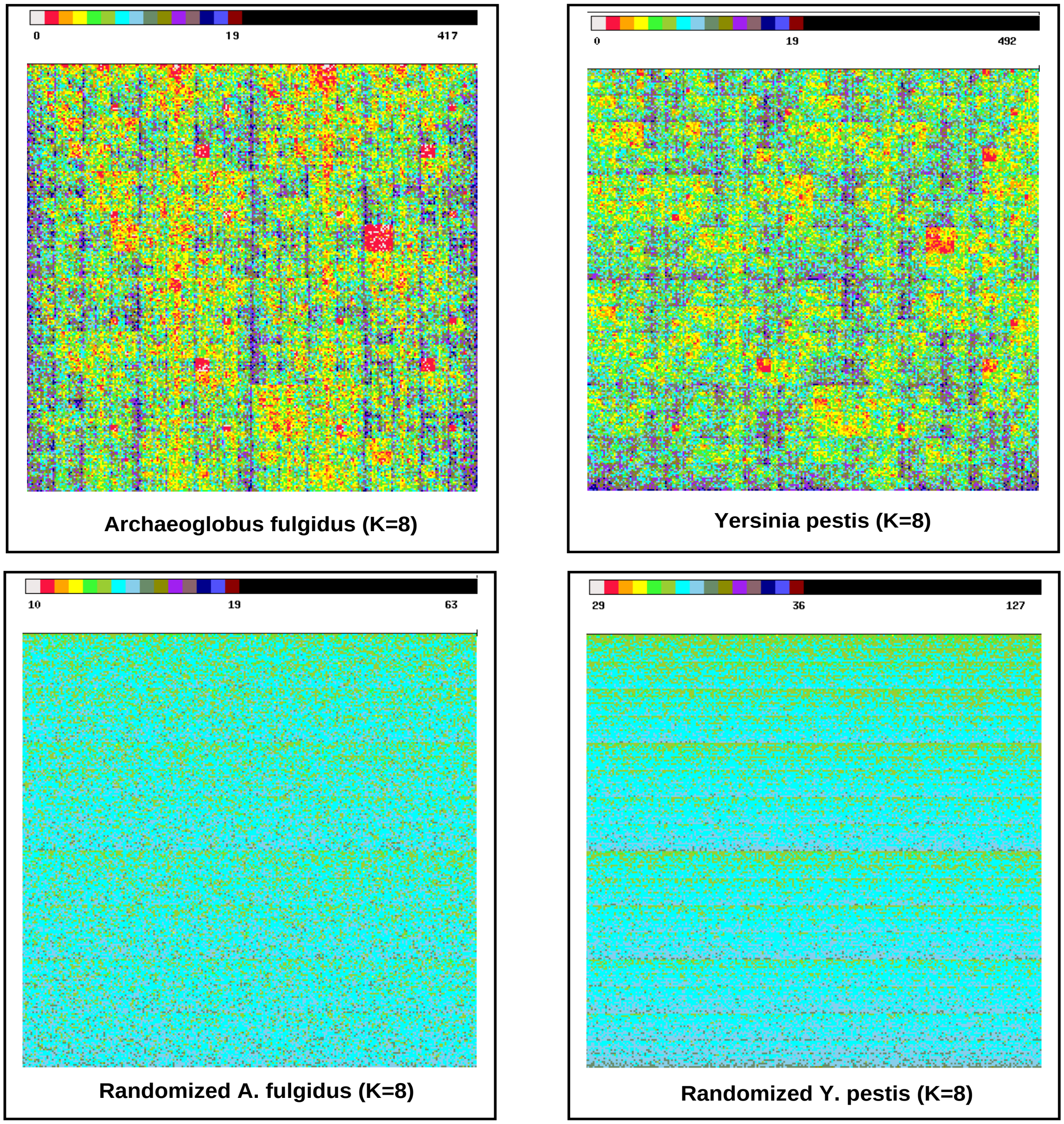
The availability of an ever growing number of complete genomes of various organisms enables one to ask many global questions. Perhaps the simplest such question is whether there are short nucleotide strings that are absent in a genome as one could not pose this question before when dealing with pieces of a genome. Take, for example, the *Archaeoglobus fulgidus* complete genome [Klenk *et al.*, 1997]. It is a circular DNA of 2 178 400 nucleotides/letters. Now count the frequency of appearance of all short strings of length $K = 7$. Since there are $4^7 = 16384$ different types of such strings, each type of strings would on average appear $2178400/16384 \approx 133$ times if the genome is a random sequence. Actually the distribution is very much biased towards smaller counts (see Fig. 2 below for histograms at $K = 8$) and there are four missing strings *gcgcgcg*, *cgcgcgc*, *gcactag* and *cactagt*. In order to clarify the situation for all available genomes we have developed a simple counting method and a visualization scheme [Hao *et al.*, 2000; Hao, 2000]. In this paper we put aside the biological implications of the findings and only explain why the patterns in the space of short symbolic strings of a given length may reflect some "temporal" events in the evolutionary history.

2. The Visualization Scheme

We call an oligonucleotide of length K a K-string. The study of K-string composition is a natural extension of $g + c$ content or CpG island analysis and the like. In order to visualize the K-string composition of a long DNA sequence a total of 4^K counters are needed. We display these counters as a 2^K by 2^K square matrix on a computer screen using a crude color code. The counters are arranged according to the location of the corresponding string in the direct product of K copies of the 2 by 2 matrix

$$M = \begin{pmatrix} g & c \\ a & t \end{pmatrix} . \tag{1}$$

It is easy to devise an algorithm that the computing time depends only linearly on the length of the DNA sequence, but not on K. Shown in Fig. 1 are what we call *composition portraits* or simply *portraits* of *A. fulgidus* [Klenk *et al.*, 1997] and *Yersinia pestis*, a recently sequenced bacterial



Archaeoglobus fulgidus (K=8)

Yersinia pestis (K=8)

Randomized A. fulgidus (K=8)

Randomized Y. pestis (K=8)

Fig. 1. Composition portraits of *A. fulgidus* and *Y. pestis*. Upper figures show the genomes; lower figures show the randomized sequences.

genome [Parkhill *et al.*, 2001]. At string length $K = 8$ a portrait shows a matrix of $256 \times 256 = 65532$ counters and a color code of 16 colors is used. In the upper figures the portraits of the original genomes are shown; in the lower figures the portraits of the corresponding randomized sequences are given. The genome sequences were randomized by using the `shuffleseq` program in the publicaly available European Molecular Biology Open Software System (EMBOSS) [The Emboss Package]. The program shuffles a given sequence keeping the number of each type of nucleotides unchanged. Different realization of the randomizing process yields similar portrait and only one realization is shown in Fig. 1.

In a portrait white color designates an avoided string and bright colors are allocated to small counts. When the counts exceed 40, the counters are shown in black. This is a useful "coarse-graining" in order to highlight the under-represented strings. The most prominent regular patterns in the two upper portraits are related to the under-representation of all strings containing a certain type of short strings. For example, there is a common set of patterns in both portraits which is caused by the under-representation of the tetranucleotide *ctag*. There are some interesting mathematical problems inspired by these regular fractal-like patterns which may be solved exactly by using combinatorial and language theory methods [Hao, 2000]. In the portraits of the randomized sequences all regular patterns disappear. The horizontal contrast in these figures is a consequence of the difference in $g + c$ versus $a + t$ content, as g and c are in the upper row of the matrix M and the counters for $g + c$-rich strings are more concentrated in the upper half of the figures.

3. Palindromic Signature of Avoided Strings

An inspection of all available genome portraits tells us that the most prominent patterns are related to the absence or under-representation of certain palindromic nucleotide strings. In DNA sequences a string is said to be palindromic if it is the same as its reverse-conjugate string; conjugation means interchanging the letters according to the Watson–Crick paring rule, i.e. interchanging $a \leftrightarrow t$ and $c \leftrightarrow g$. For example, in the *A. fulgidus* portraits at $K = 7$ the first avoided strings contain the palindromes *cgcgcg* (or *gcgcgc*), *actagt*, and *ctag*. The avoidance patterns are highly species-specific and we call them *palindromic signature* of the organism. We emphasize that the palindromic signature reflects specific features in strings that do *not* appear in a genome.

Palindromic oligonucleotides appear in many contexts of genomic analysis. In particular, many recognition sites of restriction endonucleases (of type II, to be precise), are palindromes of length 4 to 8. Restriction enzymes may be considered as part of the immune system of bacteria. A comparison of the palindromic signatures of closely related species reveals the following facts:

1. Different chromosomes of the same species have similar signatures.
2. Different strains of the same species have similar signatures.
3. Species in the same genus have similar signatures.
4. Different genera in the same family have similar signatures.

These facts hint on the close relation of palindromic signature to the activity of restriction endonucleases. There must be a time in the early years of evolution when there were no restriction enzymes at all and microbes lived happily in the primordial soup. One day a bacterium produced the first endonuclease which cut foreign DNA at, say, the *ctag* site. Many species were killed; only those containing less *ctag* strings survived. Later on bacteria developed in their defense system, e.g. the methylation enzymes to protect their own DNA from being cut by restriction endonuleases. Different conditions in various ecological niches caused different combinations of avoided or under-represented string patterns. What we see nowadays in the compositional portraits may well be a trace of these historic events and might not have actual meaning.

There are no avoidance patterns in the portraits of randomized sequences.

The species-specificity of avoidance patterns prompts one to infer phylogenetic relationships for prokaryotes from the palindromic signatures. However, the result does not make much sense. It seems that the main reason for the failure consists in that a palindromic signature provides too short a vector to represent a species. Instead of looking at avoided strings we have been trying to infer phylogeny from the K-string composition of different genomes. If all penta-peptides, i.e. $K = 5$ strings, are collected from the translated amino acid sequences of a genome, we would be able to work with compositional vectors of $20^5 = 3\,200\,000$ components. Our work along this direction has led to promising results [Qi *et al.*, 2001].

4. Distribution of Avoided Strings

The distribution of K-strings as seen in Fig. 1 may be expressed by histograms as shown in Fig. 2. The abscissa in the histograms show the counts from a minimal to a maximal number. The ordinate gives the number of counters whose counts fall in a small bin in the abscissa. The two upper figures in Fig. 2 show the histograms of the original genomes of *A. fulgidus* and *Y. pestis*, while the lower figures show that for the corresponding randomized sequences.

Due to different scales in these figures we did not put numbers on the axes. Instead the minimal, median (corresponding to the peak), and

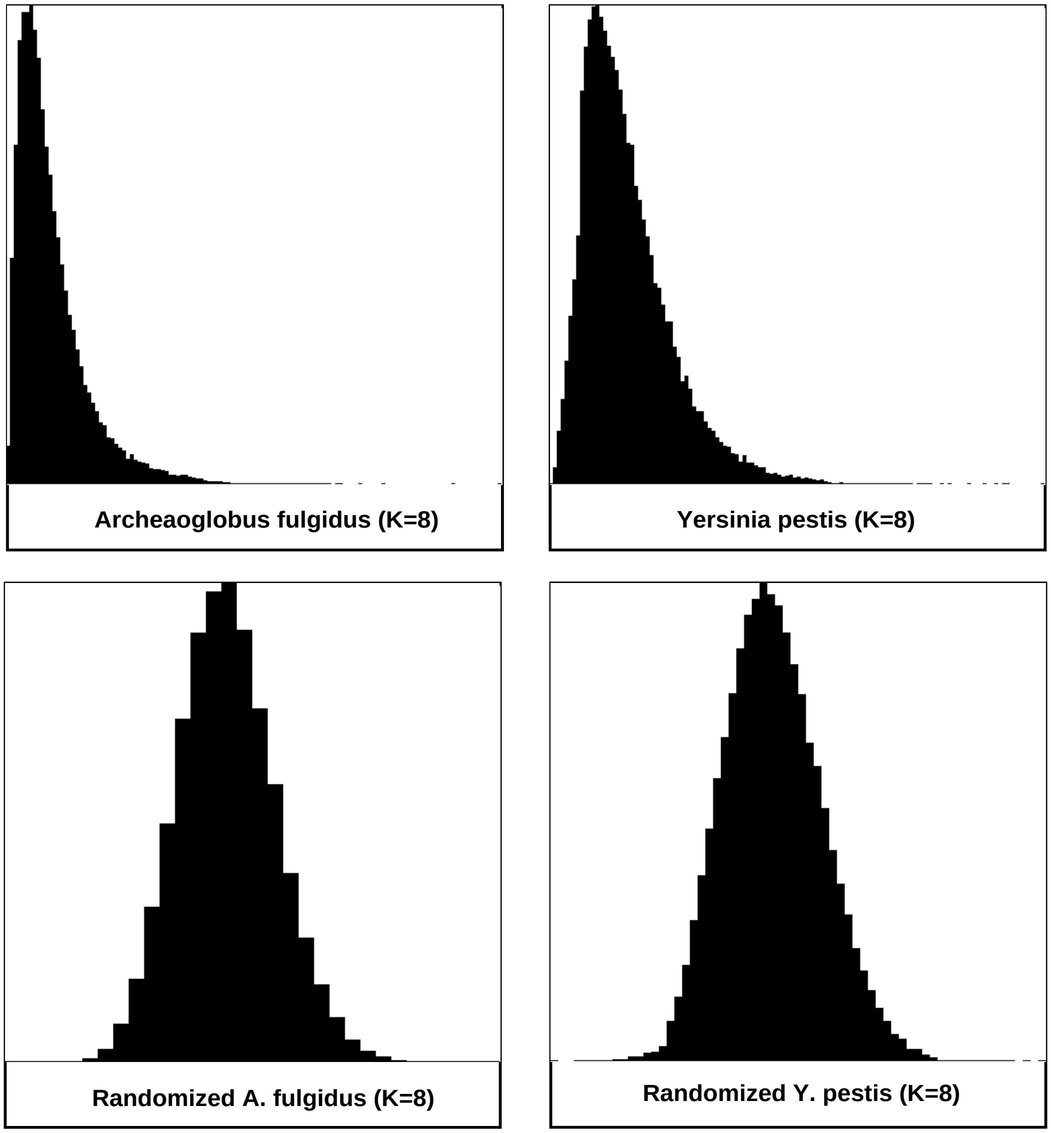


Fig. 2. Histograms of string counts in genomic sequences (upper figures) and their randomized counterparts (lower figures). For explanation of the abscissa and ordinate see text.

Table 1. Minimal, median and maximal counts in the histograms. Put in parentheses is the number of string types in that count range.

Sequence	Minimal	Median	Maximal	Average
A. fulgidus	0 (365)	$16 \sim 18$ (4544)	417 (4)	
Randomized *Aful*	10 (1)	$33 \sim 34$ (8241)	63 (1)	33
Yersinia pestis	0 (7)	$38 \sim 41$ (2972)	492 (1)	
Randomized *Ypest*	29 (1)	$68 \sim 69$ (4132)	127 (1)	71

maximal counts of the four histograms are collected in Table 1. Given in parentheses are the number of string types falling in the count range. The last column in Table 1 gives the average number of strings in each type if the sequence is assumed to be entirely random. By the way, the minimal and maximal counts also appear in Fig. 1 under the color code.

Take the *Yersinia pestis* genome as an example. This sequence of 4 653 728 nucleotides would give an average number of $4653728/4^8 = 71$ strings for each of the $4^8 = 65536$ string types at $K = 8$.

For the randomized *Y. pestis* histogram the peak is indeed centered around 70 and the minimal and maximal counts are 29 and 127, respectively. To the contrary, for the real *Y. pestis* genome the distribution is shifted towards smaller counts with a peak at the bin 38 $\sim$ 41 and the counts extend from zero to 492. This is yet another manifestation of the nonrandomness of the real genome sequences.

5. Application to Protein Sequences

Proteins are directed, nonbranching heteropolymers made of 20 different kinds of monomers — amino acids. As the definition of direct product of matrices applies to rectangular matrices as well, the visualization scheme discussed in this paper can be extended to proteins. Instead of the 2 by 2 matrix in Eq. (1), we define a 4 by 5 matrix

$$X = \begin{pmatrix} A & C & D & E & F \\ G & H & I & K & L \\ M & N & P & Q & R \\ S & T & V & W & Y \end{pmatrix}, \qquad (2)$$

where the matrix elements are the one-letter abbreviation of the amino acids [IUPAC-IUB Commission, 1970].

However, there are some peculiarities when the visualization scheme is used to display amino acid composition of proteins. As there are 20 letters the length of strings must not exceed 5 in order to have the picture displayed within a computer screen, otherwise one has to scroll the figure behind the visible window of the monitor. This technical complication may easily be dealt with when the necessity occurs.

A more essential difference consists in that protein sequences are much shorter than nucleic acids — from 50 to 5000 amino acids in most cases. Therefore, instead of looking at avoided and under-represented strings one must focus on those short polypeptides which are present in a protein. Here again one expects to extract "temporal" information from the "portraits". In the primordial soup short proteins of limited compositional variety must have predominated. These proteins must have taken only a small number of points in the compositional space made of short amino acid strings of a given length. With evolution going on these points spread slowly in the space of K-strings. Our preliminary study shows that at $K = 5$ the 3 200 000 points have not been used up yet by all the known proteins in the SWISS-PROT database.[1] In other words, for the time being the compositional space has not saturated yet and there is a good chance to infer some evolutionary history from studying the "diffusion" in this space. Our ongoing work along this line will be reported elsewhere.

6. Discussion

We mention in passing that no avoidance patterns have been observed in portraits of eukaryote genomes. The most clearly seen feature is related to contrast in $g+c$ content (for yeast chromosomes) or under-representation of cg as compared to gc (for segments of human DNA), a fact known before. This might have something to do with the fact that no homologs of restriction enzymes are known to be present in eukaryote genomes.

Before concluding this paper we would like to comment on the relation of our visualization scheme with so-called Chaos Game Representation (CGR) [Jeffrey, 1990] of DNA sequences. If drawn in black/white our compositional portraits may look like the CGR figures. In fact, recently Peter Tiňo has proved that the multifractal characteristics of both schemes can be translated from each other [Tiňo, 2001]. However, there are several essential differences between the two schemes. First, our portraits reflect frequency distribution of K-strings by definition, while in CGR the density information cannot be obtained in one run. One has to introduce "cells" and to count the number of points in each cell to allow for using color code. Second, the precision in our method is under control: it is just the string length K. In the CGR approach the precision is limited by the screen resolution and may be different in different directions. Third, the algorithm of CGR is somewhat more complicated, whereas ours is simple counting. Therefore, it seems that the CGR method may be replaced entirely by ours.

Acknowledgments

This work was partly supported by the Chinese Natural Science Foundation, the Special Funds for Major State Basic Research Projects, and the

[1] For more details refer to the References section.

Beijing Municipality "248 Project". It is a great pleasure to dedicate this paper to the celebration of the sixtieth birthday of Prof. Manuel G. Velarde.

References

Hao, B.-L. [2000] "Fractals from genomes — exact solutions of a biology-inspired problem," *Physica* **A282**, 225–246.

Hao, B.-L., Lee, H. C. & Zhang, S.-Y. [2000] "Fractals related to long DNA sequences and bacterial complete genomes," *Chaos Solit. Fract.* **11**, 825–836.

IUPAC-IUB Commision on Biochemical Nomenclature [1970] "Abbreviations and symbols for nucleic acids, polynucleotides and their constituents," *Eur. J. Biochem.* **15**, 203–208.

Jeffrey, H. J. [1990] "Chaos game representation of gene structure," *Nucleic Acid Res.* **18**, 2163–2170.

Klenk, H. P. *et al.* [1997] "The complete genome sequence of the hyperthermophilic, sulphate-reducing archaeon Archaeoglobus fulgidus," *Nature* **390**, 364–370.

Parkhill, J. *et al.* [2001] "Genome sequence of Yersinia pestis, the causative agent of plaque," *Nature* **413**, 523–527.

Qi, J., Wang, B. & Hao, B.-L. [2001] "Prokaryote phylogeny based on complete genomes," in preparation.

The EMBOSS package may be fetched by anonymous `ftp` from: ftp://ftp.uk.embnet.org/pub/EMBOSS/

The URL of SWISS-PROT database: http://www.expasy.ch/sprot/

Tiňo, P. [2002] "Multifractal properties of Hao's geometric representations of DNA sequences," *Physica* **A304**, 480–494.