Statistical Properties of DNA Sequences

S. Havlin, S. V. Buldyrev, A. L. Goldberger, R. N. Mantegna,
C.-K. Peng, M. Simons, and H. E. Stanley

April 29, 1999

Contents

Chapter 1

Statistical Properties of DNA Sequences

We present evidence supporting the idea that the DNA sequence in genes containing *noncoding* regions is correlated, and that the correlation is remarkably long range—indeed, base pairs *thousands of base pairs* distant are correlated. We do not find such a long-range correlation in the coding regions of the gene. We resolve the problem of the "non-stationarity" feature of the sequence of base pairs by applying a new algorithm called *Detrended Fluctuation Analysis (DFA)*. We address the claim of Voss that there is no difference in the statistical properties of coding and noncoding regions of DNA by systematically applying the DFA algorithm, as well as standard FFT analysis, to every DNA sequence (33 301 coding and 29 453 noncoding) in the entire GenBank database. We describe a simple model to account for the presence of long-range power-law correlations which is based upon a generalization of the classic Lévy walk. Finally, we describe briefly some recent work showing that the *noncoding* sequences have certain statistical features in common with natural languages. Specifically, we adapt to DNA the Zipf approach to analyzing linguistic texts, and the Shannon approach to quantifying the "redundancy" of a linguistic text in terms of a measurable entropy function. We demonstrate that noncoding regions in eukaryotes display a smaller entropy and larger redundancy than coding regions, further supporting the possibility that noncoding regions of DNA may carry biological information.

1.1 Long-range power-law correlations

In recent years long-range power-law correlations have been discovered in a remarkably wide variety of systems. Such long-range power-law correlations are a physical fact that in turn gives rise to the increasingly appreciated "fractal

geometry of nature" (Bunde and Havlin 1991, Bunde and Havlin 1994). So if fractals are indeed so widespread, it makes sense to anticipate that long-range power-law correlations may be similarly widespread. Indeed, recognizing the ubiquity of long-range power-law correlations can help us in our efforts to understand nature, since as soon as we find power-law correlations we can quantify them with a critical exponent. Quantification of this kind of scaling behavior for apparently unrelated systems allows us to recognize similarities between different systems, leading to underlying unifications that might otherwise have gone unnoticed.

Traditionally, investigators in many fields characterize processes by assuming that correlations decay exponentially. However, there is one major exception: at the critical point, the exponential decay turns into a power law decay (Stanley 1971)

$$C_r \sim (1/r)^{d-2+\eta}. \tag{1.1}$$

Many systems drive themselves spontaneously toward critical points (Stanley 1971, Bunde and Havlin 1991, Barabási and Stanley 1995). One of the simplest models exhibiting such "self-organized criticality" is invasion percolation, a generic model that has recently found applicability to describing anomalous behavior of rough interfaces.

In the following sections we will attempt to summarize some recent findings (see Peng et al. 1992, Li and Kaneko 1992, Voss 1994, and references cited in Buldyrev et al. 1994) concerning the possibility that—under suitable conditions—the sequence of base pairs or "nucleotides" in DNA also displays power-law correlations. The underlying basis of such power law correlations is not understood at present, but this discovery has intriguing implications for molecular evolution (Buldyrev et al. 1993), as well as potential practical applications for distinguishing coding and noncoding regions in long nucleotide chains (Ossadnik et al. 1994). It also may be related to the presence of a language in noncoding DNA (Mantegna et al. 1994).

1.2 DNA

The role of genomic DNA sequences in coding for protein structure is well known (Watson et al. 1992). The human genome contains information for approximately 100,000 different proteins, which define all inheritable features of an individual. The genomic sequence is likely the most sophisticated information database created by nature through the dynamic process of evolution. Equally remarkable is the precise transformation of information (duplication, decoding, etc) that occurs in a relatively short time interval.

The building blocks for coding this information are called *nucleotides*. Each nucleotide contains a phosphate group, a deoxyribose sugar moiety and either a *purine* or a *pyrimidine base*. Two purines and two pyrimidines are found in

DNA. The two purines are adenine (A) and guanine (G); the two pyrimidines are cytosine (C) and thymine (T).

In the genomes of high eukaryotic organisms only a small portion of the total genome length is used for protein coding (as low as 3% in the human genome). The segments of the chromosomal DNA that are spliced out during the formation of a mature mRNA are called *introns* (for intervening sequences). The coding sequences are called *exons* (for expressive sequences).

The role of introns and intergenomic sequences constituting large portions of the genome remains unknown. Furthermore, only a few quantitative methods are currently available for analyzing information which is possibly encrypted in the noncoding part of the genome.

1.3 The "DNA walk"

One interesting question that may be asked by statistical physicists would be whether the sequence of the nucleotides A,C,G, and T behaves like a one-dimensional "ideal gas", where the fluctuations of density of certain particles obey Gaussian law, or if there exist long range correlations in nucleotide content (as in the vicinity of a critical point). These result in domains of all sizes with different nucleotide concentrations. Such domains of various sizes were known for a long time but their origin and statistical properties remain unexplained. A natural language to describe heterogeneous DNA structure is long-range correlation analysis, borrowed from the theory of critical phenomena (Stanley 1971).

In order to study the scale-invariant long-range correlations of a DNA sequence, we first introduced a graphical representation of DNA sequences, which we term a *fractal landscape* or *DNA walk* (Peng et al. 1992). For the conventional one-dimensional random walk model, a walker moves either "up" $[u(i) = +1]$ or "down" $[u(i) = -1]$ one unit length for each step i of the walk. For the case of an uncorrelated walk, the direction of each step is independent of the previous steps. For the case of a correlated random walk, the direction of each step depends on the history ("memory") of the walker.

One definition of the DNA walk is that the walker steps "up" if a pyrimidine (C or T) occurs at position i along the DNA chain, while the walker steps "down" if a purine (A or G) occurs at position i. The question we asked was whether such a walk displays only short-range correlations (as in an n-step Markov chain) or long-range correlations (as in critical phenomena and other scale-free "fractal" phenomena). A different type of DNA walk was introduced earlier by Azbel (1973).

There have also been attempts to map DNA sequence onto multi-dimensional DNA walks (Li and Kaneko 1992, Berthelsen et al. 1992). However, recent work (Ossadnik et al. 1994) indicates that the original purine-pyrimidine rule provides the most robust results, probably due to the purine-pyrimidine chemical complementarity.

Figure 1.1: DNA walk displacement $y(\ell)$ (excess of purines over pyrimidines) vs nucleotide distance ℓ for (a) HUMHBB (human beta globin chromosomal region of the total length $L = 73,239$); (b) the LINE1c region of HUMHBB starting from 23,137 to 29,515; (c) the generalized Lévy walk model of length 73,326 with $\mu = 2.45$, $l_c = 10$, $\alpha_o = 0.6$, and $\epsilon = 0.2$; and (d) a segment of a Lévy walk of exactly the same length as the LINE1c sequence from step 67,048 to the end of the sequence. This sub-segment is a Markovian random walk. Note that in all cases the overall bias was subtracted from the graph such that the beginning and ending points have the same vertical displacement ($y = 0$).

1.4 Correlations and fluctuations

An important statistical quantity characterizing any walk is the root mean square fluctuation $F(\ell)$ about the average of the displacement of a quantity $\Delta y(\ell)$ defined by $\Delta y(\ell) \equiv y(\ell_0 + \ell) - y(\ell_0)$, where

$$y(\ell) \equiv \sum_{i=1}^{\ell} u(i). \tag{1.2}$$

If there is no characteristic length (i.e., if the correlations were "infinite-range"), then fluctuations will also be described by a power law

$$F(\ell) \sim \ell^{\alpha} \tag{1.3}$$

with $\alpha \neq 1/2$.

Figure 1a shows a typical example of a gene that contains a significant fraction of base pairs that do *not* code for amino acids. It is immediately apparent that the DNA walk has an extremely jagged contour which corresponds to long-range correlations.

The fact that data for intron-containing and intergenic (i.e., noncoding) sequences are linear on this double logarithmic plot confirms that $F(\ell) \sim \ell^{\alpha}$. A least-squares fit produces a straight line with slope α substantially larger than the prediction for an uncorrelated walk, $\alpha = 1/2$, thus providing direct experimental evidence for the presence of long-range correlations.

On the other hand, the dependence of $F(\ell)$ for coding sequences is not linear on the log-log plot: its slope undergoes a crossover from 0.5 for small ℓ to 1 for large ℓ. However, if a single patch is analyzed separately, the log-log plot of $F(\ell)$ is again a straight line with the slope close to 0.5. This suggests that within a large patch the coding sequence is almost uncorrelated.

1.5 Detrended fluctuation analysis (DFA)

The initial report (Peng et al. 1992) on long-range (scale-invariant) correlations only in noncoding DNA sequences has generated contradicting responses. For

Figure 1.2: Analysis of section of Yeast Chromosome III using the sliding box *Coding Sequence Finder* "CSF" algorithm. The value of the long-range correlation exponent α is shown as a function of position along the DNA chain. In this figure, the results for about 10% of the DNA are shown (from base pair #30,000 to base pair #60,000). Shown as vertical bars are the putative genes and open reading frames; denoted by the letter "G" are those genes that have been more firmly identified (March 1993 version of *GenBank*). Note that the local value of α displays **minima** where genes are suspected, while between the genes α displays **maxima**. This behavior corresponds to the fact that the DNA sequence of genes lacks long-range correlations ($\alpha = 0.5$ in the idealized limit), while the DNA sequence in between genes possesses long-range correlations ($\alpha \approx 0.6$).

details see the work of Buldyrev et al. (1994). The source of these contradicting claims may arise from the fact that, in addition to normal statistical fluctuations expected for analysis of rather short sequences, coding regions typically consist of only a few lengthy regions of alternating strand bias—and so we have non-stationarity. Hence conventional scaling analyses cannot be applied reliably to the entire sequence but only to sub-sequences.

Peng et al. (1994) have recently applied the "bridge method" to DNA, and have also developed a similar method specifically adapted to handle problems associated with non-stationary sequences which they term *detrended fluctuation analysis* (DFA).

The idea of the DFA method is to compute the dependence of the standard error of a linear interpolation of a DNA walk $F_d(\ell)$ on the size of the interpolation segment ℓ. The method takes into account differences in local nucleotide content and may be applied to the entire sequence which has lengthy patches. In contrast with the original $F(\ell)$ function, which has spurious crossovers even for ℓ much smaller than a typical patch size, the detrended function $F_d(\ell)$ shows linear behavior on the log-log plot for all length scales up to the characteristic patch size, which is of the order of a thousand nucleotides in the coding sequences. For ℓ close to the characteristic patch size the log-log plot of $F_d(\ell)$ has an abrupt change in its slope.

The DFA method clearly supports the difference between coding and non-coding sequences, showing that the coding sequences are less correlated than noncoding sequences for the length scales less than 1000, which is close to characteristic patch size in the coding regions. One source of this difference is the tandem repeats (sequences such as AAAAAA...), which are quite frequent in noncoding sequences and absent in the coding sequences.

1.6 Coding sequence finder (CSF) algorithm

To provide an "unbiased" test of the thesis that noncoding regions possess but coding regions lack long-range correlations, Ossadnik et al. (1994) analyzed several artificial uncorrelated and correlated "control sequences" of size 10^5 nucleotides using the GRAIL neural net algorithm (Uberbacher and Mural 1991). The GRAIL algorithm identified about 60 putative exons in the uncorrelated sequences, but only about 5 putative exons in the correlated sequences.

Using the DFA method, we can measure the local value of the correlation exponent α along the sequence (see Figure 1.2) and find that the local minima of α as a function of a nucleotide position usually correspond to noncoding regions, while the local maxima correspond to noncoding regions. Statistical analysis using the DFA technique of the nucleotide sequence data for yeast chromosome III (315,338 nucleotides) shows that the probability that the observed correspondence between the positions of minima and coding regions is due to random coincidence is less than 0.0014. Thus, this method—which we called the "coding sequence finder" (CSF) algorithm—can be used for finding coding regions in the newly sequenced DNA, a potentially important application of DNA walk analysis.

1.7 Systematic analysis of GenBank database

An open question in computational molecular biology is whether long-range correlations are present in both coding and noncoding DNA or only in the latter. To answer this question, Buldyrev et al. (1995) recently analyzed all 33 301 coding and all 29 453 noncoding eukaryotic sequences—each of length larger than 512 base pairs (bp)—in the present release of the GenBank to determine whether there is any statistically significant distinction in their long-range correlation properties.

Buldyrev et al. find that standard fast Fourier transform (FFT) analysis indicates that *coding* sequences have practically no correlations in the range from 10 bp to 100 bp (spectral exponent $\beta \pm 2SD = 0.00 \pm 0.04$). Here β is defined through the relation $S(f) \sim 1/f^{\beta}$, where $S(f)$ is the Fourier transform of the correlation function, and β is related to the long-range correlation exponent α by $\beta = 2\alpha - 1$ so that $\alpha = 1/2$ corresponds to $\beta = 0$ (white noise).

In contrast, for *noncoding* sequences, the average value of the spectral exponent β is positive (0.16 ± 0.05), which unambiguously shows the presence of long-range correlations. They also separately analyzed the 874 coding and 1157 noncoding sequences which have more than 4096 bp, and found a larger region of power law behavior. Buldyrev et al. calculated the probability that these two data sets (coding and noncoding) were drawn from the same distribution, and found that it is less than 10^{-10}. Buldyrev et al. also obtained independent confirmation of these findings using the DFA method, which is designed to treat

sequences with statistical heterogeneity such as DNA's known mosaic structure ("patchiness") arising from non-stationarity of nucleotide concentration. The near-perfect agreement between the two independent analysis methods, FFT and DFA, increases the confidence in the reliability of the conclusion that long-range correlation properties of coding and noncoding sequences.

¿From a practical viewpoint, the statistically significant difference in long-range power law correlations between coding and noncoding DNA regions that we observe supports the development of gene finding algorithms based on these distinct scaling properties. A recently reported algorithm of this kind (Ossadnik et al. 1994) is especially useful in the analysis of DNA sequences with relatively long coding regions, such as those in yeast chromosome III.

Very recently Arneodo et al. (1995) studied long-range correlation in DLA sequences using wavelet analysis. The wavelet transform can be made blind to "patchiness" of genomic qenomic sequences. They found the existence of long-range correlations in non-coding regimes, and no long-range correlations in coding regimes in excellent agreement with Buldyrev et al. (1995).

Finally, we note that although the scaling exponents α and β have potential use in quantifying changes in genome complexity with evolution, the current GenBank database does not allow us to address the important question of whether unique values of these exponents can be assigned to different species or to related groups of organisms. At present, the GenBank data have been collected such that particular organisms tend to be represented more frequently than others. For example, about 80% of the sequences from birds are from *Gallus gallus* (the chicken) and about 2/3 of the insect sequences are from *Drosophila melanogaster*. The results indicate the importance of sequencing not only coding but also noncoding DNA from a wider variety of species.

1.8 Linguistic analysis of noncoding and coding DNA

Long-range correlations have been found recently in human writings (Schenkel et al. 1993). A novel, a piece of music or a computer program can be regarded as a one-dimensional string of symbols. These strings can be mapped to a one-dimensional random walk model similar to the DNA walk allowing calculation of the correlation exponent α. Values of α between 0.6 and 0.9 were found for various texts.

An interesting hierarchical feature of languages was found by Zipf (1949). He observed that the frequency of words as a function of the word order ("rank") decays as a power law (with a power ζ close to -1) for more than four orders of magnitude.

In order to adapt the Zipf analysis to DNA, the concept of word must first be defined. In the case of coding regions, the words are the 64 3-tuples ("triplets")

Figure 1.3: Linguistic features of noncoding DNA. Log-log plot of a histogram of word frequency for the noncoding part of Yeast Chromosome III ($\approx$ 315,000 bp). The 6-character words are placed in rank order, where rank 1 corresponds to the most frequently used word, rank 2 to the second most frequently used word, and so forth. The straight line behavior provides evidence for a structured language in noncoding DNA. Top bar: Rainbow color code corresponds to the rank of words in the language of this sequence, which is used as a "reference language." Bottom bar: The colors are re-arranged, corresponding to the re-arrangements of their rank with respect to the reference language for the coding part.

which code for the amino acids, AAA, AAT, ... GGG. However for noncoding regions, the words are not known. Therefore Mantegna et al. (1994,1995) consider the word length n as a free parameter, and performs analyses not only for $n = 3$ but also for all values of n in the range 3 through 8. The different n-tuples are obtained for the DNA sequence by shifting progressively by 1 base a window of length n; hence, for a DNA sequence containing L base pairs, we obtain $L - n + 1$ different words.

The results of the Zipf analysis for all 40 DNA sequences analyzed are summarized by Mantegna et al. (1994). The averages for each category support the observation that ζ is consistently larger for the noncoding sequences, suggesting that the noncoding sequences bear more resemblance to a natural language than the coding sequences. Moreover, the "words" used in coding and noncoding sequences appear in quite different orders (Figure 1.3).

Related interesting statistical measures of short-range correlations in languages are the entropy and redundancy. The redundancy is a manifestation of the *flexibility* of the underlying code. To quantitatively characterize the redundancy implicit in the DNA sequence, we utilize the approach of Shannon (1951), who provided a mathematically precise definition of redundancy (Shannon 1951, Brillouin 1956). Shannon's redundancy is defined in terms of the entropy of a text—or, more precisely, the "n-entropy"

$$H(n) = -\sum_{i=1}^{4^n} p_i \log_2 p_i, \tag{1.4}$$

which is the entropy when the text is viewed as a collection of n-tuple words. Here p_i is the normalized frequency of occurrence of n-tuple i. The redundancy is defined through as $R \equiv \lim_{n\to\infty} R(n)$, where

$$R(n) \equiv 1 - H(n)/kn; \tag{1.5}$$

here $k = \log_2 4 = 2$.

Mantegna et al. (1994) also calculate the Shannon n-entropy $H(n)$ for $n = 1, 2, \cdots 6$. The maximum value of n for which it is possible to determine $H(n)$ is

$n = 6$—even for very long sequences (e.g., *C. elegans*, 2.2 million nucleotides)—due to the extremely slow convergence to the final value. For shorter sequences, reliable values of $H(n)$ are obtainable only up to a value of n less than 6.

For sufficiently high values of n (for example $n = 4$), we found that the redundancy is consistently larger for the primarily noncoding sequences. In fact, for most of the sequences consisting primarily of coding regions, we find that $R(n)$ is quite close to the value $R(n) = 0$ which we find for a control sequence of random numbers.

It appears that linearity of a Zipf plot is generally indicative of hierarchical ordering. For example, it is possible that a wide range of systems result in straight-line behavior when subjected to Zipf analysis and some understanding of the implications of Zipf analysis is now emerging (Czirok et al. 1995, Havlin 1995). An example that was the subject of some discussion is the remarkable linearity of the Zipf plot giving the annual sales of a company as a function of its sales rank. J.P. Bouchaud (1995) finds that this plot is linear for European companies, while M.H.R. Stanley et al. (1995) find linearity for American companies. Furthermore, M.H.R. Stanley et al. (1995) find a significant deviation from this apparent linearity at rank ≈ 100, and relate this feature to the log-normal distribution of sales (the "Gibrat law").

1.9 Acknowledgements

We are grateful to many individuals, including M.E. Matsa, S.M. Ossadnik, and F. Sciortino, for major contributions to those results reviewed here that represent collaborative research efforts. We also wish to thank C. Cantor, C. DeLisi, M. Frank-Kamenetskii, A. Yu. Grosberg, G. Huber, I. Labat, L. Liebovitch, G. S. Michaels, P. Munson, R. Nossal, R. Nussinov, R. D. Rosenberg, J. J. Schwartz, M. Schwartz, E. I. Shakhnovich, M. F. Shlesinger, N. Shworak, and E. N. Trifonov for valuable discussions. Partial support was provided by the National Science Foundation, National Institutes of Health (Human Genome Project), the G. Harold and Leila Y. Mathers Charitable Foundation, the National Heart, Lung and Blood Institute, the National Aeronautics and Space Administration, the Israel-USA Binational Science Foundation, Israel Academy of Sciences, and (to C-KP) by an NIH/NIMH Postdoctoral NRSA Fellowship.

Bibliography

[1] Arneodo, A., Bacry, E., Graves, P.V. and Mugy, J.F. (1995) *Phys. Rev. Lett.*, **74**, 3293.

[2] Azbel, M.Ya. (1973) *Phys. Rev. Lett.*, **31**, 589.

[3] Barabasi, A.-L. and Stanley, H.E. (1995) *Fractal Concepts in Surface Growth*. Cambridge University Press, Cambridge.

[4] Berthelsen, C.L., Glazier, J.A. and Skolnick, M.H. (1992) *Phys. Rev. A*, **45**, 8902.

[5] Bouchaud, J.-P. (1995) More Lévy distributions in physics. *Proc. 1993 International Conf. on Lévy Flights*, edited by M. F. Shlesinger, G. Zaslavsky and U. Frisch. Springer, Berlin.

[6] Brillouin, L. (1956) *Science and Information Theory*. Academic Press, New York.

[7] Buldyrev, S.V., Goldberger, A.L., Havlin, S., Peng, C.-K., Simons, M. and Stanley, H.E. (1993) *Phys. Rev. E*, **47**, 4514.

[8] Buldyrev, S.V., Goldberger, A.L., Havlin, S., Peng, C.-K. and Stanley H.E. (1994) *Fractals in Science*, edited by A. Bunde and S. Havlin. Springer-Verlag, Berlin, 49–83.

[9] Buldyrev, S.V., Goldberger, A.L., Havlin, S., Mantegna, R.N., Matsa, M.E., Peng, C.-K., Simons, M. and Stanley, H.E. (1995) Long-range correlation properties of coding and noncoding DNA sequences. *Phys. Rev. E*, **51**, xxx.

[10] Bunde, A. and Havlin, S., eds. (1991) *Fractals and Disordered Systems*. Springer-Verlag, Berlin.

[11] Bunde, A. and Havlin, S., eds. (1994) *Fractals in Science*. Springer-Verlag, Berlin.

[12] Czirók, A., Mantegna, R.N., Havlin, S. and Stanley, H.E. (1995) Correlations in binary sequences and generalized Zipf analysis. *Phys. Rev. E*, **52**, xxx.

[13] Havlin, S. (1995) Distance between Zipf plots. *Physica A*, **xx**, xxx.

[14] Li, W. and Kaneko, K. (1992) *Europhys. Lett.*, **17**, 655.

[15] Mantegna, R.N., Buldyrev, S.V., Goldberger, A.L., Havlin, S., Peng, C.-K., Simons, M. and Stanley, H.E. (1994) *Phys. Rev. Lett.*, **73**, 3169–3172.

[16] Mantegna, R.N., Buldyrev, S.V., Goldberger, A.L., Havlin, S., Peng, C.-K., Simons, M. and Stanley, H.E. (1995) Linguistic analysis of coding and noncoding DNA sequences. *Phys. Rev. E* (submitted).

[17] Ossadnik, S.M., Buldyrev, S.V., Goldberger, A.L., Havlin, S., Mantegna, R.N., Peng, C.-K., Simons, M. and Stanley, H.E. (1994) *Biophys. J.*, **67**, 64.

[18] Peng, C.-K., Buldyrev, S.V., Goldberger, A.L., Havlin, S., Sciortino, F., Simons, M. and Stanley, H.E. (1992) *Nature*, **356**, 168.

[19] Peng, C.-K., Buldyrev, S.V., Havlin, S., Simons, M., Stanley, H.E. and Goldberger, A.L. (1994) *Phys. Rev. E*, **49**, 1685.

[20] Schenkel, A., Zhang, J. and Zhang, Y-C. (1993) *Fractals*, **1**, 47.

[21] Shannon, C.E. (1951) *Bell Systems Tech. J.*, **80**, 50.

[22] Stanley, H.E. (1971) *Introduction to Phase Transitions and Critical Phenomena*. Oxford University Press, London.

[23] Stanley, M.H.R., Buldyrev, S.V., Havlin, S., Mantegna, R., Salinger, M.A., Stanley, H.E. (1995) Zipf plots and the size distribution of firms. *Eco. Lett.*, **xx**, xxx.

[24] Uberbacher, E.C. and Mural, R.J. (1991) *Proc. Natl. Acad. Sci. USA*, **88**, 11261.

[25] Voss, R. (1992) *Phys. Rev. Lett.*, **68**, 3805.

[26] Watson, J.D., Gilman, M., Witkowski, J. and Zoller, M. (1992) *Recombinant DNA*. Scientific American Books, New York.

[27] Zipf, G.K. (1949) *Human Behavior and the Principle of "Least Effort"*. Addison-Wesley, New York.

1.10 Biographies

Sergey V. Buldyrev

Sergey V. Buldyrev was born in St. Petersburg (Russia). After graduation from the St. Petersburg State University (M.S. in Mathematical Physics) he worked in the same university as a junior research fellow and teaching assistant. He studied scaling properties of polymers under Prof. T. M. Birshtein and received a Ph.D. in Physics in 1988. Since 1990, he has worked in the Center for Polymer Studies at Boston University as a senior research associate. His research intersts are in the application of statistical physics and computer simulation to various problems of material science and biology.

Ary L. Goldberger

Ary L. Goldberger is a graduate of Harvard College and Yale Medical School. He is currently Associate Professor of Medicine at Harvard Medical School and Clinical Director of the Electrocardiography and Arrhythmia Monitoring/Nonlinear Dynamics Laboratories at Boston's Beth Israel Hospital. Dr. Goldberger and his colleagues were among the first to apply concepts from nonlinear dynamics, including chaos theory and fractals, to cardiovascular pathophysiology and the problem of sudden cardiac death. Most recently, he and his colleagues have extended their analyses to the organization of DNA nucleotide sequences with the recent discovery of long-range correlation properties in non-coding DNA.

Shlomo Havlin

Shlomo Havlin was born Jerusalem on September 21, 1942. After obtaining his Ph.D. from Bar-Ilan University in 1972, he joined the Bar-Ilan faculty—serving as Chairman in 1984-1988. He was awarded the 1988 Landau Prize for Outstanding Research. Also in 1988, he won a MINERVA Fellowship. In 1992, he won the Humboldt Award. In 1993, he was elected Vice President of the Israel Physical Society. Currently he performs research in statistical mechanics with applications to a wide range of topics in chemistry, physics, and biology.

Rosario N. Mantegna

Rosario N. Mantegna is a graduate of the University of Palermo and has done post-doctoral research at the Max-Planck Institüt für Quantenoptik (collaborating with Professor H. Walther) and the Center for Polymer Studies at Boston University (with H. Eugene Stanley). He is currently Ricercatore at the Dipartimento di Energetica ed Applicazioni di Fisica in Palermo. Dr. Mantegna's main area of interest is stochastic and complex systems. This interest has led to research on atomic (quantum chaos), bistable (stochastic resonance), biological (DNA sequences) and economic (stock exchange) systems.

Chunk-Kang Peng

Chunk-Kang Peng was born in Taiwan. After graduating from the National Tsing Hua University (M.S. in Physics) and serving for two years in the Taiwan Army as a platoon leader in a tank unit, he went to Boston University for advanced study under Professor H. Eugene Stanley and received a Ph.D. in Physics in 1993. While pursuing doctoral research, he was awarded an NIH graduate fellowship, and, after receiving his Ph.D., he was awarded a NIH NRSA post-doctoral fellowship. Dr. Peng joined Professor Ary L. Goldberger's research group at Harvard Medical School/Beth Israel Hospital (Boston) as a research associate. Dr. Peng's research interests are stochastic phenomena in complex biological systems.

Michael Simons

Born in St. Petersburg, Russia, Michael Simons is an honors graduate of both Yale College and the Yale Medical School. After post-doctoral work at NIH (with Dr. R. Adelstein) and MIT (with Dr. R. Rosenberg), Dr. Simons joined the faculty of the Harvard Medical School and Beth-Israel Hospital as an Assistant Professor. His major research interests are angiogenesis, cell cycles, and molecular biology.

H. Eugene Stanley

Professor Stanley received his Ph.D. in Physics from Harvard in 1967, and an honorary doctorate from Bar-Ilan University in 1994. He is currently Director of the Center for Polymer Studies and Professor of Physics and Physiology at Boston University. He has co-authored over 400 papers and 10 books, including *Introduction to Phase Transitions and Critical Phenomena* (OUP), which was awarded the Choice Award for Outstanding Academic Book of 1971, and *Fractal Concepts in Surface Growth*, which was published in 1995. Stanley is co-editor-in-chief of Physica A, and serves on the editorial boards of Fractals, Nuclear Physics B, and Heterogeneous Chemical Reviews (HCR). He has received a Guggenheim Memorial Fellowship, a BP Venture Research Award, and the *Massachusetts Professor of the Year* award of the Council for Advancement and Support of Education (CASE). He has also been chosen for distinguished visiting professorships in France and in Japan, and was elected a Fellow of the American Physical Society in 1974 and a Fellow of AAAS in 1995.