Chapter 4

Non Gaussian statistics and the DNA sequences of prokaryotes

4.1 General background

4.1.1 Why study correlations in DNA sequences

The material substrate of the genetic heritage of every organism is the well-known DeoxyriboNucleic Acid, or DNA, a macromolecule composed by stacked pairs of four different nucleotides - Adenine (A), Guanine (G), Cytosine (C) and Thymine (T) - having specific physical and chemical properties (Figure 4.1). Protein coding occurs by transcripting the DNA code contained in segments called genes into RNA, a molecule similar to DNA, and then translating it into a chain of aminoacids. However, only a part of the gene DNA is used to build the aminoacid chain: while the code contained in exons is eventually translated into proteins, introns are only transcripted and then cut out of the process (Figure 4.2). In addition, DNA is composed by a non-coding part whose percentage is very variable across species. A gross main separation can be traced between species: the DNA of prokaryotes (virus and bacteria, not endowed with a cellular nucleus) is mainly coding, having almost no introns and very few non-coding segments; the DNA of eukaryotes (animals, plants and humans, endowed with a cellular nucleus) has a percentage of introns and non-coding part that increases with evolution, yeast non-coding part representing only about 10% of the total, while humans having only 1.5% of their DNA which is coding.

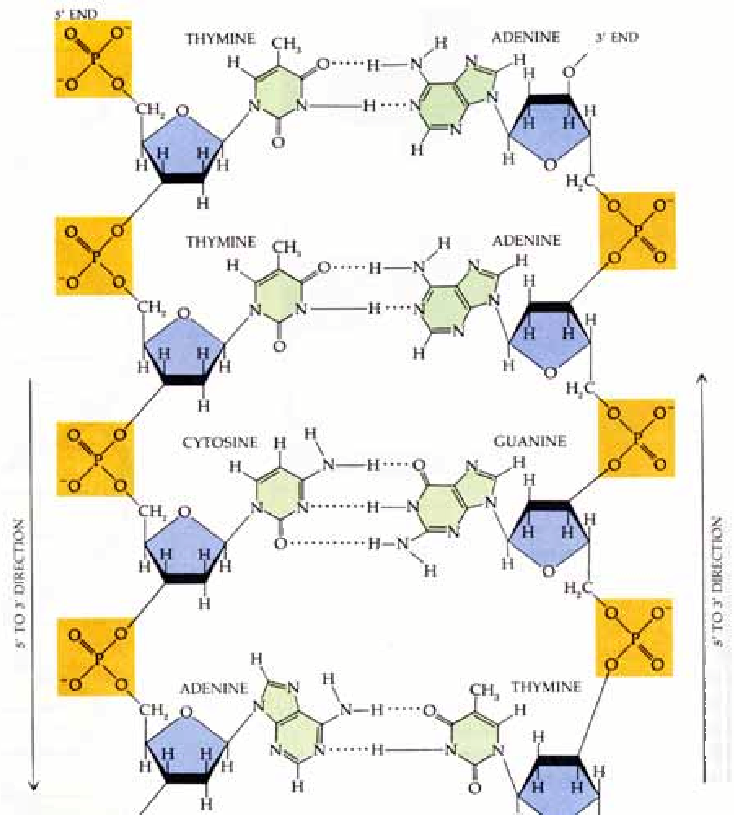

Figure 4.1: Chemical composition of the four nucleotides and their binding along the DNA helix.

DNA function, in the classical conception of molecular genetics, is essentially reduced to the transmission through generations of heritable information concerning protein synthesis. This is the basic meaning of the "central dogma of molecular genetics" proposed in 1958 by Francis Crick. The "dogma" states that each polypeptide chain is synthesized by translating into a 20 letters alphabet (aminoacids) a message written in a 4 letters one (nucleotides), on the basis of the universal 1:3 code (each aminoacid being coded by a triplet of nucleotides called codon). Information transmission is unequivocal and unidirectional both in the transcription into RNA and in the translation steps. According to this linguistic and informational view, DNA is merely a string of symbols (the 4 letters A, C, G and T) whose chemical and physical features seem to be irrelevant for the correct

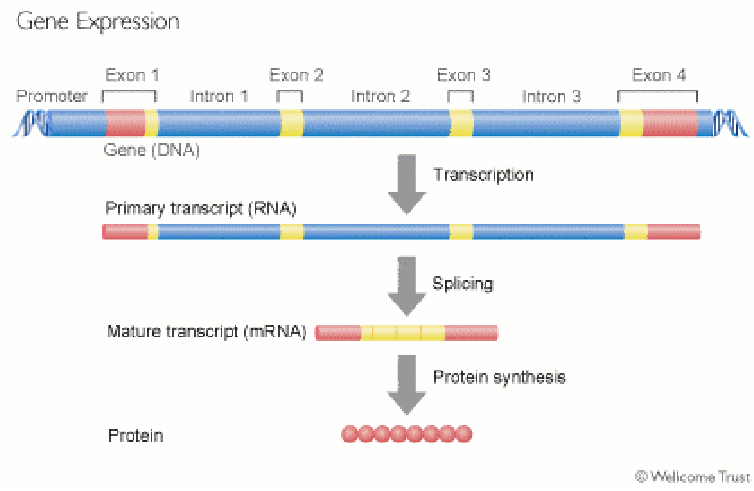

Figure 4.2: Gene expression process. While the code contained in exons is eventually translated into proteins, introns are only transcripted and then spliced out of the process.

functioning of the DNA machinery.

Since 1958, the extraordinary progress of molecular biology has improved our insight into such machinery, suggesting a much more complex organization of DNA function. In particular, non-coding sequences, covering more than 90% of the entire genome of many eukaryotes, have for a long time been considered as "junk", devoid of any biological meaning. The discovery that these sequences contain a large array of regulatory functions has changed this view and led to a new attention to their physical-chemical features (Wray et al., 2003). The emerging view is that transcriptional regulation plays a more important role than it was thought, and that, as depicted in Figure 4.3, loose but detectable physical-chemical features of non-coding tracts composition can influence regulation of genes located relatively far apart along the sequence (Wray et al., 2003).

DNA structure is also more complex than previously thought. The structural complexity is already present in prokaryotes. For example, the circular chromosome of the bacterium Escherichia Coli is 1 mm long, and is packaged inside a cylindric envelope whose dimensions are of the order of 1 μm. The mass containing all the prokaryotic genetic material, called nucleoid, also contains several proteins (in prokaryotes about 20% of the total) that

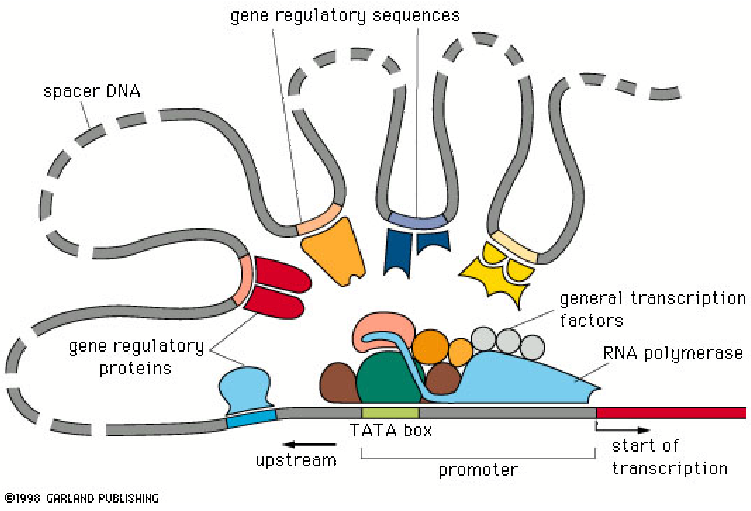

Figure 4.3: Regulatory mechanism of DNA transcription. Transcription can be modulated by regulatory proteins that bind along the sequence far apart from the gene.

take part in the nucleoid scaffolding. The high packaging ratio and the need of local rapid and precise unwrapping for transcription and replication suggests an organized, hierarchical structure of DNA coiling. Prokaryotic chromosomes are known to form loops consisting of few thousands base pairs (Figure 4.4), but knowledge about their structure at larger scales is very poor. Packaging ratio is much more dramatic in eukaryotes: the 2 m long human genome is supercoiled inside a 6 μm wide nucleus! Its nucleotide chain is tightly wound up and packaged around spool-like proteins called histones. Figure 4.5 sketches the main levels of eukaryotic DNA packaging at different scales, from the lowest one (DNA loops around octamers of histons called nucleosomes), to complexes that are like tight spool clusters, up to chromosomes.

Regulatory and structural features are deeply intertwined: the expression of a gene is modulated by the interaction with a specific protein, and such interaction occurs only if the associated non-coding regulatory DNA tracts are properly bended and physically accessible to that protein. Since the curvature and chemical properties of DNA tracts ultimately depend on their nucleotide composition, it is thought that both regulatory and structural organizations set some constraints on DNA nucleotide composition.

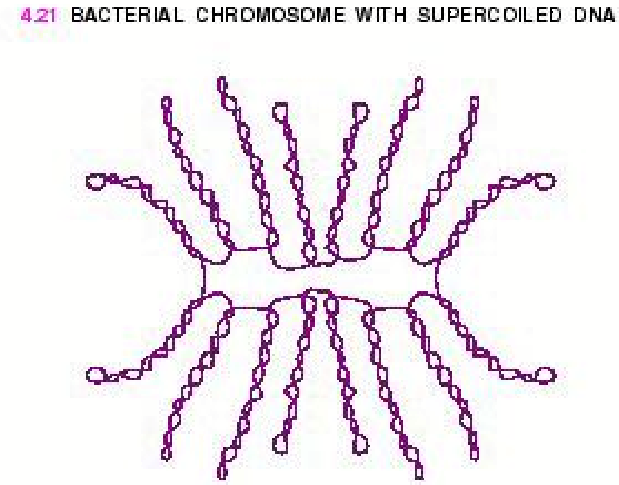

Figure 4.4: Supercoling of the bacterial circular chromosome.

As we will show for well-known structural features, these constraints are detectable by studying the correlations in the distribution of nucleotides along the sequence seen as a symbolic sequence. A way of investigating the large-scale structure of DNA sequences is thus to study the statistical correlations of its associated symbolic sequence. Even though our work does not directly link the correlation analysis with the structural and functional properties of the DNA sequences analysed, it is motivated by the above rationale, as well as most of the other works presented in this introductory section. Here we present an overview of the results emerging from the statistical analysis of correlations in DNA sequences, focusing at the end of the section to the analysis of DNA sequences of prokaryotes as an introduction to our study.

4.1.2 Correlations, structure and function in DNA sequences

Short-range correlations reflect low-level DNA structure

Herzel et al. (1998) showed that the relationship between structure and sequence composition can be striking at short-range distances. Herzel et al. (1998) computed the 16 correlation functions corresponding to all the 16 possible pairs of nucleotides:

$$C_{ab}(k) = \langle a(l)b(l-k)\rangle - \langle a(l)\rangle\langle b(l-k)\rangle, \tag{4.1}$$

where $a(l)$ and $b(l)$ correspond to dichotomic time series in which $a(l) = 1$ if the nucleotide in position l along the sequence is the one associated to a, and $a(l) = 0$ if it is a different nucleotide, and analogously for b. The average is computed over all positions along the sequence. Figure 4.6 shows three out of the 16 correlation functions computed on yeast chromosome IV. All three thin lines exhibit a clear period-3 oscillation. This is a simple consequence of the non-uniform nucleotide probabilities, which in turn is caused by the nonuniform codon usage[1]. In particular, there is typically an excess of guanine (G) in position 1 of the codon which leads to strong period-3 oscillations of the GG autocorrelation function (bottom graph of Figure 4.6). The thick lines (running average over 3 base pairs (bp)) in the upper and middle graph show another oscillating pattern with a period 10-11 bp. Herzel et al. (1998) argue that its origin is probably two-fold: 1) DNA folding around nucleosomes induces a nonuniform bias towards the weak bound A and T nucleotides that allow a better bending around the nucleosome, and in turn this bias induces a period 10-11 pattern corresponding to the DNA helic period; 2) 10-11 periodicity could also be due, in coding regions to well-known correlations in the aminoacid sequences.

Trifonov (1998) extended the above analysis to a larger scale and found 200 bp and 400 bp periodicities, which he associated with DNA folding around nucleosomes and couples of nucleosomes (the segments folding around nuclosomes typically contain about 200 bp). These relatively simple patterns are easily detectable, and suggest that correlation analysis offers an important support to the study of DNA structure.

Long-range correlations suggest large-scale organization

Many different factors stimulated researchers to investigate the presence of long-range correlations in DNA sequences: the DNA large-scale, hierarchical structure, the heterogeneity of the nucleotide distribution (Li, 1997), models of DNA evolution (Li & Kaneko, 1992), or simply the curiosity to see whether DNA has scale-invariant statistical properties similar to other self-similar phenomena in nature (Voss, 1992; Peng et al., 1992). The exponential increase of available sequenced genomes in the nineties paved the way to these studies. The first investigations gave positive though partially controversial results. Voss (1992) and Li & Kaneko (1992) computed the low

[1]Since each codon codes for one out of 20 aminoacids, and there are 64 different codon compositions, there is a redundancy in the codon code that can be biased towards a particular distribution. This bias is called codon usage.

frequency scaling behaviour of the power spectrum (Eq. 2.8) on sequences obtained by a mapping equivalent to that of Herzel et al. (1998). They found $\gamma \simeq 1$ and $0.5 < \gamma < 0.85$ respectively: though the ranges of the scaling exponent differed, probably due to a different way of averaging over the sequences, both groups agreed in claiming the presence of long-range correlations in all DNA sequences analyzed.

Peng et al. (1992, 1994) measured the long-range correlations of the dichotomic sequences $x(l)$ obtained by the following binary mapping: $x(l) = 1$ if nucleotide in position l of the DNA sequence is a purine (A or G) and $x(l) = -1$ if nucleotide in l is a pyrimidine (C or T). They estimated the long-range correlations by means of the DFA (Eq. 2.16) on the walk associated with this mapping, which they called DNA walk (see 2.2.1 for the mathematical definition of a walk). Figure 4.7 shows the results obtained by using the DFA on one coding sequence (a segment of the bacterium Escherichia Coli) and one non-coding sequence (a segment of human DNA). They interpreted the results shown in the figure by claiming that while non-coding sequences exhibit robust long-range correlations, coding ones do not show any long-range scaling behaviour. They argued that the increase in slope of the DFA analysis relative to the E. Coli sequence (crossover denoted by an arrow in Figure 4.7) does not indicate genuine long-range correlations, but rather spurious correlations generated by compositional biases having a long but finite characteristic scale. To support their argument, they also show in the figure that the behaviour of the E. Coli sequence is well reproduced by an uncorrelated biased random walk obtained by stitching together subsequences (average length 2500) that correspond to random walks with different nucleotide composition.

Peng's results were further supported by Arneodo et al. (1995), who detected long-range correlations in introns and non-coding regions (but not in coding regions) by means of a wavelet based analysis. Arneodo et al. (1995) also claimed that fluctuations in DNA statistics are Gaussian, and can be modeled as a fractional Brownian motion.

Two main points emerge from these first results. On one hand, the existence of long-range correlations suggests that DNA complex organization is reflected in constraints acting on the nucleotide composition at all scales. On the other hand, investigations had to face an heterogeneity of results that depend on the variability of DNA statistics across species and at different scales. As we will see later, the controversy about the presence of

long-range correlations in coding regions is a typical example in which the scaling behaviour is not uniform at all scales. Since this is the object of our study, we address this issue in detail in the next section. We postpone to the Discussion chapter of the thesis an overview of the results obtained recently on the complete genome of eukaryotes.

4.1.3 Modeling the statistics of prokaryotic DNA

Peng and coauthors' claim of no correlations in coding regions was contraddicted by other studies showing that long-range correlations indeed arise also in prokaryotes (containing almost all coding regions) and in coding regions of eukaryotes (Prabhu & Claverie, 1992; Allegrini et al., 1995; Herzel & Groe, 1997). The controversy probably arised from the fact that long-range correlations are absent at short scale, and arise only at large scale. Inspired by this observation, Allegrini et al. (1995) attempted to model the statistics underlying prokaryotic DNA sequences as the superposition of two processes, one deterministic with long-range correlations, and the other random with no correlations. The long-range correlated process consists of a dichotomic time series $x_c(l)$ taking the values 1 or -1, generated by a non-linear deterministic map, and characterised by an autocorrelation function (as defined by Eq. 2.3) decaying as in Eq. 2.5, which we report here for clarity:

$$C_{x_c}(t) \propto t^{-\beta} \text{ for } t \to \infty \tag{4.2}$$

with $0 < \beta < 1$. The superposition of this process with a random component resulted in the time series

$$x(t) = \begin{cases} x_c(t) & \text{with probability} \quad \epsilon \\ \text{random } \{\pm 1\} & \text{with probability} \quad 1-\epsilon \end{cases} \tag{4.3}$$

where the parameter $0 \leq \epsilon \leq 1$ controls the level of the correlated component. The map defined by Eq. (4.3) was called the Copying Mistake Map (CMM). Allegrini et al. (1995) showed that the diffusion of the integrated signal (the simulated DNA walk) $y(t)$ obtained from $x(t)$ through Eq. 2.10 follows the asymptotic behaviour

$$\langle y^2(t) \rangle = At^{2H} + Bt \text{ for } t \to \infty \tag{4.4}$$

where the constants A and B depend on ϵ and $H = 1 - \beta/2$. Allegrini and colleagues then fitted the parameters ϵ in Eq. (4.3) and β in Eq. (4.2) of the CMM in order to better reproduce the second order statistics of a sequence

of Human Cytomegalovirus mapped following the same binary association purines/pyrimidines that gives rise to the DNA walk of (Peng et al., 1992, 1994). Similarity between the DNA sequence and the associated CMM was investigated with three different measures: diffusion analysis (Eq. 2.11), DFA (Eq. 2.16) and rescaled Hurst analysis (for details about this last measure please refer to (Allegrini et al., 1995)). Figure 4.8 shows a good agreement between the real DNA sequence and the fitted CMM with all three measures: while at short scales they both exhibit a normal diffusion (slope $H \sim 0.5$) denoting the lack of correlations, at longer scales the curves' slope clearly increases towards $H = 0.75$, revealing the presence of long-range correlations.

The CMM model offers a richer interpretation for the statistics of prokaryotes DNA: while the coding part has intrinsecally no long-range correlations, the slight redundancy in the protein code allows a superposition of long-range correlations that are probably the effect of a large-scale structural organization.

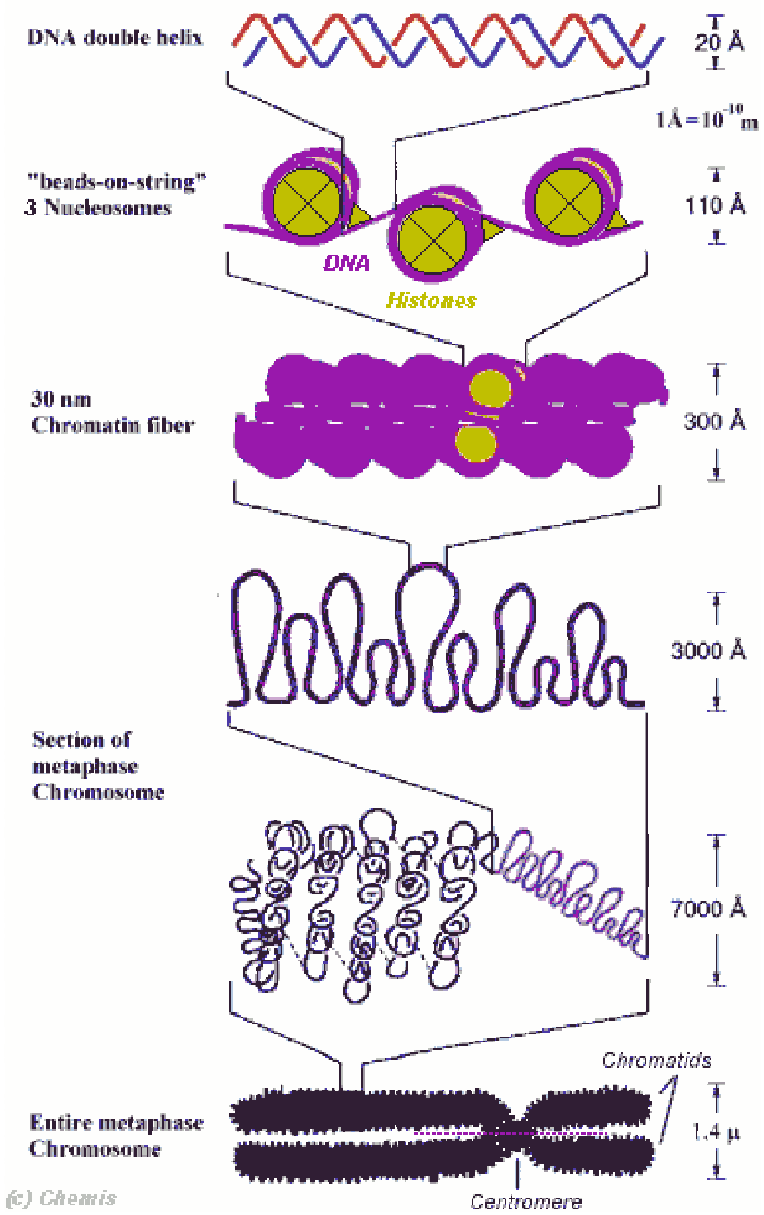

Figure 4.5: Hierarchical structure of eukaryotic DNA packaging at different scales.

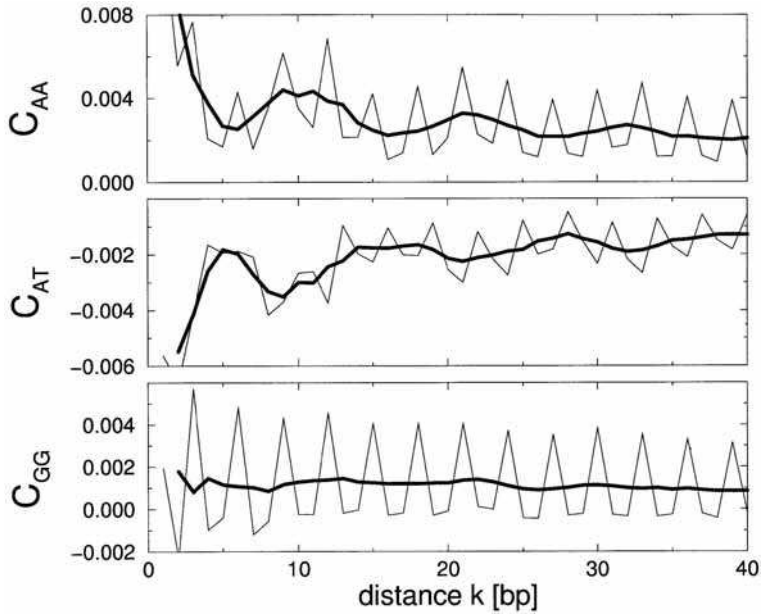

Figure 4.6: Correlation functions in yeast chromosome IV (1522191 bp). Upper graph: $C_{AA}(k)$ (thin line) and running average over 3 bp (thick line). Middle graph: $C_{AT}(k)$. Lower graph: $C_{GG}(k)$. From Herzel et al. (1998).

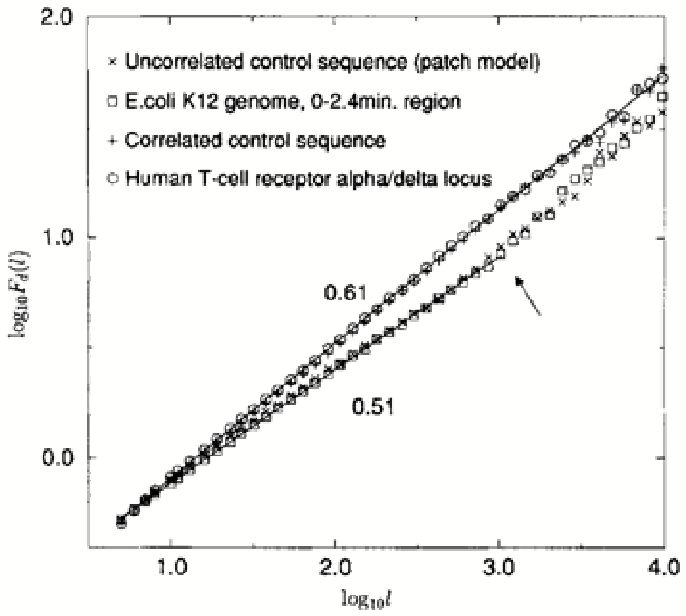

Figure 4.7: DFA analysis of the following sequences: an uncorrelated biased random walk obtained by stitching together subsequences that correspond to random walks with different nucleotide composition (X); an Escherichia Coli coding fragment (little squares); a long-range correlated control sequence (+); the non-coding human T-cell receptor alpha/delta locus (∘). The lower solid line, the best fit for E. Coli data from l=4 to 861, has slope 0.51. The upper solid line, the best fit for human data from l=4 to 8192, has slope 0.61. The arrow denoting the crossover phenomenon is explained in the text. From (Peng et al., 1994).

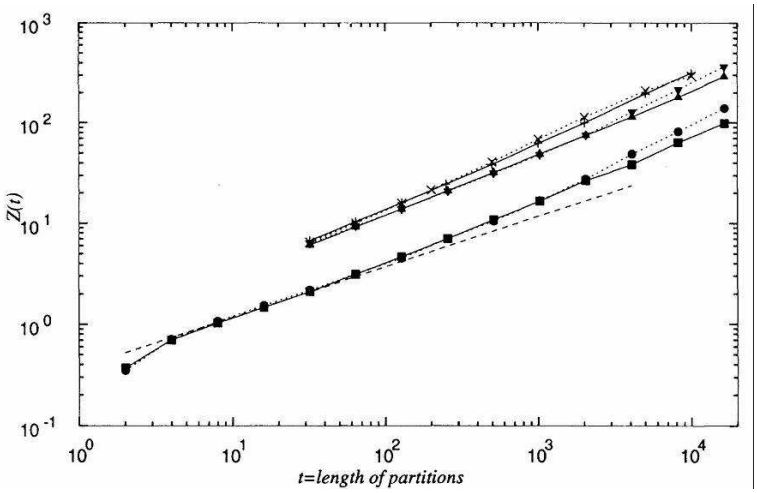

Figure 4.8: Three analyses (from top to bottom: diffusion analysis, rescaled Hurst analysis and DFA) applied to the Cytomegalovirus strain AD169 sequence (solid curves) and to the CMM (dotted curves) with $\epsilon = 1/9$ and $\beta = 0.67$. The function $Z(t)$ is defined as $\sqrt{\langle y^2(t)\rangle}$, $R(t)$ (see Allegrini et al. (1995) for a definition) and $F(t)$ for the three analyses, respectively. For comparison, a dashed line with the slope of a random walk (0.5) has been drawn. From Allegrini et al. (1995).