1 Representation and analysis of DNA sequences

Paul Dan Cristea

1.1. Introduction

Data on genome structural and functional features for various organisms is being accumulated and analyzed in laboratories all over the world, from the small university or clinical hospital laboratories to the large laboratories of pharmaceutical companies and specialized institutions, both state owned and private. This data is stored, managed, and analyzed on a large variety of computing systems, from small personal computers using several disk files to supercomputers operating on large commercial databases. The volume of genomic data is expanding at a huge and still growing rate, while its fundamental properties and relationships are not yet fully understood and are subject to continuous revision. A worldwide system to gather genomic information centered in the National Center for Biotechnology Information (NCBI) and in several other large integrative genomic databases has been put in place [1, 2]. The almost complete sequencing of the genomes of several eukaryotes, including man (*Homo sapiens* [2, 3, 4]) and "model organisms" such as mouse (*Mus musculus* [5, 6]), rat (*Rattus norvegicus* [7]), chicken (*Gallus-gallus* [8]), the nematode *Caenorhabditis elegans* [9], and the plant *Arabidopsis thaliana* [10], as well as of a large number of prokaryotes, comprising bacteria, viruses, archeia, and fungi [1, 2, 5, 11, 12, 13, 14, 15, 16, 17, 18, 19], has created the opportunity to make comparative genomic analyses at scales ranging from individual genes or control sequences to whole chromosomes. The public access to most of these data offers to scientists around the world an unprecedented chance to data mine and explore in depth this extraordinary information depository, trying to convert data into knowledge.

The standard symbolical representation of genomic information—by sequences of nucleotide symbols in DNA and RNA molecules or by symbolic sequences of amino acids in the corresponding polypeptide chains (for coding sections)—has definite advantages in what concerns storage, search, and retrieval of genomic information, but limits the methodology of handling and processing genomic information to pattern matching and statistical analysis. This methodological limitation

determines excessive computing costs in the case of studies involving feature extraction at the scale of whole chromosomes, multiresolution analysis, comparative genomic analysis, or quantitative variability analysis [20, 21, 22].

Converting the DNA sequences into digital signals [23, 24] opens the possibility to apply signal processing methods for the analysis of genomic data [23, 24, 25, 26, 27, 28, 29, 30, 31, 32] and reveals features of chromosomes that would be difficult to grasp by using standard statistical and pattern matching methods for the analysis of symbolic genomic sequences. The genomic signal approach has already proven its potential in revealing large scale features of DNA sequences maintained over distances of 10^6–10^8 base pairs, including both coding and noncoding regions, at the scale of whole genomes or chromosomes (see [28, 31, 32], and Section 1.4 of this chapter). We enumerate here some of the main results that will be presented in this chapter and briefly outline the perspectives they open.

1.1.1. Unwrapped phase linearity

One of the most conspicuous results is that the average unwrapped phase of DNA complex genomic signals varies almost linearly along all investigated chromosomes, for both prokaryotes and eukaryotes [23]. The magnitude and sign of the slope are specific for various taxa and chromosomes. Such a behavior proves a rule similar to Chargaff's rule for the distribution of nucleotides [33]—a statistics of the first order, but reveals a statistical regularity of the succession of the nucleotides—a statistics of the second order. As can be seen from equation (1.11) in Section 1.4, this rule states that the difference between the frequencies of positive and negative nucleotide-to-nucleotide transitions along a strand of a chromosome tends to be small, constant, and taxon & chromosome specific. As an immediate practical use of the unwrapped phase quasilinearity rule, the compliance of a certain contig with the large scale regularities of the chromosome to which it belongs can be used for spotting errors and exceptions.

1.1.2. Cumulated phase piecewise linearity in prokaryotes

Another significant result is that the cumulated phase has an approximately piecewise linear variation in prokaryotes, while drifting around zero in eukaryotes. The breaking points of the cumulated phase in prokaryotes can be put in correspondence with the limits of the chromosome "replichores": the minima with the origins of the replication process, and the maxima with its termini.

The existence of large scale regularities, up to the scale of entire chromosomes, supports the view that extragene DNA sequences, which do not encode proteins, still play significant functional roles. Moreover, the fact that these regularities apply to both coding and noncoding regions of DNA molecules indicates that these functionalities are also at the scale of the entire chromosomes. The unwrapped and cumulated phases can be linked to molecule potentials produced by unbalanced hydrogen bonds and can be used to describe "lateral" interaction of a given DNA segment with proteins and with other DNA segments in processes like replication,

transcription, or crossover. An example of such a process is the movement of DNA polymerase along a DNA strand during the replication process, by operating like a "Brownian machine" that converts random molecule movements into an ordered gradual advance.

1.1.3. Linearity of the cumulated phase for the reoriented exons in prokaryotes

A yet other important result is the finding that the cumulated phase becomes linear for the genomic signals corresponding to the sequences obtained by concatenating the coding regions of prokaryote chromosomes, after reorienting them in the same positive direction. This is a property of both circular and linear prokaryote chromosomes, but is not shared by most plasmids. This "hidden linearity" of the cumulated phase suggests the hypothesis of an ancestral chromosome structure, which has evolved into the current diversity of structures, under the pressure of processes linked to species separation and protection.

The rest of this chapter presents the vector and complex representations of nucleotides, codons, and amino acids that lead to the conversion of symbolic genomic sequences into digital genomic signals and presents some of the results obtained by using this approach in the analysis of large scale properties of nucleotide sequences, up to the scale of whole chromosomes.

Section 1.2 briefly describes aspects of the DNA molecule structure, relevant for the mathematical representation of nucleotides. Section 1.3 presents the vector (3D, tetrahedral) and the complex (2D, quadrantal) representations of nucleotides (Section 1.3.1), codons, and amino acids (Section 1.3.2). It is shown that both the tetrahedral and the quadrantal representations are one-to-one mappings, which contain the same information as the symbolic genomic sequences. Their main advantage is to reveal hidden properties of the genetic code and to conveniently represent significant features of genomic sequences.

Section 1.4 presents the phase analysis of genomic signals for nucleotide sequences and gives a summary of the results obtained by using this methodology. The study of complex genomic signals using signal processing methods facilitates revealing large scale features of chromosomes that would be otherwise difficult to find.

Based on the phase analysis of complex genomic signals, Section 1.5 presents a model of the "patchy" longitudinal structure of chromosomes and advances the hypothesis that it derives from a putative ancestral highly ordered chromosome structure, as a result of processes linked to species separation and specificity protection at molecular level. As mentioned above, it is suggested that this structure is related to important functions at the scale of chromosomes.

In the context of operating with a large volume of data at various resolutions and visualizing the results to make them available to humans, the problem of data representability becomes critical. Section 1.6 presents a new approach to this problem using the concept of data scattering ratio on a pixel. Representability analysis

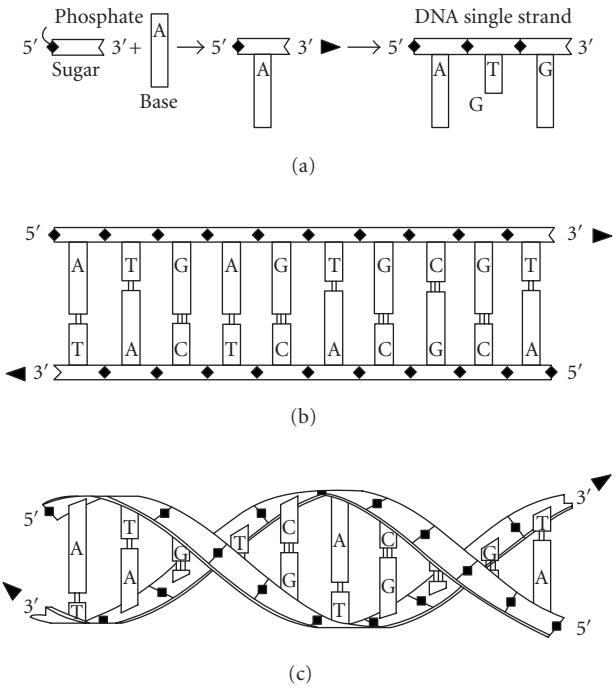


Figure 1.1. Schematic model of the DNA molecule.

is applied to several cases of standard signals and genomic signals. It is shown that the variation of genomic data along nucleotide sequences, specifically the cumulated and unwrapped phase, can be visualized adequately as simple graphic lines for low and large scales, while for medium scales (thousands to tens of thousands of base pairs), the statistical descriptions have to be used.

1.2. DNA double helix

This section gives a brief summary of the structure, properties, and functions of DNA molecules, relevant to building mathematical representations of nucleotides, codons, and amino acids and in understanding the conditions to be satisfied by the mappings of symbolic sequences to digital signals. The presentation is addressed primarily to readers with an engineering background, while readers with a medical or biological training can skip this section.

The main nucleic genetic material of cells is represented by DNA molecules that have a basically simple and well-studied structure [34]. The DNA double helix molecules comprise two antiparallel intertwined complementary strands, each a helicoidally coiled heteropolymer. The repetitive units are the nucleotides, each consisting of three components linked by strong covalent bounds: a monophosphate group linked to a sugar that has lost a specific oxygen atom—the deoxyribose, linked in turn to a nitrogenous base, as schematically shown in Figure 1.1

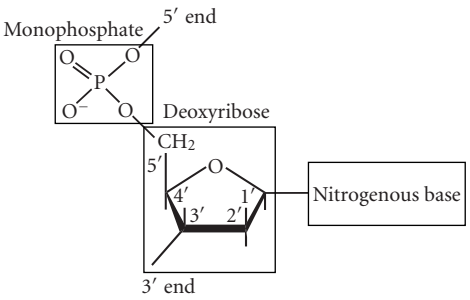


Figure 1.2. Structure of a nucleotide.

and detailed in Figure 1.2. Only four kinds of nitrogenous bases are found in DNA: thymine (T) and cytosine (C)—which are pyrimidines, adenine (A) and guanine (G)—which are purines. As can be seen in Figures 1.3 and 1.5, purine molecules consist of two nitrogen-containing fused rings (one with six atoms and the other with five), while pyridmidine molecules have only a six-membered nitrogen-containing ring. In a ribonucleic acid (RNA) molecule, apart of the replacement of deoxyribose with ribose in the helix backbone, thymine is replaced by uracil (U), a related pyrimidine. As shown in Figure 1.3, along the two strands of the DNA double helix, a pyrimidine in one chain always faces a purine in the other, and only the base pairs T−A and C−G exist. As a consequence, the two strands of a DNA helix are complementary, store the same information, and contain exactly the same number of A and T bases and the same number of C and G bases. This is the famous first Chargaff's rule [33], found by a chemical analysis of DNA molecules, before the crucial discovery of the double helix structure of DNA by Watson and Crick [34], and fully confirmed by this model. The simplified model in Figure 1.1 shows schematically how the nucleotides are structured, the single and double stranded DNA molecules, and gives a sketchy representation of the DNA secondary structure—the double helix resulting from the energy minimization condition. The figure does not show other significant features of the DNA longitudinal structure, such as the major and minor grooves. The hydrogen bonds within the complementary base pairs keep the strands together. When heating double stranded DNA at temperatures around 90°C, the hydrogen bonds melt and the two strands separate, resulting in "DNA denaturation." If lowering again the temperature, the reverse process—called "DNA renaturation"—reestablishes the double helix structure. The pair A−T contains only two hydrogen bonds, while the couple C−G contains three hydrogen bonds, so that the second link is stronger. This is reflected in an increased melting temperature for DNA molecules with a higher C-G concentration. Along each chain, there is a well-defined positive direction, given by the 5′ to 3′ direction in which the strand can grow by the addition of new nucleotides. The growth of a DNA chain is quite a complex process requiring the fulfillment of several conditions, from which we mention only the most important few. The normal process of growing a new DNA single-chain molecule is

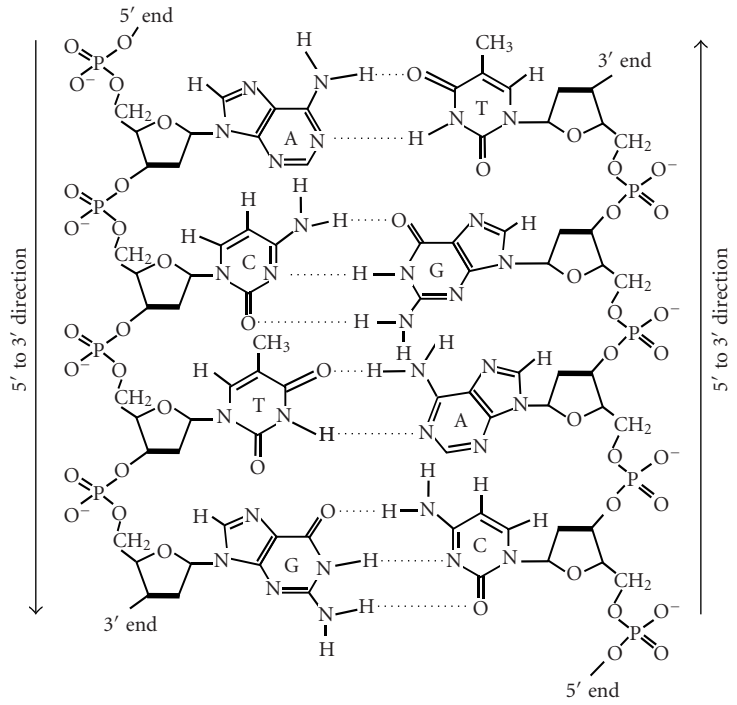


Figure 1.3. The chemical model of a double-stranded DNA molecule.

the replication, in which an existing (complementary) strand is used as a template, along which moves the DNA polymerase—the enzyme that performs the replication, adding to the growing chain nucleotides complementary to the ones in the template. A primer is also required; that is, the DNA polymerase can only prolong an already existing strand, by interacting with both the template strand and the strand to which it adds the nucleotide. The replication process consumes energy, so that the molecules that are needed by DNA polymerase to perform the addition of a nucleotide to the chain are not directly the nucleosine monophosphates, the monomers in the DNA strand, but the nucleosine triphosphates, which contain three phosphate groups and have the necessary free energy stored in the two phosphoanhydride bonds. Figure 1.4 gives the chemical model of adenosine triphosphate (ATP), the nucleosine triphosphate needed to add an adenine nucleotide to a DNA strand. The energy is released by the hydrolysis of the phosphoanhydride bonds and the loss of the two additional phosphate groups. This mechanism is so successful that nature uses ATP molecules not only for the replication of DNA but also for any other biochemical process that requires additional energy, ATP being the "molecular currency" of intracellular energy transfer. In the synthesis of nucleic acids, the ATP to AMP conversion mechanism imposes the 5′ to 3′ positive direction for the growth of DNA strands.

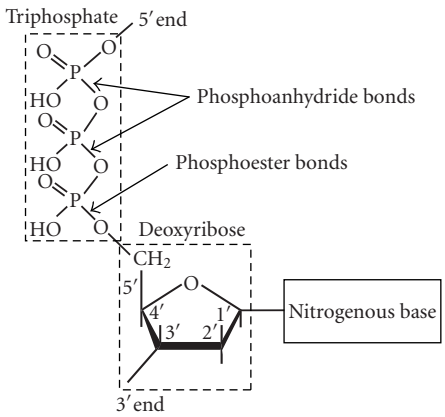


Figure 1.4. The chemical structure of ATP, precursor of the adenoside (adenosine monophosphate)—one of the nucleotides, and the most ubiquituos source of biochemical energy at molecular level.

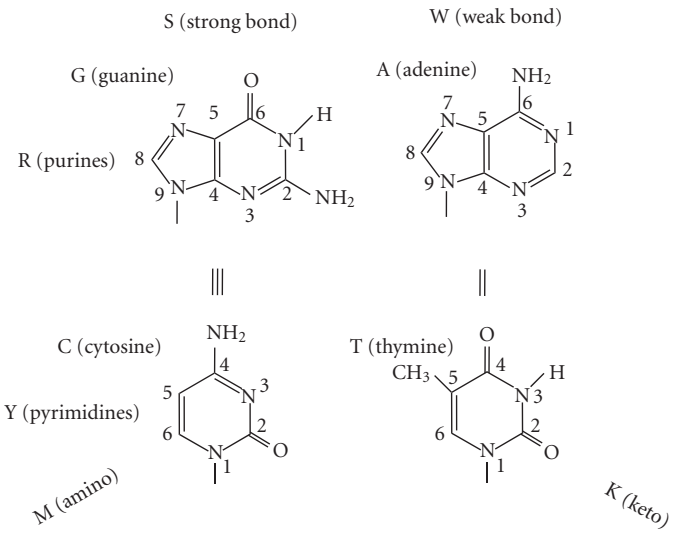


Figure 1.5. Class structure of nitrogenous bases.

The entities in the nucleotide chains that encode polypeptides, that is, specify the primary structure of proteins, are called genes. Genes are divided into exons—coding regions, interrupted by introns—noncoding regions. According to the current GenBank statistics [2], exons in the human genome account for about 3% of the total DNA sequence, introns for about 30%, and intergenic regions for the remaining 67%. Different methodologies produce different results in what concerns the number and size of the coding and noncoding regions. Based on mRNA and EST (Expressed Sequence Tags) studies, human genes contain on the

Table 1.1. The genetic code.

First position in codon		Second position in codon: T			C			A			G			Third position in codon
	T	TTT	Phe	[F]	TCT	Ser	[S]	TAT	Tyr	[Y]	TGT	Cys	[C]	T
		TTC	Phe	[F]	TCC	Ser	[S]	TAC	Tyr	[Y]	TGC	Cys	[C]	C
		TTA	Leu	[L]	TCA	Ser	[S]	TAA	*Ter*	[*end*]	TGA	*Ter*	[*end*]	A
		TTG	Leu	[L]	TCG	Ser	[S]	TAG	*Ter*	[*end*]	TGG	Trp	[W]	G
	C	CTT	Leu	[L]	CCT	Pro	[P]	CAT	His	[H]	CGT	Arg	[R]	T
		CTC	Leu	[L]	CCC	Pro	[P]	CAC	His	[H]	CGC	Arg	[R]	C
		CTA	Leu	[L]	CCA	Pro	[P]	CAA	Gln	[Q]	CGA	Arg	[R]	A
		CTG	Leu	[L]	CCG	Pro	[P]	CAG	Gln	[Q]	CGG	Arg	[R]	G
	A	ATT	Ile	[I]	ACT	Thr	[T]	AAT	Asn	[N]	AGT	Ser	[S]	T
		ATC	Ile	[I]	ACC	Thr	[T]	AAC	Asn	[N]	AGC	Ser	[S]	C
		ATA	Ile	[I]	ACA	Thr	[T]	AAA	Lys	[K]	AGA	Arg	[R]	A
		ATG	Met	[M]	ACG	Thr	[T]	AAG	Lys	[K]	AGG	Arg	[R]	G
	G	GTT	Val	[V]	GCT	Ala	[A]	GAT	Asp	[D]	GGT	Gly	[G]	T
		GTC	Val	[V]	GCC	Ala	[A]	GAC	Asp	[D]	GGC	Gly	[G]	C
		GTA	Val	[V]	GCA	Ala	[A]	GAA	Glu	[E]	GGA	Gly	[G]	A
		GTG	Val	[V]	GCG	Ala	[A]	GAG	Glu	[E]	GGG	Gly	[G]	G

average 3 and 10 exons, respectively, having an average length of 631 bp/262 bp and being separated by introns with average length 6,106 bp/5,420 bp. But there is a very large dispersion, with exon length ranging from just 1 bp/6 bp, up to 12,205 bp/17,105 bp. Minimum intron length is 17 bp/1 bp, while the maximum value reaches 482,576 bp/1,986,943 bp. Protein coding regions are rich in C and G, while intergene (noncoding) regions are rich in T and A.

Protein coding is governed by the genetic code that gives the mapping of codons—triplets of successive nucleotides in the corresponding reading frame in the exons—to the 20 amino acids found in the polypeptide chains and to the terminator that marks the end of an encoding segment. The genetic code is universal, applying to all known nuclear genetic material, DNA, mRNA, and tRNA, and encompasses animals (including humans), plants, fungi, bacteria, and viruses, with only small variations in mitochondria, certain eubacteria, cilliate, fungi, and algae [2]. From Table 1.1, which gives the standard genetic code, it can be seen that there is a large redundancy (degeneration) of the genetic code, as there are $4^3 = 64$ codons to specify only 21 distinct outputs. The redundancy is distributed unevenly among the outputs: there are amino acids encoded by one (2 instances), two (9 instances), three (one instance), four (5 instances), or six (3 instances) distinct codons, while the terminator is encoded by three codons. Most genes start with the codon ATG that also encodes the amino acid methionine.

The codon—amino acid mapping comprises two steps: (1) the *transcription*, in which a specific enzyme, called transcriptase, copies a section of the DNA template into a complementary mRNA (messenger RNA) molecule, in the presence of a mixture of the four ribonucleotides (ATP, UTP, GTP, and CTP), and (2) the *translation*, in which the actual mapping of the codons in the mRNA to amino

acids is performed by ribosomes, after *slicing*—the editing of mRNA by the excision of all introns and the joining of all exons. Quite surprisingly, the number of nucleotides in an exon is not necessarily a multiple of three, that is, an exon does not necessarily comprise an integer number of codons. The amino acids for the protein are brought to the site by tRNA (transfer RNA) molecules, each having a nucleotide triplet which binds to the complementary sequence on the mRNA. Each of the 20 amino acids is brought by a specific tRNA. In fact, there are at least 23 tRNAs for the 20 amino acids, as it will be discussed in the following in relation with the representation of the genetic code. There is a sharp contrast between the deceivingly simple structure of DNA nucleotide chains—unbranched linear code written in a four-letter alphabet, and the overwhelming complexity of the protein 3D structure built of twenty amino acids. As mentioned, there are only about 30 000 genes in the human genome, but millions of proteins, many of them transitory. Nevertheless, the nucleotide chains and the proteins are the bearers of essentially the same genetic information.

1.3. Conversion of genomic sequences into genomic signals

The conversion of genomic sequences from the symbolic form given in the public genomic databases [1, 2] into digital genomic signals allows using signal processing procedures for processing and analyzing genomic data. We have investigated a large number of mappings of symbolic genomic data to digital genomic signals and we have compared how the structure of the genomic code was highlighted by the various representations and how the features of DNA sequences were revealed by the resulting digital signals [25, 26, 27, 28, 29, 30, 31, 32]. Such a representation has to be both truthful and unbiased. The mapping is truthful if all biologically relevant characteristics of the represented objects are expressed in corresponding mathematical properties of the samples in the resulting digital signal. The mapping is unbiased if none of the features belonging to the mapping itself, but without correspondent in the properties of the initial sequence, is introduced as artifacts. The representation must also be simple enough to allow fast and computationally effective conversion and to provide an output readable for a human operator. The last request favors representations with low dimensions of the output, preferably 1D or 2D. This section briefly presents the digital representation of nucleotides starting from the essential features of DNA sequences. A detailed study of the symbolic-to-digital conversion of genomic sequences can be found in [23].

1.3.1. Nucleotide representation

As schematically shown in Figure 1.5, there are three main dichotomies of the nitrogenous bases biochemical properties that allow arranging them in classes: (1) *molecular structure*—A and G are purines (R), while C and T are pyrimidines (Y); (2) *strength of links*—bases A and T are linked by two hydrogen bonds (W—weak bond), while C and G are liked by three hydrogen bonds (S—strong bond);

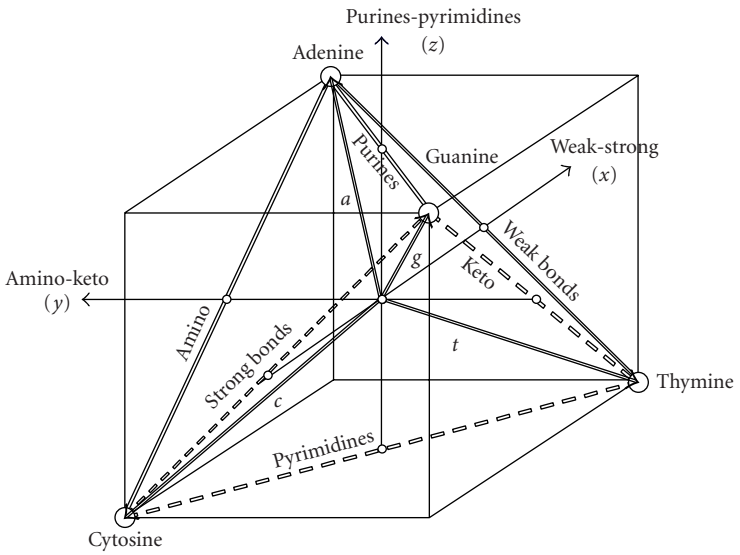


Figure 1.6. Nucleotide tetrahedron.

(3) *radical content*—A and C contain the amino (NH_3) group in the large groove (M class), while T and G contain the keto (C=O) group (K class).

To express the classification of the system of nucleotides in couples of pairs shown in Figure 1.5, we have proposed the nucleotide tetrahedral representation [24] shown in Figure 1.6. The nucleotides are mapped to four vectors symmetrically placed in the 3D space, that is, oriented towards the vertices of a regular tetrahedron. Each of the six edges corresponds to one of the classes comprising a pair of nucleotides. The representation is three dimensional and the axes express the differences "weak minus strong bonds," "amino minus keto," and "purines minus pyrimidines":

$$x = \mathrm{W} - \mathrm{S}, \qquad y = \mathrm{M} - \mathrm{K}, \qquad z = \mathrm{R} - \mathrm{Y}. \tag{1.1}$$

By choosing $\{\pm 1\}$ coordinates for the vertices of the embedding cube, the vectors that represent the four nucleotides take the simple form:

$$\begin{aligned} \vec{a} &= \vec{i} + \vec{j} + \vec{k}, \\ \vec{c} &= -\vec{i} + \vec{j} - \vec{k}, \\ \vec{g} &= -\vec{i} - \vec{j} + \vec{k}, \\ \vec{t} &= \vec{i} - \vec{j} - \vec{k}. \end{aligned} \tag{1.2}$$

This representation is fully adequate for well-defined sequences, when each entry is uniquely specified. Such sequences are given in the large integrative genomic databases, which provide a single curated standard sequence with respect to which single nucleotide polymorphisms (SNPs) or other variations are defined. But, when working with experimental data that can have ambiguous or multiple values for some entries in the sequence, caused by either noise, or by the true variability within the population for which the genome is sequenced, the IUPAC conventions [2] have to be used. Apart of the symbols for the nucleotides (A, C, G, T), IUPAC conventions include symbols for the classes mentioned at the beginning of this section (S, W, R, Y, M, K), as well as for classes comprising three nucleotides (B = {C, G, T} =∼ A, D = {A, G, T} =∼ C, H = {A, C, T} =∼ G, V = {A, C, G =∼ T}), or all four nucleotides (i.e., unspecified nucleotide, N). The corresponding vector representation is shown in Figure 1.7, in which the additional vectors are given by:

$$\begin{aligned}
\vec{w} &= \frac{\vec{a}+\vec{t}}{2} = \vec{i}, \\
\vec{s} &= \frac{\vec{c}+\vec{g}}{2} = -\vec{i}, \\
\vec{m} &= \frac{\vec{a}+\vec{c}}{2} = \vec{j}, \\
\vec{k} &= \frac{\vec{g}+\vec{t}}{2} = -\vec{j}, \\
\vec{r} &= \frac{\vec{a}+\vec{g}}{2} = \vec{k}, \\
\vec{y} &= \frac{\vec{c}+\vec{t}}{2} = -\vec{k}, \\
\vec{b} &= \frac{\vec{c}+\vec{g}+\vec{t}}{3} = -\frac{\vec{a}}{3}, \\
\vec{d} &= \frac{\vec{g}+\vec{t}+\vec{a}}{3} = -\frac{\vec{c}}{3}, \\
\vec{h} &= \frac{\vec{t}+\vec{a}+\vec{c}}{3} = -\frac{\vec{g}}{3}, \\
\vec{u} &= \frac{\vec{a}+\vec{c}+\vec{g}}{3} = -\frac{\vec{t}}{3}.
\end{aligned} \tag{1.3}$$

The dimensionality of the representation can be reduced to two, by projecting the nucleotide tetrahedron on an adequately chosen plane. This plane can be put in correspondence with the complex plane, so that a complex representation of the nucleotides is obtained. The choice of the projection plane is determined by the features that have to be conserved as being most relevant in the given context. For the study of large scale features of DNA sequences and for other similar problems, we have found that the separation within the amino-keto classes is less significant as compared to the strong-weak and purine-pyrimidine dichotomies.

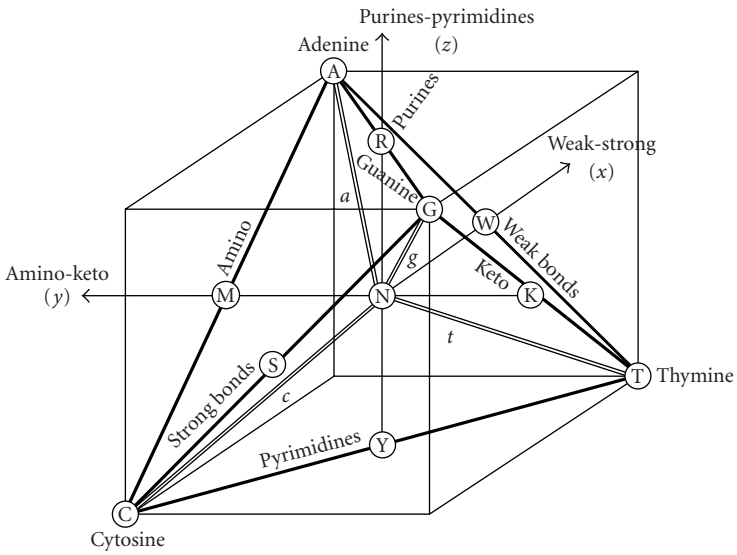


Figure 1.7. IUPAC symbols 3D representation.

This corresponds to the projection on the plane xOz and expresses the S−W and Y−R dichotomies. The resulting complex quadrantal representation of the nucleotides is given in Figure 1.8, in which the pairs of nucleotides are grouped in the six above-mentioned classes, while the corresponding analytical expressions are given in the equations:

$$\begin{aligned} a &= 1 + j, \\ c &= -1 - j, \\ g &= -1 + j, \\ t &= 1 - j. \end{aligned} \tag{1.4}$$

In this representation the complementarity of the pairs of bases A−T and C−G, respectively, is expressed by the symmetry with respect to the real axis (the representations are complex conjugates: $t = a^*, g = c^*$), while the purine/pyrimidine pairs have the same imaginary parts. We have investigated several other representations, but the complex representation given by (1.4) has shown most advantages.

It should be noted that both the vector (3D, tetrahedral) and the quadrantal (2D, complex) nucleotide representations shown above, as well as the real (1D) representation to be discussed in the following, are one-to-one mappings that allow rebuilding the initial symbolic sequence from the vector, complex or real genomic signals. The information is wholly conserved and so are all the features and properties of the initial sequences. Nevertheless, there are significant differences in what concerns the expression of the various significant features and how accessible

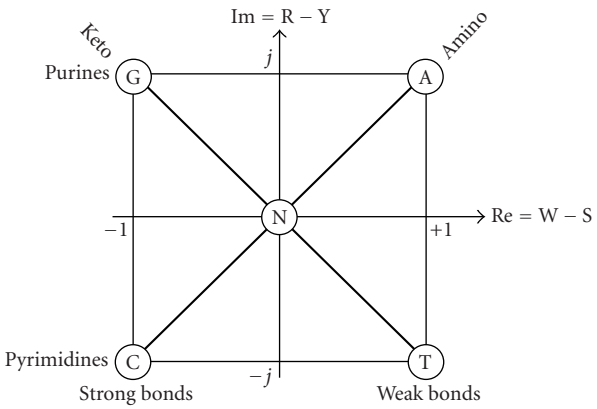


Figure 1.8. Nucleotide complex representation.

or directly readable for a human agent these features become. As in many other cases, a good representation of a system is an important part in solving various problems related to that system.

The projection of the vectors in Figure 1.7 on the same xOz plane provides the complex representation of the IUPAC symbols given in Figure 1.9 and expressed by the equations:

$$\begin{aligned} w &= 1, \\ y &= -j, \\ s &= -1, \\ r &= j, \\ k &= m = n = 0, \\ d &= \frac{1}{3}(1+j), \\ h &= \frac{1}{3}(1-j), \\ b &= \frac{1}{3}(-1-j), \\ v &= \frac{1}{3}(-1+j). \end{aligned} \tag{1.5}$$

As mentioned above, it is possible to further reduce the dimensionality of the representation of nucleotide, codon, and amino acid sequences by using a real one-dimensional mapping. The digits $\{0, 1, 2, 3\}$ can be attached to the four nucleotides. The three-base codons are interpreted as three-digit numbers written in base four, that is, the codons along the DNA strands are mapped to the numbers $\{0, 1, 2, \ldots, 63\}$. Actually, a whole DNA sequence can be seen as a very large number written in base four. Nevertheless, it corresponds better to the biological reality to interpret each codon as a distinct sample of a digital genomic signal distributed

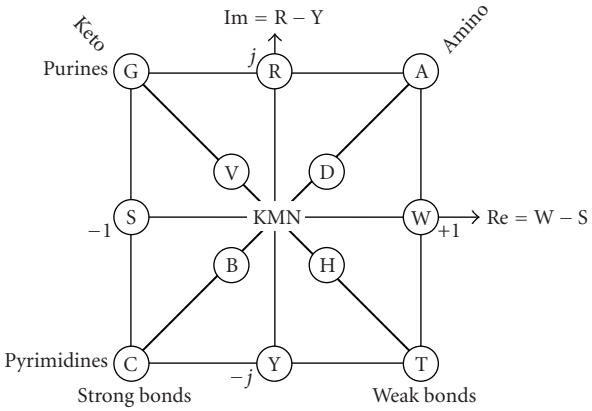


Figure 1.9. IUPAC symbols complex representation.

Table 1.2. Real representation of nucleotides to digits in base four.

Pyrimidines	Purines
Thymine = T = 0	Adenine = A = 2
Cytosine = C = 1	Guanine = G = 3

along the DNA strand. There are 4! = 24 choices for attaching the digits 0–3 to the bases A, C, G, T. The optimal choice given in Table 1.2 results from the condition to obtain the most monotonic mapping of the codons 0–63 to the amino acids plus the terminator 0–20, leading to best autocorrelated intergene genomic signals [23].

1.3.2. Codon and amino acid representation

The tetrahedral (3D), complex (2D), and real (1D) representations of nucleotides can be naturally extended for the representation of codons and amino acids.

A codon consists of a sequence of three nucleotides:

$$X = B_2B_1B_0, \quad B_i \in \{\mathrm{A}, \mathrm{C}, \mathrm{G}, \mathrm{T}\};\ i = 0, 1, 2, \tag{1.6}$$

situated in a coding area of a DNA molecule, that is, in an exon, and having the start aligned to the correct open reading frame (ORF). There are six possible ORFs, three in each direction of the DNA strand, shifted with a nucleotide from each other.

The codon can be seen as a word of three letters, taken from a four-letter alphabet. The codon can also be seen as a number written in a certain base, using the four digits B_i. For the vectorial (tetrahedral) representation of nucleotides, we have chosen the base two and the four-vector digits having the values given in

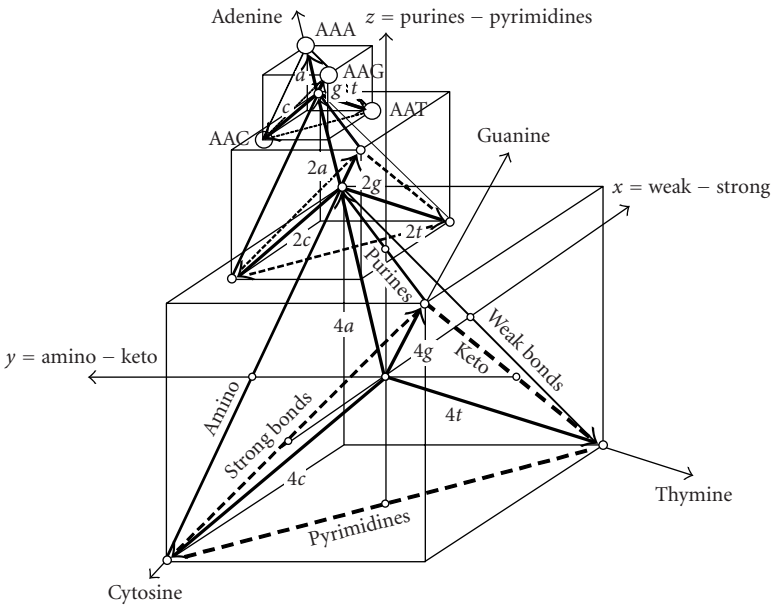


Figure 1.10. Example of the vector representation of codons.

equation (1.2). Correspondingly, the codon X is mapped to the vector:

$$\vec{x} = 2^2\vec{b}_2 + 2^1\vec{b}_1 + 2^0\vec{b}_0, \quad \vec{b}_i \in \{\vec{a}, \vec{c}, \vec{g}, \vec{t}\};\ i = 0, 1, 2. \tag{1.7}$$

This is a natural extension of the concept of *numeration system* to vectorial (and complex) numbers. The vectorial conversion procedure is repeated for each of the three nucleotides in a codon, treating them as digits of a three-digit number written in base two: the vector corresponding to the third, that is, the last nucleotide in the codon (the least significant digit) is multiplied by 1, the vector corresponding to the second base in the codon by 2, and the vector corresponding to the first base of the codon (the most significant digit) by $2^2 = 4$. This results in the codon vectorial representation illustrated in Figure 1.10 for the special cases of the codons AAA $(4\vec{a}+2\vec{a}+\vec{a})$ and AAG $(4\vec{a}+2\vec{a}+\vec{g})$—encoding lysine, and AAC $(4\vec{a}+2\vec{a}+\vec{c})$ and AAT $(4\vec{a}+2\vec{a}+\vec{t})$—encoding asparagine. Applying the same rule for all the 64 codons, the codon tetrahedral representation in Figure 1.11 is obtained [24]. The first nucleotide in a codon selects one of the four *first-order* 16-*codon tetrahedrons* that form together the *zero-order tetrahedron* of the overall genetic code, the second nucleotide selects one of the *second-order* 4-*codon tetrahedrons* that compose the already selected first-order tetrahedron and, finally, the third nucleotide identifies one of the vertices. In this way, each of the codons is attached to one of the vertices in a resulting three-level fractal-like tetrahedron structure. Taking into account the codon-to-amino acid mapping imposed by the genetic code, the amino acids encoded by the codons can be assigned to one or

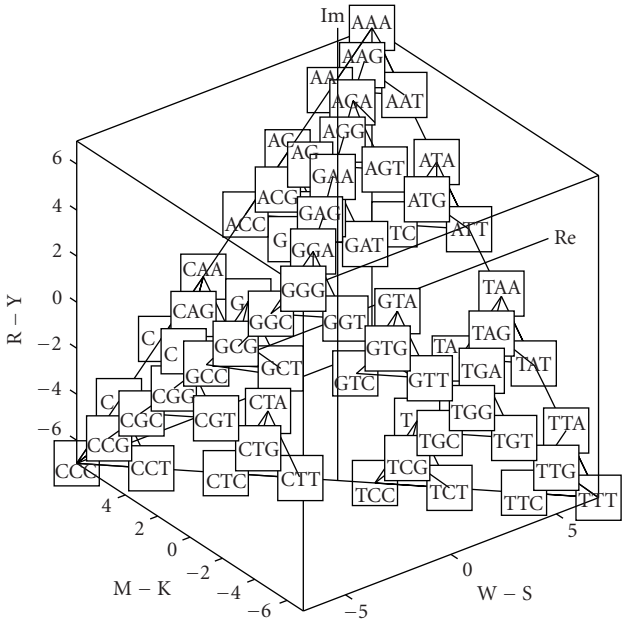


Figure 1.11. Codon tetrahedral representation.

several of the 64 vertices, as shown in Figure 1.12. It turns out that the tetrahedron representation of the genomic code, as well as the mathematical descriptions based on it, reflects better the metric structure of the genomic code. Specifically, the codons that correspond to the same amino acid are mapped in neighboring points, so that related codons are clustered. Moreover, the degeneration is basically restricted to the second-order tetrahedrons and most pairs of interchangeable nucleotides are distributed on the edges parallel to the pyrimidines and purines directions. The tetrahedron representation has also the advantage to naturally determine putative ancestral coding sequences by the simple passage to a lower-level tetrahedron. Thus, the tetrahedron representation grasps some essential features of the genetic code which appear as symmetries and regularities of the resulting 3D image. To make the nucleotide and codon sequences easy to read for a human observer, the three axes of the representation space can be assigned to the three basic color components of the RGB—red, green, blue system [35]. Consequently, each point in the representation space—each nucleotide in the case of Figure 1.6, or each IUPAC symbol in the case of Figure 1.7, corresponds to a distinct hue. This approach is useful for the fast visual exploration of DNA sequences at the nucleotide level and can be extended at the codon (Figure 1.11) and amino acid levels (Figure 1.12). The superposition of the codon tetrahedron and of the amino acid tetrahedron, as shown in Figure 1.13, is the 3D equivalent of a periodic table for the genomic code. This representation gives a better image of the regularities of the genomic code and allows sensing of some aspects of its early evolution before

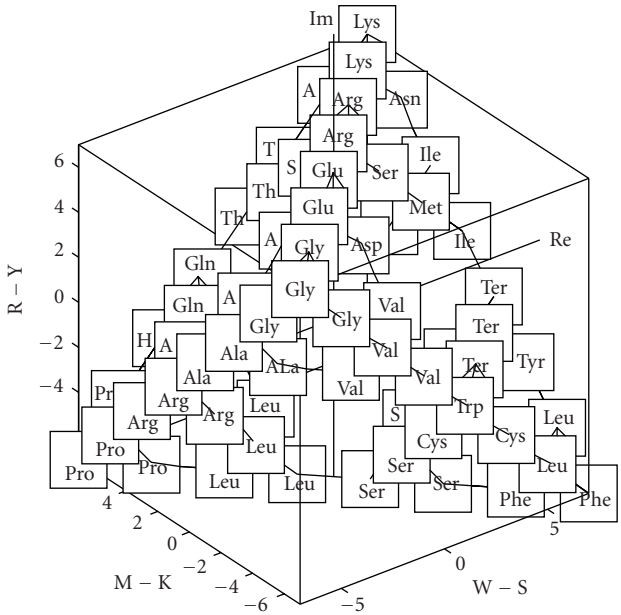


Figure 1.12. Amino acid tetrahedral representation.

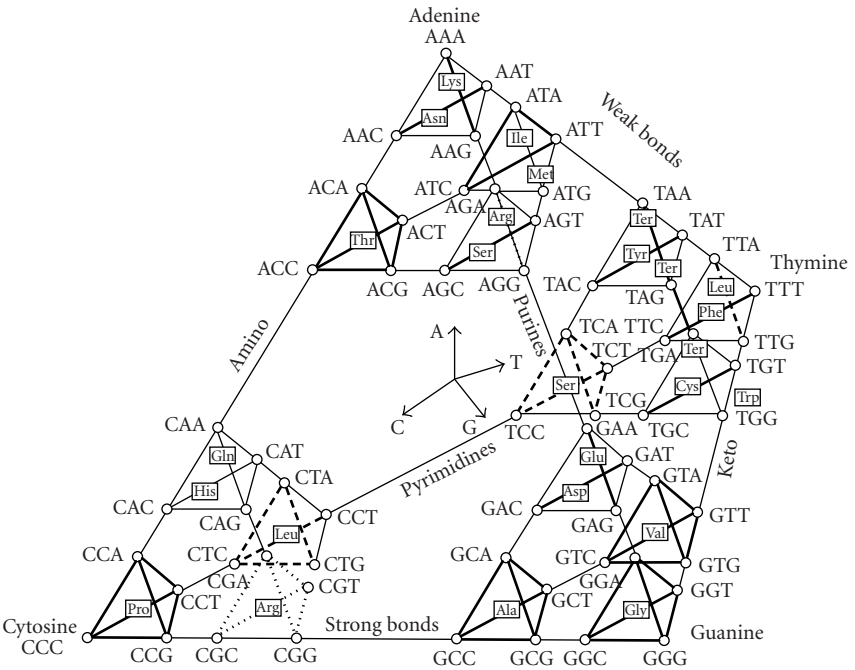
reaching the current frozen state. It is especially remarkable that the different representations of an amino acid, resulting from the redundancy of the genetic code, are mapped in neighboring points of the codon tetrahedron, with the exception of the three instances of amino acids degenerated of order six, for which none of the investigated mapping can obtain the contiguity of the representations. It must be mentioned, though, that for each of these amino acids, there are two distinct tRNA (giving the total of 23 tRNAs for the 20 amino acids), each encoded by the codons of the same cluster.

A complex representation of the codons can be obtained in a similar way, by considering the codon as a three-digit number in base two and by using the complex representations (1.4) of its nucleotides as the digits:

$$x = 2^2 b_2 + 2^1 b_1 + 2^0 b_0, \quad b_i \in \{a, c, g, t\};\ i = 0, 1, 2. \tag{1.8}$$

Again, relation (1.8) results from (1.7) by the projection on the xOz plane, and by taking Ox as the real axis and Oz as the imaginary one. Relation (1.8) can also be seen as representing the nucleotides in two (bipolar) binary systems, with mutually orthogonal complex units.

For methionine, to which corresponds the codon ATG that starts any gene, the vector representation is $\vec{M} = 4\vec{a} + 2\vec{t} + \vec{g} = 5\vec{i} + \vec{j} + 3\vec{k}$, while the complex representation is $M = 4a + 2t + g = 5 + 3j$. For the example in Figure 1.10, the corresponding complex representation is given in Figure 1.14.



Ala - alanine	Gly - glycine	Pro - proline
Arg - arginine	Ile - isoleucine	Ser - serine
Asn - asparagine	His - histidine	Thr - thereonine
Asp - aspartic acid	Leu - leucine	Trp - tryptophan
Cys - cysteine	Lys - lysine	Tyr - tyrosine
Gln - glutamine	Met - methionine	Val - valine
Glu - glutamic acid	Phe - phenylalanine	Ter - terminator

Figure 1.13. Genetic code vectorial representation.

Applying the same method for the 64 codons, the complete mapping of all the codons to the complex plane given in Figure 1.15 is obtained [23]. This is the projection on the xOz plane of the codon tetrahedron in Figure 1.11. Figure 1.16 shows the mapping of the amino acids to the complex plane and can be obtained either by applying the genetic code on the complex representation of the codons in Figure 1.11, or by projecting the amino acid tetrahedron in Figure 1.12 on the xOz plane. The superposition of the codon and amino acid complex representations in Figure 1.17 gives a complex (2D) representation of the genomic code. The clustering of the codons corresponding to the same tRNAs is obvious, and this corresponds in 17 of the cases with the clustering of all the codons corresponding to the same amino acid in a single contiguous domain. Only for arginine, leucine, and serine, each of them encoded by six codons, the corresponding domains are splitted in two noncontiguous subdomains comprising four and, respectively, two codons. It is interesting to mention that the clustering refers not

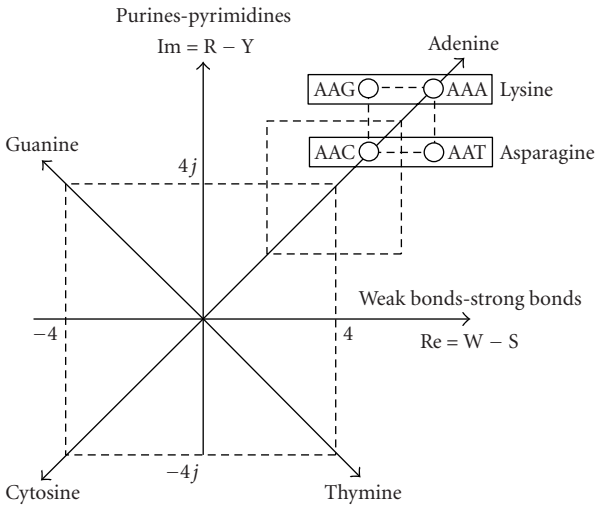


Figure 1.14. Codon complex representation.

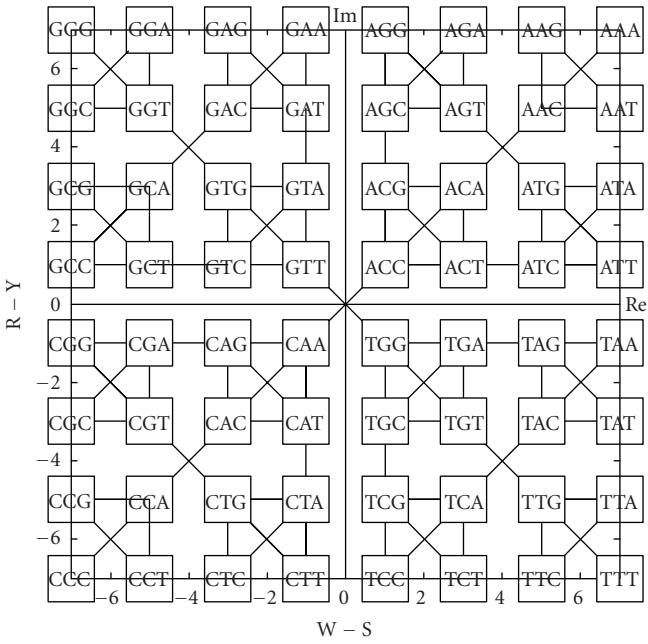


Figure 1.15. Mapping of the codons on the complex plane.

only to the codons, but also to the features of the amino acids. Amino acids with similar properties (e.g., which cluster on state transition probability) tend to be neighbors in the complex representation of the genomic code in Figure 1.17.

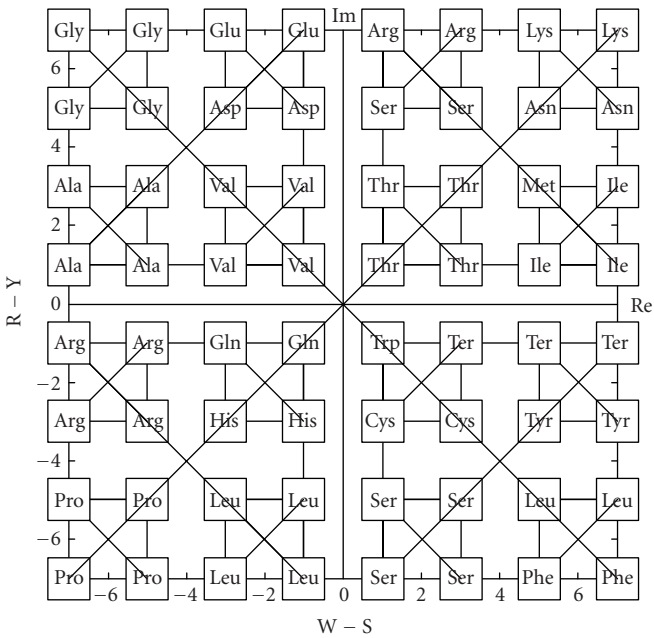


Figure 1.16. Mapping of the amino acids on the complex plane.

As mentioned above, it is possible to further reduce the dimensionality of the representation of nucleotide, codon, and amino acid sequences by using a real one-dimensional mapping. Table 1.3 gives the mapping of the digital codons to the numerical codes of the amino acids. The numerical values of the codons result from the base-four values of the nucleotides given in Table 1.2 and from the "nucleotide digits" in each codon. The numerical codes assigned to the amino acids result from the order of their first reference when gradually increasing the values of the codons from 0 to 63. By convention, the code zero is assigned to the terminator. As can be seen in the representations of the genetic code in Table 1.1 and in Figures 1.12, 1.13, 1.16, and 1.17, there are only two nondegenerated (one codon—one amino acid) mappings—for tryptophan and methionine, but nine double, one triple, five quadruple, and three sextuple degenerations, plus the three codons corresponding to the terminator. The minimum nonmonotonic dependency has only four reversals of the ascending order: for a terminator sequence and for the three instances of sextuple degeneration (leucine, serine, and arginine). An exhaustive search for all the 24 possible correspondences of the nucleotides to the digits 0–3 has shown that there does not exist a more monotonic mapping. The proposed mapping gives a piecewise constant function, with only the three mentioned reversals of the order, as shown in Table 1.3 and in Figure 1.18.

The reference to the various real and complex representations of the nucleotides can be simplified by using the pair of indices (p, q) as described in details in [23]. The index p specifies the *nucleotide permutations* and takes values from

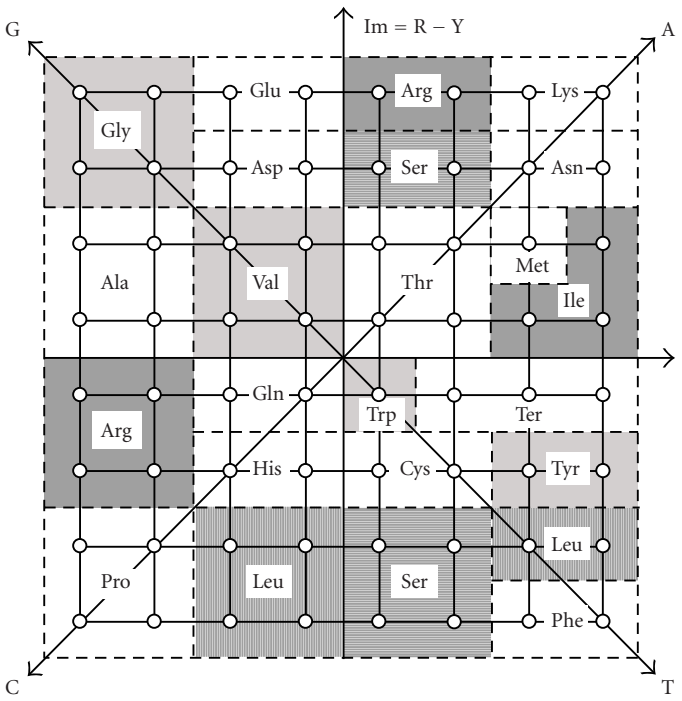


Figure 1.17. Genetic code complex representation.

1 to 24. The index q is used to specify the *representation type* and has the values: $q = 0$ for the real representation, $q = 1$ for a representation defined by the mapping of the nucleotides to pure real/pure imaginary numbers, and $q = 2$ for the mapping of nucleotides to quadrantally symmetric complex numbers, as defined by equation (1.4) and Figure 1.8 (for $p = 1$).

Despite the fact that the real representations of nucleotides described above are also one-to-one mappings, having an exact inverse, thus conserving the whole information in the initial symbolic sequence, the vectorial or complex representations are better fitted to reveal the basic features of the genomic sequences by their emphasis on the classes of nucleotides. Unfortunately, the simpler-to-handle real representations seem to be biased, as they induce some additivity of the properties of the mathematical representation, which does not have a direct correspondence in the nucleotide biochemical properties. In the following sections, we will present results obtained by using the complex (2D) and vectorial (3D) representations.

Complex representations have the advantage of expressing some of the biochemical features of the nucleotides in mathematical properties of their representations. For instance, the complementarity of the pairs of bases A−T, G−C is expressed by the fact that their representations are complex conjugates, while purines and pyrimidines have the same imaginary parts and opposite sign real parts. As already discussed, the complex representation of the codons and the amino acids

Table 1.3. Optimal correspondence of real numerical codons to amino acids.

Digital codon	Amino acid code	Long name	Short name	Symbol
10, 11, 14	0	Terminator	Ter	[end]
0, 1	1	Phenylalanine	Phe	[F]
2, 3, 16, 17, 18, 19	2	Leucine	Leu	[L]
4, 5, 6, 7, 44, 45	3	Serine	Ser	[S]
8, 9	4	Tyrosine	Tyr	[Y]
12, 13	5	Cysteine	Cys	[C]
15	6	Tryptophan	Trp	[W]
20, 21, 22, 23	7	Proline	Pro	[P]
24, 25	8	Histidine	His	[H]
26, 27	9	Glutamine	Gln	[Q]
28, 29, 30, 31, 46, 47	10	Arginine	Arg	[R]
32, 33, 34	11	Isoleucine	Ile	[I]
35	12	Methionine	Met	[M]
36, 37, 38, 39	13	Thereonine	Thr	[T]
40, 41	14	Asparagine	Asn	[N]
42, 43	15	Lysine	Lys	[K]
48, 49, 50, 51	16	Valine	Val	[V]
52, 53, 54, 55	17	Alanine	Ala	[A]
56, 57	18	Aspartic acid	Asp	[D]
58, 59	19	Glutamic Acid	Glu	[E]
60, 61, 62, 63	20	Glycine	Gly	[G]

shown in Figures 1.15 and 1.16 results simply from the projection of the codon and amino acid tetrahedrons in Figures 1.11 and 1.12 on the xOz plane. This leads naturally to the complex representation of the genetic code in Figure 1.17 and allows representing DNA sequences by complex signals at the levels of nucleotides, codons, and amino acids. It can be noticed that this complex mapping conserves the meaning of the distances between the codons, as resulting from the genetic code. Specifically, codons corresponding to the same amino acid are clustered in contiguous regions of the complex plane. From the frequency of the amino acids in the proteins, it results that the genetic code has some of the characteristics of Huffman (entropy) coding. Higher redundancy (degeneracy) in the encoding could correspond to primitive, older amino acids, while low redundancy, meaning a higher local resolution of the genetic code, could correspond to more recent amino acids. This hypothesis allows building models of ancestral proteins in the early times before the freezing of the genomic code.

Complex values can be attached in various ways to the amino acids. One modality is to assign to a certain amino acid the average value over the whole area onto which it is mapped, taking into account the relative frequencies of occurrence of the different codons that correspond to the amino acid. It has been shown that the assigning of the complex values to the nucleotides and to the amino acids can be adapted to various tasks. For instance, the optimum values for detecting the exons are different from the optimum ones for detecting the reading frames [35]. This gives the flexibility needed for targeting the approach to each application.

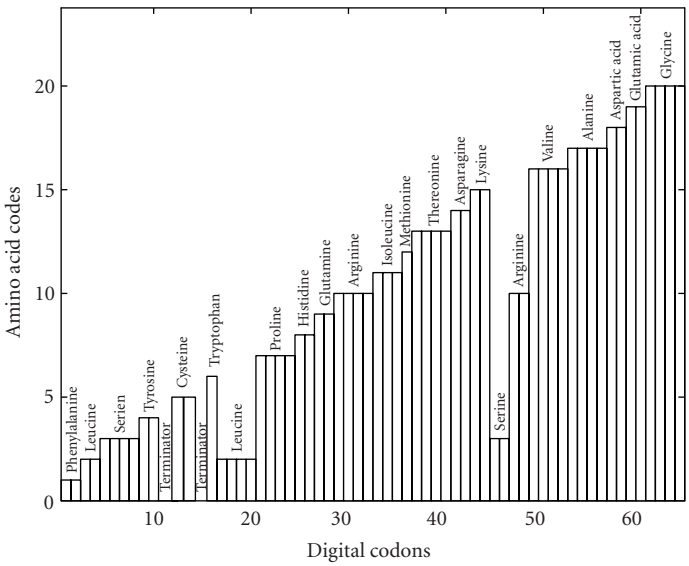


Figure 1.18. Optimal (minimally nonmonotonic) correspondence of numerical codons to amino acid codes.

For the analysis of large scale DNA features, only the nucleotide to complex mapping given in equations (1.4) and (1.5) and Figures 1.8 and 1.9 will be used.

1.4. Phase analysis of DNA sequences

All available complete genomes or available sets of contigs for eukaryote and prokaryote taxa have been downloaded from the GenBank [2] database of NIH, converted into genomic signals by using the mapping given in equation (1.4). The signals have been analyzed focussing on the extraction of large scale features of DNA sequences, up to the scale of whole chromosomes. Such properties transcend the differences between the coding (exons) and noncoding (introns) regions of DNA sequences, and refer to properties and functions of the chromosomes as whole entities. Several tools have been developed for this type of analysis, some also useful for local feature extraction, and have been presented elsewhere [23, 24, 25, 26, 27, 28, 29, 30, 31, 32]. This section is devoted to the phase analysis of the complex genomic signals, which revealed some interesting large scale DNA features that could be important for better understanding such functions of chromosomes like replication, transcription, and crossover.

1.4.1. Fundamentals of phase analysis

The *phase* of a complex number is a periodic magnitude: the complex number does not change when adding or subtracting any multiple of 2π to or from its phase. To remove the ambiguity, the standard mathematical convention restricts

the phase of a complex number to the domain $(-\pi, \pi]$ that covers only once all the possible orientations of the associated vector in the complex plane. For the genomic signals obtained by using the mapping defined in Figure 1.8 and in equation (1.4), the *phases* of the nucleotide representations can have only the values $\{-3\pi/4, \pi/4, \pi/4, 3\pi/4\}$ radians.

The *cumulated phase* is the sum of the phases of the complex numbers in a sequence from the first element in the sequence, up to the current element. For the complex representation (1.4), the cumulated phase at a certain location along a sequence of nucleotides has the value:

$$\theta_c = \frac{\pi}{4}[3(n_G - n_C) + (n_A - n_T)], \tag{1.9}$$

where n_A, n_C, n_G, and n_T are the numbers of adenine, cytosine, guanine, and thymine nucleotides in the sequence, from the first to the current location. Consequently, the slope s_c of the cumulated phase along the DNA strand at a certain location is linked to the frequencies of occurrence of the nucleotides around that location by the equation:

$$s_c = \frac{\pi}{4}[3(f_G - f_C) + (f_A - f_T)], \tag{1.10}$$

where f_A, f_C, f_G, and f_T are the nucleotide occurrence frequencies.

The *unwrapped phase* is the corrected phase of the elements in a sequence of complex numbers, in which the absolute value of the difference between the phase of each element in the sequence and the phase of its preceding element is kept smaller than π by adding or subtracting an appropriate multiple of 2π to or from the phase of the current element. The unwrapped phase eliminates the phase jumps introduced by the conventional restriction of the phase domain described above and allows observing the true global phase trends along a sequence. For the complex representation given in equation (1.4), the *positive transitions* A → G, G → C, C → T, T → A determine an increase of the unwrapped phase, corresponding to a rotation in the trigonometric sense by $\pi/2$, the *negative transitions* A → T, T → C, C → G, G → A determine a decrease, corresponding to a clockwise rotation by $-\pi/2$, while all other transitions are *neutral*. A distinction has to be made between the exactly (first type) neutral transitions A ↔ A, C ↔ C, G ↔ G, T ↔ T, for which the difference of phase is zero in each instance, so that the unwrapped phase does not change, and the "on average" (second type) neutral transitions A → C, C → A, G → T, T → G, for which the difference of phase is $\pm\pi$. Because of the bias introduced by the conventional restriction of the phase to the domain $(-\pi, \pi]$, which favors π over $-\pi$, the standard unwrapped phase function and the corresponding functions implemented in most commercial software mathematics libraries, which apply the basic convention for the phase mentioned above, attach $+\pi$ to all the "on average" neutral transitions. This would distort the unwrapped phase and even the cumulated phase, if using complex representations that include real negative numbers (which is not the case for equations (1.4)). To avoid this unwanted effect, two solutions have been used:

(1) for large genomic sequences, from millions to hundreds of millions of nucleotides, uniformly distributed small random complex numbers have been added to each nucleotide complex representation, so that phases and differences of phase close to $-\pi$ are equally probable with the phases close to π and the artificial drift of the unwrapped phase towards positive values has been eliminated, (2) primarily for medium or small sequences, for example, when studying virus genomes, but also for large and very large sequences, a custom unwrapped phase function has been used that attaches zero phase change for all neutral transitions.

The accuracy of both procedures has been thoroughly verified using artificial sequences. It has been found that any bias related to the conventional restriction of the phase domain, which could affect crisp data processed with the standard unwrapped phase function, has been eliminated.

For the complex representation (1.4), taking the precautions mentioned above, the unwrapped phase at a certain location along a sequence of nucleotides has the value:

$$\theta_u = \frac{\pi}{2}(n_+ - n_-), \tag{1.11}$$

where n_+ and n_- are the numbers of the positive and negative transitions, respectively. The slope s_u of the variation of the unwrapped phase along a DNA strand is given by the relation:

$$s_u = \frac{\pi}{2}(f_+ - f_-), \tag{1.12}$$

where f_+ and f_- are the frequencies of the positive and negative transitions.

An almost constant slope of the unwrapped phase corresponds to an almost helicoidal wrapping of the complex representations of the nucleotides along the DNA strand. The step of the helix, that is, the spatial period over which the helix completes a turn, is given by

$$L = \frac{2\pi}{s_u}. \tag{1.13}$$

As will be shown in the next subsection, such an almost linear increase of the unwrapped phase, corresponding to a counter clockwise helix, is a long-range feature of all chromosomes of *Homo sapiens*, *Mus musculus*, and of other animal eukaryotes, while an opposite winding is common in plants and prokaryotes. The trend is maintained over distances of tens of millions of bases and reveals a regularity of the second-order statistics of the distribution of the succession of the bases which is a new property, distinct of Chargaff's laws.

It must be noted that the cumulated phase is related to the statistics of the nucleotides, while the unwrapped phase id related to the statistics of the pairs of nucleotides. Thus, the phase analysis of complex genomic signals is able to reveal features of both the first- and the second-order statistics of nucleotide distributions along DNA strands.

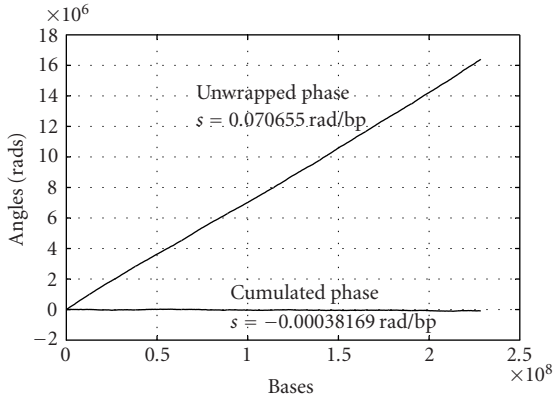


Figure 1.19. Cumulated and unwrapped phase along *Homo sapiens* chromosome 1 (phase 3, total length 228, 507, 674 bp [2]).

1.4.2. Phase analysis of eukaryote DNA sequences

Using the genomic signal approach, long-range features maintained over distances of 10^6–10^8 of base pairs, that is, at the scale of whole chromosomes, have been found in all available eukaryote genomes [31, 32]. The most conspicuous feature is an almost linear variation of the unwrapped phase found in all the investigated genomes, for both eukaryotes and prokaryotes. The slope is specific for various taxa and chromosomes.

Figure 1.19 presents the cumulated and unwrapped phase along concatenated phase-3 data for chromosome 1 of *Homo sapiens,* downloaded from GenBank [2]. Two main features of these phases are readily noticeable.

(i) The cumulated phase remains close to zero, in accordance to the second Chargaff's law for the *distribution* nucleotides—a first-order statistics, stating that the frequency of occurrence of purines and pyrimidines along eukaryote DNA molecules tend to be equal and balance each other [33].

(ii) The unwrapped phase has an almost linear variation maintained for the entire chromosome, for more than 228 millions of nucleotides, including both coding and noncoding regions. Such a behavior proves a rule similar to Chargaff's rule, but reveals a statistical regularity in the *succession* of the nucleotides—a second-order statistics, but reveals a statistical regularity in the *succession* of the nucleotides—a second-order statistics: *the difference between the frequencies of positive nucleotide-to-nucleotide transitions (*A → G, G → C, C → T, T → A*) and of negative transitions (the opposite ones) along a strand of nucleic acid tends to be small, constant, and taxon- and chromosome-specific* [28].

It is worth mentioning that less precise data tend to conform less to this rule, as can be seen from Figure 1.20 that presents the same plots as in Figure 1.19, but for all the concatenated contigs of chromosome 1 of *Homo sapiens*, comprising all the available 238,329,632 nucleotides, without filtering. As a practical use of the unwrapped phase quasilinearity rule, the compliance of a certain contig with the

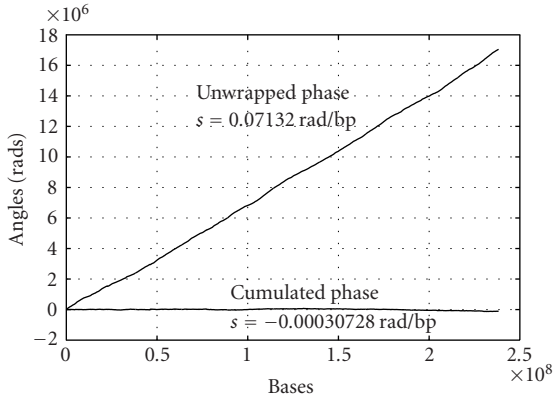


Figure 1.20. Cumulated and unwrapped phase along all concatenated contigs of *Homo sapiens* chromosome 1 (nonfiltered data, total length 238, 329, 632 bp [2, 3, 4]).

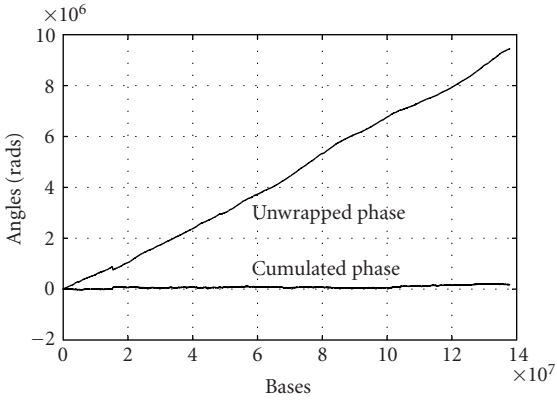


Figure 1.21. Cumulated and unwrapped phase along concatenated contigs of *Homo sapiens* chromosome 11 (older release, nonfiltered data, total length *xxx* bp [2]).

large scale regularities of the chromosome to which it belongs can be a used to spot out exceptions and errors. Figure 1.21 shows the cumulated phase and unwrapped phase along the ensemble of all concatenated contigs of *Homo sapiens* chromosome 11. The average slope of the unwrapped phase is $s_u = 0.0667$ rad/bp, while the various contigs have slopes in the range between 0.047 rad/bp = 2.7 degree/bp and 0.120 rad/bp = 6.9 degree/bp. A striking exception is found in the interval ∼ 15.17–15.38 Mbp of the concatenated string of contigs and corresponds to the contig of accession NT 029410 [2] for which the nucleotide complex representation phases are shown in Figure 1.22. On a length of about 210 Kbp, the unwrapped phase decreases linearly with a sharp average slope $s_u = -0.65$ rad/bp $= -37.2$ degree/bp, which corresponds to a large negative difference in the frequencies of positive and negative transitions $\triangle f_{pm} = f_+ - f_- = -39.4\%/\text{bp}$ and

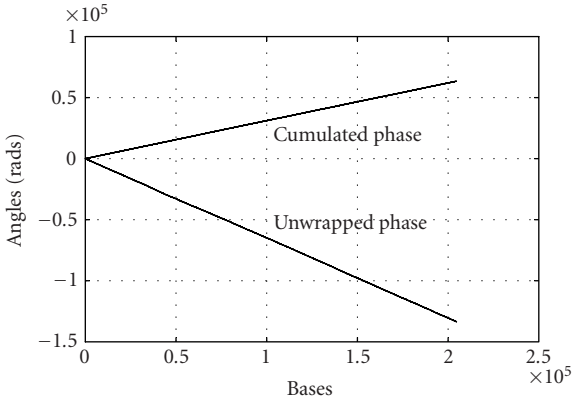


Figure 1.22. Cumulated and unwrapped phase along contig NT_029410 of *Homo sapiens* chromosome 11 (length *xxx* bp [2]).

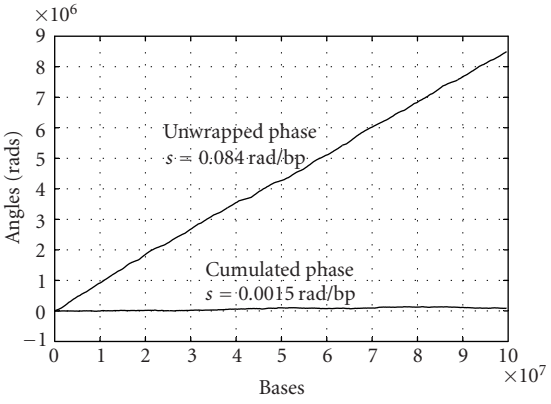


Figure 1.23. Cumulated and unwrapped phase for the available concatenated contigs of *Mus musculus* chromosome 11 (nonfiltered data, total length 99, 732, 879 bp [2, 5, 6]).

to a nucleotide average helix oriented clockwise, completing a turn for about every 9.7 bp. At the same time, the cumulated phase increases linearly with a slope $s_c = 0.325$ rad/bp = 18.6 degree/bp. This data seems to have been dropped from recent releases of chromosome 11 sequences.

Similar large scale properties can be found in all available eukaryote genomes. Figure 1.23 shows the phase diagram for the 99,732,879 nucleotides of the concatenated contigs of *Mus musculus* chromosome 11. The unwrapped phase increases also almost linearly with an average slope $s_u = 0.086$ rad/bp = 4.93 degree/bp, while the cumulated phase remains again almost constant at the scale of the diagram.

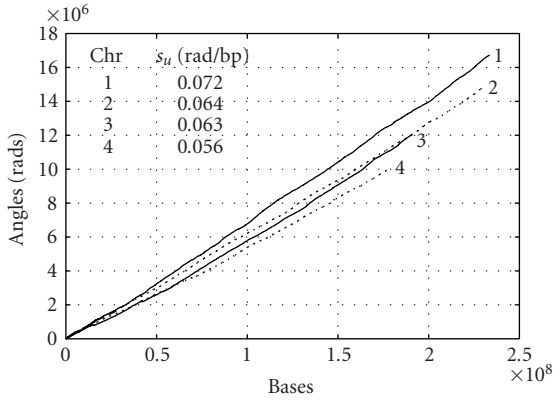


Figure 1.24. Unwrapped phase of the genomic signals for the nucleotide sequence of the concatenated contigs of *Homo sapiens* chromosomes 1–4 [2].

Such long-range regularities of the DNA molecules reveal a structuring of the genomic information at the level of whole chromosomes and contradict the assertion that genomes consist of scarce gene oases in an otherwise essentially empty, unstructured desert. Now it is accepted that the extragene regions can play significant functional roles at the level of the whole chromosome, in controlling processes like replication, transcription, crossover, and others. Along with many of the genes, the *Homo sapiens* and the *Mus musculus* genomes share twice as much other extragene DNA sequences. It is conjectured that these sequences must have important functions that explain how they were conserved over a divergent evolution of some 75 million years of the human and mouse lineages [1, 2, 3, 5, 7].

The approximately linear variation with positive slope has been found for the unwrapped phase of the genomic signals of all the chromosomes of *Homo sapiens* and *Mus musculus*. Figure 1.24 shows the results for the four largest chromosomes of *Homo sapiens*, while Figure 1.25 gives the curve for the shortest three chromosomes. Significant segments with negative slopes of the unwrapped phase have been found in *Homo sapiens* chromosomes 5, 8, 11, 17, 21, and Y. The average slope of the unwrapped phase is taxon and chromosome specific and has a functional role, most probably in controlling the movement Brownian machines like the DNA polymerase and in selecting homologous sites for the crossover exchange of genomic material. Table 1.4 shows the average slopes of the unwrapped phase for the concatenated contigs of *Homo sapiens* chromosomes currently available in the GenBank [2] data base.

1.4.3. Phase analysis of prokaryote DNA sequences

We start illustrating the phase features of prokaryote DNA sequences with the case of the well-studied *Escherichia coli*, for which the genome has been one of the first completely sequenced [14]. The most striking feature in Figure 1.26 is

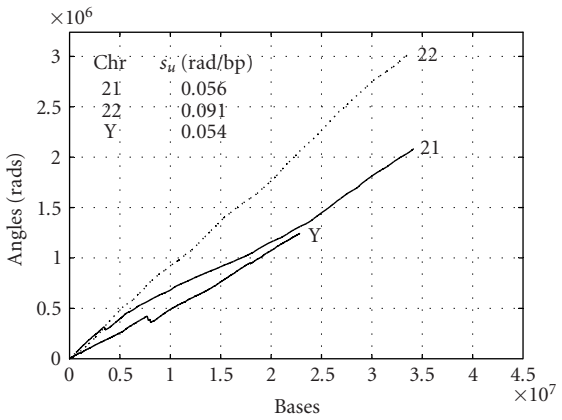


Figure 1.25. Unwrapped phase of the genomic signals for the nucleotide sequence of the concatenated contigs of *Homo sapiens* chromosomes 21, 22, Y [2].

Table 1.4. Average slopes of the unwrapped phase for the concatenated contigs of *homo sapiens* chromosomes.

Chr	s_u (rad/bp)	Chr	s_u (rad/bp)	Chr	s_u (rad/bp)	Chr	s_u (rad/bp)
1	0.072	7	0.066	13	0.057	19	0.084
2	0.064	8	0.062	14	0.066	20	0.073
3	0.063	9	0.066	15	0.072	21	0.057
4	0.056	10	0.067	16	0.075	22	0.091
5	0.060	11	0.069	17	0.078	X	0.057
6	0.062	12	0.068	18	0.060	Y	0.054

that the cumulated phase varies piecewise linearly along two domains of the circular DNA having almost equal length: a region of positive slope $s_{c+} = 0.0393$ rad/bp of length $l_+ = 2,266,409$ bp (split into two domains 1–1, 550, 413 bp and 3, 923, 226–4, 639, 221 bp) and a region of negative slope $s_{c-} = -0.0375$ rad/bp of length $l_- = 2,372,812$ bp. The quite sharp extremes of the cumulated phase are at 3, 923, 225 bp and 1, 550, 413 bp, respectively, very close to the experimentally found origin and terminus of chromosome replication. Quite similar diagrams have been obtained analyzing the difference in the occurrence frequencies of purines over pyrimidines R − Y = (A + G) − (T + C) (Figure 1.27) and of ketones over amines K − M = (G + T) − (C + A) (Figure 1.28). Figure 1.29 shows the excess of weak over strong bonds along the *Escherichia coli* DNA strand. As is well known, for prokaryotes most of the chromosome comprises encoding regions and in which cytosine and guanine are in excess over adenine and thymine.

It is rather surprising that the variation closest to (piecewise) linear is found for the cumulated phase, which has a slope dependent on a mixture of the nucleotide occurrence frequencies given by equation (1.10). Again, the variation of the unwrapped phase is almost linear for the whole chromosome (Figure 1.30) and passes without change over the points where the slope of the cumulated phase

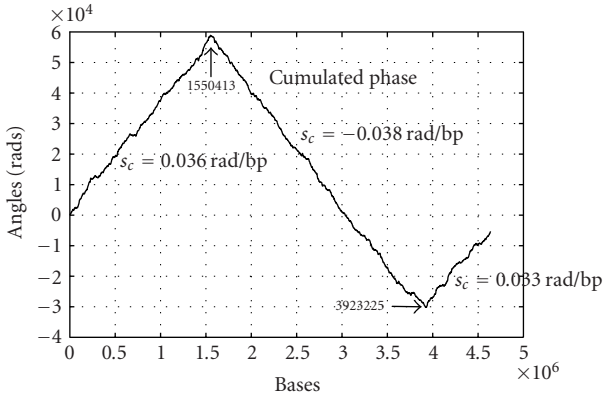


Figure 1.26. Cumulated phase for the circular chromosome of *Escherichia coli* K12 (NC_000913, complete genome, length 4, 639, 221 bp [2, 14]).

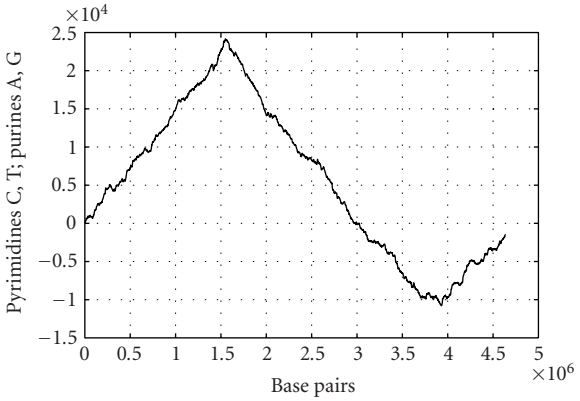


Figure 1.27. Purine over pyrimidine excess (A + G) − (T + C) along the circular chromosome of *Escherichia coli* K12 (NC_000913, complete genome, length 4, 639, 221 bp [2, 14]).

changes sign. This is a general feature, found for all chromosomes and all prokaryotes, and will be discussed in the next section of this chapter.

Figure 1.31 shows the cumulated and the unwrapped phase along the circular chromosome of *Yersinia pestis* [18] (accession number NC_003143 [2]). As in the case of *Escherichia coli*, the breaking points are most probably in relation with the origins and the termini of chromosome replichores, but we are not aware of the corresponding experimental results. It is to be noticed that, in opposition to *Escherichia coli* [14] and *Bacillus subtilis* [16] which display only one maximum and one minimum [24], the cumulated phase of *Yersinia pestis* shows four points of each type. This corresponds to the fusion of more strains into the circular chromosome of *Yersinia pestis* and could reveal aspects of the ancestral history of the pathogen. The change of sign of the cumulated phase slope at the breaking points shows that there is a cut and a macroswitch of the two DNA strands, so that the

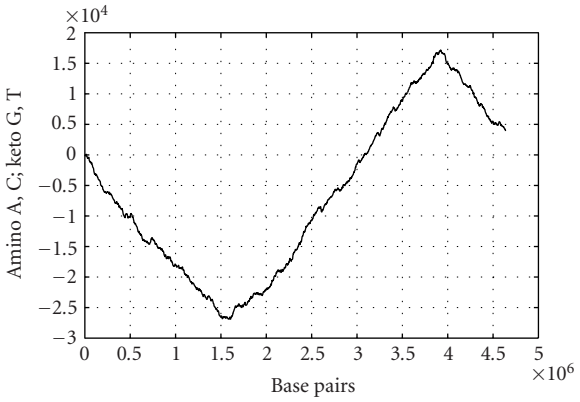


Figure 1.28. Keto over amino excess (G + T) − (C + A) along the chromosome of *Escherichia coli* (NC_000913 [2, 14]).

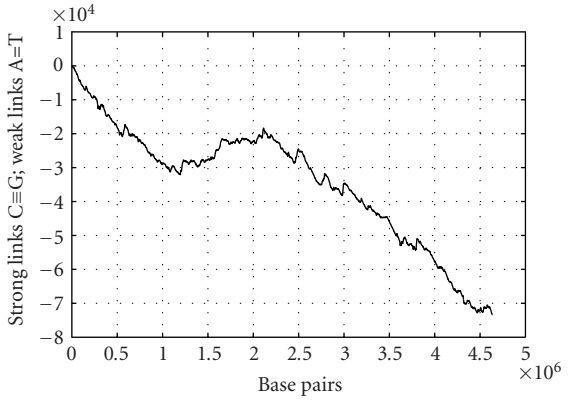


Figure 1.29. Weak bonds over strong bonds W − S = (A + T) − (C + C) along the chromosome of *Escherichia coli* (NC_000913 [2, 14]).

difference between the frequencies of occurrence of the nucleotides changes the sign. It is remarkable that, in the same points, there is little or no change in the unwrapped phase. This will be explained in the next section of this chapter based on a longitudinal model of the chromosomes' "patchy" structure.

Similar characteristics have been found for almost all other studied prokaryotes. Figure 1.32 presents the cumulated and unwrapped phase for an intracellular pathogen of humans: *Chlamydophila pneumoniae* CWL029 (NC_000922 [34]). Again the linear regions correspond to the "replichores" of bacterial circular chromosomes, and the extremes of the cumulated phase are the origin and terminus of chromosome replication. The differences in nucleotide occurrence frequencies have been explained by the differences in mutation probabilities resulting from the asymmetry of replication mechanisms for the leading and lagging strands but, most probably, this statistically ordered nonhomogenity plays a fundamental

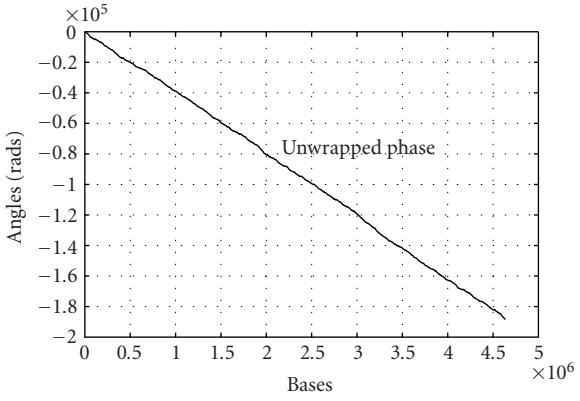


Figure 1.30. Unwrapped phase for the circular chromosome of *Escherichia coli* (NC_000913 [2, 14]).

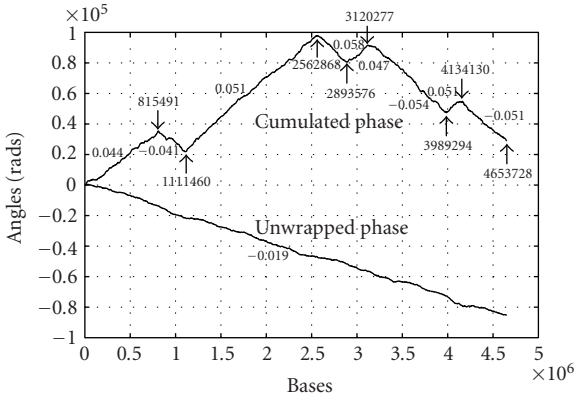


Figure 1.31. Unwrapped and cumulated phase for the circular chromosome of *Yersinia pestis* (NC_0003143, complete genome, length 4, 653, 728 bp [2, 18]).

role in the functioning of some "molecular machines," like DNA polymerase that moves along a DNA strand by converting the thermal motion in an ordered displacement.

It has been shown recently that DNA molecules have a fractal-like structure resulting from their long-range correlations [30]. The self-similarity, that is, the fractal-like structure is revealed by the linearity of the plot $\log(N)$ versus $\log(B)$, where N is the number of filled boxes of size B, while the slope gives the fractal dimension. From the analysis of the cumulated phase of the circular chromosome of *Chlamydophila pneumoniae* CWL029 in Figure 1.32, with a 1024 bp sliding window, an average fractal dimension of 1.05 has been found, only slightly higher than one, in accordance with the long correlations observed in the cumulated and unwrapped phase curves.

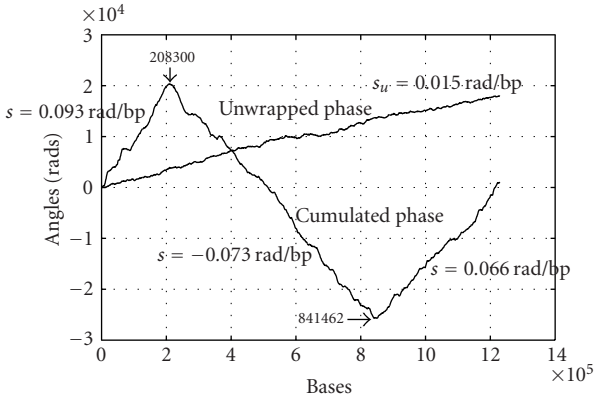


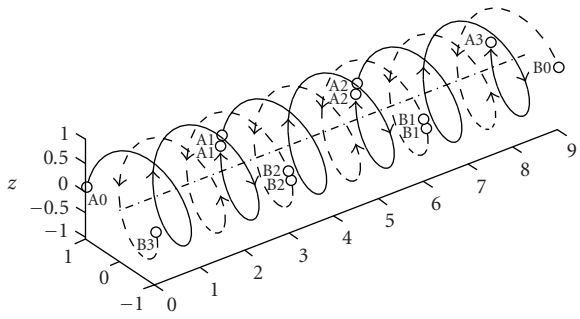
Figure 1.32. Cumulated and unwrapped phase for the circular chromosome of *Chlamydophila pneumoniae* CWL029 (NC_000922, complete genome, length 1,230,230 bp [2]).

1.5. Phase analysis of reoriented ORFs

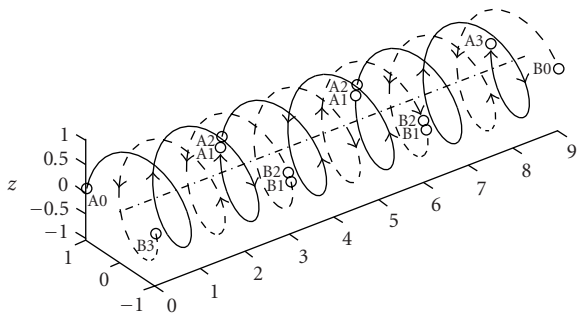
As discussed in Section 1.2, each DNA strand has a well-defined positive direction (the 5′ → 3′ sense), along which successive nucleotides can be joined to each other [29]. The two strands of a DNA double helix have opposite positive directions. DNA molecules have a very "patchy" structure with intertwined coding and noncoding segments oriented in both direct and inverse sense [36]. For most currently sequenced genomes, the information about the direct or inverse orientation of the coding regions—the ORF—has been identified and is available in the genomic databases [2].

The main point that results from the analysis of the modalities in which DNA segments can be chained together along a DNA double helix is that a direction reversal of a DNA segment is always accompanied by a switching of the antiparallel strands of its double helix. This property is a direct result of the requirement that all the nucleotides be linked to each other along the DNA strands only in the 5′ to 3′ sense.

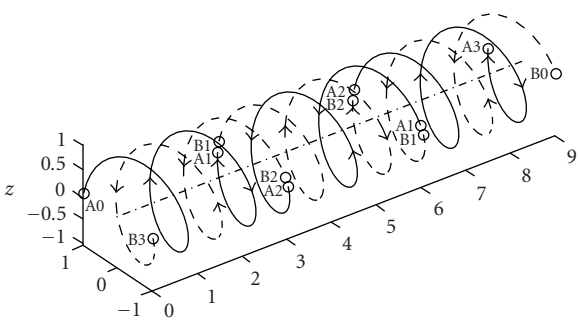
Figure 1.33 schematically shows the way in which the positive orientation restriction is satisfied when a segment of a DNA double helix is reversed and has simultaneously switched its strands. In Figure 1.33a, the chains $(A_0A_1)(A_1A_2)(A_2A_3)$ and $(B_0B_1)(B_1B_2)(B_2B_3)$ have been marked on the two strands, having the positive (5′ to 3′) directions as indicated by the arrows. The reversal of the middle segment, without the corresponding switching of its strands (Figure 1.33b), would generate the forbidden chains $(A_0A_1)(A_2A_1)(A_2A_3)$ and $(B_0B_1)(B_2B_1)(B_2B_3)$ that violate the 5′ to 3′ alignment condition. Similarly, the switching of the strands of the middle segment, without its reversal, would generate the equally forbidden chains $(A_0A_1)(B_2B_1)(A_2A_3)$ and $(B_0B_1)(A_2A_1)(B_2B_3)$, not shown in Figure 1.33. Finally, only the conjoint reversal of the middle segment and the switching of its strands (Figure 1.33c) generate the chains $(A_0A_1)(B_1B_2)(A_2A_3)$ and $(B_0B_1)(A_1A_2)(B_2B_3)$, which are compatible with the 5′ to 3′ orientation condition.



(a)



(b)



(c)

Figure 1.33. Schematic representation of a DNA segment direction reversal: (a) the two antiparallel strands have the segments ordered in the 5′ to 3′ direction indicated by arrows; (b) hypothetic reversal of the middle segment, without the switching of the strands; (c) direction reversal and strand switching for the middle segment. The 5′ to 3′ alignment condition is violated in case (b) but reestablished for (c).

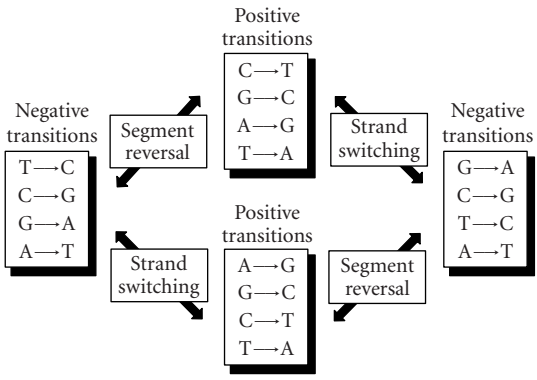


Figure 1.34. Interchange of positive and negative nucleotide-to-nucleotide transitions after *segment reversal and strand switching.*

We also mention that, in order to practically perform such a reversal, the two branches (A_1A_2) and (B_1B_2) of the DNA double helix segment should not be exactly aligned, but slightly shifted with respect to each other and the "free" nucleotides at the two ends should be complementary, to provide the necessary "sticky ends" allowing the easy reattachment of the strands. This condition does not affect the aspects discussed here.

As a consequence of the coupling of the direction reversal with the strand switching imposed by the condition to maintain the continuity of the positive directions ($5' \rightarrow 3'$) along the two strands of the DNA molecule, there is always a pair of changes when a DNA segment is inversely inserted. Thus, the sense/antisense orientation of individual DNA segments affects only the nucleotide frequencies, but conserves the frequencies of the positive and negative transitions. Figure 1.34 shows how the type of nucleotide-to-nucleotide transitions changes (positive to negative, and *vice versa*) for a segment reversal and for a strand switching. The reversal of an individual DNA segment affects only the first-order statistics of the nucleotides, while the second-order statistics remains unchanged. Thus, the cumulated phase of a genomic signal, which depends on the frequency of nucleotides along the corresponding DNA strand, changes significantly for a segment reversal, while the unwrapped phase, which depends on second-order statistical features, does not.

This model explains why the unwrapped phase has a regular, almost linear, variation even for eukaryote chromosomes [23, 24], despite their very high fragmentation and quasirandom distribution of direct and inverse DNA segments, while the cumulated phase has only a slight drift close to zero.

Figure 1.35 shows together the cumulated phase and the unwrapped phase of the genomic signal for the complete circular chromosome of *Escherichia coli* [14] (NC_000913 [2]) comprising 4,639,221 bp (also shown in Figures 1.26 and 1.30) and for the 4,290 concatenated reoriented coding regions, comprising 4,097,248 bp. All the coding regions having an inverse reading frame have been inversed

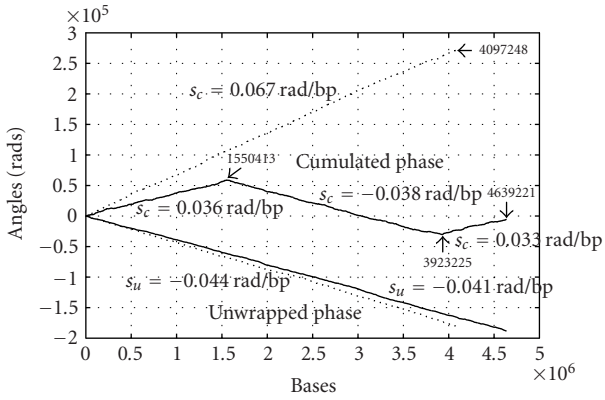


Figure 1.35. Cumulated and unwrapped phase of the genomic signals for the complete genome (4, 639, 221 bp) and the 4, 290 concatenated reoriented coding regions (4, 097, 248 bp) of *Escherichia coli* (NC_000913 [2, 14]).

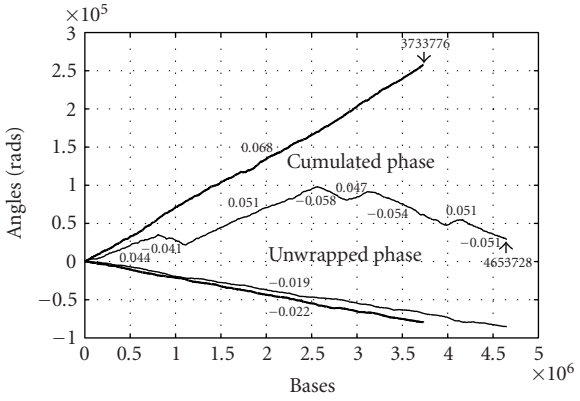


Figure 1.36. Cumulated and unwrapped phase of the genomic signals for the complete genome (4, 653, 728 bp) and the 4034 concatenated reoriented coding regions (3, 733, 776 bp) of *Yersinia pestis*[3,17] (accession number NC_003143 [2, 18]).

and complemented (i.e., A and T, on one hand, C and G, on the other, have been interchanged to account for strand switching). The disappearance of the breaking points in the cumulated phase under the effect of the reorienting is evident, while the unwrapped phase changes little.

Similarly, Figure 1.36 shows the cumulated and the unwrapped phase for the complete circular chromosome of *Yersinia pestis* strain CO92 (accession number NC_003143) with a length of 4, 853, 728 bp and for its concatenated reoriented 3, 884 coding regions comprising 3, 733, 776 bp. The slope of the cumulated and the unwrapped phases are changed not only because the intergene regions have been eliminated, but also because direct and inverse coding regions are actually

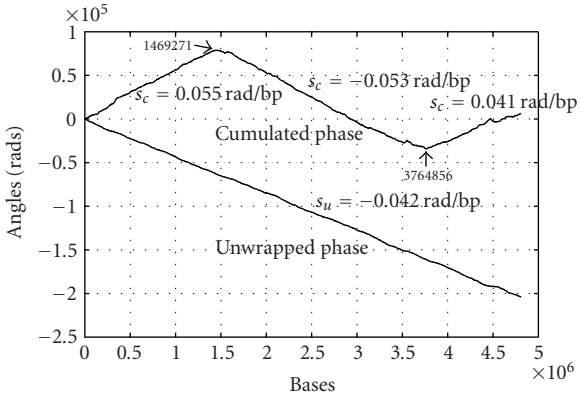


Figure 1.37. Cumulated and unwrapped phase for the circular chromosome of *Salmonella typhi* (AL_5113382, length 4, 809, 037 bp [2]).

distributed in all the four positive and four negative slope segments of the cumulated phase, certainly, with very different frequencies. The orientation of the coding regions correlates well with the slope of the cumulated phase: most direct ORF are in the positive slope regions, while most inverse ORF are in the negative slope regions.

Figure 1.37 presents the cumulated and the unwrapped phase of the complete circular chromosome *Salmonella typhi*, the multiple drug resistant strain CT18 (accession AL_513382 [2]). The locations of the breaking points, where the cumulated phase changes the sign of the slope of its variation along the DNA strand, are given in the figure. Even if locally the cumulated phase and the unwrapped phase have not a smooth variation, at the scale used in Figure 1.37, the variation is quite smooth and regular. A pixel in the lines in Figure 1.37 represents 6050 data points, but the absolute value of the difference between the maximum and minimum values of the data in the set of points represented by each pixel is smaller than the vertical pixel dimension expressed in data units. This means that the local data variation falls between the limits of the width of the line used for the plot, so that the graphic representation of data by a line is fully adequate. The conditions for signals graphical representability as lines will be presented in more detail in the next section of this chapter. As shown in the previous section for other prokaryotes, the cumulated phase has an approximately piecewise linear variation over two almost equal domains, one of positive slope (apparently divided in the intervals 1–1469271 and 3764857–4809037, but actually contiguous on the circular chromosome) and the second of negative slope (1469272–3764856), while the unwrapped phase has an almost linear variation for the entire chromosome, showing little or no change in the breaking points. The breaking points, like the extremes of the integrated skew diagrams, have been put in relation with the origins and termini of chromosome replichors [28, 37, 38]. The slope of the cumulated phase in each domain is related to the nucleotide frequency in that domain by equation

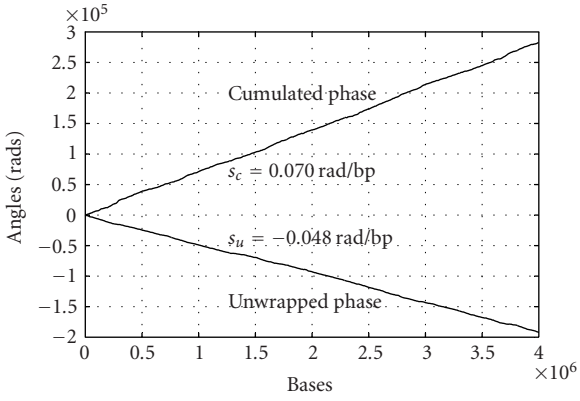


Figure 1.38. Cumulated and unwrapped phase of the concatenated 4393 reoriented coding regions (3,999,478 bp) of *Salmonella typhi* genome (AL_5113382 [2]).

(1.10). In the breaking points, apparently a macroswitching of the strands, accompanied by the reversal of one of the domain-large segments, occurs. The two domains comprise a large number of much smaller segments, oriented in the direct and the inverse sense. At the junctions of these segments, the reversal and switching of DNA helix segments, as described in the previous section, take place. The average slope of each large domain is actually determined by the density of direct and inverse small segments along that domain. Because the intergenic regions, for which the orientation is not known, have to be left out of the reoriented sequence, the new sequence is shorter than the one that contains the entire chromosome or all the available contigs given in the GenBank data base [2].

Figure 1.38 shows the cumulated and unwrapped phase of the genomic signal obtained by concatenating the 4393 reoriented coding regions of *Salmonella typhi* genome (accession AL_5113382 [2]). Each inverse coding region (inverse ORF) has been reversed and complemented, that is, the nucleotides inside the same W (adenine-thymine) or S (cytosine-guanine) class have been replaced with each other, to take into account the switching of the strands that accompanies the segment reversal. As expected from the model, the breaking points in the cumulated phase disappear and the absolute values of the slopes increase, as there is no longer interweaving of direct and inverse ORFs. The average slope s_c of the cumulated phase of a genomic signal for a domain is linked to the average slope $s_c^{(0)}$ of the concatenated reoriented coding regions by the relation:

$$s_c = \frac{\sum_{k=1}^{n_+} l_k^{(+)} - \sum_{k=1}^{n_-} l_k^{(-)}}{\sum_{k=1}^{n_+} l_k^{(+)} + \sum_{k=1}^{n_-} l_k^{(-)}} s_c^{(0)}, \tag{1.14}$$

where $\sum_{k=1}^{n_+} l_k^{(+)}$ and $\sum_{k=1}^{n_-} l_k^{(-)}$ are the total lengths of the n_+ direct and n_- inverse ORFs in the given domain.

The unwrapped phase, which is linked by equation (1.12) to the nucleotide positive and negative transition frequencies, shows little or no change when replacing the chromosome nucleotide sequence with the concatenated sequence of reoriented coding regions. As explained, the reorientation of the inverse coding regions consists in their reversal and switching of their strands. Figure 1.34 shows the effect of the segment reversal and strand switching transformations on the positive and negative nucleotide-to-nucleotide transitions for the case of the complex genomic signal representation given by equation (1.1). After an even number of segment reversal and/or strand switching transformations of a DNA segment, the nucleotide transitions do not change their type (positive or negative). As a consequence, the slope of the unwrapped phase does not change.

It is remarkable that the approximately piecewise linear variation of the cumulated phase for the whole chromosome, comprising two complementary regions—also found by skew diagrams techniques [36, 38]—is replaced with an approximately linear variation over the whole sequence, when reorienting all coding regions in the same reference direction. This result could suggest the existence of an ancestral chromosome structure with a single global statistical regularity, which has evolved into a more complex structure by the reversal of the direction for a significant part of DNA segments.

Similar results have been found in the phase analysis of many other genomic signals corresponding to circular and linear chromosomes of various prokaryotes. A special case is the aerobic hyperthermophilic crenarchaeon *Aeropyrum pernix K*, for which the genome comprising $1,669,695$ base pairs has been completely sequenced [3]. The unwrapped phase varies almost linearly, in agreement with the similar results found for all the other investigated prokaryote and eukaryote genomes, confirming the rule stated in the previous section. But the cumulated phase decreases irregularly, an untypical behavior for prokaryotes that tend to have a regular piecewise linear variation of the cumulated phase along their circular DNA molecules, as shown above. Nevertheless, the cumulated phase of the $1,553,043$ base pairs signal corresponding to the sequence obtained by concatenating the $1,839$ coding regions, after reorienting them in the same reference direction, becomes approximately linear, while the unwrapped phase remains unchanged.

We conjecture that the fine combining of DNA segments with opposite orientation, in order to generate certain well-defined values of the slope of the cumulated phase, that is, certain densities of the repartition of nucleotides, has a functional role at the level of the chromosomes, most probably in processes like replication, transcription, or cross over. The particular statistical structure of DNA molecules that generates this specific shape of the cumulated and unwrapped phases could play an important role in the mutual recognition and alignment of interacting regions of chromosomes and the separation of the species. The first- and second-order statistical regularities, resulting from the specific variation of the unwrapped and cumulated phases, can be put in correspondence with the molecule potentials produced by available hydrogen bonds and can be used to describe the interaction of a given DNA segment with proteins and with other

DNA segments in processes like replication, transcription, or crossover. An example is the movement of DNA polymerase along a DNA strand, operating like a "Brownian machine that converts random molecular movements into an ordered gradual advance during replication. The speed of movement can be expressed as a function of the temperature and the slope of the phase. These hypotheses are also sustained by the fact that the emergence of an almost linear variation of the cumulated phase after the reorientation of all coding regions is a property found in both circular and linear chromosomes of prokaryotes, but not in the plasmids.

1.6. Representability of genomic signals

1.6.1. Well-fitted screens and the data scattering ratio

When operating with large sets of data, especially data describing complex systems or processes or generated by such systems or processes, with a possibly chaotic or random dynamics, the problem of adequate representation of data is central. The final understanding of any set of data or signals lays with human operators for which the graphical representation, allowing to grasp at once features hidden in piles of numerical data, is the natural choice, at least as a guide to the precise values. As shown in the previous sections of this chapters, symbolic nucleotide sequences can be converted into digital genomic signals by using the complex (2D) quadrantal representation derived from the tetrahedral (3D) representation of nucleotides. The study of complex genomic signals, using signal processing methods, reveals large scale features of chromosomes that would be difficult to grasp by applying only the statistical or pattern matching methods currently used in the analysis of symbolic genomic data. In the context of operating with a large volume of data, at various resolutions, and visualizing the results to make them available to humans, the problem of data representability becomes critical. In the following, we present an analysis of data representability based on the concept of the data scattering ratio of a pixel. Representability diagrams are traced for several typical cases of standard signals and for some genomic signals. It is shown that the variation of genomic data along nucleotide sequences, specifically the cumulated and unwrapped phase, can be visualized adequately as simple graphic lines for low and large scales, while for medium scales (thousands to tens of thousands of base pairs) the statistical-like description must be used.

Figure 1.39 shows the plot as a line of the digital signal $s[i]$, $i \in I^S = \{1, \ldots, L\}$, where L is the length of the sequence or subsequence of data to be represented. One pixel is extracted and magnified to allow comparing the absolute value of the variation V_y of the signal for the set of samples represented by the pixel with the pixel height P_y measured in signal units. For the case in the figure, which corresponds to real data giving the unwrapped phase of the complete genome of *Bacillus subtilis*, we have $V_y < P_y$, so that the graphical representation of the data by a line with the width of a pixel is adequate in that point and, actually, for the whole sequence. The size of the screen in pixels is considered fixed, for example, the usual screen size $N_x = 1024$ by $N_y = 768$ pixels. To optimally use the screen to represent the

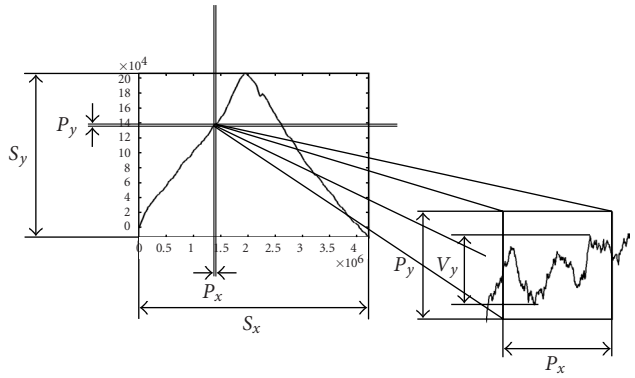


Figure 1.39. Data-fitted screen and a magnified pixel.

data, the available screen space must be fitted to the data: the horizontal screen size S_x, in number of samples, has to be made equal to the length L of the sequence (or subsequence) to be represented, while the screen vertical size S_y, in data units, must be chosen equal to the absolute value of the variation of the data in the represented sequence:

$$S_y = \max_{i \in I^S} \left(s[i]\right) - \min_{i \in I^S} \left(s[i]\right). \tag{1.15}$$

Correspondingly, the horizontal and vertical pixel sizes are given by

$$P_x = \frac{S_x}{N_x}, \qquad P_y = \frac{S_y}{N_y}, \tag{1.16}$$

in number of samples and data units, respectively.

The variation of the data for the set of samples corresponding to a pixel is

$$V_y(h) = \max_{i \in I_h^P} \left(s[i]\right) - \min_{i \in I_h^P} \left(s[i]\right), \tag{1.17}$$

where $I_h^P = \{(h-1)P_x + 1, \ldots, hP_x\}; h = 1, \ldots, N_x$.

As mentioned above, the adequateness of the representation of the set of P_x data samples by just one a pixel can be characterized by the ratio:

$$Q(h) = \frac{V_y(h)}{P_y} \tag{1.18}$$

that we will call the data scattering ratio of the pixel h.

If $Q \leq 1$, the pixel represents properly all the data samples it represents and covers. When all the pixels in a line satisfy this condition, the data can be represented adequately by a line having the width of one pixel. If Q is below two or

three units for every pixel of the sequence fitted in the screen, the data can also be represented properly by a line, but the width of the line must correspond to the maximum value of Q. When Q has larger values, but the data is densely and quite uniformly distributed, so that Q is approximately the same for all the pixels and there are no outliers, the data can be represented adequately by a couple of lines showing the maximum and minimum values of the data for each pixel. Finally, if the data is scattered and/or there are outliers, this approach is no longer practical and a statistical-like description of data is needed for their representation. The pixel can be considered a sliding window of size P_x. If the data distribution is close enough to a normal distribution, the data can be described for each such window by the mean value and the standard deviation. A line giving the mean value and two other lines or some error bars for delimiting some confidence interval expressed in terms of the standard deviation can be used to represent the data.

In the following, we analyze the representability of several types of data and signals, including genomic signals, in terms of their representability characteristic

$$\tilde{Q} = \frac{\tilde{V}_y}{P_y} = f(P_x), \tag{1.19}$$

where $\tilde{Q} = \tilde{V}_y/P_y$ is the average data scattering ratio for all the pixels in the represented line, with $\tilde{V}_y = \text{mean}_{h=1,\ldots,N_x}(V_y(h))$, while P_x is the pixel horizontal size. When drawing the representability diagram showing the representability characteristic (1.19), logarithmic scales in base 2 will be used for both abscissa and ordinate. Correspondingly, the pixel size $P_x^{(k)}$ will be increased in a geometrical scale with ratio two:

$$P_x^{(1)} = 1,\ldots, \qquad P_x^{(k)} = 2^{k-1},\ldots, \qquad P_x^{(k_{\max})} = 2^{k_{\max}-1}, \tag{1.20}$$

so that the screen horizontal size $S_x^{(k)} = N_x P_x^{(k)}$, $k = 1,\ldots,k_{\max}$, will also double at each step for a fixed N_x. The number of steps necessary to cover the whole sequence of length L is $k_{\max} = \lfloor \log_2 L/N_x \rfloor + 1$, where $\lfloor x \rfloor$ denotes the smaller integer larger than or equal to x. In this case, the largest screen equals or exceeds the length of the sequence. The number of screens necessary to represent the whole sequence at step k is

$$N_S^{(k)} = \left\lfloor \frac{L}{S_x^{(k)}} \right\rfloor = \left\lfloor \frac{L}{N_x 2^{k-1}} \right\rfloor. \tag{1.21}$$

If the length L of the sequence is not a power of two, the last screen at each step k, including the largest screen for the last step, might not be well fitted to the data and will be excluded from the diagram. When $L = 2^m$ and $N_x = 2^s$, all screens will be horizontally fitted to the data and their number $N_S^{(k)} = 2^{m-s+1-k} = 2^{k_{\max}-k}$ will form a geometrically decreasing sequence with ratio 1/2, from $2^{k_{\max}-1}$ to 1. Each screen (window) will be vertically fitted to the data, by choosing its vertical

size equal to the absolute value of the variation of the data in that screen:

$$S_y^{(k)}(j) = \max_{i \in I_j^S} (s[i]) - \min_{i \in I_j^S} (s[i]), \quad j = 1, \ldots, j_{\max}^{(k)}, \tag{1.22}$$

where $I_j^S = \{(j-1)S_x^{(k)} + 1, \ldots, jS_x^{(k)}\}$ are the indices of the samples represented in the screen j and $j_{\max}^{(k)} = N_S^{(k)}$ is the number of screens at step k.

A 3D diagram will be used to show the variation of the average data scattering ratio for the pixels in each of the screens used to cover all the length of the sequence L at various pixel sizes.

1.6.2. Representability best case: monotonic signals

In the case of monotonically increasing signals, the relation (1.13) for the vertical size of screen j becomes

$$S_y^{(k)}(j) = s\left[jS_x^{(k)}\right] - s\left[(j-1)S_x^{(k)} + 1\right], \tag{1.23}$$

so that the average screen height results:

$$\tilde{S}_y^{(k)} = \operatorname*{mean}_{j=1,\ldots,N_S^{(k)}} \left(S_y^{(k)}(j)\right) = \frac{1}{N_S^{(k)}} \sum_{j=1}^{N_S^{(k)}} \left(s\left[jS_x^{(k)}\right] - s\left[(j-1)S_x^{(k)} + 1\right]\right). \tag{1.24}$$

Using $j_{\max}^{(k)} S_x^{(k)} = L$, this expression can be rewritten as

$$\tilde{S}_y^{(k)} = \frac{2^{k-1}N_x}{L}\left(s[L] - s[1] - \sum_{j=1}^{N_S^{(k)}-1} \left(s\left[jS_x^{(k)} + 1\right] - s\left[jS_x^{(k)}\right]\right)\right), \tag{1.25}$$

where the sum contains signal variations between samples at distance one, subsampled with the step $S_x^{(k)}$. A similar expression holds for monotonically decreasing signals, so that the average screen height for monotonic signals results:

$$\tilde{S}_y^{(k)} = \frac{2^{k-1}N_x}{L}\left(s[L] - s[1] - \left(j_{\max}^{(k)} - 1\right) \operatorname{mean}\left(|d|\right)_{\downarrow S_x^{(k)}}\right), \tag{1.26}$$

where

$$\operatorname{mean}\left(|d|\right)_{\downarrow S_x^{(k)}} = \operatorname*{mean}_{j=1,\ldots,j_{\max}^{(k)}-1} \left(\left|d\left[jS_x^{(k)}\right]\right|\right) = \frac{1}{j_{\max}^{(k)} - 1} \sum_{j=1}^{j_{\max}^{(k)}-1} \left|d\left[jS_x^{(k)}\right]\right| \tag{1.27}$$

is the average absolute variation of the signal between samples at distance one, down-sampled at the step $S_x^{(k)}$.

Similarly, from (1.17) results the average variation of the data for sets of samples corresponding to pixels:

$$\tilde{V}_y^{(k)} = \frac{2^{k-1}}{L}\Big(s[L] - s[1] - (h_{\max}^{(k)} - 1)\,\text{mean}\,(|d|)_{\downarrow P_x^{(k)}}\Big), \tag{1.28}$$

where

$$\text{mean}\,(|d|)_{\downarrow P_x^{(k)}} = \underset{h=1,\ldots,h_{\max}^{(k)}-1}{\text{mean}}\,(|\,d[hP_x^{(k)}]\,|) = \frac{1}{h_{\max}^{(k)} - 1}\sum_{h=1}^{h_{\max}^{(k)}-1}|\,d[hP_x^{(k)}]\,| \tag{1.29}$$

is the average absolute variation of the signal between samples at distance one, down-sampled at the pixel step $P_x^{(k)}$.

As a consequence, the average data scattering ratio for a monotonic signal is given by the equation

$$\tilde{Q}^{(k)} = \frac{\tilde{V}_y^{(k)}}{\tilde{P}_y^{(k)}} = \frac{N_y}{N_x}\frac{s[L] - s[1] - (N_P^{(k)} - 1)\,\text{mean}\,(|d|)_{\downarrow P_x^{(k)}}}{s[L] - s[1] - (N_S^{(k)} - 1)\,\text{mean}\,(|d|)_{\downarrow S_x^{(k)}}}, \tag{1.30}$$

where $N_P^{(k)}$ is the total number of pixels to represent the sequence $s[i]$, $i = 1,\ldots,L$, for a horizontal pixel size $P_x^{(k)} = 2^{k-1}$, $N_S^{(k)} = N_P^{(k)}/N_x$ is the total number of screens necessary to represent the data at resolution k, and mean$(|d|_{\downarrow D})$ is the average of the absolute values of the signal variation between successive samples $d[i] = s[i+1] - s[i]$, down-sampled at step D. As long as the sampling density is high enough,

$$\text{mean}\,(|d|)_{\downarrow S_x^{(k)}} \approx \text{mean}\,(|d|)_{\downarrow P_x^{(k)}} \approx \frac{s[L] - s[1]}{L - 1}, \tag{1.31}$$

so that equation (1.30) becomes

$$\tilde{Q}^{(k)} = \frac{N_y}{N_x}\frac{P_x^{(k)} - 1}{P_x^{(k)} - 1/N_x}. \tag{1.32}$$

From (1.32) it results that all monotonic signals have almost the same representability characteristic drawn in Figure 1.40 as a line. The circles correspond to experimental data for various monotonic signals like linear, parabolic of various degrees, logarithmic and exponential of various bases, and so forth. Monotonic signals are the best practical case in what concerns the representability characteristic. As results from (1.32) and from the data in Figure 1.40, for large values of the pixel width P_x, the representability characteristic tends asymptotically towards the aspect ratio of the screen:

$$\tilde{Q}^{(k)} \xrightarrow[2^{k-1}\gg 1]{} \frac{N_y}{N_x}. \tag{1.33}$$

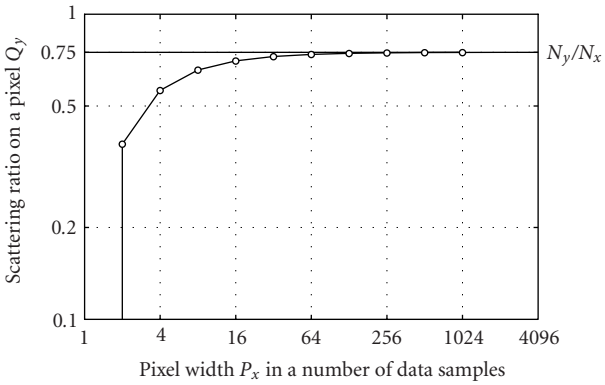


Figure 1.40. Representability diagram (pixel width Px versus average data scattering ratio on a pixel $\tilde{Q}$) for monotonic signals. For the illustration, the length of the signal has been chosen $2^{20} = 1048576$ bp, and the screen size 1024×768 pixels.

1.6.3. Representability practical worst case: uniformly distributed random signals

The theoretical *worst case* from the representability point of view is a hypothetical signal for which the variation between two successive samples is equal to the screen height. A practical worst case is provided by a random signal uniformly distributed on the screen height. The representability characteristic can also be found in closed form for his case. The average variation of the data for the set of $P_x^{(k)}$ samples corresponding to a pixel, that is, the average of the difference between the largest and the smallest values of the samples in the set of $P_x^{(k)}$ random variables uniformly distributed across the screen height expressed in pixels is given by [26]

$$\tilde{Q}^{(k)} = \frac{\tilde{V}_y^{(k)}}{\tilde{P}_y^{(k)}} = N_y \frac{P_x^k - 1}{P_x^k + 1}. \tag{1.34}$$

The representability characteristic is shown in Figure 1.41. The line has been computed analytically using the equation (1.25), while the circles represent data from a Monte Carlo simulation of the uniform distribution of the samples in a range equal to the screen height in data units. For large values of the pixel width, the representability characteristic asymptotically approaches N_y—the vertical size of the screen in pixels:

$$Q^{(k)} \xrightarrow[2^{k-1} \gg 1]{} N_y. \tag{1.35}$$

The monotonic signals and the uniformly distributed random signal provide the practical limiting cases of the framework in which the real-word signal fall.

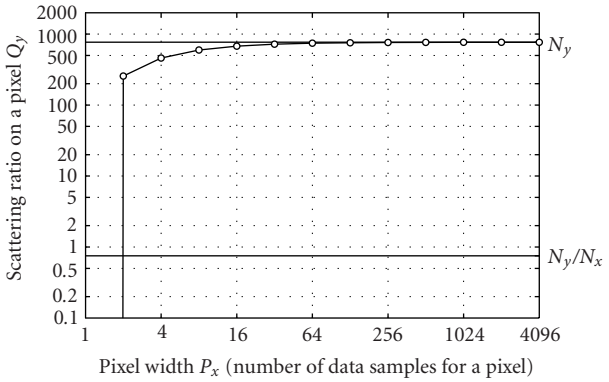


Figure 1.41. Representability diagram for a uniformly distributed random signal (length of the signal $2^{22} = 4194304$ bp, screen size 1024×768 pixels).

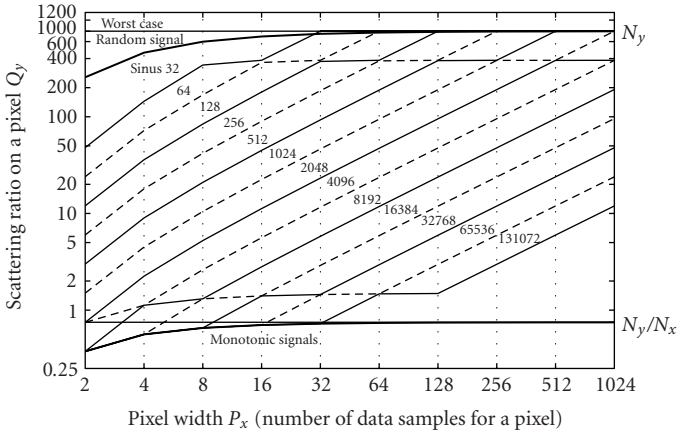


Figure 1.42. Representability diagram for sinus signals of various periods (length of signals 1048576 bp, screen size 1024×768 pixels).

1.6.4. Sine signal representability

To illustrate the behavior of nonmonotonic signals, in Figure 1.42 are given the representability characteristics of several sine functions with periods form 2^5 to 2^{17} samples. As expected, the sine signal behaves as a monotonic signal—the best case—when its period is larger than four times the width of the screen in number of samples, and as the worst case—when the period is lower than the width of the pixels. Two aliasing effects occur in the vicinity of the limiting cases, at levels of the average data scattering ratio equal to twice the best case and half the worst case, respectively. In-between these two levels, the average data scattering ratio varies almost linearly with respect to the pixel width.

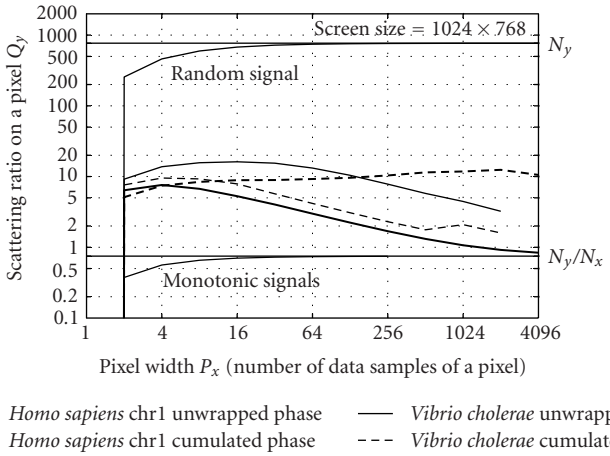


Figure 1.43. Representability diagram $\tilde{Q} = f(P_x)$ for the cumulated and unwrapped phase of the contig NT_004424 [2] of *Homo sapiens* chromosome 1 (length 6,311,978 bp [2, 3, 4]) and of the circular chromosome of *Yersinia pestis* (NC_0003143, length 4,653,728 bp [2, 18]).

1.6.5. Phase signals of genomic signals

Figure 1.43 shows the average data scattering ratio for 6,311,978 base pairs along contig of the first chromosome of *Homo sapiens* (NT 004424 [2]). The results are typical for many other prokaryote and eukaryote genomic signals. The screen size has been considered to be 1024×768 pixels. For the special case of one pixel per sample, for which the variation inside a pixel is zero, the scattering ratio cannot be represented on the logarithmic plot. This case corresponds to an error-free graphic, disregarding the smoothness of the resulting line. For pixels comprising two samples and up to about 16 samples, that is, for DNA segments comprising up to 16384 base pairs, both the cumulated and the unwrapped phase have the average data scattering ratio in the range 5–8, so that the data should be presented taking into account their dispersion. In most cases, this can be done by tracing a couple of lines showing the minimum and maximum values, respectively. When there are only several points apart from the others, the representation can be made by a line corresponding to the average value in a sliding window with the width of a pixel, accompanied by error bars. What is remarkable for the analysis of large scale DNA features is the fact than the average vertical scattering ratio of the signal for a pixel $\tilde{Q}$ becomes less than one, that is, the variation of the signal for the set of samples represented by a pixel becomes less than the pixel height, when the pixel width is larger than about 1450 samples. Obviously, the scale used to represent large scale features of genomic signals is much larger, up to hundreds of thousands of samples per pixel, so that the data can be represented adequately by a single line having the width of only one pixel.

The cumulated phase displays a relatively small variation and, when represented independently, remains with a rather significant dispersion of the samples that requires a presentation similar to the one used for statistical data.

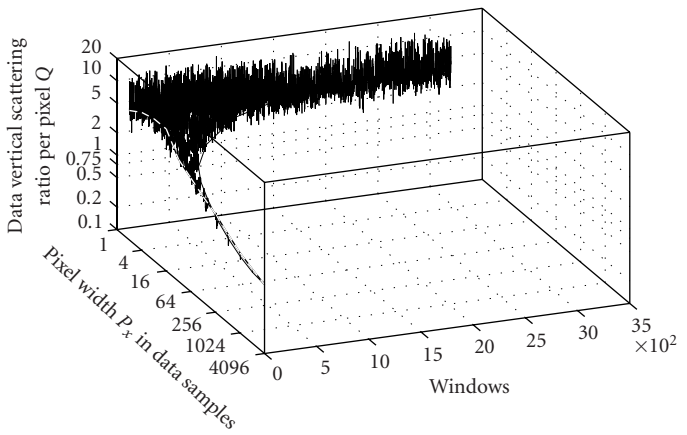


Figure 1.44. 3D representability diagram for the unwrapped phase of the contig NT_004424 [2] of *Homo sapiens* chromosome 1. The average curve in the plane $P_x - Q$ is the representability diagram shown in Figure 1.43.

Figure 1.44 gives a 3D representability diagram of the unwrapped phase of the *Homo sapiens* chromosome 1 contig NT_004424 shown in Figure 1.43. Both the average value of Q—the vertical scattering ratio of the signal on a pixel and the fluctuations of its value in the various windows decrease with the increase of the pixel width P_x.

1.7. Conclusions

This chapter presents results in the analysis of genomic information at the scale of whole chromosomes or whole genomes based on the conversion of genomic sequences into genomic signals, concentrating on the phase analysis.

The most conspicuous result is the linear variation displayed by the unwrapped phase almost along all chromosomes. This feature holds for all the investigated genomes, being shared by both prokaryotes and eukaryotes, while the magnitude and sign of the unwrapped phase slope are specific for each taxon and chromosome. Such a behavior proves a rule similar to Chargaff's rule, but reveals a statistical regularity of the *succession* of the nucleotides—a second-order statistics, not only of the *distribution* of nucleotides—a first order statistics.

This property is related to functions at the scale of whole chromosomes, such as replication, transcription, and crossover. The cumulated phase of the genomic signal of certain prokaryotes also shows a remarkable specific behavior. The comparison between the behavior of the cumulated phase and of the unwrapped phase across the putative origins and termini of the replichores suggests an interesting model for the structure of chromosomes.

The highly regular (linear) shape of the cumulated phase of reoriented ORFs strongly suggests a putative ancestral DNA longitudinal structure from which the current structures have evolved to satisfy restrictions resulting from various chromosome functions.

The analysis of data representability shows that the cumulated phase and the unwrapped phase can be represented adequately as simple graphic lines for very low and large scales, while for medium scales (thousands to tens of thousands of base pairs) statistical descriptions have to be used.

Bibliography

[1] The Genome Data Base, http://gdbwww.gdb.org/, Genome Browser, http://genome.ucsc.edu, European Informatics Institute, http://www.ebl.ac.uk, Ensembl, http://www.ensembl.org.
[2] National Center for Biotechnology Information, National Institutes of Health, National Library of Medicine, http://www.ncbi.nlm.nih.gov/genoms/, ftp://ftp.ncbi.nlm.nih.gov/genoms/, GenBank, http://www.ncbi.nlm.nih.gov/Genbank/index.html.
[3] International Human Genome Sequencing Consortium, "Initial sequencing and analysis of the human genome," *Nature*, vol. 409, no. 6822, pp. 860–921, 2001.
[4] J. C. Venter, M. D. Adams, E. W. Myers, et al., "The sequence of the human genome," *Science*, vol. 291, no. 5507, pp. 1304–1351, 2001.
[5] Y. Kawarabayasi, Y. Hino, H. Horikawa, et al., "Complete genome sequence of an aerobic hyperthermophilic crenarchaeon, Aeropyrum pernix K1," *DNA Res.*, vol. 6, no. 2, pp. 83–101, 1999.
[6] RIKEN Genome Exploration Research Group Phase II Team and the FANTOM Consortium, "Functional annotation of a full-length mouse cDNA collection," *Nature*, vol. 409, no. 6821, pp. 685–690, 2001.
[7] Rat Genome Sequencing Consortium, http://www.ncbi.nlm.nih.gov/genoms, 30 August, 2003.
[8] Genome Sequencing Center, Chicken genome, Washington University Medical School, 1 March 2004, http://www.genome.wustl.edu/projects/chicken/.
[9] The C. elegans Sequencing Consortium, "Genome sequence of the nematode C. elegans: a platform for investigating biology," *Science*, vol. 282, no. 5396, pp. 2012–2018, 1998.
[10] A. Theologis, J. R. Ecker, C. J. Palm, et al., "Sequence and analysis of chromosome 1 of the plant Arabidopsis thaliana," *Nature*, vol. 408, no. 6814, pp. 816–820, 2000.
[11] R. A. Alm, L. S. Ling, D. T. Moir, et al., "Genomic-sequence comparison of two unrelated isolates of the human gastric pathogen Helicobacter pylori," *Nature*, vol. 397, no. 6715, pp. 176–180, 1999.
[12] K. Aoki, A. Oguchi, Y. Nagai, et al., Sequence of *Staphylococcus aureus* (strain MW2), direct submission to GenBank," 6 March 2002, National Institute of Technology and Evaluation, Biotechnology Center, Tokyo, Japan, http://www.bio.nite.go.jp/.
[13] T. Baba, F. Takeuchi, M. Kuroda, et al., "Genome and virulence determinants of high virulence community-acquired MRSA," *Lancet*, vol. 359, no. 9320, pp. 1819–1827, 2002.
[14] F. R. Blattner, G. Plunkett III, C. A. Bloch, et al., "The complete genome sequence of Escherichia coli K-12," *Science*, vol. 277, no. 5331, pp. 1453–1474, 1997.
[15] J. Kawai, A. Shinagawa, K. Shibata, et al., "Functional annotation of a full-length mouse cDNA collection," *Nature*, vol. 409, no. 6821, pp. 685–690, 2001.
[16] F. Kunst, N. Ogasawara, I. Moszer, et al., "The complete genome sequence of the gram-positive bacterium Bacillus subtilis," *Nature*, vol. 390, no. 6657, pp. 249–256, 1997.
[17] J. R. Lobry, "Asymmetric substitution patterns in the two DNA strands of bacteria," *Mol. Biol. Evol.*, vol. 13, no. 5, pp. 660–665, 1996.
[18] J. Parkhill, B. W. Wren, N. R. Thomson, et al., "Genome sequence of *Yersinia pestis*, the causative agent of plague," *Nature*, vol. 413, no. 6855, pp. 523–527, 2001.
[19] T. Shimizu, K. Ohtani, H. Hirakawa, et al., "Complete genome sequence of Clostridium perfringens, an anaerobic flesh-eater," *Proc. Natl. Acad. Sci. USA*, vol. 99, no. 2, pp. 996–1001, 2002.
[20] J. M. Claverie, "Computational methods for the identification of genes in vertebrate genomic sequences," *Hum. Mol. Genet.*, vol. 6, no. 10, pp. 1735–1744, 1997.
[21] W. F. Doolittle, "Phylogenetic classification and the universal tree," *Science*, vol. 284, no. 5423, pp. 2124–2128, 1999.
[22] R. Durbin, S. Eddy, A. Krogh, and G. Mitchison, *Biological Sequence Analysis. Probabilistic Models of Proteins and Nucleic Acids*, Cambridge University Press, Cambridge, UK, 1998.

[23] P. D. Cristea, "Conversion of nucleotides sequences into genomic signals," *J. Cell. Mol. Med.*, vol. 6, no. 2, pp. 279–303, 2002.
[24] P. D. Cristea, "Genetic signal representation and analysis," in *SPIE Conference, International Biomedical Optics Symposium, Molecular Analysis and Informatics (BIOS '02)*, vol. 4623 of *Proceedings of SPIE*, pp. 77–84, San Jose, Calif, USA, January 2002.
[25] P. D. Cristea, "Genomic signals of chromosomes and of concatenated reoriented coding regions," in *SPIE Conference, Biomedical Optics (BIOS '04)*, vol. 5322 of *Proceedings of SPIE*, pp. 29–41, San Jose, Calif, USA, January 2004, Progress in Biomedical Optics and Imaging, Vol. 5, No. 11.
[26] P. D. Cristea, "Representability of genomic signals," in *Proc. 26th Annual International Conference of the IEEE Engineering in Medicine and Biology Society*, San Francisco, Calif, USA, September 2004.
[27] P. D. Cristea, "Genomic signals of reoriented ORFs," *EURASIP J. Appl. Signal Process.*, vol. 2004, no. 1, pp. 132–137, 2004, Special issue on genomic signal processing.
[28] P. D. Cristea, "Large scale features in DNA genomic signals," *Signal Process.*, vol. 83, no. 4, pp. 871–888, 2003, Special issue on genomic signal processing.
[29] P. D. Cristea, "Analysis of chromosome genomic signals," in *7th International Symposium on Signal Processing and Its Applications (ISSPA '03)*, pp. 49–52, Paris, France, July 2003.
[30] P. D. Cristea and G. A. Popescu, "Fractal dimension of human chromosome 22," in *The 1st South-East European Symposium on Interdisciplinary Approaches in Fractal Analysis (IAFA '03)*, pp. 131–134, Bucharest, Romania, May 2003.
[31] P. D. Cristea, "Genomic signals for whole chromosomes," in *SPIE Conference, International Biomedical Optics Symposium, Molecular Analysis and Informatics (BIOS '03)*, vol. 4962 of *Proceedings of SPIE*, pp. 194–205, San Jose, Calif, USA, January 2003.
[32] P. D. Cristea, "Large scale features in prokaryote and eukaryote genomic signals," in *9th International Workshop on Systems, Signals and Image Processing (IWSSIP '02)*, Manchester, UK, November 2002.
[33] E. Chargaff, "Structure and function of nucleic acids as cell constituents," *Fed. Proc.*, vol. 10, no. 3, pp. 654–659, 1951.
[34] J. D. Watson and F. H. C. Crick, "A structure for deoxyribose nucleic acid," *Nature*, vol. 171, no. 4356, pp. 737–738, 1953.
[35] D. Anastassiou, "Frequency-domain analysis of biomolecular sequences," *Bioinformatics*, vol. 16, no. 12, pp. 1073–1081, 2000.
[36] D. R. Forsdyke, "Sense in antisense?," *J. Mol. Evol.*, vol. 41, no. 5, pp. 582–586, 1995.
[37] J. M. Freeman, T. N. Plasterer, T. F. Smith, and S. C. Mohr, "Patterns of genome organization in bacteria," *Science*, vol. 279, no. 5358, pp. 1827–1830, 1998.
[38] A. Grigoriev, "Analyzing genomes with cumulative skew diagrams," *Nucleic Acids Res.*, vol. 26, no. 10, pp. 2286–2290, 1998.
[39] J. O. Andersson, W. F. Doolittle, and C. L. Nesbø, "Genomics. Are there bugs in our genome?," *Science*, vol. 292, no. 5523, pp. 1848–1850, 2001.
[40] H. Gee, "A journey into the genome: what's there," *Nature Science Update*, February 2001, http://www.nature.com/news/2001/010215/full/010215-3.html.
[41] M. Kuroda, T. Ohta, I. Uchiyama, et al., "Whole genome sequencing of meticillin-resistant Staphylococcus aureus," *Lancet*, vol. 357, no. 9264, pp. 1225–1240, 2001.

Paul Dan Cristea: Biomedical Engineering Center, University Politehnica of Bucharest, 313 Splaiul Independentei, 060042 Bucharest, Romania

Email: pcristea@ieee.org