

# Large-Scale Patterns in DNA Texts

Wentian Li

originally prepared for Scientific American. March 1999

**An intriguing statistical pattern, called one-over-f spectra, has been observed in many DNA sequences. The same pattern was also found in temporal variation of music. Are there any connections?**

The genome – the complete set of genetic material – of a living organism is made up of long DNA molecules, with four nucleotides (A, C, G, and T) arranged in a linear sequence [see “DNA”, by Gary Felsenfeld, Scientific American, October 1995]. The “code” contained in DNA sequences determines what this organism is, and to a large extent, what it does. Obtaining the complete DNA sequences is similar to getting a blueprint, that encodes all the details of a living organism. The current human genome project is attempting to identify the complete DNA sequence for humans, as well as the genomes of many other animals, plants, and bacteria [see “Hacking The Genome”, by Deborah Erickson, Scientific American, April 1992, and “Vital Data”, by Tim Beardsley, Scientific American, March, 1996].

When a DNA sequence is considered as a symbolic text, are there any statistical patterns present in the arrangement of the symbols? If some pattern exists, how can it be explained by the evolutionary process that led to the DNA sequence? Researchers have been able to address these questions only recently as sequences of more organisms have become available for analysis. A few types of patterns were already known to researchers. One was the statistical correlation between neighboring nucleotides. This correlation was easily explained by the presence of codons, the three consecutive nucleotides coding for one amino acid [see, “The Genetic Code”, by Francis H. Crick, Scientific American, October, 1962, and “The Genetic Code II”, by Marshall W. Nirenberg, Scientific American, March 1963]. These amino acids are the building blocks of protein molecules. Since each amino acid requires specific codons in DNA sequences, the nearest neighbor nucleotides are correlated.

The presence of the three-nucleotide codon in a protein-coding DNA sequence produces a global cyclic pattern of a repeating unit of three. This periodicity is a signature of the protein-coding region in a DNA sequence and has been used for computational DNA sequence analyses to automatically recognize protein-coding regions. Another pattern observed in DNA sequences is an approximate cyclic pattern of a repeating unit of around ten to eleven nucleotides. This periodicity matches the helix turns in the three-dimensional structure of the DNA molecule, and was proposed to aid the bending of the DNA molecule.

Besides these three well known patterns: nearest neighbor correlation, repeating patterns of three, and ten to eleven nucleotides, a new kind of statistical pattern in much larger length scales was discovered in 1992. Interestingly, similar patterns have been observed in a large number of temporal variations in nature, including the luminosity of stars, the electronic current in conductors, semiconductors, and superconductor devices, photocurrent in lasers and other optical devices, highway traffic, internet traffic, and even music. This statistical regularity is called one-over-f (“ $1/f$ ”, where  $f$  stands for frequency) noise or one-over-f spectrum [see “Mathematical Games”, by Martin Gardner, *Scientific American*, April 1978]. To understand one-over-f noise, it is necessary to first understand the Fourier transformation and spectral analysis [see “Fourier Transform”, by Ronald N. Bracewell, *Scientific American*, April 1992].

A Fourier transform is a way to express a numerical series by periodic signals. In the Fourier representation, a series becomes a sum of many sine or cosine functions with different periodicities. The frequency of a periodic signal is inversely proportional to the periodicity: the longer the period, the lower the frequency. The amplitude of a particular sine or cosine function measures the contribution of that periodic signal to the series. Actually, a better measure of the contribution from that sine or cosine function is the square of the amplitude, which is also called *power*. Power plotted as a function of the frequency is called a *power spectrum*, and the procedure that leads to the power spectrum, is called the *spectral analysis*.

One-over-f noise or one-over-f spectra are numerical series whose power spectra are approximately inversely proportional to the frequency, within certain frequency ranges. This name also covers power spectra that are the inverse function raised to certain powers. For example, an inverse function raised to the power of 0.5 is a power spectrum that is inversely proportional to the square root of the frequency. How does one-over-f noise differ from other noise? First, these numerical series receive considerable contributions from low-frequency (long-period) signals. Second, the increase of power as the frequency decreases follows a specific power-law function, which reflects a specific self-similarity in the original

sequence. Benoit Mandelbrot of the Thomas J Watson Research Center of IBM coined the word *fractal* to describe this type of self-similar signals or objects [see, “The Language of Fractal”, by Hartmut Jürgens, Heinz-Otto Peitgen, Dietmar Saupe, Scientific American, August 1990, and “A Multifractal Walk Down Wall Street”, by Benoit B Mandelbrot, Scientific American, February 1999], but one-over-f noise was observed and studied long before the fractal became a popular topic in scientific research.

### Expansion-Modification Systems

I came upon one-over-f spectra in DNA sequences by a few unexpected events that started when I was a graduate student. Working with Stephen Wolfram, then at the Center for Complex Systems Research at University of Illinois at Urbana-Champaign, I was given a project to examine whether a class of locally interacting computer models called *cellular automata* [see “Computer Software in Science and Mathematics”, by Stephen Wolfram, Scientific American, September 1984] were able to produce long-range self-similar correlations. Similar to DNA sequences, the object to be manipulated in a cellular automaton is a symbolic sequence. A cellular automaton updates a symbolic sequence repeatedly by a locally operating rule, until a stationary state is reached.

While pondering why cellular automata were unable to generate long-range statistical correlations in my study, an idea occurred to me that maybe the sequence should be allowed to increase, unlike cellular automata where the sequence length was fixed. Eventually I came up with a length-increasing model called the *expansion-modification system*, which contains only two operations. The first operation is *modification*, or *mutation*, which switches a symbol to another symbol. The second operation is *expansion*, or *duplication*, which turns one symbol to two symbols. Each symbol in the sequence could either be duplicated or mutated, and all symbols are updated repeatedly until a very long sequence is generated (see Illustration 1). To my delight, the expansion-modification system was able to generate sequences with one-over-f spectra.

Learning about this result, Kunihiko Kaneko of Tokyo University, who was visiting the Center for Complex Systems Research at the time, suggested that the expansion-modification model might capture the essence of the evolution of a DNA sequence: the expansion would be the nucleotide duplication process, and modification the mutation process. At the time, Kuni was reading a book by Susumo Ohno of the City of Hope Medical Center titled *Evolution by Gene Duplication*, and he was convinced that due to the prevailing presence of duplication events described in Ohno’s book, plus the prediction of one-over-f spectra by the expansion-modification model, one-over-f spectra could be found in DNA sequences.

Anxious to see whether this prediction was indeed true, Kuni and I talked to people at the GenBank

of Los Alamos National Laboratory (LANL), where Kuni continued his visit in the U.S. at the Center for Nonlinear Studies at LANL. We retrieved all the longest DNA sequences available, ran these sequences through spectral analysis. Unfortunately, we were unable to see a clear one-over-f spectrum. By that time, I had finished my Ph.D and was a postdoctoral fellow at the Santa Fe Institute in New Mexico. I learned of the work by Alan Lapedes of Los Alamos National Lab and his collaborators on distinguishing protein coding and non-coding DNA sequences, and I requested more sequences from him. It was then that I realized that protein coding and non-coding DNA sequences exhibited different statistical patterns, and non-coding sequences tend to extend correlations to longer distances than protein coding sequences. Then, in one of the human non-coding sequences, I found indications of the power roughly increasing inversely with the frequency.

Unknown to us, Ary Goldberger of Harvard Medical School, who had been studying chaos and self-similar patterns in heart beat time series [see “Chaos and Fractals in Human Physiology”, by Ary L Goldberger, David R Rigney, Bruce J West, *Scientific American*, February 1990], was wondering whether DNA sequences also exhibited self-similar patterns. Goldberger, Eugene Stanley and Sergey Buldyrev of Boston University, their student Chung-Kang Peng, as well as other collaborators used a direct measure of self-similarity by converting a DNA sequence into a trajectory that fluctuates according to the nucleotides in the DNA sequence. This type of trajectory is common for the thermal motion of small particles, and is called random walk in physics.

A random walk converted from a random sequence is also self-similar, and it has a specific scaling exponent. A deviation from this special value would signal a non-randomness in the DNA sequence. The Boston team indeed found evidence for non-trivial self-similarity in non-coding DNA sequences, but no evidence was found in coding sequences. They illustrated the self-similarity in DNA sequences by a typical approach in fractal research, stating that the enlarged version of a random walk plot that corresponds to a DNA sequence of a short length, looked similar to the original plot corresponding to the longer DNA sequence; and the enlargement could be carried out recursively.

Meanwhile, Goldberger also discussed this idea with Richard Voss, a research scientist then at the Thomas J Watson Research Center of IBM, and a pioneer in the study of one-over-f spectra and fractals. In fact, Voss was the first person to show the one-over-f spectra in music. He subsequently published a paper on one-over-f spectra in a mixture of protein coding and non-coding DNA sequences with a layout similar to his paper on one-over-f spectra in music: in the music paper, power spectra for classical music, rock-n-roll, and talk radio were displayed side by side, and in the DNA sequence paper, it was the spectra of primates, rodents, plants, and so on. The universal shape of the power spectra in both cases

was unmistakable. I remember clearly the day I saw Voss' paper in the library: the title of his paper was almost identical to mine, whereas the widespread one-over-f spectra in DNA sequences present in his paper were more striking.

### Coding, Non-coding, and Complete Sequences

Supposing that duplication or other large-scale shuffling of DNA segments was responsible for the one-over-f spectra, one can argue that they are more likely to be present in non-coding sequences because that was where these processes could occur. On the other hand, one would be less likely to find these dynamical changes in protein coding sequences. Any alteration of the nucleotides in a coding sequence could easily alter its ability to synthesize a functional protein molecule. It was thus not surprising that the first few reports of one-over-f spectra and long-range correlation were found in non-coding sequences.

Nevertheless, both coding and non-coding sequences have limited length. A typical protein sequence may contain 300 amino acids, which correspond to 900 nucleotides. Coding sequences in higher organisms are usually interrupted by pieces of non-coding sequences called *introns* [see "Split Genes", by Pierre Chambon, Scientific American, May 1981]. There are also non-coding sequences between genes, but the spacing between two genes cannot be arbitrarily long. In order to study even larger-scale patterns in DNA sequences, one needs to look at the whole genome.

These entire genome sequences have been generated by sequencing machines at an increasing rate in the last few years. The genomic sequences of at least eighteen bacteria (including the famous intestinal bacteria *Escherichia coli*), the yeast *Saccharomyces cerevisiae*, and more recently, the worm *Caenorhabditis elegans*, have been fully determined. There are sixteen chromosomes in the yeast genome, and six in the worm genome. Illustration 2 shows the power spectra of five complete representative sequences. Both the horizontal (frequency) and the vertical (power) axes are drawn on a logarithmic scale. A one-over-f power spectrum is a straight line with a negative slope of 1, and that function raised to the power of 0.5 is a straight line with a negative slope of one-half. These plots show not only the existence of one-over-f spectra, but also a surprisingly long extension of one-over-f spectra to the lowest frequency range (longest spatial distance).

The large-scale pattern characterized by the one-over-f spectra is closely related to another topic studied in DNA sequences: the heterogeneity of the nucleotide concentration along the sequence. Some regions of DNA sequences may contain more nucleotides C's and G's, others more A's and T's. Giorgio Bernardi of the Institut Jacques Monod in France and his collaborators found that many genomes including the human genome contain long regions alternatingly rich in C and G, and regions rich in A

and T. Bernardi called these compositionally homogeneous regions *isochores*, from the Greek for “equal landscapes”.

The isochores, or alternating homogeneous regions along a DNA sequence, naturally lead to a large-scale pattern in the DNA text. But the one-over- $f$  spectra require something more. The self-similarity implied by a one-over- $f$  spectrum indicates that the partitioning of a long DNA sequence into compositionally different regions can also be carried out on a shorter sequence, in particular, an isochore itself. In other words, a seemingly homogeneous DNA sequence such as an isochore can be weakly heterogeneous on a smaller scale, and the distinction between homogeneity and heterogeneity is relative. This “domains within domains” phenomenon is a direct consequence of the self-similarity and lack of intrinsic length scale in DNA sequences.

José Oliver, Ramón Román-Roldán, and their student Pedro Bernaola-Galván at the University of Granada in Spain, pursued this idea in more detail. They designed a computer algorithm to partition a DNA sequence into relatively homogeneous regions with a given level of stringency. Relaxing the stringency makes the partitioning easier, and consequently reveals subdomains within a domain. Using a series of stringency levels, the self-similarity can be studied at different length scales. They have shown convincingly that a recursive partition is possible mainly in DNA sequences with one-over- $f$  spectra. In some sense, sequences with one-over- $f$  spectra are more *complex* than those without. These researchers also proposed a measure of the complexity related to the recursive partition process, and were able to observe a distinction between protein coding and non-coding sequences in this complexity measure.

### DNA, Language, and Music

DNA sequences are frequently compared with human language: nucleotides are analogous to letters, codons to words, genes to sentences, etc. It is tempting to push this analogy further. David Searls of SmithKline Beecham Pharmaceuticals, for example, uses the linguistic analogy extensively in his work on computational gene recognition. The expansion-modification system provides another connection: it belongs to a class of formal languages, called *context-free languages*, proposed by Noam Chomsky of Massachusetts Institute of Technology [see, John Horgan, “A Word (or Two) About Linguist Noam Chomsky, Scientific American, May 1990] as an approximate grammar for human language. A context-free language is characterized by the tree-like branching in the language generation, for example, when it is used to match a left and a right parenthesis. This tree structure is also the key reason why the expansion-modification model was able to generate long-range correlations.

There was another puzzling statistical regularity in human languages which was first discovered by the late George Zipf of Harvard University. He plotted the number of times a word used in a text

as a function of the rank of that word: the most frequently used word is ranked 1, the next most frequently used word ranked 2, etc. Zipf observed that the frequency of the usage of a word was inversely proportional to the rank of the word, in many language texts he studied. This is the so-called Zipf's law.

Some researchers claimed that DNA sequences exhibit a similar Zipf's law, and, as a result, that they share a key linguistic feature with human languages. Unfortunately, this controversial claim is problematic in several respects. The definition of a word in DNA sequences in this study was not comparable to that in a human language. Also, the Zipf's plot in this study was far different from the inverse power-law observed in human languages. More importantly, although proven a long time ago by Mandelbrot, and re-stated several times after that, it was not widely known that even random texts such as, for example, the ones typed by a monkey, follow Zipf's law in a way similar to human languages. In other words, Zipf's law is not a true linguistic regularity.

With the one-over-f spectra being observed in both music and DNA sequences, it is perhaps natural to compare the two. Susumo Ohno proposed that the repetition of the same theme followed by a variation is a principle for both DNA evolution and musical composition. In DNA evolution, a duplicated gene acquires a new function after being modified. In a musical composition, the same musical motif is repeated, not exactly but with variations. Exact repetition is boring, but repetition with variation is creation.

Ohno and his wife Midori even transformed DNA sequences into musical score, and transformed Chopin and Bach's music back to pseudo DNA sequences. Besides the attempt by the Ohno's, composing what is called "DNA music" or "protein music" has become a cottage industry in recent years. Web pages appear where one can play online a piece of "human sex hormone"; or if one prefers, listen to customized DNA music based on your own DNA sequence! Repetition with a variation, expansion-modification systems, one-over-f spectrum, domains-within-domains, the dynamical process in evolution and composition: these isolated bridges between music and DNA sequences now emerge as a perhaps more coherent and more meaningful framework.

The discovery of the self-similar statistical pattern in DNA sequences was not the end of the story. Is the domains-within-domain phenomenon in DNA texts related to the hierarchical three-dimensional structure of the DNA molecule? Can we reconstruct a detailed history of global-scale changes in DNA sequences? Were these evolutionary processes acting upon different length scales? How was the seemingly purposeless duplication of DNA segments related to the duplication of genes proposed by Ohno as a driving force of biological evolution and innovation? Even though we do not know all the

answers to these questions, it is unlikely that the presence of large-scale self-similar patterns in DNA texts is accidental.

## Further Readings

- Evolution by Gene Duplication. S.Ohno. Springer-Verlag, 1970.
- Expansion-modification systems: a model for spatial  $1/f$  spectra. W.Li in *Physical Review A*, Vol.43, No.10, pages 5240-5260, 1991.
- Evolution of long-range fractal correlations and  $1/f$  noise in DNA base sequences. R.F. Voss in *Physical Review Letters*, Vol. 68, No. 25, pages 3805-3808, 1992.
- The study of correlation structures of DNA sequences – a critical review. W. Li in *Computer & Chemistry*, Vol.21, No.4, pages 257-272, 1997.
- The complexity of DNA: the measure of compositional heterogeneity in DNA sequences and measures of complexity. W.Li in *Complexity*, Vol.3, No.2, pages 33-37, 1997.

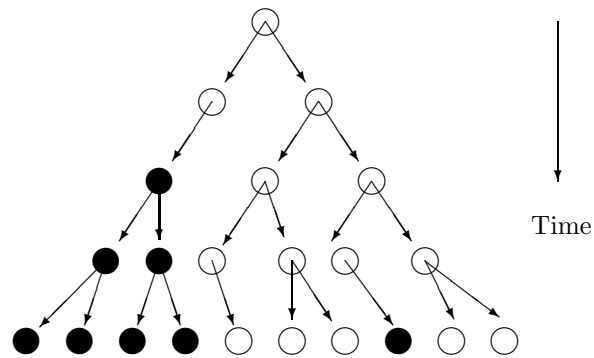


Figure 1: **An expansion-modification model:** Black and white balls represent two different symbols or two types of nucleotide in DNA sequences, and the drawing illustrates an updating of a symbolic sequence or a possible evolutionary process of a DNA sequence. In this process, each symbol experiences either a duplication to create two symbols of the same type, or a mutation to switch to a symbol of different type. For example, at the top of the drawing, a white balls becomes two white balls; then one of the white ball switches to a black ball, and another white ball continue to duplicate to two white balls. The process is repeated until a long sequence is generated.

## Power Spectra of A Few Genome Sequences

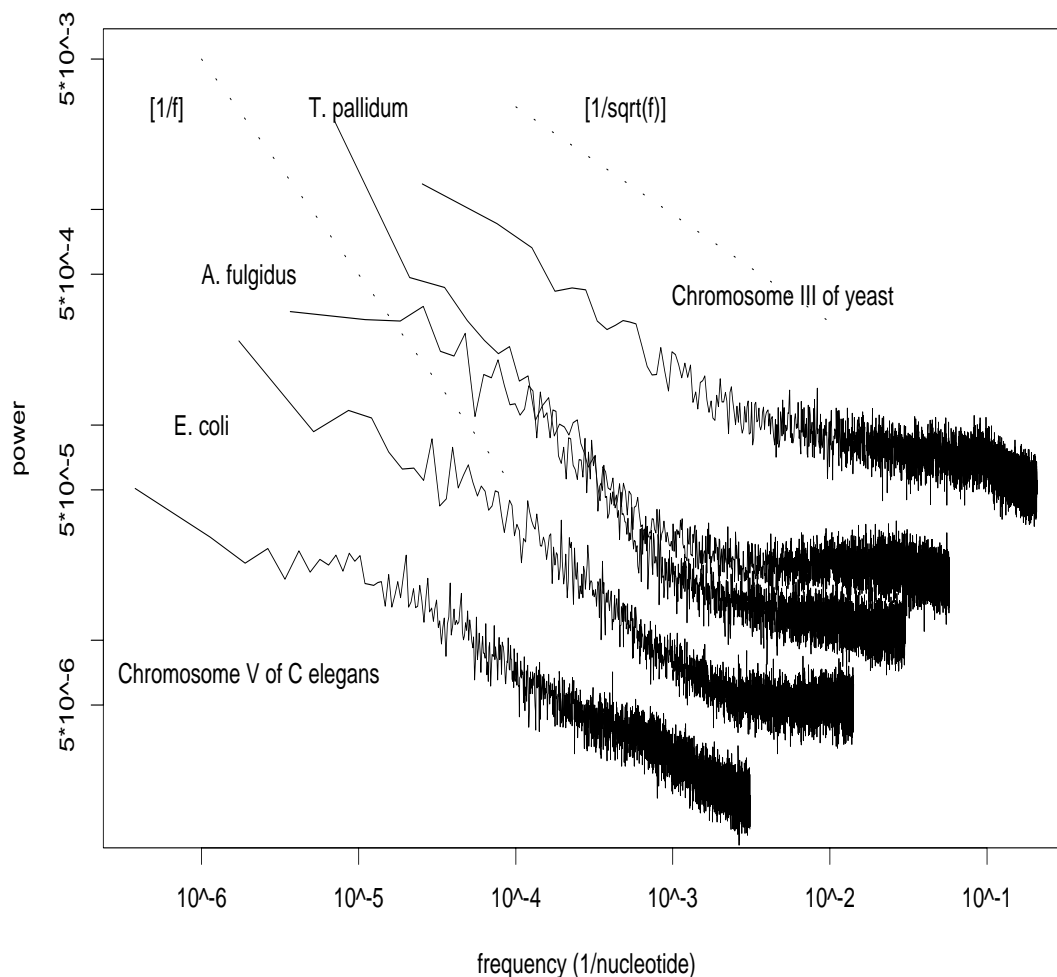


Figure 2: **1/f spectra in genome sequences:** Power spectra of five complete genome sequences: the chromosome 3 of yeast *Saccharomyces cerevisiae*, the first complete chromosome sequence being determined (in 1992); *Treponema pallidum*, the bacterium that causes syphilis, a sexually transmitted disease; *Archaeoglobus fulgidus*, a sulphate-reducing archaeon; bacterium *Escherichia coli*, one of the favorite model systems used by generations of biologists; the longest chromosome, chromosome 5, of *Caenorhabditis elegans*, a round worm which is another favorite system for biologists. Both the x axis (frequency  $f$ , in the unit of 1/nucleotide) and y axis (power) are drawn in a logarithmic scale. A spectrum of the form of  $1/f$  or any other power-law function will be a straight line in the plot. Two such functions are shown as a reference:  $1/f$  and  $1/\sqrt{f}$  (or  $1/f^{0.5}$ ). Since the variable frequency is the inverse of the length variable, this plot spans the length scale from around 10 nucleotides to 1,000,000 nucleotides.