Indian J. Phys. **74B** (**1**), 1–39 (2000)

IJP B
- an international journal

Order and fluctuations in DNA sequences*

S Chattopadhyay[1], A Som[1], S Sahoo[2] and J Chakrabarti[1,†]

[1]Department of Theoretical Physics, Indian Association for the Cultivation of Science, Calcutta-700 032, India

[2]Institute of Atomic and Molecular Sciences, Academia Sinica, P O Box 23-166, Taipei, Taiwan 10764, Republic of China

Received 24 May 1999, accepted 15 September 1999

Abstract : At the time the DNA was observed in pus cells, by the Swiss scientist Johan Friedrich Miescher, back in 1869, no one knew what it does. Quietly and independently, the Czech abbot Gregor Mendel, working in his pea farms, had discovered the experimental basis of heredity. This was in 1860. It took almost a century to establish that the two discoveries were interrelated . it was the DNA that determines heredity. The discovery of the genetic code revealed the other function of the DNA, namely its role in the synthesis of proteins and enzymes.

The genetic codes, made of the triplet codons, but with huge degeneracy, imply hidden periodicities. The Fourier analysis identifies this three period from the sharp peak at 1/3 frequency in the power spectrum. It turns out though that the genetic code, or the three periodicity, is not there in the complete DNA. Only for low level organisms, the three periodicity exists through the whole sequence. In higher organisms, the protein coding regions responsible for the three periodicity, are few and far between. Indeed, they constitute about 3% of the sequence for the humans. The function of the rest 97% remains unaccounted for. These parts constitute the 'junk' DNA.

From the power spectrum of the 'junk' DNA, when the 'white noise' is subtracted, a long-range hidden order is obtained. The sort of order, with the typical $1/f$ spectrum, is ubiquitous in the physical world. The analysis of the moments and the cumulants of the 'junk' DNA base distributions once again reveal the same long-range inverse power-law correlations of the bases. In the language of the distributions, we have long range-tails. These tails make the second moments diverge, leading to deviations from the Central Limit and to Lévy type distributions. The 'junk' DNA base organisation is then analogous to the distribution function of anomalous diffusion and of Fractional Brownian Motion.

The analysis of the coding parts of the DNA show some differences. In the short-range there exists the three periodicity peaks in the power spectrum. However, for short coding sequences the organisation of the bases are near random, characterized by the Hurst index close to 0.5 for the second moment. As we go to larger coding sequences, by splicing out the intervening 'junk' DNA, or by going to the prokaryotic (lower organisms) DNA sequences, the long-range inverse power-law correlations reappear. The Hurst index, for the second moment, deviates a bit from 0.5

With all these data on short-ranged periodicities, and long-range inverse-power-law correlations, we are ready to model the DNA sequences. How to create symbolic sequences with long-range order of bases? The Expansion-Modification algorithm creates such an order. In the Insertion Models sequences of different lengths are inserted, with the lengths distributed a la inverse power law. The Copying-Mistake Map is another model generating long-range order. Here the bases appear with the inverse power-law distribution in 'waiting times'. Simultaneously a point mutation is introduced to randomise the short-range behaviour. The relative strength of the long-range ordering and point mutation probability, is a parameter that is adjusted.

Keywords : DNA structure, genetic code, amino acids

PACS Nos. : 05.40.+j, 87.10.+e, 87.15.–v

E-mail : tpjc @ mahendra.iacs.res.in

[†] To whom all correspondence be addressed

[*] Dedicated to the memory of Kariamanikkam Srinivasa Krishnan, born December 4, 1898, on his birth centenary.

Plan of the Article

1. Introduction

"In the study of Nature, there is the need of dual viewpoint, the alternating interpenetration of biological thought with physical studies, and physical thought with biological studies".

–Jagadish Chandra Bose

In the last decade, the DNA sequences have drawn physicists anew. The works of Niels Bohr (Light and Life, *Nature* **131** (1933) 421) had earlier inspired a generation of physicists to look at the DNA to unravel its stucture and function. That the laws of living matter must follow a regular rational pattern was reassuringly emphasized by Erwin Schroedinger (What is life? Cambridge University Press, 1944). The subsequent explosion of interest led to the determinations of the structure of the DNA, and, later, the genetic code, two notable discoveries of the century.

The recent spate of interest in the subject stems, in part, from the realisation that, despite the progress, the DNA eludes understanding. While the genetic code does isolate one of the major functions, the "coding" regions are but a small part in many of the DNA. The functions of the "non-coding", sometimes called the "junk" parts, remain unknown. Amusingly enough, these "junk", "non-coding" regions are the largest component of the DNA. It is improbable they are there doing nothing.

The investigations over the last decade have brought some hints that the "junk" parts of the DNA do have a built-in organisation. These parts have long-range correlations of the inverse power-law form. Long-range order, the inverse power type, exists in many physical systems. Their precise physical origin remains ill-understood. Indeed, there is the well known result in physics, that for one dimensional systems long-range order is improbable. It is a challenge then to understand the unmistakable correlations in the "junk" DNA.

The coding regions, in many cases less than 5% of the DNA of higher organisms, have structure that is equally elusive. First, they show three periodicity, presumably due to the presence of the triplet codons. Second, over "short"-to-"intermediate" range they have the random statistical behaviour.

This review is about this intricate hidden structural organisation of the DNA. It is divided into five parts. Part I is a brief look at the DNA, the polymer, and its underlying constituents called the nucleotides, or more simply, the monomers. Part II gives a simple introduction to the spectral analysis of symbolic sequences such as the DNA. It also briefly discusses the ideas of information-entropy and order. Part III is a brief foray into random walkology on which a good bit of the modern DNA correlation analysis is patterned. In Part IV we discuss the underlying order of the DNA sequences. There have been some effort at modelling of the DNA sequences based on insights gleaned about its structure in the recent years. We outline the framework of some of the recent models. Needless to add, the modelling effort has a long way to go. Part V assesses the progress thus far.

The choice of topics has been dictated by our intent to make this review accessible to specialists from many fields. We would have liked to deal with some of the background material in more detail, but are restrained partly by limitations of space; more by limitations of our own knowledge.

There are many we would like to thank. Prof. Anjali Mookerjee and Prof. A B Roy allowed us to present part of this material to teachers from universities and colleges at the UGC sponsored school at the Sivatosh Mookerjee Science Centre, Calcutta. We are grateful to Prof. S C Mukherjee, who contributed substantially towards building up of our laboratory; to Prof. Ashesh Nandy for much of the initial impetus, and to Drs. Chaitali Mukhopadhyay, Sujata Tarafdar and Papiya Nandi for many useful discussions. The speakers and the participants at the School of Complex Systems, Jan 30 — Feb 3, 1995 [*Indian J. Phys.* **69B** (1995)] provided the initial spark; we thank them all.

2. An overview of DNA

> **"Living matter, while not eluding the laws of physics as established to date, is likely to involve other laws of physics hitherto unknown which, however, once they have been revealed, will form as integral a part of this science as the former".**
>
> *- Erwin Schroedinger*

At about the time, in the later part of the nineteenth century, when the doctrines of classical physics had reached its height, a fascinating and far reaching new discipline of research, far removed from classical physics, was silently born. The ideas were conceived by Gregor Mendel, around 1860, at the Augustinian monastery at Brno (Czechoslovakia), on experiments with breeding of pea-plants. The results were published in 1866 in the obscure Verhandlungen des naturforschenden Vereines in Brunn (The Proceedings of the Society of Natural Sciences in Brno). Mendel had studied the inheritance characters, such as plant height, colour of flower, the shape of seed, of the usual garden peas, and concluded that heredity works on clear, logical principles that are experimentally accessible and verifiable.

Curiously, Mendel's work went unnoticed for a good thirtyfour years till 1900, about the time Max-Planck was busy with his experiments on blackbody radiation, when three scientists — Hugo de Vries, Carl Correns and Erich von Tschermak independently conceived of and performed experiments that showed heredity follows clear physical principles. Studying the literature they realised they had rediscovered the ideas of Mendel conceived more than three decades earlier.

2.1. *From peas to fruit flies :*

The work of Mendel, confirmed now by De Vries, Correns and Tschermak, paved the way for the rational scientific approach to the characteristics of living organisms; how these are passed from one generation to the text. Within a decade from 1900 experiments established that these informations reside in the chromosomes and are passed on duing the process of cell division. The term gene was used to describe the objects residing in chromosomes that carry these informations. No one yet knew what these objects were. Figure 1 gives the idea of an idealized cell that, being the structural and functional unit of a living organism, carries the chromosomes.

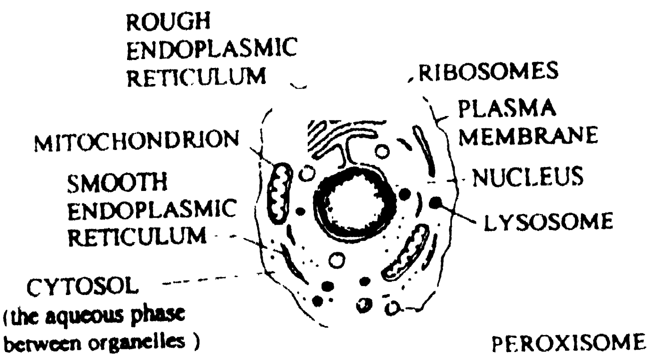

ROUGH ENDOPLASMIC RETICULUM
RIBOSOMES
PLASMA MEMBRANE
MITOCHONDRION
NUCLEUS
SMOOTH ENDOPLASMIC RETICULUM
LYSOSOME
CYTOSOL
(the aqueous phase between organelles)
PEROXISOME

Figure 1. Diagram of an idealized animal cell.

It was about this time in May 1910, came the white-eyed fruit fly from the laboratory of Thomas Hunt Morgan [1]. The fruit flies exist in many different forms, and crossing them together the "fly room" of Morgan created whole set of varieties in accord with Mendel's ideas. Careful experimental techniques developed by Morgan mapped the position of genes in the chromosomes for the characteristic features of fruit flies (Figure 2) [2].

Fruit flies, or *Drosophila melanogaster* as they are technically called, because of their variety, provided the ideal laboratory for the study of inheritance. The science of heredity that began in the pea gardens of Mendel took off on the wings of *Drosophila melanogaster.*

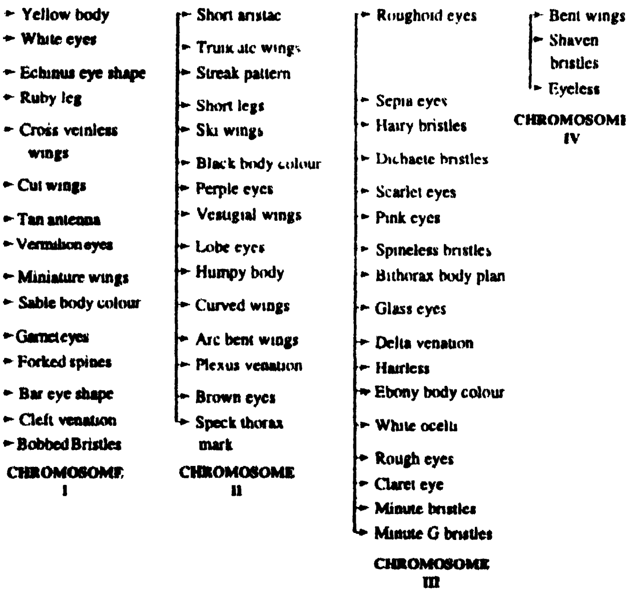


Figure 2. Positions of 50 different genes on the 4 chromosomes of the fruit fly, *Drosophila melanogaster.*

2.2. *Here comes Niels Bohr :*

Far away from garden peas and fruit flies a group of physicists, inspired by Niels Bohr, began to work on the issue of inheritance. The lecture of Bohr at an international congress in 1932, published the following year in Nature, provided the spur to physicists, trained in quantum mechanics, to work on the ideas laid out by Mendel and Morgan. The questions what genes were, are how they worked — haunted them. Max Delbrück, a nuclear physicist from Göttingen (migrated to the US in 1937), played a pivotal role in shaping the course for the next three decades [3]. In 1940 he, along with Salvador Luria and Alfred Hershey, set up the Phage group, consisting of physicists, chemists and bilogists, that led, eventually, to cracking the mystery of genes. The group was named after bacteriophages, which are viruses that infect bacteria.

2.3. *What the genes are made of :*

That chromosomes have the constituents, the genes, that determine heritage, led to intense exploration of the genetic material. The analysis of chromosomes, by chemical methods, established that are made of proteins and nucleic acids. This was known by 1920. The nucleic acid, namely deoxyribonucleic acid (DNA), or the protein, or a combination of the two, *i.e.* nucleoprotein, must transmit the data of one generation to the next. The early suspicion pointed the finger at protein. The reason being, protein was known to be a long polymer made up of 20 amino acid monomers. Since the amino acid residues (*i.e.* the monomer units of protein) appear in arbitrary order, the protein polymers could contain large amount of information. In contrast, initially the structure of the DNA was incorrectly determined. The constituents – adenine (A), guanine (G), cytosine (C) and thymine (T) that make up the DNA — were put together in a way that had little possibility of storing the vast amount of information required. By the late thirties it became clear, however, that the DNA is a polymer of A, G, C and T and, therefore, could exist in large number of variable forms suitable for storage of information, just like protein. The crucial evidence that it is the DNA that stored the genetic data came from experiments.

In 1928, Frederick Griffith studied both virulent (disease causing) and avirulent (harmless) forms in *Streptococcus pneumoniae*, the agent that causes pneumonia, and found out that the principle responsible for the transformation of bacteria from one form to the other was actually the genetic material. But he did not identify the transforming principle. Afterwards, significant experiments in this direction were carried out by Oswald Avery and coworkers (Rockefeller Institute, New York) on the same bacteria. They used degrading agents protease and ribonuclease enzymes to selectively degarde proteins and nucleic acids respectively and study the information carrying capability of the resulting genes [4]. Alternatively, in experiments carried out by Alfred Hershey and Martha Chase at Cold Spring Harbor Laboratory, radioisotope labelling of protein and the DNA were carried out. Proteins carry sulphur and can be doped with ^{35}S. The DNA carry phosphorus and were doped with ^{32}P. The information carrying agent in bacteriophage T2 was studied with these doping agents. They concluded from the results that the DNA carries the information [5].

Avery's results appeared in 1944, but remained unaccepted. Even with the Hershey-Chase experiment of 1951–52, there remained some lingering questions. The determination of the structure of the DNA by Watson and Crick in 1953 established the information carrying capability of the DNA and laid at rest these doubts. Much later, in the 1970's, with the advent of recombinant DNA technology, that injected pure DNA in plants, insects, yeast, bacteria etc., the role of DNA as the sole genetic material became experimentally established.

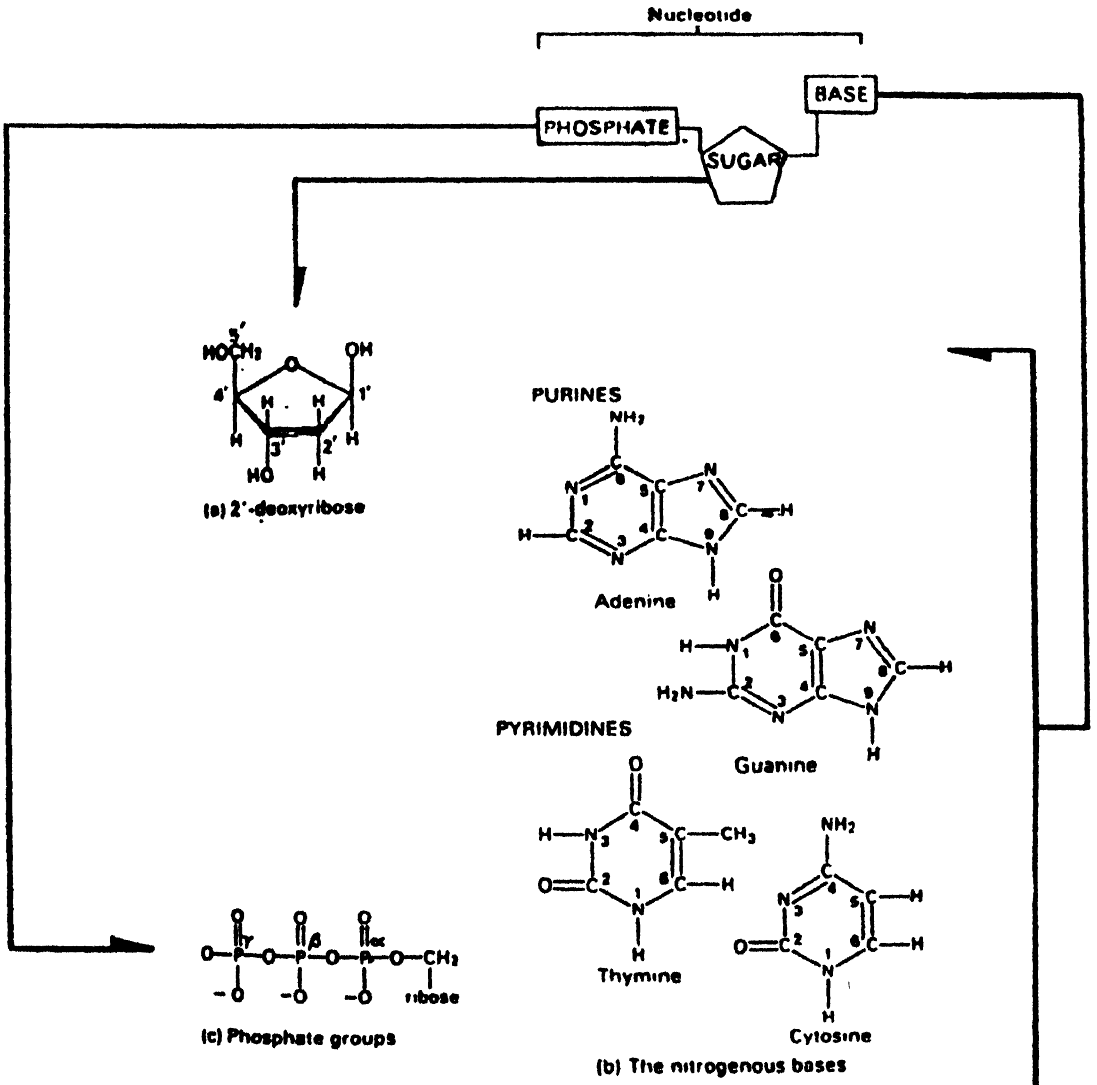
2.4. The DNA :

In close parallel with the experiments and ideas put forward by Mendel, Morgan, Griffith, Avery, Hershey, Chase and others on inheritance and the role of the DNA, another group of scientists were busy unravelling its structure. The DNA was isolated from pus cells by Johan Friedrich Miescher in 1869, and the majority of its nitrogenous bases were identified in 1894. The sugar component of the DNA came to be identified by Hammersten in 1900; the exact structure of the sugar ingredient, the deoxyribose, was obtained by Levene by 1929. By 1934, Caspersson had established its long chain polymer form capable of existing in variable configuration of the bases A, T, G and C. This variability confers it the potential to store large amount of information. That the bases A, T, G and C follow a definte compositional constraint was established in 1950 by Chargaff [6]. The X-ray diffraction studies on crystals of the DNA by Rosalind Franklin in 1952 showed the DNA to be a helix. The methodology of X-ray diffraction studies were established earlier by Maurice Wilkins. The final step came in 1953 by Watson and Crick, who put together all these informations to arrive at the double helical structure of the DNA [7].

2.5. The building blocks :

The monomers

The DNA is made up of a chain of four monomers, arranged in arbitrary order. The monomers, also called nucleotides, are



(a) 2'-deoxyribose

(c) Phosphate groups

(b) The nitrogenous bases

Figure 3. The components of any deoxyribonucleotide.

in turn made of three distinct entities : the sugar, the nitrogenous base and the phosphoric acid.

The sugar : It is made of a ring of 5 carbon atoms, labelled from 1' to 5'. The reason for the primes we explain later. It has the form of sugar called ribose, out of which at the 2' position an oxygen atom is removed. Hence the name 2'-deoxyribose (Figure 3a).

The nitrogenous bases : The nitrogenous bases come in four different types, labelled : A for Adenine, T for Thymine, G for Guanine and C for Cytosine. Hence the four monomers may be denoted by the symbols A, T, G, C. Out of the four, A and G are called purines and are both made of two rings (Figure 3b). T and C have single-ringed forms (Figure 3b). The positions of atoms in the bases are labelled from 1 onwards. It is for this reason the positions in the ribose are denoted by primes. These four bases attach on to the site 1' of the ribose sugar.

The phosphoric acid : The phosphoric acid group attaches to the 5' carbon of the ribose sugar. The phosphates that attach could be the monophosphate, the diphosphate or the triphosphate. The individual phosphate groups are labelled α, β and γ, with the convention that the α-phosphate attaches on to the deoxyribose (Figure 3c).

The polymer

The monomers put together in a chain form the polymer, also called the polynucleotide. The individual monomers attach to the other through the phosphate groups. The α-phosphate attaches to the 5' position of one ribose and 3' position of another forming the linkage (Figure 4). Of the α-, β-, γ-phosphates, the β and the γ detach during polymerisation, leaving only the α to provide the connecting links of one ribose to the next.

There is a sense of direction in the polymer. One end (phosphate at 1' carbon) is the P-terminus, the other end has the 3'-OH terminus. Thus we have the polymer running, so to speak, from 5' to 3' as the two ends are different.

The polymer can have arbitrary number of monomers in any arrangement of A, T, G and C. When we talk of the DNA sequence, we mean the sequence of A, T, G and C in this polymer chain.

2.6. *The double helix :*

That the DNA is a polymer mode of A, T, G, C monomers tied together through phosphate links was known prior to 1953. The work of Wilkins on X-ray diffraction and its application to crystals of DNA fibers by Rosalind Franklin in 1952 established that the DNA has a helical shape [8]. It was left to Watson and Crick to show that DNA consists of two polymer chains, both of A, T, G, C, in the shape of double helix. Of the two polymers, one runs from 5' to 3'; the other, the complementary polymer, runs in the opposite direction, *i.e.*, from 3' to 5'. The two polymers are held together by hydrogen bonds runing between the nitrogenous bases [9,10] (Figure 5).

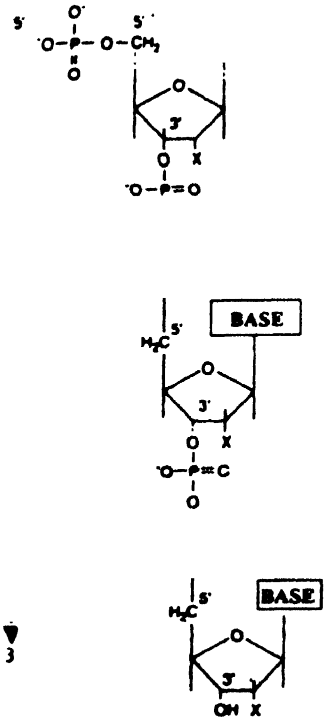


Figure 4. Structure of a trinucleotide, as it runs from 5' to 3' direction. If *X* is H, the sugar is a deoxyribose one and so the structure is DNA. If *X* is OH, the sugar is a ribose one and so the structure is RNA.

The distance between the polymer chains is such that the purines (A and G of two rings) of one polymer connects two the pyrimidines (T and C of single rings) of other. Indeed A connects through two hydrogen bonds to T; G connects through three hydrogen bonds to C. While we are not going to be discussing the energetics to the macromolecules, clearly the triple bonds between G and C imparts greater stability to chains that have higher G or C content. The A binding to T, and G binding to C of the complementary chain makes the helix satisfy the compositional contraint observed by Chargaff.

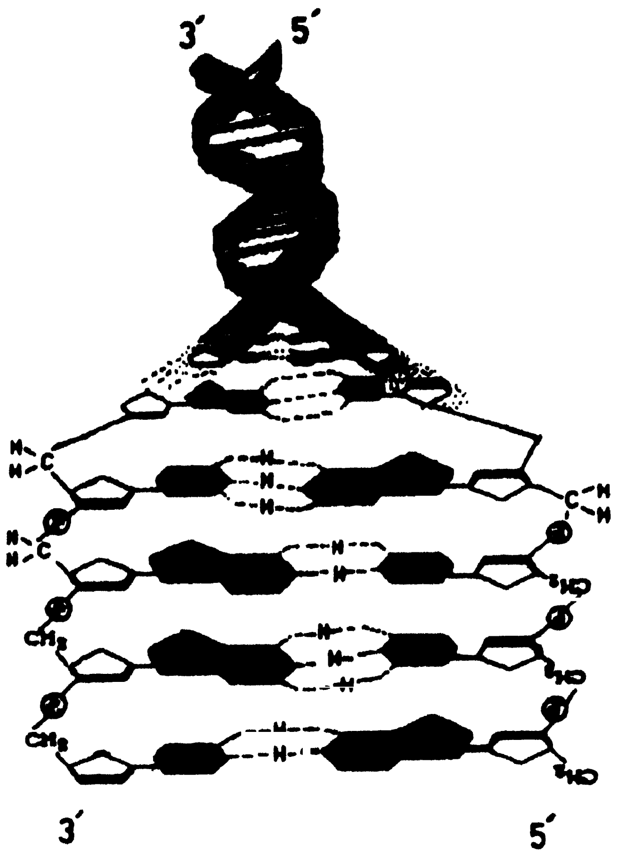


Figure 5. The two antiparallel DNA strands are connected together by non-covalent hydrogen bonding between paired bases. A and T are connected by two hydrogen bonds; while G and C are held together by three hydrogen bonds.

2.7. *The DNA organisation :*

The DNA we know, from experiments of Avery, Hershay-Chase, is the genetic material. The initial experiments were carried out with low-level organisms, such as bacteria and bacterophages. Questions remained whether in the higher organisms the DNA played the same exclusive role. The proteins present in chromosomes, could they carry information on heritage? Some of these questions were laid to rest with the advent of Recombinant DNA Technology in the seventies. Here pure DNA is introduced into the cells and its effects are observed. The experiment with recombinant DNA technology establishd the central and the exclusive role of the DNA as the genetic material.

Before we look at the major functions of the DNA, let us briefly summarize the organisation of the DNA in the cells.

The DNA occurs in the chromosomes or in the mitochondria of higher living organisms, called the eukaryotes. In the eukaryotes, the chromosomes, the mitochondria, the golgi bodies are distinct structures inside the cells. These structures are surrounded by membranes. The eukaryotes could be unicellular, or have many cells.

In contrast the prokaryotes, such as bacteria, are organisms that do not have structures such as the nucleus, mitochondria *etc.* well segregated inside the cells.

There could be several chromosomes, and in each chromosome can reside several genes (Table 1).

Table 1. The average number of genes present in each chromosome varies among species.

Name of the Organism	Total No. of Chromosomes	Total No. of Genes (Approx.)	Genes/Chromosomes (Average)
E. coli (Bacteria)	1	2,800	2,800
Baker's Yeast	16	8,750	550
Human	23	50,000	2,200

The DNA molecule, the long bi-stranded polymer, has discrete segments called genes. These discrete segments are not discontinuous but are connected to one another by intergenic DNA sequences [11]. The length of the intergenic regions vary. In lower organisms, the intergenic regions are usually short, or could be absent altogether. In higher organisms, most of the genes are well-separated with long intergenic DNA regions.

The genes are segments of the DNA located on one of the strands of the bipolymer. The strand carrying the gene is called the template strand, and the sequence is read from the 5' to the 3' direction. The template strand differs from gene to gene.

The gene itself is not one continuous segment, but is interspersed with DNA sequences that do not carry known genetic functions. The parts of the segments of genes that carry genetic information are called exons; the regions in between are called introns [12, 13] (Figure 6). A gene may be interrupted with many introns. Table 2 shows the variation in the number of introns for a few human genes. For lower organisms, the introns are shorter, or may be absent altogether.

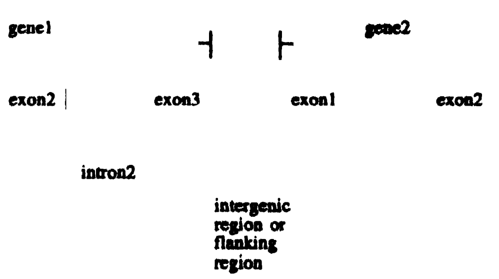


Figure 6. Any two non-overlapping genes are separated by an intergenic or flanking region. Again a gene may be divided into a number of exon (*i.e.* coding) and intron (*i.e.* non-coding) regions.

Table 2. Number and proportion of introns differs in different genes of the same organism, *e g* human

Name of the Human Gene	Total Length (kilobase)	Total No. of Introns	Proportion of Intron (% length)
Insulin	1.4	2	67
Serum albumin	18	13	88
Phenylalanine hydroxylase	90	25	97
Cystic fibrosis trans-membrane regular	250	26	98
Dystrophin	2,300	> 100	99

2.8. *The DNA functions :*

The function of the DNA was summarized in 1958 in Crick's Central Dogma. Simply stated, the DNA sequences in the genes make the RNA (ribonucleic acid) that make protein [14]. It is these proteins that allow organisms to carry out the multitude of functions necessary for living. The RNA is almost a copy of the DNA sequence, with one of the nitrogenous bases thymine is replaced by uracil, denoted by the symbol U (Figure 7). Thus the DNA is responsible for synthesis of all the proteins [15] (enzymes that catalyse reactions are proteins too).

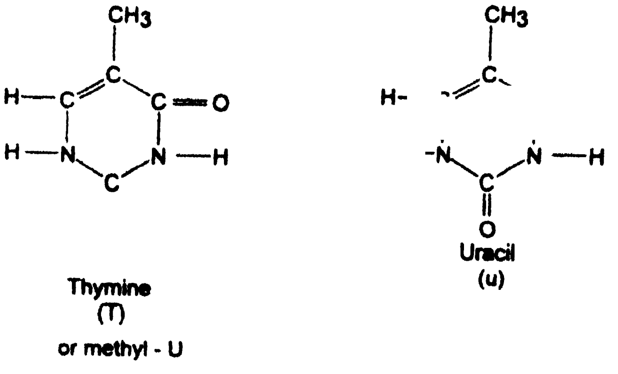


Figure 7. Uracil (U) is present in RNA; whereas Thymine (T), nothing but the methylated Uracil, is present in DNA.

The detailed chemical pathways that lead from the DNA to the RNA to the proteins is beyond the scope of the present review. These chemical pathways are summarized in Figure 8 [16].

We now discuss in brief the proteins, their structures, and the genetic code. The genetic code gives us the mapping of the monomers of the DNA, namely A, T, G and C, to the monomers, *i.e.* the amino acids of the protein polymer.

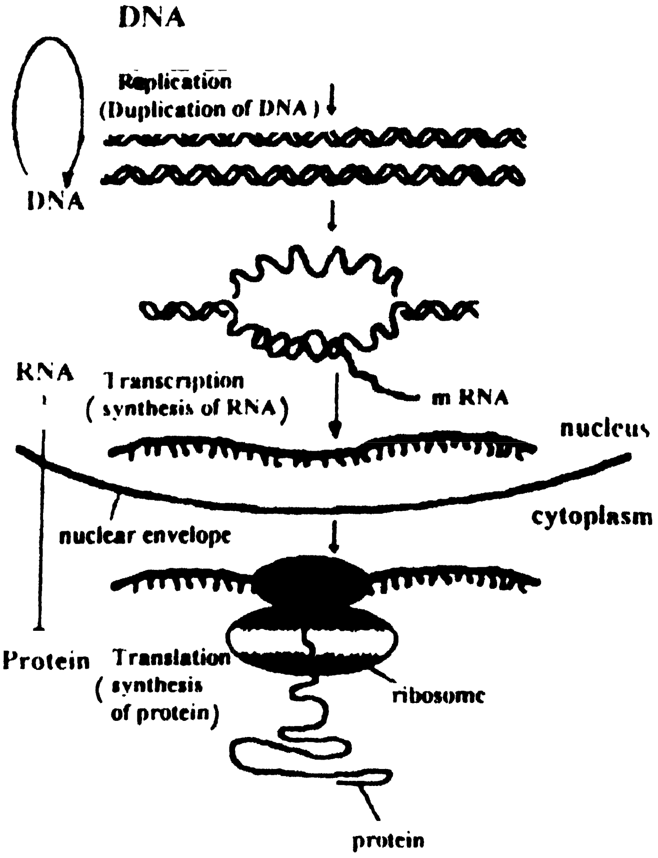


Figure 8. The Central Dogma of molecular biology . The DNA replicates its information through *replication*; the DNA gives rise to messenger RNA (mRNA) during *transcription*; in eukaryotic cells, the mRNA is *processed* by splicing and migrates from the nucleus to the cytoplasm; the ribosomes "read" the information coded in mRNA and use it for protein synthesis by *translation*

2.9. *The protein polymer :*

The protein is a polymer of monomers called amino acids, sometimes also called peptides (the polymer in this language is called the polypeptide). The monomers, *i.e.* the amino acids are twenty in number; their structures are given in Table 3. They are joined together by chemical bonds, called the peptide bonds shown in Figure 9.

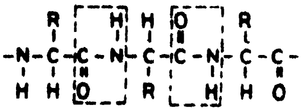


Figure 9. The peptide bond is formed by the interaction of two amino acids with the elimination of water between the NH_2 and COOH groups.

2.10. *The protein structure :*

The protein structure given in Figure 10(a) is usually referred to as the primary structure. The polymer that is protein, in its "denatured" form assumes its primary structure; usually though the structure of the polymer exists in levels of folded

Table 3. The categories, symbols and structural formulae of 20 different amino acids.

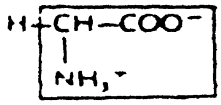
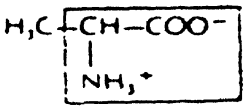
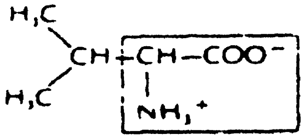
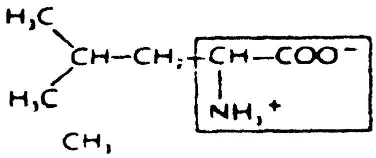
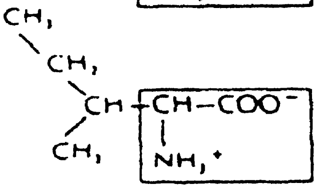
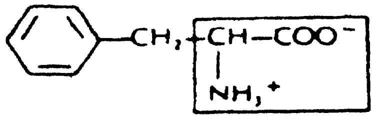
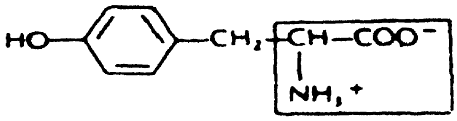
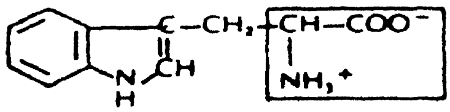
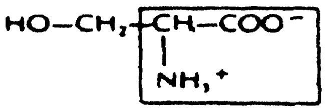
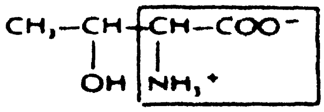
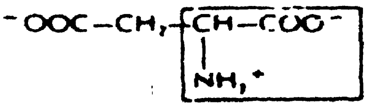
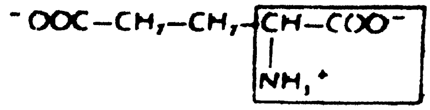
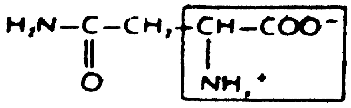
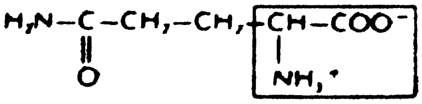
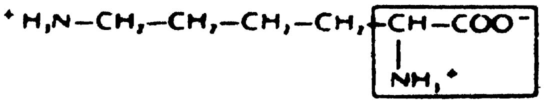
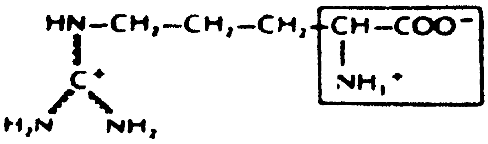
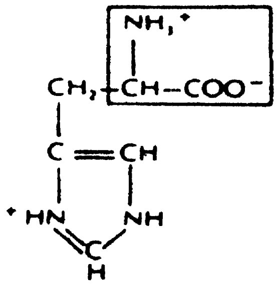
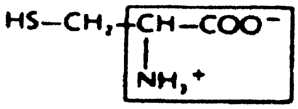
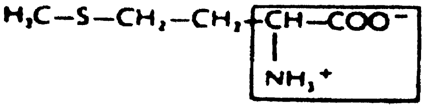
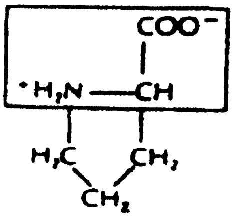
Name	Symbol	Structural Formula
Aliphatic nonpolar side chains		
Glycine	Gly (G)	
Alanine	Ala (A)	
Valine	Val (V)	
Leucine	Leu (L)	
Isoleucine	Ile (I)	
Aromatic side chains		
Phenylalanine	Phe (F)	
Tyrosine	Tyr (Y)	
Tryptophan	Trp (W)	
Hydroxyl-containing side chains		
Serine	Ser (S)	
Threonine	Thr (T)	
Acidic side chains		
Aspartate	Asp (D)	

Table 3. *(Cont'd.)*

Name	Symbol	Structural Formula
Glutamate	Glu (E)	
Amidic amino acids Asparagine	Asn (N)	
Glutamine	Gln (Q)	
Basic side chains Lysine	Lys (K)	
Arginine	Arg (R)	
Histidine	His (H)	
Sulfur-containing side chains Cysteine	Cys (C)	
Methionine	Met (M)	
Imino acid Proline	Pro (P)	

forms labelled secondary, tertiary and quaternary structures (Figure 10).

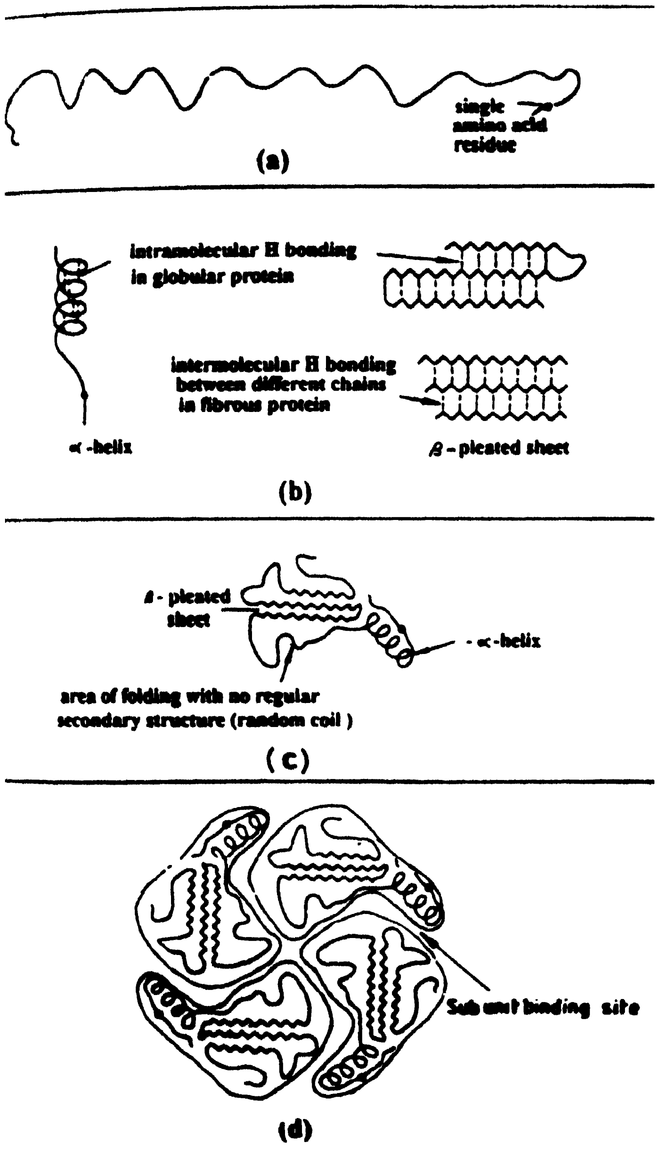

single amino acid residue
(a)
intramolecular H bonding in globular protein
intermolecular H bonding between different chains in fibrous protein
α -helix
β - pleated sheet
(b)
β - pleated sheet
α -helix
area of folding with no regular secondary structure (random coil)
(c)
Sub unit binding site
(d)

Figure 10. The structure of a protein in four hierarchies. (a) The *primary* structure of a protein describes the order of covalently linked amino acid residues. (b) The *secondary* structure, either α-helix or β-pleated sheet or a combination of both, shows the role of CO-NH hydrogen bonds, either intramolecular or intermolecular in nature. (c) The *tertiary* structure describes the way the chains with secondary structure interact through the side chains of the amino acid residues to form a 3-D shape. (d) The *quaternary* structure describes the interaction, through weak bonds, of the polypeptide subunits.

It is to be noted that the quaternary (or the tertiary, or the secondary) structure, upon heating, or upon chemical treatment with urea, denatures to the primary form made up of the sequence of amino acids. Upon renaturation, *i.e.* upon cooling for instance, it resumes spontaneously its correct tertiary structure. It is assumed, therefore, that the amino acid sequence, at the primary level (which depends on the sequence of the DNA it is made from), determines the tertiary structure of the protein. Thus built into the DNA exists the information on the amino acid sequence that in turn determines the folds of its structure [17].

2.11. The genetic code :

About 1953 when Watson and Crick put together, from the known results, the structure of the DNA, the work on the genetic code began in earnest. It continued through the fifties and was not completed until 1966. A large group of scientists— Crick, Yanofsky, Brenner, Ochoa, Nirenberg, Matthaei, Khorana, Leder and others— unravelled the genetic code.

Since the amino acid monomers are twenty in number it was clear early on that the nucleotide bases (remember they are 4 in number— A, T, G and C), have to work in combination to give rise to these twenty variety. Clearly two of them can make upto 4 × 4 = 16 varieties. Three of them can make upto 4 × 4 × 4 = 64 types. Thus, three is the least number of the DNA monomers necessary [18]. However, since three of them can make 64 different types, while the amino acids number just twenty, the genetic code has a high degeneracy (codon degeneracy) [19,20]. The genetic code, as obtained in 1966, is summarized in Table 4 [21].

Table 4. The genetic code.

		2nd	base	in	codon		
		U	C	A	G		
		Phe	Ser	Tyr	Cys	U	
	U	Phe	Ser	Tyr	Cys	C	
1st		Leu	Ser	STOP	STOP	A	3rd
		Leu	Ser	STOP	Trp	G	
		Leu	Pro	His	Arg	U	
	C	Leu	Pro	His	Arg	C	
Base		Leu	Pro	Gln	Arg	A	base
		Leu	Pro	Gln	Arg	G	
		Ile	Thr	Asn	Ser	U	
	A	Ile	Thr	Asn	Ser	C	
in		Ile	Thr	Lys	Arg	A	in
		Met	Thr	Lys	Arg	G	
		Val	Ala	Asp	Gly	U	
	G	Val	Ala	Asp	Gly	C	
codon		Val	Ala	Glu	Gly	A	codon
		Val	Ala	Glu	Gly	G	

Legend :

Amino acids specified by each codon sequence on mRNA Key for the above table :

Phe : Phenylalanine	*Ser : Serine*	*His : Histidine*	*Glu : Glutamic acid*
Leu : Leucine	*Pro : Proline*	*Gln : Glutamine*	*Cys : Cysteine*
Ile : Isoleucine	*Thr : Threonine*	*Asn : Asparagine*	*Trp : Tryptophan*
Met : Methionine	*Ala : Alanine*	*Lys : Lysine*	*Arg : Arginine*
Val : Valine	*Tyr : Tyrosine*	*Asp : Aspartic acid*	*Gly : Glysine*

A = adenine G = guanine C = cytosine T = thymine

Aside from the codes given in Table 4, there are several other features that are important to note.

(i) Stop Codons : Some triplet combinations, namely, UAA, UGA and UAG do not code for amino acids. Presence of them in the RNA stops the process of protein synthesis. These are therefore called stop codons (Note that U stands for uracil).

(ii) Start Codon : The triplet AUG that codes for the amino acid methionine also acts as the start codon. The protein synthesis begins at the position AUG occurs. In the final protein methionine may initially occur at the first position only to be removed later by further processing.

(iii) Non-universality of the Codes : The genetic code, given in Table 6, back in 1966 appeared universal. Subsequently small deviations have been observed, first in mitochondrial DNA sequences, later in some nuclear sequences as well. Some of these deviations from universality are summarized in Table 5 [22].

Table 5. Examples of some nuclear and mitochondrial non-standard codons.

Name of the Organism	Location of the Genes	Codon	Codes for	Universally codes for
Protozoa	Nucleus	UAA, UAG	Glutamine	Termination
Candida cylindracea	Nucleus	CUG	Serine	Leucine
Baker's Yeast	Mitochondria	UGA	Tryptophan	Termination
		CUN*.	Threonine	Leucine
		AUA	Methionine	Isoleucine
Drosophila melanogaster	Mitochondria	UGA	Tryptophan	Termination
		AGA	Serine	Arginine
		AUA	Methionine	Isoleucine
Mammals	Mitochondria	UGA	Tryptophan	Termination
		AGA, AGG	Termination	Arginine
		AUA	Methionine	Isoleucine

*N stands for any nucleotide.

In as far as is known, the departure from the genetic code of Table 4, are rare. The results of 1966 continue to hold for most of the coding regions.

Table 6. Variations in the length of the DNA segments among different organisms.

Name of the Organism	Genome size (kilobase)	Total No. of Chromosomes	Average Length of DNA/Chromosome (kilobase)
E. coli (Bacteria)	4,000	1	4,000
Baker's Yeast	20,000	16	1,250
Drosophila melanogaster	165,000	4	41,250
Human	3,000,000	23	130,000
Salamander	90,000,000	12	7,500,000

2.12. Experiments with the DNA :

The present knowledge about gene structure is mostly due to the enormous applicability of 'recombinant DNA technology' The DNA molecule created *invitro* by ligating together pieces of the DNA that are not normally contiguous is termed a 'recombinant DNA technology'. The r-DNA technology comprises of all the techniques involved in the construction, study and use of those molecules. At the heart of this technology are the nucleic acid enzymes acting as tools that allow the DNA and the RNA to be manipulated [23].

2.12.1. Enzymes

Restriction endonucleases are a group of enzymes which actually initialized the development of this technology and naturally deserve the most importance. A restriction endonuclease cuts DNA moleculs only at a limited number of specific nucleotide sequences (Figure 11a).

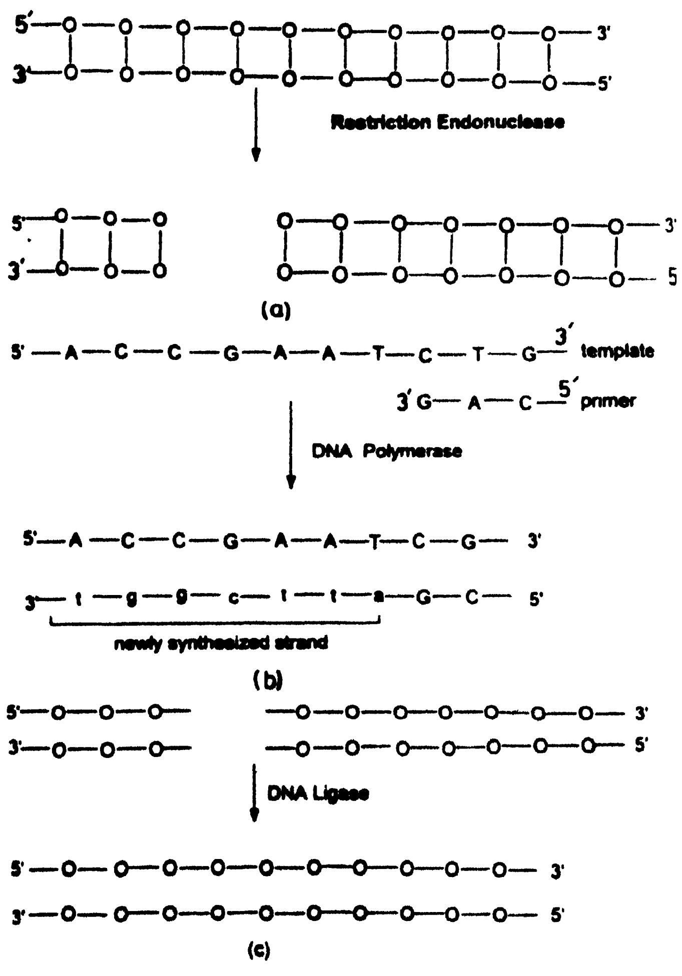


Figure 11. Three important classes of enzymes, frequently used in recombinant DNA technology. (a) A restriction *endonuclease* cleaves double-stranded DNA only at specific sites. (b) The basic reaction of a *DNA polymerase* : a new DNA strand is synthesized in the 5' to 3' direction. (c) A *DNA ligase* joins together two individual fragments of double-stranded DNA.

DNA polymerases make complementary copies of DNA templates and are useful in the production of labeled probes, DNA sequencing and also DNA amplifiction (Figure 11b).

DNA ligases are the enzymes that repair single-strand discontinuities in double-stranded DNA molecules in the cell. The purified form of this enzyme joins the DNA molecules together to form a recombinant DNA (Figure 11c).

2.12.2. Analytical techniques

A number of recombinant DNA-based analytical techniques [24] have been found to have tremendous impact in the medical sciences. 'Southern blot analysis' is one of those diagnostic techniques; it transfers bands of DNA from an agarose gel to a nitrocellulose or similar membrane and is used to detect specific sequences contained on a DNA fragment generated by restriction enzyme digestion within a mixture of all the restriction enzyme fragments of genome. It also sets the basis of 'restriction fragment length polymorphism (RFLP) linkage analysis' and 'DNA fingerprinting'.

RFLP is a mutation that gives rise to a detectable change in the pattern of fragments obtained when a DNA molecule is cut with a restriction endounclease. The restriction fragment markers that demonstrate close linkage analysis; it has become a means of screening individuals for defective genes responsible for genetic diseases.

DNA fingerprint analysis is just a variation of RFLP analysis in which the probe hybridizes to the hypervariable regions or HVRs. Its uses include forensic identification, indentification of parentage and also the evaluation of the success of bone marrow transplants.

DNA sequencing is another strong and informative DNA analytical technique that determines the order of nucleotides in the DNA molecule. DNA can be sequenced either chemically, by the Maxam and Gilbert pocedure [25], or enzymatically, by the Sanger method [26]; the latter is easier and qualitatively superior to the chemical method. The invention of the automated DNA sequencer has now provided an enormous pace in the field of research in molecular biology.

Polymerase chain reaction (PCR) is another very powerful technique [27] that enables multiple copies of a DNA molecule to be generated by enzymatic amplification of arget DNA molecule. For each round of synthesis, the amount of DNA is doubled. Thus, 30 rounds yield more that 1.0 × 10^9 copies of a region of DNA from one molecule. It uses are mainfold. Genes susceptible to mutations that cause a disease can be quickly amplified and sequenced. PCR helps to readily detect viral or bacterial infections. It has also got a lot of importance for forensic uses. Thus PCR, DNA sequencing and Southern blot analysis, acting in concert, has put the r-DNA technology at the foremost position in the present world of molecular biophysics.

2.13. The DNA habitat :

To appreciate the meaning of the mathematical analysis that the DNA sequences are subjected to in the following, we discuss briefly where and how the DNA resides. It is known that the DNA resides in the nucleus of eukaryotes or in the nucleoids of the prokaryotes. The DNA is also found in the mitochondria of all eukaryotes and in the chloroplasts of plants (eukaryotes). The mitochondrial and the chloroplast DNA synthesize proteins necessary for the function of these two bodies inside the cells. The genetic code for the mitochondrial DNA differs in a few instances from that of the nuclear DNA. Interestingly the majority of the proteins required for the mitochondrial functions are synthesized in the nucleus and transported to the mitochondria. Why the mitochondria has to work as a separate centre for protein synthesis remains unknown.

The DNA residing in the nucleus, in chromosomes, is being referred to as the nuclear DNA It is with them that we concern ourselves through this review

The DNA molecule is split into a number of segments each contained in one chromosome. The total number of chromosomes vary from one organism to another. The lengths of the DNA segments vary from chromosome to chromosome [28]. Table 6 gives some of these variations for a few samples.

The dimension of the chromoseme falls in the 10^{-6} meter range. The DNA segments that fit into them could be several centimeters in length. It is known that chromosomes contain mixture of the DNA and the proteins. These proteins (called histones) help the DNA to wind around and compactify inside the chromosomes In the eukaryotes, and in the prokaryotes, enzymes help in the process of compactification. The DNA is said to supercoil with their aid.

The process of compactification has to follow numerous constraints to allow freely the synthesis of proteins to occur. As the process of synthesis follows from one end towards the other, the DNA has to untangle at least locally [29]. The question of whether DNA compactification can allow for knots remains unanswered.

2.14. The DNA sequence :

The DNA molecule, the bistranded polymer, as we have noticed, is made up of monomers, called nucleotides A, T, G and C. The two strands are complementary, that is, the specification of nucleotide sequence of one strand completely specifies the sequence of the other. A and G in one couple to T and C in the other respectively through hydrogen bonds that keep the bistrand together. The specification of sequence in one, therefore, is sufficient.

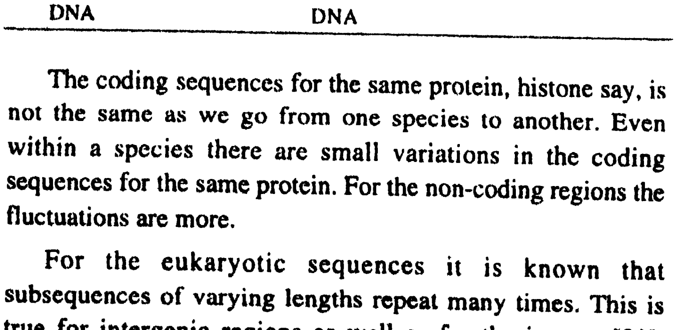
The template strand is the one that takes part in the initial stage of protein synthesis. The DNA sequence of the template strand, by convention, is read from 3' to 5' direction. The template strand synthesizes the complementary RNA molecule. The DNA sequences that are presented are of the non-template strand in the 5' to 3' direction. The reason is that the RNA strand is a copy of the non-template strand (except for thymine, T replaced by Uracil, U), and amino acid is formed from this RNA sequence. The convention, therefore, is to describe the non-template strand.

The DNA bipolymer is made up of genes and intergenic regions. The intergenic sequences usually are much larger than the genic sequences. The genes, in turn, are made up of the coding, *i.e.* the exons, and the non-coding, *i.e.* the intron regions. The intron regions for higher eukaryotic beings far exceed the exons.

The coding regions, the exons, carry the triplet codons. The codons are degenerate in the sense that many triplets give rise to the same amino acid. The second position of the codon, except for the case of serine, is nondegenerate; the first position is degenerate; and the third position has more flexibility. The exon region begins with the start codon and ends in the stop codon.

The exon region is preceded, in the immediate vicinity by promoter regions that alert biomoleculer agents responsible for the protein synthesis about the upstream coding sequence. The exons are interspersed with non-coding intron regions. The part that the introns play remains unknown. The composition of the sequence of human genome, about 6 billion base pairs long, gives a view of the relative proportions of coding (exon), non-coding (intron) and integenic regions [30]. This is given in Table 7.

Table 7. Broad subdivisions of the human genome, approximately 6,000,000 kb in length, with about 50,000–100,000 genes, split into 23 chromosomes, each containing a single, linear, double-stranded DNA molecule


Human Genome
(approx 6×10^9 bp)
Genes and generelated sequences
Intergenic or extragenic DNA
<10%
>90%
Coding DNA
Non-coding DNA

The coding sequences for the same protein, histone say, is not the same as we go from one species to another. Even within a species there are small variations in the coding sequences for the same protein. For the non-coding regions the fluctuations are more.

For the eukaryotic sequences it is known that subsequences of varying lengths repeat many times. This is true for intergenic regions as well as for the introns [31]. Table 8 gives an idea of these repeats for the human sequences.

Table 8. A few examples of repetitive human DNA.

Family	Location	Average size of Repeat Unit (bp)	Number of copies of Repeat Units
Telomeric	Telomeres	6	$2–3 \times 10^4$
Hypervariable	All chromosomes, often near telomeres	9–64	3×10^4
$(CA)_n/(TG)_n$	All chromosomes	2	7×10^6
Alu	Euchromatin	250	7×10^6
Kpn (L1)	Euchromatin	1,300	6×10^4

2.15. *Order and fluctuations in the DNA sequences :*

The DNA sequences, by convention, refer to the series of nucleotides, A, C, G, T, read on the non-template strand from 5' to 3' direction. The reason for the non-template strand has been discussed earlier.

The question that arises naturally is : What are the characteristics of these DNA sequences? For one, we know that as far as the coding sequences are concerned the genetic code is important. The triplet codons sit side by side. In cDNA (coding DNA) there does exist an order, albeit of short range. The cDNA, however, is but a small part of the DNA sequence. What happens for the introns and the intergenic regions? Does order, or correlations, exist in them? If they do, what do they physically imply?

It has been argued that the sequence carries all the physiobiological information. So far only a small part of it, namely the genetic code, has been deciphered. The information stored in the other regions remains to be understood.

In these other domains, the introns and the intergenics, are the sequences of the nucleotides (A, T, G and C) random? If they are random, perhaps they do not carry any useful information. If they are not random, how far are they from the random sequences? What are the nature of correlations? As we have noticed the sequences for the same species have small fluctuations. As we go from one species to another the fluctuations increase. The further apart the species are in the scale of evolution the larger are the fluctuations. An understanding of the fluctuations, as opposed to order, is important for evolutions. What gives rise to these fluctuations? Are they purely random, or is there a method to this madness? Clearly, any arbitrary fluctuation does not lead to a viable new organism, but some do.

3. Spectral decomposition, algorithmic complexity, entropy and order

"At the end of his life, John von Neumann challenged mathematicians to find an abstract mathematical theory for the origin and evolution of life. This fundamental problem, like most fundamental problems, is magnificently difficult.

Perhaps algorithmic information theory can help to suggest a way to proceed".

–Gregory J Chaitin

Given the nontemplate sequence in the 5' to 3' direction how does its Fourier transform (FT), or more precisely Discrete Fourier Transform (DFT), look like? What do we get from the Fourier spectra? Before we get to answer some of these questions let us lay out how we arrive at the Fourier spectra of symbolic sequences composed of symbols A, T, G and C.

3.1. From symbols to numbers :

The symbolic DNA sequences made of the nucleotide bases first need to be converted to numbers. Consider a sequence of four symbols, such as :

S (A, C, G, T) = S = GTGCACTCCCA The sequence has the length 11, *i.e.*, it has 11 symbols in all. It is made up of four subsequences : (i) the G sequence : G 0 G 0 0 0 0 0 0 0 0 (ii) the T sequence : 0 T 0 0 0 0 T 0 0 0 0 (iii) the C sequence : 0 0 0 C 0 C 0 C C C 0 (iv) the A sequence : 0 0 0 0 A 0 0 0 0 0 A

The G sequence, denoted S_G is thus

$$S_G = 1\,0\,1\,0\,0\,0\,0\,0\,0\,0\,0\,;$$

the T sequence, S_T, is

$$S_T = 0\,1\,0\,0\,0\,0\,1\,0\,0\,0\,0\,; \text{ similarly}$$

$$S_C = 0\,0\,0\,1\,0\,1\,0\,1\,1\,1\,0$$

$$S_A = 0\,0\,0\,0\,1\,0\,0\,0\,0\,0\,1.$$

The symbols, now, have been changed to numbers.

3.2. Fourier transform :

The Fourier transform (FT) method, in many cases, is basically an efficient computational tool for performing some common manipulations of data. For some other problems, FT or the related 'power spectrum' is itself of intrinsic importance. With the help of FT, a periodic function $f(x)$ of period 2π can be expanded (probably in an infinite series) in terms of sin kx and cos kx, where $k = 0, 1, 2, \ldots$. In essence, the FT separates a function into sinusoids of different frequency which sum to the original function. It distinguishes the different frequency sinusoids and their respective amplitudes [32–34].

The discrete Fourier transform (DFT) [35] is a modification of FT. DFT of any sequence is practically not a continuous function, but a sequence itself that coresponds to equispaced samples in frequency of FT of the signal. A digital computer works only with discrete data; so numerical computation of the FT requires sample values, and DFT significantly helps in implementing effective algorithms for the computation.

The fast Fourier transform (FFT) is a modified DFT algorithm that, though implemented independently by a number of workers over the last 30 years, became generally known from the work of Tukey and Cooley in 1965 [36] FFT reduces the number of computations from something on the order of M^2 to $M \log M$, M being the length of the sequence given.

3.3. Fourier transform of S :

The Fourier transform of S, given in (2), made of four symbols, is taken to mean the transform of the four subsequences S_G, S_T, S_C, and S_A. From these four separate transforms the power spectrum of S is constructed.

It is to be noticed that the assignment of numbers to sequence, such as S, is not unique. It depends on convenience. For instance one could assign +1 for purines and –1 for pyrimidines in S and construct the Fourier transform. The assignment of number depends on the feature of the sequence being studied. In our assignment we have assumed all the bases A, T, G and C to be independent, without a priori correlations.

Define the quantity $S_{m,\alpha}$, where the subscript m refers to the position along the sequence S, and α takes the values G, T, C, A

$$S_{m,\alpha} = 1, \text{ if the } \alpha \text{ symbol occupies position } m$$
$$= 0, \text{ otherwise.} \tag{1}$$

The DFT of the subsequences are defined as :

$$S_\alpha(q_n) = \frac{1}{\sqrt{M}} \sum_{m=1}^{M} S_{m,\alpha} \exp\,(-iq_n \cdot m), \tag{2}$$

where $S_\alpha\,(q_n)$ is called the DFT of $S_{m,\alpha}$, M is the total length of the sequence S measured in number of bases, q_n are related to the frequencies as discussed below. The q_n take the values determined from periodic boundary conditions.

3.4. Periodic boundary conditions :

Periodic boundary condition (PBC) means that the original series S, of (2), is extended with the condition :

$$S_{m+M,\alpha} = S_{m,\alpha}\,. \tag{3}$$

Imposing this extension on (2) determines the possible values of q_n as follows :

$$S_\alpha(q_n) = \frac{1}{\sqrt{M}} \sum^{M} S_{m+M,\alpha}\, e^{-iq_n(m+M)}. \tag{4}$$

This implies :

$$e^{-iq_n M} = 1\,. \tag{5}$$

Thus, $$q_n = \frac{2\pi}{M} \cdot n\,, \tag{6}$$

where n takes integer values 0 to M–1 in steps of 1.

It is to be remembered then that DFT of $S_{m,\alpha}$, defined in (2), with the choice of q_n, (6), implies the periodic boundary condition (3).

3.5. *The inverse transform :*

The eq. (2) gives

$$S_\alpha(q_n) = \frac{1}{\sqrt{M}}\sum_{m=1}^{M} S_{m,\alpha}\exp(-iq_n.m).$$

Now,

$$\sum_{n=0}^{M-1} S_\alpha(q_n)\exp(iq_{n'}m) = \frac{1}{\sqrt{M}}\sum_{m=1}^{M}\sum_{n=0}^{M-1} S_{m,\alpha} \times \exp\left(i(q_{n'} - q_n)m\right)$$

$$= \sqrt{M}\sum_{m=1}^{M} S_{m,\alpha}\frac{1}{M}\sum_{n=0}^{M-1}\exp\left(i(q_{n'} - q_n)m\right). \quad (7)$$

It is known that

$$\frac{1}{M}\sum_{m=1}^{M}\exp\left(i(q_{n'} - q_n)m\right) = 1, \text{ if } m = 0; \quad = 0, \text{ if } m \neq 0. \quad (8)$$

Likewise

$$\frac{1}{M}\sum_{m=1}^{M}\exp\left(i(q_{n'} - q_n)m\right) = 1, \text{ if } n = n'; \quad = 0, \text{ if } n \neq n'. \quad (9)$$

Then, considering $n = n'$, (7) becomes

$$\sum_{n=0}^{M-1} S_\alpha(q_n)\exp(iq_n m) = \sqrt{M}.S_{m,\alpha}. \quad (10)$$

So

$$S_{m,\alpha} = \frac{1}{\sqrt{M}}\sum_{n=0}^{M-1} S_\alpha(q_n)\exp(iq_n m). \quad (11)$$

This gives us the inverse or reciprocal transformation of (2).

3.6. *The reality of $S_{m,\alpha}$:*

The sequence $S_{m,\alpha}$, (11), consists of elements that are real. Thus,

$$S_{m,\alpha} = \frac{1}{\sqrt{M}}\sum_{n} S_\alpha(q_n)e^{iq_n m} = S^*_{m,\alpha}$$

$$= \frac{1}{\sqrt{M}}\sum_{n} S^*_\alpha(q_n)e^{-iq_n m}. \quad (12)$$

This is ensured if

$$S^*(q_n) = S_\alpha(2\pi - q_n). \quad (13)$$

The DFT spectrum has this symmetric form following from the reality of $S_{m,\alpha}$.

3.7. $S_{m,\alpha}(0)$:

When n is set at zero $q_n = 0$, the series (2) gives :

$$S_\alpha(0) = \frac{N_\alpha}{\sqrt{M}}. \quad (14)$$

Thus, $S_\alpha(0)$ is just a measure of the number, N_α, of symbols of type α in the sequence of M bases.

3.8. *Excluded volume effect :*

For the sequence $S_{m,\alpha}$, (2), each position has an occupant, A or C or G or T. No point of $S_{m,\alpha}$ is empty. Thus :

$$\sum_{\alpha} S_{m,\alpha} = 1 \text{ for any } m. \quad (15)$$

In terms of the DFT $S_\alpha(q_n)$, this translates into

$$\sum S_\alpha(q_n) = 0 \ : \text{for } n \neq 0. \quad (16)$$

3.9. *Frequencies and periodicities :*

The periodic boundary condition gives (6), *i.e.*

$$q_n = \frac{2\pi}{M}n,$$

where n takes values from 0 to $(M - 1)$. The frequencies, f, defined from $q_n = 2\pi f$, gives

$$f = \frac{n}{M}. \quad (17)$$

The periodicity is the inverse of frequency and is given by $\frac{1}{f}$.

3.10. *Correlations :*

The correlations are usually defined with periodic boundary conditions (PBC), sometimes they are, therefore, called the circular correlations.

For the sequence $S_{m,\alpha}$, eq. (2), the correlations $K_{\alpha\beta}$ sometimes also called auto-correlations, are defined as :

$$K_{\alpha\beta}(l) = \frac{1}{M}\sum_{m=1}^{M} S_{m,\alpha}S_{m+l,\beta}, \quad (18)$$

where, as usual, the $S_{m,\alpha}$ satisfy (3), the PBC. The PBC implies :

$$K_{\alpha\beta}(l) = K_{\alpha\beta}(M + l) \quad (19)$$

3.11. *The structure factor :*

The structure factors, $F_{\alpha\beta}$, of the sequence $S_{m,\alpha}$, eq. (2), are defined as [37] :

$$F_{\alpha\beta}(q_n) = S_\alpha(q_n)S^*_\beta(q_n) \quad (20)$$

These quantities, because of the symmetry of $S_\alpha(q_n)$ about the point $q_n = \pi$ (following from the reality condition of $S_{m,\alpha}$, (13)) are also symmetric about the point $q_n = \pi$.

Sum-rules

The structure factors, defined in (20), satisfy a set of sum-rules. These sum-rules, derivable from the definitions, are

$$(i) \ \sum_{n=0}^{M-1} F_{\alpha\beta}(q_n) = \sum_{m=1}^{M} S_{m,\alpha}S_{m,\beta} = \delta_{\alpha\beta}N_\alpha \quad (21)$$

where N_α is the number of bases of types α.

$$(ii)\ \bar{F}_{\alpha\beta} = \frac{1}{(M-1)}\sum_{n=1}^{M-1} F_{\alpha\beta}(q_n)$$

$$= \frac{1}{(M-1)}\left[\delta_{\alpha\beta} - \frac{N_\alpha N_\beta}{M}\right], \qquad (22)$$

where it is to be noticed, the sum on the left-hand-side does not include the zeroth harmonics.

3.12. The Wiener-Khinchin relation :

The structure factors, $F_{\alpha\beta}(q_n)$, are related to the correlations $K_{\alpha\beta}$ through the Wiener-Khinchin relation

$$K_{\alpha\beta}(l) = \frac{1}{M}\sum_{n=0}^{M-1} F_{\alpha\beta}(q_n)e^{-iq_n l}. \qquad (23)$$

We conclude, therefore, that given the Fourier coefficients of $S_{m,\alpha}$ we can calculate *via* the structure factors, all the autocorrelations that exist in the sequence.

3.13. The power spectrum :

As $S_{m,\alpha}$ are all real,

$$S_{m,\alpha} = S^*_{m,\alpha}. \qquad (24)$$

Now from (11),

$$S_{m,\alpha} = \frac{1}{\sqrt{M}}\sum_{n=0}^{M-1} S_\alpha(q_n)\exp(iq_n m).$$

So,

$$S^*_{m,\alpha} = \frac{1}{\sqrt{M}}\sum_{n=0}^{M-1} S^*_\alpha(q_n)\exp(-iq_n m) \qquad (25)$$

Since $q_n = 2\pi n/M$, [where $2\pi/M$ is the fundamental of period and n varies from 0 to $(M-1)$. Let $q_n = 2\pi - q_{n'}$.

So, (25) becomes

$$S^*_{m,\alpha} = \frac{1}{\sqrt{M}}\sum_{n=0}^{M-1} S^*_\alpha(2\pi - q_{n'})\exp(iq_{n'}m)\exp(2\pi im)$$

$$= \frac{1}{\sqrt{M}}\sum_{n=0}^{M-1} S^*_\alpha(2\pi - q_{n'})\exp(iq_{n'}m)$$

[since, $\exp(2\pi im) = 1$].

Thus, it can be written as

$$S^*_{m,\alpha} = \frac{1}{\sqrt{M}}\sum_{n=0}^{M-1} S^*_\alpha(2\pi - q_n)\exp(iq_n m). \qquad (26)$$

So, from (25) and (26)

$$S_\alpha(q_n) = S^*_\alpha(2\pi - q_n), \qquad (27)$$

$$S^*_\alpha(q_n) = S_\alpha(2\pi - q_n). \qquad (28)$$

Similarly,

$$S_\beta(q_n) = S^*_\beta(2\pi - q_n), \qquad (29)$$

$$S^*_\beta(q_n) = S_\beta(2\pi - q_n). \qquad (30)$$

Again from (20)

$$F_{\alpha\beta}(q_n) = S_\alpha(q_n)S^*_\beta(q_n). \qquad (31)$$

So, using (28) and (29), (31) becomes :

$$F_{\alpha\beta}(2\pi - q_n) = S_\alpha(2\pi - q_n)S^*_\beta(2\pi - q_n) \qquad (32)$$

and, on the other hand, using (27) and (30), (31) becomes :

$$F_{\beta\alpha}(2\pi - q_n) = S_\beta(2\pi - q_n)S^*_\alpha(2\pi - q_n). \qquad (33)$$

Using (32) and (33), the equation (20) becomes

$$F_{\alpha\beta}(2\pi - q_n) = S^*_\beta(q_n)S_\alpha(q_n). \qquad (34)$$

Now, comparing (20) (34),

$$F_{\alpha\beta}(q_n) = F_{\beta\alpha}(2\pi - q_n). \qquad (35)$$

And for the diagonal structure factor, *i.e.*, power spectrum

$$F_{\alpha\alpha}(q_n) = F_{\alpha\alpha}(2\pi - q_n). \qquad (36)$$

This is a symmetric function of q_n with the centre of symmetry at $q_n = \pi$. Therefore, if q_n is plotted as a function of q_n, only half of the $S(q_n)$ values are independent and the rest is just the mirror image of it.

As α stands for any of the four nucleotides (A, C, T, G) in DNA, the power spectrum, also known as spectral density, of a DNA sequency can be calculated by summing over the four possible values for A, C, T and G as follows :

$$F(q_n) = \sum_{\alpha=A,C,T,G} F_{\alpha\alpha}(q_n \qquad (37)$$

The power spectrum $F(q_n)$ is sometimes denoted by $S(f)$, where f is the frequency.

3.14. Ramdomness, algorithmic complexity, information entropy and order :

The concept of randomness and order in sequences are inversely related. The random sequence does not have any order. In the ordered case the knowledge of some of the entries can determine acurately what the others are. This is not possible for the random case. Most sequences that one finds are somewhere in between. They may have a certain degree of randomness, or order. The question we address now is : how to characterise this amount of randomness [38]?

To begin let us think of a symbolic sequence of two monomers A and B. How do we test for the randomness? The first step would be to carry out the frequency test. If the sequence is random, the proportion of A and B must be equal. Thus if $N_A(N_B)$ is the number of $A(B)$ in the sequence of N entries

$$\frac{N_A}{N} = \frac{N_B}{N}, \qquad (38)$$

for the sequence to be random.

Even though this condition must be satisfied, *i.e.*, it is necessary, it does not assure the randomness of the sequence. The series : ABABABABAB...does satisfy (30), but is not random.

The step next is to carry out the frequency test of words of length two *i.e.*, in this case : AA, AB, BA and BB. These four must appear in equal proportion. That is

$$\frac{N_{AA}}{N} = \frac{N_{AB}}{N} = \frac{N_{BA}}{N} : : \frac{N_{BB}}{N} \tag{39}$$

t be satisfied. It is, once again, necessary that the random sequence must have (39), but (39) does not assure of

The step next is to form words of length : three. In this case there are $2^3 = 8$ possibilities, they must all appear in l proportion : the word of length four, $2^4 = 16$ in number, be in equal proportion. The words of length n, 2^n in number, must all be in equal proportions.

Curiously, even if all these frequency tests are carried out, and all are satisfied, the series could still not be random. The well known counter example in the Champ's series (named after David Champernowne who first found it out), which consist of ten monomers 0,1,2,3,4,5,6,7,8,9 in the sequence 1 2 3 4 5 6 7 8 9 10 11 12 13 14 15 16 ... In view, therefore, randomness is difficult to define and test.

An alternate approch towards defining randomness and (complexity) of sequence came from the theory of information and algorithms, sometimes called the algorithmic information theory. The definition of the randomness came in the sixties by Kolmogorov [39], and Chaitin [40]. To generate the random sequence, according to the hypothesis, requires algorithm (*i.e.*, program) that is as large as the sequence itself. Stated differently, random sequences are incompressible in the algorithmic sense.

While the definition of the random sequence from the theory of algorithms is not particularly useful for our purpose here, we show now that the theory of information, the concepts of information entropy developed by Shannon, does provide some measure of order and disorder in sequences.

Before we get there, let us discuss the measure of randomness and complexity of sequences following algorithmic information theory. The degree of randomness, or complexity, of the sequence may be measured in terms of the length of the minimal algorithm needed to generate the sequence. For the completely random sequence, the length of the minimal program equals the length of the sequence. On the other extreme are ordered 'or less complex sequences that may be generated by few bits of algorithm.

Given general sequences of N bases, they may be categorised in terms of their complexity. Thus one of these may be complexity N-2, another of N-100, and so on. The exact value of the complexity below which the series is no longer random remains arbitrary. This uncertainty in quantifing randomness (or complexity) implies that the complexity of a sequence is roughly equal to the size of the minimal algorithm. Consider all the series of size N. We can plot number of sequences of complexity m aganist n [n taking the maximum value of N]. It is clear that the number of ordered sequences are few, the majority of them are fairly random.

3.15. *Information entropy :*

If we toss a coin, there are two possible outcomes. When we throw a die, the number of possible outcomes is six. This number of possibilities, *i.e.*, the available number of states, is related to the Shannon information [41].

If R_1 denotes the number of outcomes, the Shannon information entropy (SIE), denoted by I, is defined to satisfy ·

I be additive for independent events. Thus, if we have two independent events with R_1 and R_2 possible states, the total number of outcomes $R = R_1R_2$. The constraint of additivity of I requires

$$I(R_1R_2) = I(R_1) + I(R_2). \tag{40}$$

Thus, $I(R) = K \ln(R)$. Where K is the normalisation factor that we can fix conveniently.

Consider a binary sequence of 0 and 1 of N symbols. Let us say N_0 of them are 0; N_1 and 1. In this case, the number of possible outcomes R, of a series of N_0 zeroes and N_1 ones clearly is

$$R = \frac{N!}{N_0!N_1!}. \tag{41}$$

R denotes the number of independent messages that can be sent using N_0 zeroes and N_1 ones. The SIE, denoted I is just the log of the available states. Thus,

$$I = K \ln R = K[\ln N! - \ln N_0! - \ln N_1!]. \tag{42}$$

If we assume that N, N_0, and N_1 are large numbers, the Stirling's approximation to the log of factorials of large numbers may be used, *i.e.*

$$\ln N! = N(\ln N - 1). \tag{43}$$

Thus, The SIE, in this approximation is

$$I = K[N(\ln N - 1) - N_0(\ln N_0 - 1) - N_1(\ln N_1 - 1)], \tag{44}$$

with $N_0 + N_1 = N$.

It is covenient to define the average SIE as I/N, which, in this Stirling approximation, becomes :

$$i = \frac{I}{N} = -K\left[\frac{N_0}{N}\ln\frac{N_0}{N} + \frac{N_1}{N}\ln\frac{N_1}{N}\right]. \tag{45}$$

The quantities $\frac{N_0}{N}$ and $\frac{N_1}{N}$ are the frequencies, [see (38)], or proportions of zero and one. If we denote these proportions by p_0 and p_1, we get

$$i = -K\sum_j p_j \ln p_j, \tag{46}$$

where j takes values between 0 and 1.

3.16. Determination of K :

For the general binary series of N bits, the possible outcomes R is : $R = 2^N$. If the SIE for the case is normalised to N, we get

$$K \ln 2^N = N. \tag{47}$$

Thus, $K = \log_2 R$ and $I = \log_2 R$ (48)

3.17. Shannon information entropy tends towards extremum :

The SIE, we have seen, depends on the frequencies p_j, (46). Changes in p_j lead to changes in SIE. SIE is related to the number of available possibilities (or States). It has been proposed that this number tends to be a maximum. Thus :

$$-\sum p_j \ln p_j \tag{49}$$

is an extremum, subject to the constraint

$$\sum p_j = 1. \tag{50}$$

3.18. Shannon information entropy and order :

The SIE is a function of the frequencies or the properties of the various bases. For a purely random sequence all the frequencies are equal, and the SIE becomes one. While for an ordered sequence the SIE tends to zero. The SIE of sequences range between these two extremes of zero and one, and provide a measure of the randomness, complexity and order in the sequence.

3.19. Spectral analysis of complexity, short and long range order :

We have seen that the Shannon Information Entropy, SIE, denoted by the symbol I, gives a measure of the degree of complexity of the sequence. It turns out that a refined version of this measure, sometimes called the metric entropy, is given by

$$I = \lim_{n \to \text{large}} \left\{ -\frac{1}{n} \sum p_j(n) \log p_j(n) \right\}, \tag{51}$$

where $p_j(n)$ refers to words of length n and the subscript j goes over the number of such words. For the binary sequence clearly j takes 2^n values.

While this measure of complexity of the sequence has some mathematical sense, in practice the $n \to$ large limit makes this definition difficult to implement. It requires measuring frequencies of large word lengths. This is usually possible for sequences where the algorithm for its generation is known, such as the Thue-Morse sequence. If on the other hand, n is kept small, we arrive only at short range correlations of the monomers. The long range order has to be separately analysed.

3.20. Spectral measure of complexity and order :

The parallel approach to complexity or order comes from spectral analysis. This measure, sometimes called the structural entropy of the sequence, is given by

$$I_a = \sum_{n=1}^{M-1} \ln F_{\alpha\alpha}(q_n), \tag{52}$$

where $F_{\alpha\alpha}$, (20), are the diagonal elements of the structure factors of the sequence. The structure factors do satisfy some constraints, namely the sum-rules (21) and (22).

The structural entropy, under condition (21,22), is extremum. This extremisation leads to the solution $F_{\alpha\alpha}(q_n) =< F_{\alpha\alpha} >$ [Note $< F_{\alpha\alpha} >\equiv \bar{F}_{\alpha\alpha}$], for all q_n. For the random sequence we expect no peak and troughs in the spectra, *i.e.*, all the Fourier harmonics are of equal strength. Thus, the presence of the sharp peaks or troughs in the spectra denotes deviation from randomness and they are, therefore, ordered sequences.

The logarithmic dependence of I_α on $F_{\alpha\alpha}(q_n)$ makes it slow and insensitive. A more practical measure is

$$I_\alpha = -\sum^{M-1} \frac{F_{\alpha\alpha}(q_n)}{< F_{\alpha\alpha} >} . \ln \frac{F_{\alpha\alpha}(q_n)}{< F_{\alpha\alpha} >}. \tag{53}$$

For the complete sequence the structural entropy is

$$I = \sum I_\alpha . \tag{54}$$

While the short range order generally leads to sharp peaks of $F_{\alpha\alpha}$ and, therefore, can be read off from the power spectrum (37), or the strufctural entropy I (53), the long range order requires careful analysis.

3.21. The smoothed Fourier spectra and the long-range order :

The short range order leads to peaks in the power spectrum, or the structural entropy, and are easy to identify; for the long-ranged order special methods are needed. One such technique, called the method of normalised sweep, often referred to as Hurst's method, is now discussed [42].

At site m of sequence the smoothed out S_α [see (2)] is defined as follows :

$$\bar{S}_\alpha = \frac{1}{m_0} \sum^{n+m_0-} S_{\alpha,m'} \tag{55}$$

where m_0 clearly is the window over which the average is being defined. The deviation from this average δ is

$$\delta(m,\bar{m}) = \sum^{m} (S_{\alpha,m'} - \bar{S}_\alpha), \tag{56}$$

where $\bar{m}$ lies between m and $m + m_0 + 1$. The difference between the maximum of $\delta(\delta_{\max})$ and the minimum of δ $(\delta_{\min})$ determine the sweep W :

$$W_\alpha(m, m + m_0 - 1) = \delta_{\max}(m,\bar{m}) - \delta_{\min}(m,\bar{m}), \tag{57}$$

the standard deviation

$$\delta(S_\alpha) = {\frac{1}{m} \sum^{m+m_0-1} (S_{\alpha,m'} - \bar{S}_\alpha)^2}^{1/2} \tag{58}$$

If we define the normalized quantity W_α(norm) as

$$W_\alpha \; \cdot \frac{W_\alpha}{\delta}, \tag{59}$$

we can determine an average $<W_\alpha$(norm)> by varying the chosen site m over the whole sequence. This same quantity for the random sequence (of identical base composition) is denoted by

$$\langle W_\alpha(\text{random, norm})\rangle.$$

The difference between the logarithmic derivative, with respect to the window size m_0, of $<W_\alpha$ (norm) > and $<W_\alpha$ (random, norm) > determines the long-range order in the following way. Let

$$\begin{aligned}\frac{d\ln < W_\alpha(\text{norm}) >}{d\ln m_0} = \frac{d\ln < W_\alpha(\text{random, norm}) >}{dm_0} \\ = H_\alpha(\text{seq.}\, m_0) - H_\alpha(\text{random}, m_0) \\ = \Delta H_\alpha(m_0).\end{aligned} \tag{60}$$

If $$H_\alpha(\text{seq.}\, m_0) > H_\alpha(\text{random}, m_0), \tag{61}$$

the correlations are called persistent. They are antipersistent if the reverse, namely

$$H_\alpha(\text{seq.}\, m_0) < H_\alpha(\text{random}, m_0) \tag{62}$$

is true.

The $\Delta H_\alpha(m_0)$, [60] tends to zero above some m_0, called m_{corr}, if the correlations in the sequence are of the short range. If $\Delta H_\alpha(m_0)$ is not zero as m_0 is increased, we have long-range order.

4. Random walks, Fickian and fractional Brownian diffusion

"The phenomenon of Brownian motion has been known since the time that van Leeuwenhock first peered through a microscope. Although it must have been regarded as a nuisance by early microscopists, Brownian motion has played a significant role in the development of our understanding of the physical world".

–George H Weiss

Deviations from randomness bring order. Randomness leads to the characteristic distribution of the monomers (for the DNA sequences the monomers are called the bases; for the protein sequences the amino acid monomers are referred to as the residues) over the polymer chain. The measurement of the distribution of the bases characterise order and fluctuations in the sequences. The random walk approach to sequences is to study and model the distribution of the bases [43, 44].

Diffusion of the particles in a medium was reported by the Dutch physician van Leeuwenhock, and was subsequently rediscovered by Robert Brown. This diffusion process of particles is known as Brownian motion. It turns out that the normal diffusion process follow the distribution functions of the random walker, *i.e.*, Gaussian. Hence, all the moments of the distribution are finite. In contrast, there are other diffusive processes, many encountered in Biophysical systems, where the moments are not finite, these processes are, therefore, anomalous.

The random walk, underneath its randomness, hides some subtle regularities. The reason for the regularities may be ascribed to its fractal nature. For instance, the mean square displacement follow a fairly regular pattern. Random walk is a statistical fractal *i.e.*, it is generated by an algorithm that has a stochastic element in it. The regularities may be ascribed to the fractal structure. The Gaussian distribution, so characteristic of random walk, is scale invariant. This lack of the scale point to its fractal nature The regularities, such as the mean square displacement as a function of the number of steps, are ascribed to fractal correlations. The deviations, for sequences, from these regular relations are important charactenstics.

4.1. Random walks :

The mathematical theory of the random walk is based on the following simple steps. We illustrate these steps for the one dimensional walker that can move right or left with known probabilities. The steps are

(a) Find the probabilities the single step to right, and the single step to left.

(b) Fourier transform the probabilities to determine the characteristic functions of the random walker.

(c) For an arbitrary number n of steps, obtain the characteristic function by raising the single step characteristic function to the power n.

(d) All the mathematical properties, namely, moments, the distributions *etc.* are derivable from the n-step characteristic function.

Before we illustrate these mathematical steps in detail let us, for motivation, work on the one dimensional random walker.

For generality consider the random walker on the one dimensional lattice; p the probability of the step to the right, q the probability of the step to the left (Figure 12). Clearly

$$p + q = 1. \tag{63}$$

If we consider the function [45]

$$pe^{ik} + qe^{-ik}, \tag{64}$$

clearly, it corresponds to the Fourier transform of the single step probability, namely,

$$f(x) = p\delta(x-1) + q\delta(x+1). \tag{65}$$

Thus, [64] is the single step characteristic function. It is easy to verify that the probability, after 2 steps, is given by

$$\left(pe^{ik} + qe^{-ik}\right)^2 = p^2e^{2ik} + q^2e^{-2ik} + 2pq \tag{66}$$

i.e., the coefficient of $e^{\pm 2ik}$ is the probability the walker is at site ±2; the $2pq$ gives the probability the walker back to the starting point, zero, after two steps. Generalizing to n steps, the characteristic function becomes

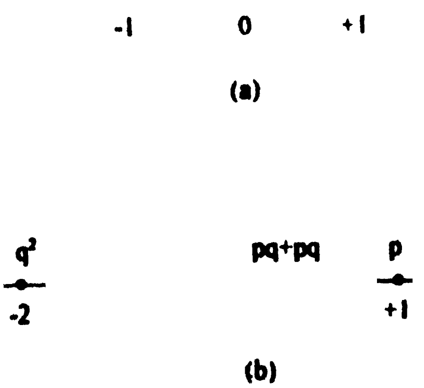


Figure 12. Demonstration of a one-dimensional random walk. (a) If a random walker starts at 0 site, after one step he will be either at +1 or at -1. (b) After one step, the probability that he ends at site +1 is p and that at site -1 is q. After 2 steps, he might end at any of the sites +2, 0, -2; therefore, the probability that he ends at +2 is p^2, that at 0 is 2pq, and that at -2 is q^2.

$$\left(pe^{ik}+qe^{-ik}\right)^n. \tag{67}$$

Thus, the probability the walker is at site $+m(n \geq m \geq 0)$ is just the cofficient of e^{imk} in the above characteristic function. Since we have the. completeness of the e^{ik} functions, namely,

$$\frac{1}{2\pi}\int_{-\pi}^{\pi} e^{ik(l-m)} = \delta(l-m), \tag{68}$$

we can obtain the probability $P_n(m)$, the random walker is at site m after n steps as

$$P_n(m)=\frac{1}{2\pi}\int_{-\pi}^{\pi}\left(pe^{ik}+qe^{-ik}\right)^n e^{imk}dk. \tag{69}$$

Evaluating the integral, using (68), for the case $p=q=\frac{1}{2}$, we get

$$_n(m)=\left(\frac{1}{2}\right)^n\cdot\frac{n!}{\left[\frac{1}{2}(n+m)!\right]\left[\frac{1}{2}(n-m)!\right]} \tag{70}$$

Note that when n is even (odd) m takes even (odd) values.

Large n :

When the number of steps n is large, *i.e.*, $n\rangle\rangle 1$, we use Stirling formula

$$n!=\sqrt{2\pi n}^{\,n+\frac{1}{2}}\exp(-n), \tag{71}$$

in order to simplify (70).

If we take log, Stirling formula takes the form

$$\ln(n)!=\frac{1}{2}\ln(2n\pi)+n.\ln(n)-n. \tag{72}$$

From (70), we get

$$\ln P_n(m)=-n.\ln(2)+\ln(n)! - \ln\frac{(n+m)}{2!}-\ln\frac{(n-m)}{2!}, \tag{73}$$

$$\begin{aligned}\ln P_n(m) &= -n.\ln(2)+\frac{1}{2}\ln(2\pi n)+n.\ln(n)\\ &\quad -n-\frac{1}{2}\ln\left[2\pi\frac{(n+m)}{2}\right]\\ &\quad -\frac{(n+m)}{2}\ln\frac{(n+m)}{2}+\frac{(n+m)}{2}\\ &\quad -\frac{1}{2}\ln\left[2\pi\frac{(n-m)}{2}\right]-\frac{(n-m)}{2}\\ &\quad \times\ln\frac{(n-m)}{2}+\frac{(n-m)}{2}.\end{aligned} \tag{74}$$

Simplifying (74), we get

$$\ln P_n(m)=\frac{1}{2}\ln\left(\frac{1}{2\pi n}\right)+\frac{1}{2}\left(\frac{m^2}{n^2}-\frac{m^2}{n}\right); \tag{75}$$

And from (75), we get

$$P_n(m)=(2n\pi)^{-\frac{1}{2}}\exp\left(\frac{m^2}{2n^2}-\frac{m^2}{2n}\right). \tag{76}$$

Hence, as $n\rangle\rangle 1$, one obtains the gaussian form for $P_n(m)$ as

$$P_n(m)=(2n\pi)^{-\frac{1}{2}}\exp\ ^{-m^2} \tag{77}$$

4.2. *Continuum limit :*

If the lattice spacing is a and τ denotes time interval between the steps, then

$$x=ma \tag{78}$$

is the net displacement in time

$$t=n\tau. \tag{79}$$

The probability of displacement between x and $x+dx$ in time t denoted $P(x,t)$ and satisfies

$$P(x,t)dx=P_n(m)dm, \tag{80}$$

as $n\to\infty$, $\tau\to\infty$ and $a\to\infty$.

Thus, $$P(x,t)dx=\frac{1}{2\pi n}\exp\left(-\frac{m^2}{2n}\right).dm \tag{81}$$

Therefore, $$P(x,t)=(4\pi Dt)^{-\frac{1}{2}}\exp\left(-\frac{x^2}{4Dt}\right), \tag{82}$$

with $$D::\frac{^{\ldots}}{2\tau}.$$

The probability distribution function, $P(x,t)$, for the case $p=q=\frac{1}{2}$ in the continuum limit, is a Gaussian.

The generalisation to the case of the d-dimensional lattice with $2d$ possible step directions, to the nearest neighbour, is straightforward. The single step characteristic function is

$$P_1(k)=\sum_{i=1}\left(p_i e^{ik_i}+q_i e^{-ik_i}\right), \tag{83}$$

where $\sum_i (p_i+q_i)=1$

The n-step probability, as earlier, is

$$P_n(k)=\left[P_1(k)\right]^n. \tag{84}$$

The probability distribution function $P_n(x)$ in real space after n steps, may be obtained from the Fourier transform of $P_n(k)$.

In the continuum limit, for the case when p_i and q_i all are equal, we get the d-dimensional Gaussian function.

4.3. *The chain rule :*

It is to be noted that the probability distribution functions in real space satisfy the chain rule

$$P_{n+1}(m)=\sum_{m_0} P(m-m_0)P_n(m_0) \tag{85}$$

and the law of conservation of probability $\sum_m P_n(m)=1$.

In the Fourier transformed space, the characteristic functions satisfy the chain rule. Written out the chain rule takes the form

$$P(k,t)=P(k,t_1)P(k,t-t_1). \tag{86}$$

4.4. *The moments of the distributions :*

The moments denoted, μ_l, of the probability distributions are averages of powers of displacements from the starting point.

$$\mu_l=\langle x^l\rangle, \tag{87}$$

where the brackets arounds x^l mean the average value of. The average of a function $g(x)$ for the probability distribution $P(x)$ is given by

$$\langle g(x)\rangle=\int g(x)P(x)dx. \tag{88}$$

Thus, $$\mu_l=\langle x^l\rangle=\int x^l P(x)dx. \tag{89}$$

Since the characteristic function

$$P(k)=\int P(x)e^{ikx}dx=\int\sum\frac{i^l k^l x^l}{l!}P(x).dx$$
$$=\sum\frac{i^l k^l}{l!}\langle x^l\rangle, \tag{90}$$

thus, $$\mu_l=l!\,i^{-l}\ \text{[coefficient of } k^l \text{ in } P(k)]. \tag{91}$$

Expressed differently,

$$\mu_l=(-1)^l\frac{\partial^l P(k)}{\partial k^l}\Bigg|_{k\to 0} \tag{92}$$

For the Gaussian distribution, all moments are finite.

Clearly then the characteristic function of the single step defines all the important parameters of walk. Raised to power n, it yields the n-step characteristic function. The Fourier transform of the n-step characteristic function determines the spatial distributions. Derivatives of the n-step-characteristic function determine all the moments of the distributions.

4.5. *Generating function of random walk :*

The probability function $P_n(x)$, *i.e.*, the distribution after n steps, is related by the chain rule to the probability distribution function of a single step $P_1(x)$ as follows :

$$P_n(x)=\int P_1(x-x').P_{n-1}(x')dx'. \tag{93}$$

In terms of the characteristic function, we know that

$$P_n(k)=\left[P_1(k)\right]^n. \tag{94}$$

Thus, $$P_n(x)=\frac{1}{2\pi}\int\left[P_1(k)\right]^n e^{-ik.x}dk. \tag{95}$$

The random-walk generating function $G(x,z)$ determine many important properties of the process, and is defined as

$$G(x,z)=\sum_{n=0}^{\infty} z^n P_n(x). \tag{96}$$

Using (94), we have

$$G(x,z)=\frac{1}{2\pi}\int\frac{e^{-ik.x}}{1-zP_1(k)}dk. \tag{97}$$

Similar generating functions may be defined for walks on periodic lattice points are at

$$l=l_i a_i. \tag{98}$$

Periodicity of the lattice implies

$$(l_1,l_2,l_3,\ldots\ldots)=(l_1+N,l_2,\ldots\ldots)$$
$$=(l_1,l_2+N,\ldots\ldots)=\cdots, \tag{99}$$

where N is the period in each direction. The chain rule on the lattice means

$$P_{n+1}(l)=\sum_{l'}P_1(l-l')P_n(l'), \tag{100}$$

where P_n is the probability distribution function after n step on the lattice. On the lattice the characteristic function are defined with the periodic boundary conditions, *i.e.*, the k are restricted to $k=2\pi m/N$. Thus, in analogy with (95), we have

$$P_n(l)=\frac{1}{N^{1/2}}\sum\left[P_1(k)\right]^n\exp(-ik.l). \tag{101}$$

The lattice walk generating function is

$$G(l,z)=\sum P_n(l).z^n=\frac{1}{N^{1/2}}\sum\frac{e^{-ikl}}{1-zP_1(\bar{k})}.\quad(102)$$

4.6. The Central Limit theorem :

No matter what the moments of the distributions are, provided the first and the second moments are finite, these distributions, to the first approximation, are Gaussian asymptotically. This is the statement of the Central Limit theorem.

Proof : We know, from the definition of the characteristic function

$$\begin{aligned}P_1(k)&=\int e^{ikx}.P_1(x).dx\\&=\int(1+ikx-1/2.k^2x^2+\ldots)P_1(x).dx\\&=1+ik<x>-1/2.k^2<x^2>+\ldots,\end{aligned}\quad(103)$$

where the brackets <> denote the averages values.

Note that $P_1(0)=1$ and $P_1(k)\leq 1$. For other values of k, the Central Limit theorem approximates $P_1(k)$ keeping only the first two moments, which are assumed to be finite. The integral (103) is thus assumed to be dominated by the region of small k. In this approximation if $<x^2>$ is replaced by σ^2, the variance, defined as

$$\sigma^2=<x^2>-<x>^2,\quad(104)$$

then,
$$P_n(k)=\exp\left\{n\left(ik<x>-\frac{\sigma^2k^2}{2}\right.\right.\quad(105)$$

Taking the Fourier transform of (105), we get

$$P_n(x)=\frac{1}{\sigma\sqrt{2\pi n}}\exp\left.\frac{-(x-n<x>)^2}{2\sigma^2n}\right\}.\quad(106)$$

This is the lowest order approximation as per the Central Limit theorem.

4.7. General solution of the chain rule :

The Central Limit theorem assumes the existence of the first and the second moments of the probability distribution functions. Provided these moments are finite, the distributions in the asymptotic region become Gaussian.

Are there distribution functions that do not approach the Central Limit? The answer to this question came from the work of Paul Levy. Consider the characteristic function of the form

$$f_i(k)=\exp\{-b_i|k|^\alpha\},\quad(107)$$

where $0<\alpha\leq 2$.

When α becomes 2, $f_i(k)$ is the characteristic function of the Gaussian distribution. For $\alpha<2$, some of the moments, in particular the second moment, is divergent. If we look at the probability distribution function in real space, *i.e.*, the Fourier transform of $f_i(k)$, only for a few values of α the analytic form of the pdf exist. In general, the asymptotic structure, *i.e.*, the distribution at large x has the power law form

$$f(x)\sim|x|\rightarrow\frac{\alpha b}{\pi|x|^{\alpha+1}}\cdot\Gamma(\alpha)\sin\pi\alpha/2.\quad(108)$$

These distributions, with divergent second moments, do not clearly approch the Central Limit. The Fourier Transform exists in closed form only for a few special cases :

(1) For $\alpha=1$, we have the Cauchy distribution

$$f(x)=\frac{1}{\pi}\frac{1}{(x^2+b^2)}.\quad(109)$$

(2) $\alpha=2$ corresponds to the Gauss distribution and

(3) $\alpha=\frac{2}{3}$ leads to Zolotarev distribution, which has the form :

$$F(x)=\frac{1}{\sqrt{2\pi}}\frac{1}{x}W_{\frac{1}{2}\cdot\frac{1}{6}}\left(\frac{4}{27}\frac{b^{2/3}}{x^2}\right).\exp\left\{-\frac{2}{27}\frac{b^2}{x^2}\right.\quad(110)$$

where $W_{\frac{1}{2}\cdot\frac{1}{6}}(x)$ is a Whittaker function.

The distributions like (107) were first obtained by Cauchy. That for $\alpha>2$, they are not positive definite was not recognised by him. The constraint that the probability distribution be positive for all x keeps $\alpha\leq 2$.

4.8. Continuous time random walk (CTRW) :

If the time between sucessive steps are not fixed but vary with a certain probability density we have continuous time random walk. Mathematically, if T_i is the time of the i-th step, then [46]

$$t_i=T_{i+1}-T_i$$

is identically distributed independent random variable.

If the probability density for the time interval between sucessive steps is called $I_1(t)$, then $I_n(t)$ is the probability density for the time at which the n-th step is taken. Clearly, the chain rule yields :

$$I_n(t)=\int_0^t I_1(T).I_{n-1}(t-T)dT.\quad(111)$$

Since the above is weitten in terms of integral over time, the Laplace transformation are appropriate over the usual Fourier decomposition. The Laplace transforms of $I_n(t)$ is denoted $I_n(s)$ and it satisfies

$$I_n(s)=[I_1(s)]^n.\quad(112)$$

We want to calculate the probability density $P(x,t)$ that the walker is at x in time t. Let us denote by $J(t)$ the probability that the time between successive steps exceed or equal t, then

$$P(x,t)=\sum_{n=0}^{\infty}p_n(x)\int_0^t I_n(T)J(t-T)dT. \qquad (113)$$

The Laplace transform of $J(t)$ is :

$$J(s)=\int_0^{\infty}e^{-st}dt\int_t^{\infty}J(T).dT=\frac{1-I_1(s)}{} \qquad (114)$$

Thus, if $P(x,s)$ is the Laplace transform of $P(x,t)$, then

$$P(x,s)=\frac{1-I_1(s)}{s}\sum_{n=0}p_n(x).I_1^n(s). \qquad (115)$$

Comparing with (96), we find that the form is like that of the generating function. The z which did not appear to have any physical meaning in now related to the waiting time distribution. Thus,

$$P(x,s)=\frac{1-I_1(s)}{2\pi s}\int\frac{e^{-ikx}}{1-p(k)I_1(s)}, \qquad (116)$$

just as in (97).

We can use the expression above that does not depend on the structure of $P(x)$ but only on course properties such as the moments.

4.9. Diffusion-Fickian and fractional Brownian :

The Chain rule (85) is fundamental to the problem of random walk. Under certain assumpations, this rule

$$P_{n+1}(x)=\int_{-\infty}^{+\infty}p_n(x-y).p(y).dy \qquad (117)$$

may be written as a differential equation of the probability functions.

Suppose that the steps are taken at regular, small time intervals ΔT. The walker has taken a large number n of step so that

$$n\Delta T=t \qquad (118)$$

is finite. If we expand $P_{n+1}(x)$, we have

$$P_{n+1}(x)\approx P(x)+\Delta t.\frac{\partial p(x,t)}{\partial t}. \qquad (119)$$

If the jumps are in small steps, then

$$P_n(x-y)\approx P(x,t)-y\frac{\partial p(x,t)}{\partial x}+\frac{1}{2}y^2\frac{\partial^2 p(x,t)}{\partial x^2} \qquad (120)$$

Assume that the moments of $p(x)$ are of the following type

$$\frac{1}{\Delta T}\int_{-\infty}^{+\infty}xp(x).dx=v;\quad \frac{1}{2\Delta T}\int_{-\infty}^{+\infty}x^2p(x).dx=D; \qquad (121)$$

the other higher moments are negligibly small.

For the case where the probability of walk to the left and right are identical, the first moment, proportional to v, is also zero. Thus the Chain rule leads to the simple differential form for $p(x,t)$

$$\frac{\partial p}{\partial t}=D\frac{\partial^2 p}{\partial x^2}. \qquad (122)$$

If we compare with the usual Brownian diffusion, we recover an identical equation for the density distribution. The usual diffusion, (sometimes referred to as the normal diffussion, or Fickian diffusion) is governed by the Fick's law on the density distribution.

$$J=-D\nabla_p, \qquad (123)$$

supplemented by the equation of continuity

$$\frac{\partial p}{\partial t}+\nabla.J=0 \qquad (124)$$

where, the current density J is given by

$$J=p.v_d, \qquad (125)$$

v_d being the drift velocity.

Putting the above two eqns. together, we get

$$\frac{\partial p}{\partial t}=D\frac{\partial^2 p}{\partial x^2}. \qquad (126)$$

Thus, the density distribution function for normal diffusion has the same form as the probability distribution when the first moment vanish, the second moment is finite, the higher moments negligible. The solution of (126) gives usual Gaussian Function

$$P(x,t)=\frac{1}{(4\pi Dt)^{\frac{1}{2}}}\exp\left[-\frac{x^2}{4Dt}\right]. \qquad (127)$$

Comparing with (82), we have the usual random walk distribution excecuted by Brownian motion. In this case the second moments, as a function of the number, n, of steps. goes as

$$<x^2>\ \sim n. \qquad (128)$$

The normal Brownian motion is characterised by the above scaling behaviour of the 2nd moment. If we write

(129)

H, the Hurst index, is 1/2 for normal diffusion.

What happens if the second moment is not finite. The second moment is related to the correlation function K, as

$$<x^2(l)>\ \propto\int_0^l dl\int_0^{l'}dl''K(l''). \qquad (130)$$

Thus, the finiteness, or the divergence of the second moment is related to how the correlations behave as a function of distance. If the correlations do not fall rapidly, and the integral on the r.h.s. of (130) is not convergent, the second moment diverges. This brings us to the Levŷ type probability distribution. These are allowed solutions of the Chain rule. In this case, the $<x^2>$ goes as

$$<x^2>\ \sim n^{2H} \qquad (131)$$

where H differs from $\frac{1}{2}$. The case where the Hurst index differs from $\frac{1}{2}$ is called the Fractional Brownian Motion (FBM).

5. Measurements on the DNA : order, fluctuations and modelling

"The wonderful features which are constantly revealed in physiological investigations and differ so strikingly from what is known of inorganic matter, have led many biologists to doubt that a real understanding of the nature of life is possible on a purely physical basis... I think that we all agree with Newton that the real basis of science is the conviction that Nature under the same conditions will always exhibit the same regularities. Therefore, if we were able to push the analysis of the mechanisms of the living organisms as far as that of atomic phenomena, we should scarcely expect to find any features differing from the properties of inorganic matter".

-Niels Bohr

We have seen in Chapter 1 tha the DNA, the long chain of biopolymer, carries information from one generation to the text. These biopolymers synthesize proteins necessary for living. They are made of monomers denoted by A, T, G and C. If the sequence of the monomers in one of the strands of the polymers is known, the sequence in the other strand is obtained by replacing A by T, and G by C, or vice versa. The order of the sequence of these four monomers determine all the information there is in the DNA. The sequences, by

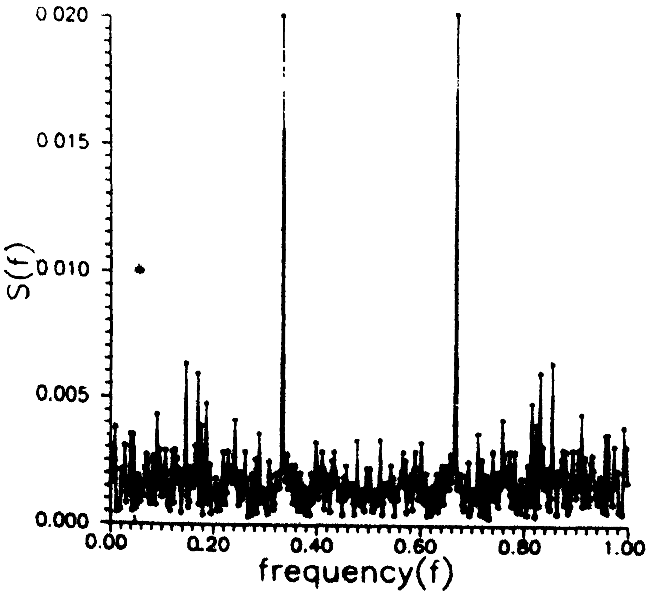


Figure 13(a). Frequency (f) of nucleotides is plotted against Power Spectrum, S(f). The peak at f = 1/3 appears to be the maximum one for the exons of different genes; while for the respective introns, no such peak at f = 1/3 is noted. This figure shows the exons of alpha-globin gene from goat (GenBank J00043).

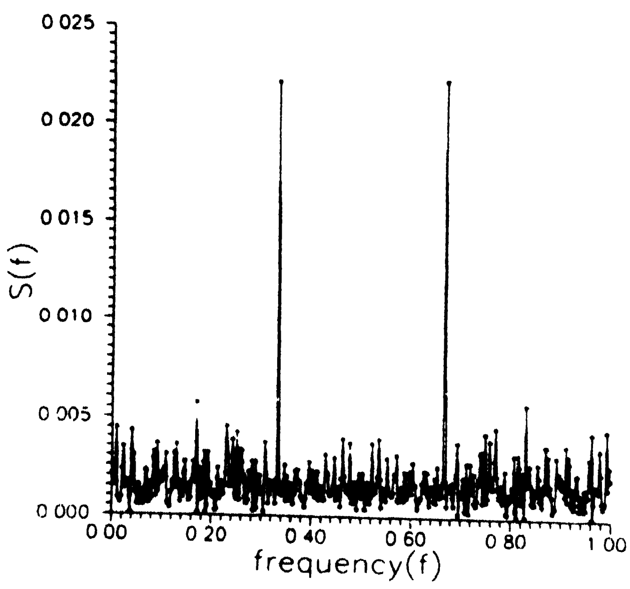


Figure 13(b). Frequency (f) of nucleotides is plotted against Power Spectrum, S(f). The peak at f = 1/3 appears to be the maximum one for the exons of different genes; while for the respective introns, no such peak at f = 1/3 is noted. This figure shows the exons of alpha-globin gene from Horse (GenBank M17902).

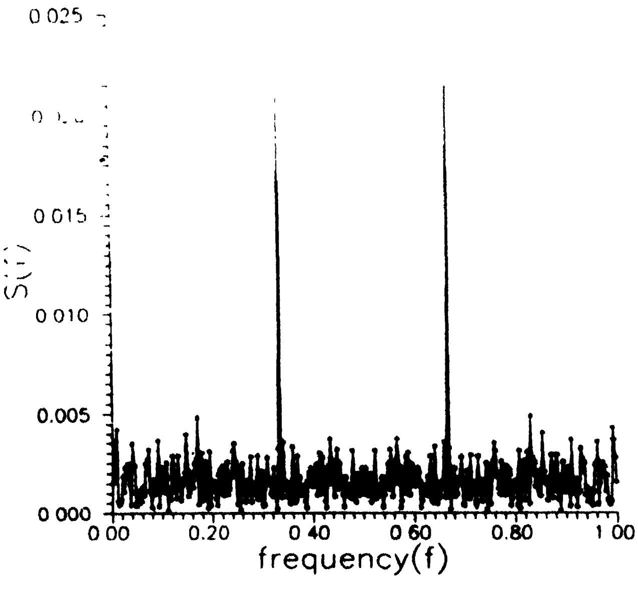


Figure 13(c). Frequency (f) of nucleotides is plotted against Power Spectrum, S(f). The peak at f = 1/3 appears to be the maximum one for the exons of different genes; while for the respective introns, no such peak at f = 1/3 is noted. This figure shows the exons of alpha-globin gene from Rhesus Monkey (GenBank J04495).

convention, are given for the non-template strand from the 5' to the 3' direction.

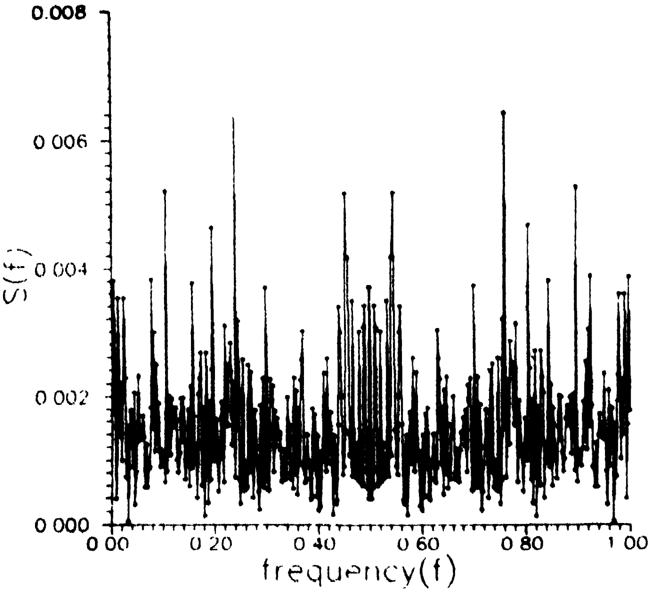


Figure 13(d). Frequency (f) of nucleotides is plotted against Power Spectrum, S(f). The peak at f = 1/3 appears to be the maximum one for the exons of different genes; while for the respective introns, no such peak at f = 1/3 is noted. This figure shows the introns of alpha-globin gene from Xenopus (GenBank X14260).

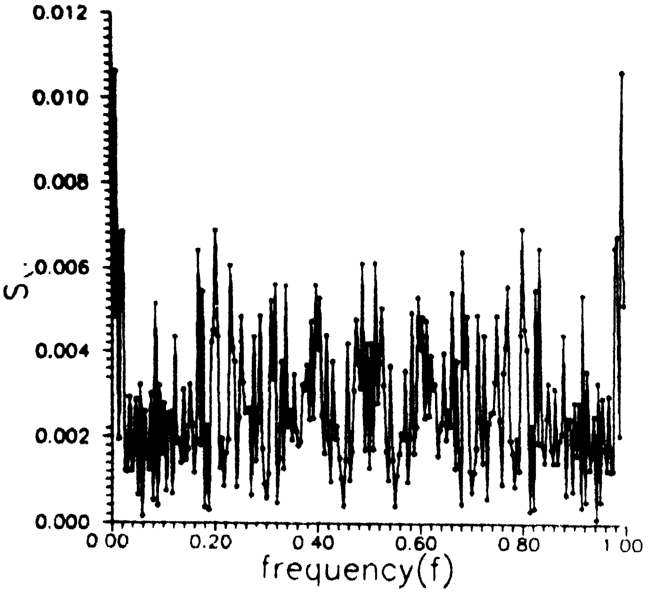


Figure 13(e). Frequency (f) of nucleotides is plotted against Power Spectrum, S(f). The peak at f = 1/3 appears to be the maximum one for the exons of different genes; while for the respective introns, no such peak at f = 1/3 is noted. This figure shows the introns of alpha-globin gene from Chicken (GenBank V00140).

The sequences may be roughly divided into three distinct parts. First there are the genes that code for proteins, and there is the intergenic DNA. Inside the genes the sequences divide into the exons and the introns. The exons are the ones coding for proteins, the introns come in between the exo ' regions For prokaryotic organisms (roughly the lower organisms) the DNA sequences, almost in its entirety, code for proteins. The eukaryotic genes, on the other hand, consist mostly of the intergenic regions and the introns. The protein coding parts, the exons, are few and far between.

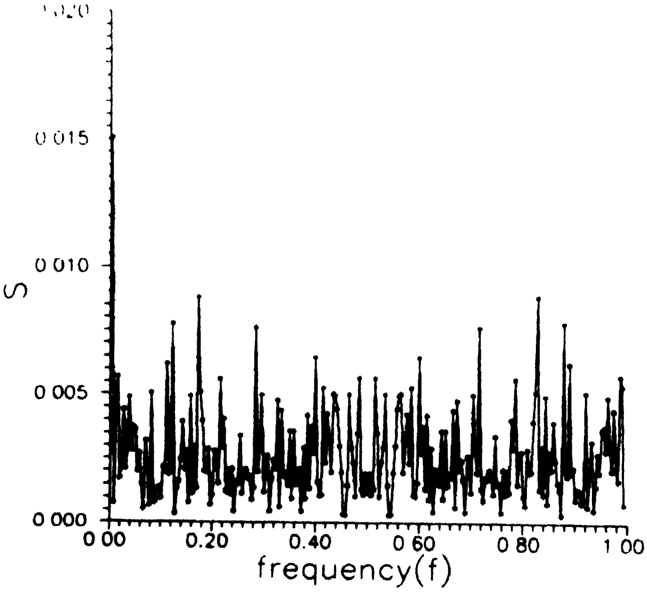


Figure 13(f). Frequency (f) of nucleotides is plotted against Power Spectrum, S(f) The peak at f = 1/3 appears to be the maximum one for the exons of different genes; while for the respective introns, no such peak at f = 1/3 is noted This figure shows the introns of alpha-globin gene from Orangutan (GenBank M12157).

5.1. Short range order— the peak at f = 1/3 :

Analysis of periodicities, or short-range correlations, are made *via* Fourier spectrum analysis (2). In the protein coding regions the triplet codons are arranged one after another leading to the important three periodicity [47]. In the power spectrum (37) of the sequence, we expect therefore to see peaks at 1/3 frequency. On the other hand for the introns and the intergenic regions no such peaks are expected. Figure 13 gives us the power spectrum of exons and introns. The $f = 1/3$ peak distinguishes the exon regions from the rest of the sequence. Identification of protein coding regions for long sequences is an important exercise for the DNA sequences. Clearly the Fourier spectrum, or more precisely the peak at $f = 1/3$, can distinguish between the coding and the non-coding regions [48]. Such a program has some difficulty in identifying the beginning or the end of the exon region. The beginning of the exon and the end of the exon may, however, be identified from the start and the stop codons. It is to be remembered that there are pseudogenes that have $f = 1/3$ peak, but do not code for proteins.

For the exon regions interspersed between introns the problem of identifying the beginning and the end remains. For these segments there are no start codons, nor the stop codons. The $f = 1/3$ peak analysis for these regions merely provide the rough location of the exon segment. These exon regions may be identified by chemical identification of the corresponding m-RNA (see Chapter 1).

5.2. *Other periodicities :*

Aside from the usual 3 period corresponding to the codons, there are other important periodicities reported for genomes of organisms. Some of these observations are summarized in Table 9 [49].

Table 9. The power spectrum, upon increased averaging, gives rise to distinguishing peaks at different periods in different categories of living organisms

Category	Peaks at Period ($\tau = 1/f$)		
	3	6	9
Primate	present	–	present
Rodent	present	-	--
Mammal	present	–	..
Vertebrate	present	–	present
Invertebrate	present	present	present
Plant	present	-	–
Bacteria	present		–
Virus	present	-	-
Organelle	present	–	-
Bacteriophage	present	-	–

5.3. *Repetitive segments :*

For higher eukaryotes the exons are a small part of the sequence. This coding region may be about 5% of the length. When we take account of the introns and other segments (such as the promoters, leaders, trailers and other regulatory sites) we are still left with about 80% of the sequence that remains unused. For lower eukaryotes the complete sequence, or most of it, are used.

It is now known that a good part of the DNA are made up of repetitive segments, *i.e.* segments, almost identical to one another, repeated many times over the sequence. These repetitive segments fall into two types :

(i) Highly repetitive DNA : These may be repeated several hundred to several million times in the sequence.

(ii) Moderately repetitive DNA : These are repeated upto several hundred times.

Indeed, the repetition frequency is almost continuous. Some of these repeated sequences do have functions. For the others no functions have been discovered so far. The Fourier spectrum can potentially identify the repetition periods and the frequencies.

To summarize, the DNA of higher eukaryotes contains subsequences that repeat as many as a million times in identical or very similar copies.

5.4. *The mosaic model :*

The mosaic model of genetic structure says that the sequences consist of more of less independent units stretching roughly over thousand bases. The protein coding sequences have lengths of that order. The model simply says that the composition of the mosaic units can vary from one to another giving rise to a deviation from random correlations amongst the mosaic units [50].

5.5. *The scale dependence of the $f = 1/3$ peak :*

It may be interesting to recall the scale dependence of the three periodicity for some well known sequence such as the Thue Morse (TM). The TM sequence is generated by the substitutions [51] :

$$A -> AB,$$

$$B -> BA.$$

It is well studied that three periodicity does play a dominant role in the structure of the TM sequence.

The structure factor F_{AA} in TM sequence at the point $q = 2\pi/3$ may be studied as follows. Take a window of size l and measure F_{AA} ($q = 2\pi/3$). Increase the window size and find how F_{AA} scales with the window size l. The results, derived theoretically, yield :

$$F_{AA}\left(q = \frac{2\pi}{3}\right) \propto l^{1-\alpha}, \tag{132}$$

where $\alpha = 2 - \frac{\ln 3}{\ln 2}$; therefore, $1 - \alpha \approx 0.585$.

In practice, it is convenient to study the normalized structure factors (20) defined as

$$F_{AA}^{NOR} = \frac{F_{AA}}{\overline{F}_{AA}}, \tag{133}$$

where $\overline{F}_{AA}$, the average value, is defined in (22) [52].

Similar analyses for the DNA sequences for the four bases A, T, G and C are shown in Figure 14. The monotonic forms of the curves point towards a long-range organisation in these sequences.

5.6. *Wee frequency enhancement :*

The analysis of the diagonal structure factors, $F_{\alpha\alpha}(q_n)$ of the spectra, reveal interesting structures. Some of these diagonal elements, in the wee region, *i.e.* the low frequency region, appear to be well above, almost 10 times, the mean level. This is generally not true for all the diagonal elements, but only for some of them. This result clearly hints towards an overall long-range organisation in the sequences [53].

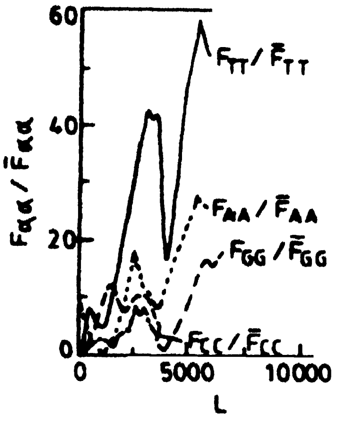


Figure 14. The normalised structure factors (22, 133) for A, C, G and T, calculated from the genome of bacteriophage PHIX174, are plotted against window size L.

The structural entropies of the sequences have been studied as well. They indicate a clear deviation from the purely random order. The distribution of the structural harmonics may be compared with the expected distribution, for the same nucleotide composition, of a random sequence. For the random sequence the distributions are of the Rayleigh type. They are supposed to fall off exponentially. In the real sequences, however, the exponential fall-off is not observed.

5.7. *The Hurst analysis :*

The Hurst Analysis, presented in (55–62), has been over the DNA sequences. The results are presented in Figure 15. The deviations of the measurements from random realizations of identical composition point towards a long-range order in the DNA sequences [53].

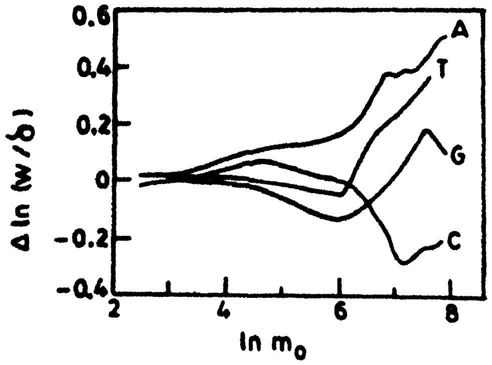


Figure 15. The Hurst's curves (55–62) for A, C, G and T in the genome of bacteriophage PHIX174.

5.8. *The 1/f behaviour of the power spectrum :*

The long-rang behaviour of correlations, *K*, are directly related to the low-frequency dependence of the power-spectrum (sometimes also referred to as the spectral density), denoted here by *S*(*f*). If the power-spectrum is given as (37) :

$$F(q_n) \equiv S(f) \sim \frac{1}{f^{\beta}}. \tag{134}$$

with $\beta > 0$, then the corresponding correlations are given as [49,54],

$$K(x) \sim \frac{1}{x^{\nu}}. \tag{135}$$

The β and ν are related as (23)

$$\nu = 1 - \beta. \tag{136}$$

The case of $\beta = 0$ gives a random sequence. Since ν ranges between 0 and 1, it is not possible to define a range for the correlation function. It is non-zero for all x.

It is important in the analysis of $1/f$ spectra to take out the white-noise. This may be done by comparing the sequence with the decimal figures of π of of same length [49]. The subtraction of the white-noise remains somewhat ambiguous [55], and we shall discuss this point subsequently. The results of the analysis is summarized in Table 12.

5.9. *The DNA walk :*

The DNA walk is an alternate approach towards understanding the organisation of the bases in the DNA sequences. While the short-range periodicities are most clearly appreciated in the Fourier spectra, the longer range correlations amongst the bases and their distributions become transparent in the walk models (See Chapter 3).

The basic idea is to associate the sequence with walk. The ways to do it are many and the choice depends on what features of the sequence one is interested in studying. For the DNA sequences consisting of 4 bases, we illustrate below some possible choices.

(i) One-dimensional Walk : There are many possibilities here. Any two of the bases could be taken to signal the walk step of one step, +1, to the right; the other two, the walk step –1 to the left.

The most popular of these choices is the purine-pyrimidine (PuPy) [56] walk where one moves in opposite directions depending on purine (*A*,*G*) or pyrimidine (*C*,*T*). We shall discuss the PuPy walk in some detail in the ensuing pages.

(ii) Two-dimensional Walk : Once again there are many choices here. Any two of the four bases may move us in +1 step along the x-axis; the other two could be chosen to move us +1 step along the *y*-axis. Since the two-dimensional walk is but a minor variant of the walk in 1-*d* we do not consider it here.

(iii) Four-dimensional Walk : The walk in 4-*d* is unique and directed. Here the *A*, *C*, *G* and *T* are all independent axes along which the sequence makes the walk. A moves +1 in the *A* direction, *C* moves +1 in the *C* direction, *G* moves +1 in the *G* direction and *T* moves +1 in the *T* direction. The sequence is uniquely mapped to the walk, unlike in the lower dimensions where the mapping is not unique [49,55].

We shall discuss this 4-*d* walk as it is of interest to us. Since it treats all the bases independently, it does not introduce spurious correlations unlike in the lower dimensional walks.

5.10. Perspective on the DNA walk :

The basic strategy of the walk models may be summarized as follows :

(i) Plot the walk and measure the averages of quantities such as moments, displacements *etc.* and find out how they scale with the number of steps.

(ii) Compare the scaling behaviour to that of random walk (131).

(iii) If there are significant deviations in scaling there exists correlations in the sequences.

(iv) It is necessary to find out if the scaling properties change with the number of steps. Such changes would imply existence of hidden scales in the sequences. They imply deviations from purely fractal behaviour.

(v) Characterise the correlations and the deviations from randomness and fractality. Find out how these relate to the physiological characteristics of the organisms.

In practice there are several pitfalls that one has to look out for. These are

(a) The proportions of *A*, *C*, *G* and *T* in most sequences are different. Thus, while comparing with the "random" sequence, the randomness needs to be clearly defined. The differences in the proportions of the bases are referred to as the strand bias.

(b) All the sequences we are dealing with have finite lengths. The exon sequences are typically of length of the order of few hundred to a thousand. The effect of the finite size needs to be carefully analysed.

(c) The repetitions, the mosaic structure, sometimes also called the "patchiness", of the sequences need to be carefully kept in mind in investigating the nature of the correlations.

The walk is particularly suited for investigation of long-range correlations that may not appear as peaks in spectral analysis. These long-ranged organisation may not be due to the repetitions in the DNA sequences. The plots of walks, the analysis of the moments, cumulants, and their scaling properties reveal the existence of hidden scales of the sequences.

5.11. The one-dimensional PuPy walk :

The purine-pyrimidine (PuPy) walk has been the subject of major investigation over the last few years. The reasons for this focus on the PuPy walk is mainly because of its simple mathematical framework. If the DNA sequences have to be modelled on 1-*d* walk there are many possible choices, the PuPy walk is really as good as any (55).

The basic steps of the PuPy walk are as follows — The walker steps right, $u(i) = +1$, if pyrimidine (*C* or *T*) occurs at the *i*-th position along the DNA chain; the step is to the left, $u(i) = -1$, if purine (*A* or *G*). The positive steps correspond to concentration of pyrimidines; the negative steps to purines.

The statistical quantity of interest for this is the root mean square fluctuation from the average displacement. The quantity $F(l)$ is thus

$$F^2(l) = < \Delta x(l)^2 > - < \Delta x(l) >^2, \qquad (137)$$

where $\Delta x(l) = x(l_0 + 1) - x(l_0)$ and $x(l) = \sum_{i=0}^{l} u(i)$.

The averaging <> indicates that l_0 has to be varied through the sequence.

The mean square fluctuations $F(l)$ may be related to the correlation functions defined in (18). The correlations above mean are defined as

$$K(l) = \frac{1}{M}\sum u(l_0)u(l_0 + l) - \left(\frac{1}{M}\sum u(l_0)\right)^2. \qquad (138)$$

The relation is

$$F^2(l) = \sum_{i=1}^{l}\sum_{j=1}^{l} K(j - l). \qquad (139)$$

The measurements of $F(l)$ can distinguish between the following possibilities :

(i) If the bases are randomly arranged, $K(l)$, the correlations, are zero except for $K(0)$, which is equal to 1. Thus

$$F^2(l) \sim l, \qquad (140)$$

as expected of the random sequence.

(ii) If there are short-ranged correlations extending upto a length of ξ, then

$$K(l) \sim \exp\left| -\frac{l}{\xi} \right|. \qquad (141)$$

However, asymptotically *i.e.* as $l \to \infty$, the correlations are random; thus

$$F^2(l) \sim l, \qquad (142)$$

for large $l \gg \xi$.

(iii) When there are no characteristic length in the walk, the correlations $K(l)$ are likely to be power laws, and the $F^2(l)$ also follows the power law behaviour (131)

$$F(l) \sim l^{\alpha}, \qquad (143)$$

where α is deviated from 1/2. Note the value of α = 1/2 characterizes Brownian motion, *i.e.* random walk.

The results of the analysis of the scaling of $F(l)$ vary for the exons and the introns. For the exon regions, α is found to be near to 1/2. For the introns there is a significant difference, α differs from 1/2.

The data for the introns and intergenic regions show that α is substantially more than 1/2, indicating the long-range correlations in these regions. For exon regions the log-log plot is not linear; the slope changes from 0.5 for small 1 to 1 for large 1. The exons are interrupted by long intron regions. If the fluctuation analysis is confined to a single patch of the exon region (as opposed to splicing the patches together to form the complete gene or coding region), the value for α is near to 0.5. The value varies from one protein coding region to another. This indicates that the exon patches, despite the short-ranged periodicities, have somewhat lower longer range correlations, compared to the introns, upto the size of the patch.

The results of the PuPy walk has to be interpreted with caution. The reasons are :

(i) The long-range correlation studies from the power spectrum give somewhat different results. They show $\frac{1}{f^{\beta}}$ behaviour for the power spectrum [49].

(ii) The components of the correlation matrix $K_{\alpha\beta}$ obtained in the PuPy and the power spectrum are different.

What is the meaning of the long-range order? Does it have some physical implications in terms of observable biophysical effects?

An independent check on the order in the coding regions came from the GRAIL neural network approach. The GRAIL neural algorithm is trained to identify the protein coding regions in the DNA sequences. The GRAIL was fed random uncorrelated sequences generated artificially. It was also fed artificially produced long-ranged ordered sequences of sizes of about 10^5 bases. Amongst the random sequences several were identified by GRAIL as exon sequences. A less number, amongst the ordered ones, were picked up by GRAIL as candidate exons [58].

The value of the exponent α may be calculated for small windows of nucleotide bases. This is done in an approach called DFA (detrended fluctuation analysis). The minimum value of α usually falls on the coding regions; the maxima on the introns. Based on these results software has been developed that identifies the approximate region of the exons as opposed to the introns. This software has had reasonable success.

The intron regions contain tandem repeats such as AAAAAA. Such repeats do not occur in the coding regions. This may be one of the reasons for the organisational difference between them. Yet, as we move away by about 1000 bases, which is typical size of the protein coding regions of a gene, the $F(l)$ typically undergo crossovers, indicating changes in the proportions of the four nucleotides in the sequences. Beyond these approximate 1000 bases, the existence of the long-range order in exons seems to be indicated in some mesearements. Thus upto about 1000 bases, that is within one protein coding region, the arrangement is almost random.

5.12. *Detrended fluctuations :*

In the analysis of long-range order, it is important to eliminate the effects of hidden underlying bias in the sequences. For the PuPy the major bias is that purine and pyrimidine do not occur in equal proportions in the sequences This hidden bias of compositional complexity needs to be eliminated to establish the existence of the real long-range order [59].

The effect of different $G + C$ content, *i.e.* purine density at different parts of the sequence may be eliminated by detrending.

For this purpose the DNA, sequence is divided into smaller segments. The total number of symbols M, divided now into $\frac{M}{l}$ subsequences each of length l. The subsequences are lebeled by the index s. The bias in the box s is

$$B_l(s)=\frac{1}{l}\ \overset{sl}{\underset{n=(s-1)l+1}{}}\ x(n). \qquad (144)$$

The detrended variable $x_l(n,s)$ is defined as

$$x_l(n,s)=x(n)-nB_l(s) \text{ for } (s-1)l+1\le n\le sl. \qquad (145)$$

The variance over the segment is

$$\sigma_2(s,l)=\frac{1}{l}\sum_{n=(s-1)l+1}^{sl} x_1(n,s). \qquad (146)$$

The fluctuation $F_d^2(l)$ is the average of over $\frac{M}{l}$ segments and depends on the segment size l. Once again, from the behaviour of $F_d^2(l)$, namely

$$F_d^2(l)\sim l^{2H} \qquad (147)$$

allows the evalution of H. If H is close to 0.5 we have random walk. For H > 0.5 we have long-range correlated sequences.

For introns H has been shown to exceed 0.5. For intronless sequences H is near 0.5 below a certain characteristic length, and exceeds 0.5 for larger lengths [60].

5.13. *Four-dimensional walk :*

The walk in 4-*d* treats all the bases A, T, G and C independently and, therefore, does not introduce spurious correlations in the system. Since the bases A, T, G and C signal +1 step move in the A, T, G and C directions, this is a directed walk that never turns back [55].

The characteristic function of a single step is (83) :

$$P_1(k) = p_A e^{ik_1} + p_T e^{ik_2} + p_G e^{ik_3} + p_C e^{ik_4}, \quad (148)$$

where p_i is the probability of a step in the i-th direction. Using the law of convolution (84), the characteristic function for n steps is

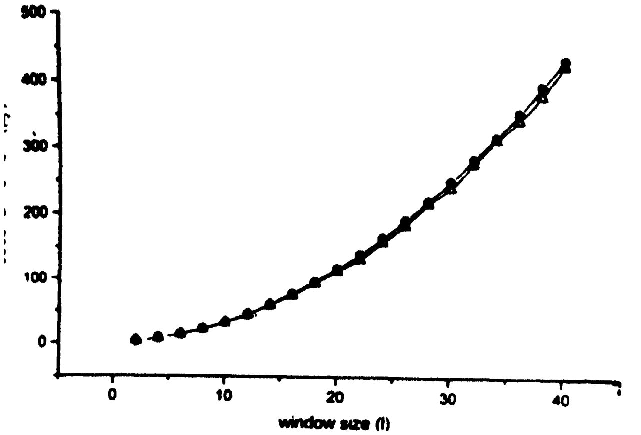


Figure 16. Window size (l) is plotted against corresponding average second moment (μ_2) for the beta-globin gene from *Xenopus* (GenBank Y00501). We have two curves, one for the experimental values while the other for the theoretical values drawn from the analytical calculations for a sequence with same base composition. The (a1) shows the plots of line connecting points for the exons; here solid circle (●) symbol represents the theoretical values and hollow triangle (Δ) symbol represents the experimental values.

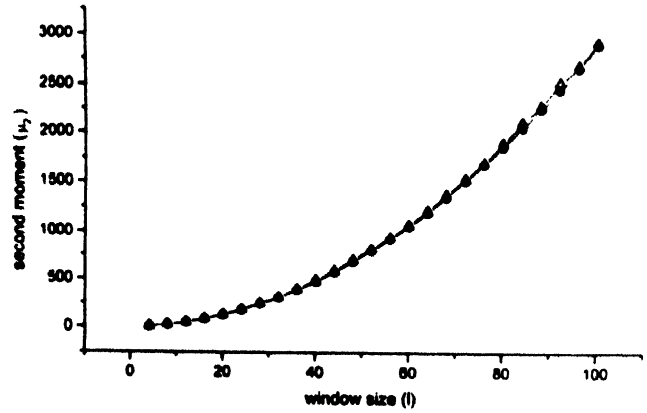


Figure 16. Window size (l) is plotted against corresponding average second moment (μ_2) for the beta-globin gene from *Xenopus* (GenBank Y00501). We have two curves, one for the experimental values while the other for the theoretical values drawn from the analytical calculations for a sequence with same base composition. The (b1) shows the plots of line connecting points for the introns; here solid circle (●) symbol represents the theoretical values and hollow triangle (Δ) symbol represents the experimental values.

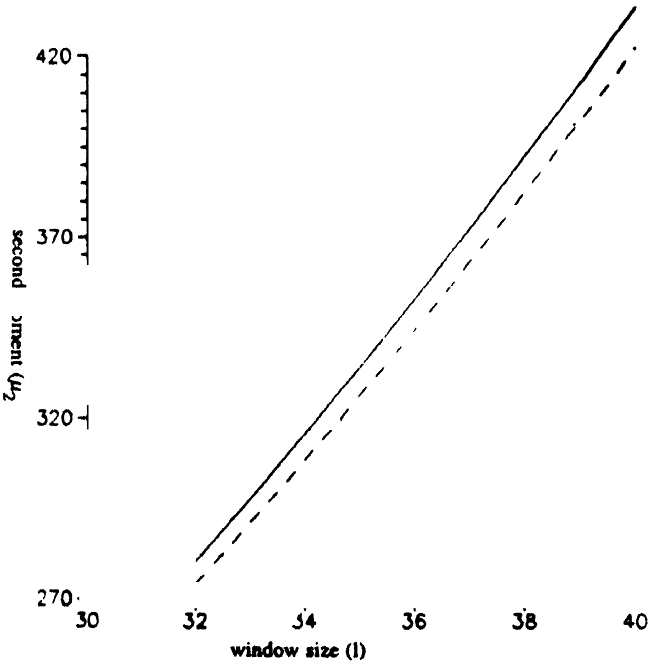


Figure 16. Window size (l) is plotted against corresponding average second moment (μ_2) for the beta-globin gene from *Xenopus* (GenBank Y00501). We have two curves, one for the experimental values while the other for the theoretical values drawn from the analytical calculations for a sequence with same base composition. The (a2) shows the power law fitted plots for parts of the exons; here the solid lines represent the theoretical curves and the dashed lines represent the experimental curves. In each case, the theoretical curve shows the moments for the random sequence with base composition (*i.e.* the proportion of A, C, G, T) same as the DNA sequence. The deviation of the experimental curve from the theoretical one indicates correlation.

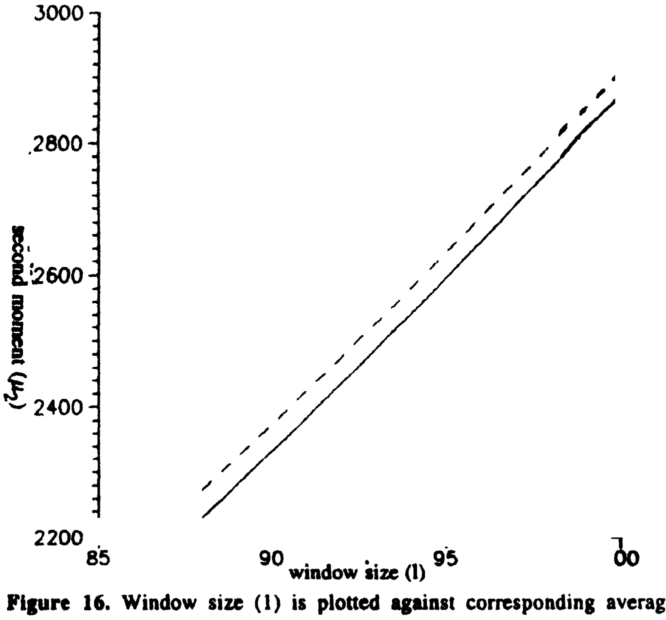


Figure 16. Window size (l) is plotted against corresponding average second moment (μ_2) for the beta-globin gene from *Xenopus* (GenBank Y00501). We have two curves, one for the experimental values while the other for the theoretical values drawn from the analytical calculations for a sequence with same base composition. The (b2) shows the power law fitted plots for parts of the introns; here the solid lines represent the theoretical curves and the dashed lines represent the experimental curves. In each case, the theoretical curve shows the moments for the random sequence with base composition (*i.e.* the proportion of A, C, G, T) same as the DNA sequence. The deviation of the experimental curve from the theoretical one indicates correlation.

$$P_n(k) = \left(p_A e^{ik_1} + p_T e^{ik_2} + p_G e^{ik_3} + p_C e^{ik_4}\right)^n. \qquad (149)$$

The probabilities p_A, p_T, p_G, p_C are obtainable for the sequence. They are the proportions of A, T, G and C respectively. Thus;

$$p_i = \frac{\text{No. of times the symbol } i \text{ appears in the sequence}}{\text{Total no. of bases}}.$$

The moments of the distributions are calculable from the characteristic function of n-steps, (91–92). The first and second moments are given by

$$\mu_1 = l(l-1).p_l^2 + l.p_l, \qquad (150)$$

$$\mu_2 = l\left[(l-1)(p_A^2 + p_T^2 + p_G^2 + p_C^2) + 1\right], \qquad (151)$$

where l stands for the window size.

Figure 16 shows the typical plot of μ_2 as a function of the window size l. The deviation from the theoretical expression (150–151) points to the internal organisation in the DNA sequences.

The Figure 16 shows that for small values of the scale the sequences may be steeper than the theoretical prediction. The second moment

$$\mu_2 = l\left[(l-1)(p_A^2 + p_T^2 + p_G^2 + p_C^2) + 1\right], \qquad (152)$$

for a window of size l is just the square of the vector distance between the end points. This is averaged over the whole sequence. Thus a steeper slope indicates an increase in persistence. These persistence may continue through the sequence, or may crossover into antipersistence at a higher scale. Figure 17 illustrates the meaning of persistent/ antipersistent behaviour.

We look at the local values of the second moment as we move along the sequence, the typical behaviour is illustrated in Figure 18. The tandem repeats lead to sudden hugh rise in the local moments. Otherwise, they are distributed as shown in Figure 18.

5.14. Base organisation in DNA :

To summarise the results of the measurements on the DNA sequences, we have :

(i) Introns and Intergene Regions : No universal short-range periodicity. Existence of the long-range order is noted. The autocorrelations for these segments show an inverse power law decay. The typical form of this decay has the structure

$$K_{aa}(x) \sim \frac{1}{x^{\alpha}}. \qquad (153)$$

(ii) Exons : For these segments there is the short-ranged periodicity typified by the sharp peak of the power spectrum at $f = 1/3$. This is presumably due to the triplet codons sitting along the exon segments.

As we go to distances larger than 3 bases, the exon bases enter random fluctuations with no significant correlations. Further out, that is, as we move from one gene onto the next, there begins, once again, an inverse power law structure of the autocorrelations.

The above characterisation of the various segments of the DNA points towards the complexity of the sequences. As the sequences are subdivided into segments these segments do not show randomness. The sequences behave somewhat differently at different scales [see Chapter 2].

There are some indications that the long-range behaviour, in particular the exponent β in (134), may be characteristic of phyla to which the DNA belongs [49] (See Table 10).

Table 10. The variation among different categories of living organisms in the value of β, averaged over a number of sequences from each category

Category	Avg β Value
Primate	0.77
Rodent	0.81
Mammal	0 84
Vertebrate	0.87
Invertebrate	1
Plant	0.86
Virus	0 82
Organelle	0.71
Bacteria	1 16
Bacteriophage	1 02

The long-range order, *i.e.* non-zero values of β in (134), is sensitive to the method of analysis. Clearly log $S(f)$ *vs* log f in (134) when plotted ought to give rise to a straight line with slope of $-\beta$. In practice the above plot is rarely linear. The subtraction of white noise may take the plot a bit more linear. A linear fit in the low frequency region can then give us the value of β. However, there is no consensus on the frequency region where the fit is to be carried out. Further, the data for β, presented in the literature [49,56], averages over phyla. The variations in β from the averages appear as meaningful as the averages themselves. More important is to find an unambiguous method for determining β. We note here that the averaged β values obtained for the sequences from Genbank are [56] :

(i) for exons : $\beta = 0.00 \pm 0.04$,

(ii) for non-coding regions : $\beta = 0.16 \pm 0.05$.

5.15. DNA modelling :

The results above place constraints on modelling of the DNA sequences. The evolution of the sequences as we go from the prokaryotes to the eukaryotes requires careful understanding. The modelling has to account for this evolutionary pattern and identify the underlying physical laws.

Broadly there have been three different approches to the DNA modellings. All of these models carry out sequence

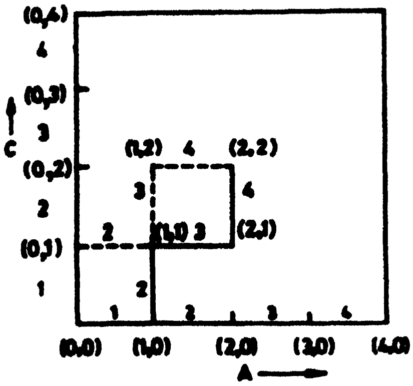


Figure 17. The second moment, μ_2 (152), for the persistent nature of A or C (*i.e.* one A or C is followed by another A or C respectively), is greater than that for thè antipersistent nature of A or C (*i.e.* an A is followed by a C, or a C is followed by an A). The μ_2 for a 4-step persistent walk has a value of 16, while the μ_2 for a 4-step antipersistent walk has a value of 8.

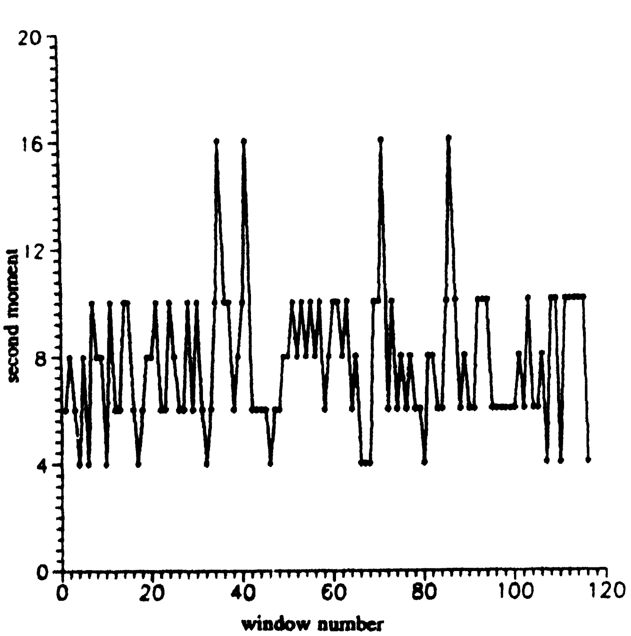


Figure 18. The second moment is plotted against the number of windows for a fixed window size 4. The sequence selected is the total intron region of the α-globin gene from *Xenopus* (GenBank X14260).

mainpulations, *i.e.*, update a sequence by means of some rules.

(I) Markov chain models

Here the assumption is that the last base "emits" the one that follows in the chain. In more sophisticated versions of these models the last *n* bases "emit" the base that follows, the so called *n*-step Markov process.

The difficulty in the Markov Chain models is that they rarely have long-range order that is so characteristic of the DNA sequences. The inverse power law of correlations are difficult to generate.

(II) Cellular automata approach

Here local law of updating is proposed that is supposed to mimic the rules of the DNA mutations. Once again, starting from the local updating rules the long-range correlation structure shows, usually the exponential fall off, as opposed to the inverse power-law behaviour [61].

There are some candidate cellular automata models that do have possibilities of inverse power law correlations [62]. These propagate local effects with a low randomness level. These "edge-of-chaos" models have been studied widely. There also exist models of cellular automatas with solitonic configurations and sometimes leads to $1/f$ type spectrum. These topics take us beyond the scope of the present review.

(III) Inflationary models

These are models that generate the sequence starting from a single or a few bases. We have touched upon the well known Thue-Morse sequence in our discussions of the $f = 1/3$ peak. There are methods of sequence generation that provide the correlation structure shown in the DNA sequences. Here we discuss some of these inflationary models.

(i) Expansion-modification system [63]

The expansion modification is an algorithm to generate a sequence that has long-range inverse power law correlation of symbols. In its simplest form the algorithm is :

Step i	1		0	
Step $i+1$	11	0	00	1
Probability	$1-p$	p	$1-p$	p

Clearly the algorithm generates a binary sequence of 0 and 1. The algorithm, stated in words, changes 1 to 11 with probability $1-p$, or to 0 with probability p. Similarly 0 is changed to 00 with probability $1-p$, or to 1 with probability p.

When p is small, the sequence generated show a $\frac{1}{f^\beta}$ power spectrum with β close to 1. By changing the value of p we can get other values of β. Thus by choosing appropriately, it is possible to maintain the value of β of the starting sequence.

It is known that β depends on the category of the organism as shown from the analysis of the Genebank data. The control parameter p, sometimes called the mutation rate, in that sense, signifies the differences in the evolution of gene categories.

(ii) Insertion models [64]

The "model" of evolution by "mutation" of bases as proposed in the expansion modification system fails to address the issue of repetitions observed in the DNA sequences. These repetitions suggest that the dominant mechanism of evolution could be the process of duplication of segments, followed by the insertions of these duplicates at various sites along the sequence. The general observation is that the duplicates may undergo some mutations, while the original segment (from which the duplicates are made) remain unmutated and in tact. In other words, the working original is left untouched, while the duplicated copies are subject to evolutionary process. In the Insertion models the idea is to begin with segments that are repeated many times in the DNA sequences. Take for instance the LINE1 segment which has a length of 6139 bases. This segment codes for protein. In the human genome the LINE1 or its variations appear about 107000 times. If we look at the genome of Chimpanzee, which is quite close to the humans in the evolutionary scale, the LINE1 or its fragments appear 51000 times. Thus the difference in the LINE1 content between human and Chimpanzee is large. This difference has come about in a short time in the scale of evolution.

Since the LINE1 is a protein coding segment the insertions of LINE1 so many times in the sequence imply the followings :

1. The corresponding protein is required in large quantities for the biological systems such as humans and Chimpanzee. Chearly humans require more of this protein.
2. Since these are coding regions, the bases are arranged in more-or-less random order. The fluctuations $F(l)$ scale as

$$F^2(l) \sim l^{1/2}. \tag{154}$$

There are similar other segments which repeat over the sequences. ALU for instance is a protein coding stretch of 290 bases that is found several times in the sequences.

The idea of insertion model is to develop a complex sequence by insertions of simple random segments. The algorithm is summarized in the following two assumptions :

(a) The probability of finding a repeated segment of length l in the sequence goes as

$$P(l) \sim \frac{1}{l\mu}, \tag{155}$$

where μ is an adjustable parameter that is related to the exponent of the inverse power law of the long-range order.

(b) The segments do not have any long-range order in them. These parts may have short range order or periodicities and may be modelled by n-step-Markov processes, with small values of n.

This insertion model of evolution leads to the Levy walk that has inverse power law correlations in the long-range. The correlation exponent α is related to the control parameter μ as follows :

$$\begin{aligned} \alpha &= 1 \text{ for } \mu \le 2 \\ &= 2 - \mu/2 \quad 2 < \mu < 3 \\ &= 1/2 \quad \mu \ge 3. \end{aligned} \tag{156}$$

Thus the case of μ between 2 and 3 is of interest. In this case the distribution $P(l)$ has a diverging second moment The first moment, however, is finite. Note that the distribution $P(l)$, given in (155) does not have any special length scale.

As l increases, the fluctuations $F(l)$, in the log-log plot of $F(l)$ *vs* l, increases and asymptotically, for very large values of l, reaches the value of $\alpha = 0.9$.

A generalized version of this model allows for deletions and also insertion of intron elements according to the distribution law (155). Starting with a statistically uncorrelated stretch of myosin heavy chain (MHC) coding sequence, the delation-insertion model has tried to chalk out the evolutionary pattern of the MHC gene. The value of α increases as we go up in the evolutionary chain.

(iii) Copying mistake map (CMM) [65]

The expansion modification system creates a sequence with long-range inverse power law correlation of arbitrary exponent On the other hand the insertion models and its generalisations create long range order by insertions and random delations of random segments distributed as an inverse power law of the length of the segments. Both these methods of sequence generation model aspects of the DNA sequences They are, however incomplete and do not have an unified approach to the different parts, *i.e.*, the coding and the noncoding regions of sequences.

The Copying Mistake Map (CMM) by contrast takes an unified view of the DNA sequences. The CMM is based on the following observations :

(a) The fluctuations $F(l)$, defined in (137), scale differently than the usual random walk. The usual random walk leads to the Gaussian probability distribution. Since the correlations have inverse power law behaviour, the corresponding probability distribution, must have long tails typical of the Levy. The "diffusion" is thus of the anomalous variety.

(b) This anomalous diffusion and long range order is generated in CMM by modeling the DNA sequences a la' continuous "time" random walk. The waiting "time" distribution is chosen to generate the long-range correlations.

(c) Simultaneously, a point mutation, of the white noise variety, works to randomise the sequences.

(d) The strengths of these two opposing ingredients, the one that brings order; the other that randomises, are adjusted to fit the DNA sequences.

The precise physical understanding of the waiting "time" in the DNA continuous "time" walk remains unclear. These waiting times are chosen to have the inverse power law form with finite first moment.

In the general diffusive process, the diffusive variable x is related to the stochastic variable, such as velocity v, that causes diffusion as :

$$\langle x^2(t)\rangle = \langle x^2(0)\rangle + 2\langle v^2\rangle \int_0^t dt' \int_0^{t'} dt'' K(t''), \quad (157)$$

where $K(t)$ are the correlation of the stochastic variable

$$K(t) = \frac{\langle v(0)\, v(t)\rangle}{\langle v^2\rangle}. \quad (158)$$

For the processes that are stationary, the correlations depend only on the time difference

$$t - 0 = t.$$

If the correlations are of the normal type, there exists a time scale τ such that

$$\tau = \int_0^\infty K(t)dt. \quad (159)$$

The correlations usually decay quickly so that τ is finite. The eq (157) in this case simply result

$$\langle x^2(t)\rangle = \langle x^2(0)\rangle + 2Dt, \quad (160)$$

where the diffusion coefficient

$$D = \langle v^2\rangle \tau. \quad (161)$$

The Central Limit theorem for this case works and one has asymptotic Gaussian form.

If on the other hand τ is not finite, we have the anomalous case. A simple way to realize this is when the correlations have inverse-power law behaviour

$$K(t) \sim \frac{1}{t^\beta} \quad \text{with } 0 < \beta < 1. \quad (162)$$

The correlation time t diverges; the Central Limit is no longer realizable.

It turns out that the correlation function is related to the waiting time distribution $w(t)$ for changes in the stochastic variable $v(t)$. The correlation K is related to w as follows :

$$K(t) = \frac{\int_t^\infty (T-t)\omega(T)dT}{\int_t^\infty T\omega(T).dT}. \quad (163)$$

Thus if $\quad \omega(t) \sim \frac{1}{t^\mu}, \quad (164)$

with $2 < \mu < 3$, β in (162) ranges between 0 to 1, since

$$\beta = \mu - 2.$$

Thus the waiting time distribution of the form (164) leads to an inverse-power correlation, resulting in anomalous diffusion. The connection with anomalous diffusion becomes transparent with

$$\langle x^2\rangle = t^{2H} \quad (165)$$

where $H = 2 - \mu/2$.

Thus H ranges form 1/2 to 1, indicating deviation form the normal diffusion exponent.

It is known that above behaviour arise from characteristic function of the Levy form :

$$P(k,t) = \exp\left\{-b|k|^\alpha .t\right\}, \quad (166)$$

with $\alpha = \mu - 1$. It is known how to generate this type of waiting time distribution by using deterministic maps. We shall not concern here with this exact form of the update algorithm required to generate the inverse power distribution (164).

For the DNA sequences the bases are the stochastic variable and may be chosen to assume ±1 values depending on purine or pyrimidine (say). The analogy with diffusion means that in this case there are the possible choices of velocities, v_t. However the long-range correlation of bases ensured by the waiting time is not sufficient to generate the DNA sequences. A further noise that randomises is introduced as follows :

v_t is updated a la deterministic map that produces (164) with probability ε

v_i is updated to $\pm v_i$ with probability $1 - \varepsilon$.

The second moment $<x^2>$ under the action of the deterministic plus the random updates takes the form :

$$\langle x^2(l)\rangle = Al^{2H} + Bl, \quad (167)$$

where the factor H now is solely determined by the deterministic map that has the power law waiting distribution The difference between the intron containing and intronless sequences are obtained by varying $1 - \varepsilon$, the copying mistake probability. The ratio

$$\frac{A}{B} : \frac{\varepsilon}{1-\varepsilon}, \quad (168)$$

when B is larger than A, suppresses the long-range effects.

5.16. *Facts and Physics of evolution :*

It is important to remind ourselves of the simple facts about the DNA evolution.

(i) In the prokaryotic genomes, in the majority of cases, the bases are all used for protein coding The introns and the intergenic regions are, by and large, absent This 'economy' of base arrangement help these biological entities to reproduce on very short time scales (The reproduction entails the duplication of the genomes). It is also known that the prokaryotic genes contain less of repetitions. These coding sequences are characterised by long-range inverse power law correlations.

(ii) As we move "up" in the scale of biological evolution into higher and higher eukaryotes we have

(a) The appearance of the introns and the intergenic regions marked by long-range power law correlations.

(b) The relatively short coding regions have almost random arrangement of base pairs. However when the introns are spliced out, and all the exons are put together, the data reveals the appearance of the ubiquitous long-range correlations.

(c) The appearance of the introns and intergenic stretches do slow down the process of genome duplication and consequently the time it takes to reproduce.

(d) The large amount of base repeats in these organisms usually point towards the increased necessity of certain proteins and enzymes. Since these proteins are required in larger amounts, they have to be produced more relative to the others. Hence the repeats.

(e) The structural complexity of the sequences is indicated when it is subdivided into smaller segments. The segments do not have similar statistical character. While the introns have order, the small exon elements appear disordered, but with a peak at $f = 1/3$ in the power spectrum.

5.17. The DNA tertiary structure :

The DNA, or more precisely the nuclear DNA, occurs in chromosomes in conjunction with proteins. The DNA, stretched out, could be as long as a meter in length, such as in the humans. Inside the chromosome, however, they remain coiled into a region of 10^{-6} meters.

There has been an effort towards understanding the long-range correlations in terms of the constraints on the DNA coiling, *i.e.*, the tertiary structure of the chain. The intial few steps of size reduction is brought about by the proteins (called histones) around which the DNA complex binds into spiral.

In a recent work properties of polymers we studied under the following conditions [66] :

(i) that the polymer be confined to the minimal volume,

(ii) that they remain knot-free. This follows from the requirements of the duplication and the transcriptions of the genes. This kind of packaging of polymers, called the crumpled globule structure, seems to require a long-range correlation of the bases with the Hurst index of 2/3. Interestingly, the Hurst index of the intron and the intergenic regions of the DNA sequences are near this value.

6. An assessment

"In Nature's infinite book of secrecy
A little can I read".
–William Shakespeare

There has been an upsurge of interest amongst physicists in the DNA in recent years. The discovery of the structure of the DNA and the subsequent deciphering of the genetic code mark two high points of research of this century. Yet, there are parts and features of the DNA that remain beyond our grasp. These parts, amusingly enough, constitute the bulk of the DNA. They have been called the "junk" DNA, and swept under the rug. The recent interest of physicists stems from the belief that the "junk" DNA is ready now for another attempt at deciphering.

It began in the early part of the nineties when the analysis of the power spectrum of the DNA showed that it goes like $\frac{1}{f^{\beta}}$, where f is the frequency. Simultaneously, the scaling behaviour of the second moment of the DNA distribution showed that the Hurst index deviates from 0.5. In this analysis the DNA sequence was considered to be a sequence of purines and pyrimidines. The sequence was thought of as a walk on these purines and pyrimidines. Normally, if purines and pyrimidines are distributed randomly over the sequence, the mean square displacement from the origin (the starting point) should go as the number of steps raised to the power twice of the Hurst index of 0.5. If we detect a deviation from 0.5 for the Hurst index, the sequence has long-range correlations.

The detection of deviation from the value of 0.5 of the Hurst index came first for the "junk" parts of the DNA. These "junk" parts, made of the introns and the intergenic regions, therefore, have inverse-power-law correlations over the long-range.

For the exons, *i.e.*, the coding regions, there are short-ranged periodicity, of period 3, arising from the triplet codons. Over intermediate ranges, the exon sequences, curiously enough, show random arrangement of purine-pyrimidine bases. As we go further away, putting the exon segments together by splicing out the introns, the long-range order does seem to return. The order, however, is weaker.

Usually, that is in good majority of circumstances studied in the physical world, we are familiar with correlations that die off exponentially. We know, however, that near the second order phase transition, correlations exist over all length scales. Near this sort of transition, the systems have no preferred scale, and therefore, are scale invariant. Interestingly, in one dimensional systems this type of behaviour is not common, even unexpected.

The inverse-power-law correlations, seen in the DNA, are observed in many other natural phenomena. The fractal property is generally held to be responsible. For fractals imply the absence of any intrinsic scale. Inverse-power-law correlated systems do lack the decay lengths so characteristic of exponential correlations. The DNA, in that sense, shows fractal nature.

The regularities shown in random walk are well studied. All the moments of the underlying distributions are precisely predictable. The random walk is a fractal, albeit a statistical fractal. The Hurst index for the second moment is $\frac{1}{2}$. The deviation from this value for the Hurst index for the DNA implies :

(i) that the DNA base distribution is a statistical fractal;

(ii) that the base distribution has long-range tails.

This kind of distribution does not follow the Central Limit. Instead, the general solution of the chain-rule of probability distribution functions given by Levy is appropriate. The walk executed in purine-pyrimidine bases appear more like anomalous diffusion or Fractional Brownian Motion.

If the DNA are indeed statistical fractals, and have fairly regular features, it must be possible to generate sequences with features of the DNA. Most of the DNA modelling are efforts in this direction. The symbols are manipulated and sequences generated to have features statistically similar to the DNA sequences.

The progress in understanding the regularities in the DNA have been rapid in recent years. There is a lot, however, that remains to be understood.

The presence of the long-range order requires careful analysis. The base composition of the sequences are not uniform. That is, the bases A, C, G and T, occur in different proportions in the sequences. An added unavoidable feature is the finite length of the sequence. The exon sequences are short, running upto about 500 bases. The introns are somewhat longer, upto several thousands. The intergenic sequences are the longest.

To define the long-range correlations it is necessary to subtract out the correlations in the random sequences of the same base compositions (or the strand bias as it is technically called). An unambiguous elimination of this background determines how much of the long-range correlation that remains.

The sequences have large number of repeats of identical or nearly identical subsequences. The long-range correlations, according to some analysis, is due to these base repeats. The physical meaning of the correlations continues to elude complete understanding.

Much of the effort has gone into the analysis of purine-pyrimidine walk. That DNA is just a walk on purine-pyrimidine is but half the story. The correlations obtained from the power spectrum, and the ones from purine-pyrimidine walk, refer to two quite different aspects of the sequences. Taken together, they do not completely define the long-range statistical properties of the sequences. The complete walk in A, T, G and C is required for model building. In the full 4-*d* walk in A, T, G and C space, the diagonal as well the off-diagonal correlations are required. Without them, the model building effort will remain incomplete.

The notion that the DNA is fractal-like, once again, is just a part of the story. In practice the DNA sequences have many scales.

There have been efforts at relating the tertiary structure of the DNA to the long-range order. The Hurst index of order is close to the number calculated for knot-free coiling of DNA into crumpled globules. The physical meaning and the purpose of the order require further attention.

The story that began almost a century-and-a-half ago in the laboratory of Meischer and the pea farm of Mendel has come a long way. There is but a good bit that remains untold.

References

[1] T H Morgan *Science* **32** 120 (1910)

[2] I Shine and S Wrobel *Thomas Hunt Morgan : Pioneer of Genetics* (Lexington : University Press of Kentucky) (1976)

[3] R Olby *Trends in Biochemical Sciences* **11** 303 (1986)

[4] O T Avery, C M MacLeod and M McCarty *J Experim Medicine* **79** 137 (1944)

[5] A D Hershey and M Chase *J Gen Physiol* **36** 39 (1952)

[6] E Chargaff *Experientia* **6** 201 (1950)

[7] R Olby *The Path to the Double Helix* (London Macmillan) (1974)

[8] H R Wilson *Trends in Biochemical Sciences* **13** 275 (1988)

[9] J D Watson and F H C Crick *Nature* **171** 737 (1953)

[10] J D Watson and F H C Crick *Nature* **171** 964 (1953)

[11] P Chambon *Sci. Am.* **244** 48 (1981)

[12] W Gilbert *Nature* **271** 501 (1978)

[13] W Gilbert *Science* **228** 823 (1985)

[14] F H C Crick *Nature* **227** 561 (1970)

[15] H Temin *Sci. Am.* **226** 24 (1972)

[16] F H C Crick *Science* **204** 264 (1979)

[17] L Stryer *Biochemistry* (New York Freeman) (1995)

[18] F H C Crick, F R S L Barrett, S Brenner and R J Watts Tobin *Nature* **192** 1227 (1961)

[19] F H C Crick *Sci Am* **207** 66 (1962)

[20] M W Nirenberg *Sci Am.* **208** 80 (1963)

[21] F H C Crick *Sci. Am.* **215** 55 (1966)

[22] B D Hall *Nature* **282** 129 (1979)

[23] T A Brown *Gene Cloning An Introduction* (London Chapman & Hall) (1990)

[24] V L Davidson and D B Sittman *Biochemistry* (New Delhi B Y Waverly) (1994)

[25] A Maxam and W Gilbert *Proc. Natl. Acad Sci* (USA) **74** 560 (1977)

[26] F Sanger, S Nicklen and A R Coulson *Proc. Natl Acad Sci* (USA) **74** 5463 (1977)

[27] T J White *Trends in Genetics* **5** 185 (1989)

[28] B Lewin *Genes V* (Oxford . Oxford University Press) (1994)

[29] M D Frank-Kamenetskii *Phys. Rep.* **288** 13 (1997)

[30] T Strachan *The Human Genome* (Oxford : BIOS Scientific Publishers) (1992)

[31] T A Brown *Genetics A Molecular Approach* (London : Chapman & Hall) (1992)

[32] D B Percival and A T Walden *Spectral Analysis for Physical Applications* (Cambridge : Cambridge University Press) (1993)

[33] G B Folland *Fourier Analysis and its Applications* (USA : Wadswork and Books) (1992)

[34] J K Bhattacharjee and A K Mallik *Modelling of Complex System* (New Delhi : Narosa) (1997)

[35] S L Marple *Digital Spectral Analysis with Applications* (New Jersey : Prentice Hall) (1987)

[36] W H Press, S A Teukolsky, W T Vetterling and B R Flannery *Numerical Recipes* (Cambridge : Cambridge University Press) (1992); *Selected Papers on Fast Fourier Transform in Digital Signal Processing* (IEEE Press) (1972)

[37] V R Chechetkin and A Yu Turygin *J. Phys.* **27** 4875 (1994)

[38] S K Das *Indian J Phys* **69B** 601 (1995)

[39] A N Kolmogorov *IEEE Trans. on Information Theory* **IT-14** 662 (1968)

[40] G J Chaitin *J Assoc. Comp Mach.* **13** 547 (1966)

[41] C E Shannon and W Wearver *The Mathematical Theory of Communication* (Illinois : University of Illinois Press) (1949)

[42] H E Hurst *Trans. Am. Soc. Civ. Eng.* **116** 770 (1951); J Feder *Fractals* (New York : Plenum) (1988)

[43] S Chandrasekhar *Rev. Mod. Phys.* **15** 1 (1943)

[44] M C Wang and G E Uhlenback *Rev. Mod. Phys.* **17** 323 (1945)

[45] E W Montroll and J L Bebowitz *Fluctuation Phenomena* **VII** 64 (eds.) E W Montroll and J L Lebowitz (Amsterdam : North-Holland) (1979)

[46] E W Montroll and M F Shlesinger in *Nonequilibrium Phenomena II From Stochastics to Hydrodynamics* (eds.) J L Lebowitz and E W Montroll (Amsterdam : North-Holland) 1 (1984)

[47] J W Fickett *Nucl. Acids Res.* **10** 5303 (1982); A A Tsonis, J B Elsner and P A Tsonis *J. Theor. Biol.* **151** 323 (1991); F H C Crick, S Brenner, A Klug and G Pieczenik *Orig. Life* **7** 389 (1976); M Eigen, W Gardiner, P Schuster and R Winkler-Oswatitsch *Sci. Am.* **222(4)** 88 (1981); E N Trifonov and J L Susman *Proc. Natl. Acad. Sci.* **77** 3816 (1980); J C W Shepherd *Proc. Natl. Acad. Sci.* **78** 1596 (1981); B D Silverman and R Linsker *J. Theor. Biol.* **118** 295 (1986); S Tavare and B W Giddings in *Mathematical Methods for DNA Sequences* (ed.) M S Waterman 116 CRC Press (1989); D G Argnes and C J Michel *J. Theor. Biol.* **143** 307 (1990)

[48] S Tiwari, S Ramachandran, A Bhattacharya and R Ramaswamy *Comput. Appli. Biosci.* **13** 263 (1997); P Lio, S Ruffo and M Buiatti *J Theor. Biol.* **171** 215 (1994)

[49] R Voss *Phys. Rev. Lett.* **68** 3805 (1992); S V Buldyrev, A L Goldberger, S Havlin, C K Peng, M Simon, F Sciortino and H Stanley *Phys Rev. Lett* **71** 1776 (1993); R Voss *Phys Rev. Lett.* **71** 1777 (1993)

[50] S Nee *Nature* **357** 450 (1992); S Karlin and V Brendel *Science* **259** 677 (1993)

[51] Z Cheng, R Savit and R Marlin *Phys Rev.* **B37** 4375 (1988); Z Cheng and R Savit *Phys. Rev.* **A44** 6379 (1991)

[52] V R Chechetkin and A Yu Truygin *Phys. Lett.* **A199** 75 (1995)

[53] V R Chechetkin, L A Knizhnikova and A Yu Turygin *J Biomol Struct Dyn.* **12** 271 (1994)

[54] W Li and K Keneko *Europhys Lett* **17** 655 (1992)

[55] S Chattopadhyay, A Som, S Sahoo and J Chakrabarti (Submitted for Publication)

[56] C K Peng, S V Buldyrev, A L Goldberger, S Havlin, F Sciortino, M Simons and H E Stanley *Nature* **356** 168 (1992), S V Buldyrev, A L Goldberger, S Havlin, R N Mantegna, M E Matsa, C K Peng, M Simons and H E Stanley *Phys Rev* **E51** 5084 (1995)

[57] J Maddox *Nature* **358** 103 (1992), V V Prabhu and J M Clavier *Nature* **359** 782 (1992); C A Chatzidimitriou-Dreismann and D Larhammer *Nature* **361** 212 (1993)

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

[59] C K Peng, S V Buldyrev, S Havlin, M Simons, H E Stanley and A L Goldberger *Phys Rev.* **E49** 1685 (1994)

[60] B J West and W Deering *Phys. Reports* **246** 1 (1994), W Li, T G Marr and K Kaneko *Physica* **D75** 392 (1994), S V Buldyrev, A L Goldberger, S Havlin, C K Peng and H E Stanley *Fractals in Biology and Medicine : From DNA to the Heartbeat* in *Fractals in Science* 49 (Berlin : Springer-Verlag) (1994)

[61] C Burks and D Farmer *Physica* **D10** 157 (1984); **D45** (1990)

[62] C G Langton *Physica* **D47** 12 (1990)

[63] W Li *Europhys. Lett.* **10** 395 (1989); W Li *Phys. Rev* **A43** 5240 (1991)

[64] S V Buldyrev, A L Goldberger, S Havlin, C K Peng, M Simons and H E Stanley *Phys Rev.* **E47** 4514 (1993); S V Buldyrev, A L Goldberger, S Havlin, C K Peng, H E Stanley, M Stanley and M Simons *Biophys. J.* **65** 2675 (1993)

[65] P Allegrini, M Barbi, P Grigonini and B J West *Phys. Rev.* **E52** 5281 (1995); P Allegrini, P Grigolini and B J West *Phys. Lett.* **A211** 217 (1996); P Allegrini, M Buiatti, P Grigonini and B J West *Phys. Rev.* **E57** 4558 (1998)

[66] A Yu Grosberg, Y Rabin, S Havlin and A Neer *Europhys. Lett.* **23** 373 (1993); A Yu Grosberg and A R Khokhlov *Statistical Physics of Macromolecules* (Moscow : Nauka) (1989)

About the Reviewers

Dr. Jayprokas Chakrabarti - Presently holding the position of Reader in the Department of Theoretical Physics, Indian Association for the Cultivation of Science.

Dr. Satyabrata Sahoo - Completed Ph.D. in the Department of Theoretical Physics, Indian Association for the Cultivation of Science; presently undergoing post-doctoral research in the Institute of Atomic and Molecular Sciences, Academia Sinica, Taipei, Taiwan.

Sujay Chattopadhyay - Completed M.Tech. in Biotechnology from Indian Institute of Technology, Kharagpur; presently undergoing doctoral research in the Department of Theoretical Physics, Indian Association for the Cultivation of Science under the supervision of Dr. J Chakrabarti.

Anup Som - Completed M.Sc. in Physics from Jadavpur University, Calcutta; presently undergoing doctoral research in the Department of Theoretical Physics, Indian Association for the Cultivation of Science under the supervision of Dr. J Chakrabarti.